ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευφυείς Μέθοδοι Σύνοψης Πατεντών και Μελέτη Εφαρμογής σε Εργασίες Αναζήτησης και Ταξινόμησης

Του φοιτητή
Καλκαντέρα Λεωνίδα
Αρ. Μητρώου: 12/2024

Επιβλέπων
Σαλαμπάσης Μιχάλης
Βαθμίδα Καθηγητή

Ημερομηνία 20/01/2026

Τίτλος Δ.Ε.: Ευφυείς Μέθοδοι Σύνοψης Πατεντών και Μελετη εφαρμογης τους σε Εργασίες Αναζήτησης και Ταξινόμησης

Κωδικός Δ.Ε.: 25287

Ονοματεπώνυμο φοιτητή: Καλκαντέρας Λεωνίδας

Ονοματεπώνυμο εισηγητή: Σαλαμπάσης Μιχάλης

Ημερομηνία ανάληψης Δ.Ε.: 14/07/2025

Ημερομηνία περάτωσης Δ.Ε.: 20/01/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Knowledge isn't free. You have to pay attention.»

— Richard P. Feynman

This page is intentionally left blank.

Πρόλογος

Η επιλογή της συγκεκριμένης διπλωματικής εργασίας προέκυψε από την ανάγκη μελέτης και αξιοποίησης σύγχρονων ευφυών μεθόδων επεξεργασίας της φυσικής γλώσσας σε μεγάλης κλίμακας και ιδιαίτερα απαιτητικά σύνολα κειμένου, όπως είναι τα έγγραφα πατεντών. Οι πατέντες αποτελούν μία από τις σημαντικότερες πηγές τεχνολογικής και επιστημονικής πληροφορίας, ωστόσο ο όγκος, η πολυπλοκότητα και η εκτενής δομή τους καθιστούν δύσκολη την αποτελεσματική ανάγνωση, ανάλυση και αναζήτηση της πληροφορίας που περιέχουν.

Στο πλαίσιο αυτό, η αυτόματη σύνοψη πατεντών αναδεικνύεται ως κρίσιμο εργαλείο για τη μείωση της γνωστικής επιβάρυνσης του χρήστη και τη διευκόλυνση διαδικασιών όπως η αναζήτηση, η σύγκριση και η ταξινόμηση τεχνικών εγγράφων. Παράλληλα, η ραγδαία εξέλιξη των γλωσσικών μοντέλων και των νευρωνικών μεθόδων ανάκτησης πληροφορίας δημιουργεί νέες δυνατότητες για πιο αποδοτικά και ακριβή συστήματα επεξεργασίας πατεντών.

Η εργασία αυτή φιλοδοξεί να συμβάλει στην κατανόηση των δυνατοτήτων και των περιορισμών των σύγχρονων προσεγγίσεων σύνοψης, αναδεικνύοντας τη σημασία της ποιότητας και της δομής της πληροφορίας ως ενδιάμεσου βήματος ανάμεσα στον άνθρωπο και τα συστήματα αναζήτησης. Μέσα από αυτή τη σκοπιά, η σύνοψη αντιμετωπίζεται όχι μόνο ως εργαλείο συμπύκνωσης κειμένου, αλλά και ως μέσο υποστήριξης της αποτελεσματικής πρόσβασης στη γνώση.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται την εφαρμογή ευφυών μεθόδων σύνοψης πατεντών και τη διερεύνηση της συμβολής τους σε εργασίες αναζήτησης και ταξινόμησης τεχνικών εγγράφων. Οι πατέντες αποτελούν ιδιαίτερα εκτενή και σύνθετα κείμενα, γεγονός που δυσχεραίνει την αποτελεσματική επεξεργασία και αξιοποίηση της πληροφορίας που περιέχουν. Στο πλαίσιο αυτό, η αυτόματη σύνοψη αναδεικνύεται ως κρίσιμο εργαλείο τόσο για τη μείωση της πολυπλοκότητας και του όγκου του κειμένου όσο και για τη βελτίωση της πρόσβασης στη γνώση.

Αξιοποιήθηκε σύνολο 2.592 πατεντών (XML), το οποίο προεπεξεργάστηκε και μετατράπηκε σε δομημένα αρχεία SGML, ώστε να είναι εφικτή η συστηματική χρήση επιμέρους πεδίων (π.χ. TITLE, ABSTRACT, DESCRIPTION, CLAIMS) στις διαδικασίες σύνοψης και αναζήτησης. Εφαρμόστηκαν τέσσερις κατηγορίες σύνοψης (extractive, abstractive, instructive και hybrid) και υλοποιήθηκε πλήρες πειραματικό pipeline (σύνοψη, ευρετηρίαση/ανάκτηση, εξαγωγή μετρικών, αποθήκευση/οπτικοποίηση), με μεγάλο αριθμό εκτελέσεων και καταγραφή παραμέτρων, χρόνων και παραγόμενων αρχείων.

Η κλίμακα της αξιολόγησης αντικατοπτρίζεται στον αριθμό των παραγόμενων αρχείων αποτελεσμάτων. Για τα τελικά σενάρια που διατηρήθηκαν στην παρούσα εργασία εκτελέστηκαν **319** πειράματα ανάκτησης στους δύο τελικούς δείκτες, με σταθερή παραγωγή τεσσάρων αρχείων αποτελεσμάτων ανά πείραμα (συνολικά **1.276 αρχεία**). Παράλληλα, για τις **18 μεθόδους** σύνοψης παρήχθησαν πέντε αρχεία ανά μέθοδο (συνολικά **90 αρχεία**). Συνεπώς, το τελικό σύνολο των παραγόμενων αρχείων ανέρχεται σε **1.366 αρχεία**, χωρίς να συνυπολογίζονται ενδιάμεσα outputs και βοηθητικά logs. Σημειώνεται ότι οι συνολικές εκτελέσεις και τα παραγόμενα αρχεία στο σύνολο του έργου ήταν περισσότερα από όσα τελικά παρουσιάζονται, καθώς για την εργασία διατηρήθηκαν και αναλύθηκαν τα κυριότερα/αντιπροσωπευτικότερα σενάρια.

Για τα υπολογιστικά απαιτητικά στάδια του pipeline, και ειδικότερα για τη νευρωνική σύνοψη και τη νευρωνική ανάκτηση, αξιοποιήθηκε επιτάχυνση GPU (A100) όπου ήταν διαθέσιμη, ώστε να καταστεί εφικτή η εκτέλεση μεγάλου αριθμού runs σε πρακτικούς χρόνους. Ενδεικτικά, μόνο για τα πειράματα που διατηρήθηκαν ως τελικά στην παρούσα εργασία απαιτήθηκαν περίπου **225 ώρες GPU** για τις εκτελέσεις ανάκτησης και περίπου **110 ώρες GPU** για την παραγωγή περιλήψεων, ενώ επιπλέον GPU χρόνος διατέθηκε για τη δημιουργία των αντίστοιχων δεικτών (***indexing***) ανάλογα με τη μέθοδο ευρετηρίασης. Η αξιοποίηση GPU υψηλών επιδόσεων (π.χ. A100) δεν είναι μόνο τεχνικό ζήτημα, αλλά συνεπάγεται και *οικονομικό κόστος* κατά την κλιμάκωση των πειραμάτων. Σε στάδια όπου η επιτάχυνση GPU δεν ήταν αναγκαία, η εκτέλεση έγινε σε CPU με παράλληλη επεξεργασία (8–12 cores).

Για την αξιολόγηση της επίδρασης της σύνοψης στην αναζήτηση, υλοποιήθηκαν δύο διαφορετικές προσεγγίσεις ανάκτησης: (α) κλασική στατιστική ανάκτηση βασισμένη σε BM25 (Pyserini) και (β) νευρωνική ανάκτηση βασισμένη σε ColBERT. Σημειώνεται ότι, παρότι το fine-tuning αναμένεται να βελτιώσει τη νευρωνική ανάκτηση, δεν εφαρμόστηκε λόγω περιορισμένης διαθεσιμότητας επισημασμένων δεδομένων, και ο ColBERT αξιολογήθηκε zero-shot (χωρίς πρόσθετη εκπαίδευση). Η αξιολόγηση βασίστηκε στις μετρικές MAP και Recall@100, αποτυπώνοντας ποιότητα κατάταξης και κάλυψη στα πρώτα 100 αποτελέσματα. Παράλληλα, εξετάστηκε το trade-off συμπίεσης–πιστότητας–απόδοσης σε σχέση με το υπολογιστικό κόστος (χρόνος σύνοψης/ανάκτησης). Ενδεικτικά, στα τελικά σενάρια δύο από τις υψηλότερες επιδόσεις (χωρίς συνεκτίμηση χρόνου/κόστους εκτέλεσης) προέκυψαν με instructive σύνοψη Qwen2_14B_Instruct (MAP=0,4999, Recall@100=0,8218) και Qwen2_7B_Instruct (MAP=0,5056, Recall@100=0,8198), σε 2.592 queries.

Τα αποτελέσματα δείχνουν ότι η σύνοψη δεν επηρεάζει την ανάκτηση με ενιαίο τρόπο, αλλά η ωφέλειά της εξαρτάται από το (i) ποια πεδία χρησιμοποιούνται ως query και (ii) ποιος μηχανισμός ανάκτησης εφαρμόζεται. Σε σενάρια όπου το query εμπλουτίζεται με δομικά πεδία υψηλής πληροφορίας (όπως TITLE και ABSTRACT) παρατηρείται συστηματική βελτίωση τόσο στο MAP όσο και στο Recall@100 σε σχέση με τη χρήση μεμονωμένων, πιο "βαριών" πεδίων (π.χ. μόνο DESCRIPTION ή μόνο CLAIMS), γεγονός που υποδηλώνει ότι η κατάλληλη σύνθεση query fields είναι κρίσιμη για την αξιοποίηση των περιλήψεων. Παράλληλα, σε συγκρίσεις "ORIGINAL έναντι σύνοψης" με αυστηρά αντιστοιχισμένες συνθήκες (ίδια query fields και ίδιες ρυθμίσεις ανάκτησης), οι υβριδικές και instructive προσεγγίσεις τείνουν να εμφανίζουν πιο σταθερά θετική επίδραση στον BM25, ενώ στον ColBERT οι μεταβολές είναι γενικά πιο συγκρατημένες και συχνά εκφράζονται εντονότερα ως βελτίωση κάλυψης (Recall@100) παρά ως βελτίωση κατάταξης (MAP). Συνολικά, η εργασία τεκμηριώνει ότι οι περιλήψεις μπορούν να ενισχύσουν την αναζήτηση πατεντών, όταν εντάσσονται σε συνεπή πειραματική ρύθμιση (κατάλληλος σχεδιασμός queries, δίκαιες συγκρίσεις baseline–summarized και παρακολούθηση κόστους), αναδεικνύοντας παράλληλα τη σημασία της συστηματικής καταγραφής και διαχείρισης πολλών πειραματικών εκτελέσεων.

«Intelligent Methods for Patent Summarization and a Study of Their Application to Information Retrieval and Ranking Tasks»

«Leonidas Kalkanteras»

Abstract

This thesis addresses the application of intelligent patent summarization methods and investigates their contribution to technical document retrieval and classification tasks. Patents constitute particularly extensive and complex texts, which hinders the effective processing and utilization of the information they contain. In this context, automatic summarization emerges as a critical tool for both reducing text complexity and volume, and improving access to knowledge.

A dataset of 2,592 patents (XML) was utilized, preprocessed, and converted into structured SGML files to enable the systematic use of individual fields (e.g., TITLE, ABSTRACT, DESCRIPTION, CLAIMS) in summarization and retrieval processes. Four summarization categories were applied (extractive, abstractive, instructive, and hybrid), and a complete experimental pipeline was implemented (summarization, indexing/retrieval, metric extraction, storage/visualization), involving a large number of runs and logging parameters, times, and generated files.

The scale of the evaluation is reflected in the number of generated result files. For the final scenarios retained in this work, 319 retrieval experiments were executed on the two final indexes, consistently producing four result files per experiment (totaling 1,276 files). Concurrently, for the 18 summarization methods, five files were produced per method (totaling 90 files). Consequently, the final set of generated files amounts to 1,366 files, excluding intermediate outputs, auxiliary logs, and exploratory runs. It is noted that the total runs and generated files throughout the project exceeded those presented here, as only the most representative/significant scenarios were retained and analyzed.

For the computationally demanding pipeline stages, specifically neural summarization and neural retrieval, GPU acceleration (A100) was utilized where available, to enable the execution of a large number of runs within practical timeframes. Indicatively, for the experiments retained as final in this work alone, approximately 225 GPU hours were required for retrieval runs and about 110 GPU hours for summary generation, while additional GPU time was required for index creation, varying by indexing method.

The utilization of high-performance GPUs (e.g., A100) is not merely a technical issue but also incurs real monetary cost during experiment scaling. In stages where GPU acceleration was not necessary, execution was performed on CPUs using parallel processing (8–12 cores).

To evaluate the impact of summarization on search, two different retrieval approaches were implemented: (a) classical statistical retrieval based on BM25 (Pyserini) and (b) neural retrieval based on ColBERT. It is noted that, although fine-tuning is expected to improve neural retrieval, it was not applied due to limited availability of labeled data, and ColBERT was evaluated in a zero-shot setting (without additional training). The evaluation relied on MAP and Recall@100 metrics, capturing ranking quality and coverage within the top 100 results. Concurrently, the compression–fidelity–performance trade-off was examined in relation to computational cost (summarization/retrieval time). Indicatively, in the final

scenarios, two of the highest performances (without factoring in execution time/cost) were achieved with instructive summarization using Qwen2_14B_Instruct (MAP=0.4999, Recall@100=0.8218) and Qwen2_7B_Instruct (MAP=0.5056, Recall@100=0.8198), across 2,592 queries.

The results indicate that summarization does not affect retrieval uniformly; its benefit depends on (i) which fields are used as queries and (ii) which retrieval mechanism is applied. In scenarios where the query is enriched with high-information structural fields (such as TITLE and ABSTRACT), a systematic improvement is observed in both MAP and Recall@100 compared to the use of isolated, "heavier" fields (e.g., only DESCRIPTION or only CLAIMS), suggesting that the appropriate composition of query fields is critical for leveraging summaries. Furthermore, in "ORIGINAL vs. summary" comparisons under strictly matched conditions (identical query fields and retrieval settings), hybrid and instructive approaches tend to exhibit a more consistently positive effect on BM25, whereas in ColBERT, changes are generally more moderate and are often expressed more strongly as improved coverage (Recall@100) rather than improved ranking (MAP). Overall, the thesis documents that summaries can enhance patent search when integrated into a consistent experimental setting (appropriate query design, fair baseline–summarized comparisons, and cost monitoring), while also highlighting the importance of systematically logging and managing multiple experimental runs.

Keywords: Patent Summarization, Information Retrieval, Classification, BM25, ColBERT, Late Interaction, Query Enrichment, Tag-aware Summarization, MAP and Recall

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στους γονείς μου για τη στήριξη, και την ενθάρρυνση που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου. Ευχαριστώ επίσης τον αδελφό μου για τη συμπαράσταση και την κατανόηση που έδειξε σε αυτή τη διαδρομή. Και οπωσδήποτε ένα πολύ μεγάλο ευχαριστώ στην αγαπημένη μου φίλη ΣΔ για την επιμονή της να παρακολουθήσω αυτό το μεταπτυχιακό…

Ιδιαίτερες ευχαριστίες οφείλω στον επιβλέποντα καθηγητή μου κ. Μιχάλη Σαλαμπάση και στην υποψήφια διδάκτορα κα Ελένη Καματέρη για την εμπιστοσύνη που μου έδειξαν, δίνοντάς μου την ευκαιρία να εκπονήσω την παρούσα διπλωματική εργασία καθώς και για τη συνεργασία και τις πολύτιμες συμβουλές τους κατά την υλοποίησή της.

Περιεχόμενα

Κατάλογος Απεικονίσεων

Κατάλογος Πινάκων

Κεφάλαιο 1ο: Εισαγωγή

1.1 Αντικείμενο και κίνητρο της εργασίας

Η ραγδαία αύξηση του όγκου της διαθέσιμης ψηφιακής πληροφορίας έχει αναδείξει την ανάγκη για την ανάπτυξη αποδοτικών μεθόδων επεξεργασίας, αναζήτησης και αξιοποίησης μεγάλων κειμενικών συνόλων. Στο πλαίσιο αυτό, η επεξεργασία φυσικής γλώσσας και οι ευφυείς μέθοδοι ανάλυσης κειμένου διαδραματίζουν καθοριστικό ρόλο στη διευκόλυνση της πρόσβασης στη γνώση.

Τα έγγραφα πατεντών αποτελούν ένα χαρακτηριστικό παράδειγμα τέτοιων απαιτητικών κειμενικών συνόλων. Πρόκειται για εκτενή, δομημένα και συχνά ιδιαίτερα τεχνικά κείμενα, τα οποία περιγράφουν λεπτομερώς νέες εφευρέσεις και τεχνολογικές λύσεις. Η αποτελεσματική μελέτη και αξιοποίησή τους απαιτεί σημαντικό χρόνο και εξειδίκευση, γεγονός που καθιστά αναγκαία την υποστήριξη της διαδικασίας αυτής μέσω αυτοματοποιημένων εργαλείων.

Κίνητρο της παρούσας εργασίας αποτελεί η διερεύνηση του τρόπου με τον οποίο οι σύγχρονες ευφυείς μέθοδοι σύνοψης μπορούν να συμβάλουν στη βελτίωση της κατανόησης και της αξιοποίησης των πατεντών, καθώς και στη διευκόλυνση εργασιών αναζήτησης και ταξινόμησης τεχνικών εγγράφων.

1.2 Πατέντες ως πηγή πληροφορίας

Οι πατέντες αποτελούν μία από τις σημαντικότερες και πιο αξιόπιστες πηγές τεχνολογικής και επιστημονικής πληροφορίας. Κάθε έγγραφο πατέντας περιλαμβάνει λεπτομερή περιγραφή της εφεύρεσης, το τεχνικό υπόβαθρο, τις αξιώσεις προστασίας και συχνά αναφορές σε προ υπάρχουσες τεχνολογίες.

Σε αντίθεση με άλλες μορφές επιστημονικών δημοσιεύσεων, οι πατέντες χαρακτηρίζονται από αυστηρή δομή και νομική ακρίβεια, στοιχεία που εξασφαλίζουν τη σαφήνεια αλλά ταυτόχρονα αυξάνουν την πολυπλοκότητα του κειμένου. Ο μεγάλος αριθμός πατεντών που δημοσιεύονται σε παγκόσμιο επίπεδο καθιστά πρακτικά αδύνατη τη χειροκίνητη μελέτη και σύγκριση τους σε βάθος.[1]

Ως αποτέλεσμα, οι πατέντες αποτελούν ένα ιδανικό αλλά απαιτητικό πεδίο εφαρμογής για τεχνικές επεξεργασίας φυσικής γλώσσας, σύνοψης και ανάκτησης πληροφορίας.

1.3 Πρόβλημα σύνοψης και αναζήτησης πατεντών

Η αναζήτηση πληροφορίας σε συλλογές πατεντών παρουσιάζει ιδιαίτερες δυσκολίες, οι οποίες οφείλονται τόσο στον όγκο όσο και στη φύση του περιεχομένου τους. Οι χρήστες καλούνται συχνά να εντοπίσουν σχετικές πατέντες μέσα από χιλιάδες έγγραφα, τα οποία περιέχουν εκτενείς τεχνικές περιγραφές και εξειδικευμένη ορολογία.[2]

Η αυτόματη σύνοψη πατεντών έρχεται να αντιμετωπίσει αυτό το πρόβλημα, προσφέροντας συμπυκνωμένες αναπαραστάσεις του περιεχομένου τους. Μέσω της σύνοψης, καθίσταται δυνατή η ταχύτερη κατανόηση της βασικής ιδέας μιας πατέντας, η σύγκριση πολλαπλών εγγράφων και η υποστήριξη διαδικασιών αναζήτησης και ταξινόμησης.

Ωστόσο, η σύνοψη πατεντών αποτελεί ένα ιδιαίτερα απαιτητικό πρόβλημα, καθώς απαιτεί τη διατήρηση της τεχνικής ακρίβειας και της σημασιολογικής πληρότητας, χωρίς την αλλοίωση του αρχικού νοήματος.

1.4 Στόχοι και συνεισφορές της διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη και αξιολόγηση σύγχρονων ευφυών μεθόδων σύνοψης πατεντών και η διερεύνηση της επίδρασής τους σε εργασίες αναζήτησης και ταξινόμησης τεχνικών εγγράφων.

Οι βασικές συνεισφορές της εργασίας συνοψίζονται στα εξής:

- η εφαρμογή και σύγκριση διαφορετικών κατηγοριών μεθόδων σύνοψης πατεντών,
- η αξιολόγηση της χρησιμότητας των παραγόμενων περιλήψεων σε συστήματα αναζήτησης,
- η διερεύνηση διαφορετικών προσεγγίσεων ευρετηρίασης και ανάκτησης πληροφορίας,
- η ανάπτυξη συστήματος διαχείρισης και οπτικοποίησης των πειραματικών αποτελεσμάτων.

1.5 Δομή της εργασίας

Η εργασία οργανώνεται σε οκτώ κεφάλαια. Στο Κεφάλαιο 1 παρουσιάζεται το αντικείμενο, το κίνητρο και οι στόχοι της εργασίας. Στο Κεφάλαιο 2 αναλύεται το θεωρητικό υπόβαθρο και οι βασικές έννοιες που σχετίζονται με τη σύνοψη κειμένου και την ανάκτηση πληροφορίας. Το Κεφάλαιο 3 περιγράφει τα δεδομένα που χρησιμοποιήθηκαν και τη διαδικασία προεπεξεργασίας τους. Στο Κεφάλαιο 4 παρουσιάζονται οι μέθοδοι σύνοψης πατεντών που εφαρμόστηκαν, ενώ στο Κεφάλαιο 5 αναλύονται οι τεχνικές ευρετηρίασης και αναζήτησης. Το Κεφάλαιο 6 περιγράφει το σύστημα διαχείρισης και οπτικοποίησης των αποτελεσμάτων. Στο Κεφάλαιο 7 παρουσιάζονται τα πειραματικά αποτελέσματα και η αξιολόγησή τους, ενώ το Κεφάλαιο 8 συνοψίζει τα συμπεράσματα της εργασίας και προτείνει κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2ο: Σχετικό Υπόβαθρο και Θεωρία

2.1 Ανάκτηση Πληροφορίας σε μεγάλα κειμενικά σύνολα

Η ανάκτηση πληροφορίας (Information Retrieval – IR) αποτελεί έναν από τους βασικούς τομείς της επιστήμης της πληροφορικής και ασχολείται με τον εντοπισμό και την επιστροφή σχετικών εγγράφων από μεγάλες συλλογές κειμένου, με βάση τις ανάγκες ενός χρήστη. Σε αντίθεση με την κλασική "αναζήτηση" σε μικρά σύνολα δεδομένων, τα σύγχρονα συστήματα IR καλούνται να λειτουργούν σε συνθήκες μεγάλης κλίμακας, ποικιλίας μορφών κειμένου, και υψηλών απαιτήσεων ακρίβειας, καθώς τα αποτελέσματα πρέπει όχι μόνο να εντοπίζονται αλλά και να κατατάσσονται σωστά ως προς τη συνάφεια. [1]

2.1.1 Βασικές έννοιες: έγγραφα, ερωτήματα και κατάταξη

Σε ένα σύστημα IR, το "έγγραφο" αντιπροσωπεύει μια μονάδα αναζήτησης (π.χ. πατέντα, άρθρο, αναφορά), ενώ το "ερώτημα" (query) εκφράζει την πληροφοριακή ανάγκη του χρήστη. Το κεντρικό πρόβλημα δεν είναι απλώς η εύρεση εγγράφων που περιέχουν λέξεις του ερωτήματος, αλλά η εκτίμηση του βαθμού συνάφειας (relevance) και η παραγωγή μιας κατάταξης (ranking), όπου τα πιο σχετικά αποτελέσματα εμφανίζονται πρώτα. Στα μεγάλα κειμενικά σύνολα, οι λίστες αποτελεσμάτων είναι συνήθως πολύ μεγάλες, συνεπώς η ποιότητα του ranking στα πρώτα k αποτελέσματα είναι καθοριστική.

2.1.2 Ιδιαιτερότητες μεγάλων συλλογών και γιατί τα "keywords" δεν αρκούν

Σε μεγάλες συλλογές, όπως πατέντες, επιστημονικά άρθρα ή νομικά κείμενα, εμφανίζονται συστηματικά φαινόμενα που δυσκολεύουν την ανάκτηση με απλή λεξιλογική αντιστοίχιση. Ενδεικτικά:

- συνωνυμία: διαφορετικές λέξεις/φράσεις περιγράφουν την ίδια έννοια , ,
- πολυσημία: η ίδια λέξη μπορεί να έχει διαφορετική σημασία , ,
- τεχνική ορολογία: εξειδικευμένοι όροι, ακρωνύμια και παραλλαγές ,
- μεγάλο μήκος εγγράφων: η σχετική πληροφορία είναι διάσπαρτη και όχι συγκεντρωμένη.

Στην περίπτωση των πατεντών, η δυσκολία ενισχύεται από την τυποποιημένη γλώσσα, την εκτενή ανάπτυξη περιγραφών/αξιώσεων και την τάση χρήσης διαφορετικών διατυπώσεων για την ίδια τεχνική ιδέα.

2.1.3 Μέτρα αξιολόγησης στην ανάκτηση πληροφορίας

Η αξιολόγηση των συστημάτων IR βασίζεται σε μετρικές που συνδέονται με την έννοια της συνάφειας, συνήθως μέσω ενός "ground truth" (π.χ. qrels) που δηλώνει ποια έγγραφα θεωρούνται σχετικά για κάθε ερώτημα. Κλασικές μετρικές είναι:

- Precision@k: ποσοστό σχετικών εγγράφων στα πρώτα k αποτελέσματα , ,
- Recall@k: ποσοστό των συνολικά σχετικών εγγράφων που ανακτήθηκαν στα πρώτα k , ,
- MAP: μέσος όρος της precision σε σημεία ανάκτησης σχετικών εγγράφων ,
- nDCG@k: μετρική που λαμβάνει υπόψη τη θέση των σχετικών εγγράφων και την "ποιότητα" της κατάταξης.

Στα συστήματα πατεντών, όπου συχνά ενδιαφέρει η μέγιστη ανάκληση (recall) χωρίς σημαντική πτώση στην ακρίβεια, η ανάλυση πολλαπλών μετρικών είναι απαραίτητη για ισορροπημένα συμπεράσματα. [3], [4]

2.2 Σύνοψη κειμένου: extractive, abstractive, instructive και hybrid προσεγγίσεις

Η σύνοψη κειμένου αποτελεί βασικό αντικείμενο της επεξεργασίας φυσικής γλώσσας και στοχεύει στη δημιουργία σύντομων αναπαραστάσεων μεγάλων κειμένων, διατηρώντας την ουσία του αρχικού περιεχομένου. Στο πλαίσιο της παρούσας εργασίας, η σύνοψη αντιμετωπίζεται όχι μόνο ως εργαλείο "συμπίεσης", αλλά και ως μέσο που μπορεί να επηρεάσει άμεσα την απόδοση της ανάκτησης πληροφορίας, ειδικά όταν τα αρχικά έγγραφα είναι μεγάλα και η κρίσιμη πληροφορία είναι διάσπαρτη. [5], [6]

2.2.1 Extractive σύνοψη

Οι extractive μέθοδοι βασίζονται στην επιλογή και συνδυασμό τμημάτων του αρχικού κειμένου, συνήθως σε επίπεδο προτάσεων ή αποσπασμάτων. Η επιλογή μπορεί να στηρίζεται σε στατιστικά κριτήρια (π.χ. σημαντικότητα όρων) ή σε σημασιολογικά κριτήρια (π.χ. ομοιότητα προτάσεων με κεντρική αναπαράσταση του κειμένου). Το βασικό πλεονέκτημα είναι η υψηλή πιστότητα, καθώς η σύνοψη αποτελείται από αυτούσιες φράσεις του κειμένου και άρα ελαχιστοποιεί τον κίνδυνο αλλοίωσης τεχνικών/νομικών λεπτομερειών. Ωστόσο, επειδή τα αποσπάσματα προέρχονται από διαφορετικά σημεία, το αποτέλεσμα μπορεί να εμφανίζει μειωμένη συνοχή ή επαναλήψεις.[7]

2.2.2 Abstractive σύνοψη

Οι abstractive μέθοδοι παράγουν νέο κείμενο, συμπυκνώνοντας το νόημα του αρχικού εγγράφου. Τυπικά υλοποιούνται με νευρωνικές αρχιτεκτονικές ακολουθίας-σε-ακολουθία (encoder-decoder) ή με μοντέλα παραγωγής κειμένου, που μαθαίνουν να δημιουργούν συνεκτικές περιλήψεις. Παράγουν συχνά πιο "αναγνώσιμες" περιλήψεις, όμως εισάγουν τον κίνδυνο παραγωγής περιεχομένου που δεν τεκμηριώνεται επαρκώς από την είσοδο (hallucination), κάτι ιδιαίτερα κρίσιμο για πατέντες. Για αυτό, σε τεχνικά κείμενα απαιτείται προσεκτικός έλεγχος ισορροπίας ανάμεσα σε συνοχή και πιστότητα.[8]

2.2.3 Instructive σύνοψη

Η instructive σύνοψη στηρίζεται στην παροχή ρητών οδηγιών προς το μοντέλο, ώστε η παραγόμενη περίληψη να ακολουθεί συγκεκριμένη μορφή ή να εστιάζει σε συγκεκριμένα σημεία. Στην πράξη, η είσοδος δεν είναι μόνο το κείμενο της πατέντας αλλά και ένα "prompt" που ορίζει στόχο, ύφος ή δομή. [9] Αυτό επιτρέπει προσαρμογή σε διαφορετικά σενάρια χρήσης (π.χ. τεχνική περίληψη έναντι συνοπτικής περιγραφής για μη ειδικούς) και ευνοεί τη συστηματική εξαγωγή συγκεκριμένων τύπων πληροφορίας. Η πρόκληση βρίσκεται στη σταθερότητα των αποτελεσμάτων, καθώς μικρές διαφοροποιήσεις στο prompt ή στο περιεχόμενο της εισόδου μπορεί να επηρεάζουν την έξοδο. [10]

2.2.4 Hybrid προσεγγίσεις σύνοψης

Οι hybrid προσεγγίσεις συνδυάζουν στάδια, με στόχο να εκμεταλλευτούν τα πλεονεκτήματα διαφορετικών κατηγοριών. Μια συνηθισμένη λογική είναι: πρώτα επιλογή σημαντικών αποσπασμάτων (extractive) και έπειτα αναδιατύπωση/οργάνωση (abstractive ή instructive). Αυτό είναι ιδιαίτερα χρήσιμο σε μεγάλα έγγραφα, όπου ένα παραγωγικό μοντέλο δυσκολεύεται να "δει" όλο το κείμενο, αλλά μπορεί να δουλέψει αποδοτικά πάνω σε ένα επιλεγμένο υποσύνολο. Στις πατέντες, οι hybrid μέθοδοι στοχεύουν σε ισορροπία ανάμεσα σε τεχνική ακρίβεια και αναγνωσιμότητα, μειώνοντας τον κίνδυνο απώλειας κρίσιμων λεπτομερειών. [11], [12]

2.2.5 Ποιότητα σύνοψης και κριτήρια αξιολόγησης

Η αξιολόγηση της σύνοψης μπορεί να γίνει είτε ποσοτικά είτε ποιοτικά. Ποσοτικές μετρικές συγκρίνουν την παραγόμενη σύνοψη με αναφορά (reference) ή εκτιμούν ομοιότητα/συμπίεση (π.χ. λόγος συμπίεσης, σημασιολογική ομοιότητα). Ωστόσο, σε πατέντες, η ποιοτική διάσταση είναι εξίσου κρίσιμη: διατήρηση τεχνικών όρων, αποφυγή παραποίησης, επάρκεια κάλυψης κύριων ιδεών, και σαφήνεια. Η επιλογή μεθόδου σύνοψης επηρεάζει άμεσα αυτά τα κριτήρια και, κατ' επέκταση, μπορεί να επηρεάσει και την απόδοση σε αναζήτηση/ανάκτηση.

2.3 Γλωσσικά μοντέλα και αρχιτεκτονικές transformers σε πατέντες

Τα σύγχρονα γλωσσικά μοντέλα βασίζονται κατά κύριο λόγο σε αρχιτεκτονικές transformers, οι οποίες αξιοποιούν μηχανισμούς προσοχής (attention) για να αποτυπώσουν σχέσεις μεταξύ όρων και προτάσεων σε μεγάλα συμφραζόμενα. Η επικράτησή τους οφείλεται στην αποτελεσματικότητα στη μάθηση σημασιολογικών αναπαραστάσεων, τόσο για κατανόηση (encoding) όσο και για παραγωγή κειμένου (generation).

2.3.1 Encoder, decoder και encoder-decoder μοντέλα

Τα encoder μοντέλα (π.χ. BERT-τύπου) στοχεύουν κυρίως στην παραγωγή αναπαραστάσεων (embeddings) κατάλληλων για κατηγοριοποίηση, ομοιότητα και retrieval. Τα decoder μοντέλα (GPT-τύπου) είναι κατάλληλα για παραγωγή κειμένου και instruction-following. Τα encoder-decoder (T5/PEGASUS-τύπου) χρησιμοποιούνται ευρέως στη σύνοψη, καθώς κωδικοποιούν το κείμενο εισόδου και παράγουν περίληψη ως έξοδο. [13]

2.3.2 Μακροσκελή κείμενα και μηχανισμοί long-context

Οι πατέντες είναι συχνά πολύ μεγάλες για να εισαχθούν αυτούσιες σε τυπικά μοντέλα με περιορισμένο μήκος εισόδου. Αυτό οδηγεί σε ανάγκη είτε για chunking είτε για μοντέλα long-context που τροποποιούν τον attention μηχανισμό ώστε να κλιμακώνει καλύτερα (π.χ. αραιή προσοχή ή global attention). Οι τεχνικές αυτές επιτρέπουν την καλύτερη αξιοποίηση μεγάλων πεδίων όπως DESCRIPTION και CLAIMS, χωρίς να χάνεται ολική εικόνα.

2.3.3 Domain adaptation σε πατέντες

Η γλώσσα των πατεντών διαφέρει από το γενικό κείμενο ως προς την ορολογία, τη δομή και τη νομικοτεχνική διατύπωση. Για αυτό, έχουν προκύψει μοντέλα που προεκπαιδεύονται ή προσαρμόζονται σε τεχνικά/επιστημονικά corpora, με στόχο καλύτερες αναπαραστάσεις και πιο αξιόπιστη σύνοψη. Η domain προσαρμογή αποκτά ιδιαίτερη σημασία όταν ο στόχος δεν είναι μόνο "ωραία" περίληψη, αλλά και διατήρηση κρίσιμων τεχνικών πληροφοριών που επηρεάζουν downstream διαδικασίες όπως το retrieval. [14]

2.4 Τεχνικές ευρετηρίασης και αναζήτησης: στατιστικές και νευρωνικές προσεγγίσεις

Η ευρετηρίαση (indexing) και η αναζήτηση αποτελούν τον μηχανισμό που επιτρέπει σε ένα σύστημα IR να απαντά γρήγορα και αξιόπιστα σε ερωτήματα. Η εξέλιξη του πεδίου έχει οδηγήσει από καθαρά στατιστικές προσεγγίσεις σε νευρωνικές τεχνικές που αξιοποιούν embeddings, καθώς και σε υβριδικά σχήματα που συνδυάζουν τα δύο.

2.4.1 Στατιστικές προσεγγίσεις και λεξιλογική αντιστοίχιση BM25

Οι κλασικές μέθοδοι βασίζονται σε αναπαράσταση "bag-of-words" και στη λεξιλογική επικάλυψη μεταξύ ερωτήματος και εγγράφου. Σε αυτή την κατηγορία, μοντέλα όπως το BM25 αξιοποιούν τη

συχνότητα όρων και κανονικοποιήσεις ως προς το μήκος εγγράφου για να αποδώσουν score συνάφειας. Η προσέγγιση αυτή είναι ερμηνεύσιμη, αποδοτική και ιδιαίτερα ισχυρή ως baseline. Παράλληλα, όμως, παρουσιάζει περιορισμό σε παραφράσεις και σημασιολογικές ισοδυναμίες, κάτι που στις πατέντες είναι συχνό φαινόμενο λόγω διαφορετικών διατυπώσεων της ίδιας τεχνικής έννοιας. [15], [16]

2.4.2 Νευρωνικές προσεγγίσεις ανάκτησης: late-interaction και ColBERT

Οι νευρωνικές μέθοδοι ανάκτησης χαρτογραφούν ερωτήματα και έγγραφα σε διανυσματικές αναπαραστάσεις, ώστε η συνάφεια να εκτιμάται ως εγγύτητα σε σημασιολογικό χώρο. Στην κατηγορία late interaction, όπως στην αρχιτεκτονική ColBERT, η αναπαράσταση δεν συμπυκνώνεται σε ένα ενιαίο embedding ανά κείμενο: το ερώτημα και το έγγραφο κωδικοποιούνται σε contextualized embeddings ανά token και η συνάφεια υπολογίζεται μέσω λεπτομερούς αντιστοίχισης σε επίπεδο tokens (π.χ. MaxSim/aggregation). Αυτή η προσέγγιση διατηρεί περισσότερο σημασιολογικό "σήμα" από το αρχικό κείμενο και μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη σε τεχνικά έγγραφα, όπου συγκεκριμένοι όροι ή μικρές φράσεις είναι καθοριστικοί για τη σχετικότητα. [17]

2.4.3 Chunking, passage retrieval και document-level aggregation

Σε μεγάλα έγγραφα, η ανάκτηση σε επίπεδο πλήρους εγγράφου μπορεί να "θολώνει" τη συνάφεια όταν μόνο μικρό μέρος του κειμένου είναι σχετικό. Για αυτό χρησιμοποιείται passage-level ανάκτηση (ανάκτηση σε chunks/αποσπάσματα), όπου το έγγραφο τεμαχίζεται σε μικρότερες μονάδες. Στη συνέχεια, απαιτείται συνάθροιση (aggregation) των passage scores για παραγωγή τελικού score ανά έγγραφο. Κλασικές στρατηγικές είναι το μέγιστο score (MaxP), το άθροισμα scores (SumP) ή παραλλαγές τους. Η λογική αυτή είναι ιδιαίτερα σημαντική για πατέντες, όπου DESCRIPTION και CLAIMS είναι εκτενή και η πληροφορία στόχος συχνά εμφανίζεται σε απομακρυσμένα σημεία.

2.4.4 Υβριδικά σχήματα ανάκτησης

Υβριδικές προσεγγίσεις συνδυάζουν στατιστικά και νευρωνικά αποτελέσματα, είτε σε επίπεδο ανάκτησης υποψηφίων (candidate generation) είτε σε επίπεδο τελικής κατάταξης. Μια συνήθης πρακτική είναι η ένωση λιστών αποτελεσμάτων και ο συνδυασμός βαθμολογιών, ώστε να αξιοποιηθεί η "ανθεκτικότητα" της λεξιλογικής αντιστοίχισης μαζί με τη σημασιολογική κάλυψη των embeddings. Στο πλαίσιο πατεντών, τέτοιες προσεγγίσεις είναι ιδιαίτερα χρήσιμες όταν η απόδοση εξαρτάται ταυτόχρονα από τεχνικούς όρους (όπου το BM25 είναι ισχυρό) και από σημασιολογικές παραφράσεις (όπου οι νευρωνικές μέθοδοι υπερτερούν).

2.5 Σύνοψη Κεφαλαίου

Συνοψίζοντας, η ανάκτηση πληροφορίας σε μεγάλα κειμενικά σύνολα απαιτεί αποτελεσματική κατάταξη, αξιόπιστη αξιολόγηση και στρατηγικές αντιμετώπισης μεγάλου μήκους εγγράφων. Παράλληλα, η σύνοψη κειμένου προσφέρει διαφορετικές προσεγγίσεις συμπίεσης και αναδιατύπωσης, με κρίσιμες διαφορές σε πιστότητα, συνοχή και ελεγχόμενη παραγωγή περιεχομένου. Στην παρούσα εργασία, οι έννοιες αυτές συνδυάζονται: οι παραγόμενες περιλήψεις εξετάζονται όχι μόνο ως κείμενα, αλλά και ως είσοδος που μπορεί να επηρεάσει την ευρετηρίαση και την ανάκτηση πατεντών. Το θεωρητικό υπόβαθρο του κεφαλαίου λειτουργεί ως βάση για τα επόμενα κεφάλαια, όπου παρουσιάζονται η προεπεξεργασία δεδομένων, οι μέθοδοι σύνοψης και οι τεχνικές ανάκτησης που εφαρμόστηκαν, καθώς και η πειραματική αξιολόγηση των αποτελεσμάτων.

Η ανάκτηση πληροφορίας αποτελεί έναν από τους βασικούς τομείς της επιστήμης της πληροφορικής και ασχολείται με τον εντοπισμό και την επιστροφή σχετικών εγγράφων από μεγάλες συλλογές

κειμένου, με βάση τις ανάγκες ενός χρήστη. Στα σύγχρονα πληροφοριακά συστήματα, η ανάκτηση πληροφορίας καλείται να διαχειριστεί τεράστιους όγκους δεδομένων, τα οποία συχνά χαρακτηρίζονται από ετερογένεια, πολυπλοκότητα και ποικιλία μορφών.

Σε μεγάλα κειμενικά σύνολα, όπως συλλογές πατεντών, επιστημονικών άρθρων ή νομικών εγγράφων, η απλή αναζήτηση λέξεων-κλειδιών δεν επαρκεί για την κάλυψη των πληροφοριακών αναγκών του χρήστη. Η πολυπλοκότητα της γλώσσας, η χρήση συνωνύμων και η εκτενής ανάπτυξη τεχνικών εννοιών καθιστούν αναγκαία την αξιοποίηση πιο εξελιγμένων μεθόδων ανάκτησης.
Η ανάκτηση πληροφορίας δεν περιορίζεται μόνο στην επιστροφή εγγράφων, αλλά επεκτείνεται και στην κατάταξή της με βάση τη συνάφεια, έτσι ώστε τα πιο σχετικά αποτελέσματα να παρουσιάζονται πρώτα. Η αποτελεσματικότητα ενός συστήματος ανάκτησης εξαρτάται σε μεγάλο βαθμό από τον τρόπο αναπαράστασης των εγγράφων, τη μορφή των ερωτημάτων και τις μεθόδους σύγκρισης μεταξύ τους.

Κεφάλαιο 3ο: Δεδομένα και Προεπεξεργασία

3.1 Περιγραφή του αρχικού συνόλου δεδομένων

Για τις ανάγκες της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε σύνολο 2.592 εγγράφων πατεντών σε μορφή XML. Τα έγγραφα περιέχουν δομημένη πληροφορία με τεχνικό περιεχόμενο (τίτλος, abstract, περιγραφή, claims) καθώς και μεταδεδομένα (ταξινομήσεις IPCR/CPC, citations). Η XML μορφή είναι κατάλληλη για τυπική αποθήκευση, όμως δεν είναι πρακτική ως άμεση είσοδος στα επόμενα στάδια (σύνοψη, ευρετηρίαση, ανάκτηση), επειδή απαιτεί ενιαίο σχήμα πεδίων, σταθεροποίηση επιλογής περιεχομένου και συστηματικό καθαρισμό κειμένου.

3.2 Στόχος και σχεδιαστικές αρχές προεπεξεργασίας

Στόχος της προεπεξεργασίας ήταν η δημιουργία ενιαίων αρχείων κειμένου τύπου SGML-like, ώστε όλα τα επόμενα βήματα της εργασίας να βασίζονται σε κοινή, σταθερή αναπαράσταση. Η διαδικασία οργανώθηκε με δύο βασικές αρχές:

- Σταθερή επιλογή περιεχομένου (ώστε τα πεδία να είναι συγκρίσιμα μεταξύ εγγράφων).
- Ανθεκτικότητα σε πραγματικά δεδομένα (ώστε η μετατροπή να μην αποτυγχάνει σε ασυνέπειες XML/SGML-like).

3.3 Εξαγωγή πεδίων και τελικό σχήμα tags (SGML-like schema)

3.3.1 Κανόνες επιλογής περιεχομένου (language & source fallback)

Για την εξαγωγή των πεδίων εφαρμόστηκαν σαφείς κανόνες επιλογής:

- Γλώσσα: προτεραιότητα σε περιεχόμενο με lang="EN". Εάν δεν υπάρχει, γίνεται αποδοχή περιεχομένου χωρίς attribute γλώσσας ως fallback.
- Πηγή: για Abstract/Description/Claims δίνεται προτεραιότητα σε load-source="patent-office" και στη συνέχεια fallback σε load-source="docdb".

3.3.2 Τελικό σχήμα tags που χρησιμοποιήθηκε στα πειράματα

Κάθε έγγραφο αποθηκεύεται ως.txt και περιβάλλεται από ρίζα <DOC>... </DOC>. Η τελική αναπαράσταση που χρησιμοποιήθηκε ως query-corpus περιλαμβάνει σταθερή σειρά tags και διαχωρίζει καθαρά (α) μεταδεδομένα/blocks και (β) καθαρό κείμενο προς σύνοψη και retrieval. Συνοπτικά, το τελικό σύνολο tags περιλαμβάνει:

- <DOCNO>
- <TITLE>
- <ABSTRACT>
- <IPCR-CLASSIFICATIONS>
- <CPC-CLASSIFICATIONS>
- <CITATIONS>
- <DESCRIPTION>
- <BRIEF-DESCRIPTION>

- <BRIEF-DESCRIPTION-OF-THE-DRAWINGS>
- <DETAILED-DESCRIPTION>
- <FIRST-CLAIM>
- <CLAIMS>

Η παρουσία ταξινομήσεων (IPCR/CPC), citations και claims στο ίδιο ενιαίο "document envelope" επιτρέπει όχι μόνο την ανάκτηση με βάση το κείμενο, αλλά και μελλοντικές επεκτάσεις όπου απαιτείται αξιολόγηση/ανάλυση πεδίων ή φιλτράρισμα με βάση μεταδεδομένα.

3.4 Καθαρισμός κειμένου και ομογενοποίηση μορφής

Μετά την εξαγωγή των πεδίων εφαρμόστηκε ενιαία συνάρτηση καθαρισμού στα βασικά text fields (TITLE, ABSTRACT, DESCRIPTION, CLAIMS), με στόχο τη μείωση θορύβου και την παραγωγή κειμένου κατάλληλου για σύνοψη και ανάκτηση. Σε υψηλό επίπεδο, ο καθαρισμός περιλαμβάνει:

- Κανονικοποίηση encoding και απομάκρυνση μη έγκυρων bytes.
- Αφαίρεση markers απαρίθμησης (ιδιαίτερα σε claims/λίστες) και αφαίρεση αναφορών σε σχήματα/εικόνες.
- Αφαίρεση URLs/emails και υπολειμματικών XML/HTML tags.
- Διόρθωση hyphenation μεταξύ γραμμών.
- Κανονικοποίηση παύλων/στίξης, ομογενοποίηση whitespace/newlines και περιορισμό υπερβολικής στίξης.
- Αφαίρεση πολύ θορυβωδών tokens/μονοψήφιων artifacts και τελικό trim.

Η κοινή αυτή διαδικασία καθαρισμού διασφαλίζει ότι τα επόμενα στάδια (σύνοψη, BM25 indexing, ColBERT indexing/retrieval) εφαρμόζονται πάνω σε δεδομένα σταθερής ποιότητας και μορφής, ώστε οι συγκρίσεις μεταξύ διαφορετικών runs να είναι δίκαιες.

3.5 Ανθεκτική ανάγνωση πεδίων και σταθεροποίηση parsing (Plan A / Plan B)

Στα πραγματικά δεδομένα, SGML/XML-like αρχεία μπορεί να περιέχουν μη-ασφαλείς χαρακτήρες ή ασυνέπειες που προκαλούν parsing failures. Για αυτό υιοθετήθηκε μηχανισμός ανθεκτικότητας δύο επιπέδων:

- Plan A: προσπάθεια parsing με XML parser, αφού προηγηθεί προεπεξεργασία που κάνει escape σε "γυμνά" &, προστατεύει ακολουθίες τύπου]]> και τυλίγει συγκεκριμένα tags σε CDATA ώστε να μειωθούν ParseError σε πεδία με έντονο θόρυβο/σύμβολα.
- Plan B: fallback ανάκτηση πεδίων με regex απευθείας από raw string, όταν ο parser αποτυγχάνει ή όταν απαιτείται διατήρηση εσωτερικής δομής.

Στο τελικό "χτίσιμο" των SGML αρχείων γίνεται και διακριτή μεταχείριση πεδίων:

α. "raw blocks" που διατηρούνται ως έχουν (π.χ. IPCR-CLASSIFICATIONS, CPC-CLASSIFICATIONS, CITATIONS) και

β. text fields που γράφονται με ασφαλές escaping, ώστε το παραγόμενο SGML να παραμένει σταθερό ως προς parsing και επαναληπτική χρήση.

3.6 Διάσπαση του DESCRIPTION σε επιπλέον tags και λόγος σχεδίασης

Το σημαντικότερο δομικό αποτέλεσμα του τελικού σταδίου προεπεξεργασίας είναι ότι το DESCRIPTION δεν παραμένει ως ένα ενιαίο, αδιαχώριστο block. Αντίθετα, αναδομείται και παράγονται επιπλέον tags που αντιστοιχούν σε τεχνικά χρήσιμες υποενότητες (BRIEF/BACKGROUND/SUMMARY, BRIEF DESCRIPTION OF THE DRAWINGS, DETAILED DESCRIPTION), με ευριστικούς κανόνες που στοχεύουν σε σταθερότητα ακόμη και σε ετερογενείς περιγραφές πατεντών.

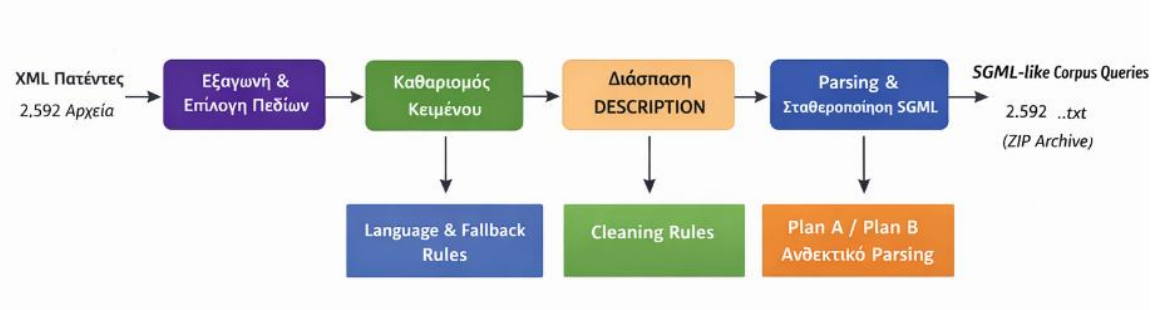
Ο λόγος αυτής της διάσπασης είναι διπλός:

1. Τεχνική οργάνωση του περιεχομένου: διαφορετικά "κομμάτια" του DESCRIPTION έχουν διαφορετικό ρόλο (συνοπτικό υπόβαθρο έναντι λεπτομερούς τεχνικής υλοποίησης). Με τη διάσπαση, το downstream pipeline μπορεί να αξιοποιήσει πιο στοχευμένα τμήματα κειμένου.
2. Επαναχρησιμοποίηση ως query-corpus: τα τελικά SGML αρχεία δεν εξυπηρετούν μόνο τα πειράματα της παρούσας διπλωματικής, αλλά είναι σχεδιασμένα ώστε να μπορούν να αξιοποιηθούν σε μελλοντικές μελέτες (π.χ. query enrichment ή στοχευμένη ανάκτηση ανά υπο-ενότητα), χωρίς να απαιτείται εκ νέου parsing/αναδόμηση του DESCRIPTION.

3.7 Παραγόμενο corpus queries και αναπαραγωγιμότητα

Η τελική έξοδος της προεπεξεργασίας είναι ένα σύνολο 2.592 δομημένων αρχείων.txt, με κοινό schema tags και καθαρισμένο κείμενο στα βασικά πεδία. Για λόγους φορητότητας και αναπαραγωγιμότητας, τα τελικά SGML queries πακετάρονται σε ένα μοναδικό ZIP artifact εξόδου, ενώ η επιλογή των αρχείων μπορεί να ελεγχθεί (π.χ. με target ids ή όριο πλήθους) ώστε να διευκολύνονται γρήγορα tests ή πλήρεις εκτελέσεις.

Επιπλέον, η διαδικασία καταγράφει επιτυχίες/αποτυχίες και συνεχίζει ακόμη και όταν ένα μεμονωμένο αρχείο αποτύχει, ώστε να μην ακυρώνεται ολόκληρη η παραγωγή του query corpus.



Διάγραμμα 1: Ροή προ επεξεργασίας εγγράφων πατεντών από αρχική μορφή XML σε δομημένα αρχεία τύπου SGML-like, κατάλληλα για σύνοψη και ανάκτηση πληροφορίας

Κεφάλαιο 4ο: Μέθοδοι Σύνοψης Πατεντών

4.1 Γενική αρχιτεκτονική συστήματος σύνοψης

Η διαδικασία σύνοψης πατεντών στην παρούσα διπλωματική βασίζεται σε ενιαία αρχιτεκτονική, ανεξάρτητη από τη μέθοδο σύνοψης που εφαρμόζεται. Ο σχεδιασμός αυτός επιτρέπει τη δίκαιη σύγκριση διαφορετικών προσεγγίσεων, καθώς όλες λειτουργούν πάνω στο ίδιο σύνολο δεδομένων και ακολουθούν κοινή ροή επεξεργασίας.

Η αρχιτεκτονική του συστήματος σύνοψης ξεκινά από τα δομημένα αρχεία τύπου SGML-like που προέκυψαν από τη διαδικασία προεπεξεργασίας. Κάθε έγγραφο πατέντας αποτελείται από διακριτά πεδία, όπως τίτλος, περίληψη, περιγραφή και αξιώσεις, τα οποία μπορούν να αξιοποιηθούν μεμονωμένα ή συνδυαστικά, ανάλογα με τις απαιτήσεις της εκάστοτε μεθόδου σύνοψης.

Σε πρώτο στάδιο, επιλέγεται το κειμενικό περιεχόμενο που θα αποτελέσει είσοδο για τη σύνοψη. Στη συνέχεια, εφαρμόζεται η αντίστοιχη προσέγγιση σύνοψης (extractive, abstractive, instructive ή hybrid), η οποία παράγει μία ή περισσότερες περιλήψεις για κάθε έγγραφο. Οι παραγόμενες περιλήψεις αποθηκεύονται σε δομημένη μορφή και χρησιμοποιούνται τόσο για ποιοτική ανάλυση όσο και ως είσοδος στα επόμενα στάδια ευρετηρίασης και αναζήτησης πληροφορίας.

Η χρήση κοινής αρχιτεκτονικής διασφαλίζει ότι οι διαφοροποιήσεις στα αποτελέσματα οφείλονται αποκλειστικά στη μέθοδο σύνοψης και όχι σε διαφορές στη μορφή ή την ποιότητα των δεδομένων εισόδου.

4.2 Extractive σύνοψη

Η extractive σύνοψη βασίζεται στην επιλογή και συνδυασμό τμημάτων του αρχικού κειμένου, χωρίς τη δημιουργία νέου περιεχομένου. Στην περίπτωση των πατεντών, η προσέγγιση αυτή αξιοποιεί την εγγενή δομή των εγγράφων, επιλέγοντας προτάσεις ή αποσπάσματα που θεωρούνται περισσότερο αντιπροσωπευτικά του συνολικού περιεχομένου. [7]

Οι extractive μέθοδοι παρουσιάζουν το πλεονέκτημα της υψηλής πιστότητας προς το αρχικό κείμενο, καθώς δεν αλλοιώνουν την τεχνική ορολογία και το νομικό περιεχόμενο των πατεντών. Για τον λόγο αυτό, θεωρούνται ιδιαίτερα κατάλληλες για τεχνικά και νομικά έγγραφα, όπου η ακρίβεια είναι κρίσιμη.

Στην παρούσα εργασία επιλέχθηκαν extractive μοντέλα που βασίζονται σε σύγχρονες σημασιολογικές αναπαραστάσεις προτάσεων, ώστε η επιλογή των αποσπασμάτων να μην στηρίζεται αποκλειστικά σε στατιστικά κριτήρια, αλλά και στο περιεχόμενο και το νόημα του κειμένου.

Για λόγους κατανόησης της extractive προσέγγισης, παρουσιάζεται ενδεικτικό παράδειγμα σύνοψης που προέκυψε από πραγματικό έγγραφο πατέντας και βασίζεται στο πεδίο <DESCRIPTION>.

Παράδειγμα Extractive σύνοψη (PatentSBERTa_V2):

<DESCRIPTION>The invention provides compounds that inhibit activity of one or more Pim kinases and are useful for the treatment of cancers and autoimmune disorders.
These compounds exhibit improved therapeutic effects and pharmacokinetic properties compared to previously disclosed inhibitors. </DESCRIPTION>

Σχόλιο:

Το παράδειγμα καταδεικνύει ότι η extractive σύνοψη διατηρεί αυτούσια τεχνικά τμήματα του αρχικού κειμένου, εξασφαλίζοντας υψηλή πιστότητα στο περιεχόμενο της πατέντας.

4.3 Abstractive σύνοψη

Η abstractive σύνοψη στοχεύει στη δημιουργία νέου κειμένου, το οποίο αποτυπώνει το νόημα του αρχικού εγγράφου με πιο συνοπτικό και φυσικό τρόπο. Οι μέθοδοι αυτές αξιοποιούν νευρωνικά γλωσσικά μοντέλα, τα οποία έχουν εκπαιδευτεί να παράγουν κείμενο με βάση τη σημασιολογική κατανόηση της εισόδου.

Σε αντίθεση με τις extractive προσεγγίσεις, οι abstractive μέθοδοι προσφέρουν μεγαλύτερη ευελιξία και συνοχή στη διατύπωση, ωστόσο ενδέχεται να εισαγάγουν αποκλίσεις από το αρχικό κείμενο. Στο πλαίσιο των πατεντών, η πρόκληση έγκειται στη διατήρηση της τεχνικής ακρίβειας, χωρίς απώλεια κρίσιμων πληροφοριών. [8]

Τα μοντέλα που επιλέχθηκαν στην παρούσα εργασία ανήκουν στην κατηγορία των transformer-based αρχιτεκτονικών και έχουν σχεδιαστεί ή προσαρμοστεί για μεγάλα και πολύπλοκα κείμενα, όπως αυτά των πατεντών.

Παράδειγμα Abstractive σύνοψη (T5_Small_HUPD):

<DESCRIPTION> The present invention relates to new compounds and their tautomers and pharmaceutically acceptable salts, esters, metabolites or prodrugs thereof, compositions of the new compounds together with pharmaceutically unacceptable carriers, and uses of these compounds, either alone or in combination with at least one additional therapeutic agent, in the prophylaxis or treatment of cancer and other cellular proliferation disorders. In particular, mutational activation of several well known oncogenes in hematopoietic malignancies is thought to exert its effects at least in part through Pim( ) </DESCRIPTION>

Σχόλιο:

Η abstractive σύνοψη συμπυκνώνει το περιεχόμενο σε νέα διατύπωση, διατηρώντας τον βασικό σκοπό της εφεύρεσης χωρίς να ακολουθεί τη δομή του αρχικού κειμένου.

4.4 Instructive σύνοψη

Η instructive σύνοψη βασίζεται στη χρήση ρητών οδηγιών ή ερωτημάτων, τα οποία καθοδηγούν το γλωσσικό μοντέλο ως προς το είδος και το περιεχόμενο της παραγόμενης σύνοψης. Οι προσεγγίσεις αυτές έχουν αναδειχθεί ιδιαίτερα με την εμφάνιση μεγάλων γλωσσικών μοντέλων, τα οποία μπορούν να ακολουθούν σύνθετες εντολές. [10]

Στο πλαίσιο της παρούσας εργασίας, οι instructive μέθοδοι επιτρέπουν την προσαρμογή της σύνοψης στις ανάγκες του εκάστοτε σεναρίου χρήσης, όπως η εστίαση σε τεχνικά χαρακτηριστικά ή η συνοπτική

παρουσίαση της εφεύρεσης. Το πλεονέκτημα της προσέγγισης αυτής είναι η ευελιξία, ενώ η πρόκληση έγκειται στη σταθερότητα και την επαναληψιμότητα των αποτελεσμάτων.

Η instructive σύνοψη αντιμετωπίζεται ως αυτόνομη κατηγορία μεθόδων και όχι ως παραλλαγή της abstractive, καθώς διαφοροποιείται ουσιαστικά στον τρόπο ελέγχου της διαδικασίας παραγωγής σύνοψης.

Παράδειγμα Instructive σύνοψη (Mistral_7B):

<DESCRIPTION>The crude material was purified via silica gel column chromatography, resulting in high-yield synthesis of substituted cyclohexenyl and aminopyridyl compounds. </DESCRIPTION>

Σχόλιο:

Το παράδειγμα αναδεικνύει ότι η instructive σύνοψη εξαρτάται έντονα από τη διατύπωση της οδηγίας και ενδέχεται να εστιάσει σε επιμέρους τεχνικές διαδικασίες αντί στη συνολική περιγραφή της εφεύρεσης.

4.5 Hybrid προσεγγίσεις σύνοψης

Οι hybrid προσεγγίσεις συνδυάζουν στοιχεία από περισσότερες της μίας κατηγορίες σύνοψης, με στόχο την αξιοποίηση των πλεονεκτημάτων κάθε μεθόδου. Συνήθως, περιλαμβάνουν ένα αρχικό στάδιο extractive επιλογής σημαντικών τμημάτων και ένα επόμενο στάδιο αναδιατύπωσης ή εμπλουτισμού μέσω νευρωνικών ή instructive τεχνικών.

Στην παρούσα εργασία, οι hybrid προσεγγίσεις αξιοποιούνται για να επιτευχθεί ισορροπία μεταξύ τεχνικής ακρίβειας και αναγνωσιμότητας. Η επιλογή τους κρίνεται ιδιαίτερα κατάλληλη για πατέντες, όπου απαιτείται τόσο πιστότητα στο αρχικό κείμενο όσο και συνοπτική παρουσίαση του περιεχομένου.

Παράδειγμα Hybrid σύνοψη (PatentSBERTa_V2 + Mistral_7B):

<DESCRIPTION> This patent describes synthesizing various compounds, including those containing polycyclic aryl structures, such as benzodioxozolo, and nitrogen-containing groups that can be quaternized. Notable examples of these compounds include cyclohexenyl Boc amino alcohol nitro pyridyl compound VIII, which can be modified to produce substituted aminocyclohexylpyridyl amides Xc. Another compound synthesized is methyl (-difluoro-hydroxyphenyl)-fluoropicolinate. The process involves using solvents like tetrahydrofuran and tetrahydropyran, and quaternization agents like lower alkyl halides and dialkyl sulfates. The synthesis method also includes purification steps, such as silica gel column chromatography. </DESCRIPTION>

Σχόλιο:

Η hybrid προσέγγιση συνδυάζει την τεχνική ακρίβεια extractive μεθόδων με τη συνεκτική διατύπωση νευρωνικών μοντέλων.

4.6 Σύγκριση μεθόδων και μοντέλων

Για την υλοποίηση των τεσσάρων προσεγγίσεων σύνοψης επιλέχθηκε σύνολο σύγχρονων γλωσσικών μοντέλων, τα οποία καλύπτουν διαφορετικές αρχιτεκτονικές και φιλοσοφίες σχεδίασης. Ο Πίνακας 4.1 παρουσιάζει συνοπτικά τα μοντέλα που χρησιμοποιήθηκαν ανά κατηγορία σύνοψης.

Η επιλογή ποικιλίας γλωσσικών μοντέλων και τεχνικών σύνοψης αποσκοπεί στην κάλυψη διαφορετικών φιλοσοφιών και επιπέδων επεξεργασίας κειμένου. Τα μοντέλα που χρησιμοποιήθηκαν διαφέρουν ως προς την αρχιτεκτονική, το μέγεθος και τον τρόπο παραγωγής σύνοψης, επιτρέποντας τη μελέτη της συμπεριφοράς τόσο κλασικών όσο και σύγχρονων νευρωνικών προσεγγίσεων. Με τον τρόπο αυτό καθίσταται δυνατή η συγκριτική αξιολόγηση των τεσσάρων κατηγοριών σύνοψης σε κοινό σύνολο δεδομένων, καθώς και η διερεύνηση της καταλληλότητάς τους για την επεξεργασία πατεντών, οι οποίες χαρακτηρίζονται από εκτενή και τεχνικά απαιτητικό περιεχόμενο.

Ο πίνακας αυτός λειτουργεί ως σημείο αναφοράς για τα επόμενα κεφάλαια, όπου παρουσιάζονται παραδείγματα περιλήψεων και πειραματικά αποτελέσματα. Η σύγκριση των μεθόδων δεν περιορίζεται μόνο στην ποιότητα της σύνοψης, αλλά επεκτείνεται και στη συμβολή τους στις διαδικασίες ευρετηρίασης και αναζήτησης πληροφορίας.

Πίνακας 4.1: Γλωσσικά μοντέλα και τεχνικές που χρησιμοποιήθηκαν στην παρούσα εργασία, οργανωμένα ανά κατηγορία προσέγγισης σύνοψης.

Extractive Models	Abstractive Models	Instructive Models	Hybrid Models
all-mpnet-base-v2	BART_large_CNN	Llama3_8B_Instruct	PatentSBERTa_V2 + TextRank
bge-base-en-v1.5	BigBird_Pegasus	Mistral_7B_Instruct	PatentSBERTa_V2 + Mistral_7B
e5-base-v2	LED_Base_BigPatent	Qwen2_7B_Instruct	—
google_bert_for_patents	LongT5_TGlobal_Base	Qwen2_14B_Instruct	—
PatentSBERTa_V2	PEGASUS_BigPatent	—	—
SBERT_all-MiniLM-L6-v2	T5_Small_HUPD	—	—

4.6.1 Σύντομη περιγραφή των χρησιμοποιούμενων μοντέλων

Συνολικά, οι μέθοδοι και τα μοντέλα σύνοψης που παρουσιάζονται στο παρόν κεφάλαιο αξιοποιούνται όχι μόνο για βελτίωση αναγνωσιμότητας, αλλά και ως ενδιάμεσο στάδιο που μπορεί να επηρεάσει την ευρετηρίαση και την ανάκτηση πληροφορίας: η σύνοψη λειτουργεί ως συμπυκνωμένη αναπαράσταση του εγγράφου και, ανάλογα με το πόσο καλά διατηρεί κρίσιμους όρους/έννοιες, μπορεί να αλλάξει τη συμπεριφορά της αναζήτησης και την τελική κατάταξη αποτελεσμάτων. , ,

4.6.1.1 Extractive μοντέλα

Τα extractive μοντέλα που χρησιμοποιήθηκαν βασίζονται σε σημασιολογικές αναπαραστάσεις προτάσεων και επιλέγουν τμήματα του αρχικού κειμένου χωρίς αναδιατύπωση. Η λογική αυτή συνδέεται άμεσα με τα sentence-embedding μοντέλα τύπου SBERT, όπου η ομοιότητα προτάσεων σε έναν κοινό διανυσματικό χώρο επιτρέπει την επιλογή των πιο αντιπροσωπευτικών αποσπασμάτων.

- **SBERT_all-MiniLM-L6-v2**
  Το μοντέλο βασίζεται στο Sentence-BERT (SBERT), δηλαδή σε siamese/bi-encoder αρχιτεκτονική που παράγει embeddings προτάσεων για σύγκριση σημασιολογικής ομοιότητας. Η έκδοση MiniLM έχει στόχο χαμηλό υπολογιστικό κόστος με ικανοποιητική ποιότητα αναπαραστάσεων, άρα είναι πρακτική για μεγάλης κλίμακας εξαγωγή προτάσεων από πατέντες. Στην παρούσα εργασία χρησιμοποιείται για επιλογή «πιο κεντρικών» προτάσεων/αποσπασμάτων μέσω ομοιότητας, χωρίς αναδιατύπωση. [18]

- **all-mpnet-base-v2**
  Το MPNet είναι pretraining σχήμα που συνδυάζει ιδέες masked και permuted modeling, στοχεύοντας σε ισχυρότερες αναπαραστάσεις σε σχέση με κλασικά BERT-τύπου σχήματα. Η έκδοση “all-mpnet-base-v2” στη στοίβα των sentence-transformers αξιοποιείται για embeddings προτάσεων και, συνεπώς, για επιλογή αποσπασμάτων με βάση σημασιολογική συνάφεια. Πλεονέκτημά του είναι ότι συχνά αποδίδει πιο «λεπτές» σημασιολογικές διακρίσεις, χρήσιμες σε τεχνική ορολογία πατεντών. [19]

- **bge-base-en-v1.5**
  Το BGE ανήκει σε σύγχρονα embedding models γενικής χρήσης, τα οποία έχουν βελτιστοποιηθεί για retrieval/similarity, ώστε να αποδίδουν ισχυρές διανυσματικές αναπαραστάσεις για κείμενα και ερωτήματα. Στη σύνοψη extractive λειτουργεί ως «κριτής ομοιότητας» για την επιλογή προτάσεων που καλύπτουν καλύτερα το συνολικό περιεχόμενο. Πρακτικό πλεονέκτημα είναι ότι παραμένει αποδοτικό σε μαζική επεξεργασία, με σταθερή συμπεριφορά σε διαφορετικά υπο-πεδία τεχνικού λεξιλογίου. [20]

- **e5-base-v2**
  Η οικογένεια E5 (“Text Embeddings by Weakly-Supervised Contrastive Pre-training”) στοχεύει σε embeddings που δουλεύουν καλά τόσο για queries όσο και για documents, μέσω contrastive εκπαίδευσης μεγάλης κλίμακας. Αυτό το χαρακτηριστικό είναι χρήσιμο όταν το extractive στάδιο «προσεγγίζει» και λογική ανάκτησης: επιλέγονται προτάσεις που μοιάζουν περισσότερο με μία συνοπτική αναπαράσταση του εγγράφου ή/και με ερωτήματα αξιολόγησης. Στην πράξη, βοηθά σε περιπτώσεις παραφράσεων όπου το απλό keyword overlap είναι φτωχό.[21]

- **google_bert_for_patents**
  Πρόκειται για BERT-τύπου encoder που έχει προσαρμοστεί/εκπαιδευτεί πάνω σε πατεντιακά corpora, με στόχο να αποτυπώνει καλύτερα τη δομή και την ορολογία των πατεντών σε σχέση με γενικής χρήσης μοντέλα. Σε extractive σύνοψη δεν «γράφει» νέο κείμενο, αλλά χρησιμοποιείται ως βάση για αναπαραστάσεις (embeddings) που επιτρέπουν καλύτερη επιλογή προτάσεων σε τεχνικά/νομικά συμφραζόμενα. Πλεονέκτημα είναι η αυξημένη πιστότητα σε domain όρους, ακρωνύμια και τυποποιημένες διατυπώσεις.

- **PatentSBERTa_V2**
  Το PatentSBERTa αποτελεί domain-adapted transformer με έμφαση σε αναπαραστάσεις κατάλληλες για tasks πατεντών (π.χ. ομοιότητα, ταξινόμηση, retrieval), αξιοποιώντας προσαρμογή σε πατεντιακά δεδομένα. Στο extractive πλαίσιο λειτουργεί ως «σημασιολογικός μετρητής» για την επιλογή αποσπασμάτων με υψηλή τεχνική πυκνότητα, μειώνοντας τον κίνδυνο να επιλεγούν γενικόλογες προτάσεις. Πλεονέκτημα είναι ότι τείνει να διατηρεί καλύτερα την τεχνική ακρίβεια στην επιλογή, επειδή οι αναπαραστάσεις του είναι πιο συμβατές με τη γλώσσα των claims/description. [22]

4.6.1.2 Abstractive μοντέλα

Τα abstractive μοντέλα βασίζονται σε αρχιτεκτονικές encoder-decoder και είναι σχεδιασμένα για τη δημιουργία νέου κειμένου με βάση τη σημασιολογική κατανόηση της εισόδου.

- **BART_large_CNN**
  Το BART είναι encoder–decoder transformer που προεκπαιδεύεται ως denoising autoencoder και στη συνέχεια fine-tune-άρεται για παραγωγικά tasks, όπως σύνοψη. Η έκδοση "bart-large-cnn" είναι κλασικό baseline για news-style summarization, άρα παρέχει ισχυρό σημείο αναφοράς, αλλά μπορεί να δυσκολεύεται σε πολύ μεγάλα και νομικο-τεχνικά κείμενα (λόγω μήκους και domain). Πλεονέκτημά του είναι η καλή συνοχή και αναγνωσιμότητα σε «σύντομη» είσοδο. [23]

- **PEGASUS_BigPatent**
  Το PEGASUS προεκπαιδεύεται με στόχο σύνοψη, μέσω gap-sentence generation, και έχει δείξει ισχυρά αποτελέσματα σε abstractive summarization. Η εκδοχή fine-tuned σε BigPatent αξιοποιεί μεγάλο, πατεντιακό dataset για να μάθει τυπικές δομές περίληψης πατεντών. Πλεονέκτημα είναι ότι τείνει να παράγει περίληψη πιο κοντά στο «ύφος» των abstracts πατεντών, σε σχέση με μοντέλα εκπαιδευμένα σε γενικό κείμενο. [24]

- **T5_Small_HUPD**
  Το T5 εισάγει ενοποιημένο text-to-text framework, όπου κάθε task (και η σύνοψη) εκφράζεται ως μετατροπή κειμένου σε κείμενο. Όταν fine-tune-άρεται σε πατεντιακά δεδομένα όπως το HUPD, αποκτά πιο κατάλληλες κατανομές λεξιλογίου και δομών για πατέντες, παρότι παραμένει «small» σε μέγεθος. Πλεονέκτημα είναι η αποδοτικότητα και η σταθερή συμπεριφορά σε μαζική παραγωγή περιλήψεων, με κόστος ότι οι περιλήψεις μπορεί να είναι λιγότερο πλούσιες σε λεπτομέρεια. [25]

- **BigBird_Pegasus**
  Το BigBird εισάγει sparse attention ώστε να κλιμακώνει σε πολύ μεγαλύτερα μήκη ακολουθιών από το κλασικό quadratic attention. Σε συνδυασμό με PEGASUS-style summarization, προκύπτει μοντέλο κατάλληλο για μεγάλα κείμενα όπως DESCRIPTION/CLAIMS, όπου τα συμβατικά encoder–decoder μοντέλα «κόβονται» σε μήκος εισόδου. Πλεονέκτημα είναι ότι μπορεί να «δει» περισσότερο περιεχόμενο ανά έγγραφο, μειώνοντας την ανάγκη επιθετικού chunking.[26]

- **LED_Base_BigPatent**
  Ο Longformer εισάγει attention μηχανισμούς για long documents και η παραλλαγή LED (Longformer Encoder-Decoder) στοχεύει άμεσα σε παραγωγικά tasks μεγάλου context, όπως σύνοψη. Αυτό είναι κρίσιμο στις πατέντες, όπου η πληροφορία είναι διάσπαρτη και η περίληψη πρέπει να συμπτύξει πολλά τμήματα (description, claims). Πλεονέκτημα είναι η καλύτερη κάλυψη περιεχομένου χωρίς υπερβολικό τεμαχισμό, με τίμημα συνήθως μεγαλύτερο υπολογιστικό κόστος από μικρότερα μοντέλα. [27]

- **LongT5_TGlobal_Base**
  Το LongT5 επεκτείνει τη λογική του T5 με αποδοτικότερη επεξεργασία μακρών ακολουθιών, εισάγοντας μηχανισμούς που μειώνουν το κόστος attention σε μεγάλα μήκη. Στη σύνοψη πατεντών αυτό επιτρέπει πιο «ολιστική» σύνοψη μεγάλων sections, ειδικά όταν θέλουμε να περιορίσουμε την απώλεια πληροφορίας που προκαλεί το chunking. Πλεονέκτημα είναι ότι

συνδυάζει το text-to-text φορμάτ του T5 με long-context δυνατότητες, άρα είναι φυσική επιλογή για εκτενή τεχνικά κείμενα. [28]

4.6.1.3 Instructive μοντέλα

Τα instructive μοντέλα ανήκουν στην κατηγορία μεγάλων γλωσσικών μοντέλων που ακολουθούν ρητές οδηγίες χρήστη (instruction-following) για την παραγωγή σύνοψης.

- **Llama3_8B_Instruct**
  Τα instruct μοντέλα της οικογένειας Llama εκπαιδεύονται ώστε να ακολουθούν οδηγίες και να παράγουν κείμενο προσαρμοσμένο σε prompt, κάτι που διευκολύνει τη στοχευμένη σύνοψη (π.χ. συγκεκριμένη δομή, bullets, έμφαση σε novelty). Στο πλαίσιο της εργασίας, ο ρόλος τους είναι να παράγουν περίληψη «ελεγχόμενη» από οδηγία, και όχι απλώς μια γενική abstractive σύνοψη. Πλεονέκτημα είναι η ευελιξία μορφής/στόχου, με βασική πρόκληση τη σταθερότητα (ίδιο prompt δεν εγγυάται πάντα ίδιο ύφος/βάθος). [29]

- **Mistral_7B_Instruct**
  Τα instruct-tuned LLMs (όπως Mistral 7B) αξιοποιούνται για παραγωγή σύνοψης με βάση ρητές απαιτήσεις και μπορούν να δώσουν πιο «ανθρωποκεντρική» διατύπωση από κλασικά summarization fine-tunes. Σε πατέντες, αυτό βοηθά όταν ζητάμε να αναδειχθεί σκοπός, τεχνικό πρόβλημα και συμβολή, αλλά χρειάζεται προσοχή στον κίνδυνο εισαγωγής μη τεκμηριωμένων λεπτομερειών. Πλεονέκτημα είναι η καλή ισορροπία κόστους/ποιότητας σε σχέση με πολύ μεγαλύτερα LLMs, ειδικά σε batch παραγωγή. [30]

- **Qwen2_7B_Instruct**
  Το Qwen2 είναι οικογένεια LLMs που υποστηρίζει instruction-following και μπορεί να προσαρμόσει το παραγόμενο κείμενο σε σαφείς οδηγίες μορφής και περιεχομένου. Για σύνοψη πατεντών αξιοποιείται όταν θέλουμε συγκεκριμένη δομή εξόδου (π.χ. “Problem–Solution–Advantages”) ή όταν θέλουμε να δώσουμε έμφαση σε συγκεκριμένα πεδία (claims, novelty, εφαρμογές). Πλεονέκτημα είναι η ισχυρή ικανότητα prompt adherence, που κάνει τη σύνοψη πιο «ελεγχόμενη» σε σχέση με καθαρά abstractive pipelines. [31]

- **Qwen2_14B_Instruct**
  Η μεγαλύτερη παραλλαγή (14B) συνήθως ενισχύει την ικανότητα κατανόησης σύνθετων τεχνικών συμφραζομένων και την ποιότητα διατύπωσης, ιδιαίτερα όταν το prompt ζητά πολλαπλές όψεις της εφεύρεσης. Στις πατέντες αυτό μπορεί να βελτιώσει τη συνοχή και την κάλυψη, ειδικά όταν η είσοδος περιέχει πυκνή τεχνική πληροφορία. Το κόστος είναι αυξημένοι πόροι και μεγαλύτερος χρόνος παραγωγής σε σχέση με 7B. [31]

4.6.1.4 Hybrid προσεγγίσεις

Οι hybrid προσεγγίσεις συνδυάζουν extractive και νευρωνικές τεχνικές, με στόχο τη διατήρηση τεχνικής ακρίβειας και τη βελτίωση της αναγνωσιμότητας.

- **PatentSBERTa_V2 + TextRank**

  Το TextRank είναι γράφημα-βασισμένη μέθοδος εξαγωγής προτάσεων, όπου η σημαντικότητα προκύπτει μέσω αλγορίθμου τύπου PageRank πάνω σε γράφημα ομοιότητας προτάσεων. Σε συνδυασμό με PatentSBERTa embeddings, το γράφημα ομοιότητας αποκτά domain-προσαρμοσμένη σημασιολογία, άρα οι «κεντρικές» προτάσεις τείνουν να είναι πιο τεχνικά χρήσιμες για πατέντες. Πλεονέκτημα είναι η υψηλή πιστότητα (extractive) με πιο «συστηματική» επιλογή κάλυψης περιεχομένου. [32], [22]

- **PatentSBERTa_V2 + Mistral_7B**

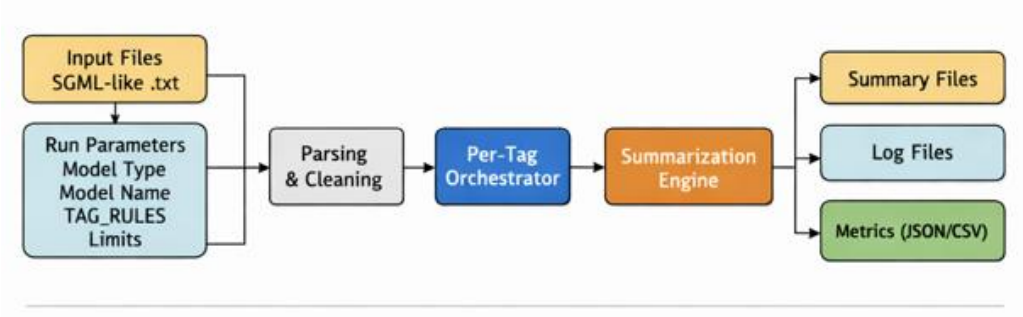
  Η λογική είναι δύο σταδίων: πρώτα extractive επιλογή αποσπασμάτων (με PatentSBERTa) ώστε να συγκεντρωθεί το πιο σχετικό υλικό, και έπειτα instructive/abstractive αναδιατύπωση με Mistral για καλύτερη συνοχή και αναγνωσιμότητα. Αυτό μειώνει την πιθανότητα το LLM να «χαθεί» σε πολύ μεγάλο κείμενο, αφού δουλεύει πάνω σε συμπιεσμένο υποσύνολο. Πλεονέκτημα είναι η ισορροπία τεχνικής ακρίβειας και ευαναγνωσιμότητας, με βασική προϋπόθεση ο σωστός έλεγχος του τι περνά στο στάδιο αναδιατύπωσης. [22], [30]

4.7 Ροή υλοποίησης και παραμετροποίηση εκτελέσεων σύνοψης

Σε αυτό το σημείο περιγράφεται το ενιαίο pipeline εκτέλεσης σύνοψης, το οποίο εφαρμόζεται με τον ίδιο βασικό τρόπο σε όλες τις κατηγορίες μεθόδων (extractive, abstractive, instructive, hybrid), αλλά με εναλλαγή του αντίστοιχου module/μοντέλου. Κάθε εκτέλεση (run) ορίζεται από ένα σύνολο παραμέτρων που καθορίζουν πλήρως το αποτέλεσμα και επιτρέπουν την αναπαραγωγή του πειράματος.

Η ροή ξεκινά με έναν κοινό κορμό ρυθμίσεων, όπου ορίζονται η ταυτότητα του run (τύπος μεθόδου και όνομα μοντέλου), τα όρια εισόδου (μέγιστο μήκος σε tokens), καθώς και παράμετροι που επηρεάζουν άμεσα την παραγωγή της σύνοψης (ρυθμίσεις decoding στα generative μοντέλα ή παράμετροι επιλογής σε extractive/hybrid). Κεντρικό στοιχείο της παραμετροποίησης αποτελεί ένα σύνολο κανόνων ανά πεδίο/ενότητα πατέντας (TAG_RULES), το οποίο λειτουργεί ως συμβόλαιο παραγωγής: για κάθε tag ορίζεται πότε ενεργοποιείται η σύνοψη (μέσω threshold) και ποιο είναι το επιθυμητό εύρος μεγέθους της (ελάχιστο/μέγιστο σε λέξεις ή/και tokens). Με τον τρόπο αυτό, το pipeline δεν αντιμετωπίζει το έγγραφο ως ένα ενιαίο κείμενο, αλλά ως δομημένο αντικείμενο με διαφορετικές απαιτήσεις ανά ενότητα (π.χ. ABSTRACT έναντι DESCRIPTION ή CLAIMS).

Στη συνέχεια, κάθε αρχείο εισόδου (SGML-like.txt) περνά από ασφαλές parsing ώστε να εξαχθούν τα πεδία που θα συνοψιστούν (π.χ. ABSTRACT, DESCRIPTION, BRIEF-DESCRIPTION, FIRST-CLAIM, CLAIMS), ενώ ο τίτλος (TITLE) διατηρείται αυτούσιος και εξαιρείται συστηματικά από τη διαδικασία σύνοψης. Μετά την εξαγωγή εφαρμόζεται προκαταρκτικός καθαρισμός κειμένου (εξομάλυνση κενών, απομάκρυνση θορύβου, προαιρετική αφαίρεση boilerplate), ώστε το υλικό που εισάγεται στον μηχανισμό παραγωγής να είναι σταθερό και συγκρίσιμο μεταξύ runs. Από αυτό το σημείο και μετά ενεργοποιείται το κύριο τεχνικό στάδιο: παραγωγή σύνοψης ανά tag, με κοινή σύμβαση εισόδου/εξόδου (λαμβάνει κείμενο tag και επιστρέφει σύνοψη tag), αλλά διαφορετικό μηχανισμό ανά κατηγορία script.



Διάγραμμα 2: Ενιαίο pipeline εκτέλεσης σύνοψης. Κάθε run ορίζεται πλήρως από τις παραμέτρους του και παράγει επανα-λήψιμο πακέτο αποτελεσμάτων.

Στα extractive scripts, το τεχνικό κέντρο βάρους βρίσκεται στη μοντελοποίηση της σημαντικότητας προτάσεων και στην ελεγχόμενη επιλογή τους. Το κείμενο του tag τμηματοποιείται σε προτάσεις (με περιορισμούς στο μέγιστο μήκος πρότασης ώστε να αποφεύγονται ακραίες περιπτώσεις), και στη συνέχεια παράγονται sentence embeddings (Sentence-BERT) για όλες τις προτάσεις, με κανονικοποίηση και batching για σταθερότητα και αποδοτικότητα. Η βαθμολόγηση υλοποιείται μέσω cosine similarity κάθε πρότασης ως προς το κεντροειδές (centroid) του συνόλου προτάσεων, ως πρακτική προσέγγιση του "αντιπροσωπευτικού" περιεχομένου. Με βάση τα scores εφαρμόζεται επιλογή προτάσεων με στόχο μήκους (target_min/target_max σε λέξεις) και πρακτικές αποφυγής πλεονασμού. Επιπλέον, όταν ένα tag υπερβαίνει το threshold_words, ενεργοποιείται ιεραρχική (block-wise) ροή: το κείμενο χωρίζεται σε διαδοχικά blocks περίπου σταθερού μεγέθους, από κάθε block επιλέγονται λίγες κορυφαίες προτάσεις, και έπειτα οι επιλεγμένες προτάσεις επαναβαθμολογούνται συνολικά ώστε να εξαχθεί μία ενιαία σύνοψη με καλύτερη κάλυψη κατά μήκος του κειμένου. Αυτό είναι κρίσιμο ειδικά σε DESCRIPTION/CLAIMS όπου η απλή κατάταξη συχνά ευνοεί μόνο ένα τοπικό τμήμα.

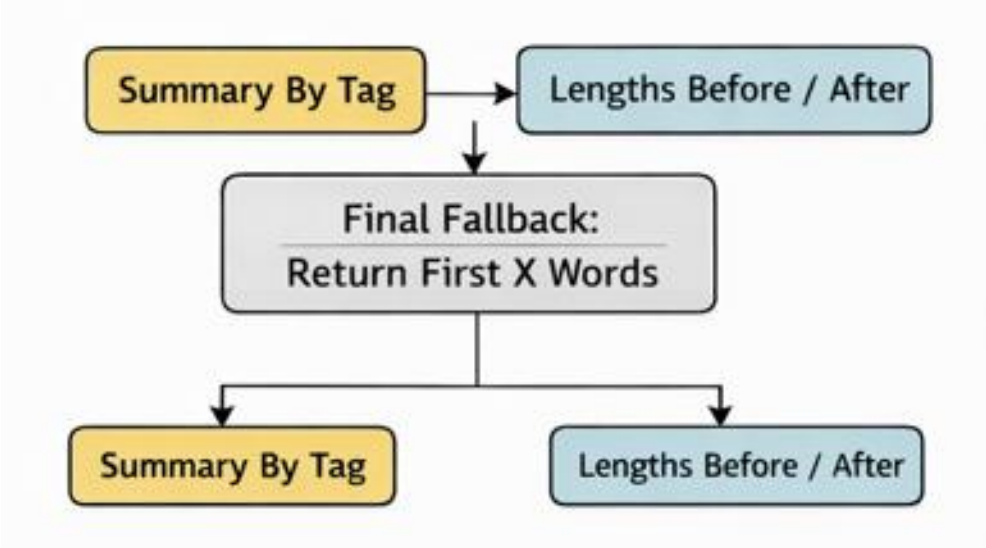
Η κατηγορία Hybrid στην παρούσα εργασία χρησιμοποιείται ως ομπρέλα για υλοποιήσεις που συνδυάζουν υποσυστήματα σύνοψης με δύο τρόπους. Πρώτον, περιλαμβάνει τις κλασικές υλοποιήσεις δύο σταδίων, όπου εφαρμόζεται αρχικά ένα στάδιο επιλογής/συμπίεσης περιεχομένου (extractive) και στη συνέχεια ένα γενετικό στάδιο (abstractive ή instructive) για αναδιατύπωση και συνοχή κειμένου. Δεύτερον, περιλαμβάνει και μια ειδική παραλλαγή "ενός σταδίου" που παραμένει extractive, αλλά διαφοροποιείται στον μηχανισμό βαθμολόγησης: αξιοποιεί Sentence-BERT embeddings για να υπολογίσει ομοιότητες μεταξύ προτάσεων, κατασκευάζει γράφο προτάσεων με βάρη ακμών ίσα με τις ομοιότητες, και εφαρμόζει PageRank/TextRank ώστε να εκτιμηθεί η κεντρικότητα κάθε πρότασης. Έτσι, το κριτήριο επιλογής μετακινείται από το "κοντά στο κεντροειδές" στο "κεντρική στο γράφο επειδή συνδέεται ισχυρά με άλλες σημαντικές προτάσεις". Για ανθεκτικότητα σε πραγματικά δεδομένα, ενσωματώνεται robust fallback: σε περιπτώσεις μη σύγκλισης ή εκφυλισμένου γράφου, η βαθμολόγηση υποχωρεί σε απλούστερη εκτίμηση (π.χ. άθροισμα ομοιοτήτων ανά πρόταση), ώστε να διατηρείται σταθερή παραγωγή σύνοψης. Όπως και στα υπόλοιπα per-tag σενάρια, η διαδικασία ελέγχεται από πολυσταδιακή "manager" λογική (κανονικός τμηματισμός, επιθετικότερη προετοιμασία όταν οι προτάσεις είναι προβληματικές, και τελικό fallback που εγγυάται μη-κενή έξοδο).

Στα abstractive και instructive scripts, το τεχνικό επίκεντρο μεταφέρεται στη δρομολόγηση του κειμένου προς generative μοντέλα και στον έλεγχο μήκους εξόδου. Στο abstractive σενάριο (Seq2Seq), για κάθε tag που περνά τα thresholds, το κείμενο καθαρίζεται και γίνεται tokenization με truncation σε προκαθορισμένο όριο (max_input_length_tokens), ώστε να αποφεύγονται out-of-memory ή υπερβολικοί χρόνοι σε ακραία μεγάλα πεδία. Η παραγωγή υλοποιείται με model.generate και παραμέτρους που κλειδώνουν το στυλ αποκωδικοποίησης (π.χ. beams, penalties επανάληψης, no-repeat ngram, early stopping). Κρίσιμο σημείο είναι ότι το script δεν αφήνει το μήκος εξόδου ελεύθερο: μετατρέπει τους στόχους των TAG_RULES σε δυναμικά όρια tokens και τα προσαρμόζει με βάση το πραγματικό μήκος εισόδου, ώστε να αποφεύγεται έξοδος που είναι δυσανάλογη ή συγκρίσιμη με το input. Μετά τη γενετική έξοδο ακολουθεί αποκωδικοποίηση και post-processing ώστε να αφαιρεθούν artifacts και να παραχθεί καθαρό κείμενο σύνοψης.

Το instructive σενάριο (Causal LM με instruction-following) διατηρεί την ίδια λογική ελέγχου εισόδου/εξόδου, αλλά αλλάζει ο τρόπος κατασκευής του input: για κάθε tag δημιουργείται prompt που καθοδηγεί ρητά το μοντέλο ως προς στόχο και ύφος, και το input διαμορφώνεται μέσω chat template/tokenizer. Υπάρχει επίσης βαλβίδα ασφαλείας στο tokenization: αν prompt και κείμενο ξεπεράσουν ένα σκληρό όριο context, εφαρμόζεται hard truncation πριν την παραγωγή. Η παραγωγή

ρυθμίζεται συνήθως για σταθερότητα (greedy decoding) με σαφές max_new_tokens και repetition penalty, ώστε να αποφεύγονται φαινόμενα επανάληψης ή υπερβολικής έκτασης.

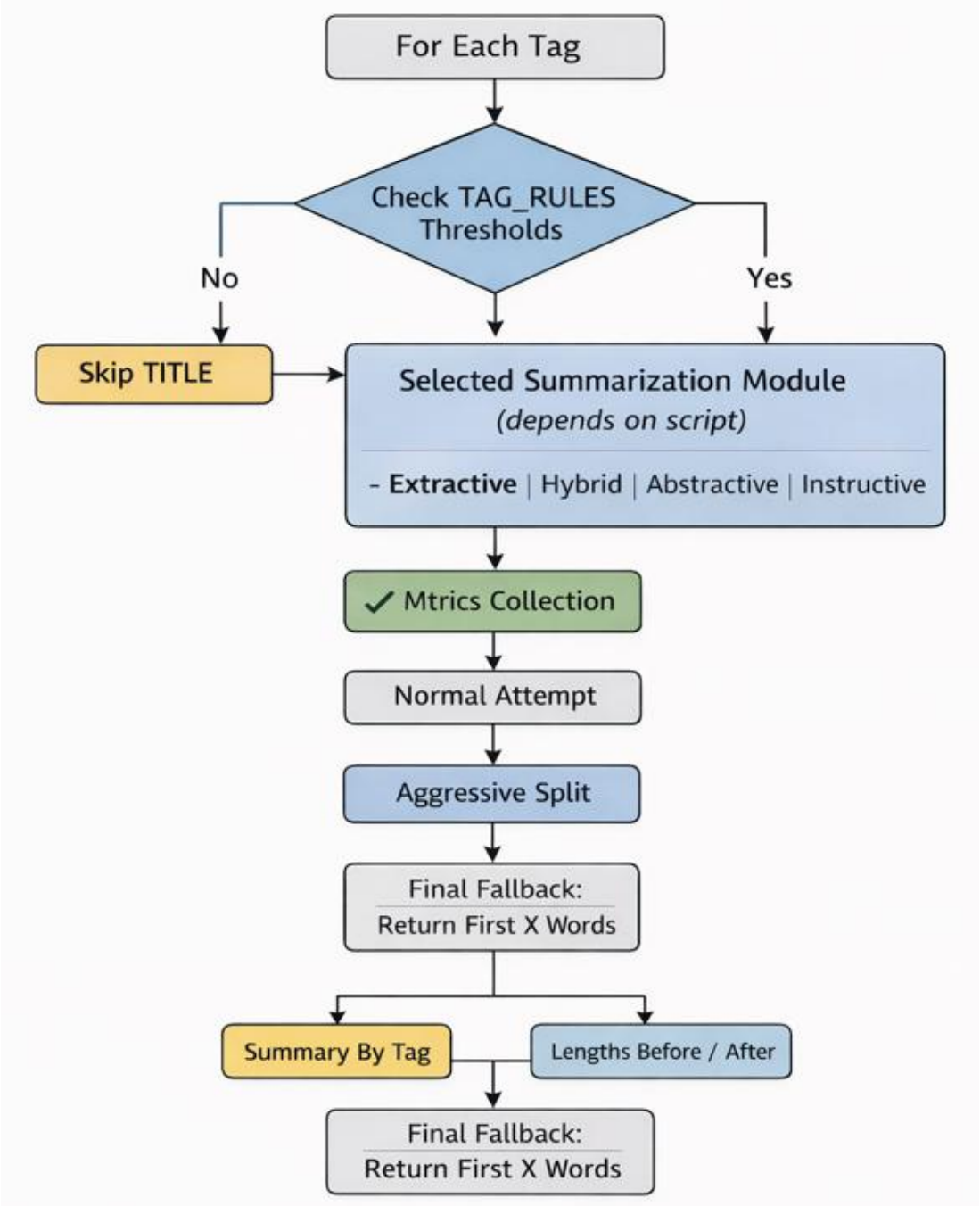
Σε επίπεδο υλοποίησης, εφαρμόζεται τελική δικλείδα ασφαλείας (final fallback) ώστε να εγγυάται ότι η έξοδος δεν είναι ποτέ κενή. Αν κάποιος μηχανισμός σύνοψης (extractive, generative ή hybrid) αποτύχει να επιστρέψει αξιοποιήσιμο κείμενο για ένα tag, το σύστημα επιστρέφει ένα ελάχιστο, σταθερό υποκατάστατο (π.χ. τις πρώτες X λέξεις του αρχικού κειμένου του tag). Παράλληλα, για κάθε tag καταγράφονται τα μήκη πριν και μετά (σε λέξεις ή/και tokens), ώστε να υπολογίζονται με συνέπεια δείκτες συμπίεσης (compression ratio/percent) και να διασφαλίζεται η συμμόρφωση με τους στόχους των TAG_RULES. Με αυτόν τον μηχανισμό, τα παραγόμενα artifacts διατηρούν πάντα πλήρη και προβλέψιμη δομή, χωρίς κενά πεδία που θα επηρέαζαν τα επόμενα στάδια της ροής.



Διάγραμμα 3: Τελική διαμόρφωση outputs ανά tag και μηχανισμός final fallback για εγγύηση μη-κενής σύνοψης.

Τέλος, ανεξάρτητα από την κατηγορία μεθόδου, η παραγωγή σύνοψης ολοκληρώνεται με ενιαίο τρόπο αποθήκευσης: δημιουργούνται δομές summaries_by_tag και μήκη πριν/μετά σε λέξεις και tokens, ενώ καταγράφονται τα metadata του run ώστε το αποτέλεσμα να είναι πλήρως ιχνηλάσιμο. Παράλληλα, στο ίδιο βήμα συλλέγονται δείκτες που συνδέονται άμεσα με την παραγωγή (π.χ. compression ratio/percent), καθώς και ελαφριά σημασιολογική αξιολόγηση (SBERT similarity) που λειτουργεί ως συνεπές σήμα σύγκρισης μεταξύ διαφορετικών runs.

Το παρακάτω Διάγραμμα 4 απεικονίζει την εννοιολογική δρομολόγηση ανά tag στο κοινό πλαίσιο εκτέλεσης. Στην πράξη, κάθε εκτέλεση αντιστοιχεί σε ξεχωριστό script (extractive, abstractive, instructive ή hybrid), το οποίο ενεργοποιεί έναν μόνο μηχανισμό παραγωγής στο στάδιο “Selected Summarization Module”, διατηρώντας σταθερά τα υπόλοιπα βήματα (TAG_RULES/thresholds, fallbacks, καταγραφή μετρικών και αποθήκευση). Οι μηχανισμοί fallback (normal attempt → aggressive split → final fallback) εξασφαλίζουν μη-κενή έξοδο και συνεπή δομή αποτελεσμάτων για όλα τα tags.



Διάγραμμα 4: Δρομολόγηση ανά tag με βάση TAG_RULES, thresholds και μηχανισμούς ανθεκτικότητας (fallback) για σταθερή παραγωγή σύνοψης.

Κεφάλαιο 5ο: Ευρετηρίαση και Αναζήτηση Πληροφορίας

Η αναζήτηση πληροφορίας σε συλλογές πατεντών αποτελεί ιδιαίτερα απαιτητική διαδικασία, λόγω του μεγάλου όγκου των εγγράφων, της τεχνικής τους πολυπλοκότητας και της εξειδικευμένης ορολογίας που περιέχουν. Στο πλαίσιο της παρούσας εργασίας, η διαδικασία της αναζήτησης δεν αντιμετωπίζεται ανεξάρτητα, αλλά ως άμεση συνέχεια του σταδίου της σύνοψης, αξιοποιώντας τις παραγόμενες περιλήψεις ως εναλλακτικές ή συμπληρωματικές αναπαραστάσεις των αρχικών εγγράφων.

Στο παρόν κεφάλαιο παρουσιάζονται οι τεχνικές ευρετηρίασης και αναζήτησης που εφαρμόστηκαν, καθώς και ο τρόπος με τον οποίο διαφορετικές μορφές σύνοψης επηρεάζουν την αποτελεσματικότητα της ανάκτησης πληροφορίας. Εξετάζονται τόσο κλασικές στατιστικές προσεγγίσεις όσο και σύγχρονες νευρωνικές μέθοδοι, με στόχο τη συγκριτική αξιολόγηση της συμπεριφοράς τους σε δεδομένα πατεντών.

5.1 Στόχος της αναζήτησης μετά τη σύνοψη

Ο βασικός στόχος της αναζήτησης πληροφορίας μετά το στάδιο της σύνοψης είναι η διερεύνηση του κατά πόσο οι παραγόμενες περιλήψεις μπορούν να λειτουργήσουν ως αποτελεσματικές αναπαραστάσεις των αρχικών εγγράφων πατεντών. Μέσω της σύνοψης επιδιώκεται η μείωση του μήκους και της πολυπλοκότητας των κειμένων, διατηρώντας παράλληλα τις κρίσιμες τεχνικές και σημασιολογικές πληροφορίες που είναι απαραίτητες για την επιτυχή ανάκτηση σχετικών εγγράφων.

Στο πλαίσιο αυτό, η αναζήτηση δεν αποσκοπεί μόνο στην ανάκτηση εγγράφων βάσει λέξεων-κλειδιών, αλλά στη μελέτη της επίδρασης διαφορετικών προσεγγίσεων σύνοψης στην ποιότητα των αποτελεσμάτων. Διαφορετικοί τύποι σύνοψης ενδέχεται να τονίζουν διαφορετικές πτυχές των πατεντών, γεγονός που μπορεί να επηρεάσει τόσο την ακρίβεια όσο και την πληρότητα της αναζήτησης.

Η διαδικασία της ευρετηρίασης και της αναζήτησης εφαρμόζεται σε σύνολα εγγράφων που προκύπτουν από τις διάφορες μεθόδους σύνοψης, καθώς και, όπου απαιτείται, στα αρχικά μη συνοψισμένα κείμενα, προκειμένου να καταστεί δυνατή η συγκριτική αξιολόγηση των αποτελεσμάτων. Με τον τρόπο αυτό, διερευνάται αν και σε ποιο βαθμό η σύνοψη μπορεί να βελτιώσει την αποδοτικότητα και τη διαχειρισιμότητα της αναζήτησης πατεντών.

5.2 BM25 με Pyserini (Κλασική Στατιστική Προσέγγιση)

Η πρώτη προσέγγιση ανάκτησης πληροφορίας που υιοθετήθηκε στην παρούσα εργασία βασίζεται στο κλασικό στατιστικό μοντέλο **BM25**, το οποίο αποτελεί μία από τις πιο διαδεδομένες και καθιερωμένες μεθόδους ανάκτησης εγγράφων σε μεγάλα κειμενικά σύνολα. Το BM25 αξιολογεί τη συνάφεια ενός εγγράφου ως προς ένα ερώτημα με βάση τη συχνότητα εμφάνισης των όρων, το μήκος του εγγράφου και τη συνολική κατανομή των όρων στο σύνολο δεδομένων, χωρίς να αξιοποιεί σημασιολογικές αναπαραστάσεις.

Για την υλοποίηση της μεθόδου χρησιμοποιήθηκε το πλαίσιο **Pyserini**, το οποίο αποτελεί Python διεπαφή του συστήματος Anserini και βασίζεται στη μηχανή αναζήτησης Lucene. Η επιλογή αυτή επιτρέπει την αποδοτική ευρετηρίαση και αναζήτηση μεγάλου αριθμού εγγράφων, διασφαλίζοντας ταυτόχρονα αναπαραγώγιμα και αξιόπιστα πειραματικά αποτελέσματα.

5.2.1 Ευρετηρίαση με BM25 σε επίπεδο τμημάτων (Indexing)

Η ευρετηρίαση με BM25 υλοποιείται μέσω Pyserini/Anserini (Lucene) και ακολουθεί μια "passage-based" λογική: τα μεγάλα πεδία της πατέντας (κυρίως DESCRIPTION και CLAIMS) τεμαχίζονται σε τμήματα σταθερού μεγέθους, τα οποία εισάγονται στο ευρετήριο ως ανεξάρτητα "documents" στο επίπεδο Lucene. Στόχος είναι να διατηρηθεί υψηλή ανακλητικότητα (recall) σε μεγάλα κείμενα, αποφεύγοντας το φαινόμενο να "χάνονται" σχετικά σημεία μέσα σε πολύ εκτενή έγγραφα.

Πριν από την κατασκευή του index, το script σταθεροποιεί το περιβάλλον εκτέλεσης εγκαθιστώντας Pyserini και διασφαλίζοντας συμβατό Java runtime (Java 21), καθώς η Java πλευρά του Lucene/Anserini είναι κρίσιμη για αξιόπιστη λειτουργία σε Colab. Επιπλέον γίνεται ρητή ρύθμιση μεταβλητών περιβάλλοντος (JAVA_HOME και JVM_PATH προς libjvm.so) ώστε να αποφεύγονται σφάλματα φόρτωσης της JVM από το pyjnius.

Στο στάδιο προετοιμασίας δεδομένων, κάθε SGML-like αρχείο (.txt) γίνεται parsing με ασφαλή χειρισμό ειδικών χαρακτήρων και περιεχομένου: εφαρμόζεται escaping στο σύμβολο "&" όταν δεν αποτελεί ήδη έγκυρο entity και γίνεται wrapping των κειμενικών tags (TITLE, ABSTRACT, DESCRIPTION, CLAIMS) σε CDATA ώστε το XML parsing να είναι ανθεκτικό σε "θορυβώδη" πραγματικά δεδομένα. Από κάθε αρχείο εξάγονται τα πεδία TITLE, ABSTRACT, DESCRIPTION και CLAIMS, με το doc_id να αντιστοιχεί στο stem του filename.

Ο τεμαχισμός (chunking) εφαρμόζεται σε επίπεδο λέξεων με παραμετροποίηση στόχου μήκους και επικάλυψης: κάθε chunk έχει περίπου 220 λέξεις με επικάλυψη 15 λέξεων, ώστε να διατηρείται συνέχεια συμφραζομένων μεταξύ διαδοχικών τμημάτων. Δεν επιβάλλεται όριο πλήθους chunks (MAX_CHUNKS = None), άρα για DESCRIPTION/CLAIMS λαμβάνονται όλα τα τμήματα που προκύπτουν. Για κάθε chunk παράγεται μοναδικό αναγνωριστικό μορφής doc_id#DESC_### ή doc_id#CLMS_###.

Κεντρική επιλογή της στρατηγικής είναι ότι κάθε chunk εμπλουτίζεται με σταθερό "context prefix" από TITLE+ABSTRACT (όταν υπάρχουν), το οποίο επαναλαμβάνεται σε όλα τα chunks του ίδιου εγγράφου. Έτσι, ακόμη και όταν το retrieval "πέφτει" σε passage του DESCRIPTION ή των CLAIMS, το Lucene scoring βλέπει μαζί και τα πιο διακριτικά πεδία (TITLE/ABSTRACT), αυξάνοντας την πιθανότητα να ευνοηθούν passages που συνδέονται θεματικά με τον πυρήνα του εγγράφου. Αν δεν παραχθεί κανένα chunk (π.χ. κενά DESCRIPTION/CLAIMS), το script δημιουργεί fallback εγγραφή #BASIC_000 με βάση το διαθέσιμο TITLE+ABSTRACT ώστε να μην αποκλείεται πλήρως το έγγραφο από το index.

Για να τροφοδοτηθεί ο Pyserini indexer, το script μετατρέπει το corpus σε συλλογή JSONL τύπου {"id":..., "contents":...} μέσα σε υποφάκελο jsonl_docs/. Η παραγωγή γίνεται σε batches (π.χ. υποσύνολα 10.000 αρχείων) με progress reporting και logging ανά batch (chunks που παράχθηκαν, χρόνος), ώστε να είναι ελέγξιμη και επαναλήψιμη η διαδικασία.

Η κατασκευή του Lucene index γίνεται με pyserini.index.lucene πάνω σε JsonCollection, με ενεργοποιημένη αποθήκευση θέσεων/διανυσμάτων όρων και raw κειμένου (storePositions, storeDocvectors, storeRaw). Η επιλογή αυτή διευκολύνει μεταγενέστερη ανάλυση/διαγνωστικό έλεγχο και διατηρεί το index "πλούσιο" ως προς την πληροφορία που κρατά. Το index παράγεται αρχικά σε τοπικό προσωρινό path στο Colab (για ταχύτητα I/O) και στη συνέχεια μεταφέρεται στο Google Drive με rsync, ώστε να διασφαλιστεί ακεραιότητα και πλήρης αντιγραφή της δομής.

Τέλος, για λόγους αναπαραγωγιμότητας, αποθηκεύεται snapshot των εξαρτήσεων (pip freeze) ως ξεχωριστό αρχείο requirements με timestamp, και εφαρμόζεται "safe shutdown" στρατηγική (sync + αναμονή) ώστε να μειωθεί ο κίνδυνος απώλειας δεδομένων λόγω καθυστέρησης συγχρονισμού Drive.

5.2.2 Διαδικασία Ανάκτησης (Retrieval & Aggregation)

Η ανάκτηση με BM25 εκτελείται μέσω LuceneSearcher του Pyserini πάνω στο passage-based index. Η εκτέλεση είναι οργανωμένη ως πειραματικό loop που μπορεί να τρέχει για διαφορετικά σύνολα εισόδου (π.χ. summarized corpus ανά μοντέλο σύνοψης), κρατώντας κοινή μορφή outputs και κοινό σύστημα καταγραφής. Στο retrieval στάδιο, το κρίσιμο τεχνικό ζήτημα είναι ότι τα αποτελέσματα του Lucene επιστρέφονται σε επίπεδο chunks, ενώ η αξιολόγηση (qrels) και η σύγκριση των runs είναι σε επίπεδο "base document". Για αυτό ενσωματώνεται ρητά document-level aggregation.

Πριν την ανάκτηση, τα query κείμενα κατασκευάζονται από SGML-like topics με ελεγχόμενη σύνθεση πεδίων. Το script επιτρέπει να οριστεί σειρά ένταξης tags (INCLUDE_ORDER), σύνολο εξαιρέσεων (EXCLUDE_TAGS), προαιρετικά τοπικά caps ανά tag (TAG_CAP) και συνολικό budget λέξεων (TOTAL_WORD_BUDGET). Έτσι, κάθε query γίνεται "συνεπές" ως προς το μέγιστο μήκος και τη συμβολή κάθε ενότητας (π.χ. TITLE, ABSTRACT, DESCRIPTION, CLAIMS), ενώ καταγράφονται και στατιστικά μήκους (μέσος/ελάχιστος/μέγιστος αριθμός λέξεων μετά την εφαρμογή των caps).

Η αναζήτηση εκτελείται ως BM25 retrieval σε επίπεδο chunks, αλλά το script λαμβάνει ειδική πρόνοια για μεγάλα queries: αν προκύψει σφάλμα λόγω υπερβολικού μήκους, εφαρμόζεται truncation (π.χ. έως ~900 tokens/λέξεις με βάση απλό tokenization) και επαναλαμβάνεται η αναζήτηση. Αυτό λειτουργεί ως μηχανισμός ανθεκτικότητας ώστε η εκτέλεση να μην αποτυγχάνει σε "βαριά" queries.

Η συγκέντρωση αποτελεσμάτων (aggregation) μετατρέπει τη λίστα (chunk_id, score) σε (doc_id, aggregated_score) ομαδοποιώντας όλα τα chunks που ανήκουν στο ίδιο base document (αφαίρεση του suffix μετά το #). Υποστηρίζονται διαφορετικές στρατηγικές aggregation (max/sum/avg), με default επιλογή "max" (MaxP) ως πρακτική προσέγγιση passage retrieval: ένα έγγραφο θεωρείται ισχυρό αν διαθέτει έστω ένα πολύ ισχυρό passage. Μετά το aggregation εφαρμόζεται ταξινόμηση κατά score και παράγεται η τελική λίστα εγγράφων ανά query.

Για να διασφαλιστεί ότι μετά το aggregation υπάρχει επαρκής αριθμός μοναδικών εγγράφων (ιδίως όταν πολλά hits προέρχονται από πολλά chunks του ίδιου doc), το script δεν ζητά απλώς index_k hits σε chunk-level. Αντίθετα, αυξάνει προσαρμοστικά το πλήθος των chunk hits (π.χ. index_k * mult για πολλαπλά mult), μέχρι να συγκεντρωθούν τουλάχιστον index_k μοναδικά doc-level αποτελέσματα ή μέχρι να εξαντληθούν τα όρια. Έτσι, σταθεροποιείται η ποσότητα αξιολογήσιμων αποτελεσμάτων ανά query.

Η εκτέλεση επιταχύνεται με παράλληλη επεξεργασία queries μέσω threading backend (joblib), αξιοποιώντας ότι η Java πλευρά του Pyserini δεν "μπλοκάρει" με τον ίδιο τρόπο όπως καθαρός Python υπολογισμός. Τα αποτελέσματα συλλέγονται σε λίστες doc_ids και scores ανά query, και στη συνέχεια εφαρμόζεται re-sort και αφαίρεση duplicates ως επιπλέον ασφάλεια πριν από την αποθήκευση.

Στο στάδιο αξιολόγησης, το script υπολογίζει μετρικές σε πολλαπλά cut-offs (π.χ. @10/@50/@100) με macro averaging, συμπεριλαμβάνοντας precision, recall, f1, καθώς και δείκτες όπως MAP, nDCG και PRESS. Παράγονται αρχεία αποτελεσμάτων σε δομές κατάλληλες για downstream αξιολόγηση (π.χ. JSON σε μορφή pytrec_eval input) και δημιουργούνται logs που καταγράφουν πλήρως τις παραμέτρους του run (dataset, budgets, index_k/SAVE_K, AGG_MODE, πλήθος queries, runtime).

Ένα επιπλέον τεχνικό χαρακτηριστικό είναι ότι η ανάκτηση συνδέεται με τα στατιστικά σύνοψης: προφορτώνονται flat metrics της σύνοψης ανά (docno, tag) και υπολογίζονται σταθμισμένοι μέσοι όροι

(weighted averages) συμπίεσης και SBERT similarity ανά query, με βάρη τις λέξεις που συνεισέφερε κάθε tag στο query. Με αυτόν τον τρόπο, το retrieval run "κουβαλά" μαζί του συνοπτικές πληροφορίες για το πόσο συμπιεσμένο/σημασιολογικά συνεπές ήταν το υλικό που χρησιμοποιήθηκε ως query, επιτρέποντας ενιαία καταγραφή σε επίπεδο πειραμάτων. Τέλος, τα συγκεντρωτικά αποτελέσματα γίνονται append σε κεντρικό Excel (μία γραμμή ανά run) ώστε να υποστηριχθεί η μεταγενέστερη ανάλυση και σύγκριση στο Κεφάλαιο 7.

Η χρήση του BM25 με τον παραπάνω τρόπο λειτουργεί ως ισχυρό σημείο αναφοράς (baseline) για τη σύγκριση με τις νευρωνικές προσεγγίσεις ανάκτησης που παρουσιάζονται στην επόμενη ενότητα. Παρά τα πλεονεκτήματά του, το BM25 παραμένει εγγενώς εξαρτημένο από τη λεξιλογική επικάλυψη μεταξύ ερωτήματος και εγγράφου, γεγονός που περιορίζει την ικανότητά του να συλλαμβάνει σημασιολογικές σχέσεις και παραφράσεις.

Για τον λόγο αυτό, στην επόμενη ενότητα παρουσιάζεται η νευρωνική ανάκτηση πληροφορίας με χρήση της αρχιτεκτονικής ColBERT, η οποία επιχειρεί να αντιμετωπίσει τους περιορισμούς των στατιστικών μεθόδων μέσω πυκνών σημασιολογικών αναπαραστάσεων.

5.3 Νευρωνική Ανάκτηση Πληροφορίας (ColBERT / Late-Interaction Retrieval)

Η νευρωνική ανάκτηση πληροφορίας αποτελεί τη σύγχρονη εξέλιξη των κλασικών στατιστικών μεθόδων, αξιοποιώντας μοντέλα βαθιάς μάθησης για να αναπαριστούν ερωτήματα και έγγραφα σε διανυσματικούς χώρους. Σε αντίθεση με προσεγγίσεις όπως ο BM25, όπου η αντιστοίχιση βασίζεται κυρίως σε επιφανειακή επικάλυψη όρων, οι νευρωνικές μέθοδοι στοχεύουν στη σύλληψη σημασιολογικών σχέσεων μεταξύ κειμένων, ακόμη και όταν δεν υπάρχει ακριβής λεξιλογική ταύτιση. Στην παρούσα εργασία η νευρωνική ανάκτηση δεν υλοποιείται ως "ένα ενιαίο embedding ανά κείμενο" (τυπικό dense bi-encoder), αλλά ως late-interaction retrieval με αντιστοίχιση σε επίπεδο tokens.

Συγκεκριμένα, υιοθετείται η αρχιτεκτονική ColBERT (Contextualized Late Interaction over BERT), η οποία συνδυάζει contextualized embeddings ανά token με αποδοτική αναζήτηση μεγάλης κλίμακας. Κάθε έγγραφο και κάθε ερώτημα κωδικοποιείται σε σύνολο token embeddings και η συνάφεια υπολογίζεται μέσω late interaction (π.χ. MaxSim/aggregation), επιτρέποντας λεπτομερή αντιστοίχιση λέξεων/φράσεων και όχι μόνο σύγκριση δύο συνολικών διανυσμάτων.

5.3.1 Δημιουργία Νευρωνικού Ευρετηρίου (Indexing)

Η ευρετηρίαση ColBERT δημιουργεί νευρωνικό index τύπου Voyager (pylate.indexes.Voyager), όπου κάθε indexed item αντιστοιχεί σε chunk embedding ColBERT v2 (embedding_size=128). Το script οργανώνει το run με timestamp Europe/Athens, χτίζει τον index τοπικά (για I/O ταχύτητα) και στο τέλος μεταφέρει το build στο Drive σε ξεχωριστό φάκελο run.

Η προετοιμασία των κειμένων για indexing ακολουθεί την ίδια λογική SGML parsing με ασφαλή escape/CDATA, με στόχο να εξαχθούν TITLE/ABSTRACT/DESCRIPTION/CLAIMS χωρίς σφάλματα και με σταθερή συμπεριφορά σε θορυβώδη περιεχόμενα.

Στο ColBERT indexing, το κείμενο δεν εισάγεται αυτούσιο ως ένα μεγάλο document: εφαρμόζεται chunking σε DESCRIPTION και CLAIMS με στόχο λέξεων και overlap, ώστε κάθε chunk να χωράει στο document_length του ColBERT (π.χ. 256 tokens). Επιπλέον, εφαρμόζονται "ασφαλιστικές δικλείδες" με MAX_CHUNKS_DESC/MAX_CHUNKS_CLAIMS ώστε να αποφεύγονται ακραίες περιπτώσεις υπερβολικά πολλών chunks ανά έγγραφο.

Κάθε chunk εμπλουτίζεται προαιρετικά με ελαφρύ context (TITLE + ABSTRACT) ως prefix, ώστε τα επιμέρους passages να διατηρούν βασικό θεματικό πλαίσιο.

Η παραγωγή embeddings γίνεται με ColBERT encode σε batch mode, με is_query=False για document embeddings και ρύθμιση batch_size (π.χ. 2048) ώστε να αξιοποιείται αποδοτικά η συσκευή και να μειώνεται το overhead. Τα embeddings μετατρέπονται σε float32 πριν εισαχθούν στον index και το indexing εκτελείται τμηματικά σε batches αρχείων (π.χ. ανά 10.000) με καταγραφή Added/Skipped και χρόνων ανά batch.

5.3.2 Διαδικασία Ανάκτησης (Retrieval)

Το retrieval script ColBERT ακολουθεί ενιαία λογική "ιχνηλασιμότητας run" και αναπαραγωγιμότητας: κάθε εκτέλεση δημιουργεί timestamped φάκελο αποτελεσμάτων, καταγράφει snapshot περιβάλλοντος (εκδόσεις βιβλιοθηκών, CPU/GPU, CUDA), και γράφει αναλυτικό log με τις τελικές ρυθμίσεις (paths, flags, cut-offs, vector-search παραμέτρους), ώστε διαφορετικά runs να είναι άμεσα συγκρίσιμα. Επιπλέον υποστηρίζει επιλογή dataset εισόδου (π.χ. παλαιό corpus έναντι summarized corpus ή εκτέλεση και στα δύο), επιτρέποντας να διατηρείται σταθερή η υπόλοιπη ροή και να απομονώνεται το effect της σύνοψης στην ανάκτηση.

Σύνθεση query κειμένου (tag-based query building)

Η ανάκτηση ξεκινά από SGML-like queries, όπου το τελικό query-text δεν θεωρείται μονολιθικό αλλά συναρμολογείται από επιλεγμένα tags (π.χ. TITLE, ABSTRACT, DESCRIPTION κ.λπ.). Η σύνθεση ελέγχεται από ρυθμίσεις INCLUDE_ORDER/EXCLUDE_TAGS, καθώς και από TAG_CAP (όριο λέξεων ανά tag) και TOTAL_WORD_BUDGET (συνολικό όριο λέξεων), ώστε η είσοδος να μένει εντός προκαθορισμένου μεγέθους και να μπορεί να στοχεύει πειραματικά διαφορετικές "όψεις" του query (π.χ. πιο περιγραφικό ή πιο συνοπτικό). Το script κρατά επίσης "contributions map" (πόσες λέξεις συνέβαλε κάθε tag), το οποίο αξιοποιείται αργότερα για σταθμισμένη συσχέτιση με summarization-metrics.

Query embeddings και batch retrieval στον Voyager index

Τα queries μετατρέπονται σε embeddings με ColBERT encoding σε λειτουργία query (is_query=True), ώστε να είναι συμβατά με τον μηχανισμό late interaction. Η διαδικασία γίνεται batch-wise (encoding και search σε batches), για να επιτευχθεί σταθερό throughput, και η κλήση προς τον Voyager index εκτελεί HNSW αναζήτηση με παραμέτρους όπως ef_search και index_k (πόσους υποψηφίους ανασύρει το Stage-1). Στη λογική του script διαχωρίζεται σαφώς το "δίχτυ" από τη "σακούλα": το index_k καθορίζει πόσο βαθιά ψάχνει το Stage-1 (άρα το μέγιστο θεωρητικό recall), ενώ το SAVE_K καθορίζει πόσα τελικά αποτελέσματα αποθηκεύονται/αξιολογούνται (π.χ. για MAP@k, nDCG@k).

Stage-1: token-level hits και aggregation σε document-level κατάταξη

Ο Voyager επιστρέφει token-level αντιστοιχίσεις (λίστες εγγράφων και scores ανά token). Το script συγκεντρώνει αυτά τα hits σε προσωρινά document scores (Stage-1 aggregation) με επιλογή AGG_MODE:

- ***raw_sum_desc: άθροισμα όλων των token-level scores ανά έγγραφο (συχνά ευνοεί recall).***
- ***max_sum_desc: για κάθε token κρατά το καλύτερο hit ανά έγγραφο και έπειτα αθροίζει (πιο "συγκρατημένη" συγκέντρωση).***

Επειδή το indexing είναι chunk-based, το ίδιο doc_id μπορεί να εμφανίζεται πολλαπλές φορές (μέσω διαφορετικών chunk ids). Για να γίνει αξιολόγηση σε "document-level" και όχι σε "chunk-level",

εφαρμόζεται δεύτερη συμπύκνωση από chunks σε base-docs (base_id), με πολιτικές όπως sum/max/mean/topn_sum και προαιρετική κανονικοποίηση (π.χ. /√N στα topn_sum), ώστε να ελέγχεται η επίδραση του πλήθους chunks στο τελικό score. Αυτό το βήμα είναι κρίσιμο για να μην υπερεκπροσωπούνται μεγάλα έγγραφα λόγω πολλών chunks και για να διατηρείται συνεπές ranking στο επίπεδο πατέντας.

Stage-2: per-token reranking σε περιορισμένο σύνολο υποψηφίων

Πάνω στη doc-level λίστα του Stage-1 εφαρμόζεται Stage-2 reranking σε περιορισμένο πλήθος υποψηφίων (M=STAGE2_M). Η λογική είναι "per-token best accumulation": για κάθε query token κρατείται το καλύτερο (μέγιστο) score ανά έγγραφο μέσα στους top-M υποψηφίους, και αυτά τα per-token best scores αθροίζονται ώστε να παραχθεί τελικό rerank score. Με αυτόν τον περιορισμό (top-M) ο υπολογιστικός φόρτος παραμένει ελεγχόμενος, ενώ το ranking γίνεται πιο "token-faithful" χωρίς να απαιτείται πλήρης επεξεργασία σε όλο το candidate set. Αν παρουσιαστεί εκφυλισμός/κενή είσοδος, υπάρχει fallback στο Stage-1 order ώστε να μην αποτυγχάνει η εκτέλεση.

Post-processing: dedup, ασφαλής ταξινόμηση και ακριβές top-k

Μετά το Stage-2, το script εφαρμόζει (α) re-sort για ασφάλεια, (β) deduplication κρατώντας την πρώτη εμφάνιση (υψηλότερο score), και (γ) slicing σε top_k για την τελική λίστα αποτελεσμάτων. Αυτό είναι απαραίτητο ώστε τα metrics να υπολογίζονται αυστηρά πάνω στα k καλύτερα documents και να μην αλλοιώνονται από διπλοεγγραφές ή από μη-σταθερή σειρά επιστροφής.

Προαιρετικό FILTER_LABEL: αναδιάταξη αποτελεσμάτων χωρίς αφαίρεση υποψηφίων

Το script υποστηρίζει προαιρετικό FILTER_LABEL, το οποίο δεν "πετάει" έγγραφα, αλλά κάνει reordering: προωθεί πρώτα τα αποτελέσματα που μοιράζονται το ίδιο label με το query και έπειτα συμπληρώνει με τα υπόλοιπα, ώστε να μην μειώνεται το πλήθος αποτελεσμάτων. Αυτό το design επιτρέπει πειράματα "label-aware ranking bias" χωρίς να αλλάζει το recall ceiling που δίνει το Stage-1.

Metrics, αποθήκευση outputs και σύνδεση με summarization statistics

Η αξιολόγηση γίνεται σε πολλαπλά cut-offs (EVAL_KS όπως 10/50/100) και περιλαμβάνει macro Precision/Recall/F1, καθώς και macro MAP@k και nDCG@k. Επιπλέον υπολογίζεται PRESS@k πάνω σε καλιμπραρισμένα scores (sigmoid), ώστε να αποτυπώνεται σφάλμα πρόβλεψης σε μορφή loss-like μέτρου. Τα αποτελέσματα αποθηκεύονται σε structured JSON (π.χ. pytrec_eval input, συνολικά metrics, per-query metrics) και συνοδεύονται από log με όλες τις παραμέτρους του run.

Τέλος, για runs που βασίζονται σε summarized corpus, το retrieval script μπορεί να "φορτώσει" summarization stats (compression ratio, SBERT-F1 ανά (docno, tag)) και να υπολογίσει σταθμισμένους μέσους όρους ανά query, χρησιμοποιώντας ως βάρη τη συνεισφορά (λέξεις) κάθε tag στο query-text. Με αυτόν τον τρόπο, στο ίδιο artifact του retrieval run συνυπάρχουν τόσο retrieval metrics (P/R/MAP/nDCG/PRESS) όσο και συνοπτικά signals ποιότητας σύνοψης,

5.4 Διαφορετικές στρατηγικές ευρετηρίασης (indexing)

Η ευρετηρίαση πατεντών αποτελεί κρίσιμο στάδιο, καθώς οι πατέντες είναι εκτενή έγγραφα και συχνά μόνο ένα μικρό υποτμήμα τους είναι άμεσα σχετικό με ένα ερώτημα. Για τον λόγο αυτό, στην παρούσα εργασία υλοποιήθηκαν και δοκιμάστηκαν πολλαπλές στρατηγικές δημιουργίας index για το ίδιο σύνολο δεδομένων και για BM25 και για Colbert με στόχο να διερευνηθεί η επίδραση (α) του επιπέδου τμηματοποίησης (document-level έναντι chunk-level), (β) της χρήσης συμφραζομένων (π.χ. προσθήκη

TITLE/ABSTRACT ως context) και (γ) των παραμέτρων chunking (μέγεθος, overlap, ανώτατο πλήθος chunks) στην αποτελεσματικότητα της ανάκτησης.

5.4.1 Στρατηγικές ευρετηρίασης (indexing) για BM25 (Pyserini/Lucene)

Στο πλαίσιο της στατιστικής ανάκτησης πληροφορίας (BM25), δοκιμάστηκαν διαφορετικές στρατηγικές ευρετηρίασης με χρήση Pyserini/Lucene, με στόχο να διερευνηθεί η επίδραση της τμηματοποίησης (chunking), της ποσότητας κειμένου που εισάγεται στο index και του τρόπου ενσωμάτωσης των βασικών πεδίων (TITLE/ABSTRACT) στη συνολική απόδοση ανάκτησης. Και στις τρεις εκδοχές, τα έγγραφα πατεντών διαβάζονται από SGML-like αρχεία, εξάγονται τα πεδία TITLE, ABSTRACT, DESCRIPTION και CLAIMS, γίνεται τμηματοποίηση σε επικαλυπτόμενα chunks λέξεων και παράγεται συλλογή JSONL, η οποία ευρετηριάζεται σε Lucene index.

5.4.1.1 Στρατηγική 1: Πλήρες chunking DESCRIPTION/CLAIMS με prefix TITLE+ABSTRACT σε κάθε chunk

Η πρώτη στρατηγική εφαρμόζει πλήρη passage-based ευρετηρίαση: τα πεδία DESCRIPTION και CLAIMS τεμαχίζονται σε chunks σταθερού μεγέθους (TARGET_WORDS=220) με επικάλυψη (OVERLAP_WORDS=15) χωρίς ανώτατο όριο στον αριθμό chunks ανά πεδίο (MAX_CHUNKS_DESC=None, MAX_CHUNKS_CLAIMS=None). Κάθε chunk ευρετηριάζεται ως ξεχωριστό Lucene document, με μοναδικό chunk_id (π.χ. doc_id#DESC_###, doc_id#CLMS_###), ενώ το base doc_id παραμένει ανακτήσιμο μέσω του prefix πριν το #, ώστε στο retrieval να γίνει aggregation σε document-level.

Κρίσιμο στοιχείο είναι το prefixing του TITLE+ABSTRACT ως base_context σε κάθε chunk DESCRIPTION/CLAIMS. Τεχνικά, αυτό "αντιγράφει" στο επίπεδο κάθε passage τα πιο σημασιολογικά/διακριτικά πεδία, αυξάνοντας την πιθανότητα οι όροι του query να ταιριάξουν με το chunk ακόμη και όταν η σχετική πληροφορία στο DESCRIPTION/CLAIMS είναι διάσπαρτη. Η στρατηγική τείνει να μεγιστοποιεί recall σε μεγάλα έγγραφα, αλλά αυξάνει τον όγκο του index (πολλά passages) και εισάγει έντονο redundancy (επαναλαμβανόμενο TA σε όλα τα chunks), που μπορεί να επιβαρύνει χώρο και χρόνο αναζήτησης.

5.4.1.2 Στρατηγική 2: Περιορισμένο chunking (έως 2 chunks ανά πεδίο) με prefix TITLE+ABSTRACT

Η δεύτερη στρατηγική διατηρεί την ίδια λογική chunking και το ίδιο prefixing TITLE+ABSTRACT σε κάθε chunk, αλλά επιβάλλει αυστηρό όριο στον αριθμό passages που εισάγονται στο index: έως 2 chunks για DESCRIPTION και έως 2 chunks για CLAIMS (MAX_CHUNKS_DESC=2, MAX_CHUNKS_CLAIMS=2).

Η τεχνική στόχευση είναι να μειωθεί δραστικά το πλήθος των indexed μονάδων, άρα και ο όγκος του Lucene index, βελτιώνοντας αποδοτικότητα σε χρόνο/χώρο και περιορίζοντας θόρυβο από υπερβολικά πολλά passages. Το trade-off είναι ότι το recall μπορεί να μειωθεί όταν η σχετική πληροφορία βρίσκεται "βαθύτερα" στο κείμενο και δεν καλύπτεται από τα πρώτα λίγα chunks. Σε αυτή την περίπτωση, ακόμη και αν το aggregation λειτουργεί σωστά, το σύστημα δεν έχει πρόσβαση σε passages που δεν έχουν ευρετηριαστεί, άρα το ceiling της ανάκτησης περιορίζεται από τη στρατηγική index coverage.

5.4.1.3 Στρατηγική 3: Διακριτό chunk TITLE+ABSTRACT και ανεξάρτητα chunks DESCRIPTION/CLAIMS (χωρίς prefix)

Η τρίτη στρατηγική αλλάζει τη φιλοσοφία οργάνωσης στο index, αποφεύγοντας την επανάληψη του TITLE+ABSTRACT σε κάθε passage. Δημιουργείται πρώτα ένα ξεχωριστό chunk μόνο για TITLE+ABSTRACT (π.χ. doc_id#TA_000, με MAX_CHUNKS_TA=1) και στη συνέχεια

δημιουργούνται chunks από DESCRIPTION και CLAIMS χωρίς prefix, συνήθως με όρια έως 2 chunks ανά πεδίο (MAX_CHUNKS_DESC=2, MAX_CHUNKS_CLAIMS=2).

Ο διαχωρισμός αυτός προσεγγίζει "field-like" συμπεριφορά σε passage indexing: το TA chunk λειτουργεί ως συνοπτική υψηλού επιπέδου αναπαράσταση της πατέντας, ενώ τα DESC/CLMS chunks παραμένουν καθαρά τεχνικά passages. Τεχνικά οφέλη είναι η μείωση redundancy (μικρότερη επανάληψη TA) και η καλύτερη στόχευση του matching ανάλογα με το είδος του query (high-level queries μπορεί να ταιριάξουν κυρίως στο TA chunk, ενώ τεχνικά queries σε DESC/CLMS). Ωστόσο, επειδή το TITLE/ABSTRACT δεν "ενισχύει" πλέον όλα τα passages μέσω prefix, ορισμένα passages στα μεγάλα πεδία μπορεί να χάσουν υποστηρικτικό συμφραζόμενο για BM25 scoring, άρα η συμπεριφορά εξαρτάται περισσότερο από την καθαρή λεξιλογική επικάλυψη μέσα στο ίδιο chunk.

5.4.2 Στρατηγικές ευρετηρίασης για ColBERT (Neural / Late-Interaction Indexing)

5.4.2.1 Στρατηγική 1: Ευρετηρίαση μόνο με TITLE + ABSTRACT (document-level)

Η πρώτη στρατηγική υλοποιεί την πιο συνοπτική αναπαράσταση της πατέντας στο νευρωνικό ευρετήριο, εισάγοντας αποκλειστικά τα πεδία TITLE και ABSTRACT ως ένα ενιαίο κείμενο. Με τον τρόπο αυτό, η μονάδα ανάκτησης (retrieval unit) ταυτίζεται πρακτικά με το έγγραφο και δεν απαιτείται chunk-level ανασύνθεση, άρα το ranking λειτουργεί εξαρχής σε document-level.

Σε επίπεδο ColBERT, η επιλογή TITLE+ABSTRACT είναι ιδιαίτερα "καθαρή" επειδή (α) περιορίζει θόρυβο από εκτενή sections και (β) εκμεταλλεύεται πλήρως τον διαθέσιμο token budget για περιεχόμενο υψηλής διακριτικότητας. Η στρατηγική λειτουργεί ως baseline, όμως μπορεί να μειώσει την ικανότητα ανάκτησης λεπτομερών τεχνικών στοιχείων που εμφανίζονται μόνο σε DESCRIPTION/CLAIMS, ιδίως όταν τα queries περιέχουν εξειδικευμένους όρους ή περιγραφές διαδικασιών που δεν συνοψίζονται στο abstract.

5.4.2.2 Στρατηγική 2: Chunking σε DESCRIPTION/CLAIMS με prefix TITLE+ABSTRACT ως συμφραζόμενο

Η δεύτερη στρατηγική μεταβαίνει σε chunk-level indexing για τα μεγάλα πεδία DESCRIPTION και CLAIMS, ώστε να αντιμετωπιστεί ο περιορισμός μέγιστου μήκους εισόδου ανά μονάδα ευρετηρίασης (π.χ. document_length=256 tokens). Τα chunks παράγονται με παράθυρα λέξεων (TARGET_WORDS=220) και επικάλυψη (OVERLAP_WORDS=15), ενώ εφαρμόζονται όρια πλήθους chunks ανά ενότητα (MAX_CHUNKS_DESC=2, MAX_CHUNKS_CLAIMS=2) για έλεγχο μεγέθους index και χρόνου retrieval.

Κεντρικό χαρακτηριστικό είναι το prefixing: κάθε chunk των DESCRIPTION/CLAIMS εμπλουτίζεται με base_context από TITLE+ABSTRACT ως prefix. Τεχνικά, αυτό αυξάνει την πιθανότητα να επιβιώσουν στο token-level matching όροι "υψηλού επιπέδου" που περιγράφουν το θέμα της πατέντας, όμως ανταλλάσσεται με κόστος: μέρος του token budget του chunk καταναλώνεται από το prefix, άρα μειώνεται ο χώρος για "καθαρό" περιεχόμενο DESCRIPTION/CLAIMS μέσα στο ίδιο chunk. Στο retrieval, επειδή η μονάδα ανάκτησης είναι το chunk, τα αποτελέσματα απαιτούν document-level aggregation (chunk hits → base doc), ώστε να παραχθεί τελικό ranking ανά πατέντα.

5.4.2.3 Στρατηγική 3: Ξεχωριστό chunk TITLE_ABSTRACT και περιορισμένα chunks για DESCRIPTION/CLAIMS χωρίς overlap

Η τρίτη στρατηγική διαφοροποιείται σε δύο σημεία: (α) δημιουργεί ανεξάρτητο chunk για TITLE+ABSTRACT (π.χ. #TITLE_ABSTRACT) και (β) εφαρμόζει chunking σε DESCRIPTION/CLAIMS χωρίς επικάλυψη (OVERLAP_WORDS=0), με στόχο να περιοριστεί το

redundancy μεταξύ διαδοχικών τμημάτων. Σε αυτή τη ρύθμιση, το TARGET_WORDS ορίζεται σε 230, ενώ ο αριθμός chunks περιορίζεται σε MAX_CHUNKS_DESC=1 και MAX_CHUNKS_CLAIMS=1.

Η τεχνική λογική είναι ότι το TA chunk λειτουργεί ως "συνοπτικό anchor" στο index, ενώ οι μεγάλες ενότητες εκπροσωπούνται από λίγα, μη επικαλυπτόμενα τμήματα ώστε να διατηρείται ισορροπία μεταξύ κάλυψης περιεχομένου και μεγέθους ευρετηρίου. Η απουσία overlap μειώνει τις σχεδόν διπλές αντιστοιχίσεις στο token-level retrieval, κάτι που περιορίζει την ανάγκη για έντονη deduplication και σταθεροποιεί το aggregation σε document-level. Αντίστροφα, μπορεί να αυξηθεί ο κίνδυνος "boundary loss", δηλαδή κρίσιμες φράσεις που πέφτουν στα όρια κοπής να μη συμπεριληφθούν στο ίδιο chunk με το απαραίτητο συμφραζόμενο.

5.4.2.4 Στρατηγική 4: Πολυ-ενότητα index με TA/DESC/CLMS chunks χωρίς prefix

Η τέταρτη στρατηγική υιοθετεί διακριτό διαχωρισμό ανά ενότητα: δημιουργεί (i) ένα TA chunk σύνοψης με TITLE+ABSTRACT (TA_000), (ii) πολλαπλά chunks για DESCRIPTION (DESC_000, DESC_001, …) και (iii) πολλαπλά chunks για CLAIMS (CLMS_000, CLMS_001, …), χωρίς να προσθέτει το TITLE/ABSTRACT ως prefix στα chunks των μεγάλων ενοτήτων. Οι παράμετροι chunking παραμένουν TARGET_WORDS=220 και OVERLAP_WORDS=15, με όρια MAX_CHUNKS_TA=1, MAX_CHUNKS_DESC=2 και MAX_CHUNKS_CLAIMS=2.

Η βασική τεχνική ιδέα είναι ότι το token budget κάθε chunk αξιοποιείται σχεδόν αποκλειστικά για "καθαρό" περιεχόμενο της αντίστοιχης ενότητας, ενώ το συνοπτικό πλαίσιο της πατέντας διατηρείται ως ανεξάρτητο TA chunk και όχι ως επαναλαμβανόμενο prefix. Αυτό μειώνει το replication του TITLE/ABSTRACT στο index, περιορίζει το redundancy και επιτρέπει το retrieval να ανασύρει είτε "summary-like" σήματα (TA) είτε λεπτομερείς τεχνικές αντιστοιχίσεις (DESC/CLMS). Στο τελικό ranking, τα chunk-level hits απαιτούν document-level aggregation, ενώ ο διαχωρισμός ενότητας βοηθά να ελέγχεται καλύτερα το πώς συνεισφέρουν διαφορετικά sections στο τελικό score.

5.5 Επιλογή τελικής στρατηγικής ευρετηρίασης για τα κύρια πειράματα

Συνολικά, οι στρατηγικές ευρετηρίασης για BM25 και ColBERT διαφοροποιούνται κυρίως ως προς τη μονάδα ανάκτησης (document-level έναντι chunk-level), τον βαθμό επανάληψης συμφραζομένων (prefixing TITLE+ABSTRACT) και την κάλυψη των μεγάλων ενοτήτων (DESCRIPTION/CLAIMS) μέσα στο ευρετήριο. Στο BM25, οι επιλογές αυτές επηρεάζουν άμεσα τη λεξιλογική στατιστική του index (term frequencies, μήκη κειμένων και redundancy), ενώ στο ColBERT επηρεάζουν το token budget ανά chunk, τον αριθμό των υποψηφίων που απαιτούνται στο Stage-1 και τη σταθερότητα της document-level aggregation. Κατά συνέπεια, οι πιο "πλούσιες" εκδοχές chunking ευνοούν την ανάκληση αλλά αυξάνουν κόστος σε χώρο και χρόνο, ενώ οι πιο "συμπιεσμένες" εκδοχές βελτιώνουν αποδοτικότητα αλλά μειώνουν την πιθανότητα ανάκτησης βαθύτερων λεπτομερειών. Με βάση αυτές τις παρατηρήσεις, παρακάτω τεκμηριώνεται η επιλογή της τελικής στρατηγικής που χρησιμοποιήθηκε στα κύρια πειράματα, λαμβάνοντας υπόψη τόσο τα retrieval metrics όσο και τους πρακτικούς περιορισμούς κλίμακας και εκτέλεσης.

Πριν την εκτέλεση των κύριων πειραμάτων πραγματοποιήθηκαν δειγματοληπτικές δοκιμές με εναλλακτικές στρατηγικές ευρετηρίασης, ώστε να επιλεγεί μία σταθερή και αποτελεσματική βασική ρύθμιση ανά σύστημα ανάκτησης. Για το BM25 (Pyserini/Lucene) επιλέχθηκε πλήρες chunking των DESCRIPTION/CLAIMS (TARGET_WORDS=220, OVERLAP_WORDS=15) χωρίς περιορισμό στον αριθμό chunks, με προσθήκη TITLE+ABSTRACT ως prefix σε κάθε chunk, ώστε να μεγιστοποιείται η κάλυψη των μεγάλων ενοτήτων και να ευνοείται η ανάκληση σε έγγραφα όπου η σχετική πληροφορία είναι διάσπαρτη. Επειδή η ευρετηρίαση γίνεται σε passage επίπεδο, τα αποτελέσματα της

ανάκτησης συγκεντρώνονται στη συνέχεια σε document-level μέσω aggregation, ώστε το τελικό ranking να αφορά πατέντες και όχι μεμονωμένα chunks.

Αντίστοιχα, για το ColBERT επιλέχθηκε chunking των DESCRIPTION/CLAIMS με τα ίδια βασικά μεγέθη (TARGET_WORDS=220, OVERLAP_WORDS=15), με ενσωμάτωση TITLE+ABSTRACT ως συμφραζόμενο (prefix) σε κάθε chunk και με περιορισμό έως δύο chunks ανά πεδίο (MAX_CHUNKS_DESC=2, MAX_CHUNKS_CLAIMS=2), ώστε να επιτυγχάνεται ισορροπία μεταξύ κάλυψης περιεχομένου και υπολογιστικού κόστους (μέγεθος index και κόστος Stage-1 ανάκτησης/aggregation). Οι επιλογές αυτές χρησιμοποιήθηκαν ως κοινή βάση ευρετηρίασης για όλα τα πειράματα που ακολούθησαν, ενώ οι υπόλοιπες διαφοροποιήσεις μεταξύ runs υλοποιούνται σε επίπεδο παραμέτρων ανάκτησης και αξιολόγησης.

Κεφάλαιο 6ο: Σύστημα Διαχείρισης και Οπτικοποίησης

Το κεφάλαιο αυτό παρουσιάζει το σύστημα που αναπτύχθηκε για την καταγραφή, οργάνωση και οπτικοποίηση των πειραματικών αποτελεσμάτων της εργασίας. Περιγράφεται ο τρόπος με τον οποίο συγκεντρώνονται οι εκτελέσεις, τα παραγόμενα αρχεία και οι μετρικές αξιολόγησης, καθώς και η διαδικασία παρουσίασης των αποτελεσμάτων με τρόπο που διευκολύνει τη σύγκριση μεταξύ διαφορετικών σεναρίων σύνοψης και ανάκτησης.

6.1 Ανάγκη διαχείρισης πειραμάτων και αποτελεσμάτων

Η παρούσα διπλωματική εργασία περιλαμβάνει εκτεταμένη πειραματική διαδικασία, καθώς υλοποιούνται και αξιολογούνται πολλαπλές προσεγγίσεις σύνοψης (extractive, abstractive, instructive και hybrid), διαφορετικά γλωσσικά μοντέλα ανά προσέγγιση, καθώς και εναλλακτικές τεχνικές ευρετηρίασης και ανάκτησης πληροφορίας (BM25 και ColBERT). Η συνδυαστική φύση αυτών των παραμέτρων οδηγεί σε μεγάλο αριθμό πειραμάτων, όπου κάθε εκτέλεση παράγει πολλαπλά παραγόμενα δεδομένα, όπως περιλήψεις, αρχεία ευρετηρίου, λίστες αποτελεσμάτων ανά query και μετρικές αξιολόγησης.

Σε αυτό το πλαίσιο, η οργανωμένη διαχείριση των πειραμάτων αποτελεί αναγκαία προϋπόθεση για τη διασφάλιση της αναπαραγωγιμότητας, τη συστηματική σύγκριση των μεθόδων και την εξαγωγή αξιόπιστων συμπερασμάτων. Χωρίς μηχανισμό συγκέντρωσης και τυποποιημένης καταγραφής, οι διαφοροποιήσεις σε ρυθμίσεις, εκδόσεις μοντέλων ή στρατηγικές indexing είναι δύσκολο να παρακολουθηθούν, ενώ η αποτίμηση της πραγματικής επίδρασης κάθε επιλογής (π.χ. τύπος σύνοψης ή μέθοδος ανάκτησης) γίνεται επιρρεπής σε σφάλματα και ασυνέπειες.

Επιπλέον, η αξιολόγηση δεν περιορίζεται σε μία μόνο διάσταση. Για κάθε πείραμα απαιτείται να συνδυαστούν ποσοτικά αποτελέσματα (π.χ. μετρικές ανάκτησης) με ποιοτική εξέταση των παραγόμενων περιλήψεων, καθώς και με λειτουργικές παρατηρήσεις που σχετίζονται με την αποδοτικότητα (χρόνος εκτέλεσης, όγκος index, αριθμός chunks). Συνεπώς, είναι απαραίτητη μία ενοποιημένη διαδικασία αποθήκευσης και ανάκτησης των πειραματικών δεδομένων, ώστε να είναι δυνατή η συγκριτική ανάλυση μεταξύ διαφορετικών σεναρίων με σταθερό και διαφανή τρόπο.

Για την κάλυψη αυτής της ανάγκης αναπτύχθηκε σύστημα διαχείρισης πειραμάτων, το οποίο συγκεντρώνει, οργανώνει και παρουσιάζει τα αποτελέσματα των εκτελέσεων. Το σύστημα επιτρέπει την αποθήκευση των παραγόμενων αρχείων και μετρικών σε δομημένη μορφή, τη γρήγορη αναζήτηση και φιλτράρισή τους, καθώς και την οπτικοποίηση βασικών δεικτών απόδοσης. Με αυτόν τον τρόπο διασφαλίζεται ότι τα πειράματα μπορούν να αναλυθούν με συνεπή μεθοδολογία και ότι τα συμπεράσματα της εργασίας βασίζονται σε ελεγχόμενη και τεκμηριωμένη πειραματική διαδικασία.

6.2 Αρχιτεκτονική web εφαρμογής

Η διαχείριση των πειραματικών αποτελεσμάτων υλοποιήθηκε μέσω web εφαρμογής, όπου η επιχειρησιακή λογική υλοποιείται σε PHP και η αποθήκευση των δεδομένων πραγματοποιείται σε βάση MySQL. Η επιλογή αυτής της αρχιτεκτονικής επιτρέπει την κεντρική καταγραφή όλων των εκτελέσεων και τη συνεπή οργάνωση των αποτελεσμάτων, ανεξάρτητα από το αν τα πειράματα αφορούν διαφορετικές μεθόδους σύνοψης ή διαφορετικές τεχνικές ανάκτησης (BM25 ή ColBERT).

Η εφαρμογή λειτουργεί ως ενδιάμεσο επίπεδο μεταξύ του πειραματικού pipeline και της τελικής παρουσίασης των αποτελεσμάτων. Από τη μία πλευρά, λαμβάνει τις βασικές πληροφορίες κάθε run (ρυθμίσεις, παραμέτρους και μετρικές) και τις αποθηκεύει σε δομημένη μορφή στη MySQL. Από την άλλη

πλευρά, παρέχει διεπαφή προβολής που επιτρέπει την αναζήτηση, το φιλτράρισμα και την ταξινόμηση των runs, έτσι ώστε ο χρήστης να μπορεί να συγκρίνει γρήγορα διαφορετικές εκτελέσεις και να εντοπίζει τις επιλογές που οδηγούν σε καλύτερη απόδοση.

Κεντρική σχεδιαστική επιλογή είναι ότι κάθε πείραμα αποτυπώνεται ως μία εγγραφή αποτελεσμάτων, η οποία περιλαμβάνει τόσο περιγραφικά χαρακτηριστικά (π.χ. τύπος σύνοψης, μοντέλο, είδος retrieval, στρατηγική index) όσο και ποσοτικές μετρικές. Με αυτόν τον τρόπο, η σύγκριση μεταξύ διαφορετικών σεναρίων πραγματοποιείται με ενιαία και αναπαραγώγιμη λογική, χωρίς να απαιτείται χειροκίνητη συλλογή και αντιπαραβολή δεδομένων από διαφορετικά αρχεία.

6.3 Αποθήκευση και οργάνωση δεδομένων

Η αποθήκευση των αποτελεσμάτων οργανώθηκε σε έναν βασικό πίνακα MySQL, ο οποίος συγκεντρώνει σε ενιαία μορφή τις πληροφορίες που απαιτούνται για την περιγραφή και την αξιολόγηση κάθε πειράματος. Η δομή του πίνακα περιλαμβάνει πεδία ταυτοποίησης του run, πεδία που αποτυπώνουν τις επιλογές ρυθμίσεων, καθώς και σύνολο μετρικών που χρησιμοποιούνται για την αξιολόγηση της απόδοσης.

Σε επίπεδο ταυτοποίησης, χρησιμοποιούνται κλειδιά που επιτρέπουν την ιχνηλασιμότητα των σταδίων του pipeline, όπως το index_key, το retrieval_key και το summary_key. Τα πεδία αυτά επιτρέπουν να καταγράφεται με σαφήνεια ποια ευρετηρίαση, ποια διαδικασία ανάκτησης και ποια παραγωγή περιλήψεων αντιστοιχούν σε κάθε εκτέλεση, διευκολύνοντας την αναπαραγωγή των αποτελεσμάτων και τη συσχέτισή τους με τα αντίστοιχα αρχεία.

Παράλληλα, καταγράφονται οι βασικές επιλογές παραμετροποίησης, όπως ο τύπος ευρετηρίασης και ανάκτησης (index_type, vector_search), ο τύπος και το μοντέλο σύνοψης (summary_type, summary_model), καθώς και στοιχεία που σχετίζονται με το dataset και τη στρατηγική συνάθροισης βαθμολογιών (dataset_used, agg_mode). Με τον τρόπο αυτό, κάθε εγγραφή περιγράφει πλήρως το "σενάριο" εκτέλεσης.

Τέλος, στον ίδιο πίνακα αποθηκεύονται συγκεντρωτικά οι μετρικές αξιολόγησης ανάκτησης για πολλαπλές τιμές k (Precision@k, Recall@k, MAP, nDCG), καθώς και δείκτες που σχετίζονται με τη σύνοψη (compression ratio, compression percent και, όπου εφαρμόζεται, semantic similarity). Συμπληρωματικά, καταγράφονται χρόνοι εκτέλεσης (summarized_duration, retrieval_duration), ώστε να είναι δυνατή η αξιολόγηση της αποδοτικότητας κάθε προσέγγισης και η συζήτηση πιθανών συμβιβασμών μεταξύ ποιότητας και κόστους. Ο Πίνακας 6.1 συνοψίζει τη λογική ομαδοποίηση των πεδίων.

Πίνακας 6.1: Λογική ομαδοποίηση πεδίων του πίνακα πειραματικών αποτελεσμάτων

Ομάδα πεδίων	Πεδία (ενδεικτικά)	Σκοπός
Ταυτοποίηση πειράματος	id, index_key, retrieval_key, summary_key	Μοναδική αναφορά σε κάθε run και δυνατότητα σύνδεσης των σταδίων (σύνοψη–ευρετηρίαση–ανάκτηση).
Ρυθμίσεις / Παράμετροι πειράματος	index_type, vector_search, summary_type,	Περιγράφει το σενάριο εκτέλεσης ώστε να υποστηρίζεται φιλτράρισμα και δίκαιη σύγκριση μεταξύ runs.

	summary_model, dataset_used, agg_mode	
Χαρακτηριστικά εκτέλεσης	queries_processed, query_length, tags_used_token, summarized_files, index_size_details	Καταγράφει κλίμακα και συνθήκες εκτέλεσης (πλήθος queries, είσοδοι/πεδία, μέγεθος και παραγόμενα δεδομένα).
Μετρικές ανάκτησης	p_at_10, p_at_50, p_at_100, recall_at_10, recall_at_50, recall_at_100, map_score, ndgg_at_10, ndgg_at_50, ndgg_at_100, press_at_10, press_at_50, press_at_100	Ποσοτική αξιολόγηση retrieval για διαφορετικά k, ώστε να συγκρίνεται η απόδοση μεταξύ μεθόδων/μοντέλων.
Μετρικές σύνοψης	compress_ratio, compress_percent, sbert_similarity	Περιγράφει ιδιότητες των περιλήψεων (συμπίεση και, όπου εφαρμόζεται, σημασιολογική ομοιότητα).
Χρόνοι εκτέλεσης	summarized_duration, retrieval_duration	Σύγκριση αποδοτικότητας (runtime) ανά run και αξιολόγηση κόστους/οφέλους κάθε επιλογής.

6.4 Οπτικοποίηση αποτελεσμάτων

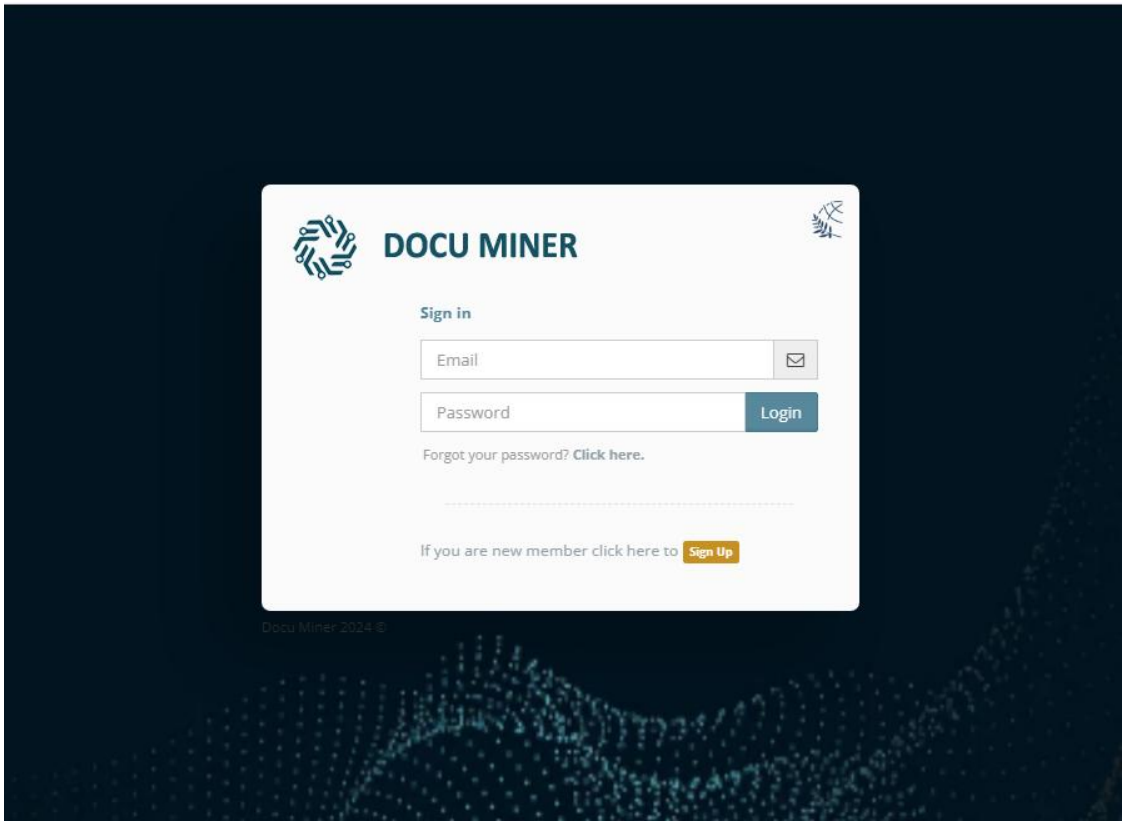
Η web εφαρμογή DOCU MINER παρέχει γραφικό περιβάλλον για τη διερεύνηση και συγκριτική αξιολόγηση των πειραμάτων, επιτρέποντας στον χρήστη να μεταβεί από την απλή αποθήκευση μετρικών σε ενεργή ανάλυση και σύγκριση σεναρίων. Η πρόσβαση στο σύστημα πραγματοποιείται μέσω μηχανισμού αυθεντικοποίησης (login), ο οποίος εξασφαλίζει ελεγχόμενη πρόσβαση στο περιβάλλον πειραμάτων και στα αποθηκευμένα αποτελέσματα (**Εικόνα 1**).

Μετά την είσοδο, η εφαρμογή οργανώνει λειτουργικά τις ενότητες της εργασίας σε δύο βασικούς άξονες: (α) μεθοδολογίες σύνοψης και (β) μέθοδοι ανάκτησης πληροφορίας. Στο αριστερό μενού παρουσιάζονται οι κατηγορίες σύνοψης (abstractive, extractive, instructive, hybrid) και τα επιμέρους μοντέλα ανά κατηγορία, καθώς και οι επιλογές για retrieval (BM25/Pyserini και ColBERT evaluation) (**Εικόνα 2**). Η δομή αυτή αντικατοπτρίζει τη λογική του pipeline της εργασίας και επιτρέπει γρήγορη πλοήγηση από την πλευρά της σύνοψης προς την πλευρά της ανάκτησης.

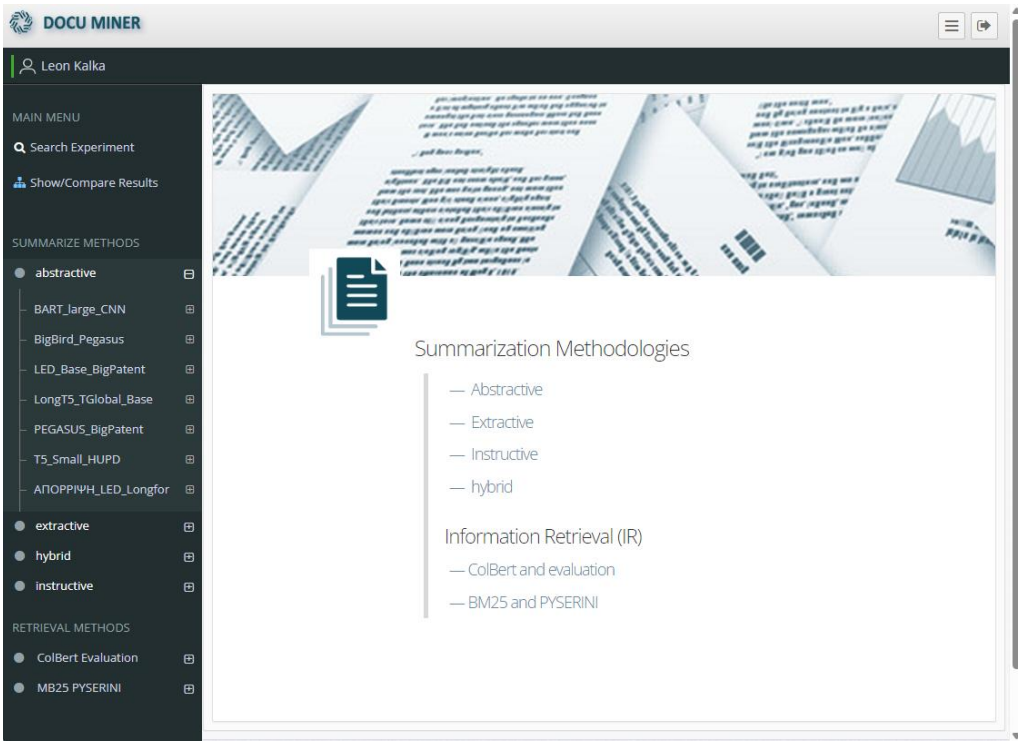
Κεντρική λειτουργία για τη συγκριτική ανάλυση αποτελεί η σελίδα “Show/Compare Results”, όπου τα αποτελέσματα προβάλλονται σε πίνακα και υποστηρίζονται μηχανισμοί φιλτραρίσματος, ταξινόμησης και εξαγωγής (**Εικόνα 3**). Ο πίνακας παρουσιάζει τις βασικές μεταβλητές κάθε run (π.χ. index_type, summary_type, summary_model, tags_used_token), καθώς και τις μετρικές αξιολόγησης (Precision@k,

Recall@k, MAP, nDCG και άλλες). Η δυνατότητα ταξινόμησης ανά στήλη επιτρέπει, για παράδειγμα, να εντοπιστούν γρήγορα τα καλύτερα runs ως προς MAP ή nDCG, ενώ τα φίλτρα (Index Type, Summary Type, Summary Model, Tags Used) επιτρέπουν περιορισμό του χώρου αναζήτησης σε συγκεκριμένες κατηγορίες πειραμάτων. Επιπλέον, παρέχονται λειτουργίες εμφάνισης/απόκρυψης στηλών, επιλογής πλήθους γραμμών ανά σελίδα και εξαγωγής σε Excel, διευκολύνοντας την περαιτέρω ανάλυση εκτός εφαρμογής.

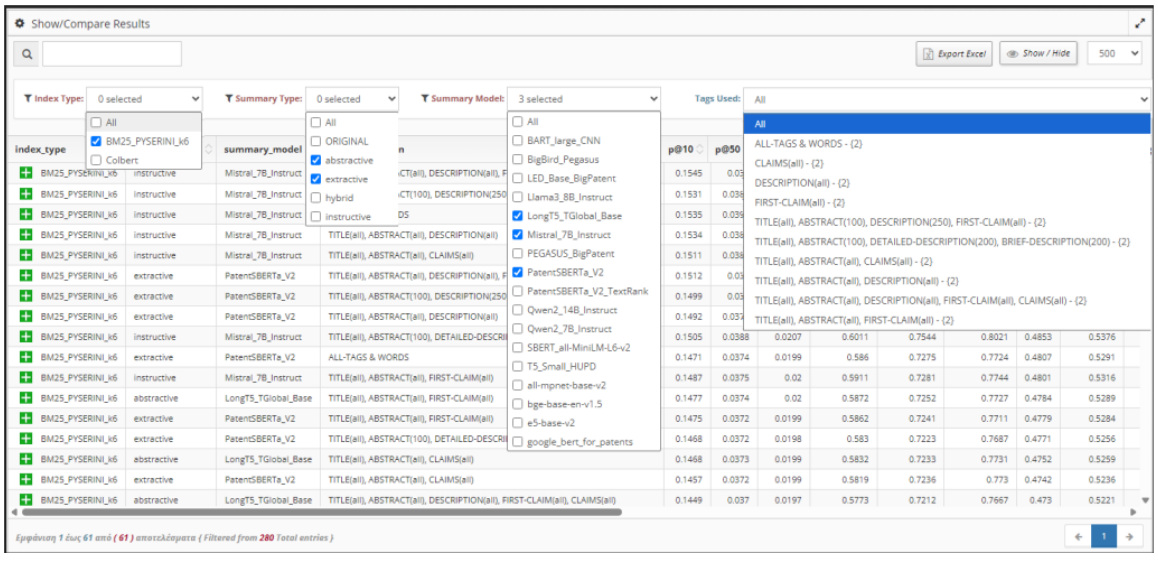
Συνολικά, η οπτικοποίηση λειτουργεί ως επιχειρησιακή "γέφυρα" μεταξύ των αρχείων που παράγει το πειραματικό pipeline και της τελικής αποτίμησης της απόδοσης: οι ρυθμίσεις, τα σενάρια και οι μετρικές γίνονται άμεσα συγκρίσιμα μέσα από ένα ενιαίο περιβάλλον, μειώνοντας την ανάγκη χειροκίνητης συλλογής δεδομένων και περιορίζοντας τα σφάλματα καταγραφής ή ερμηνείας.



Εικόνα 1: Οθόνη εισόδου (login) της εφαρμογής DOCU MINER.



Εικόνα 2: Κεντρική σελίδα πλοήγησης και μενού επιλογών μεθοδολογιών σύνοψης και retrieval.

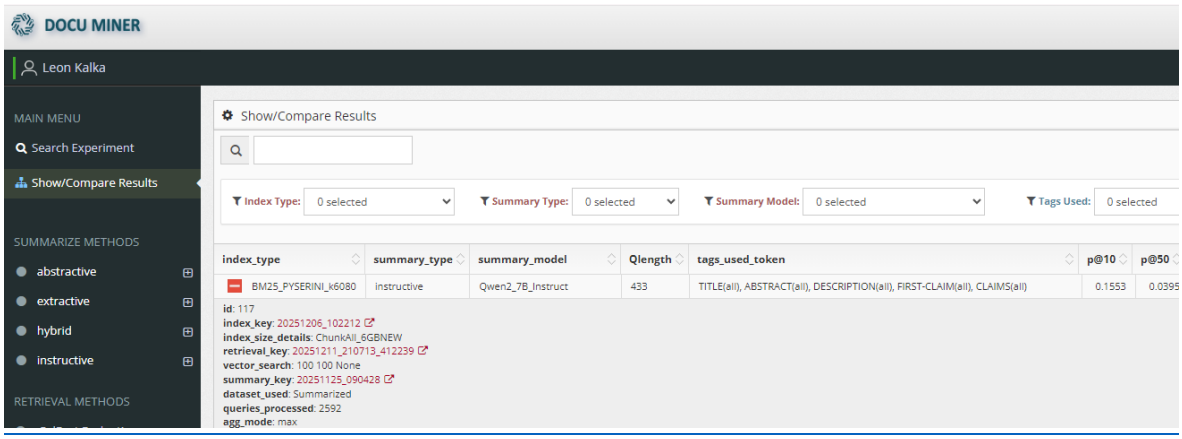


Εικόνα 3: Σελίδα “Show/Compare Results” για φιλτράρισμα και συγκριτική προβολή πειραμάτων και μετρικών.

6.4.1 Λειτουργία drill-down και ιχνηλασιμότητα αρχείων αποτελεσμάτων

Η εφαρμογή υποστηρίζει λειτουργία drill-down, όπου κάθε εγγραφή (run) μπορεί να αναπτυχθεί ώστε να εμφανίσει συμπληρωματικές πληροφορίες που δεν προβάλλονται στον βασικό πίνακα. Στην επεκτεινόμενη προβολή παρουσιάζονται συνοπτικά μεταδεδομένα ταυτοποίησης και βασικές ρυθμίσεις/παράμετροι της εκτέλεσης, διευκολύνοντας τον γρήγορο έλεγχο και τη διασταύρωση των πειραμάτων

Ιδιαίτερο χαρακτηριστικό της υλοποίησης είναι η χρήση ενεργών συνδέσμων (links) στα αντίστοιχα keys. Τα keys λειτουργούν ως δείκτες που οδηγούν στα πραγματικά αποθηκευμένα αρχεία μετρήσεων και αποτελεσμάτων, επιτρέποντας άμεση μετάβαση από τη συγκεντρωτική καταγραφή στο πρωτογενές υλικό της εκτέλεσης (π.χ. outputs, logs, metrics). Με αυτόν τον τρόπο υλοποιείται πρακτικά η ιχνηλασιμότητα (traceability) των πειραμάτων: κάθε εγγραφή της βάσης δεν αποτελεί απλώς μια σειρά αριθμών και μετρικών, αλλά συνδέεται άμεσα με τα αντίστοιχα αρχεία που παρήχθησαν κατά την εκτέλεση. Η δυνατότητα αυτή μειώνει σημαντικά τον χρόνο ελέγχου, διευκολύνει την επιβεβαίωση αποτελεσμάτων και υποστηρίζει τη διαδικασία τεκμηρίωσης για τη συγγραφή της εργασίας.

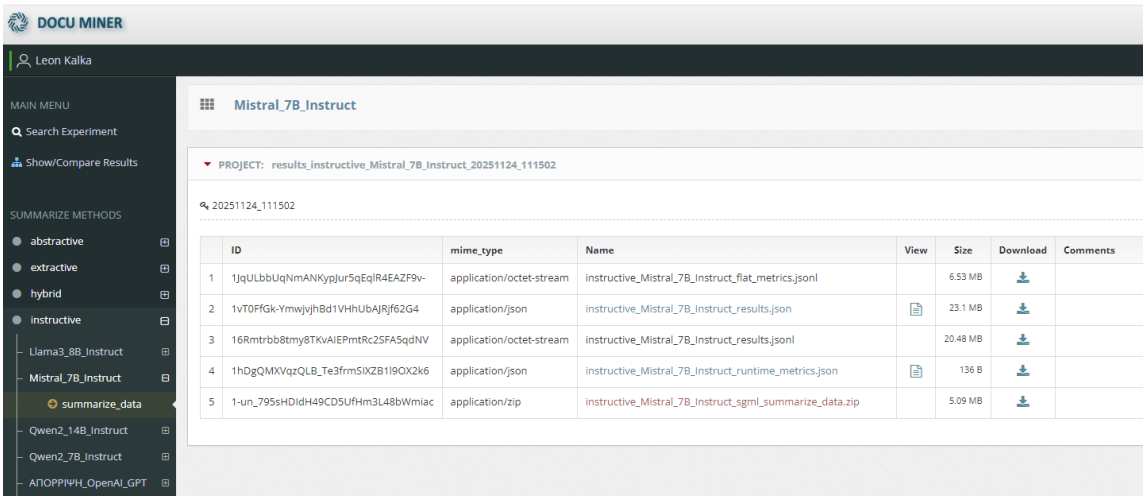


Εικόνα 4: Drill-down προβολή εγγραφής πειράματος και ενεργοί σύνδεσμοι (keys) προς τα αποθηκευμένα αρχεία μετρήσεων και αποτελεσμάτων.

6.4.2 Προβολή πακέτου αρχείων ανά εκτέλεση και σύνδεση με Google Drive

Πέρα από τη συγκεντρωτική προβολή των runs, η εφαρμογή παρέχει σελίδα “Experiments” που λειτουργεί ως κατάλογος αρχείων (file catalog) για κάθε εκτέλεση. Όταν ο χρήστης επιλέξει ένα key από το drill-down (π.χ. summary_key), μεταφέρεται σε ειδική σελίδα όπου εμφανίζεται το πλήρες “πακέτο” αρχείων που παρήχθησαν από τη συγκεκριμένη διαδικασία, μαζί με βασικά μεταδεδομένα (αναγνωριστικό αρχείου, τύπος MIME, όνομα αρχείου και μέγεθος). Η σχεδίαση αυτή επιτρέπει να αντιμετωπίζεται κάθε run όχι μόνο ως μία εγγραφή με μετρικές, αλλά και ως ένα σύνολο artifacts που είναι διαθέσιμα για έλεγχο, επαλήθευση και περαιτέρω ανάλυση.

Κάθε αρχείο στη λίστα είναι ενεργός σύνδεσμος: με επιλογή “View” ή “Download” η εφαρμογή εκτελεί ανακατεύθυνση στο αντίστοιχο αντικείμενο του Google Drive, όπου βρίσκεται το αρχείο αποθηκευμένο. Με αυτόν τον τρόπο, η εφαρμογή διατηρεί κεντρικά τη δομή και τις συσχετίσεις των εκτελέσεων, ενώ η αποθήκευση των πρωτογενών αρχείων παραμένει στο Drive. Η προσέγγιση αυτή υποστηρίζει πρακτικά την ιχνηλασιμότητα και την αναπαραγωγιμότητα: από ένα συγκεντρωτικό αποτέλεσμα στη βάση, ο χρήστης μπορεί να μεταβεί άμεσα στο πλήρες υλικό της εκτέλεσης (αποτελέσματα, μετρικές, logs και συμπιεσμένα δεδομένα).



Εικόνα 5: Σελίδα “Experiments” για ένα summary_key: λίστα των παραγόμενων αρχείων ανά εκτέλεση και ενεργοί σύνδεσμοι που οδηγούν στο Google Drive για προβολή ή λήψη.

6.4.3 Τυποποίηση παραγόμενων artifacts ανά στάδιο εκτέλεσης

Για να υποστηριχθεί η συστηματική αξιολόγηση και η αναπαραγωγιμότητα, κάθε στάδιο του πειραματικού pipeline παράγει ένα σταθερό “πακέτο” αρχείων (artifacts) με προκαθορισμένη δομή και ονοματολογία. Η τυποποίηση αυτή επιτρέπει: (α) την ενιαία καταγραφή των εκτελέσεων στη βάση, (β) την άμεση αντιστοίχιση μεταξύ μιας εγγραφής run και των παραγόμενων αρχείων, και (γ) την αυτοματοποίηση της προβολής τους από την εφαρμογή μέσω του καταλόγου αρχείων (Experiments). Έτσι, ανεξάρτητα από το μοντέλο ή τη μέθοδο που χρησιμοποιείται, το σύστημα διατηρεί κοινό “συμβόλαιο” εξόδων, το οποίο καθιστά τις εκτελέσεις συγκρίσιμες και εύκολα ελέγξιμες.

Στο πλαίσιο της παρούσας εργασίας, διακρίνονται δύο βασικές κατηγορίες πακέτων: (α) πακέτο σύνοψης (summarization) και (β) πακέτο ανάκτησης (retrieval). Το πακέτο σύνοψης περιλαμβάνει πέντε αρχεία ανά εκτέλεση, ενώ το πακέτο ανάκτησης περιλαμβάνει τέσσερα αρχεία ανά εκτέλεση. Η διάκριση αυτή αντανακλά τη διαφορετική φύση των δύο σταδίων: η σύνοψη παράγει πρωτογενές κείμενο (summaries) και συνοδευτικές μετρικές/δεδομένα, ενώ η ανάκτηση παράγει κατατάξεις αποτελεσμάτων και αρχεία αξιολόγησης/αναπαραγωγής (evaluation artifacts).

6.4.3.1 Πακέτο αρχείων σύνοψης (Summarization) ανά εκτέλεση

Κάθε εκτέλεση σύνοψης παράγει σταθερά πέντε αρχεία, τα οποία καλύπτουν τόσο το παραγόμενο κείμενο (περιλήψεις) όσο και την ποσοτική τεκμηρίωση της μεταβολής του μήκους πριν και μετά τη σύνοψη. Η λογική είναι ότι για κάθε run διατηρείται πλήρες ίχνος: (α) τι κείμενο χρησιμοποιήθηκε ως είσοδος, (β) τι σύνοψη παρήχθη, και (γ) πόσο “συμπιέστηκε” ανά πεδίο και συνολικά.

1. Αρχείο αποτελεσμάτων περιλήψεων (results)

   Περιλαμβάνει τις παραγόμενες περιλήψεις ανά έγγραφο, συνδεδεμένες με το αντίστοιχο αναγνωριστικό πατέντας/αρχείου. Επιπλέον, καταγράφει συστηματικά τα μήκη του κειμένου πριν τη σύνοψη (ανά tag/πεδίο, σε λέξεις και tokens) και τα αντίστοιχα μήκη μετά τη σύνοψη, ώστε να προκύπτουν άμεσα δείκτες συμπίεσης (compression ratio / compression percent). Η ακριβής δομή του αρχείου, τα πεδία/μεταδεδομένα και ένα ενδεικτικό παράδειγμα εγγραφής παρουσιάζονται στο **ΠΑΡΑΡΤΗΜΑ Α.** Αποτελεί το βασικό artifact του σταδίου σύνοψης, καθώς

χρησιμοποιείται: (α) για ποιοτική επιθεώρηση του παραγόμενου κειμένου, (β) για ποσοτική τεκμηρίωση της μείωσης του όγκου, και (γ) ως είσοδος σε επόμενα στάδια (indexing/retrieval) όταν αξιολογείται η επίδραση της σύνοψης στην ανάκτηση.

2. Αρχείο συγκεντρωτικών μετρικών σύνοψης (flat metrics)

   Περιλαμβάνει "επιπεδοποιημένες" (flat) μετρικές σε γραμμική μορφή, συνήθως ανά έγγραφο και ανά tag. Εκεί αποτυπώνονται συνοπτικά οι βασικοί δείκτες πριν/μετά (μήκη σε words/tokens), οι δείκτες συμπίεσης, και όπου εφαρμόζεται επιπλέον μετρική ποιότητας/ομοιότητας (π.χ. σημασιολογική ομοιότητα). Ο ρόλος του είναι να επιτρέπει γρήγορο reporting και ένταξη των μετρικών σε συγκριτικούς πίνακες, χωρίς να απαιτείται ανάγνωση του πλήρους results.

3. Αρχείο χρονικών/λειτουργικών μετρικών (runtime metrics)

   Καταγράφει στοιχεία αποδοτικότητας της εκτέλεσης, όπως χρόνο σύνοψης, ρυθμό επεξεργασίας και, όπου είναι διαθέσιμο, πρόσθετες ενδείξεις κόστους (π.χ. αριθμός επεξεργασμένων εγγράφων, batches). Είναι κρίσιμο για σύγκριση μοντέλων όχι μόνο ως προς την ποιότητα της σύνοψης, αλλά και ως προς τον απαιτούμενο χρόνο/υπολογιστικό κόστος.

4. Αρχείο αποτελεσμάτων σε μορφή JSONL (results.jsonl)

   Παρέχει την ίδια πληροφορία αποτελεσμάτων σε line-oriented μορφή (μία εγγραφή ανά γραμμή). Η μορφή αυτή είναι πρακτική για μεγάλα σύνολα δεδομένων, επειδή επιτρέπει streaming/τμηματική φόρτωση και ταχύτερη επεξεργασία σε scripts και pipelines.

5. Συμπιεσμένο πακέτο παραγόμενων αρχείων σύνοψης (sgml_summarize_data.zip)

   Περιλαμβάνει τα παραγόμενα αρχεία σύνοψης σε μορφή SGML-like, δηλαδή το περιεχόμενο όπως διαμορφώθηκε με tags μετά τη σύνοψη. Χρησιμεύει ως "αρχειακή" μορφή της εξόδου, ώστε: (α) να διατηρείται αυτούσιο δείγμα των παραγόμενων κειμένων, (β) να γίνεται εύκολη μεταφορά/αρχειοθέτηση ως ενιαίο αντικείμενο, και (γ) να υποστηρίζεται επαλήθευση ή επαναχρησιμοποίηση σε μεταγενέστερα στάδια.

6.4.3.2 Πακέτο αρχείων ανάκτησης (Retrieval) ανά εκτέλεση

Αντίστοιχα, κάθε εκτέλεση ανάκτησης παράγει τέσσερα αρχεία, τα οποία καλύπτουν: (α) το τελικό αποτέλεσμα αξιολόγησης, (β) αναλυτική πληροφορία ανά query, (γ) αρχείο-στιγμιότυπο για αναπαραγωγή αξιολόγησης, και (δ) πλήρη καταγραφή ρυθμίσεων/συνθηκών εκτέλεσης.

- **Αρχείο συγκεντρωτικών αποτελεσμάτων αξιολόγησης (retrieval_eval_metrics.json)**

  Αποτελεί το κύριο «επίσημο» αποτέλεσμα του run. Περιλαμβάνει συγκεντρωτικές μετρικές σε πολλαπλά cutoffs (π.χ. macro_precision@10/50/100, macro_recall@10/50/100, macro_map@k, macro_ndcg@k, καθώς και συνολικά αθροίσματα τύπου all_relevant_found / all_relevant_total).
  Επιπλέον, λειτουργεί και ως "config snapshot", αφού ενσωματώνει κρίσιμες παραμέτρους του run (π.χ. τύπος μεθόδου, μοντέλο, dataset, tags που χρησιμοποιήθηκαν, όρια/προϋπολογισμούς ανά tag, top_k, eval_ks, κ.ά.). Σε αρκετές εκτελέσεις καταγράφονται και στατιστικά που σχετίζονται με τη σύνοψη (π.χ. δείκτες συμπίεσης), ώστε το retrieval αποτέλεσμα να συνδέεται τεκμηριωμένα με το summarization input που χρησιμοποιήθηκε.

- **Αρχείο ανά-query αποτελεσμάτων και μετρικών (retrieval_combined_metrics_results.json)**

Περιλαμβάνει ανά query μια ενιαία εγγραφή που συνοψίζει τις μετρικές του (π.χ. precision@k, recall@k, f1@k, press@k, relevant_found@k και relevant_total). Είναι το αρχείο που υποστηρίζει πιο εύκολα drill-down ανάλυση: από τα συνολικά macro αποτελέσματα μπορεί να εντοπιστεί ποια queries αποδίδουν καλά/άσχημα και γιατί (π.χ. λίγα σχετικά διαθέσιμα, χαμηλό relevant_found στα πρώτα k, κ.λπ.). Επειδή είναι ήδη «συγκεντρωμένο ανά query», μειώνει την ανάγκη συνδυασμών πολλών πηγών κατά το reporting.

- **Αρχείο εισόδου αξιολόγησης για αναπαραγωγή (retrieval_pytreceval_input.json)**

  Αποθηκεύει το ranking αποτέλεσμα σε μορφή κατάλληλη για αναπαραγωγή της αξιολόγησης χωρίς επανεκτέλεση ανάκτησης. Τυπικά περιλαμβάνει για κάθε query μια αντιστοίχιση doc_id-score για τα ανακτηθέντα έγγραφα. Αυτό το αρχείο είναι κρίσιμο για διαφάνεια και επαληθευσιμότητα: ακόμη και αν αλλάξει το περιβάλλον ή οι βιβλιοθήκες, το «στιγμιότυπο» των κατατάξεων παραμένει διαθέσιμο ώστε να ξαναυπολογιστούν οι μετρικές ή να γίνει ανεξάρτητος έλεγχος.

- **Αρχείο καταγραφής εκτέλεσης (log_result_*.txt)**

  Περιλαμβάνει αναλυτικό αποτύπωμα της εκτέλεσης: timestamps, πληροφορίες περιβάλλοντος (εκδόσεις, platform), paths, επιλογές παραμετροποίησης (π.χ. index/query settings, tag configuration, budgets), καθώς και χρόνους (elapsed) ανά στάδιο. Το log είναι το βασικό εργαλείο για debugging και για διερεύνηση αποκλίσεων μεταξύ runs (π.χ. γιατί ένα run είχε διαφορετικό χρόνο ή διαφορετική συμπεριφορά), αλλά και για τεκμηρίωση της πειραματικής διαδικασίας.

6.4.3.3 Σχέση πλήθους artifacts με τον συνολικό όγκο αποθετηρίου

Η προσέγγιση «ένα πακέτο ανά run» οδηγεί σε προβλέψιμη αύξηση τόσο του πλήθους όσο και του συνολικού όγκου των παραγόμενων αρχείων, καθώς το πλήθος των artifacts αυξάνεται γραμμικά με τον αριθμό εκτελέσεων. Με βάση τα πειράματα που διατηρήθηκαν στην παρούσα εργασία, έχουν πραγματοποιηθεί 319 πειράματα ανάκτησης πάνω στους δύο τελικούς δείκτες που επιλέχθηκαν, με 4 παραγόμενα αρχεία ανά πείραμα, δηλαδή 319 × 4 = 1276 αρχεία. Αντίστοιχα, για τις 18 μεθόδους σύνοψης με 5 αρχεία ανά μέθοδο προκύπτουν 18 × 5 = 90 αρχεία. Συνεπώς, μόνο για τα τελικά/διατηρημένα σενάρια παράγονται συνολικά 1.276 + 90 = 1.366 αρχεία, χωρίς να συνυπολογιστούν τυχόν ενδιάμεσα outputs ή βοηθητικά logs.

Ο παραπάνω αριθμός αποτελεί συντηρητική εκτίμηση, καθώς θα ήταν σημαντικά μεγαλύτερος αν συμπεριλαμβάνονταν και οι δείκτες (indexes (BM25 & Colbert) που δεν εμφανίζονται εδώ κατά τη διαδικασία επιλογής των τελικών ρυθμίσεων.

Η κλίμακα αυτή αναδεικνύει ότι η διαχείριση των πειραμάτων δεν μπορεί να βασιστεί αποτελεσματικά σε χειροκίνητη περιήγηση φακέλων και μεμονωμένο άνοιγμα αρχείων. Όταν κάθε εκτέλεση αντιστοιχεί σε πολλαπλά αρχεία διαφορετικού τύπου (αποτελέσματα, μετρικές, logs, συμπιεσμένα δεδομένα), η αναζήτηση του σωστού συνόλου για ένα συγκεκριμένο run, η σύγκριση με άλλα runs και ο έλεγχος αναπαραγωγιμότητας γίνεται χρονοβόρα και επιρρεπής σε λάθη. Για τον λόγο αυτό κρίθηκε αναγκαία η ανάπτυξη ειδικής εφαρμογής διαχείρισης, η οποία ενοποιεί τα πειράματα σε επίπεδο “run”, διατηρεί τις συσχετίσεις μεταξύ των σταδίων του pipeline και επιτρέπει άμεση ανάκτηση του πλήρους πακέτου αρχείων που αντιστοιχεί σε κάθε εκτέλεση. Με αυτόν τον τρόπο, η εφαρμογή λειτουργεί ως ενιαίο σημείο πρόσβασης και ελέγχου, μετατρέποντας έναν μεγάλο και δύσκολα διαχειρίσιμο όγκο artifacts σε οργανωμένη πληροφορία, κατάλληλη για συστηματική ανάλυση και τεκμηρίωση.

6.5 Παραδείγματα χρήσης

Η εφαρμογή υποστηρίζει πρακτικά σενάρια που προκύπτουν άμεσα από τους στόχους της εργασίας, δηλαδή τη σύγκριση μεθόδων σύνοψης και την αξιολόγηση της επίδρασής τους στην ανάκτηση πληροφορίας.

- **Παράδειγμα 1ο : Σύγκριση μοντέλων μέσα στην ίδια κατηγορία σύνοψης.**

  Ο χρήστης επιλέγει συγκεκριμένο τύπο index (π.χ. BM25) και περιορίζει τα αποτελέσματα σε μία κατηγορία σύνοψης (π.χ. instructive). Στη συνέχεια φιλτράρει ανά summary_model, ώστε να συγκρίνει διαφορετικά γλωσσικά μοντέλα υπό ίδιες συνθήκες. Με ταξινόμηση ως προς MAP ή nDCG μπορεί να εντοπίσει ποιο μοντέλο αποδίδει καλύτερα, ενώ παράλληλα εξετάζει και τον αριθμό tags/λέξεων που χρησιμοποιήθηκαν, ώστε να έχει σαφή εικόνα του κόστους/κέρδους.

- **Παράδειγμα 2ο : Σύγκριση "ORIGINAL" έναντι περιλήψεων.**

  Με φίλτρο στο summary_type (π.χ. ORIGINAL έναντι extractive/abstractive), ο χρήστης μπορεί να αξιολογήσει αν και πότε οι περιλήψεις οδηγούν σε βελτίωση των μετρικών ανάκτησης. Το σενάριο αυτό υποστηρίζει άμεσα τον βασικό ερευνητικό στόχο της διπλωματικής, δηλαδή τη διερεύνηση της συμβολής της σύνοψης σε retrieval.

- **Παράδειγμα 3ο : Εξαγωγή αποτελεσμάτων για αναφορά και τελικό report.**

  Με τη λειτουργία εξαγωγής σε Excel, τα φιλτραρισμένα αποτελέσματα μπορούν να εξαχθούν και να χρησιμοποιηθούν για την παραγωγή πινάκων, γραφημάτων ή συνοπτικών συγκρίσεων σε επόμενα κεφάλαια (Αποτελέσματα και Συζήτηση).

Κεφάλαιο 7ο: Πειραματικά Αποτελέσματα και Αξιολόγηση

7.1 Πειραματικό πλαίσιο και οργάνωση αποτελεσμάτων

Το κεφάλαιο αυτό βασίζεται σε ένα σύνολο πειραματικών εκτελέσεων, όπου κάθε εκτέλεση αντιστοιχεί σε ένα πλήρες σενάριο: συγκεκριμένη ρύθμιση δεδομένων, πιθανή εφαρμογή μεθόδου σύνοψης, επιλογή μηχανισμού ανάκτησης και ορισμός παραμέτρων αναζήτησης. Για την αποτίμηση της απόδοσης χρησιμοποιούνται καθιερωμένες μετρικές ανάκτησης σε πολλαπλές τιμές cutoff, ώστε να αποτυπώνεται τόσο η ακρίβεια όσο και η ικανότητα εντοπισμού σχετικών εγγράφων. Οι τιμές που παρουσιάζονται στους πίνακες αποτελούν συγκεντρωτικά αποτελέσματα σε επίπεδο εκτέλεσης, προκύπτοντας από τον υπολογισμό μετρικών πάνω στο σύνολο των queries του αντίστοιχου σεναρίου.

Η παρουσίαση οργανώνεται με τρόπο που να επιτρέπει συγκρίσεις ανά κατηγορία πειραμάτων: αρχικά εξετάζονται σενάρια αναφοράς χωρίς εφαρμογή σύνοψης, στη συνέχεια αναλύεται η επίδραση διαφορετικών μεθοδολογιών σύνοψης στην ανάκτηση, και τέλος γίνεται σύγκριση μεταξύ εναλλακτικών μηχανισμών ανάκτησης και ρυθμίσεων. Παράλληλα, όπου υπάρχει εφαρμογή σύνοψης, αξιολογείται και το trade-off μεταξύ συμπίεσης και διατήρησης σημασιολογικής πληροφορίας, ώστε τα συμπεράσματα να μην περιορίζονται μόνο στην ανάκτηση αλλά να καλύπτουν τη συνολική συμπεριφορά του pipeline.

Στην παρούσα μελέτη, δίνεται προτεραιότητα σε ένα προφίλ σύνοψης που διατηρεί υψηλή πιστότητα νοήματος, ενώ παράλληλα παραμένει πρακτικά αξιοποιήσιμο σε εργασίες αναζήτησης και δεν επιβαρύνει υπερβολικά τον χρόνο εκτέλεσης. Με τον τρόπο αυτό, η "καλή" σύνοψη δεν αξιολογείται μόνο ως προς το πόσο μειώνει το κείμενο, αλλά κυρίως ως προς το κατά πόσο διατηρεί τη σημασιολογική πληροφορία και επιτρέπει στο σύστημα ανάκτησης να συνεχίσει να εντοπίζει αποτελεσματικά τα σχετικά έγγραφα. Για τον λόγο αυτό, στις συγκρίσεις που ακολουθούν, οι δείκτες πιστότητας (π.χ. σημασιολογική ομοιότητα) και το υπολογιστικό κόστος (χρόνος) αντιμετωπίζονται ως πρωτεύοντα κριτήρια, ενώ οι μετρικές ανάκτησης (MAP και Recall@100) χρησιμοποιούνται ως δείκτες χρησιμότητας των συνόψεων σε σενάρια αναζήτησης.

7.2 Δομή πειραματικού υλικού

Η ενότητα αυτή συνοψίζει τη δομή του πειραματικού υλικού που θα αναλυθεί στις επόμενες υποενότητες. Ο Πίνακας (7.1) λειτουργεί ως "χάρτης" των εκτελέσεων, παρουσιάζοντας πώς κατανέμονται τα runs ανά μηχανισμό ανάκτησης και ανά κατηγορία σεναρίου (με ή χωρίς σύνοψη), καθώς και ποιες βασικές ρυθμίσεις χρησιμοποιούνται. Με αυτόν τον τρόπο, οι συγκρίσεις που ακολουθούν βασίζονται σε σαφή ομαδοποίηση και κοινό πλαίσιο αναφοράς.

Πίνακας 7.1: Συνοπτική αποτύπωση πειραματικών εκτελέσεων.

Κατηγορία σεναρίου	Μηχανισμός ανάκτησης	Χρήση σύνοψης	Τύπος σύνοψης	Διαφορετικά μοντέλα σύνοψης	Πλήθος πειραμάτων
Χωρίς σύνοψη (baseline)	BM25 (Pyserini)	Όχι	ORIGINAL	—	12
Χωρίς σύνοψη (baseline)	ColBERT	Όχι	ORIGINAL	—	6
Με σύνοψη	BM25 (Pyserini)	Ναι	abstractive	6	72
Με σύνοψη	BM25 (Pyserini)	Ναι	extractive	6	72
Με σύνοψη	BM25 (Pyserini)	Ναι	hybrid	2	13
Με σύνοψη	BM25 (Pyserini)	Ναι	instructive	4	48

Με σύνοψη	ColBERT	Ναι	abstractive	6	33
Με σύνοψη	ColBERT	Ναι	extractive	6	38
Με σύνοψη	ColBERT	Ναι	hybrid	2	4
Με σύνοψη	ColBERT	Ναι	instructive	4	21

7.3 Χρόνοι εκτέλεσης σύνοψης

Ο παρακάτω πίνακας συγκεντρώνει τους συνολικούς χρόνους εκτέλεσης για την παραγωγή συνόψεων, ανά μεθοδολογία και μοντέλο. Για λόγους συγκρισιμότητας, σε όλες τις περιπτώσεις επεξεργάστηκαν 2.592 queries και ο χρόνος αποτυπώνεται σε μορφή HH:MM:SS.

Πίνακας 7.2: Χρόνοι εκτέλεσης ανά μέθοδο και μοντέλο σύνοψης

Τύπος σύνοψης	Μοντέλο σύνοψης	Queries που επεξεργάστηκαν	Χρόνος εκτέλεσης (HH:MM:SS)
instructive	Qwen2_14B_Instruct	2592	22:11:05
instructive	Mistral_7B_Instruct	2592	16:38:42
instructive	Llama3_8B_Instruct	2592	14:52:53
instructive	Qwen2_7B_Instruct	2592	13:26:38
abstractive	PEGASUS_BigPatent	2592	10:08:14
abstractive	LongT5_TGlobal_Base	2592	9:59:54
abstractive	LED_Base_BigPatent	2592	8:39:42
abstractive	BigBird_Pegasus	2592	4:41:44
abstractive	BART_large_CNN	2592	2:15:18
abstractive	T5_Small_HUPD	2592	1:37:24
extractive	google_bert_for_patents	2592	1:15:10
hybrid	PatentSBERTa_V2_TextRank	2592	0:54:55
extractive	PatentSBERTa_V2	2592	0:49:08
extractive	all-mpnet-base-v2	2592	0:47:44
extractive	bge-base-en-v1.5	2592	0:43:14
extractive	e5-base-v2	2592	0:42:48
extractive	SBERT_all-MiniLM-L6-v2	2592	0:27:05

Συνολικά, παρατηρείται ότι οι instructive προσεγγίσεις (LLM-based) απαιτούν τον μεγαλύτερο χρόνο, οι abstractive μέθοδοι κινούνται σε ενδιάμεσους χρόνους, ενώ οι extractive/hybrid είναι σημαντικά ταχύτερες. Αυτός ο πίνακας λειτουργεί ως βάση για την επόμενη σύγκριση "ποιότητας–κόστους", όπου οι χρόνοι συνεκτιμώνται μαζί με τις μετρικές απόδοσης και τα χαρακτηριστικά συμπίεσης.

7.4 Αξιολόγηση μεθόδων σύνοψης ως προς χρόνο, συμπίεση και απόδοση ανάκτησης

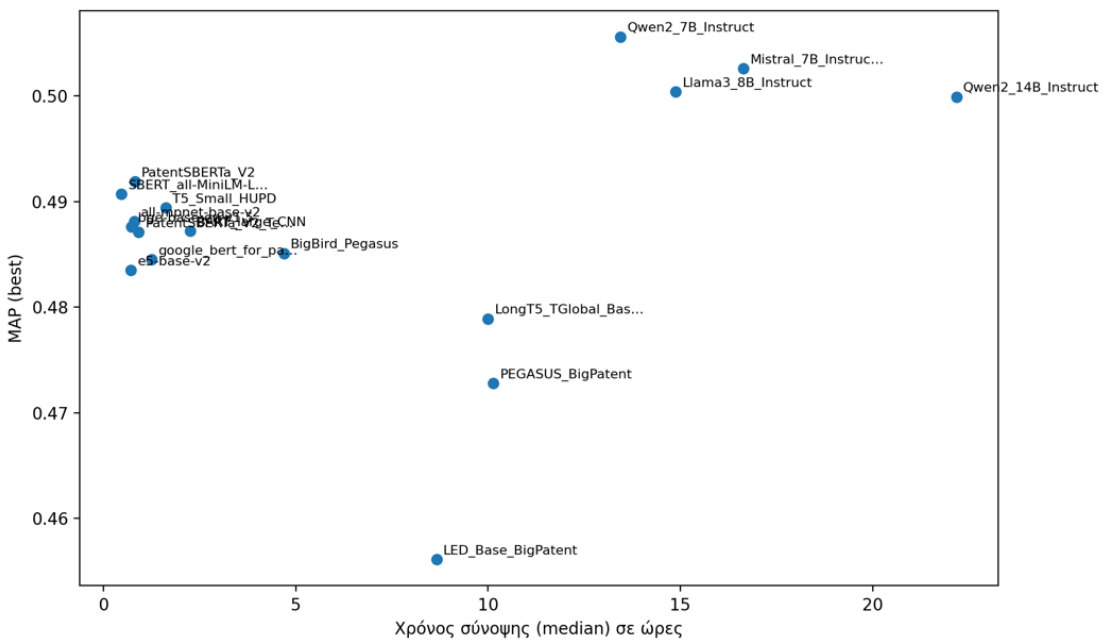
Ο πίνακας (**7.3**) συγκρίνει τις μεθοδολογίες σύνοψης ως προς το υπολογιστικό κόστος (χρόνος εκτέλεσης) και ταυτόχρονα αποτυπώνει δείκτες συμπίεσης και διατήρησης σημασιολογικής πληροφορίας.

Παράλληλα, παρουσιάζεται η επίδραση της κάθε μεθόδου στην ανάκτηση, με κύριους δείκτες το MAP και το Recall@100. Για να αποφευχθεί μεροληψία από μεμονωμένες ρυθμίσεις, δίνονται τόσο διάμεσες τιμές (median) «οι τιμές με την ένδειξη median αντιστοιχούν στη διάμεσο της αντίστοιχης μετρικής, υπολογισμένη πάνω στα runs της ίδιας ομάδας», όσο και οι καλύτερες επιδόσεις (best) που παρατηρήθηκαν ανά μέθοδο/μοντέλο, μαζί με τον μηχανισμό ανάκτησης στον οποίο εμφανίστηκε το αντίστοιχο μέγιστο.

Πίνακας 7.3: Αξιολόγηση μεθόδων σύνοψης ως προς χρόνο, συμπίεση και απόδοση ανάκτησης

Τύπος σύνοψης	Μοντέλο σύνοψης	Queries	Χρόνος σύνοψης (median)	Compress ratio (median)	SBERT similarity (median)	N runs (retrieval)	MAP (median)	Recall@100 (median)	MAP (best)	Engine (best MAP-Recall@100)	Recall@100 (best)	MAP στο best Recall
instructive	Qwen2_14B_Instruct	2592	**22:11:05**	16,1402	0,6906	17	0,435	0,7623	0,4999	BM25 (Pyserini)	**0,8218**	0,4999
instructive	Mistral_7B_Instruct	2592	**16:38:42**	14,5512	0,7307	17	0,4667	0,7824	**0,5026**	BM25 (Pyserini)	**0,8186**	0,4985
instructive	Llama3_8B_Instruct	2592	**14:52:53**	16,4978	0,7235	18	0,4516	0,7655	**0,5004**	BM25 (Pyserini)	**0,8145**	**0,5004**
instructive	Qwen2_7B_Instruct	2592	**13:26:38**	16,3605	0,6999	17	0,4475	0,7813	**0,5056**	BM25 (Pyserini)	**0,8241**	0,4912
hybrid	PatentSBERTa_V2 & Mistral_7B	2592	00:49:08+ 09:42:43 = **10:31:51**	1,9825	0,8007	2	0,4537	0,7576	**0,5052**	BM25 (Pyserini)	**0,8172**	**0,5052**
abstractive	PEGASUS_BigPatent	2592	**10:08:14**	17,7911	0,6944	16	0,444	0,7443	0,4728	BM25 (Pyserini)	0,7762	0,4716
abstractive	LongT5_TGlobal_Base	2592	**09:59:54**	11,045	0,779	15	0,4505	0,7499	0,4789	BM25 (Pyserini)	0,7826	0,4789
abstractive	LED_Base_BigPatent	2592	**08:39:42**	18,1309	0,631	21	0,384	0,6927	0,4561	BM25 (Pyserini)	0,7724	0,4561

abstractive	BigBird_Pegasus	2592	**04:41:44**	21,4475	0,8206	16	0,4545	0,7578	0,4851	BM25 (Pyserini)	0,7894	0,4801
abstractive	BART_large_CNN	2592	**02:15:18**	28,8364	0,8566	19	0,449	0,7518	0,4872	BM25 (Pyserini)	0,7932	0,4872
abstractive	T5_Small_HUPD	2592	**01:37:24**	18,9854	0,8298	17	0,4501	0,754	0,4894	BM25 (Pyserini)	0,7979	0,4894
extractive	google_bert_for_patents	2592	**01:15:10**	17,0395	0,9292	15	0,4468	0,7413	0,4845	BM25 (Pyserini)	0,785	0,4841
hybrid	PatentSBERTa_V2_TextRank	2592	**00:54:55**	5,4184	0,9179	15	0,4739	0,7785	0,4871	BM25 (Pyserini)	0,7965	0,486
extractive	PatentSBERTa_V2	2592	**00:49:08**	10,2626	0,835	35	0,3869	0,6848	0,4919	BM25 (Pyserini)	0,7944	0,4894
extractive	all-mpnet-base-v2	2592	**00:47:44**	11,3278	0,84	15	0,4722	0,7733	0,4881	BM25 (Pyserini)	0,7929	0,4881
extractive	bge-base-en-v1.5	2592	**00:43:14**	12,1595	0,9058	15	0,4681	0,7684	0,4876	BM25 (Pyserini)	0,794	0,4876
extractive	e5-base-v2	2592	**00:42:48**	24,9486	0,952	15	0,4421	0,7421	0,4835	BM25 (Pyserini)	0,7808	0,4835
extractive	SBERT_all-MiniLM-L6-v2	2592	**00:27:57**	11,1712	0,8317	15	0,474	0,7779	0,4907	BM25 (Pyserini)	0,7959	0,4907



Γράφημα 1: Χρόνος σύνοψης (median) vs MAP(best) ανά μοντέλο (BM25)

Η συγκεκριμένη υβριδική ρύθμιση «σκιασμένη με κίτρινο χρώμα» (PatentSBERTa_V2 σε συνδυασμό με Mistral_7B) ξεχωρίζει ως το καλύτερο σενάριο ισορροπίας μεταξύ χρόνου και αποτελεσματικότητας. Με συνολικό χρόνο **10:31:51** για την επεξεργασία **2.592** queries, επιτυγχάνει υψηλή απόδοση ανάκτησης, με **MAP 0,5052** και **Recall@100 0,8172**, γεγονός που δείχνει ότι διατηρεί ισχυρή ικανότητα εντοπισμού σχετικών εγγράφων. Παράλληλα, ο δείκτης συμπίεσης **1,9825** «περίπου **50%** μείωση του περιεχομένου» σε συνδυασμό με SBERT similarity **0,8007** «Τιμή γύρω στο 0,80 θεωρείται σχετικά υψηλή, άρα η σύνοψη παραμένει κοντά στο νόημα του αρχικού κειμένου.» υποδηλώνει ότι η σύνοψη μειώνει το μέγεθος του κειμένου χωρίς σημαντική απώλεια σημασιολογικής πληροφορίας. Το αποτέλεσμα αυτό προκύπτει στο σενάριο ανάκτησης BM25 (Pyserini), ενισχύοντας τη θέση της συγκεκριμένης υβριδικής επιλογής ως αποδοτικής λύσης σε επίπεδο pipeline.

7.5 Επίδραση των συνόψεων στην αποτελεσματικότητα ανάκτησης (ποσοστιαία μεταβολή έναντι baseline)

Η παρούσα ενότητα εξετάζει τη χρησιμότητα των περιλήψεων στις εργασίες αναζήτησης μέσω σύγκρισης με αντίστοιχες εκτελέσεις baseline, υπό ίδιες συνθήκες ανάκτησης (ίδιος μηχανισμός και ίδια διαμόρφωση πεδίων query). Η σύγκριση εκφράζεται ως ποσοστιαία μεταβολή και αποτυπώνει την επίδραση της σύνοψης στην απόδοση ανάκτησης, με βασικές μετρικές το MAP και το Recall@100.

Ο Πίνακας 7.4 δεν αποτελεί πλήρη παράθεση όλων των εκτελέσεων, αλλά μια συμπυκνωμένη "προβολή μέγιστου οφέλους": για κάθε συνδυασμό μηχανισμού ανάκτησης και QUERY TAGS USED εμφανίζεται η μέθοδος και το μοντέλο σύνοψης που επιτυγχάνει τη μεγαλύτερη ποσοστιαία βελτίωση έναντι του baseline. Με αυτόν τον τρόπο αναδεικνύονται οι περιπτώσεις όπου η σύνοψη λειτουργεί ως παράγοντας ενίσχυσης της αναζήτησης, ενώ το εύρος του οφέλους διαφοροποιείται ανάλογα με το περιεχόμενο και την πυκνότητα των πεδίων που συνθέτουν το query.

Η ταυτόχρονη παρουσίαση της μεταβολής σε MAP και Recall@100 επιτρέπει να φανεί αν το όφελος αφορά κυρίως την ποιότητα κατάταξης (MAP), την κάλυψη σχετικών εγγράφων (Recall@100), ή και τα δύο. Επιπλέον, όταν οι δύο μετρικές δεν κινούνται παράλληλα, αυτό υποδηλώνει διαφορετικό τύπο επίδρασης της σύνοψης στο retrieval και βοηθά στην ερμηνεία του πότε μια σύνοψη είναι "χρήσιμη" για την αναζήτηση και με ποιον τρόπο.

Πίνακας 7.4: Επίδραση της σύνοψης στην ανάκτηση σε σύγκριση με baseline (ΔMAP, ΔRecall@100)

Μηχανισμός ανάκτησης	Tags Query Used	Τύπος σύνοψης	Μοντέλο σύνοψης	MAP baseline	MAP summarized	%ΔMAP (Baseline→Summary)	Recall@100 baseline	Recall@100 summarized	%ΔRecall@100 (Baseline→Su
BM25 (Pyserini)	FIRST-CLAIM(all)	hybrid	PatentSBERTa_V2_ TextRank	0,4006	0,4545	13,45%	0,6814	0,7552	10,83%
ColBERT	DESCRIPTION(all)	instructive	Mistral_7B_Instruct	0,3020	0,3348	10,86%	0,5974	0,6274	5,02%
BM25 (Pyserini)	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)	instructive	Mistral_7B_Instruct	0,4504	0,4857	7,84%	0,7520	0,8105	7,78%
ColBERT	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)	abstractive	BART_large_CNN	0,3846	0,4132	7,44%	0,6671	0,7037	5,49%
ColBERT	CLAIMS(220), DESCRIPTION(220)	instructive	Mistral_7B_Instruct	0,3361	0,3580	6,52%	0,6460	0,6769	4,78%
BM25 (Pyserini)	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)	instructive	Qwen2_7B_Instruct	0,4800	0,5056	5,33%	0,7887	0,8198	3,94%
BM25 (Pyserini)	TITLE(all), ABSTRACT(all), CLAIMS(all)	instructive	Llama3_8B_Instruct	0,4783	0,4998	4,50%	0,7782	0,8106	4,16%
BM25 (Pyserini)	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)	hybrid	PatentSBERTa_V2_ TextRank	0,4565	0,4739	3,81%	0,7585	0,7743	2,08%
BM25 (Pyserini)	TITLE(all), ABSTRACT(all), DESCRIPTION(all)	instructive	Mistral_7B_Instruct	0,4796	0,4977	3,77%	0,7887	0,8119	2,94%
BM25 (Pyserini)	DESCRIPTION(all)	hybrid	PatentSBERTa_V2_ TextRank	0,4568	0,4732	3,59%	0,7588	0,7752	2,16%
BM25 (Pyserini)	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)	instructive	Mistral_7B_Instruct	0,4840	0,5008	3,47%	0,7985	0,8077	1,15%
BM25 (Pyserini)	CLAIMS(all)	instructive	Mistral_7B_Instruct	0,4610	0,4667	1,24%	0,7630	0,7910	3,67%
BM25 (Pyserini)	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)	abstractive	T5_Small_HUPD	0,4776	0,4816	0,84%	0,7799	0,7832	0,42%

ColBERT	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)	abstractive	T5_Small_HUPD	0,4056	0,4082	0,64%	0,6886	0,6896	0,15%
BM25 (Pyserini)	ALL TAGS & WORDS	instructive	Mistral_7B_Instruct	0,4959	0,4985	0,52%	0,7974	0,8186	2,66%
ColBERT	CLAIMS(all)	extractive	PatentSBERTa_V2	0,3029	0,3043	0,46%	0,6143	0,5728	-6,76%
BM25 (Pyserini)	TITLE(all), ABSTRACT(all)	instructive	Mistral_7B_Instruct	0,4748	0,4749	0,02%	0,7816	0,7824	0,10%

Headers Πίνακα 7.4 (baseline vs σύνοψη, με ΔMAP και ΔRecall@100)

- Μηχανισμός ανάκτησης: Ο retrieval engine που εκτέλεσε την αναζήτηση (BM25/Pyserini ή ColBERT). Κάθε γραμμή αφορά σύγκριση μέσα στον ίδιο μηχανισμό.
- Tags Query Used: Η διαμόρφωση των πεδίων που χρησιμοποιήθηκαν ως query (δηλαδή ποια tags "μπαίνουν" στο ερώτημα). Οι ενδείξεις (all), (100), (250) κ.λπ. δείχνουν το όριο λέξεων/tokens που χρησιμοποιήθηκαν από κάθε tag.
- Τύπος σύνοψης: Η κατηγορία της μεθοδολογίας σύνοψης που εφαρμόστηκε (extractive, abstractive, hybrid, instructive).
- Μοντέλο σύνοψης: Το συγκεκριμένο μοντέλο/αλγόριθμος που παρήγαγε τη σύνοψη στο run (π.χ. Mistral_7B_Instruct, PatentSBERTa_V2_TextRank, BART_large_CNN κ.λπ.).
- MAP baseline: Η τιμή του MAP στο baseline run, δηλαδή στο σενάριο "χωρίς σύνοψη" με τις ίδιες συνθήκες ανάκτησης και τα ίδια query tags.
- MAP summarized: Η τιμή του MAP στο αντίστοιχο run "με σύνοψη", όπου το ORIGINAL κείμενο έχει αντικατασταθεί/συμπυκνωθεί από σύνοψη, αλλά οι υπόλοιπες συνθήκες παραμένουν αντίστοιχες.
- %ΔMAP (Baseline→Summary): Η ποσοστιαία μεταβολή του MAP όταν περνάμε από baseline σε summarized. Υπολογίζεται ως:
- %ΔMAP = 100 × (MAP_summarized − MAP_baseline) / MAP_baseline. Θετική τιμή σημαίνει βελτίωση της ακρίβειας κατάταξης με σύνοψη, αρνητική σημαίνει επιδείνωση.
- Recall@100 baseline: Η τιμή του Recall@100 στο baseline run (χωρίς σύνοψη), δηλαδή πόσο καλά ανακτώνται σχετικά έγγραφα μέσα στα πρώτα 100 αποτελέσματα.
- Recall@100 summarized: Η τιμή του Recall@100 στο summarized run (με σύνοψη), υπό αντίστοιχες συνθήκες και ίδια query tags.
- %ΔRecall@100 (Baseline→Summary): Η ποσοστιαία μεταβολή του Recall@100 από baseline σε summarized. Υπολογίζεται ως:
- %ΔRecall@100 = 100 × (Recall@100_summarized − Recall@100_baseline) / Recall@100_baseline.
  Θετική τιμή σημαίνει ότι η σύνοψη βοηθά στην κάλυψη (φέρνει περισσότερα σχετικά έγγραφα μέσα στα top-100), ενώ αρνητική ότι μειώνει την κάλυψη.

Από τον Πίνακα 7.4 προκύπτει ότι, σε αρκετές διαμορφώσεις πεδίων query, η χρήση σύνοψης μπορεί να βελτιώσει μετρήσιμα την απόδοση ανάκτησης έναντι του baseline, τόσο ως προς το MAP όσο και ως προς το Recall@100. Οι μεγαλύτερες αυξήσεις εμφανίζονται όταν το query στηρίζεται σε συγκεκριμένα, "πυκνά" πεδία (π.χ. FIRST-CLAIM ή στοχευμένα τμήματα DESCRIPTION/CLAIMS), όπου η σύνοψη φαίνεται να ενισχύει τη διακριτική πληροφορία που αξιοποιεί ο μηχανισμός ανάκτησης. Παράλληλα, σε σενάρια με πιο εκτεταμένη περιγραφή των πεδίων (πολλά tags μαζί), οι βελτιώσεις τείνουν να είναι μικρότερες αλλά πιο σταθερές, υποδηλώνοντας ότι το όφελος της σύνοψης εξαρτάται από το πόσο "συμπυκνωμένο" ή "θορυβώδες" είναι το αρχικό query.

7.6 Επίδραση εμπλουτισμού των query tags με TITLE και ABSTRACT στην απόδοση ανάκτησης

Στο πλαίσιο της αξιολόγησης των πειραμάτων ανάκτησης, δίνεται ιδιαίτερη έμφαση όχι μόνο στο μοντέλο ή στον μηχανισμό αναζήτησης, αλλά και στον τρόπο διαμόρφωσης του ίδιου του query. Για τον λόγο αυτό, η παρούσα ενότητα επικεντρώνεται στον εμπλουτισμό των ερωτημάτων με τα πεδία TITLE και ABSTRACT και εξετάζει κατά πόσο αυτή η προσθήκη μπορεί να βελτιώσει μετρήσιμα την απόδοση ανάκτησης σε σχέση με αντίστοιχα σενάρια όπου χρησιμοποιείται μόνο ένα βασικό πεδίο (π.χ. DESCRIPTION, CLAIMS, FIRST-CLAIM ή DETAILED-DESCRIPTION+BRIEF-DESCRIPTION). Με απλά λόγια, το ίδιο "σενάριο" αναζήτησης εκτελείται σε δύο εκδοχές: (α) με query που περιέχει μόνο το αρχικό πεδίο και (β) με query όπου στο ίδιο αρχικό πεδίο προστίθενται τα TITLE και ABSTRACT. Στόχος είναι να διαπιστωθεί αν η προσθήκη αυτών των πεδίων οδηγεί σε μετρήσιμη βελτίωση στην ποιότητα ανάκτησης, δηλαδή αν ο μηχανισμός αναζήτησης εντοπίζει πιο σωστά και πιο ψηλά στη λίστα τα σχετικά έγγραφα.

Ο λόγος που επιλέγονται ειδικά τα TITLE και ABSTRACT είναι ότι, στη δομή μιας πατέντας, αποτελούν τα πιο "πυκνά" και περιγραφικά πεδία: συνήθως συμπυκνώνουν το θέμα της εφεύρεσης με υψηλή πληροφοριακή αξία και μικρό θόρυβο. Αντίθετα, πεδία όπως DESCRIPTION ή CLAIMS είναι πολύ εκτενή και συχνά περιέχουν επαναλήψεις, νομική/τυπική διατύπωση και λεπτομέρειες που δεν είναι πάντα διακριτικές για την αναζήτηση. Έτσι, ο εμπλουτισμός του query μπορεί να λειτουργήσει σαν "άγκυρα" σημασιολογικού προσανατολισμού: το TITLE και το ABSTRACT εισάγουν τους βασικούς όρους και την κεντρική έννοια, ενώ το αρχικό πεδίο (π.χ. DESCRIPTION) διατηρεί τη λεπτομέρεια που μπορεί να χρειάζεται σε πιο εξειδικευμένες αναζητήσεις.

Μεθοδολογικά, η σύγκριση σχεδιάζεται ώστε να είναι δίκαιη και ελεγχόμενη: για κάθε μηχανισμό ανάκτησης (BM25 ή ColBERT) και για κάθε μοντέλο/τύπο σύνοψης όπου υπάρχουν διαθέσιμες εκτελέσεις και για τις δύο εκδοχές query, διατηρούνται σταθερά όλα τα υπόλοιπα στοιχεία του πειράματος (ίδιο index, ίδιο σύνολο queries, ίδια ρύθμιση παραμέτρων ανάκτησης, ίδιο summary_type και ίδιο summary_model). Η μοναδική διαφορά είναι η σύνθεση του query: από "μονό-πεδίο" μεταβαίνουμε σε "εμπλουτισμένο" query με TITLE+ABSTRACT. Η βελτίωση αποτυπώνεται με τις ίδιες βασικές μετρικές που χρησιμοποιούνται σε όλο το κεφάλαιο: MAP, που εκφράζει την ποιότητα της κατάταξης των σχετικών εγγράφων, και Recall@100, που εκφράζει την κάλυψη σχετικών εγγράφων μέσα στα πρώτα 100 αποτελέσματα.

Η επιλογή να παρουσιαστεί η βελτίωση ως ποσοστιαία μεταβολή (percent increase) δεν είναι απλώς θέμα παρουσίασης. Επιτρέπει να συγκριθούν σενάρια με διαφορετικές "αφετηρίες" (baseline επιδόσεις) σε κοινή κλίμακα: μια απόλυτη αύξηση 0,01 στο MAP μπορεί να έχει διαφορετική σημασία αν το αρχικό MAP είναι 0,48 ή 0,30. Με την ποσοστιαία μεταβολή, φαίνεται πιο καθαρά πότε ο εμπλουτισμός προσφέρει ουσιαστικό κέρδος και πότε η επίδραση είναι μικρή ή αμελητέα. Παράλληλα, η ταυτόχρονη

εξέταση MAP και Recall@100 δείχνει και τον "τύπο" της βελτίωσης: σε ορισμένες περιπτώσεις ο εμπλουτισμός μπορεί να αυξάνει κυρίως το Recall (φέρνει περισσότερα σχετικά στο top-100), ενώ σε άλλες μπορεί να βελτιώνει περισσότερο το MAP (τα σχετικά ανεβαίνουν ψηλότερα στη σειρά).

Η ανάλυση οργανώνεται σε τέσσερα υπο-σενάρια, καθένα από τα οποία αντιστοιχεί σε διαφορετική λογική "πυρήνα" του query. Στο DESCRIPTION, ελέγχεται αν ένα γενικά περιγραφικό αλλά εκτενές πεδίο ωφελείται από τον "συμπυκνωτή" TITLE+ABSTRACT. Στο CLAIMS, εξετάζεται αν η πιο τυπική και νομική γλώσσα των αξιώσεων γίνεται πιο αναζητήσιμη όταν πλαισιώνεται από συνοπτική περιγραφή. Στο FIRST-CLAIM, δοκιμάζεται αν η εστιασμένη πληροφορία του πρώτου claim ενισχύεται περαιτέρω ή αν είναι ήδη αρκετά διακριτική. Τέλος, στο DETAILED-DESCRIPTION+BRIEF-DESCRIPTION, αξιολογείται μια ενδιάμεση προσέγγιση που επιχειρεί να ισορροπήσει μεταξύ "πλούσιου περιεχομένου" και "σχετικής συνοπτικότητας", πριν και μετά την προσθήκη TITLE+ABSTRACT.

Σε επίπεδο ερμηνείας, η συγκεκριμένη ενότητα έχει και πρακτική αξία: οδηγεί σε ένα απλό, εφαρμόσιμο συμπέρασμα για το "πώς να σχηματίζουμε queries" όταν έχουμε διαθέσιμη δομημένη πληροφορία. Αν αποδειχθεί ότι η προσθήκη TITLE+ABSTRACT βελτιώνει σταθερά τα αποτελέσματα, τότε προτείνεται ως κανόνας σχεδιασμού queries για μελλοντικές εκτελέσεις. Αν, αντίθετα, η βελτίωση εμφανίζεται μόνο σε ορισμένα πεδία ή μόνο σε συγκεκριμένους μηχανισμούς ανάκτησης, τότε προκύπτει μια πιο στοχευμένη οδηγία: ο εμπλουτισμός δεν είναι πάντα χρήσιμος, αλλά εξαρτάται από το πεδίο που χρησιμοποιείται ως βάση και από τη φύση του retriever.

7.6.1 Επίδραση προσθήκης TITLE και ABSTRACT σε DESCRIPTION queries

Στην παρούσα υποενότητα εξετάζεται η επίδραση του εμπλουτισμού ενός query που βασίζεται αποκλειστικά στο DESCRIPTION, με την προσθήκη των πεδίων TITLE και ABSTRACT. Ο Πίνακας που ακολουθεί αποτυπώνει την απόδοση σε MAP και Recall@100 για τις δύο εκδοχές, καθώς και την ποσοστιαία μεταβολή τους, ώστε να φανεί αν ο εμπλουτισμός λειτουργεί ως σταθερά ωφέλιμη ενίσχυση ή αν το όφελος εξαρτάται από το μοντέλο σύνοψης και τον μηχανισμό ανάκτησης.

Πίνακας 7.5: Επίδραση προσθήκης TITLE και ABSTRACT σε DESCRIPTION-based queries (MAP, Recall@100)

			DESCRIPTION (all)			TITLE (all), ABSTRACT (all), DESCRIPTION (all)				
index type	summary type	summary model	Qlength	recall@100	MAP	Qlength	recall@100	MAP	% Αυξησης recall@100	% Αυξησης MAP
BM25	instructive	Qwen2_7B_Instruct	142	**0,7784**	0,4358	247	**0,82**	0,4943	5,34%	13,42%
BM25	instructive	Qwen2_14B_Instruct	142	**0,7511**	0,3942	248	**0,8129**	0,4899	8,23%	**24,28%**
BM25	instructive	Mistral_7B_Instruct	140	**0,7822**	0,4527	245	**0,8119**	0,4977	3,80%	9,94%
BM25	instructive	Llama3_8B_Instruct	141	**0,7263**	0,4093	246	**0,8019**	0,489	10,41%	**19,47%**

BM25	hybrid	PatentSBERTa_V2_TextRank	417	**0,7752**	0,4732	435	**0,7942**	0,4866	2,45%	2,83%
BM25	abstractive	BigBird_Pegasus	78	**0,7161**	0,4197	183	**0,7894**	0,4801	10,24%	14,39%
BM25	abstractive	BART_large_CNN	50	**0,5562**	0,327	156	**0,7889**	0,4813	41,84%	**47,19%**
BM25	abstractive	T5_Small_HUPD	77	**0,6196**	0,3607	182	**0,7885**	0,4766	27,26%	32,13%
BM25	extractive	PatentSBERTa_V2	242	**0,748**	0,4562	342	**0,7871**	0,4879	5,23%	6,95%
BM25	extractive	SBERT_all-MiniLM-L6-v2	240	**0,7417**	0,4506	338	**0,7856**	0,485	5,92%	7,63%
BM25	extractive	bge-base-en-v1.5	217	**0,7235**	0,4381	300	**0,7824**	0,4773	8,14%	8,95%
BM25	extractive	all-mpnet-base-v2	237	**0,7368**	0,446	334	**0,781**	0,4831	6,00%	8,32%
BM25	abstractive	PEGASUS_BigPatent	172	**0,7413**	0,4365	240	**0,7747**	0,4698	4,51%	7,63%
BM25	abstractive	LongT5_TGlobal_Base	213	**0,6889**	0,4034	317	**0,7684**	0,4646	11,54%	**15,17%**
BM25	extractive	google_bert_for_patents	141	**0,6517**	0,3939	203	**0,756**	0,4643	16,00%	**17,87%**
BM25	abstractive	LED_Base_BigPatent	169	**0,7012**	0,3857	246	**0,7522**	0,4364	7,27%	13,14%
BM25	extractive	e5-base-v2	80	**0,631**	0,3827	128	**0,7503**	0,4574	18,91%	**19,52%**
Colbert	abstractive	BART_large_CNN	512	0,4642	0,2459	512	**0,6911**	0,4142	48,88%	**68,44%**

Με βάση τον πίνακα 7.5, το κεντρικό συμπέρασμα είναι ότι ο εμπλουτισμός των DESCRIPTION queries με TITLE και ABSTRACT οδηγεί σε συστηματική βελτίωση τόσο στο MAP όσο και στο Recall@100. Η βελτίωση **εμφανίζεται σε όλα τα μοντέλα** που εξετάζονται, με κέρδη από ήπια έως πολύ υψηλά ταυτόχρονα και την ποιότητα κατάταξης (MAP) και την κάλυψη σχετικών εγγράφων (Recall@100), και δεν περιορίζεται σε ένα συγκεκριμένο είδος σύνοψης ή σε ένα μόνο μοντέλο. Συνολικά, ο εμπλουτισμός του query με TITLE και ABSTRACT λειτουργεί ως πρακτικός τρόπος να "ισχυροποιείται" το περιεχόμενο του ερωτήματος και να αυξάνεται η αποτελεσματικότητα της αναζήτησης, σε σύγκριση με τη χρήση μόνο του DESCRIPTION.

7.6.2 Επίδραση προσθήκης TITLE και ABSTRACT σε CLAIMS queries

Επίδραση του εμπλουτισμού ενός query που βασίζεται αποκλειστικά στο CLAIMS, με την προσθήκη των πεδίων TITLE και ABSTRACT. Ο Πίνακας που ακολουθεί αποτυπώνει την απόδοση σε MAP και Recall@100 για τις δύο εκδοχές.

Πίνακας 7.6: Επίδραση προσθήκης TITLE και ABSTRACT σε CLAIM-based queries (MAP, Recall@100)

			CLAIMS(all)			TITLE(all), ABSTRACT(all), CLAIMS(all)				
index type	summary type	summary model	Qlength	recall@100	MAP	Qlength	recall@100	MAP	% Αύξησης recall@100	% Αύξησης MAP
BM25	instructive	Qwen2_14B_Instruct	161	0,762	0,415	265	0,819	0,496	7,41%	**19,51%**
BM25	instructive	Qwen2_7B_Instruct	158	0,781	0,448	263	0,813	0,499	4,04%	11,42%
BM25	instructive	Llama3_8B_Instruct	159	0,786	0,458	264	0,811	0,5	3,18%	9,13%
BM25	instructive	Mistral_7B_Instruct	156	0,791	0,467	260	0,81	0,495	2,38%	6,06%
BM25	abstractive	BART_large_CNN	77	0,718	0,425	182	0,788	0,481	9,73%	13,17%
BM25	abstractive	T5_Small_HUPD	106	0,75	0,447	210	0,787	0,482	4,89%	7,88%
BM25	abstractive	BigBird_Pegasus	97	0,727	0,428	201	0,786	0,479	8,16%	11,91%
BM25	extractive	SBERT_all-MiniLM-L6-v2	132	0,724	0,427	229	0,783	0,479	8,11%	12,10%
BM25	extractive	all-mpnet-base-v2	126	0,719	0,425	221	0,782	0,481	8,70%	13,11%
BM25	abstractive	LongT5_TGlobal_Base	140	0,722	0,423	243	0,782	0,476	8,35%	12,54%
BM25	extractive	PatentSBERTa_V2	133	0,723	0,426	231	0,782	0,475	8,23%	11,48%
BM25	hybrid	PatentSBERTa_V2_TextRank	141	0,721	0,428	242	0,782	0,476	8,39%	11,31%
BM25	extractive	bge-base-en-v1.5	98	0,711	0,418	176	0,779	0,479	9,51%	14,53%
BM25	extractive	google_bert_for_patents	53	0,681	0,397	113	0,775	0,472	13,76%	**18,82%**
BM25	abstractive	PEGASUS_BigPatent	164	0,722	0,416	233	0,765	0,457	5,94%	9,90%
BM25	extractive	e5-base-v2	38	0,653	0,368	83	0,761	0,464	16,60%	**26,04%**
BM25	abstractive	LED_Base_BigPatent	162	0,693	0,384	239	0,745	0,436	7,52%	13,54%

Με βάση τον πίνακα 7.6, το κεντρικό συμπέρασμα είναι ότι ο εμπλουτισμός των CLAIM queries με TITLE και ABSTRACT οδηγεί σε συστηματική βελτίωση τόσο στο MAP όσο και στο Recall@100. Η βελτίωση εμφανίζεται σε όλα τα μοντέλα που εξετάζονται.

7.6.3 Επίδραση προσθήκης TITLE και ABSTRACT σε FIRST-CLAIM queries

Επίδραση του εμπλουτισμού ενός query που βασίζεται αποκλειστικά στο FIRST-CLAIM, με την προσθήκη των πεδίων TITLE και ABSTRACT. Ο Πίνακας που ακολουθεί αποτυπώνει την απόδοση σε MAP και Recall@100 για τις δύο εκδοχές.

Πίνακας 7.7: Επίδραση προσθήκης TITLE και ABSTRACT σε FIRST CLAIM-based queries (MAP, Recall@100)

			FIRST-CLAIM(all)			TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)				
index type	summary type	summary model	Qlength	recall@100	MAP	Qlength	recall@100	MAP	% Αύξησης recall@100	% Αύξησης MAP
BM25	instructive	Qwen2_14B_Instruct	165	0,7477	0,435	205	0,7859	0,4797	5,11%	**10,28%**
BM25	instructive	Qwen2_7B_Instruct	163	0,7542	0,4447	204	0,7855	0,4807	4,15%	8,10%
BM25	instructive	Mistral_7B_Instruct	165	0,7545	0,4514	207	0,7839	0,4807	3,90%	6,49%
BM25	abstractive	BigBird_Pegasus	128	0,7382	0,439	193	0,7837	0,4783	6,16%	8,95%
BM25	instructive	Llama3_8B_Instruct	164	0,7488	0,4452	206	0,7837	0,4785	4,66%	7,48%
BM25	abstractive	T5_Small_HUPD	134	0,754	0,4501	199	0,7832	0,4816	3,87%	7,00%
BM25	abstractive	LongT5_TGlobal_Base	192	0,7429	0,4394	204	0,7826	0,4789	5,34%	8,99%
BM25	abstractive	BART_large_CNN	118	0,7518	0,449	193	0,7812	0,4801	3,91%	6,93%
BM25	extractive	SBERT_all-MiniLM-L6-v2	212	0,7526	0,4497	207	0,7804	0,4781	3,69%	6,32%
BM25	extractive	PatentSBERTa_V2	212	0,7522	0,4514	209	0,7803	0,4785	3,74%	6,00%
BM25	hybrid	PatentSBERTa_V2_TextRank	285	0,7552	0,4545	224	0,7803	0,4781	3,32%	5,19%
BM25	extractive	bge-base-en-v1.5	198	0,7518	0,4505	189	0,7774	0,4768	3,41%	5,84%
BM25	extractive	google_bert_for_patents	159	0,7413	0,4457	160	0,7731	0,4754	4,29%	6,66%
BM25	extractive	e5-base-v2	133	0,7421	0,4417	144	0,7668	0,4723	3,33%	6,93%
BM25	abstractive	PEGASUS_BigPatent	146	0,71	0,4122	141	0,759	0,4572	6,90%	**10,92%**
BM25	abstractive	LED_Base_BigPatent	156	0,6572	0,353	161	0,7311	0,4198	11,24%	**18,92%**

Με βάση τον πίνακα 7.7, το κεντρικό συμπέρασμα είναι ότι ο εμπλουτισμός των FIRST-CLAIM queries με TITLE και ABSTRACT οδηγεί σε συστηματική βελτίωση τόσο στο MAP όσο και στο Recall@100. Η βελτίωση εμφανίζεται σε όλα τα μοντέλα που εξετάζονται.

7.6.4 Επίδραση προσθήκης TITLE και ABSTRACT σε DETAILED/BRIEF-DESCRIPTION queries

Επίδραση του εμπλουτισμού ενός query που βασίζεται αποκλειστικά στο DETAIL_DESCRIPTION/BRIEF DESCRIPTION, με την προσθήκη των πεδίων TITLE και ABSTRACT. Ο Πίνακας που ακολουθεί αποτυπώνει την απόδοση σε MAP και Recall@100 για τις δύο εκδοχές.

Πίνακας 7.8: Επίδραση προσθήκης TITLE και ABSTRACT σε DETAILED/BRIEF-DESCRIPTION-based queries (MAP, Recall@100)

			DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)			TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)				
index type	**summary type**	**summary model**	**Qlength**	**recall@100**	**MAP**	**Qlength**	**recall@100**	**MAP**	**% Αύξησης recall@100**	**% Αύξησης MAP**
BM25	instructive	Qwen2_7B_Instruct	283	0,778	0,436	373	0,82	0,483	5,40%	10,84%
BM25	instructive	Mistral_7B_Instruct	278	0,782	0,453	368	0,811	0,486	3,70%	7,31%
BM25	instructive	Qwen2_14B_Instruct	283	0,751	0,394	373	0,809	0,475	7,83%	**20,50%**
BM25	instructive	Llama3_8B_Instruct	279	0,726	0,409	368	0,794	0,473	9,45%	**15,54%**
BM25	hybrid	PatentSBERTa_V2_TextRank	443	0,774	0,474	445	0,779	0,478	0,54%	0,93%
BM25	abstractive	BigBird_Pegasus	154	0,716	0,42	244	0,778	0,47	8,67%	12,01%
BM25	extractive	SBERT_all-MiniLM-L6-v2	407	0,745	0,45	426	0,778	0,477	4,44%	6,00%
BM25	extractive	PatentSBERTa_V2	411	0,751	0,453	429	0,776	0,478	3,36%	5,36%
BM25	abstractive	PEGASUS_BigPatent	343	0,741	0,436	410	0,769	0,46	3,71%	5,50%
BM25	extractive	bge-base-en-v1.5	378	0,727	0,439	404	0,768	0,468	5,64%	6,53%
BM25	abstractive	BART_large_CNN	99	0,555	0,327	189	0,762	0,463	37,20%	41,54%
BM25	abstractive	T5_Small_HUPD	151	0,619	0,361	242	0,761	0,458	22,94%	**27,01%**
BM25	abstractive	LongT5_TGlobal_Base	363	0,692	0,408	400	0,752	0,455	8,75%	11,36%
BM25	abstractive	LED_Base_BigPatent	337	0,702	0,386	406	0,739	0,424	5,30%	9,85%
BM25	extractive	google_bert_for_patents	266	0,654	0,394	311	0,731	0,447	11,77%	13,37%
BM25	extractive	e5-base-v2	158	0,631	0,383	202	0,723	0,442	14,53%	**15,58%**

Με βάση τον πίνακα 7.8, το κεντρικό συμπέρασμα είναι ότι ο εμπλουτισμός των DETAILED/BRIEF - DESCRIPTION queries με TITLE και ABSTRACT οδηγεί σε συστηματική βελτίωση τόσο στο MAP όσο και στο Recall@100. Η βελτίωση εμφανίζεται σε όλα τα μοντέλα που εξετάζονται.

Συνολικά, τα αποτελέσματα της Ενότητας 7.6 αναδεικνύουν με συνέπεια ότι ο εμπλουτισμός των queries με τα πεδία TITLE και ABSTRACT αποτελεί μια αποτελεσματική στρατηγική ενίσχυσης της ανάκτησης. Σε όλες τις συγκριτικές περιπτώσεις που εξετάστηκαν, η μετάβαση από "μονο-πεδίο" διατυπώσεις (π.χ. DESCRIPTION, CLAIMS, FIRST-CLAIM ή συνδυασμούς περιγραφικών πεδίων) σε αντίστοιχες διατυπώσεις που περιλαμβάνουν επιπλέον TITLE και ABSTRACT οδηγεί σε συστηματική βελτίωση τόσο της ποιότητας κατάταξης (MAP) όσο και της κάλυψης σχετικών αποτελεσμάτων (Recall@100). Το εύρημα αυτό είναι ιδιαίτερα σημαντικό, καθώς δείχνει ότι η προσθήκη συνοπτικών αλλά υψηλής διακριτικής ισχύος πληροφοριών από τον τίτλο και την περίληψη του εγγράφου ενισχύει την εκφραστικότητα του ερωτήματος και βελτιώνει την ικανότητα του μηχανισμού ανάκτησης να εντοπίζει και να κατατάσσει αποτελεσματικότερα τα σχετικά έγγραφα, ανεξάρτητα από τον τύπο σύνοψης ή το μοντέλο που χρησιμοποιείται.

7.7 Trade-off συμπίεσης, πιστότητας και απόδοσης ανάκτησης ανά μηχανισμό retrieval και τύπο σύνοψης

Στην ενότητα αυτή παρουσιάζεται μια συγκεντρωτική αποτίμηση του πρακτικού "trade-off" που προκύπτει όταν οι περιλήψεις χρησιμοποιούνται ως υποκατάστατο ή ως συμπυκνωμένη αναπαράσταση των αρχικών κειμένων στην αναζήτηση. Η ανάλυση οργανώνεται ταυτόχρονα σε τρεις άξονες: (α) το βαθμό συμπίεσης που επιτυγχάνει η σύνοψη (δείκτης και ποσοστό συμπίεσης), (β) την πιστότητα ως προς το αρχικό νόημα μέσω σημασιολογικής ομοιότητας (SBERT similarity), και (γ) την επίδοση ανάκτησης με βάση τις μετρικές MAP και Recall@100. Παράλληλα ενσωματώνεται και ο χρονικός παράγοντας, ώστε να αποτυπωθεί όχι μόνο "τι κερδίζουμε" σε ποιότητα και μέγεθος, αλλά και "τι κοστίζει" σε χρόνο παραγωγής σύνοψης και εκτέλεσης ανάκτησης. Οι τιμές παρουσιάζονται ομαδοποιημένες ανά μηχανισμό ανάκτησης (BM25 ή ColBERT) και ανά τύπο σύνοψης (extractive, hybrid, abstractive, instructive), ενώ χρησιμοποιείται η διάμεσος (median) για κάθε μέτρηση, ώστε η σύγκριση να είναι ανθεκτική σε ακραίες τιμές και να εκφράζει μια αντιπροσωπευτική "κεντρική" συμπεριφορά των runs μέσα σε κάθε ομάδα.

Πίνακας 7.9: Συγκεντρωτικός πίνακας trade-off συμπίεσης–πιστότητας–ανάκτησης (median) ανά μηχανισμό ανάκτησης και τύπο σύνοψης

Μηχ/σμός ανάκτησης	Τύπος σύνοψης	Δείκτης συμπίεσης (median)	Ποσοστό συμπίεσης (median)	SBERT similarity (median)	MAP (median)	Recall@100 (median)	Χρόνος σύνοψης (median)	Χρόνος ανάκτ/ης (median)	(N) runs	(N) Index	(N) μοντέλων σύνοψης
BM25	instructive	12,7472	92,155	0,7493	0,4802	0,79265	15:45:48	00:03:12	48	1	4
BM25	hybrid	5,4184	81,54	0,8483	0,4781	0,7819	00:54:55	00:02:52	13	1	2
BM25	extractive	11,60435	91,38	0,90705	0,4719	0,77275	00:45:29	00:02:39	72	1	6
BM25	abstractive	14,64985	93,17	0,7619	0,4573	0,76305	06:40:43	00:02:08	72	1	6
ColBERT	hybrid	1,6748	36,44	0,9167	0,381	0,6846	00:54:55	02:20:02	4	2	2

ColBERT	extractive	10,8725	90,8	0,89625	0,3776	0,6771	00:49:08	02:16:55	38	4	6
ColBERT	abstractive	19,3864	94,84	0,7078	0,3804	0,6682	04:41:44	02:15:11	33	2	6
ColBERT	instructive	20,1821	95,05	0,6467	0,3291	0,6394	14:52:53	02:18:33	21	1	4

Headers Πίνακα 7.9 (trade-off συμπίεσης, σημασιολογικής πιστότητας και απόδοσης ανάκτησης)

- Μηχανισμός ανάκτησης: Ο retrieval engine που χρησιμοποιήθηκε στα runs της ομάδας (BM25/Pyserini ή ColBERT). Επιτρέπει σύγκριση της συμπεριφοράς της σύνοψης ανά μηχανισμό.
- Τύπος σύνοψης: Η κατηγορία μεθοδολογίας σύνοψης (extractive, abstractive, hybrid, instructive). Κάθε γραμμή συγκεντρώνει runs του ίδιου τύπου μέσα στον ίδιο μηχανισμό ανάκτησης.
- Δείκτης συμπίεσης (median): Η διάμεσος του compression ratio, δηλαδή "πόσες φορές μικραίνει" το κείμενο μετά τη σύνοψη. Τιμή 11,60 σημαίνει ότι η σύνοψη είναι περίπου 11,6 φορές μικρότερη από το αρχικό κείμενο (ως τυπική/κεντρική τιμή της ομάδας).
- Ποσοστό συμπίεσης (median): Η διάμεσος του ποσοστού μείωσης μεγέθους σε σχέση με το αρχικό. Τιμή 91,38 σημαίνει περίπου 91,38% λιγότερο κείμενο (άρα απομένει περίπου 8,62% του αρχικού).
- SBERT similarity (median): Η διάμεσος σημασιολογικής ομοιότητας (0–1) μεταξύ σύνοψης και αρχικού κειμένου βάσει embeddings (Sentence-BERT). Όσο πιο κοντά στο 1, τόσο πιο πιστά διατηρείται το νόημα.
- MAP (median): Η διάμεσος τιμή του MAP από τα retrieval runs της ομάδας. Αποτυπώνει την "τυπική" επίδοση κατάταξης (όχι το καλύτερο run).
- Recall@100 (median): Η διάμεσος τιμή του Recall@100 από τα runs της ομάδας. Δείχνει την "τυπική" κάλυψη σχετικών εγγράφων μέσα στα πρώτα 100 αποτελέσματα.
- Χρόνος σύνοψης (median): Η διάμεσος διάρκεια παραγωγής/εξαγωγής σύνοψης για τα runs της ομάδας. Εκφράζει το "τυπικό" υπολογιστικό κόστος σύνοψης ανά τύπο.
- Χρόνος ανάκτησης (median): Η διάμεσος διάρκεια εκτέλεσης retrieval για τα runs της ομάδας. Αναδεικνύει το "τυπικό" κόστος ανάκτησης και τις διαφορές ταχύτητας μεταξύ BM25 και ColBERT.
- (N) runs: Πόσες πειραματικές εκτελέσεις (runs) περιλαμβάνονται σε αυτή την ομάδα, πάνω στις οποίες υπολογίστηκαν οι διάμεσοι.
- (N) Index: Πόσες διαφορετικές διαμορφώσεις index (π.χ. διαφορετικά index folders/ρυθμίσεις) συμμετέχουν στα runs της ομάδας.
- (N) μοντέλων σύνοψης: Πόσα διαφορετικά summary models συνεισφέρουν runs μέσα στην ομάδα (π.χ. πόσα διαφορετικά μοντέλα υπάρχουν μέσα στο "abstractive" για τον ίδιο μηχανισμό).

Ο πίνακας δεν αποτυπώνει μεμονωμένες "καλές" περιπτώσεις, αλλά τη συνολική, τυπική συμπεριφορά ανά κατηγορία πειραμάτων. Η χρήση median διασφαλίζει ότι τα συμπεράσματα δεν επηρεάζονται δυσανάλογα από outliers, ειδικά σε χρόνους και σε ακραία runs.

Η συμπίεση, η πιστότητα και η απόδοση ανάκτησης δεν μεταβάλλονται ανεξάρτητα. Κάθε τύπος σύνοψης εμφανίζει διαφορετικό προφίλ ισορροπίας μεταξύ "πόσο μικραίνει" το κείμενο, "πόσο κοντά" παραμένει στο νόημα, και "τι επίδοση" δίνει στην αναζήτηση. Συνεπώς, η επιλογή μεθοδολογίας σύνοψης δεν είναι καθολική αλλά εξαρτάται από τον στόχο χρήσης.

Ο χρονικός παράγοντας διαχωρίζει καθαρά το κόστος παραγωγής σύνοψης από το κόστος retrieval. Έτσι, η αξιολόγηση δεν περιορίζεται σε ποιοτικά αποτελέσματα, αλλά μετατρέπεται σε πιο ρεαλιστική σύγκριση για σενάρια όπου ο χρόνος εκτέλεσης αποτελεί κρίσιμο περιορισμό.

Η ομαδοποίηση ανά μηχανισμό ανάκτησης δείχνει ότι το trade-off πρέπει να ερμηνεύεται σε συνάρτηση με το retrieval engine. Με άλλα λόγια, η ίδια κατηγορία σύνοψης μπορεί να έχει διαφορετική πρακτική "αξία" όταν αλλάζει ο μηχανισμός ανάκτησης, άρα τα συμπεράσματα πρέπει να διαβάζονται πάντα εντός του αντίστοιχου πλαισίου.

7.8 Εξάρτηση της ωφέλειας των περιλήψεων από τα query fields

Η ενότητα αυτή απομονώνει έναν κρίσιμο παράγοντα που συχνά χάνεται όταν κοιτάμε μόνο "συνολικά" averages: το αν η σύνοψη βοηθά την ανάκτηση εξαρτάται έντονα από το τι ακριβώς δίνεται ως ερώτημα. Αντί να αντιμετωπίζεται η σύνοψη ως καθολικά ωφέλιμη ή καθολικά ουδέτερη, ο πίνακας οργανώνει τα αποτελέσματα ανά πακέτο πεδίων query (QUERY TAGS USED) και ποσοτικοποιεί τη συμπεριφορά της σύνοψης σε συνθήκες δίκαιης σύγκρισης. Η σύγκριση γίνεται αποκλειστικά σε ζευγάρια baseline–summarized που αντιστοιχούν στο ίδιο retrieval πλαίσιο (ίδιος μηχανισμός και ίδια ρύθμιση ανάκτησης), ώστε οι διαφορές σε MAP και Recall@100 να αποδίδονται στη σύνοψη και όχι σε αλλαγές παραμέτρων ή index. Με τον τρόπο αυτό, η ενότητα τεκμηριώνει πότε η σύνοψη "ταιριάζει" λειτουργικά με τη μορφή του query και πότε η επίδρασή της είναι περιορισμένη ή και αρνητική, αναδεικνύοντας ότι η χρησιμότητα της σύνοψης είναι συνάρτηση της πληροφορίας που εμπεριέχει το αρχικό query.

Πίνακας 7.10: Συχνότητα και μέγεθος βελτίωσης (ΔMAP, ΔRecall@100) ανά QUERY TAGS USED σε ζευγαρωμένες συγκρίσεις baseline–summarized

QUERY TAGS USED	N ζευγών	N (ΔMAP>0)	Ποσοστό βελτίωσης MAP (%)	Διάμεση ποσοστιαία μεταβολή	Διάμεση μεταβολή MAP (ΔMAP)	N (ΔRecall@100>0)	Ποσοστό βελτίωσης Recall@100 (%)	Διάμεση ποσοστιαία μεταβολή Recall@100 (%)
FIRST-CLAIM(all)	17	16	94,12	11,26	0,0451	16	94,12	10,33
TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)	17	14	82,35	4,35	0,0196	14	82,35	2,83
TITLE(all), ABSTRACT(all), CLAIMS(all)	17	10	58,82	0,1	0,0005	13	76,47	0,54

TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)	18	15	83,33	1,64	0,0078	12	66,67	0,65
TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)	35	17	48,57	-0,06	-0,0003	21	60	1,24
TITLE(all), ABSTRACT(all)	17	2	11,76	-0,17	-0,0008	9	52,94	0,01
TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)	36	15	41,67	-0,28	-0,0011	16	44,44	-0,22
TITLE(all), ABSTRACT(all), DESCRIPTION(all)	17	10	58,82	0,35	0,0017	7	41,18	-0,2
ALL TAGS & WORDS	17	1	5,88	-3,97	-0,0197	4	23,53	-2,18
DESCRIPTION(all)	24	3	12,5	-9,26	-0,0374	5	20,83	-3,65
DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)	17	1	5,88	-8,08	-0,0369	3	17,65	-4,1
CLAIMS(all)	19	2	10,53	-7,83	-0,0352	3	15,79	-5,37
CLAIMS(220), DESCRIPTION(220)	10	3	30	-5,37	-0,0181	1	10	-1,08

Headers Πίνακα 7.10 (ομαδοποίηση ανά QUERY TAGS USED)

- QUERY TAGS USED: Η διαμόρφωση των πεδίων που χρησιμοποιήθηκαν ως query (δηλαδή ποια tags «μπήκαν» στο ερώτημα). Τα (all), (100) κ.λπ. δηλώνουν το όριο περιεχομένου που αξιοποιήθηκε από κάθε tag (π.χ. all ή συγκεκριμένο μήκος), αλλά στον τίτλο/κείμενο μπορείς να αναφέρεις μόνο τα ονόματα των tags.
- N ζευγών: Πόσες έγκυρες συγκρίσεις baseline–summarized έγιναν για αυτή τη διαμόρφωση query tags. Κάθε «ζευγάρι» υπάρχει μόνο όταν έχει βρεθεί baseline run με τα ίδια κοινά χαρακτηριστικά (ώστε η σύγκριση να είναι δίκαιη).
- N (ΔMAP>0): Σε πόσα από τα ζευγάρια η σύνοψη οδήγησε σε αύξηση του MAP σε σχέση με το baseline (δηλαδή MAP_summarized > MAP_baseline).
- Ποσοστό βελτίωσης MAP (%): Το ποσοστό των ζευγών για τα οποία ισχύει ΔMAP>0. Δείχνει «πόσο συχνά» η σύνοψη βοηθά στο MAP για το συγκεκριμένο QUERY TAGS USED.

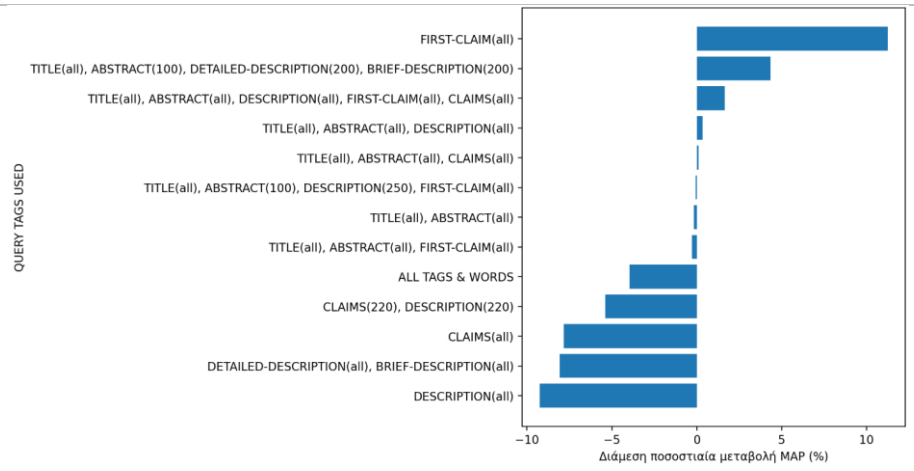
- Διάμεση ποσοστιαία μεταβολή MAP (%): Η διάμεσος της ποσοστιαίας μεταβολής του MAP μέσα στα ζευγάρια αυτής της ομάδας. Η διάμεσος είναι η «κεντρική» τιμή (50% των περιπτώσεων πάνω, 50% κάτω) και προτιμάται γιατί δεν επηρεάζεται εύκολα από ακραίες τιμές.
- Διάμεση μεταβολή MAP (ΔMAP): Η διάμεσος της απόλυτης διαφοράς MAP (σε μονάδες MAP, όχι σε ποσοστό) μεταξύ summarized και baseline.
- N (ΔRecall@100>0): Σε πόσα ζευγάρια η σύνοψη οδήγησε σε αύξηση του Recall@100 (Recall@100_summarized > Recall@100_baseline).
- Ποσοστό βελτίωσης Recall@100 (%): Το ποσοστό των ζευγών για τα οποία ισχύει ΔRecall@100>0. Δείχνει «πόσο συχνά» η σύνοψη βοηθά στην κάλυψη σχετικών εγγράφων μέσα στα πρώτα 100 αποτελέσματα.
- Διάμεση ποσοστιαία μεταβολή Recall@100 (%): Η διάμεσος της ποσοστιαίας μεταβολής του Recall@100 μέσα στην ομάδα, ως «τυπική» (robust) εικόνα της επίδρασης.

Η ωφέλεια των περιλήψεων δεν είναι ομοιόμορφη ως προς τα query fields: για ορισμένα QUERY TAGS USED η βελτίωση εμφανίζεται σε πολύ μεγάλο ποσοστό των ζευγών, ενώ σε άλλα η βελτίωση είναι σπάνια και η διάμεση μεταβολή είναι αρνητική.

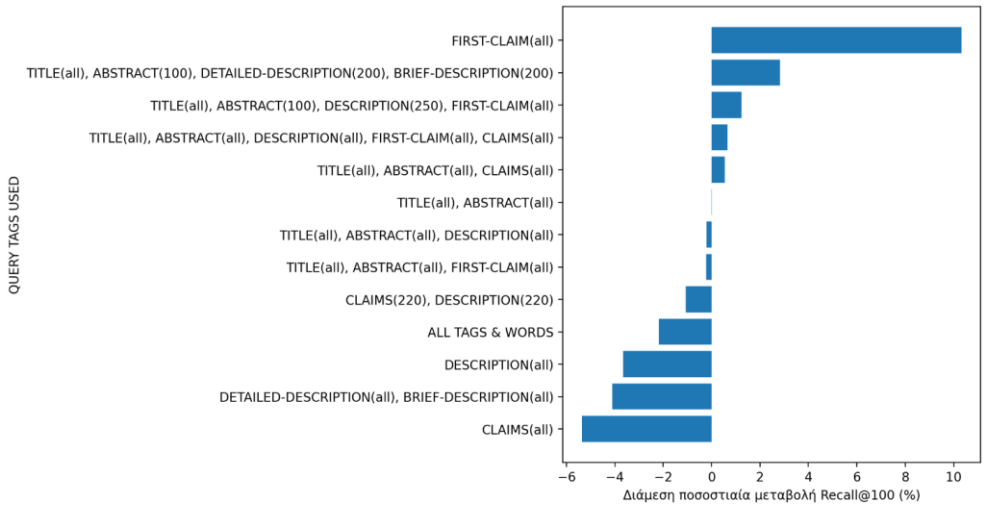
Το FIRST-CLAIM(all) ξεχωρίζει ως το πιο “συστηματικά” ευνοημένο σενάριο: 16/17 ζευγάρια βελτιώνουν MAP και 16/17 ζευγάρια βελτιώνουν Recall@100, με διάμεση ποσοστιαία μεταβολή 11,26% στο MAP και 10,33% στο Recall@100.

Τα σύνθετα queries που συνδυάζουν περισσότερα πεδία (π.χ. TITLE, ABSTRACT, DETAILED-DESCRIPTION, BRIEF-DESCRIPTION) παρουσιάζουν επίσης υψηλή συχνότητα βελτίωσης, αλλά με σαφώς μικρότερη διάμεση ποσοστιαία μεταβολή σε σχέση με το FIRST-CLAIM.

Αντίθετα, όταν το query βασίζεται κυρίως σε DESCRIPTION(all) ή σε CLAIMS(all), η σύνοψη τείνει να μην βοηθά: τα ποσοστά βελτίωσης είναι χαμηλά και η διάμεση ποσοστιαία μεταβολή είναι αρνητική τόσο στο MAP όσο και στο Recall@100 (π.χ. DESCRIPTION(all): 3/24 βελτιώνουν MAP και 5/24 βελτιώνουν Recall@100, με διάμεσες ποσοστιαίες μεταβολές -9,26% και -3,65% αντίστοιχα).



Γράφημα 2: : Επίδραση σύνθεσης query tags στη μεταβολή MAP (median %Δ)



Γράφημα 3: : Επίδραση σύνθεσης query tags στη μεταβολή Recall@100 (median %Δ)

Συνολικά, ο πίνακας τεκμηριώνει ότι το "αν αξίζει" η σύνοψη ως ενίσχυση της ανάκτησης εξαρτάται από το πληροφοριακό περιεχόμενο και τη δομή του query: όπου το query έχει πιο συμπαγές και σημασιολογικά πυκνό πεδίο (όπως FIRST-CLAIM), η σύνοψη λειτουργεί συχνότερα ενισχυτικά, ενώ σε ήδη εκτενή πεδία (όπως DESCRIPTION ή CLAIMS) η συμπύκνωση μπορεί να αφαιρεί χρήσιμα σήματα για την ανάκτηση.

7.9 Σύγκριση ORIGINAL έναντι σύνοψης με σταθερά query fields

Σε πολλά σενάρια χρήσης, το κρίσιμο ερώτημα δεν είναι αν μια σύνοψη είναι "καλή" ως κείμενο, αλλά αν μπορεί να αντικαταστήσει το ORIGINAL περιεχόμενο χωρίς να υποβαθμίσει την ανάκτηση, όταν τα query fields παραμένουν ίδια. Η παρούσα ενότητα συνοψίζει τι συμβαίνει όταν αντικαθίσταται το ORIGINAL κείμενο με σύνοψη, διατηρώντας ακριβώς την ίδια διαμόρφωση QUERY TAGS USED, και συγκρίνει την επίδραση σε MAP και Recall@100. Για να διασφαλιστεί δίκαιη σύγκριση, χρησιμοποιούνται μόνο matched baseline–summarized pairs με κοινά στοιχεία (ίδιος μηχανισμός, ίδια διαμόρφωση index, ίδιες ρυθμίσεις ανάκτησης, ίδια query tags και ίδιο aggregation mode όπου εφαρμόζεται). Αντί να παρουσιαστούν όλες οι μεμονωμένες συγκρίσεις, ο πίνακας αποτυπώνει τόσο τη συχνότητα βελτίωσης (ποσοστό ζευγών με θετικό Δ) όσο και το "τυπικό" μέγεθος επίδρασης μέσω διαμέσων (median) ποσοστιαίων και απόλυτων μεταβολών.

Πίνακας 7.11: Συνοπτική σύγκριση ORIGINAL έναντι περιλήψεων με αντιστοίχιση ίδιων query tags (MAP και Recall@100) ανά μηχανισμό ανάκτησης και τύπο σύνοψης

Μηχανισμός ανάκτησης	Τύπος σύνοψης	N pairs	N Index	N query tags	N models	pct pairs MAP improve	median pct ΔMAP	median ΔMAP	pct pairs Recall improve	median pct ΔRecall	median ΔRecall
BM25 (Pyserini)	instructive	48	1	12	4	64,58	1,51	0,01	89,58	2,92	0,02

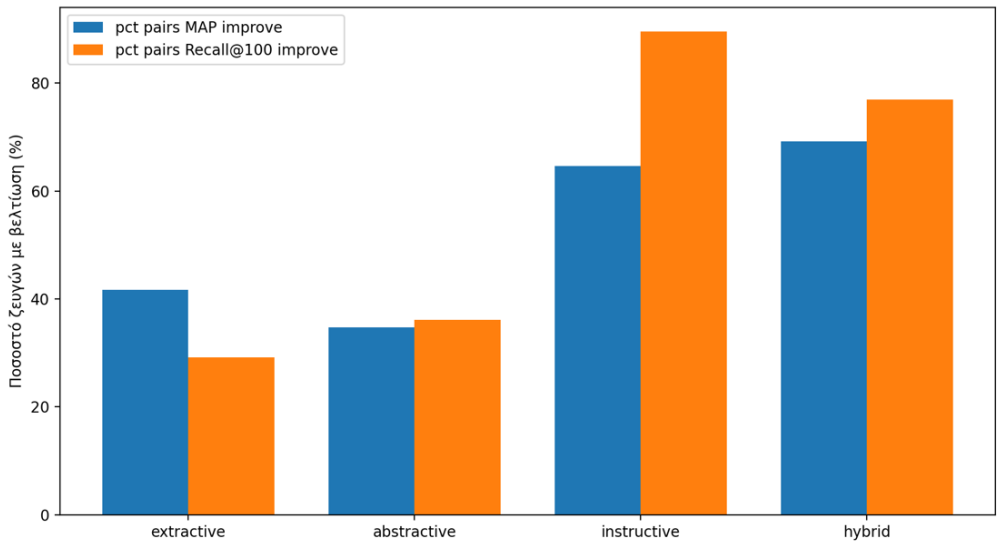
BM25 (Pyserini)	hybrid	13	1	12	2	69,23	1,46	0,01	76,92	0,7	0,01
ColBERT	hybrid	3	1	2	2	33,33	-1,69	-0,01	66,67	3,27	0,02
ColBERT	instructive	12	1	3	4	25	-6,75	-0,02	50	0,77	0,01
ColBERT	abstractive	15	1	4	6	40	-1,09	0	46,67	0	0
BM25 (Pyserini)	abstractive	72	1	12	6	34,72	-2,72	-0,01	36,11	-1,76	-0,01
ColBERT	extractive	21	1	5	6	23,81	-1,13	0	33,33	-0,53	0
BM25 (Pyserini)	extractive	72	1	12	6	41,67	-0,48	0	29,17	-0,99	-0,01

Headers Πίνακα 7.11 (ομαδοποίηση ανά Μηχανισμό ανάκτησης και Τύπο σύνοψης)

- Μηχανισμός ανάκτησης: Ο retrieval engine στον οποίο εκτελέστηκαν τα runs (BM25/Pyserini ή ColBERT). Χρησιμοποιείται για να φανεί αν η επίδραση της σύνοψης αλλάζει ανά μηχανισμό.
- Τύπος σύνοψης: Η κατηγορία προσέγγισης σύνοψης (extractive, abstractive, hybrid, instructive). Ομαδοποιεί διαφορετικά μοντέλα κάτω από κοινή μεθοδολογία.
- N pairs: Πλήθος matched ζευγών ORIGINAL–summarized που βρέθηκαν για τη συγκεκριμένη ομάδα (ίδιος μηχανισμός και ίδιος τύπος σύνοψης), πάνω στα οποία υπολογίστηκαν όλες οι μεταβολές.
- N Index: Πόσες διαφορετικές διαμορφώσεις index συμμετέχουν στα ζευγάρια της ομάδας. Αν είναι 1, όλα τα ζευγάρια προέρχονται από το ίδιο index.
- N query tags: Πόσες διαφορετικές διαμορφώσεις QUERY TAGS USED καλύπτονται μέσα στην ομάδα (δηλαδή πόσα διαφορετικά "πακέτα" πεδίων query εμφανίζονται στα matched ζευγάρια).
- N models: Πόσα διαφορετικά summary models συνεισφέρουν runs μέσα στην ομάδα (π.χ. πόσα διαφορετικά μοντέλα υπάρχουν μέσα στο "abstractive").
- pct pairs MAP improve: Ποσοστό των ζευγών όπου το MAP με σύνοψη είναι μεγαλύτερο από το MAP του ORIGINAL (MAP_summarized > MAP_original). Δείχνει «πόσο συχνά» η σύνοψη βοηθά στο MAP.
- median pct ΔMAP: Διάμεσος της ποσοστιαίας μεταβολής του MAP μέσα στην ομάδα. Δείχνει την "τυπική" ποσοστιαία επίδραση, με τρόπο ανθεκτικό σε ακραίες τιμές.
- median ΔMAP: Διάμεσος της απόλυτης μεταβολής του MAP (ΔMAP = MAP_summarized − MAP_original). Δείχνει την "τυπική" μετατόπιση σε μονάδες MAP.
- pct pairs Recall improve: Ποσοστό των ζευγών όπου το Recall@100 με σύνοψη είναι μεγαλύτερο από το Recall@100 του ORIGINAL (Recall@100_summarized > Recall@100_original).
- median pct ΔRecall: Διάμεσος της ποσοστιαίας μεταβολής του Recall@100 μέσα στην ομάδα. Αποτυπώνει την "τυπική" ποσοστιαία επίδραση στη κάλυψη σχετικών εγγράφων στα πρώτα 100 αποτελέσματα.

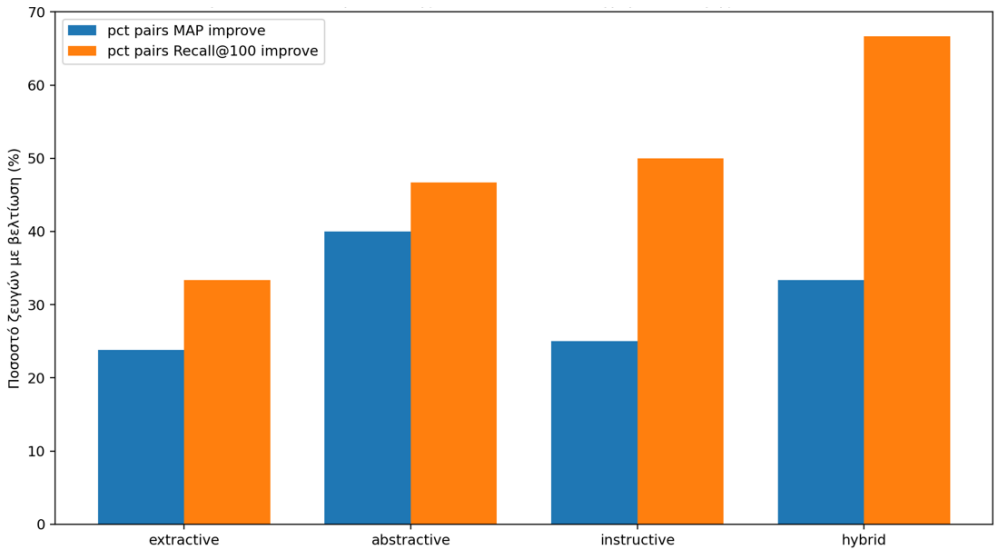
- median ΔRecall: Διάμεσος της απόλυτης μεταβολής του Recall@100 (ΔRecall = Recall@100_summarized − Recall@100_original). Δείχνει τη "τυπική" αύξηση ή μείωση σε απόλυτες μονάδες Recall@100.

Στον BM25 (Pyserini) παρατηρείται ότι οι hybrid και instructive προσεγγίσεις εμφανίζουν πιο συστηματικά θετική επίδραση στα matched-tag σενάρια. Ειδικότερα, στο hybrid το ποσοστό βελτίωσης στο MAP είναι 69,23% με διάμεσο ποσοστιαίο κέρδος 1,46% και διάμεσο ΔMAP 0,01, ενώ στο Recall@100 η βελτίωση εμφανίζεται στο 76,92% των ζευγών (διάμεσο ποσοστιαίο κέρδος 0,7% και διάμεσο ΔRecall 0,01). Στο instructive, το MAP βελτιώνεται στο 64,58% των ζευγών (median_pct_ΔMAP 1,51%, median_ΔMAP 0,01) και το Recall@100 στο 89,58% (median_pct_ΔRecall 2,92%, median_ΔRecall 0,02). Αντίθετα, στο BM25 οι abstractive και extractive κατηγορίες παρουσιάζουν αρνητική ή οριακή διάμεση επίδραση: για abstractive, median_pct_ΔMAP −2,72 και median_ΔMAP −0,01, ενώ για extractive median_pct_ΔMAP −0,48 και median_ΔMAP 0.



Γράφημα 4: Matched-pairs, ποσοστό ζευγών με βελτίωση (BM25) ανά κατηγορία σύνοψης (MAP vs Recall@100)

Στον ColBERT, η συνολική εικόνα είναι πιο συγκρατημένη και ως προς το MAP οι διάμεσοι δείκτες είναι αρνητικοί ή μηδενικοί στις περισσότερες κατηγορίες. Ενδεικτικά, στο instructive το pct_pairs_MAP_improve είναι 25 με median_pct_ΔMAP −6,75 και median_ΔMAP −0,02, ενώ στο Recall@100 η βελτίωση εμφανίζεται στο 50 με median_pct_ΔRecall 0,77 και median_ΔRecall 0,01. Το hybrid στον ColBERT εμφανίζει μικτή συμπεριφορά: στο MAP το median_pct_ΔMAP είναι −1,69 (median_ΔMAP −0,01), αλλά στο Recall@100 το pct_pairs_Recall_improve είναι 66,67 με median_pct_ΔRecall 3,27 και median_ΔRecall 0,02.



Γράφημα 5: Matched-pairs, ποσοστό ζευγών με βελτίωση (Colbert) ανά κατηγορία σύνοψης (MAP vs Recall@100)

Συνολικά, τα αποτελέσματα τεκμηριώνουν ότι, όταν κρατάμε σταθερά τα query fields, η επίδραση της σύνοψης εξαρτάται ισχυρά από τον μηχανισμό ανάκτησης και από τον τύπο σύνοψης, με τον BM25 να παρουσιάζει πιο συστηματικά θετικές ενδείξεις στις hybrid/instructive κατηγορίες σε σχέση με τον ColBERT.

Τα αποτελέσματα του Κεφαλαίου 7 συνοψίζουν τις κύριες τάσεις και τις συγκρίσεις μεταξύ μεθόδων. Οι αναλυτικές μετρήσεις σε επίπεδο εκτέλεσης, συμπεριλαμβανομένων όλων των επιμέρους δεικτών και των αντίστοιχων παραμέτρων, παρέχονται στο **ΠΑΡΑΡΤΗΜΑ Β**, ώστε να είναι διαθέσιμες για έλεγχο, διασταύρωση και επαναχρησιμοποίηση.

Κεφάλαιο 8ο: Συμπεράσματα και μελλοντική έρευνα.

8.1 Σύνοψη στόχου και λογικής αξιολόγησης

Η εργασία εξετάζει τη σύνοψη πατεντών ως διαδικασία παραγωγής δομημένων κειμενικών αναπαραστάσεων (ανά tag/ενότητα) που μπορούν να χρησιμοποιηθούν συστηματικά σε μεγάλης κλίμακας πειραματικές ροές. Το βασικό ερώτημα δεν είναι αν ένα συγκεκριμένο retrieval engine είναι «καλύτερο», αλλά υπό ποιες συνθήκες οι διαφορετικές κατηγορίες σύνοψης (extractive, abstractive, instructive, hybrid) παράγουν περιεχόμενο που διατηρεί λειτουργικά σήματα συνάφειας, ώστε να υποστηρίζει downstream διεργασίες, όπως η ανάκτηση πληροφορίας.

Για να αποτιμηθεί με αντικειμενικό τρόπο η λειτουργική αξία των summaries, χρησιμοποιήθηκε ως μέτρο η επίδρασή τους σε δείκτες ανάκτησης (MAP και Recall@100) και μάλιστα σε «δίκαιες» ζευγαρωμένες συγκρίσεις baseline–summarized με ίδια πλαίσια αναζήτησης (matched pairs). Με αυτόν τον σχεδιασμό, οι διαφορές στις μετρικές αποδίδονται στη σύνοψη και όχι σε αλλαγές index/παραμέτρων.

8.2 Κύρια συμπεράσματα ανά κατηγορία σύνοψης

8.2.1 Extractive: υψηλή σταθερότητα, αλλά περιορισμένη «τυπική» ωφέλεια σε matched retrieval

Οι extractive μέθοδοι χαρακτηρίζονται από προβλέψιμη συμπεριφορά και χαμηλό ρίσκο αλλοίωσης, καθώς διατηρούν αυτούσιες προτάσεις/όρους από το αρχικό κείμενο. Σε επίπεδο matched συγκρίσεων, η επίδραση των extractive δεν εμφανίζεται ως συστηματικά θετική: για BM25 το ποσοστό ζευγών με βελτίωση MAP είναι 41,67%, ενώ η διάμεση μεταβολή MAP είναι 0 και η διάμεση ποσοστιαία μεταβολή MAP είναι ελαφρά αρνητική (-0,48%). Αντίστοιχα, στο Recall@100 το ποσοστό βελτίωσης είναι 29,17% με διάμεση μεταβολή -0,01.

Η ερμηνεία που προκύπτει από αυτό το μοτίβο είναι ότι η extractive σύνοψη λειτουργεί πιο πολύ ως «εξομάλυνση/συμπύκνωση» χωρίς να αλλάζει αποφασιστικά το retrieval σήμα σε δίκαιες συγκρίσεις. Άρα, αποτελεί ισχυρή επιλογή όταν προτεραιότητα είναι η ασφάλεια και η επαναληψιμότητα, όχι η μεγιστοποίηση κέρδους στις μετρικές ανάκτησης.

Σε όρους υπολογιστικού κόστους, οι extractive ροές παραμένουν από τις πιο αποδοτικές (ενδεικτικά χρόνοι κάτω της τάξης της 1–1,5 ώρας για 2.592 queries στα μοντέλα της κατηγορίας).

8.2.2 Abstractive: ισχυρή συμπίεση, αλλά σε matched σύγκριση τείνει να μη βελτιώνει τυπικά το retrieval

Οι abstractive μέθοδοι προσφέρουν μεγάλη συμπίεση και παραγωγή πιο συνεκτικού κειμένου, όμως στα matched αποτελέσματα η «τυπική» (median) μεταβολή στις μετρικές δεν εμφανίζεται θετική. Για BM25, το ποσοστό ζευγών με βελτίωση MAP είναι 34,72% και η διάμεση μεταβολή MAP είναι -0,01 (με διάμεση ποσοστιαία μεταβολή -2,72%). Αντίστοιχα, στο Recall@100 το ποσοστό βελτίωσης είναι 36,11% με διάμεση μεταβολή -0,01.

Αυτό το αποτέλεσμα είναι συνεπές με την ιδιαιτερότητα των πατεντών: η abstractive αναδιατύπωση μπορεί να μειώσει ή να μετασχηματίσει λεπτομέρειες και ειδικούς όρους, οι οποίοι σε αρκετά σενάρια λειτουργούν ως κρίσιμα «σήματα» συνάφειας. Άρα, στην παρούσα εργασία οι abstractive περιλήψεις αποδείχθηκαν πιο κατάλληλες ως «αναγνωστικές»/περιγραφικές αναπαραστάσεις παρά ως σταθερά

ωφέλιμο υποκατάστατο του πρωτογενούς κειμένου για retrieval, όταν η σύγκριση γίνεται αυστηρά με αντιστοίχιση ίδιων query tags.

Σε επίπεδο χρόνων, η abstractive κατηγορία εμφανίζει μεγάλη διακύμανση ανά μοντέλο, από περίπου 1,5 ώρα έως περίπου 10 ώρες για 2.592 queries, γεγονός που αναδεικνύει ότι η επιλογή abstractive μοντέλου επηρεάζει έντονα την πρακτική δυνατότητα μεγάλης κλίμακας πειραμάτων.

8.2.3 Instructive: καλύτερη εικόνα σε matched BM25, αλλά με υψηλό κόστος χρόνου και εξάρτηση από τη μορφή query

Η instructive κατηγορία αναδεικνύεται ως η πιο «ενεργή» στη διαμόρφωση περιεχομένου, καθώς καθοδηγείται από οδηγίες (prompts) και παράγει στοχευμένο κείμενο. Στις matched συγκρίσεις για BM25, τα instructive παρουσιάζουν σαφώς θετικότερη εικόνα από extractive/abstractive: ποσοστό βελτίωσης MAP 64,58% με διάμεση μεταβολή MAP +0,01 και διάμεση ποσοστιαία μεταβολή +1,51%. Στο Recall@100, το ποσοστό βελτίωσης είναι 89,58% με διάμεση μεταβολή +0,02 και διάμεση ποσοστιαία μεταβολή +2,92%.

Το συμπέρασμα εδώ δεν είναι ότι η instructive σύνοψη «βελτιώνει πάντα» την ανάκτηση, αλλά ότι, όταν αξιολογείται σε matched συνθήκες, εμφανίζει πιο συχνά θετικές μεταβολές και μάλιστα με πιο καθαρή επίδραση στο Recall@100. Αυτό συμβαδίζει με την ιδέα ότι η instructive παραγωγή μπορεί να «συμπυκνώνει» τα χρήσιμα σήματα σε πιο προσβάσιμη μορφή, αρκεί το query να βασίζεται σε πεδία που επιδέχονται τέτοια συμπύκνωση.

Το κόστος, όμως, είναι ιδιαίτερα υψηλό: οι χρόνοι σύνοψης για instructive μοντέλα κυμαίνονται περίπου από 13 έως 22 ώρες (σε πλήρη επεξεργασία 2.592 queries), καθιστώντας την κατηγορία την πιο απαιτητική υπολογιστικά.

8.2.4 Hybrid: η πιο σταθερά θετική κατηγορία σε matched BM25, με πολύ χαμηλό χρόνο παραγωγής

Η hybrid κατηγορία, όπως ορίστηκε στην εργασία, καλύπτει συνδυαστικές προσεγγίσεις που εισάγουν επιπλέον μηχανισμό επιλογής/ιεράρχησης (π.χ. graph-based κεντρικότητα με TextRank πάνω σε semantic similarities). Στις matched συγκρίσεις για BM25, οι hybrid εμφανίζουν το υψηλότερο ποσοστό βελτίωσης MAP (69,23%) και θετική διάμεση μεταβολή MAP (+0,01). Παράλληλα στο Recall@100 παρουσιάζουν ποσοστό βελτίωσης 76,92% με διάμεση μεταβολή +0,01.

Ένα πρόσθετο πρακτικό πλεονέκτημα της hybrid κατηγορίας είναι ότι επιτυγχάνει αυτά τα θετικά σήματα με πολύ χαμηλό χρόνο παραγωγής. Ο συνολικός χρόνος σύνοψης που καταγράφεται για το hybrid μοντέλο της εργασίας είναι κάτω από 1 ώρα για 2.592 queries, το οποίο είναι συγκρίσιμο ή καλύτερο από αρκετές extractive υλοποιήσεις και κατά τάξεις μεγέθους ταχύτερο από instructive/βαριά abstractive.

8.3 Το πιο ισχυρό εύρημα: η ωφέλεια της σύνοψης εξαρτάται από το «πακέτο» query tags

Ανεξάρτητα από κατηγορία σύνοψης, ο βασικός παράγοντας που «οδηγεί» την επιτυχία είναι ο τρόπος σχηματισμού του query (δηλαδή ποια tags συμμετέχουν). Σε ζευγαρωμένες συγκρίσεις baseline–summarized ανά QUERY TAGS USED, προκύπτουν πολύ καθαρά μοτίβα:

Όταν το query βασίζεται σε FIRST-CLAIM(all), παρατηρείται σχεδόν καθολική βελτίωση: 16/17 ζεύγη βελτιώνουν MAP (94,12%) με διάμεση μεταβολή ΔMAP = 0,0451 και διάμεση ποσοστιαία μεταβολή

MAP 11,26%. Αντίστοιχα, 16/17 ζεύγη βελτιώνουν Recall@100 (94,12%) με διάμεση ποσοστιαία μεταβολή 10,33%.

Αντίθετα, όταν το query βασίζεται σε DESCRIPTION(all) ή σε μεγάλα/θορυβώδη πεδία χωρίς στοχευμένη συμπύκνωση, η επίδραση τείνει να γίνεται αρνητική στη διάμεσο: για DESCRIPTION(all) το ποσοστό βελτίωσης MAP είναι 12,5% με διάμεση ΔMAP = -0,0374 και διάμεση ποσοστιαία μεταβολή -9,26%.

Το συμπέρασμα, με όρους σχεδίασης πειραμάτων σύνοψης, είναι ότι η αξιολόγηση «σύνοψης → retrieval» δεν μπορεί να διαβαστεί ορθά χωρίς να αναφερθεί στο ποια πεδία χρησιμοποιούνται ως query. Ο εμπλουτισμός του query (π.χ. με TITLE+ABSTRACT) δεν είναι καθολικός κανόνας, αλλά εξαρτάται από το βασικό πεδίο και από το retrieval πλαίσιο.

8.3.1 Πρόσθετο εύρημα: ο εμπλουτισμός query με TITLE+ABSTRACT βελτιώνει συστηματικά την ανάκτηση σε "θορυβώδη" tags

Ένα από τα πλέον πρακτικά ευρήματα της ανάλυσης είναι ότι η απόδοση ανάκτησης δεν εξαρτάται μόνο από το αν χρησιμοποιείται σύνοψη ή όχι, αλλά και από το πώς σχηματίζεται το ερώτημα σε επίπεδο tags. Όταν ένα μεγάλο και πληροφοριακά «διάχυτο» πεδίο όπως το DESCRIPTION χρησιμοποιείται αυτούσιο ως μοναδικό query, η ανάκτηση τείνει να υποβαθμίζεται λόγω θορύβου και απώλειας εστίασης: το query περιλαμβάνει μεγάλο πλήθος γενικών ή δευτερευόντων όρων, με αποτέλεσμα να αποδυναμώνεται το σήμα των πραγματικά διακριτικών τεχνικών φράσεων.

Αντίθετα, όταν στο ίδιο query προστίθεται ως σταθερό συμφραζόμενο ο πυρήνας TITLE+ABSTRACT, παρατηρείται σαφής βελτίωση τόσο στο MAP@k όσο και στο Recall@k. Η βελτίωση αυτή ερμηνεύεται ως "query conditioning": το TITLE και το ABSTRACT λειτουργούν ως συμπυκνωμένη δήλωση τεχνικού σκοπού/καινοτομίας, η οποία περιορίζει τη σημασιολογική ασάφεια του DESCRIPTION και αυξάνει την πιθανότητα να ανακτηθούν έγγραφα που ευθυγραμμίζονται με τον κεντρικό τεχνικό άξονα της πατέντας. Με όρους retrieval, η προσθήκη TITLE+ABSTRACT ενισχύει τους πιο διακριτικούς όρους και παρέχει υψηλού επιπέδου πλαίσιο, ενώ το DESCRIPTION συνεισφέρει τις λεπτομέρειες, οδηγώντας σε καλύτερο ranking και αυξημένη κάλυψη σχετικών εγγράφων.

Συνεπώς, πέρα από τη σύγκριση των ομάδων σύνοψης, προκύπτει ένας πρακτικός κανόνας σχεδιασμού queries για πατέντες: για πεδία μεγάλου μήκους (DESCRIPTION/CLAIMS) είναι προτιμότερο να μη χρησιμοποιούνται "μονοθεματικά" ως query, αλλά να συνοδεύονται από TITLE+ABSTRACT, είτε στην αρχική μορφή είτε στη συνοπτική τους εκδοχή, ώστε να διατηρείται σταθερό και υψηλής ποιότητας σήμα συνάφειας.

8.4 Συνολική σύνθεση συμπερασμάτων (με έμφαση στη σύνοψη)

Συνδυάζοντας τα ευρήματα του Κεφαλαίου 7, προκύπτει ένα συνεκτικό σύνολο συμπερασμάτων που αφορά πρωτίστως τη σύνοψη και τον τρόπο με τον οποίο τα παραγόμενα κείμενα αξιοποιούνται ως queries:

- Δεν υπάρχει καθολικά ωφέλιμη κατηγορία σύνοψης. Οι κατηγορίες (extractive, abstractive, instructive, hybrid) εμφανίζουν διαφορετικό προφίλ ρίσκου/κόστους/ωφέλειας, κάτι που αποτυπώνεται καθαρά σε ζευγαρωμένες (matched) συγκρίσεις και σε διάμεσους (median) δείκτες. Η αξιολόγηση δείχνει πότε η τυπική επίδραση είναι θετική, ουδέτερη ή αρνητική, άρα η επιλογή μεθόδου πρέπει να γίνεται με βάση τον στόχο και όχι ως προεπιλογή.

- Σε matched αξιολόγηση, οι πιο σταθερά θετικές ενδείξεις εμφανίζονται στις hybrid και instructive προσεγγίσεις (ιδίως ως προς Recall@100), ενώ οι extractive και abstractive παρουσιάζουν μικρότερη ή/και αρνητική διάμεση επίδραση. Το εύρημα αυτό δεν σημαίνει ότι οι extractive/abstractive είναι «ακατάλληλες», αλλά ότι λειτουργούν καλύτερα ως συντηρητικές/αναγνωστικές αναπαραστάσεις, με περιορισμένη τυπική βελτίωση στο retrieval όταν η σύγκριση γίνεται με ίδια query tags.
- Η πρακτική ωφέλεια της σύνοψης είναι έντονα tag-dependent. Η FIRST-CLAIM(all) αναδεικνύεται ως το πιο ευνοϊκό πεδίο για λειτουργική σύνοψη, με πολύ υψηλή συχνότητα βελτίωσης και ουσιαστικό μέγεθος μεταβολής σε MAP και Recall. Αντίθετα, μεγάλα και «θορυβώδη» πεδία όπως DESCRIPTION(all) τείνουν να μην ωφελούνται τυπικά όταν χρησιμοποιούνται μονοθεματικά (δηλαδή ως μοναδικό query πεδίο), επειδή η πληροφορία είναι διάχυτη και το σήμα συνάφειας αποδυναμώνεται.
- Πέρα από την επιλογή κατηγορίας σύνοψης, κρίσιμος παράγοντας είναι ο σχηματισμός του query μέσω συνδυασμού tags. Ένα σημαντικό πρακτικό εύρημα είναι ότι, όταν ένα μεγάλο tag (π.χ. DESCRIPTION) χρησιμοποιείται μαζί με ένα συμπυκνωμένο "πλαίσιο" (TITLE+ABSTRACT), παρατηρείται αισθητή βελτίωση σε MAP@k και Recall@k σε σχέση με την περίπτωση όπου το tag χρησιμοποιείται μόνο του. Ο εμπλουτισμός με TITLE+ABSTRACT λειτουργεί ως σταθερό υψηλού επιπέδου σήμα που «προσανατολίζει» το ερώτημα, ενώ το DESCRIPTION προσθέτει τεχνικές λεπτομέρειες. Έτσι, η απόδοση δεν εξαρτάται μόνο από το αν το κείμενο είναι summary ή original, αλλά και από το πώς συντίθενται τα query components.
- Το κόστος χρόνου είναι κεντρικό κριτήριο επιλογής, όχι δευτερεύον. Οι instructive προσεγγίσεις έχουν τη μεγαλύτερη υπολογιστική απαίτηση (πολλές ώρες ανά πλήρη batch), ενώ οι hybrid και extractive είναι κατά πολύ ταχύτερες. Αυτό επηρεάζει άμεσα τη δυνατότητα εκτέλεσης πολλών runs, τον αριθμό των παραμετροποιήσεων που μπορεί να δοκιμαστούν, και τελικά το εύρος πειραματικής κάλυψης της εργασίας.
- Με βάση τα παραπάνω, η σύνοψη πατεντών στην παρούσα εργασία δεν αντιμετωπίζεται ως ενιαία τεχνική, αλλά ως οικογένεια επιλογών που πρέπει να συνδυάζεται με tag-aware πολιτικές (τι συνοψίζεται, πώς, και με ποιον σχηματισμό query). Το συνολικό συμπέρασμα είναι ότι τα καλύτερα αποτελέσματα προκύπτουν όταν:
- η μέθοδος σύνοψης επιλέγεται με βάση το προφίλ πεδίου και τον στόχο (πιστότητα έναντι συμπύκνωσης), και
- τα queries δεν είναι μονοθεματικά σε μεγάλα πεδία, αλλά ενισχύονται με TITLE+ABSTRACT ώστε να διατηρείται σταθερό και διακριτικό retrieval σήμα.

8.5 Περιορισμοί

Τα συμπεράσματα είναι άμεσα συνδεδεμένα με: (α) το συγκεκριμένο dataset/ορισμό συνάφειας (qrels), (β) τη μορφή των queries (πακέτα tags και όρια λέξεων), και (γ) το γεγονός ότι η αξιολόγηση είναι λειτουργική μέσω retrieval. Για τον λόγο αυτό, μια «αμιγώς γλωσσική» αξιολόγηση (π.χ. factual consistency (αν κάθε πρόταση της σύνοψης είναι 100% σωστή ως προς το πρωτότυπο) ή ποιότητα γραφής) δεν αποτελεί τον κύριο άξονα της παρούσας ανάλυσης, παρότι οι δείκτες συμπίεσης/ομοιότητας που έχουν καταγραφεί συμπληρώνουν το προφίλ κάθε μεθόδου.

8.6 Μελλοντικές κατευθύνσεις (με επίκεντρο τη σύνοψη και τα tags)

Με βάση τα αποτελέσματα, οι πιο άμεσες επεκτάσεις που «πατούν» πάνω στα ευρήματα είναι:

- **Tag-aware σύνοψη με στόχο queries**: αφού η ωφέλεια εξαρτάται από τα query tags, μια φυσική συνέχεια είναι η στοχευμένη παραγωγή περιλήψεων ανά πεδίο, με διαφορετικούς κανόνες ανά tag (π.χ. FIRST-CLAIM διαφορετική πολιτική από DESCRIPTION).
- **Κανόνες εμπλουτισμού query (TITLE+ABSTRACT) ως μέρος της μεθοδολογίας σύνοψης**: το Κεφ. 7 ήδη μεταφράζει τα αποτελέσματα σε εφαρμόσιμη οδηγία σχεδιασμού queries, με εξάρτηση από πεδίο και retriever. Αυτό μπορεί να ενσωματωθεί ως «συνοδευτικός κανόνας» σε μελλοντικές ροές σύνοψης+ανάκτησης.
- **Ελεγχόμενες υβριδικές αλυσίδες**: τα θετικά σήματα των hybrid, μαζί με το πολύ χαμηλό κόστος χρόνου, δείχνουν ότι αξίζει συστηματική διερεύνηση περισσότερων υβριδικών συνδυασμών, με έλεγχο του τι διατηρείται ως όρος/τεχνική φράση.
- **Label-aware** αξιοποίηση ταξινομήσεων (CPC/IPCR) σε query και retrieval: Ένα επιπλέον επόμενο βήμα είναι να διερευνηθεί πώς οι ταξινομήσεις τεχνολογικού πεδίου (<CPC-CLASSIFICATIONS>, <IPCR-CLASSIFICATIONS>) μπορούν να ενσωματωθούν στη ροή σύνοψης και ανάκτησης ως μηχανισμός περιορισμού/προσανατολισμού του candidate set. Πρακτικά, αυτό μπορεί να υλοποιηθεί είτε ως (α) φίλτρο προ-ανάκτησης (retrieval μέσα σε υποσύνολο εγγράφων που μοιράζονται ίδιες ή γειτονικές κλάσεις), είτε ως (β) reranking bias (ενίσχυση σκορ για έγγραφα με κοινές ταξινομήσεις). Η κατεύθυνση αυτή είναι ιδιαίτερα σχετική για πατέντες, επειδή οι κλάσεις CPC/IPCR κωδικοποιούν τεχνολογική εγγύτητα που δεν αποτυπώνεται πάντα μόνο στο λεξιλόγιο. Ως εκ τούτου, αναμένεται να λειτουργεί συμπληρωματικά προς τη tag-based σύνοψη: όταν η σύνοψη συμπυκνώνει περιεχόμενο, η χρήση labels μπορεί να διατηρεί το "τεχνολογικό πλαίσιο" και να μειώνει την ανάκτηση άσχετων τεχνολογικών πεδίων.
- **Υβριδικό reranking με γλωσσικό μοντέλο (score-level fusion):** Προκαταρκτικά πειράματα με συνδυασμό BM25 και unigram Language Model (Dirichlet smoothing) έδειξαν βελτίωση του MAP, γεγονός που καθιστά την κατεύθυνση αυτή αξιόλογη για συστηματική διερεύνηση. Η τελική βαθμολογία προκύπτει ως σταθμισμένος μέσος των δύο μεθόδων, με παραμέτρους προς διερεύνηση το βάρος α (BM25 vs LM), την παράμετρο smoothing μ, και το βάθος reranking. Το ενδιαφέρον της προσέγγισης έγκειται στη συμπληρωματικότητα των δύο μεθόδων: το BM25 εστιάζει στην ακριβή αντιστοίχιση όρων, ενώ το γλωσσικό μοντέλο υπολογίζει πόσο πιθανό είναι το query να "παραχθεί" από το έγγραφο.

8.7 Τελικό συμπέρασμα

Η εργασία καταλήγει ότι η σύνοψη πατεντών δεν πρέπει να αντιμετωπίζεται ως καθολικά ωφέλιμη ή καθολικά ουδέτερη για downstream χρήσεις. Τα αποτελέσματα δείχνουν ότι η χρησιμότητα της σύνοψης είναι συνάρτηση (i) της κατηγορίας μεθόδου, (ii) του κόστους παραγωγής, και κυρίως (iii) του τρόπου σχηματισμού query (ποια tags χρησιμοποιούνται). Σε matched συνθήκες, οι hybrid και instructive προσεγγίσεις εμφανίζουν τα πιο συνεπή θετικά σήματα (ιδίως σε Recall@100), ενώ οι extractive και abstractive τείνουν να έχουν μικρότερη ή αρνητική τυπική επίδραση. Παράλληλα, η FIRST-CLAIM(all) αναδεικνύεται ως το πιο ευνοϊκό πεδίο για λειτουργική σύνοψη, σε αντίθεση με μεγάλα πεδία όπως DESCRIPTION(all), όπου η σύνοψη δεν μεταφράζεται τυπικά σε βελτίωση των μετρικών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S. Taduri, K. H. Law, J. P. Kesan, and R. D. Sriram, "Utilization of Bio-Ontologies for Enhancing Patent Information Retrieval," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Milwaukee, WI, USA: IEEE, Jul. 2019, pp. 91–96. doi: 10.1109/COMPSAC.2019.10189.

[2] P. J. Terragno, "Patents as technical literature," *IEEE Trans. Profess. Commun.*, vol. PC-22, no. 2, pp. 101–104, Jun. 1979, doi: 10.1109/TPC.1979.6500290.

[3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st ed. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[4] G. Roda, J. Tait, F. Piroi, and V. Zenz, "CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, vol. 6241, C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, Eds., in Lecture Notes in Computer Science, vol. 6241. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 385–409. doi: 10.1007/978-3-642-15754-7_47.

[5] E. Kamateri, R. Chikkamath, M. Salampasis, L. Andersson, and M. Endres, "Enhancing patent retrieval using automated patent summarization," Jul. 22, 2025, *arXiv*: arXiv:2507.16371. doi: 10.48550/arXiv.2507.16371.

[6] A. Nenkova and K. McKeown, "Automatic Summarization," *Foundations and Trends® in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, Jun. 2011, doi: 10.1561/1500000015.

[7] D. Bankira, S. Panda, S. Ranjan, H. S. Ali, S. Parida, and N. Walubita, "Automatic Extractive text Summarization for Ho Language," in *2023 OITS International Conference on Information Technology (OCIT)*, Raipur, India: IEEE, Dec. 2023, pp. 915–919. doi: 10.1109/OCIT59427.2023.10430990.

[8] K. G. Widowati, N. Budiman, K. Foejiono, and K. Purwandari, "Abstractive Text Summarization Using BERT for Feature Extraction and Seq2Seq Model for Summary Generation," in *2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, Karawang, Indonesia: IEEE, Nov. 2023, pp. 226–230. doi: 10.1109/ICMERALDA60125.2023.10458190.

[9] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," 2022, *arXiv*. doi: 10.48550/ARXIV.2203.02155.

[10] K. Zhu *et al.*, "PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts," in *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, Salt Lake City UT USA: ACM, Nov. 2023, pp. 57–68. doi: 10.1145/3689217.3690621.

[11] M. S. Ansary, "A Hybrid Approach for Automatic Extractive Summarization," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Dhaka, Bangladesh: IEEE, Feb. 2021, pp. 11–15. doi: 10.1109/ICICT4SD50815.2021.9396855.

[12] C. M. Souza, D. S. Bastos, L. A. Souza Filho, and M. R. G. Meireles, "A Study of Training Approaches of a Hybrid Summarisation Model Applied to Patent Dataset," *J. Info. Know. Mgmt.*, vol. 22, no. 05, p. 2350030, Oct. 2023, doi: 10.1142/S0219649223500302.

[13] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020, doi: 10.1007/s11431-020-1647-3.

[14] J. Du and Y. Gao, “Domain Adaptation and Summary Distillation for Unsupervised Query Focused Summarization,” *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 3, pp. 1044–1055, Mar. 2024, doi: 10.1109/TKDE.2023.3296441.

[15] S. Dhokane, C. Deshmukh, A. Bollabattin, S. Karande, B. Karangale, and P. S. Varade, “BM25 Implementation For Information Retrieval: Candidate Shortlister For Recruitment Process,” in *2024 Intelligent Systems and Machine Learning Conference (ISML)*, Hyderabad, India: IEEE, May 2024, pp. 722–727. doi: 10.1109/ISML60050.2024.11007378.

[16] C. R. Pitoyo, S. Sulistyo, and I. Hidayah, “Comparison of Cosine Similarity and BM25 in Book Procurement Prioritization Based on Multi-Source Data Integration,” in *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, Jakarta, Indonesia: IEEE, Jan. 2025, pp. 90–94. doi: 10.1109/ICoCSETI63724.2025.11019421.

[17] B. Alkan, B. B. Tekin, A. Karamanlıoğlu, and İ. Karakaya, “Analysis of Retrieval Performance for Methods Fine-Tuned with ColBERT Architecture,” in *2024 Medical Technologies Congress (TIPTEKNO)*, Mugla, Turkiye: IEEE, Oct. 2024, pp. 1–4. doi: 10.1109/TIPTEKNO63488.2024.10755364.

[18] “optimum/sbert-all-MiniLM-L6-with-pooler · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/optimum/sbert-all-MiniLM-L6-with-pooler

[19] “sentence-transformers/all-mpnet-base-v2 · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[20] “BAAI/bge-base-en-v1.5 · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/BAAI/bge-base-en-v1.5

[21] “intfloat/e5-base-v2 · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/intfloat/e5-base-v2

[22] “AAUBS/PatentSBERTa_V2 · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/AAUBS/PatentSBERTa_V2

[23] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” 2019, *arXiv*. doi: 10.48550/ARXIV.1910.13461.

[24] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 2019, *arXiv*. doi: 10.48550/ARXIV.1912.08777.

[25] “HUPD/hupd-t5-small · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/HUPD/hupd-t5-small

[26] M. Zaheer *et al.*, “Big Bird: Transformers for Longer Sequences,” 2020, doi: 10.48550/ARXIV.2007.14062.

[27] “andreiujica/led-base-big-patent · Hugging Face.” Accessed: Jan. 15, 2026. [Online]. Available: https://huggingface.co/andreiujica/led-base-big-patent

[28] M. Guo *et al.*, “LongT5: Efficient Text-To-Text Transformer for Long Sequences,” 2021, *arXiv*. doi: 10.48550/ARXIV.2112.07916.

[29] A. Grattafiori *et al.*, “The Llama 3 Herd of Models,” 2024, *arXiv*. doi: 10.48550/ARXIV.2407.21783.

[30] A. Q. Jiang *et al.*, “Mistral 7B,” 2023, *arXiv*. doi: 10.48550/ARXIV.2310.06825.

[31] A. Yang *et al.*, “Qwen2 Technical Report,” 2024, *arXiv*. doi: 10.48550/ARXIV.2407.10671.

[32] “TextRank: Bringing Order into Text - ACL Anthology.” Accessed: Jan. 15, 2026. [Online]. Available: https://aclanthology.org/W04-3252/

ΠΑΡΑΡΤΗΜΑ Α : Δείγμα εξόδου σύνοψης σε μορφή JSON και σχήμα δεδομένων

Α.1 Σκοπός του παραρτήματος

Στο παρόν παράρτημα παρουσιάζεται ενδεικτικό δείγμα του μορφότυπου εξόδου που παράγεται μετά την εκτέλεση της σύνοψης. Η έξοδος αποθηκεύεται σε JSON και λειτουργεί ως ενιαίο "πακέτο" πληροφοριών ανά πατέντα: περιλαμβάνει τα βασικά μεταδεδομένα του εγγράφου, τα μεγέθη των αρχικών πεδίων (ανά tag), τις παραγόμενες περιλήψεις (ανά tag), δείκτες συμπίεσης (compression) και δείκτες ομοιότητας (Sentence-BERT similarity), καθώς και τις παραμέτρους του run (μοντέλο, ρυθμίσεις παραγωγής κ.λπ.). Η δομή αυτή επιλέχθηκε ώστε κάθε εκτέλεση σύνοψης να είναι ιχνηλάσιμη, συγκρίσιμη και επαναχρησιμοποιήσιμη σε επόμενα στάδια (π.χ. δημιουργία SGML-like, ευρετηρίαση, retrieval, ανάλυση αποτελεσμάτων).

Α.2 Δομή του JSON αρχείου

Το JSON αρχείο έχει μορφή λίστας (array) από αντικείμενα. Κάθε αντικείμενο αντιστοιχεί σε μία πατέντα/έγγραφο και περιλαμβάνει σταθερό σύνολο κλειδιών:

- docno: μοναδικό αναγνωριστικό πατέντας
- title: τίτλος πατέντας
- IPCR-CLASSIFICATIONS, CPC-CLASSIFICATIONS: ταξινομήσεις (όπως εξάγονται από το κείμενο)
- CITATIONS: παραπομπές/αναφορές
- sources_by_tag: μεγέθη αρχικών πεδίων ανά tag (σε λέξεις/tokens)
- summaries_by_tag: περιλήψεις ανά tag (κείμενο)
- summary_lengths_by_tag: μεγέθη περιλήψεων + δείκτες συμπίεσης ανά tag
- sbert_Similarity_by_tag: δείκτες ομοιότητας ανά tag (P/R/F1)
- meta: μεταδεδομένα run (mode, model, ρυθμίσεις και ορισμοί δεικτών)

Σημείωση: Το TITLE καταγράφεται στα sources_by_tag αλλά δεν συνοψίζεται (δεν εμφανίζεται στο summaries_by_tag), καθώς διατηρείται αυτούσιο ως υψηλής διακριτικής ισχύος πεδίο.

Α.3 Περιγραφή πεδίων (schema)

Αναγνωριστικά και μεταδεδομένα εγγράφου

- docno: μοναδικό κλειδί ταύτισης του εγγράφου σε όλα τα στάδια (σύνοψη, παραγωγή δομημένων αρχείων, retrieval, αξιολόγηση).
- title: διατηρείται για αναγνωσιμότητα, ως σταθερό συμφραζόμενο (context) και για πιθανό εμπλουτισμό queries.

Θεματικές ταξινομήσεις και παραπομπές

- IPCR-CLASSIFICATIONS, CPC-CLASSIFICATIONS: δομημένη πληροφορία ταξινόμησης (χρήσιμη για φίλτρα, αναλύσεις ή μελλοντική tag-aware ανάκτηση).
- CITATIONS: πρόσθετο τεκμηριωτικό πεδίο με αναφορές σε άλλα έγγραφα.

sources_by_tag (μέγεθος αρχικού κειμένου ανά tag)

- Περιγραφή: Λεξικό όπου κάθε tag (π.χ. ABSTRACT, DESCRIPTION, CLAIMS) αποθηκεύεται με μετρήσεις μεγέθους.
- Πεδία ανά tag:
- words: πλήθος λέξεων στο αρχικό πεδίο.
- tokens: πλήθος tokens στο αρχικό πεδίο (στο συγκεκριμένο export ισούται με words, καθώς καταγράφεται ως proxy μέτρηση tokenization στη ροή).
- Χρησιμότητα: ποσοτικοποιεί το μέγεθος εισόδου ανά πεδίο και τεκμηριώνει ποια tags υπερβαίνουν thresholds και άρα συνοψίζονται.

summaries_by_tag (παραγόμενες περιλήψεις ανά tag)

- Περιγραφή: Λεξικό με κλειδιά tags και τιμές το παραγόμενο συνοπτικό κείμενο.
- Χαρακτηριστικά: Η παραγωγή είναι tag-wise, ώστε:
- να εφαρμόζονται διαφορετικά budgets ανά tag,
- να επιτρέπεται σύγκριση "ίδιου πεδίου με ίδιο πεδίο" μεταξύ runs,
- να υποστηρίζονται query σενάρια ως πακέτα tags.

summary_lengths_by_tag (μήκη περιλήψεων και δείκτες συμπίεσης)

- Περιγραφή: Για κάθε tag καταγράφονται το μέγεθος της σύνοψης και δείκτες συμπίεσης σε σχέση με το αρχικό κείμενο.
- Πεδία ανά tag:
- words, tokens: μήκος παραγόμενης σύνοψης.
- compression_ratio_words, compression_ratio_tokens: λόγος source_length / summary_length.
- compression_percent_words, compression_percent_tokens: 1 - (summary_length/source_length).
- Χρησιμότητα: αποτυπώνει πόσο "επιθετική" είναι η σύνοψη ανά tag και επιτρέπει συγκρίσεις κόστους/απόδοσης μεταξύ μεθόδων.

sbert_Similarity_by_tag (ομοιότητα πηγής–σύνοψης ανά tag)

- Περιγραφή: Sentence-level αντιστοίχιση με Sentence-BERT cosine similarity, αποθηκευμένη ως P/R/F1 ανά tag.
- Πεδία ανά tag:
- P, R, F1: δείκτες ομοιότητας/κάλυψης σε επίπεδο προτάσεων.

- Χρησιμότητα: ελαφρύ, υπολογιστικά αποδοτικό σήμα που συνοδεύει κάθε run και βοηθά στη σκιαγράφηση του “προφίλ” μεταβολής περιεχομένου, χωρίς να αποτελεί τον κύριο άξονα αξιολόγησης.

meta (ιχνηλασιμότητα run και ρυθμίσεις παραγωγής)

- Περιγραφή: Μεταδεδομένα που τεκμηριώνουν πλήρως τον τρόπο παραγωγής της σύνοψης.
- Τυπικά περιεχόμενα:
- mode: κατηγορία μεθόδου (π.χ. instructive).
- model, bb_name: ταυτότητα μοντέλου.
- Υπο-αντικείμενο ρυθμίσεων (ενδεικτικά): batch_size, max_input_length_tokens, max_new_tokens, repetition_penalty κ.ά.
- compression_defs και sbert_Similarity.note: ορισμοί/σημειώσεις για τους δείκτες που αποθηκεύονται.

A.4 Ενδεικτικό απόσπασμα JSON

Παρατίθεται αποσπασματικό δείγμα μίας εγγραφής.


```
[
  {
    "docno": "EP-2792413-A1",
    "title": "Filtration systems and methods for filtering particles of   predetermined substance",
    "IPCR-CLASSIFICATIONS":  "\nB01D    46/00                20060101ALI20160720BHEP\nB03C      3/00
20060101AFI20160720BHEP\nB03C       7/06                   20060101ALI20160720BHEP\nB01L       3/00
20060101ALI20160720BHEP\n",
    "CPC-CLASSIFICATIONS": "\nB03C   1/02        20130101 LI20160205BHEP\nB03C   5/00        20130101
LI20160205BHEP\nB03C     5/02              20130101  LI20160205BHEP\nB01L2400/0415          20130101
LA20160205BHEP\nB01L2200/0652          20130101  LA20160205BHEP\nB03C     3/00              20130101
LI20160205BHEP\nB01L     3/502753       20130101  LI20160205BHEP\nB03C     1/00              20130101
LI20160205BHEP\nB01D  35/00        20130101 FI20140918BHEP\n",
    "CITATIONS": "\nUS-20140246321-A1\nWO-2014078507-A1\n",
    "sources_by_tag": {
      "TITLE": {
        "words": 10,
        "tokens": 10
      },
      "ABSTRACT": {
        "words": 52,
        "tokens": 52
      },
      "DESCRIPTION": {
        "words": 2147,
        "tokens": 2147
      },
      "DETAILED-DESCRIPTION": {
        "words": 1751,
        "tokens": 1751
      },
      "BRIEF-DESCRIPTION": {
        "words": 396,
        "tokens": 396
      },
      "FIRST-CLAIM": {
        "words": 47,
        "tokens": 47
      },
      "CLAIMS": {
        "words": 535,
        "tokens": 535
      }
    },
    "summaries_by_tag": {
```

```
      "ABSTRACT": "The filtration system includes filter media including plurality of apertures defined therein, and an array of micropillars. Each micropillar is substantially aligned with one of the plurality of apertures and is configured to be repelled by particles of predetermined substance entrained in flow channeled through the filtration system.",
      "DESCRIPTION": "Here is a concise summary of approximately 225 words focusing on technical novelty: The described filtration systems utilize an innovative approach to separate particles of similar size based on their properties rather than size alone. This is achieved through the use of micropillars coated with specific substances that interact with target particles, inducing attractive or repulsive forces. These interactions allow for the separation of particles without relying solely on size-based filtration. The micropillars can be fabricated from materials that exhibit unique properties, such as bending under electrical or thermal stimulation, enabling precise control over particle interaction. Coatings on the micropillars can be tailored to respond to specific particle types, ensuring targeted capture and release. The filtration systems employ novel mechanisms for removing captured particles, including mechanical manipulation, electrostatic attraction, and magnetic induction. Additionally, the systems incorporate sensors and control systems to monitor and adjust operating conditions, optimizing performance and efficiency. These innovations enable the development of high-performance filtration systems capable of separating particles of similar size based on",
      "DETAILED-DESCRIPTION": "Here is a concise summary of approximately 120 words focusing on technical novelty: A novel filtration system utilizes an array of micropillars that can selectively remove particles of a predetermined substance from a fluid stream based on properties other than size. The micropillars are coated with a substance that induces a reaction with specific particles, allowing for separation of particles with similar sizes. The micropillars can be made from materials like diphenylalanine peptide nanotubes or polyvinylidene fluoride, enabling bending under electrical or magnetic stimulation. This allows for precise manipulation of the micropillars to separate particles. Additionally, the system incorporates sensors and control mechanisms to detect changes in luminosity and adjust the micropillar's behavior accordingly. These innovations enable efficient and targeted removal of specific particles from complex mixtures.",
      "BRIEF-DESCRIPTION": "Here's a concise summary of approximately 120 words focusing on technical novelty: A novel filtration system separates particles of similar size using arrays of micropillars integrated with filter media featuring multiple apertures. In one embodiment, micropillars repel particles of specific substances, allowing them to bypass the filter medium. In another, micropillars attract and collect targeted particles, which can then be transferred to a repository. A third approach involves stimulating micropillars to selectively remove desired particles from the flowing stream. This innovative design addresses limitations of traditional filtration systems, where single-layered filters often fail to capture similarly-sized particles, leading to clogging issues. By leveraging micro-pillared structures, this technology offers improved particle separation efficiency without requiring frequent cleaning or replacement of filter media.",
      "FIRST-CLAIM": "filtration system (200) comprising: repository (240); and an array (220) of micropillars (224), wherein each said micropillar in said array is configured to attract particles (132) of predetermined substance entrained in fluid flow (134) channeled through said filtration system and transfer the attracted particles to said repository.",
      "CLAIMS": "Here's a concise summary of the technical novelty: A novel filtration system uses an array of micropillars that can attract and transfer particles of a specific substance from a fluid stream to a repository. The micropillars have a unique property where they bend towards the repository when the particle load reduces the light intensity detected by a light sensor, indicating sufficient collection. This bending action allows efficient removal of particles from the micropillars once collected. Additionally, the micropillars may feature coatings with charges opposite those of the target particles, enhancing their attractiveness. The system also includes a rotating cylindrical body with micropillars arranged around it, allowing for continuous collection and release of particles. A corresponding method involves positioning the micropillars in the fluid flow path and applying stimuli to selectively remove targeted particles. The micropillars' ability to change shape in response to particle presence and absence enables efficient separation and retrieval of desired substances."
    },
    "summary_lengths_by_tag": {
      "ABSTRACT": {
        "words": 48,
        "tokens": 48,
        "compression_ratio_words": 1.08,
        "compression_percent_words": 0.08,
        "compression_ratio_tokens": 1.08,
        "compression_percent_tokens": 0.08
      },
      "DESCRIPTION": {
        "words": 168,
        "tokens": 168,
        "compression_ratio_words": 12.78,
        "compression_percent_words": 0.92,
        "compression_ratio_tokens": 12.78,
        "compression_percent_tokens": 0.92
      },
      "DETAILED-DESCRIPTION": {
        "words": 127,
        "tokens": 127,
        "compression_ratio_words": 13.79,
```

```
        "compression_percent_words": 0.93,
        "compression_ratio_tokens": 13.79,
        "compression_percent_tokens": 0.93
      },
      "BRIEF-DESCRIPTION": {
        "words": 121,
        "tokens": 121,
        "compression_ratio_words": 3.27,
        "compression_percent_words": 0.69,
        "compression_ratio_tokens": 3.27,
        "compression_percent_tokens": 0.69
      },
      "FIRST-CLAIM": {
        "words": 47,
        "tokens": 47,
        "compression_ratio_words": 1.0,
        "compression_percent_words": 0.0,
        "compression_ratio_tokens": 1.0,
        "compression_percent_tokens": 0.0
      },
      "CLAIMS": {
        "words": 154,
        "tokens": 154,
        "compression_ratio_words": 3.47,
        "compression_percent_words": 0.71,
        "compression_ratio_tokens": 3.47,
        "compression_percent_tokens": 0.71
      }
    },
    "sbert_Similarity_by_tag": {
      "ABSTRACT": {
        "P": 1.0,
        "R": 0.9072,
        "F1": 0.9513
      },
      "DESCRIPTION": {
        "P": 0.6899,
        "R": 0.5673,
        "F1": 0.6226
      },
      "DETAILED-DESCRIPTION": {
        "P": 0.7227,
        "R": 0.5563,
        "F1": 0.6287
      },
      "BRIEF-DESCRIPTION": {
        "P": 0.6901,
        "R": 0.5712,
        "F1": 0.625
      },
      "FIRST-CLAIM": {
        "P": 1.0,
        "R": 1.0,
        "F1": 1.0
      },
      "CLAIMS": {
        "P": 0.7577,
        "R": 0.5498,
        "F1": 0.6372
      }
    },
    "meta": {
      "mode": "instructive",
      "model": "Llama3_8B_Instruct",
      "bb_name": "meta-llama/Meta-Llama-3-8B-Instruct",
      "len_margin": 10,
      "instructive": {
        "batch_size": 32,
        "max_input_length_tokens": 8192,
        "do_sample": false,
        "num_beams": 1,
        "max_new_tokens": 200,
        "repetition_penalty": 1.2
      },
```

```
      "compression_defs": {
        "ratio": "source_length / summary_length (None αν source==0 ή summary==0)",
        "percent": "1 - (summary_length / source_length) (None αν source==0 ή summary==0)"
      },
      "sbert_Similarity": {
        "note": "Sentence-level P/R/F1 with Sentence-BERT cosine (always stored). Lightweight alter-
native to BERTScore."
      }
    }
  },
  ……..
```


A.5 Πώς αξιοποιείται το JSON στην υπόλοιπη ροή

Η συγκεκριμένη δομή υποστηρίζει άμεσα τις ανάγκες της εργασίας, επειδή: απομονώνει τη σύνοψη σε επίπεδο tags (άρα επιτρέπει πειράματα με διαφορετικά “πακέτα” πεδίων), καταγράφει τα budgets έμμεσα μέσω των μηκών και των δεικτών συμπίεσης, διατηρεί ιχνηλασιμότητα (meta) και επιτρέπει συγκρίσιμες αναλύσεις μεταξύ runs, παρέχει ενιαία είσοδο για τη δημιουργία των τελικών δομημένων αρχείων που χρησιμοποιούνται στα σενάρια queries και στα επόμενα στάδια ευρετηρίασης/ανάκτησης.

ΠΑΡΑΡΤΗΜΑ Β : Αναλυτικά Αποτελέσματα

Στο παρόν παράρτημα παρατίθεται ο πλήρης κατάλογος των πειραματικών εκτελέσεων (runs) και των αντίστοιχων μετρήσεων αξιολόγησης που παρήχθησαν από το σύστημα πειραματισμού και καταγράφηκαν κατά την εκτέλεση των σεναρίων σύνοψης και ανάκτησης. Για κάθε run παρουσιάζονται, σε ενιαία μορφή, τα βασικά αναγνωριστικά της εκτέλεσης (ευρετήριο, ρυθμίσεις ανάκτησης, ρυθμίσεις σύνοψης και σενάριο tags), καθώς και οι δείκτες αποτελεσματικότητας (π.χ. MAP, Precision/Recall σε διαφορετικά k, nDCG, PRES) και οι σχετικοί δείκτες συμπίεσης/ομοιότητας και χρόνου εκτέλεσης. Επειδή ο πίνακας περιλαμβάνει μεγάλο αριθμό στηλών, η παρουσίαση γίνεται σε επιμέρους blocks με σταθερή δομή· η συσχέτιση των blocks επιτυγχάνεται μέσω του μοναδικού αναγνωριστικού run (id), το οποίο επιτρέπει την ανασύνθεση του πλήρους προφίλ κάθε εκτέλεσης.

B.1 Index & Retrieval/Summary keys

id	index_key	index_type	index_size_details	retrieval_key	vector_search	summary_key
1	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_171802_101183	100 100 None	20251119_160543
2	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172035_676116	100 100 None	20251121_113033
3	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172302_721203	100 100 None	20251121_215810
4	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172539_289566	100 100 None	20251122_102002
5	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172736_273452	100 100 None	20251122_105620
6	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172927_896728	100 100 None	20251122_134230
7	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173141_559014	100 100 None	20251113_153943
8	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173439_000752	100 100 None	20251113_155037
9	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173740_775574	100 100 None	20251113_155506
10	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174000_337031	100 100 None	20251113_170624
11	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174300_678816	100 100 None	20251113_095150
12	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174557_812855	100 100 None	20251113_143829

13	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174854_297909	100 100 None	20251124_111502
14	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175138_496118	100 100 None	20251124_205730
15	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175435_056585	100 100 None	20251125_090428
16	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175801_005912	100 100 None	20251125_224259
17	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_180151_782115	100 100 None	20251107_110523
18	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_191634_152595	100 100 None	20251119_160543
19	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_191842_077677	100 100 None	20251121_113033
20	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192055_574819	100 100 None	20251121_215810
21	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192333_287095	100 100 None	20251122_102002
22	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192521_163835	100 100 None	20251122_105620
23	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192704_761855	100 100 None	20251122_134230
24	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192906_087576	100 100 None	20251113_153943
25	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193205_133542	100 100 None	20251113_155037
26	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193509_469072	100 100 None	20251113_155506
27	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193706_783044	100 100 None	20251113_170624
28	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193954_982016	100 100 None	20251113_095150
29	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194250_622564	100 100 None	20251113_143829
30	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194547_010483	100 100 None	20251124_111502
31	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194813_482264	100 100 None	20251124_205730
32	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195051_647240	100 100 None	20251125_090428
33	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195359_120204	100 100 None	20251125_224259
34	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195732_681500	100 100 None	20251107_110523
35	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200134_597903	100 100 None	20251119_160543
36	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200423_769314	100 100 None	20251121_113033
37	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200715_619260	100 100 None	20251121_215810

38	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201018_039214	100 100 None	20251122_102002
39	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201228_566070	100 100 None	20251122_105620
40	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201430_885133	100 100 None	20251122_134230
41	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201657_965821	100 100 None	20251113_153943
42	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202025_952252	100 100 None	20251113_155037
43	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202400_594773	100 100 None	20251113_155506
44	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202626_287471	100 100 None	20251113_170624
45	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202944_791277	100 100 None	20251113_095150
46	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203306_343076	100 100 None	20251113_143829
47	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203629_576784	100 100 None	20251124_111502
48	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203951_600387	100 100 None	20251124_205730
49	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204340_081879	100 100 None	20251125_090428
50	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204811_310344	100 100 None	20251125_224259
51	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_205327_055371	100 100 None	20251107_110523
52	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172922_704527	100 100 None	20251119_160543
53	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173111_864446	100 100 None	20251121_113033
54	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173249_023758	100 100 None	20251121_215810
55	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173427_827024	100 100 None	20251122_102002
56	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173621_679109	100 100 None	20251122_105620
57	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173758_792736	100 100 None	20251122_134230
58	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173937_279061	100 100 None	20251113_153943
59	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174116_634076	100 100 None	20251113_155037
60	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174255_287625	100 100 None	20251113_155506
61	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174423_165700	100 100 None	20251113_170624
62	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174555_339037	100 100 None	20251113_095150

63	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174751_526055	100 100 None	20251113_143829
64	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174932_862747	100 100 None	20251124_111502
65	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175119_933528	100 100 None	20251124_205730
66	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175310_440875	100 100 None	20251125_090428
67	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175502_402365	100 100 None	20251125_224259
68	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175700_855024	100 100 None	20251107_110523
69	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_191703_054789	100 100 None	20251119_160543
70	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_191852_951546	100 100 None	20251121_113033
71	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192058_280320	100 100 None	20251121_215810
72	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192319_359705	100 100 None	20251122_102002
73	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192423_061269	100 100 None	20251122_105620
74	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192540_145401	100 100 None	20251122_134230
75	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192702_444201	100 100 None	20251113_153943
76	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192949_883963	100 100 None	20251113_155037
77	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193240_931261	100 100 None	20251113_155506
78	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193418_864340	100 100 None	20251113_170624
79	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193650_277102	100 100 None	20251113_095150
80	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193931_483575	100 100 None	20251113_143829
81	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194213_545427	100 100 None	20251124_111502
82	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194418_556520	100 100 None	20251124_205730
83	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194630_352212	100 100 None	20251125_090428
84	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194910_861231	100 100 None	20251125_224259
85	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195213_311356	100 100 None	20251107_110523
86	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195626_347406	100 100 None	20251119_160543
87	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195749_533319	100 100 None	20251121_113033

88	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195912_221848	100 100 None	20251121_215810
89	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200035_965510	100 100 None	20251122_102002
90	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200152_257929	100 100 None	20251122_105620
91	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200308_264039	100 100 None	20251122_134230
92	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200429_084357	100 100 None	20251113_153943
93	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200556_835790	100 100 None	20251113_155037
94	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200726_367490	100 100 None	20251113_155506
95	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200844_877619	100 100 None	20251113_170624
96	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201010_252395	100 100 None	20251113_095150
97	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201137_983632	100 100 None	20251113_143829
98	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201306_694239	100 100 None	20251124_111502
99	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201438_335421	100 100 None	20251124_205730
100	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201617_055049	100 100 None	20251125_090428
101	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201758_834447	100 100 None	20251125_224259
102	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201950_647414	100 100 None	20251107_110523
103	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202630_335625	100 100 None	20251119_160543
104	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202921_374643	100 100 None	20251121_113033
105	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203208_764310	100 100 None	20251121_215810
106	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203504_252469	100 100 None	20251122_102002
107	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203712_325882	100 100 None	20251122_105620
108	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203916_499043	100 100 None	20251122_134230
109	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204141_354973	100 100 None	20251113_153943
110	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204459_770086	100 100 None	20251113_155037
111	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204821_480772	100 100 None	20251113_155506
112	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_205044_025529	100 100 None	20251113_170624

113	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_205357_803822	100 100 None	20251113_095150
114	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_205711_821439	100 100 None	20251113_143829
115	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_210025_448047	100 100 None	20251124_111502
116	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_210339_176170	100 100 None	20251124_205730
117	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_210713_412239	100 100 None	20251125_090428
118	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_211125_887466	100 100 None	20251125_224259
119	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_211617_870642	100 100 None	20251107_110523
120	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_172926_428053	100 100 None	20251119_160543
121	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173308_974241	100 100 None	20251121_113033
122	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_173647_627459	100 100 None	20251121_215810
123	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174000_361391	100 100 None	20251122_102002
124	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174244_475735	100 100 None	20251122_105620
125	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174533_889558	100 100 None	20251122_134230
126	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_174827_808381	100 100 None	20251113_153943
127	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175132_497525	100 100 None	20251113_155037
128	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175426_452947	100 100 None	20251113_155506
129	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175630_710901	100 100 None	20251113_170624
130	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_175902_291953	100 100 None	20251113_095150
131	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_180211_243102	100 100 None	20251113_143829
132	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_180518_980113	100 100 None	20251124_111502
133	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_180950_021732	100 100 None	20251124_205730
134	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_181444_935472	100 100 None	20251125_090428
135	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_182026_278786	100 100 None	20251125_224259
136	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_182659_771686	100 100 None	20251107_110523
137	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_191743_631734	100 100 None	20251119_160543

138	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192043_041476	100 100 None	20251121_113033
139	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192359_683620	100 100 None	20251121_215810
140	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192630_474888	100 100 None	20251122_102002
141	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_192822_021717	100 100 None	20251122_105620
142	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193026_256448	100 100 None	20251122_134230
143	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193243_040928	100 100 None	20251113_153943
144	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193500_482737	100 100 None	20251113_155037
145	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193704_391381	100 100 None	20251113_155506
146	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_193828_583632	100 100 None	20251113_170624
147	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194012_183692	100 100 None	20251113_095150
148	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194229_324655	100 100 None	20251113_143829
149	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194446_571748	100 100 None	20251124_111502
150	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_194824_877140	100 100 None	20251124_205730
151	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195231_283335	100 100 None	20251125_090428
152	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_195728_034878	100 100 None	20251125_224259
153	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200306_536962	100 100 None	20251107_110523
154	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200604_333189	100 100 None	20251119_160543
155	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200759_678090	100 100 None	20251121_113033
156	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_200959_960816	100 100 None	20251121_215810
157	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201223_646392	100 100 None	20251122_102002
158	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201439_878803	100 100 None	20251122_105620
159	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201656_395396	100 100 None	20251122_134230
160	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_201912_053776	100 100 None	20251113_153943
161	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202124_728964	100 100 None	20251113_155037
162	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202331_152380	100 100 None	20251113_155506

163	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202518_042133	100 100 None	20251113_170624
164	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202708_980099	100 100 None	20251113_095150
165	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_202925_588646	100 100 None	20251113_143829
166	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203139_232610	100 100 None	20251124_111502
167	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203357_008526	100 100 None	20251124_205730
168	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203621_399606	100 100 None	20251125_090428
169	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_203838_236553	100 100 None	20251125_224259
170	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_204059_084333	100 100 None	20251107_110523
171	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_232537_728106	100 100 None	NO_SUMMARY_KEY
172	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_233800_318274	100 100 None	NO_SUMMARY_KEY
173	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_234519_664805	100 100 None	NO_SUMMARY_KEY
174	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_234825_157156	100 100 None	NO_SUMMARY_KEY
175	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_235317_537832	100 100 None	NO_SUMMARY_KEY
176	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_000049_423971	100 100 None	NO_SUMMARY_KEY
177	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_000358_722060	100 100 None	NO_SUMMARY_KEY
178	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_000652_483000	100 100 None	NO_SUMMARY_KEY
179	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_001650_753522	100 100 None	NO_SUMMARY_KEY
180	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_231026_194667	100 100 None	20251119_160543
181	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_231238_193879	100 100 None	20251121_113033
182	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_231458_156957	100 100 None	20251121_215810
183	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_231723_989330	100 100 None	20251122_102002
184	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_231913_387962	100 100 None	20251122_105620
185	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_232056_493580	100 100 None	20251122_134230
186	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_232259_310082	100 100 None	20251113_153943
187	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_232540_366559	100 100 None	20251113_155037

188	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_232828_452414	100 100 None	20251113_155506
189	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_233030_942849	100 100 None	20251113_170624
190	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_233317_810187	100 100 None	20251113_095150
191	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_233557_049417	100 100 None	20251113_143829
192	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_233836_775725	100 100 None	20251124_111502
193	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_234109_003662	100 100 None	20251124_205730
194	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_234351_843554	100 100 None	20251125_090428
195	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_234706_658214	100 100 None	20251125_224259
196	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251211_235047_388806	100 100 None	20251107_110523
197	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_003513_895842	100 100 None	NO_SUMMARY_KEY
198	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_081050_629143	100 100 None	NO_SUMMARY_KEY
199	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_081936_526634	100 100 None	NO_SUMMARY_KEY
200	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_083604_718458	100 100 None	20251119_160543
201	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_083913_145271	100 100 None	20251121_113033
202	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_084241_800087	100 100 None	20251121_215810
203	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_084632_281207	100 100 None	20251122_102002
204	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_084820_266807	100 100 None	20251122_105620
205	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_085027_464548	100 100 None	20251122_134230
206	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_085244_759268	100 100 None	20251113_153943
207	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_085715_969791	100 100 None	20251113_155037
208	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_090157_912071	100 100 None	20251113_155506
209	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_090439_522211	100 100 None	20251113_170624
210	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_090852_245038	100 100 None	20251113_095150
211	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_091316_090219	100 100 None	20251113_143829
212	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_091742_650111	100 100 None	20251124_111502

213	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_092111_968846	100 100 None	20251124_205730
214	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_092448_130862	100 100 None	20251125_090428
215	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_092910_985827	100 100 None	20251125_224259
216	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251212_093415_863164	100 100 None	20251107_110523
217	20251206_102212	BM25_PYSERINI_k6080	ChunkAll_6GBNEW	20251215_083818_258332	100 100 None	20251214_205923
218	20251015_150601	Colbert	chunk0_8GB	20251109_092023	100 100 1000	20251107_003442
219	20251026_121417	Colbert	Chunk1_40GB	20251107_013447	100 100 1000	NO_SUMMARY_KEY
220	20251026_121417	Colbert	Chunk1_40GB	20251107_142806	100 100 1000	20251107_003442
221	20251026_121417	Colbert	Chunk1_40GB	20251107_193528	100 100 1000	20251107_003442
222	20251026_121417	Colbert	Chunk1_40GB	20251107_013447	100 100 1000	20251107_003442
223	20251026_121417	Colbert	Chunk1_40GB	20251107_161619	100 100 1000	20251107_110523
224	20251026_121417	Colbert	Chunk1_40GB	20251107_123015	100 100 1000	20251107_110523
225	20251025_144025	Colbert	Chunk2_63GB	20251115_132657	100 100 1000	NO_SUMMARY_KEY
226	20251025_144025	Colbert	Chunk2_63GB	20251110_011555	100 200 1000	20251107_003442
227	20251025_144025	Colbert	Chunk2_63GB	20251108_171344	100 100 1000	20251107_003442
228	20251025_144025	Colbert	Chunk2_63GB	20251116_124935	100 100 1000	20251113_095150
229	20251025_144025	Colbert	Chunk2_63GB	20251116_115704	100 100 1000	20251113_095150
230	20251025_144025	Colbert	Chunk2_63GB	20251109_221253	100 100 1000	20251109_152911
231	20251025_144025	Colbert	Chunk2_63GB	20251116_012723	100 100 1000	20251115_232158
232	20251025_144025	Colbert	Chunk2_63GB	20251109_022937	100 100 1000	20251107_003442
233	20251025_144025	Colbert	Chunk2_63GB	20251108_144723	100 100 1000	20251107_003442
234	20251025_144025	Colbert	Chunk2_63GB	20251109_194936	100 100 1000	20251109_152911
235	20251025_144025	Colbert	Chunk2_63GB	20251111_115124	100 100 1000	20251111_012933
236	20251025_144025	Colbert	Chunk2_63GB	20251108_195725	100 100 1000	20251107_003442
237	20251025_144025	Colbert	Chunk2_63GB	20251128_105259	100 100 1000	NO_SUMMARY_KEY

238	20251025_144025	Colbert	Chunk2_63GB	20251116_182807	100 100 1000	20251113_095150
239	20251025_144025	Colbert	Chunk2_63GB	20251116_210824	100 100 1000	20251113_095150
240	20251025_144025	Colbert	Chunk2_63GB	20251116_214512	100 100 1000	20251113_095150
241	20251025_144025	Colbert	Chunk2_63GB	20251117_000811	100 100 1000	20251113_095150
242	20251117_140801	Colbert	Chunk2_67GB	20251117_205610	100 100 1000	20251113_095150
243	20251025_144025	Colbert	Chunk2_63GB	20251114_105420	100 100 1000	20251113_095150
244	20251025_144025	Colbert	Chunk2_63GB	20251113_201038	100 100 1000	20251113_153943
245	20251025_144025	Colbert	Chunk2_63GB	20251113_224817	100 100 1000	20251113_155037
246	20251025_144025	Colbert	Chunk2_63GB	20251113_225402	100 100 1000	20251113_155506
247	20251025_144025	Colbert	Chunk2_63GB	20251113_201717	100 100 1000	20251113_170624
248	20251025_144025	Colbert	Chunk2_63GB	20251114_110302	100 100 1000	20251113_143829
249	20251025_144025	Colbert	Chunk2_63GB	20251115_135405	100 100 1000	NO_SUMMARY_KEY
250	20251025_144025	Colbert	Chunk2_63GB	20251113_115634	100 100 1000	20251113_095150
251	20251025_144025	Colbert	Chunk2_63GB	20251114_134149	100 100 1000	20251113_153943
252	20251025_144025	Colbert	Chunk2_63GB	20251114_161122	100 100 1000	20251113_155037
253	20251025_144025	Colbert	Chunk2_63GB	20251114_215024	100 100 1000	20251113_155506
254	20251025_144025	Colbert	Chunk2_63GB	20251114_142524	100 100 1000	20251113_170624
255	20251025_144025	Colbert	Chunk2_63GB	20251114_134217	100 100 1000	20251113_143829
256	20251025_144025	Colbert	Chunk2_63GB	20251115_215142	100 100 1000	NO_SUMMARY_KEY
257	20251025_144025	Colbert	Chunk2_63GB	20251115_155427	100 100 1000	20251113_095150
258	20251025_144025	Colbert	Chunk2_63GB	20251115_161741	100 100 1000	20251113_153943
259	20251025_144025	Colbert	Chunk2_63GB	20251115_165413	100 100 1000	20251113_155037
260	20251025_144025	Colbert	Chunk2_63GB	20251115_193108	100 100 1000	20251113_155506
261	20251025_144025	Colbert	Chunk2_63GB	20251115_193120	100 100 1000	20251113_170624
262	20251025_144025	Colbert	Chunk2_63GB	20251115_192056	100 100 1000	20251113_143829

263	20251025_144025	Colbert	Chunk2_63GB	20251120_143140	100 100 1000	20251119_160543
264	20251025_144025	Colbert	Chunk2_63GB	20251123_011438	100 100 1000	20251122_102002
265	20251025_144025	Colbert	Chunk2_63GB	20251120_150953	100 100 1000	20251119_160543
266	20251025_144025	Colbert	Chunk2_63GB	20251120_172205	100 100 1000	20251119_160543
267	20251025_144025	Colbert	Chunk2_63GB	20251120_173723	100 100 1000	20251119_160543
268	20251025_144025	Colbert	Chunk2_63GB	20251120_224551	100 100 1000	20251119_160543
269	20251025_144025	Colbert	Chunk2_63GB	20251120_225041	100 100 1000	20251119_160543
270	20251025_144025	Colbert	Chunk2_63GB	20251123_215742	100 300 3000	20251122_102002
271	20251025_144025	Colbert	Chunk2_63GB	20251120_114932	100 100 1000	20251119_160543
272	20251025_144025	Colbert	Chunk2_63GB	20251122_215822	100 100 1000	20251122_102002
273	20251025_144025	Colbert	Chunk2_63GB	20251122_222548	100 100 1000	20251122_105620
274	20251025_144025	Colbert	Chunk2_63GB	20251122_223841	100 100 1000	20251121_215810
275	20251025_144025	Colbert	Chunk2_63GB	20251123_010124	100 100 1000	20251122_134230
276	20251025_144025	Colbert	Chunk2_63GB	20251123_010100	100 100 1000	20251121_113033
277	20251025_144025	Colbert	Chunk2_63GB	20251120_203046	100 100 1000	20251119_160543
278	20251025_144025	Colbert	Chunk2_63GB	20251123_095901	100 100 1000	20251122_134230
279	20251025_144025	Colbert	Chunk2_63GB	20251123_095854	100 100 1000	20251121_113033
280	20251025_144025	Colbert	Chunk2_63GB	20251123_122816	100 100 1000	20251122_105620
281	20251025_144025	Colbert	Chunk2_63GB	20251123_122752	100 100 1000	20251122_102002
282	20251025_144025	Colbert	Chunk2_63GB	20251123_145045	100 100 1000	20251121_215810
283	20251025_144025	Colbert	Chunk2_63GB	20251120_201942	100 100 1000	20251119_160543
284	20251025_144025	Colbert	Chunk2_63GB	20251123_145932	100 100 1000	20251122_134230
285	20251025_144025	Colbert	Chunk2_63GB	20251123_171725	100 100 1000	20251121_113033
286	20251025_144025	Colbert	Chunk2_63GB	20251123_171730	100 100 1000	20251122_105620
287	20251025_144025	Colbert	Chunk2_63GB	20251123_193517	100 100 1000	20251122_102002

288	20251025_144025	Colbert	Chunk2_63GB	20251123_193534	100 100 1000	20251121_215810
289	20251025_144025	Colbert	Chunk2_63GB	20251126_192801	100 100 1000	20251124_205730
290	20251025_144025	Colbert	Chunk2_63GB	20251126_132654	100 100 1000	20251124_111502
291	20251025_144025	Colbert	Chunk2_63GB	20251126_214601	100 100 1000	20251125_224259
292	20251025_144025	Colbert	Chunk2_63GB	20251126_192813	100 100 1000	20251125_090428
293	20251025_144025	Colbert	Chunk2_63GB	20251126_215816	100 100 1000	20251124_205730
294	20251025_144025	Colbert	Chunk2_63GB	20251127_001907	100 100 1000	20251124_111502
295	20251025_144025	Colbert	Chunk2_63GB	20251127_001107	100 100 1000	20251125_224259
296	20251025_144025	Colbert	Chunk2_63GB	20251126_215927	100 100 1000	20251125_090428
297	20251025_144025	Colbert	Chunk2_63GB	20251127_105634	100 100 1000	20251124_205730
298	20251025_144025	Colbert	Chunk2_63GB	20251127_002321	100 100 1000	20251124_111502
299	20251025_144025	Colbert	Chunk2_63GB	20251127_135054	100 100 1000	20251125_224259
300	20251025_144025	Colbert	Chunk2_63GB	20251127_105628	100 100 1000	20251125_090428
301	20251025_144025	Colbert	Chunk2_63GB	20251127_135600	100 100 1000	20251124_111502
302	20251025_144025	Colbert	Chunk2_63GB	20251127_161617	100 100 1000	20251124_205730
303	20251025_144025	Colbert	Chunk2_63GB	20251127_161543	100 100 1000	20251125_224259
304	20251025_144025	Colbert	Chunk2_63GB	20251127_185837	100 100 1000	20251125_090428
305	20251025_144025	Colbert	Chunk2_63GB	20251127_190155	100 100 1000	20251124_111502
306	20251025_144025	Colbert	Chunk2_63GB	20251127_212402	100 100 1000	20251124_205730
307	20251025_144025	Colbert	Chunk2_63GB	20251127_212331	100 100 1000	20251125_090428
308	20251025_144025	Colbert	Chunk2_63GB	20251127_234424	100 100 1000	20251125_224259
309	20251025_144025	Colbert	Chunk2_63GB	20251127_234947	100 100 1000	20251124_205730
310	20251025_144025	Colbert	Chunk2_63GB	20251128_105341	100 100 1000	NO_SUMMARY_KEY
311	20251025_144025	Colbert	Chunk2_63GB	20251203_110325	100 100 1000	20251107_110523
312	TRAIN	Colbert	100_1000	20251203_123953	100 100 1000	20251122_102002

313	20251025_144025	Colbert	Chunk2_63GB	20251203_134430	100 100 1000	20251122_134230
314	20251025_144025	Colbert	Chunk2_63GB	20251203_170403	100 100 1000	20251122_105620
315	20251025_144025	Colbert	Chunk2_63GB	20251203_194433	100 100 1000	20251122_134230
316	20251025_144025	Colbert	Chunk2_63GB	20251203_200920	100 100 1000	20251122_102002
317	20251025_144025	Colbert	Chunk2_63GB	20251203_223826	100 100 1000	20251121_113033
318	20251025_144025	Colbert	Chunk2_63GB-BM25-Hybrid	20251212_194354	100 100 1000	20251122_102002
319	20251025_144025	Colbert	Chunk2_63GB	20251215_094211	100 100 1000	20251214_205923

B.2 Summary configuration & query scenario

id	summary_type	summary_model	dataset_used	queries_processed	query_length	tags_used_token
1	abstractive	LED_Base_BigPatent	Summarized	2592	328	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
2	abstractive	PEGASUS_BigPatent	Summarized	2592	313	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
3	abstractive	LongT5_TGlobal_Base	Summarized	2592	371	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
4	abstractive	BART_large_CNN	Summarized	2592	228	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)

5	abstractive	BigBird_Pegasus	Summarized	2592	256	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
6	abstractive	T5_Small_HUPD	Summarized	2592	260	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
7	extractive	all-mpnet-base-v2	Summarized	2592	393	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
8	extractive	bge-base-en-v1.5	Summarized	2592	371	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
9	extractive	e5-base-v2	Summarized	2592	221	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
10	extractive	google_bert_for_patents	Summarized	2592	292	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
11	extractive	PatentSBERTa_V2	Summarized	2592	398	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
12	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	396	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRSTCLAIM(all)
13	instructive	Mistral_7B_Instruct	Summarized	2592	331	TITLE(all), ABSTRACT(100),

						DESCRIPTION(250), FIRST-CLAIM(all)
14	instructive	Llama3_8B_Instruct	Summarized	2592	331	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
15	instructive	Qwen2_7B_Instruct	Summarized	2592	331	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
16	instructive	Qwen2_14B_Instruct	Summarized	2592	333	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
17	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	418	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
18	abstractive	LED_Base_BigPatent	Summarized	2592	246	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
19	abstractive	PEGASUS_BigPatent	Summarized	2592	240	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
20	abstractive	LongT5_TGlobal_Base	Summarized	2592	317	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
21	abstractive	BART_large_CNN	Summarized	2592	156	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
22	abstractive	BigBird_Pegasus	Summarized	2592	183	TITLE(all), ABSTRACT(all), DESCRIPTION(all)

23	abstractive	T5_Small_HUPD	Summarized	2592	182	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
24	extractive	all-mpnet-base-v2	Summarized	2592	334	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
25	extractive	bge-base-en-v1.5	Summarized	2592	300	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
26	extractive	e5-base-v2	Summarized	2592	128	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
27	extractive	google_bert_for_patents	Summarized	2592	203	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
28	extractive	PatentSBERTa_V2	Summarized	2592	342	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
29	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	338	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
30	instructive	Mistral_7B_Instruct	Summarized	2592	245	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
31	instructive	Llama3_8B_Instruct	Summarized	2592	246	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
32	instructive	Qwen2_7B_Instruct	Summarized	2592	247	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
33	instructive	Qwen2_14B_Instruct	Summarized	2592	248	TITLE(all), ABSTRACT(all), DESCRIPTION(all)

34	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	435	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
35	abstractive	LED_Base_BigPatent	Summarized	2592	998	ALL TAGS & WORDS
36	abstractive	PEGASUS_BigPatent	Summarized	2592	992	ALL TAGS & WORDS
37	abstractive	LongT5_TGlobal_Base	Summarized	2592	1194	ALL TAGS & WORDS
38	abstractive	BART_large_CNN	Summarized	2592	472	ALL TAGS & WORDS
39	abstractive	BigBird_Pegasus	Summarized	2592	601	ALL TAGS & WORDS
40	abstractive	T5_Small_HUPD	Summarized	2592	610	ALL TAGS & WORDS
41	extractive	all-mpnet-base-v2	Summarized	2592	1277	ALL TAGS & WORDS
42	extractive	bge-base-en-v1.5	Summarized	2592	1149	ALL TAGS & WORDS
43	extractive	e5-base-v2	Summarized	2592	498	ALL TAGS & WORDS
44	extractive	google_bert_for_patents	Summarized	2592	774	ALL TAGS & WORDS
45	extractive	PatentSBERTa_V2	Summarized	2592	1309	ALL TAGS & WORDS
46	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	1296	ALL TAGS & WORDS
47	instructive	Mistral_7B_Instruct	Summarized	2592	921	ALL TAGS & WORDS
48	instructive	Llama3_8B_Instruct	Summarized	2592	928	ALL TAGS & WORDS
49	instructive	Qwen2_7B_Instruct	Summarized	2592	931	ALL TAGS & WORDS
50	instructive	Qwen2_14B_Instruct	Summarized	2592	936	ALL TAGS & WORDS
51	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	2128	ALL TAGS & WORDS
52	abstractive	LED_Base_BigPatent	Summarized	2592	161	TITLE(all), ABSTRACT(all), FIRSTCLAIM(all)
53	abstractive	PEGASUS_BigPatent	Summarized	2592	141	TITLE(all), ABSTRACT(all), FIRSTCLAIM(all)

54	abstractive	LongT5_TGlobal_Base	Summarized	2592	204	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
55	abstractive	BART_large_CNN	Summarized	2592	193	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
56	abstractive	BigBird_Pegasus	Summarized	2592	193	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
57	abstractive	T5_Small_HUPD	Summarized	2592	199	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
58	extractive	all-mpnet-base-v2	Summarized	2592	205	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
59	extractive	bge-base-en-v1.5	Summarized	2592	189	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
60	extractive	e5-base-v2	Summarized	2592	144	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
61	extractive	google_bert_for_patents	Summarized	2592	160	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
62	extractive	PatentSBERTa_V2	Summarized	2592	209	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
63	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	207	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
64	instructive	Mistral_7B_Instruct	Summarized	2592	207	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)

65	instructive	Llama3_8B_Instruct	Summarized	2592	206	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
66	instructive	Qwen2_7B_Instruct	Summarized	2592	204	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
67	instructive	Qwen2_14B_Instruct	Summarized	2592	205	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
68	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	224	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
69	abstractive	LED_Base_BigPatent	Summarized	2592	169	DESCRIPTION(all)
70	abstractive	PEGASUS_BigPatent	Summarized	2592	172	DESCRIPTION(all)
71	abstractive	LongT5_TGlobal_Base	Summarized	2592	213	DESCRIPTION(all)
72	abstractive	BART_large_CNN	Summarized	2592	50	DESCRIPTION(all)
73	abstractive	BigBird_Pegasus	Summarized	2592	78	DESCRIPTION(all)
74	abstractive	T5_Small_HUPD	Summarized	2592	77	DESCRIPTION(all)
75	extractive	all-mpnet-base-v2	Summarized	2592	237	DESCRIPTION(all)
76	extractive	bge-base-en-v1.5	Summarized	2592	217	DESCRIPTION(all)
77	extractive	e5-base-v2	Summarized	2592	80	DESCRIPTION(all)
78	extractive	google_bert_for_patents	Summarized	2592	141	DESCRIPTION(all)
79	extractive	PatentSBERTa_V2	Summarized	2592	242	DESCRIPTION(all)
80	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	240	DESCRIPTION(all)
81	instructive	Mistral_7B_Instruct	Summarized	2592	140	DESCRIPTION(all)
82	instructive	Llama3_8B_Instruct	Summarized	2592	141	DESCRIPTION(all)
83	instructive	Qwen2_7B_Instruct	Summarized	2592	142	DESCRIPTION(all)
84	instructive	Qwen2_14B_Instruct	Summarized	2592	142	DESCRIPTION(all)

85	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	417	DESCRIPTION(all)
86	abstractive	LED_Base_BigPatent	Summarized	2592	156	FIRST-CLAIM(all)
87	abstractive	PEGASUS_BigPatent	Summarized	2592	146	FIRST-CLAIM(all)
88	abstractive	LongT5_TGlobal_Base	Summarized	2592	192	FIRST-CLAIM(all)
89	abstractive	BART_large_CNN	Summarized	2592	118	FIRST-CLAIM(all)
90	abstractive	BigBird_Pegasus	Summarized	2592	128	FIRST-CLAIM(all)
91	abstractive	T5_Small_HUPD	Summarized	2592	134	FIRST-CLAIM(all)
92	extractive	all-mpnet-base-v2	Summarized	2592	209	FIRST-CLAIM(all)
93	extractive	bge-base-en-v1.5	Summarized	2592	198	FIRST-CLAIM(all)
94	extractive	e5-base-v2	Summarized	2592	133	FIRST-CLAIM(all)
95	extractive	google_bert_for_patents	Summarized	2592	159	FIRST-CLAIM(all)
96	extractive	PatentSBERTa_V2	Summarized	2592	212	FIRST-CLAIM(all)
97	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	212	FIRST-CLAIM(all)
98	instructive	Mistral_7B_Instruct	Summarized	2592	165	FIRST-CLAIM(all)
99	instructive	Llama3_8B_Instruct	Summarized	2592	164	FIRST-CLAIM(all)
100	instructive	Qwen2_7B_Instruct	Summarized	2592	163	FIRST-CLAIM(all)
101	instructive	Qwen2_14B_Instruct	Summarized	2592	165	FIRST-CLAIM(all)
102	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	285	FIRST-CLAIM(all)
103	abstractive	LED_Base_BigPatent	Summarized	2592	434	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
104	abstractive	PEGASUS_BigPatent	Summarized	2592	435	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
105	abstractive	LongT5_TGlobal_Base	Summarized	2592	429	TITLE(all), ABSTRACT(all),

						DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
106	abstractive	BART_large_CNN	Summarized	2592	309	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
107	abstractive	BigBird_Pegasus	Summarized	2592	350	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
108	abstractive	T5_Small_HUPD	Summarized	2592	364	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
109	extractive	all-mpnet-base-v2	Summarized	2592	436	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
110	extractive	bge-base-en-v1.5	Summarized	2592	418	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
111	extractive	e5-base-v2	Summarized	2592	256	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
112	extractive	google_bert_for_patents	Summarized	2592	338	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
113	extractive	PatentSBERTa_V2	Summarized	2592	439	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)

114	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	438	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
115	instructive	Mistral_7B_Instruct	Summarized	2592	430	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
116	instructive	Llama3_8B_Instruct	Summarized	2592	431	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
117	instructive	Qwen2_7B_Instruct	Summarized	2592	433	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
118	instructive	Qwen2_14B_Instruct	Summarized	2592	435	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
119	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	448	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
120	abstractive	LED_Base_BigPatent	Summarized	2592	239	TITLE(all), ABSTRACT(all), CLAIMS(all)
121	abstractive	PEGASUS_BigPatent	Summarized	2592	233	TITLE(all), ABSTRACT(all), CLAIMS(all)
122	abstractive	LongT5_TGlobal_Base	Summarized	2592	243	TITLE(all), ABSTRACT(all), CLAIMS(all)

123	abstractive	BART_large_CNN	Summarized	2592	182	TITLE(all), ABSTRACT(all), CLAIMS(all)
124	abstractive	BigBird_Pegasus	Summarized	2592	201	TITLE(all), ABSTRACT(all), CLAIMS(all)
125	abstractive	T5_Small_HUPD	Summarized	2592	210	TITLE(all), ABSTRACT(all), CLAIMS(all)
126	extractive	all-mpnet-base-v2	Summarized	2592	221	TITLE(all), ABSTRACT(all), CLAIMS(all)
127	extractive	bge-base-en-v1.5	Summarized	2592	176	TITLE(all), ABSTRACT(all), CLAIMS(all)
128	extractive	e5-base-v2	Summarized	2592	83	TITLE(all), ABSTRACT(all), CLAIMS(all)
129	extractive	google_bert_for_patents	Summarized	2592	113	TITLE(all), ABSTRACT(all), CLAIMS(all)
130	extractive	PatentSBERTa_V2	Summarized	2592	231	TITLE(all), ABSTRACT(all), CLAIMS(all)
131	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	229	TITLE(all), ABSTRACT(all), CLAIMS(all)
132	instructive	Mistral_7B_Instruct	Summarized	2592	260	TITLE(all), ABSTRACT(all), CLAIMS(all)
133	instructive	Llama3_8B_Instruct	Summarized	2592	264	TITLE(all), ABSTRACT(all), CLAIMS(all)

134	instructive	Qwen2_7B_Instruct	Summarized	2592	263	TITLE(all), ABSTRACT(all), CLAIMS(all)
135	instructive	Qwen2_14B_Instruct	Summarized	2592	265	TITLE(all), ABSTRACT(all), CLAIMS(all)
136	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	242	TITLE(all), ABSTRACT(all), CLAIMS(all)
137	abstractive	LED_Base_BigPatent	Summarized	2592	162	CLAIMS(all)
138	abstractive	PEGASUS_BigPatent	Summarized	2592	164	CLAIMS(all)
139	abstractive	LongT5_TGlobal_Base	Summarized	2592	140	CLAIMS(all)
140	abstractive	BART_large_CNN	Summarized	2592	77	CLAIMS(all)
141	abstractive	BigBird_Pegasus	Summarized	2592	97	CLAIMS(all)
142	abstractive	T5_Small_HUPD	Summarized	2592	106	CLAIMS(all)
143	extractive	all-mpnet-base-v2	Summarized	2592	126	CLAIMS(all)
144	extractive	bge-base-en-v1.5	Summarized	2592	98	CLAIMS(all)
145	extractive	e5-base-v2	Summarized	2592	38	CLAIMS(all)
146	extractive	google_bert_for_patents	Summarized	2592	53	CLAIMS(all)
147	extractive	PatentSBERTa_V2	Summarized	2592	133	CLAIMS(all)
148	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	132	CLAIMS(all)
149	instructive	Mistral_7B_Instruct	Summarized	2592	156	CLAIMS(all)
150	instructive	Llama3_8B_Instruct	Summarized	2592	159	CLAIMS(all)
151	instructive	Qwen2_7B_Instruct	Summarized	2592	158	CLAIMS(all)
152	instructive	Qwen2_14B_Instruct	Summarized	2592	161	CLAIMS(all)
153	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	141	CLAIMS(all)
154	abstractive	LED_Base_BigPatent	Summarized	2592	77	TITLE(all), ABSTRACT(all)

155	abstractive	PEGASUS_BigPatent	Summarized	2592	69	TITLE(all), ABSTRACT(all)
156	abstractive	LongT5_TGlobal_Base	Summarized	2592	105	TITLE(all), ABSTRACT(all)
157	abstractive	BART_large_CNN	Summarized	2592	105	TITLE(all), ABSTRACT(all)
158	abstractive	BigBird_Pegasus	Summarized	2592	105	TITLE(all), ABSTRACT(all)
159	abstractive	T5_Small_HUPD	Summarized	2592	105	TITLE(all), ABSTRACT(all)
160	extractive	all-mpnet-base-v2	Summarized	2592	97	TITLE(all), ABSTRACT(all)
161	extractive	bge-base-en-v1.5	Summarized	2592	84	TITLE(all), ABSTRACT(all)
162	extractive	e5-base-v2	Summarized	2592	48	TITLE(all), ABSTRACT(all)
163	extractive	google_bert_for_patents	Summarized	2592	62	TITLE(all), ABSTRACT(all)
164	extractive	PatentSBERTa_V2	Summarized	2592	99	TITLE(all), ABSTRACT(all)
165	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	98	TITLE(all), ABSTRACT(all)
166	instructive	Mistral_7B_Instruct	Summarized	2592	105	TITLE(all), ABSTRACT(all)
167	instructive	Llama3_8B_Instruct	Summarized	2592	105	TITLE(all), ABSTRACT(all)
168	instructive	Qwen2_7B_Instruct	Summarized	2592	105	TITLE(all), ABSTRACT(all)
169	instructive	Qwen2_14B_Instruct	Summarized	2592	105	TITLE(all), ABSTRACT(all)

170	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	105	TITLE(all), ABSTRACT(all)
171	ORIGINAL		Source Corpus	2592	422	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
172	ORIGINAL		Source Corpus	2592	449	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
173	ORIGINAL		Source Corpus	2592	228	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
174	ORIGINAL		Source Corpus	2592	436	TITLE(all), ABSTRACT(all), CLAIMS(all)
175	ORIGINAL		Source Corpus	2592	448	DESCRIPTION(all)
176	ORIGINAL		Source Corpus	2592	424	CLAIMS(all)
177	ORIGINAL		Source Corpus	2592	124	FIRST-CLAIM(all)
178	ORIGINAL		Source Corpus	2592	107	TITLE(all), ABSTRACT(all)
179	ORIGINAL		Source Corpus	2592	18383	ALL TAGS & WORDS
180	abstractive	LED_Base_BigPatent	Summarized	2592	406	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
181	abstractive	PEGASUS_BigPatent	Summarized	2592	410	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)

182	abstractive	LongT5_TGlobal_Base	Summarized	2592	400	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
183	abstractive	BART_large_CNN	Summarized	2592	189	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
184	abstractive	BigBird_Pegasus	Summarized	2592	244	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
185	abstractive	T5_Small_HUPD	Summarized	2592	242	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
186	extractive	all-mpnet-base-v2	Summarized	2592	424	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
187	extractive	bge-base-en-v1.5	Summarized	2592	404	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
188	extractive	e5-base-v2	Summarized	2592	202	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)

189	extractive	google_bert_for_patents	Summarized	2592	311	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
190	extractive	PatentSBERTa_V2	Summarized	2592	429	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
191	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	426	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
192	instructive	Mistral_7B_Instruct	Summarized	2592	368	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
193	instructive	Llama3_8B_Instruct	Summarized	2592	368	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
194	instructive	Qwen2_7B_Instruct	Summarized	2592	373	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
195	instructive	Qwen2_14B_Instruct	Summarized	2592	373	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)

196	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	445	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
197	ORIGINAL		Source Corpus	2592	449	TITLE(all), ABSTRACT(100), DETAILED-DESCRIPTION(200), BRIEF-DESCRIPTION(200)
198	ORIGINAL		Source Corpus	2592	450	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
199	ORIGINAL		Source Corpus	2592	450	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRSTCLAIM(all), CLAIMS(all)
200	abstractive	LED_Base_BigPatent	Summarized	2592	337	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
201	abstractive	PEGASUS_BigPatent	Summarized	2592	343	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
202	abstractive	LongT5_TGlobal_Base	Summarized	2592	363	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
203	abstractive	BART_large_CNN	Summarized	2592	99	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
204	abstractive	BigBird_Pegasus	Summarized	2592	154	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)

205	abstractive	T5_Small_HUPD	Summarized	2592	151	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
206	extractive	all-mpnet-base-v2	Summarized	2592	405	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
207	extractive	bge-base-en-v1.5	Summarized	2592	378	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
208	extractive	e5-base-v2	Summarized	2592	158	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
209	extractive	google_bert_for_patents	Summarized	2592	266	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
210	extractive	PatentSBERTa_V2	Summarized	2592	411	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
211	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	407	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
212	instructive	Mistral_7B_Instruct	Summarized	2592	278	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
213	instructive	Llama3_8B_Instruct	Summarized	2592	279	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
214	instructive	Qwen2_7B_Instruct	Summarized	2592	283	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)
215	instructive	Qwen2_14B_Instruct	Summarized	2592	283	DETAILED-DESCRIP-TION(all), BRIEF-DE-SCRIPTION(all)

216	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	443	DETAILED-DESCRIPTION(all), BRIEF-DESCRIPTION(all)
217	hybrid	PatentSBERTa_V2_Mistral_7B	Summarized	2592	318	TITLE(all), ABSTRACT(all), DESCRIPTION(all), FIRST-CLAIM(all), CLAIMS(all)
218	extractive	PatentSBERTa_V2	Summarized	2592	256	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
219	ORIGINAL		Source Corpus	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
220	extractive	PatentSBERTa_V2	Summarized	2592	128	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
221	extractive	PatentSBERTa_V2	Summarized	2592	256	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
222	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
223	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	64	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
224	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
225	ORIGINAL		Source Corpus	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
226	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100),

						DESCRIPTION(250), FIRST-CLAIM(all)
227	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
228	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
229	extractive	PatentSBERTa_V2	Summarized	2592	512	DESCRIPTION(all), TITLE(all), ABSTRACT(all)
230	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(200), FIRST-CLAIM(all)
231	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), CLAIMS(200)
232	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), FIRST-CLAIM(200), DESCRIPTION(200)
233	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(all), CLAIMS(400)
234	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
235	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
236	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), CLAIMS(250), DESCRIPTION(250)

237	ORIGINAL		Source Corpus	2592	512	DESCRIPTION(all)
238	extractive	PatentSBERTa_V2	Summarized	2592	512	DESCRIPTION(all)
239	extractive	PatentSBERTa_V2	Summarized	2592	512	CLAIMS(all)
240	extractive	PatentSBERTa_V2	Summarized	2592	512	ABSTRACT(all)
241	extractive	PatentSBERTa_V2	Summarized	2592	512	FIRST-CLAIM(all)
242	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
243	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
244	extractive	all-mpnet-base-v2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
245	extractive	bge-base-en-v1.5	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
246	extractive	e5-base-v2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
247	extractive	google_bert_for_patents	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
248	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)

249	ORIGINAL		Source Corpus	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
250	extractive	PatentSBERTa_V2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
251	extractive	all-mpnet-base-v2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
252	extractive	bge-base-en-v1.5	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
253	extractive	e5-base-v2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
254	extractive	google_bert_for_patents	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
255	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
256	ORIGINAL		Source Corpus	2592	512	CLAIMS(220), DESCRIPTION(220)
257	extractive	PatentSBERTa_V2	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
258	extractive	all-mpnet-base-v2	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
259	extractive	bge-base-en-v1.5	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
260	extractive	e5-base-v2	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
261	extractive	google_bert_for_patents	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)

262	extractive	SBERT_all-MiniLM-L6-v2	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
263	abstractive	LED_Base_BigPatent	Summarized	2592	512	DESCRIPTION(all)
264	abstractive	BART_large_CNN	Summarized	2592	512	DESCRIPTION(all)
265	abstractive	LED_Base_BigPatent	Summarized	2592	512	CLAIMS(all)
266	abstractive	LED_Base_BigPatent	Summarized	2592	512	ABSTRACT(all)
267	abstractive	LED_Base_BigPatent	Summarized	2592	512	FIRST-CLAIM(all)
268	abstractive	LED_Base_BigPatent	Summarized	2592	512	DESCRIPTION(250), ABSTRACT(100), FIRST-CLAIM(all)
269	abstractive	LED_Base_BigPatent	Summarized	2592	512	ABSTRACT(all), DETAILED-DESCRIPTION(220), BRIEF-DESCRIPTION(all)
270	abstractive	BART_large_CNN	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
271	abstractive	LED_Base_BigPatent	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
272	abstractive	BART_large_CNN	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
273	abstractive	BigBird_Pegasus	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
274	abstractive	LongT5_TGlobal_Base	Summarized	2592	512	TITLE(all), ABSTRACT(100),

						DESCRIPTION(250), FIRST-CLAIM(all)
275	abstractive	T5_Small_HUPD	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
276	abstractive	PEGASUS_BigPatent	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
277	abstractive	LED_Base_BigPatent	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
278	abstractive	T5_Small_HUPD	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
279	abstractive	PEGASUS_BigPatent	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
280	abstractive	BigBird_Pegasus	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
281	abstractive	BART_large_CNN	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
282	abstractive	LongT5_TGlobal_Base	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
283	abstractive	LED_Base_BigPatent	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
284	abstractive	T5_Small_HUPD	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
285	abstractive	PEGASUS_BigPatent	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)

286	abstractive	BigBird_Pegasus	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
287	abstractive	BART_large_CNN	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
288	abstractive	LongT5_TGlobal_Base	Summarized	2592	512	TITLE(all), ABSTRACT(all), FIRST-CLAIM(all)
289	instructive	Llama3_8B_Instruct	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
290	instructive	Mistral_7B_Instruct	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
291	instructive	Qwen2_14B_Instruct	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
292	instructive	Qwen2_7B_Instruct	Summarized	2592	512	DESCRIPTION(220), CLAIMS(220)
293	instructive	Llama3_8B_Instruct	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
294	instructive	Mistral_7B_Instruct	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
295	instructive	Qwen2_14B_Instruct	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
296	instructive	Qwen2_7B_Instruct	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)

297	instructive	Llama3_8B_Instruct	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
298	instructive	Mistral_7B_Instruct	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
299	instructive	Qwen2_14B_Instruct	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
300	instructive	Qwen2_7B_Instruct	Summarized	2592	512	CLAIMS(220), DESCRIPTION(220)
301	instructive	Mistral_7B_Instruct	Summarized	2592	512	DESCRIPTION(250), FIRST-CLAIM(all)
302	instructive	Llama3_8B_Instruct	Summarized	2592	512	DESCRIPTION(250), FIRST-CLAIM(all)
303	instructive	Qwen2_14B_Instruct	Summarized	2592	512	DESCRIPTION(250), FIRST-CLAIM(all)
304	instructive	Qwen2_7B_Instruct	Summarized	2592	512	DESCRIPTION(250), FIRST-CLAIM(all)
305	instructive	Mistral_7B_Instruct	Summarized	2592	512	DESCRIPTION(all)
306	instructive	Llama3_8B_Instruct	Summarized	2592	512	DESCRIPTION(all)
307	instructive	Qwen2_7B_Instruct	Summarized	2592	512	DESCRIPTION(all)
308	instructive	Qwen2_14B_Instruct	Summarized	2592	512	DESCRIPTION(all)
309	instructive	Llama3_8B_Instruct	Summarized	2592	512	FIRST-CLAIM(all), CLAIMS(all)
310	ORIGINAL		Source Corpus	2592	512	CLAIMS(all)
311	hybrid	PatentSBERTa_V2_TextRank	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
312	abstractive	BART_large_CNN	Summarized	923	512	TITLE(all), ABSTRACT(100),

						DESCRIPTION(250), FIRST-CLAIM(all)
313	abstractive	T5_Small_HUPD	Summarized	2592	512	TITLE(all), ABSTRACT(all), CLAIMS(400)
314	abstractive	BigBird_Pegasus	Summarized	2592	512	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
315	abstractive	T5_Small_HUPD	Summarized	2592	512	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
316	abstractive	BART_large_CNN	Summarized	2592	512	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
317	abstractive	PEGASUS_BigPatent	Summarized	2592	512	TITLE(all), ABSTRACT(all), DESCRIPTION(all)
318	abstractive	BART_large_CNN	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)
319	hybrid	PatentSBERTa_V2_Mistral_7B	Summarized	2592	512	TITLE(all), ABSTRACT(100), DESCRIPTION(250), FIRST-CLAIM(all)

B.3 Precision/Recall metrics

id	p_at_10	p_at_50	p_at_100	recall_at_10	recall_at_50	recall_at_100
1	0.1383	0.0362	0.0198	0.5527	0.7019	0.7641
2	0.1442	0.0371	0.0201	0.5732	0.7164	0.7756
3	0.1455	0.037	0.0202	0.5794	0.7206	0.7817
4	0.1497	0.0379	0.0205	0.5944	0.7356	0.7912
5	0.1489	0.0379	0.0204	0.5915	0.7354	0.7882
6	0.1508	0.0382	0.0206	0.5988	0.741	0.7967
7	0.1502	0.0378	0.0204	0.5967	0.734	0.7903
8	0.1494	0.0379	0.0205	0.5942	0.7357	0.7912
9	0.1465	0.0374	0.0201	0.5847	0.7262	0.7801
10	0.1476	0.0375	0.0202	0.588	0.7274	0.78
11	0.1499	0.0382	0.0205	0.5963	0.7412	0.7944
12	0.1503	0.038	0.0205	0.5973	0.7393	0.7915
13	0.1531	0.0388	0.0209	0.6097	0.7539	0.8077
14	0.1525	0.0387	0.0208	0.6081	0.7496	0.8054
15	0.1554	0.0391	0.0211	0.6178	0.759	0.8169
16	0.1544	0.0392	0.021	0.614	0.7586	0.8109
17	0.149	0.0383	0.0206	0.5933	0.7421	0.7965
18	0.1351	0.0355	0.0195	0.5384	0.6868	0.7522
19	0.1433	0.0369	0.0201	0.5698	0.7141	0.7747
20	0.1421	0.0364	0.0198	0.566	0.7073	0.7684
21	0.1473	0.0378	0.0204	0.585	0.7322	0.7889

22	0.1478	0.0377	0.0204	0.5877	0.7306	0.7894
23	0.147	0.0379	0.0204	0.5839	0.7331	0.7885
24	0.1478	0.0375	0.0202	0.586	0.7292	0.781
25	0.1471	0.0375	0.0203	0.5837	0.7269	0.7824
26	0.1404	0.0358	0.0194	0.5593	0.6943	0.7503
27	0.1415	0.0358	0.0195	0.5631	0.6968	0.756
28	0.1493	0.0378	0.0203	0.5929	0.7343	0.7871
29	0.1488	0.0377	0.0203	0.5909	0.7323	0.7856
30	0.1534	0.0391	0.021	0.6107	0.7579	0.8119
31	0.1501	0.0386	0.0207	0.5985	0.7481	0.8019
32	0.1537	0.0393	0.0212	0.6127	0.7649	0.82
33	0.1534	0.0392	0.021	0.6104	0.7602	0.8129
34	0.149	0.0382	0.0205	0.5918	0.7408	0.7942
35	0.1356	0.0355	0.0194	0.542	0.6895	0.7501
36	0.1421	0.0369	0.0201	0.5653	0.7125	0.7738
37	0.1377	0.0356	0.0194	0.5488	0.6921	0.7499
38	0.1447	0.0371	0.0201	0.5759	0.721	0.7775
39	0.1452	0.0374	0.0202	0.5793	0.7268	0.7834
40	0.1456	0.0375	0.0203	0.5781	0.7269	0.7824
41	0.1467	0.0372	0.02	0.5827	0.7238	0.7757
42	0.1455	0.0372	0.0201	0.5781	0.7216	0.7763
43	0.1424	0.0362	0.0195	0.5676	0.703	0.7532
44	0.1404	0.0355	0.0193	0.5596	0.6909	0.7496
45	0.1471	0.0375	0.0202	0.586	0.7291	0.7822
46	0.1479	0.0373	0.0202	0.5876	0.7251	0.78

47	0.1535	0.0393	0.0212	0.6132	0.7631	0.8186
48	0.1502	0.0387	0.0208	0.5981	0.7504	0.8053
49	0.1527	0.0395	0.0213	0.6091	0.7678	0.8241
50	0.152	0.0393	0.0212	0.6059	0.7624	0.8213
51	0.1481	0.0378	0.0204	0.5882	0.7354	0.7881
52	0.129	0.0343	0.0189	0.5165	0.6664	0.7311
53	0.1396	0.0358	0.0196	0.5549	0.6947	0.759
54	0.1477	0.0375	0.0203	0.5874	0.7272	0.7826
55	0.1485	0.0377	0.0202	0.5904	0.7296	0.7812
56	0.1481	0.0377	0.0203	0.5888	0.7304	0.7837
57	0.1486	0.0377	0.0203	0.5909	0.7302	0.7832
58	0.1476	0.0375	0.0202	0.5866	0.7263	0.7802
59	0.1476	0.0374	0.0201	0.5872	0.7259	0.7774
60	0.1445	0.0368	0.0198	0.5749	0.7129	0.7668
61	0.1465	0.0371	0.02	0.5829	0.7212	0.7731
62	0.1475	0.0374	0.0202	0.5863	0.7263	0.7803
63	0.1478	0.0375	0.0202	0.5876	0.7276	0.7804
64	0.1488	0.0377	0.0203	0.5913	0.7301	0.7839
65	0.148	0.0378	0.0203	0.5883	0.7316	0.7837
66	0.1487	0.0378	0.0203	0.5914	0.7317	0.7855
67	0.1486	0.0378	0.0204	0.5904	0.7324	0.7859
68	0.148	0.0374	0.0202	0.5881	0.7255	0.7803
69	0.1224	0.0325	0.0181	0.4869	0.6329	0.7012
70	0.1347	0.0351	0.0192	0.5356	0.6778	0.7413
71	0.1242	0.0322	0.0178	0.4977	0.6285	0.6889

72	0.1004	0.0261	0.0144	0.3991	0.5063	0.5562
73	0.1279	0.0336	0.0185	0.5116	0.6542	0.7161
74	0.111	0.0288	0.0159	0.444	0.5624	0.6196
75	0.1374	0.0353	0.019	0.5466	0.6861	0.7368
76	0.1358	0.0347	0.0187	0.5391	0.6738	0.7235
77	0.1174	0.03	0.0163	0.4685	0.5832	0.631
78	0.1185	0.0307	0.0168	0.4736	0.5994	0.6517
79	0.1404	0.0355	0.0193	0.5592	0.6906	0.748
80	0.1393	0.0352	0.0191	0.5543	0.684	0.7417
81	0.142	0.0373	0.0203	0.5652	0.7214	0.7822
82	0.1305	0.0346	0.0188	0.5179	0.67	0.7263
83	0.1393	0.0368	0.0201	0.5573	0.7156	0.7784
84	0.1296	0.0354	0.0194	0.5179	0.6886	0.7511
85	0.1449	0.0369	0.02	0.5759	0.7171	0.7752
86	0.1112	0.0303	0.017	0.4454	0.5887	0.6572
87	0.1267	0.0333	0.0184	0.5043	0.6465	0.71
88	0.1361	0.0352	0.0192	0.5399	0.6813	0.7429
89	0.1382	0.0357	0.0194	0.5484	0.6918	0.7518
90	0.1353	0.035	0.0191	0.5377	0.6804	0.7382
91	0.1394	0.0358	0.0195	0.5523	0.6947	0.754
92	0.1384	0.0358	0.0194	0.5489	0.6945	0.7525
93	0.139	0.0358	0.0194	0.5508	0.6929	0.7518
94	0.136	0.0351	0.0192	0.541	0.682	0.7421
95	0.1366	0.0352	0.0192	0.5414	0.6835	0.7413
96	0.139	0.0358	0.0194	0.5512	0.6937	0.7522

97	0.1382	0.0358	0.0195	0.5478	0.6933	0.7526
98	0.1391	0.0358	0.0195	0.5511	0.6937	0.7545
99	0.1382	0.0356	0.0194	0.5478	0.689	0.7488
100	0.1384	0.0358	0.0195	0.5489	0.694	0.7542
101	0.136	0.0355	0.0194	0.5389	0.6885	0.7477
102	0.1398	0.036	0.0195	0.5544	0.6985	0.7552
103	0.1413	0.0364	0.02	0.564	0.7062	0.7724
104	0.1442	0.037	0.0201	0.5728	0.7162	0.7762
105	0.1449	0.0371	0.02	0.5773	0.7231	0.7753
106	0.1489	0.0379	0.0205	0.5929	0.736	0.7932
107	0.149	0.0379	0.0204	0.5932	0.7341	0.7878
108	0.1513	0.0384	0.0207	0.6022	0.7434	0.7979
109	0.1498	0.038	0.0205	0.5961	0.7384	0.7929
110	0.1499	0.0381	0.0205	0.5955	0.7399	0.794
111	0.1485	0.0376	0.0202	0.5927	0.7287	0.7808
112	0.1492	0.0376	0.0203	0.5944	0.7298	0.785
113	0.1513	0.0382	0.0205	0.6015	0.7435	0.7937
114	0.1504	0.0383	0.0206	0.5979	0.7434	0.7959
115	0.1545	0.0391	0.021	0.6148	0.7605	0.8115
116	0.1534	0.0391	0.021	0.6107	0.7595	0.8145
117	0.1553	0.0395	0.0212	0.6181	0.7687	0.8198
118	0.1554	0.0393	0.0213	0.6183	0.7636	0.8218
119	0.1489	0.0382	0.0205	0.5914	0.741	0.7945
120	0.1348	0.0354	0.0193	0.5398	0.6843	0.7448
121	0.1403	0.0361	0.0198	0.5578	0.6993	0.7649

122	0.1468	0.0375	0.0203	0.5833	0.7252	0.7824
123	0.1475	0.0376	0.0204	0.5879	0.7298	0.7875
124	0.1482	0.0376	0.0203	0.5897	0.7278	0.7858
125	0.1475	0.0379	0.0204	0.5876	0.735	0.7869
126	0.1459	0.0375	0.0202	0.5826	0.7253	0.782
127	0.1466	0.0373	0.0201	0.5842	0.7267	0.7788
128	0.1429	0.0365	0.0197	0.5725	0.7073	0.7613
129	0.145	0.0369	0.02	0.5796	0.7174	0.7748
130	0.1458	0.0374	0.0202	0.582	0.7252	0.7822
131	0.1472	0.0375	0.0202	0.5855	0.7275	0.7828
132	0.1511	0.0388	0.0209	0.6038	0.753	0.8098
133	0.1528	0.039	0.0209	0.6094	0.757	0.8106
134	0.1532	0.039	0.021	0.6103	0.7579	0.8129
135	0.1539	0.0391	0.0212	0.6138	0.7593	0.8188
136	0.1461	0.0376	0.0202	0.5814	0.7297	0.7819
137	0.1197	0.0325	0.018	0.4797	0.63	0.6927
138	0.1284	0.0339	0.0187	0.5118	0.6582	0.722
139	0.1318	0.0343	0.0187	0.5244	0.6668	0.7221
140	0.1294	0.0338	0.0184	0.5186	0.6607	0.7177
141	0.1326	0.0345	0.0188	0.5305	0.6694	0.7265
142	0.1383	0.0358	0.0194	0.5512	0.695	0.7502
143	0.1315	0.0341	0.0186	0.5261	0.6622	0.7194
144	0.1288	0.0337	0.0184	0.514	0.6555	0.7112
145	0.114	0.0307	0.0169	0.4573	0.5947	0.6529
146	0.1243	0.0321	0.0176	0.4956	0.6265	0.6811

147	0.132	0.0343	0.0187	0.5288	0.6664	0.7227
148	0.1321	0.0343	0.0187	0.5284	0.6673	0.7241
149	0.1436	0.0375	0.0204	0.5762	0.7315	0.791
150	0.1419	0.0371	0.0203	0.5693	0.7222	0.7856
151	0.1398	0.0367	0.0201	0.5604	0.7143	0.7813
152	0.1339	0.0357	0.0197	0.5386	0.6972	0.7623
153	0.1314	0.0341	0.0186	0.5251	0.6643	0.7214
154	0.1251	0.033	0.0181	0.5031	0.6418	0.7026
155	0.1372	0.0354	0.0193	0.5464	0.6866	0.7474
156	0.1463	0.0375	0.0202	0.5813	0.7244	0.7819
157	0.1461	0.0375	0.0202	0.5807	0.724	0.7817
158	0.1461	0.0375	0.0202	0.5809	0.7246	0.782
159	0.1462	0.0375	0.0202	0.5809	0.724	0.782
160	0.1454	0.0374	0.0202	0.5773	0.7222	0.7801
161	0.1447	0.037	0.02	0.5755	0.7173	0.7724
162	0.1382	0.0353	0.0192	0.5497	0.6863	0.7447
163	0.1408	0.036	0.0196	0.5608	0.6998	0.7581
164	0.1455	0.0373	0.0202	0.5781	0.7215	0.7809
165	0.146	0.0373	0.0202	0.5802	0.7203	0.7805
166	0.1464	0.0375	0.0203	0.5814	0.7244	0.7824
167	0.1463	0.0375	0.0203	0.5808	0.7239	0.7822
168	0.1463	0.0375	0.0203	0.5809	0.7242	0.7825
169	0.1463	0.0375	0.0202	0.581	0.7242	0.7819
170	0.1456	0.0373	0.0202	0.5781	0.7218	0.7821
171	0.1495	0.0383	0.0207	0.594	0.7426	0.7985

172	0.147	0.0378	0.0204	0.5842	0.731	0.7887
173	0.1475	0.0375	0.0202	0.586	0.7269	0.7799
174	0.147	0.0375	0.0201	0.5841	0.7261	0.7782
175	0.141	0.0361	0.0197	0.5606	0.7001	0.7588
176	0.1417	0.0365	0.0197	0.5642	0.7064	0.763
177	0.1246	0.0323	0.0176	0.4941	0.6266	0.6814
178	0.1467	0.0375	0.0202	0.5824	0.725	0.7816
179	0.1519	0.0383	0.0206	0.6033	0.7434	0.7974
180	0.1319	0.0349	0.0191	0.5271	0.6752	0.7388
181	0.1404	0.0365	0.0199	0.5583	0.7034	0.7688
182	0.1391	0.0359	0.0195	0.5536	0.6957	0.7521
183	0.1416	0.0361	0.0197	0.5619	0.6994	0.762
184	0.1428	0.0369	0.0201	0.5684	0.7162	0.7775
185	0.1404	0.0362	0.0197	0.5572	0.699	0.7611
186	0.1454	0.037	0.02	0.577	0.7173	0.7733
187	0.1438	0.0368	0.0199	0.5706	0.7132	0.7684
188	0.1351	0.0345	0.0187	0.5385	0.6695	0.7228
189	0.1364	0.0346	0.0189	0.5423	0.6729	0.7305
190	0.1468	0.0373	0.0201	0.583	0.7244	0.7763
191	0.1459	0.0372	0.0201	0.5812	0.722	0.7779
192	0.1505	0.0389	0.021	0.6011	0.7561	0.8105
193	0.1465	0.0379	0.0205	0.5817	0.7354	0.7943
194	0.1499	0.0391	0.0212	0.5987	0.7613	0.8198
195	0.1495	0.0388	0.0209	0.5948	0.7527	0.8093
196	0.146	0.0373	0.0202	0.5811	0.7209	0.7785

197	0.138	0.0355	0.0195	0.5482	0.6875	0.752
198	0.141	0.0361	0.0197	0.5606	0.7004	0.7585
199	0.147	0.0378	0.0204	0.5842	0.7311	0.7887
200	0.1223	0.0325	0.0181	0.4867	0.633	0.7016
201	0.1346	0.0351	0.0192	0.5355	0.6778	0.7413
202	0.1255	0.0325	0.0179	0.5024	0.6325	0.6916
203	0.1004	0.0261	0.0143	0.3991	0.5066	0.5554
204	0.1279	0.0336	0.0185	0.5116	0.6545	0.7155
205	0.111	0.0288	0.0159	0.444	0.5627	0.6191
206	0.1379	0.0354	0.0191	0.5494	0.6884	0.7396
207	0.1362	0.0349	0.0189	0.5408	0.6767	0.7274
208	0.1174	0.03	0.0163	0.4682	0.5834	0.6311
209	0.1185	0.0309	0.0168	0.4744	0.6017	0.6536
210	0.1402	0.0358	0.0194	0.5575	0.6946	0.7511
211	0.1398	0.0353	0.0192	0.5575	0.6878	0.7448
212	0.142	0.0373	0.0202	0.5652	0.7217	0.7816
213	0.1305	0.0346	0.0188	0.5181	0.6702	0.7257
214	0.1393	0.0368	0.0201	0.5573	0.7159	0.7778
215	0.1296	0.0354	0.0194	0.5179	0.6889	0.7505
216	0.1451	0.037	0.02	0.5767	0.7184	0.7743
217	0.1545	0.0392	0.0211	0.6166	0.7621	0.8172
218	0.1029	0.0278	0.0159	0.4162	0.5495	0.6235
219	0.1181	0.0314	0.0174	0.4719	0.6143	0.6791
220	0.1115	0.0295	0.0168	0.4497	0.5795	0.6559
221	0.1139	0.0306	0.0173	0.4573	0.6024	0.6748

222	0.118	0.0314	0.0174	0.4723	0.6141	0.6804
223	0.1105	0.0288	0.0162	0.4465	0.5673	0.6339
224	0.118	0.0314	0.0174	0.4718	0.6136	0.6803
225	0.1159	0.0304	0.0171	0.4672	0.5955	0.6671
226	0.114	0.0306	0.0173	0.4577	0.6005	0.6763
227	0.1157	0.0315	0.0178	0.4637	0.6169	0.6937
228	0.1141	0.0312	0.0175	0.4595	0.6142	0.6846
229	0.109	0.0297	0.0168	0.4397	0.5858	0.6576
230	0.1176	0.0317	0.0176	0.4714	0.6228	0.687
231	0.1132	0.0308	0.0173	0.4573	0.6044	0.6761
232	0.118	0.0314	0.0176	0.4735	0.6161	0.6885
233	0.1209	0.0317	0.0174	0.4862	0.6214	0.6816
234	0.1222	0.032	0.0176	0.4905	0.6287	0.687
235	0.1223	0.032	0.0176	0.491	0.6288	0.6867
236	0.1119	0.0303	0.0172	0.4479	0.5944	0.6711
237	0.0964	0.0267	0.0153	0.3897	0.5227	0.5974
238	0.1009	0.0278	0.0157	0.4069	0.5487	0.6145
239	0.0959	0.0259	0.0146	0.3862	0.5091	0.5728
240	0.1155	0.0296	0.0164	0.4658	0.5865	0.6452
241	0.1092	0.029	0.0161	0.4357	0.568	0.6304
242	0.1097	0.0305	0.0174	0.4408	0.5987	0.6792
243	0.1141	0.0312	0.0176	0.4594	0.6146	0.6848
244	0.1141	0.0312	0.0175	0.4591	0.6125	0.6849
245	0.1152	0.0311	0.0174	0.4641	0.6118	0.6797
246	0.1168	0.0315	0.0175	0.4689	0.619	0.6835

247	0.1154	0.0312	0.0174	0.4632	0.612	0.6779
248	0.1164	0.0312	0.0175	0.4679	0.6119	0.6818
249	0.1234	0.032	0.0176	0.4949	0.6277	0.6886
250	0.123	0.0318	0.0175	0.4931	0.6234	0.6845
251	0.1228	0.0318	0.0176	0.492	0.6245	0.6868
252	0.1222	0.0318	0.0175	0.4889	0.6229	0.6833
253	0.1175	0.031	0.017	0.4732	0.6093	0.6675
254	0.1213	0.0314	0.0173	0.486	0.6162	0.6759
255	0.1225	0.0317	0.0175	0.4913	0.623	0.6848
256	0.1049	0.029	0.0166	0.421	0.567	0.646
257	0.1035	0.0291	0.0165	0.4168	0.5691	0.6455
258	0.1032	0.0292	0.0164	0.4166	0.5714	0.6397
259	0.1047	0.0293	0.0164	0.4223	0.5744	0.6426
260	0.0894	0.0249	0.0141	0.3606	0.4898	0.554
261	0.0957	0.0268	0.015	0.3852	0.5274	0.5879
262	0.1037	0.029	0.0165	0.4188	0.5689	0.6415
263	0.0925	0.0261	0.0149	0.3728	0.5133	0.5848
264	0.0788	0.0211	0.0118	0.3171	0.4157	0.4642
265	0.0888	0.0256	0.0145	0.3589	0.5035	0.5664
266	0.0792	0.0228	0.013	0.3241	0.4559	0.5159
267	0.083	0.0239	0.0136	0.3325	0.467	0.5319
268	0.0994	0.0284	0.0162	0.4021	0.5597	0.634
269	0.0862	0.0262	0.0154	0.3499	0.5195	0.6023
270	0.1262	0.0333	0.0183	0.5067	0.6496	0.7121
271	0.1072	0.0296	0.0167	0.4335	0.5838	0.6543

272	0.1253	0.033	0.0181	0.5034	0.6444	0.7037
273	0.1242	0.0325	0.0178	0.4983	0.6366	0.6956
274	0.1121	0.0304	0.017	0.4524	0.5971	0.6682
275	0.1258	0.0325	0.018	0.5053	0.6364	0.7016
276	0.1154	0.0311	0.0173	0.4632	0.6091	0.6754
277	0.0979	0.0282	0.0161	0.3957	0.554	0.6304
278	0.1186	0.0313	0.0174	0.4744	0.6133	0.6779
279	0.103	0.0292	0.0166	0.4151	0.5728	0.6474
280	0.1152	0.0309	0.0171	0.4631	0.605	0.6681
281	0.1149	0.0303	0.0167	0.4622	0.5938	0.6518
282	0.0953	0.0268	0.0155	0.3852	0.5301	0.6087
283	0.103	0.0282	0.016	0.4206	0.5608	0.6308
284	0.1238	0.032	0.0176	0.4968	0.6279	0.6896
285	0.1161	0.0303	0.0168	0.4638	0.5948	0.6556
286	0.1233	0.0319	0.0176	0.495	0.6267	0.6886
287	0.1242	0.0321	0.0176	0.4985	0.6292	0.6889
288	0.1231	0.0319	0.0176	0.4935	0.6256	0.6872
289	0.0985	0.0281	0.0164	0.3967	0.5544	0.6391
290	0.1141	0.0312	0.0174	0.4584	0.6118	0.6808
291	0.0843	0.0259	0.0152	0.3424	0.5143	0.5997
292	0.0953	0.0281	0.0161	0.3866	0.555	0.6318
293	0.1165	0.0317	0.0178	0.4697	0.6233	0.695
294	0.1197	0.0322	0.0179	0.4819	0.6326	0.6979
295	0.1138	0.0312	0.0175	0.4609	0.6155	0.6852
296	0.1159	0.0316	0.0179	0.4681	0.6203	0.6977

297	0.0995	0.0283	0.0163	0.4017	0.557	0.6384
298	0.1129	0.031	0.0174	0.4543	0.6052	0.6769
299	0.086	0.026	0.0152	0.3483	0.5127	0.599
300	0.0963	0.0282	0.0162	0.3896	0.5572	0.6349
301	0.1154	0.031	0.0172	0.4607	0.606	0.6694
302	0.1064	0.0294	0.0166	0.4264	0.5752	0.6491
303	0.0991	0.0286	0.0162	0.4024	0.5645	0.6394
304	0.1071	0.03	0.0168	0.4311	0.59	0.6571
305	0.1056	0.0289	0.0161	0.425	0.5651	0.6274
306	0.0859	0.0246	0.014	0.3464	0.4836	0.5516
307	0.0879	0.0256	0.0147	0.3572	0.5054	0.5794
308	0.0749	0.0231	0.0136	0.3068	0.4608	0.5382
309	0.1082	0.0303	0.017	0.4344	0.593	0.6642
310	0.095	0.0273	0.0157	0.3829	0.5356	0.6143
311	0.1147	0.0309	0.0176	0.4603	0.6071	0.6889
312	0.1232	0.0365	0.0213	0.3509	0.5004	0.5822
313	0.1233	0.0319	0.0175	0.4957	0.6249	0.6848
314	0.1245	0.0322	0.0177	0.5006	0.6323	0.6902
315	0.1263	0.0324	0.0179	0.5063	0.6365	0.6968
316	0.126	0.0323	0.0177	0.5072	0.6356	0.6911
317	0.119	0.0314	0.0173	0.4772	0.6169	0.6772
318	0.1476	0.038	0.0204	0.5873	0.7396	0.7916
319	0.1242	0.0324	0.0179	0.4996	0.6361	0.698

B.4 nDCG & PRES metrics

id	ndgg_at_10	ndgg_at_50	ndgg_at_100	press_at_10	press_at_50	press_at_100
1	0.4939	0.5386	0.5521	8.6172	48.192	98.0229
2	0.52	0.5631	0.5758	8.5581	48.1469	97.9922
3	0.5223	0.5646	0.578	8.5453	48.1485	97.9778
4	0.5356	0.5781	0.5903	8.5029	48.1057	97.9514
5	0.5335	0.5765	0.588	8.5107	48.1042	97.9592
6	0.5375	0.5803	0.5923	8.492	48.0894	97.9401
7	0.5374	0.5786	0.5907	8.4983	48.1088	97.9576
8	0.5357	0.5784	0.5904	8.5056	48.1034	97.9526
9	0.5304	0.573	0.5847	8.5348	48.1302	97.9852
10	0.5331	0.5751	0.5865	8.5239	48.1271	97.9841
11	0.5382	0.5818	0.5934	8.501	48.0921	97.9468
12	0.5387	0.5812	0.5926	8.4975	48.0979	97.9549
13	0.5515	0.5951	0.6068	8.4691	48.0575	97.9114
14	0.5465	0.5894	0.6014	8.4745	48.0657	97.9199
15	0.5525	0.5951	0.6077	8.4458	48.0439	97.8888
16	0.5459	0.5895	0.6009	8.4559	48.0424	97.9021
17	0.5339	0.5789	0.5907	8.5103	48.0863	97.9394
18	0.4845	0.5291	0.5433	8.6487	48.225	98.0525
19	0.5174	0.5606	0.5737	8.5667	48.1535	97.993
20	0.5124	0.5548	0.568	8.5787	48.1788	98.0163
21	0.529	0.5731	0.5853	8.527	48.1088	97.96

22	0.529	0.572	0.5846	8.5223	48.1174	97.9627
23	0.5247	0.5695	0.5816	8.5301	48.1069	97.9588
24	0.5311	0.5739	0.5851	8.522	48.1244	97.9841
25	0.5262	0.5689	0.581	8.529	48.1255	97.9736
26	0.504	0.5443	0.5566	8.5962	48.2114	98.0591
27	0.511	0.5509	0.5637	8.5853	48.2099	98.0513
28	0.5362	0.5786	0.59	8.5072	48.11	97.9674
29	0.5332	0.5755	0.5872	8.5119	48.1147	97.967
30	0.5496	0.5937	0.6054	8.466	48.047	97.9021
31	0.5387	0.5842	0.5958	8.4986	48.07	97.93
32	0.5467	0.5924	0.6044	8.4633	48.0326	97.8842
33	0.5427	0.5879	0.5993	8.4664	48.0412	97.9013
34	0.5344	0.5789	0.5905	8.5099	48.0906	97.9475
35	0.4837	0.5276	0.5408	8.6436	48.2231	98.0595
36	0.512	0.5558	0.5691	8.5787	48.1566	97.9949
37	0.4967	0.5397	0.5522	8.6226	48.22	98.0634
38	0.523	0.5663	0.5786	8.5531	48.1438	97.9914
39	0.5241	0.5685	0.5809	8.5484	48.131	97.9771
40	0.521	0.5654	0.5774	8.5441	48.1244	97.9736
41	0.5247	0.5667	0.578	8.5332	48.1387	97.9969
42	0.5212	0.564	0.576	8.5449	48.1391	97.9876
43	0.5099	0.5501	0.5612	8.576	48.1924	98.049
44	0.5087	0.5485	0.5612	8.5958	48.2243	98.0668
45	0.5291	0.5721	0.5836	8.5286	48.1228	97.9782
46	0.5289	0.5701	0.5821	8.5208	48.1337	97.9825

47	0.5514	0.5963	0.6084	8.4648	48.0358	97.885
48	0.5337	0.5795	0.5914	8.4979	48.0657	97.918
49	0.5436	0.5912	0.6036	8.4726	48.026	97.8714
50	0.5365	0.5837	0.5964	8.48	48.0369	97.8807
51	0.5336	0.5775	0.5891	8.5188	48.108	97.9627
52	0.4651	0.5101	0.5242	8.7097	48.2857	98.1112
53	0.5034	0.5453	0.5592	8.604	48.2075	98.0373
54	0.5289	0.571	0.583	8.5227	48.124	97.9732
55	0.5307	0.5724	0.5836	8.5146	48.117	97.9771
56	0.5286	0.5712	0.5828	8.5192	48.1162	97.9693
57	0.5316	0.5735	0.5851	8.5142	48.1166	97.972
58	0.5279	0.57	0.5817	8.5239	48.1267	97.9813
59	0.5277	0.5693	0.5806	8.5238	48.1309	97.9886
60	0.5209	0.5625	0.5742	8.5549	48.162	98.0177
61	0.526	0.5672	0.5786	8.5348	48.1461	98.0043
62	0.5285	0.5706	0.5824	8.5251	48.129	97.9806
63	0.5285	0.5705	0.582	8.5216	48.1236	97.9802
64	0.5317	0.5733	0.585	8.5119	48.1158	97.9689
65	0.5288	0.5718	0.5832	8.52	48.1112	97.9693
66	0.5312	0.5734	0.5851	8.513	48.1104	97.9654
67	0.5302	0.5728	0.5845	8.5142	48.1088	97.9639
68	0.5289	0.57	0.582	8.52	48.131	97.9802
69	0.4318	0.4753	0.4902	8.7761	48.3762	98.1869
70	0.4834	0.5259	0.5396	8.6533	48.2429	98.0766
71	0.4479	0.4868	0.5	8.7579	48.3905	98.2229

72	0.3622	0.394	0.4047	8.9961	48.6968	98.5637
73	0.4623	0.5049	0.5184	8.7206	48.3214	98.15
74	0.4012	0.4361	0.4485	8.8899	48.5609	98.407
75	0.4918	0.5334	0.5444	8.6257	48.2359	98.0995
76	0.4843	0.5243	0.5352	8.6424	48.2638	98.1258
77	0.4241	0.458	0.4684	8.8259	48.5014	98.3662
78	0.4322	0.4695	0.4809	8.8154	48.4629	98.3233
79	0.5032	0.5426	0.555	8.5958	48.2237	98.0708
80	0.4979	0.5368	0.5493	8.6075	48.2414	98.0863
81	0.5037	0.5503	0.5634	8.5725	48.0972	97.8966
82	0.4579	0.5029	0.5152	8.6914	48.2409	98.0508
83	0.488	0.5354	0.5491	8.5997	48.1224	97.9113
84	0.4443	0.495	0.5087	8.6961	48.192	97.9813
85	0.5206	0.5627	0.5753	8.5507	48.1551	97.9984
86	0.3947	0.4376	0.4525	8.8885	48.4838	98.2972
87	0.4559	0.4979	0.5117	8.733	48.3354	98.162
88	0.4869	0.529	0.5423	8.6385	48.2421	98.0792
89	0.496	0.5387	0.5517	8.6183	48.2168	98.0555
90	0.4855	0.5281	0.5407	8.6467	48.2491	98.092
91	0.4978	0.5402	0.553	8.6059	48.2075	98.0501
92	0.4963	0.5399	0.5525	8.6164	48.2094	98.0551
93	0.4981	0.5406	0.5533	8.6098	48.2114	98.0562
94	0.4889	0.531	0.5441	8.6397	48.2457	98.0829
95	0.492	0.5345	0.547	8.6339	48.2378	98.0836
96	0.4988	0.5414	0.554	8.6098	48.2102	98.0555

97	0.4963	0.5399	0.5527	8.6176	48.2106	98.0531
98	0.4992	0.5417	0.5549	8.6086	48.2087	98.0454
99	0.493	0.5352	0.5481	8.618	48.2215	98.0609
100	0.4929	0.5362	0.5493	8.6164	48.211	98.0481
101	0.4824	0.527	0.5399	8.6405	48.2234	98.0637
102	0.5019	0.5449	0.5571	8.6016	48.199	98.0489
103	0.5063	0.5486	0.563	8.5869	48.1796	98.0027
104	0.5196	0.5624	0.5753	8.5585	48.1508	97.9914
105	0.5221	0.5656	0.577	8.5511	48.143	98.0
106	0.536	0.5791	0.5915	8.5111	48.1034	97.9479
107	0.535	0.5773	0.5891	8.5099	48.1069	97.96
108	0.5408	0.5833	0.5951	8.487	48.0824	97.9347
109	0.5381	0.5807	0.5926	8.5021	48.0995	97.953
110	0.5372	0.5807	0.5924	8.5014	48.0937	97.946
111	0.5342	0.5753	0.5866	8.5146	48.122	97.9829
112	0.5344	0.5751	0.5871	8.5084	48.122	97.9736
113	0.5418	0.584	0.595	8.4874	48.0902	97.9499
114	0.5391	0.5827	0.5942	8.4959	48.0867	97.9425
115	0.5545	0.5983	0.6095	8.4551	48.0431	97.9036
116	0.5513	0.5961	0.6081	8.4664	48.0447	97.897
117	0.5573	0.6026	0.6138	8.4469	48.0233	97.8818
118	0.5524	0.5961	0.6088	8.4462	48.0346	97.8749
119	0.5345	0.5792	0.5908	8.5115	48.0917	97.9475
120	0.4845	0.5279	0.5409	8.6518	48.2297	98.0684
121	0.5051	0.5475	0.5617	8.5974	48.1967	98.019

122	0.526	0.5685	0.5809	8.5317	48.1259	97.9747
123	0.5297	0.5725	0.585	8.5251	48.1185	97.9627
124	0.5295	0.5712	0.5837	8.5177	48.1193	97.9654
125	0.5308	0.5755	0.5868	8.5251	48.1034	97.9611
126	0.5287	0.5721	0.5843	8.5414	48.1275	97.9763
127	0.5277	0.5702	0.5816	8.5336	48.1333	97.9876
128	0.5135	0.5541	0.5658	8.5709	48.1762	98.0321
129	0.521	0.5627	0.5751	8.5499	48.1547	98.0004
130	0.5237	0.5669	0.5793	8.5422	48.1321	97.9782
131	0.5282	0.5711	0.5831	8.5278	48.1267	97.979
132	0.5455	0.5906	0.603	8.4885	48.0622	97.9083
133	0.5512	0.5956	0.6072	8.4722	48.049	97.9067
134	0.5509	0.5953	0.6074	8.4683	48.0509	97.8997
135	0.5495	0.5931	0.606	8.4609	48.047	97.8846
136	0.5252	0.5698	0.581	8.5391	48.1197	97.981
137	0.4283	0.473	0.4867	8.8033	48.3747	98.2013
138	0.4614	0.505	0.5188	8.7163	48.3047	98.1317
139	0.47	0.5123	0.5243	8.6825	48.2849	98.1349
140	0.4691	0.5113	0.5237	8.7057	48.3102	98.1563
141	0.4751	0.5168	0.5292	8.6739	48.2767	98.122
142	0.4957	0.5387	0.5505	8.6168	48.211	98.0649
143	0.4732	0.514	0.5264	8.6812	48.2744	98.0973
144	0.4637	0.5059	0.518	8.7116	48.3099	98.1515
145	0.4118	0.4531	0.4656	8.8584	48.4547	98.2859
146	0.4439	0.4828	0.4948	8.7571	48.3928	98.2395

147	0.4731	0.5145	0.5267	8.6752	48.2602	98.0806
148	0.4748	0.5163	0.5286	8.6783	48.2806	98.1227
149	0.5171	0.5637	0.5767	8.5639	48.1248	97.9619
150	0.5082	0.5539	0.5677	8.5806	48.1426	97.9736
151	0.498	0.5444	0.5589	8.6024	48.1659	97.9856
152	0.467	0.5143	0.5285	8.6611	48.2134	98.033
153	0.4751	0.5162	0.5286	8.6856	48.2916	98.1322
154	0.4449	0.4865	0.4998	8.7493	48.35	98.1861
155	0.4967	0.5387	0.5517	8.6277	48.2314	98.0714
156	0.5239	0.5671	0.5794	8.5371	48.1257	97.9746
157	0.5235	0.5668	0.5792	8.5387	48.1269	97.975
158	0.5239	0.5673	0.5796	8.5387	48.1253	97.9746
159	0.5236	0.5668	0.5793	8.5383	48.1273	97.9746
160	0.521	0.5648	0.5772	8.5464	48.1308	97.9804
161	0.5185	0.5613	0.5733	8.5526	48.149	97.9998
162	0.4973	0.5384	0.5511	8.6177	48.2311	98.0714
163	0.5091	0.5509	0.5636	8.5919	48.1996	98.0418
164	0.5211	0.5645	0.5773	8.5453	48.1335	97.9781
165	0.5231	0.5656	0.5785	8.5398	48.1358	97.9781
166	0.5241	0.5673	0.5798	8.5363	48.1261	97.9731
167	0.5237	0.5669	0.5794	8.5367	48.1261	97.9738
168	0.5237	0.5669	0.5794	8.5371	48.1257	97.9731
169	0.5235	0.5667	0.5791	8.5371	48.1257	97.9742
170	0.5215	0.5646	0.5775	8.5441	48.1331	97.9762
171	0.5332	0.5775	0.5896	8.5049	48.0836	97.9339

172	0.5269	0.571	0.5834	8.5297	48.1104	97.9572
173	0.5277	0.57	0.5815	8.5247	48.124	97.9802
174	0.5284	0.571	0.5822	8.5297	48.124	97.9868
175	0.5031	0.545	0.5578	8.59	48.1935	98.0338
176	0.51	0.5527	0.5649	8.5826	48.1772	98.0264
177	0.4449	0.4846	0.4965	8.7368	48.2939	98.0427
178	0.5243	0.5672	0.5794	8.5332	48.1244	97.9755
179	0.546	0.5875	0.5993	8.4812	48.0851	97.9355
180	0.4718	0.5162	0.53	8.6813	48.2561	98.0863
181	0.5071	0.5506	0.5646	8.5958	48.1772	98.0093
182	0.5022	0.5446	0.5569	8.609	48.2044	98.0505
183	0.5089	0.5499	0.5635	8.5841	48.1932	98.0276
184	0.5164	0.5607	0.574	8.5725	48.1555	97.9903
185	0.5042	0.5465	0.56	8.5958	48.192	98.0291
186	0.5199	0.562	0.5741	8.5461	48.152	98.0012
187	0.5154	0.5581	0.5702	8.5616	48.1582	98.0051
188	0.4869	0.5257	0.5374	8.649	48.2771	98.1333
189	0.4917	0.5308	0.5434	8.6358	48.2713	98.1135
190	0.5256	0.5678	0.5792	8.5317	48.1333	97.993
191	0.5251	0.5676	0.5797	8.541	48.1376	97.9864
192	0.5376	0.5841	0.5959	8.4951	48.0532	97.9048
193	0.5218	0.5683	0.5811	8.5352	48.1049	97.9468
194	0.534	0.5828	0.5956	8.5014	48.0431	97.8846
195	0.5265	0.574	0.5863	8.5049	48.0614	97.9083
196	0.5242	0.5669	0.5792	8.5402	48.1333	97.9848

197	0.4954	0.5368	0.5508	8.6203	48.2231	98.0482
198	0.503	0.545	0.5577	8.59	48.1928	98.0346
199	0.5273	0.5714	0.5838	8.5297	48.1104	97.9572
200	0.4318	0.4753	0.4903	8.7765	48.3762	98.1866
201	0.4833	0.5259	0.5396	8.6537	48.2429	98.0766
202	0.453	0.4919	0.5047	8.7447	48.3735	98.2114
203	0.3622	0.394	0.4045	8.9961	48.6961	98.5655
204	0.4623	0.5049	0.5182	8.7206	48.3206	98.1512
205	0.4012	0.4362	0.4483	8.89	48.5608	98.4096
206	0.4919	0.5335	0.5445	8.6207	48.2305	98.0925
207	0.4857	0.526	0.5371	8.6378	48.2546	98.1127
208	0.4239	0.458	0.4684	8.8263	48.5005	98.3704
209	0.4328	0.4707	0.4819	8.8146	48.4559	98.3175
210	0.5003	0.5412	0.5535	8.5977	48.2122	98.0587
211	0.4987	0.5378	0.5502	8.602	48.2344	98.0781
212	0.5036	0.5503	0.5632	8.5725	48.0964	97.8978
213	0.458	0.503	0.5151	8.691	48.241	98.0529
214	0.4879	0.5353	0.5488	8.5997	48.1216	97.9126
215	0.4442	0.4951	0.5085	8.6961	48.1912	97.9825
216	0.5212	0.5636	0.5757	8.5492	48.1512	98.0
217	0.5562	0.5999	0.6118	8.4547	48.0385	97.8927
218	0.3786	0.4174	0.4335	8.9697	48.0671	93.9017
219	0.43	0.4715	0.4853	8.8192	48.4021	97.9386
220	0.4118	0.4497	0.4663	8.7812	42.973	75.9607
221	0.4194	0.4617	0.4773	8.8537	47.3203	90.8725

222	0.4288	0.4706	0.4849	8.8196	48.4113	98.0286
223	0.4111	0.4465	0.4609	8.3014	36.5046	62.3287
224	0.4287	0.4706	0.4849	8.82	48.4125	98.0294
225	0.4277	0.4654	0.4809	8.8368	47.9276	94.38
226	0.4214	0.463	0.4794	8.8591	48.3555	97.2061
227	0.4252	0.4701	0.4865	8.8353		93.4771
228	0.4198	0.4653	0.4806	8.8491	47.6401	93.3024
229	0.3949	0.4374	0.453	8.9081	48.2239	95.9906
230	0.4292	0.4731	0.487	8.8162	47.6981	93.7467
231	0.4189	0.4622	0.4777	8.8589	47.6711	93.3275
232	0.4308	0.4725	0.4882	8.8098	47.6129	93.193
233	0.445	0.4848	0.4978	8.7899	47.6955	96.9727
234	0.4493	0.4895	0.5021	8.7773	48.2738	97.2064
235	0.4491	0.4892	0.5017	8.7758	48.2459	97.0475
236	0.4088	0.4518	0.4683	8.8789	48.1256	95.3954
237	0.3436	0.3827	0.3988	9.0354	48.5249	96.811
238	0.3629	0.4048	0.4191	8.9909	48.5698	97.9254
239	0.3478	0.3835	0.3973	9.0412	48.6944	98.3283
240	0.4315	0.4667	0.4795	8.8449	48.4909	98.0254
241	0.3988	0.4374	0.4509	8.9063	48.4347	97.6092
242	0.3967	0.4423	0.4596	8.903	48.3738	97.0931
243	0.4197	0.4653	0.4806	8.8495	47.6396	93.2946
244	0.4203	0.4653	0.4809	8.8497	47.6798	93.5395
245	0.4226	0.4659	0.4807	8.8426	47.8522	94.3769
246	0.4298	0.4741	0.4879	8.8311	48.3136	97.3489

247	0.4212	0.4648	0.4791	8.8448	48.2028	96.487
248	0.4244	0.4665	0.4815	8.8266	47.6479	93.3613
249	0.4527	0.4915	0.5046	8.7641	48.2077	96.8441
250	0.4512	0.4896	0.5028	8.7689	48.2909	97.2516
251	0.4499	0.4887	0.5021	8.7713	48.2967	97.2881
252	0.4472	0.4867	0.4997	8.7773	48.3221	97.4729
253	0.4337	0.4735	0.4861	8.8246	48.4113	97.8528
254	0.4419	0.4798	0.4928	8.7865	48.3863	97.7744
255	0.45	0.4888	0.5021	8.7747	48.2973	97.2743
256	0.3781	0.4207	0.4376	8.9501	48.2028	95.2092
257	0.3783	0.4231	0.4395	8.9649	48.4183	96.9323
258	0.3739	0.4195	0.4343	8.9671	48.4341	97.1447
259	0.3749	0.4194	0.4341	8.9533	48.4806	97.5942
260	0.321	0.359	0.3728	9.1061	48.7467	98.4265
261	0.3409	0.3823	0.3955	9.0427	48.6425	98.227
262	0.3789	0.4231	0.4389	8.9624	48.4267	96.9707
263	0.327	0.3685	0.3838	9.075	48.6861	98.3317
264	0.281	0.3096	0.32	9.2122	48.9419	98.7506
265	0.3148	0.3571	0.3706	9.1123	48.7124	98.4161
266	0.2912	0.3291	0.3419	9.2083	48.8541	98.5738
267	0.2963	0.3354	0.3494	9.1693	48.7355	98.0661
268	0.3526	0.399	0.415	9.0054	48.4169	96.7133
269	0.3082	0.3576	0.3756	9.1373	48.3879	95.4672
270	0.4616	0.5039	0.5174	8.738	48.3326	98.1399
271	0.3906	0.4351	0.4504	8.927	48.2553	96.1403

272	0.4592	0.501	0.5139	8.7465	48.2166	97.0652
273	0.4528	0.4937	0.5065	8.7572	48.1788	96.657
274	0.4118	0.4543	0.4695	8.8711	47.8083	94.0992
275	0.4599	0.4988	0.513	8.7403	48.1859	96.7217
276	0.4229	0.4659	0.4801	8.8431	48.1187	95.8633
277	0.3466	0.3935	0.4099	9.0207	48.471	97.0317
278	0.428	0.4685	0.4824	8.8142	48.4149	97.9915
279	0.3707	0.4167	0.4327	8.9692	48.3345	96.3601
280	0.4147	0.4563	0.4698	8.8484	48.4287	97.8918
281	0.4167	0.4558	0.4683	8.8507	48.4713	98.0873
282	0.3412	0.3834	0.4004	9.0465	48.5626	97.3417
283	0.3777	0.4194	0.4346	8.97	48.5697	98.1308
284	0.4549	0.4935	0.5068	8.7611	48.2984	97.3883
285	0.4263	0.4646	0.4777	8.8386	48.4572	97.9955
286	0.4529	0.4917	0.5051	8.7665	48.3012	97.4003
287	0.455	0.4933	0.5062	8.757	48.2935	97.3957
288	0.4514	0.4902	0.5035	8.7679	48.2962	97.3382
289	0.3447	0.3908	0.4091	9.015	48.5363	97.5524
290	0.4081	0.4528	0.4677	8.8582	48.3615	97.3345
291	0.2866	0.3363	0.3546	9.1565	48.6591	97.792
292	0.3322	0.3809	0.3976	9.0464	48.54	97.612
293	0.4269	0.4719	0.4874	8.8304	47.9296	94.8588
294	0.4374	0.4816	0.4958	8.7975	47.8689	94.7599
295	0.4156	0.4608	0.476	8.8579	47.9593	94.9217
296	0.4265	0.4711	0.4878	8.8362	47.9306	94.8728

297	0.3488	0.3939	0.4114	9.0045	48.5272	97.5673
298	0.4063	0.4509	0.4663	8.8708	48.3757	97.3348
299	0.2921	0.3395	0.358	9.1402	48.654	97.7646
300	0.3376	0.3861	0.4028	9.037	48.5275	97.5675
301	0.4143	0.4568	0.4705	8.8459	48.4003	97.6752
302	0.3765	0.4201	0.436	8.9358	48.4841	97.7535
303	0.3486	0.396	0.4121	9.0085	48.5453	97.949
304	0.3774	0.4242	0.4386	8.9287	48.4579	97.8191
305	0.3805	0.4219	0.4352	8.9436	48.5416	98.2053
306	0.3022	0.3419	0.3564	9.1415	48.7639	98.4605
307	0.3067	0.3499	0.3658	9.1208	48.7157	98.4357
308	0.2528	0.2977	0.3143	9.2511	48.842	98.5751
309	0.3856	0.4318	0.4471	8.9178	48.4127	97.4205
310	0.3421	0.3869	0.4039	9.0488	48.3549	95.6419
311	0.4202	0.4631	0.4807	8.8426	47.5195	92.501
312	0.3292	0.3787	0.3986	8.7681	48.1727	97.8152
313	0.4514	0.4893	0.5022	8.766	48.285	97.2426
314	0.4578	0.4966	0.5092	8.7549	48.3077	97.4908
315	0.4605	0.499	0.5121	8.737	48.3159	97.6286
316	0.4617	0.4997	0.512	8.7394	48.3224	97.6597
317	0.4324	0.4735	0.4865	8.8091	48.3417	97.4305
318	0.5233	0.5689	0.5803	2.5411	12.6958	25.335
319	0.4528	0.4929	0.5064	8.7577	48.3194	97.5202

B.5 MAP, compression, similarity, durations

id	map_score	compress_ratio	compress_percent	sbert_similarity	summarized_duration	retrieval_duration
1	0.4446	18.1324	94.49	0.631	08:39:42	00:02:31
2	0.4728	17.7911	94.38	0.6944	10:08:14	00:02:26
3	0.4733	11.7719	91.51	0.79	09:59:54	00:02:35
4	0.4859	28.8364	96.53	0.8799	02:15:18	00:01:56
5	0.4837	25.1685	96.03	0.8514	04:41:44	00:01:50
6	0.4872	23.364	95.72	0.8525	01:37:24	00:02:12
7	0.4867	11.451	91.27	0.8453	00:47:44	00:02:56
8	0.4858	12.1595	91.78	0.909	00:43:14	00:03:00
9	0.4817	24.3522	95.89	0.9579	00:42:48	00:02:18
10	0.4845	16.3192	93.87	0.9334	01:15:10	00:02:58
11	0.4894	10.8725	90.8	0.835	00:49:08	00:02:56
12	0.4894	11.2988	91.15	0.8317	00:27:57	00:02:55
13	0.5008	16.4252	93.91	0.7973	16:38:42	00:02:43
14	0.495	18.3498	94.55	0.7845	14:52:53	00:02:55
15	0.4996	16.3605	93.89	0.777	13:26:38	00:03:25
16	0.4926	16.1402	93.8	0.7665	22:11:05	00:03:49
17	0.486	12.8827	92.24	0.8407	00:54:55	00:02:46
18	0.4364	23.3223	95.71	0.6044	08:39:42	00:02:06
19	0.4698	22.2815	95.51	0.6621	10:08:14	00:02:12
20	0.4646	14.4379	93.07	0.746	09:59:54	00:02:36
21	0.4813	42.1988	97.63	0.8386	02:15:18	00:01:47

22	0.4801	35.3472	97.17	0.8105	04:41:44	00:01:42
23	0.4766	33.5587	97.02	0.8075	01:37:24	00:02:00
24	0.4831	14.0078	92.86	0.8079	00:47:44	00:02:58
25	0.4773	15.2044	93.42	0.8833	00:43:14	00:03:03
26	0.4574	38.845	97.43	0.9305	00:42:48	00:01:55
27	0.4643	22.7088	95.6	0.9044	01:15:10	00:02:47
28	0.4879	13.1848	92.42	0.7961	00:49:08	00:02:54
29	0.485	13.7519	92.73	0.7916	00:27:57	00:02:55
30	0.4977	22.2337	95.5	0.7514	16:38:42	00:02:25
31	0.489	26.6658	96.25	0.7387	14:52:53	00:02:37
32	0.4943	22.0238	95.46	0.7276	13:26:38	00:03:06
33	0.4899	21.8474	95.42	0.7176	22:11:05	00:03:32
34	0.4866	16.0431	93.77	0.7991	00:54:55	00:03:07
35	0.4335	22.1992	95.5	0.605	08:39:42	00:02:48
36	0.4644	21.5823	95.37	0.6531	10:08:14	00:02:50
37	0.4505	11.045	90.95	0.7389	09:59:54	00:03:01
38	0.4765	26.898	96.28	0.8103	02:15:18	00:02:09
39	0.4762	22.5526	95.57	0.7757	04:41:44	00:02:01
40	0.4735	20.5764	95.14	0.7822	01:37:24	00:02:26
41	0.4753	11.3278	91.17	0.82	00:47:44	00:03:27
42	0.4727	12.8621	92.23	0.8921	00:43:14	00:03:32
43	0.4608	28.178	96.45	0.9443	00:42:48	00:02:24
44	0.4628	18.4235	94.57	0.9156	01:15:10	00:03:17
45	0.4812	10.7059	90.66	0.8091	00:49:08	00:03:20
46	0.4796	11.1712	91.05	0.8067	00:27:57	00:03:22

47	0.4985	13.2478	92.45	0.729	16:38:42	00:03:21
48	0.4813	14.6458	93.17	0.7004	14:52:53	00:03:47
49	0.4912	13.8385	92.77	0.6956	13:26:38	00:04:30
50	0.484	13.8313	92.77	0.6821	22:11:05	00:05:14
51	0.4865	9.7382	89.73	0.8483	00:54:55	00:03:56
52	0.4198	1.8148	44.9	0.7036	08:39:42	00:01:39
53	0.4572	1.8405	45.67	0.8115	10:08:14	00:01:31
54	0.4789	1.2374	19.19	0.9765	09:59:54	00:01:33
55	0.4801	1.6338	38.79	0.9804	02:15:18	00:01:47
56	0.4783	1.5108	33.81	0.9741	04:41:44	00:01:31
57	0.4816	1.3962	28.38	0.9766	01:37:24	00:01:32
58	0.4778	1.2736	21.48	0.9864	00:47:44	00:01:32
59	0.4768	1.3988	28.51	0.9865	00:43:14	00:01:31
60	0.4723	2.3017	56.55	0.9848	00:42:48	00:01:22
61	0.4754	1.9261	48.08	0.9828	01:15:10	00:01:26
62	0.4785	1.2185	17.93	0.9874	00:49:08	00:01:51
63	0.4781	1.2493	19.96	0.987	00:27:57	00:01:36
64	0.4807	1.2046	16.98	0.9663	16:38:42	00:01:41
65	0.4785	1.2125	17.53	0.9618	14:52:53	00:01:45
66	0.4807	1.2097	17.33	0.964	13:26:38	00:01:46
67	0.4797	1.1964	16.42	0.959	22:11:05	00:01:52
68	0.4781	1.2625	20.79	0.9927	00:54:55	00:01:45
69	0.3857	31.1983	96.79	0.572	08:39:42	00:01:47
70	0.4365	28.6314	96.51	0.6119	10:08:14	00:02:04
71	0.4034	20.2861	95.07	0.6224	09:59:54	00:02:20

72	0.327	116.6248	99.14	0.5388	02:15:18	00:01:02
73	0.4197	75.1337	98.67	0.5852	04:41:44	00:01:16
74	0.3607	74.5991	98.66	0.5731	01:37:24	00:01:21
75	0.446	18.1885	94.5	0.739	00:47:44	00:02:46
76	0.4381	19.6796	94.92	0.8464	00:43:14	00:02:50
77	0.3827	59.5032	98.32	0.9101	00:42:48	00:01:37
78	0.3939	31.6473	96.84	0.8748	01:15:10	00:02:30
79	0.4562	17.216	94.19	0.7218	00:49:08	00:02:40
80	0.4506	17.8896	94.41	0.7159	00:27:57	00:02:41
81	0.4527	36.6689	97.27	0.5918	16:38:42	00:02:03
82	0.4093	69.8917	98.57	0.5572	14:52:53	00:02:10
83	0.4358	36.7661	97.28	0.5552	13:26:38	00:02:39
84	0.3942	36.5554	97.26	0.5382	22:11:05	00:03:01
85	0.4732	20.1865	95.05	0.7414	00:54:55	00:03:23
86	0.353	4.3319	76.92	0.7078	08:39:42	00:01:22
87	0.4122	4.2869	76.67	0.8	10:08:14	00:01:21
88	0.4394	2.5311	60.49	0.9391	09:59:54	00:01:22
89	0.449	5.9533	83.2	0.9397	02:15:18	00:01:14
90	0.439	5.04	80.16	0.9224	04:41:44	00:01:15
91	0.4501	4.4068	77.31	0.9391	01:37:24	00:01:19
92	0.4493	2.7471	63.6	0.9702	00:47:44	00:01:26
93	0.4505	2.9203	65.76	0.9807	00:43:14	00:01:28
94	0.4417	6.659	84.98	0.9866	00:42:48	00:01:17
95	0.4457	4.4464	77.51	0.9825	01:15:10	00:01:24
96	0.4514	2.4801	59.68	0.9695	00:49:08	00:01:26

97	0.4497	2.7181	63.21	0.9684	00:27:57	00:01:27
98	0.4514	2.8655	65.1	0.9262	16:38:42	00:01:30
99	0.4452	3.1353	68.11	0.9163	14:52:53	00:01:37
100	0.4447	2.8897	65.39	0.9172	13:26:38	00:01:40
101	0.435	2.824	64.59	0.9088	22:11:05	00:01:50
102	0.4545	2.4392	59.0	0.9792	00:54:55	00:01:34
103	0.4561	14.5425	93.12	0.6364	08:39:42	00:02:49
104	0.4716	13.8655	92.79	0.697	10:08:14	00:02:46
105	0.4734	11.4521	91.27	0.779	09:59:54	00:02:54
106	0.4872	24.1571	95.86	0.8566	02:15:18	00:02:07
107	0.4851	20.3424	95.08	0.8308	04:41:44	00:02:02
108	0.4894	18.9854	94.73	0.8298	01:37:24	00:02:23
109	0.4881	11.6853	91.44	0.84	00:47:44	00:03:17
110	0.4876	12.5717	92.05	0.9058	00:43:14	00:03:19
111	0.4835	24.9486	95.99	0.9554	00:42:48	00:02:21
112	0.4841	17.0395	94.13	0.9292	01:15:10	00:03:12
113	0.4919	11.0594	90.96	0.8294	00:49:08	00:03:13
114	0.4907	11.5234	91.32	0.827	00:27:57	00:03:12
115	0.5026	12.8458	92.22	0.7861	16:38:42	00:03:12
116	0.5004	12.9586	92.28	0.7642	14:52:53	00:03:33
117	0.5056	12.6486	92.09	0.7601	13:26:38	00:04:11
118	0.4999	12.4413	91.96	0.7472	22:11:05	00:04:51
119	0.4871	15.9001	93.71	0.8034	00:54:55	00:03:14
120	0.436	5.4093	81.51	0.6624	08:39:42	00:03:36
121	0.4574	5.0723	80.29	0.7307	10:08:14	00:03:34

122	0.4758	5.0348	80.14	0.8513	09:59:54	00:03:08
123	0.4805	8.1574	87.74	0.8855	02:15:18	00:02:39
124	0.4793	6.9256	85.56	0.8648	04:41:44	00:02:45
125	0.4818	6.6439	84.95	0.8593	01:37:24	00:02:49
126	0.4806	5.7422	82.59	0.9083	00:47:44	00:03:00
127	0.4786	7.5156	86.69	0.9375	00:43:14	00:02:49
128	0.4637	17.8462	94.4	0.952	00:42:48	00:02:00
129	0.4716	12.409	91.94	0.9349	01:15:10	00:02:27
130	0.4747	5.3327	81.25	0.9046	00:49:08	00:03:04
131	0.4788	5.4639	81.7	0.9098	00:27:57	00:03:03
132	0.495	4.5649	78.09	0.8101	16:38:42	00:04:25
133	0.4998	4.4958	77.76	0.7774	14:52:53	00:04:50
134	0.4986	4.5228	77.89	0.7779	13:26:38	00:05:37
135	0.4961	4.4157	77.35	0.7627	22:11:05	00:06:28
136	0.4762	5.4184	81.54	0.9179	00:54:55	00:03:10
137	0.384	6.9353	85.58	0.6521	08:39:42	00:02:57
138	0.4162	6.2676	84.04	0.7027	10:08:14	00:03:14
139	0.4228	8.6342	88.42	0.7342	09:59:54	00:02:29
140	0.4246	21.4394	95.34	0.6997	02:15:18	00:01:50
141	0.4283	14.7572	93.22	0.7041	04:41:44	00:02:03
142	0.4466	13.1064	92.37	0.7147	01:37:24	00:02:15
143	0.4249	11.11	91.0	0.8385	00:47:44	00:02:16
144	0.4179	15.4651	93.53	0.894	00:43:14	00:02:02
145	0.3679	37.4623	97.33	0.9358	00:42:48	00:01:23
146	0.3969	25.99	96.15	0.8964	01:15:10	00:01:42

147	0.4258	10.2626	90.26	0.8313	00:49:08	00:02:16
148	0.4271	10.6176	90.58	0.8419	00:27:57	00:02:16
149	0.4667	7.0041	85.72	0.6926	16:38:42	00:03:37
150	0.458	7.0723	85.86	0.642	14:52:53	00:04:05
151	0.4475	6.9811	85.68	0.6414	13:26:38	00:04:55
152	0.4151	6.6386	84.94	0.62	22:11:05	00:05:37
153	0.4278	10.7346	90.68	0.8465	00:54:55	00:02:20
154	0.3977	1.6561	39.62	0.6939	08:39:42	00:01:54
155	0.4515	1.6606	39.78	0.8128	10:08:14	00:01:59
156	0.4748	1.0289	2.81	0.9912	09:59:54	00:02:22
157	0.4744	1.156	13.49	0.9912	02:15:18	00:02:15
158	0.4749	1.0377	3.63	0.9906	04:41:44	00:02:15
159	0.4746	1.0272	2.65	0.9915	01:37:24	00:02:14
160	0.4728	1.2165	17.8	0.9843	00:47:44	00:02:11
161	0.4695	1.6067	37.76	0.9778	00:43:14	00:02:05
162	0.451	3.6868	72.88	0.9623	00:42:48	00:01:45
163	0.4618	2.4106	58.52	0.964	01:15:10	00:01:49
164	0.4725	1.1309	11.57	0.987	00:49:08	00:02:15
165	0.474	1.163	14.02	0.986	00:27:57	00:02:12
166	0.4749	1.0236	2.31	0.9906	16:38:42	00:02:16
167	0.4747	1.0243	2.37	0.99	14:52:53	00:02:23
168	0.4746	1.04	3.85	0.9903	13:26:38	00:02:15
169	0.4744	1.0341	3.3	0.9902	22:11:05	00:02:19
170	0.4724	1.1289	11.42	0.9914	00:54:55	00:02:14
171	0.484					00:03:23

172	0.4796					00:03:45
173	0.4776					00:01:40
174	0.4783					00:02:12
175	0.4568					00:04:05
176	0.461					00:02:05
177	0.4006					00:01:16
178	0.4748					00:01:25
179	0.4959					00:12:04
180	0.4237	28.6593	96.51	0.5787	08:39:42	00:02:08
181	0.4604	28.1266	96.44	0.627	10:08:14	00:02:19
182	0.4547	9.4495	89.42	0.7493	09:59:54	00:02:24
183	0.4627	20.2677	95.07	0.8188	02:15:18	00:01:47
184	0.47	18.3962	94.56	0.7745	04:41:44	00:01:42
185	0.458	16.1935	93.82	0.7878	01:37:24	00:02:01
186	0.4722	10.1277	90.13	0.8252	00:47:44	00:02:40
187	0.4681	11.7785	91.51	0.8928	00:43:14	00:02:47
188	0.4421	27.9352	96.42	0.9314	00:42:48	00:02:01
189	0.4468	17.3508	94.24	0.9089	01:15:10	00:02:45
190	0.4775	9.6018	89.59	0.8154	00:49:08	00:02:38
191	0.4771	10.0405	90.04	0.8111	00:27:57	00:02:38
192	0.4857	10.954	90.87	0.7307	16:38:42	00:02:31
193	0.4729	13.3874	92.53	0.707	14:52:53	00:02:41
194	0.4828	12.5678	92.04	0.6994	13:26:38	00:03:13

195	0.4749	13.0613	92.34	0.6906	22:11:05	00:03:39
196	0.4783	2.7279	63.34	0.9568	00:54:55	00:02:28
197	0.4504					00:02:55
198	0.4565					00:07:13
199	0.48					00:06:36
200	0.3857	35.1675	97.16	0.5509	08:39:42	00:03:01
201	0.4364	33.7964	97.04	0.5845	10:08:14	00:03:23
202	0.4083	12.4763	91.98	0.6582	09:59:54	00:03:45
203	0.3269	39.4379	97.46	0.6262	02:15:18	00:01:43
204	0.4196	29.8556	96.65	0.6201	04:41:44	00:02:02
205	0.3606	26.5477	96.23	0.6405	01:37:24	00:02:12
206	0.4451	13.098	92.37	0.7729	00:47:44	00:04:26
207	0.4394	15.0072	93.34	0.8657	00:43:14	00:04:37
208	0.3825	36.3905	97.25	0.918	00:42:48	00:02:37
209	0.3941	22.0946	95.47	0.8901	01:15:10	00:04:07
210	0.4532	12.4346	91.96	0.7589	00:49:08	00:04:18
211	0.4501	13.0131	92.32	0.7536	00:27:57	00:04:22
212	0.4526	14.5512	93.13	0.6308	16:38:42	00:03:22
213	0.4093	20.4414	95.11	0.5897	14:52:53	00:03:32
214	0.4356	16.7955	94.05	0.5892	13:26:38	00:04:18
215	0.3941	17.4928	94.28	0.5767	22:11:05	00:05:00
216	0.4739	3.5635	71.94	0.9441	00:54:55	00:05:27
217	0.5052	1.8778	46.75	0.8137	09:42:43	00:02:52
218	0.337	1.2601	20.64	0.9927	00:53:48	01:05:31

219	0.3852					02:30:00
220	0.3704	1.2599	20.63	0.9927	00:53:48	00:38:00
221	0.3773	1.2599	20.63	0.9927	00:53:48	01:20:23
222	0.384	1.2599	20.63	0.9927	00:53:48	02:30:00
223	0.3694	1.2624	20.79	0.9927	00:54:55	00:20:45
224	0.3839	1.2624	20.79	0.9927	00:54:55	02:30:00
225	0.3846					02:21:04
226	0.3788	12.8773	92.23	0.8407	00:53:48	04:24:42
227	0.3832	13.7529	92.73	0.8292	00:53:48	02:15:17
228	0.3774	10.8725	90.8	0.835	00:49:08	02:23:13
229	0.3512	13.1962	92.42	0.796	00:49:08	02:16:44
230	0.3841	26.9594	96.29	0.8365	00:51:38	02:13:36
231	0.3772	8.6037	88.38	0.8123	00:47:39	02:21:22
232	0.3869	11.2539	91.11	0.8617	00:53:48	02:20:38
233	0.3975	5.4384	81.61	0.918	00:53:48	02:16:23
234	0.4032	1.2922	22.61	0.9881	00:51:38	02:15:18
235	0.4025	1.3029	23.25	0.9912	01:31:25	02:14:56
236	0.3647	16.2814	93.86	0.7808	00:53:48	02:13:10
237	0.302					02:23:35
238	0.3215	17.216	94.19	0.7218	00:49:08	02:23:31
239	0.3043	10.2626	90.26	0.8313	00:49:08	02:13:17
240	0.3872	1.1444	12.62	0.9868	00:49:08	02:20:05
241	0.3542	2.4801	59.68	0.9695	00:49:08	02:16:34
242	0.3538	10.8725	90.8	0.835	00:49:08	02:22:49

243	0.3774	10.8725	90.8	0.835	00:49:08	02:24:54
244	0.3778	11.451	91.27	0.8453	00:47:44	02:21:17
245	0.3793	12.1595	91.78	0.909	00:43:14	02:15:13
246	0.3857	24.3522	95.89	0.9579	00:42:48	02:17:30
247	0.3765	16.3192	93.87	0.9334	01:15:10	02:23:42
248	0.3804	11.2988	91.15	0.8317	00:27:57	02:22:12
249	0.4056					02:19:34
250	0.4046	1.2185	17.93	0.9874	00:49:08	02:15:30
251	0.4035	1.2736	21.48	0.9864	00:47:44	02:15:07
252	0.401	1.3988	28.51	0.9865	00:43:14	02:15:13
253	0.3885	2.3017	56.55	0.9848	00:42:48	02:17:06
254	0.3945	1.9261	48.08	0.9828	01:15:10	02:18:34
255	0.4039	1.2493	19.96	0.987	00:27:57	02:15:43
256	0.3361					02:16:10
257	0.3381	13.9273	92.82	0.7659	00:49:08	02:17:50
258	0.3328	14.8818	93.28	0.7784	00:47:44	02:16:37
259	0.3321	17.2063	94.19	0.8639	00:43:14	02:21:04
260	0.2815	49.0236	97.96	0.9195	00:42:48	02:18:07
261	0.3002	28.33	96.47	0.8835	01:15:10	02:23:50
262	0.3379	14.5116	93.11	0.7656	00:27:57	02:14:19
263	0.2848	31.1983	96.79	0.572	08:39:42	02:12:30
264	0.2459	116.6248	99.14	0.5388	02:15:18	02:12:02
265	0.2742	6.9353	85.58	0.6521	08:39:42	02:18:03
266	0.2562	1.6704	40.13	0.6939	08:39:42	02:09:56

267	0.2595	4.3319	76.92	0.7078	08:39:42	02:11:35
268	0.3099	18.1309	94.48	0.631	08:39:42	02:13:45
269	0.2716	28.5542	96.5	0.5789	08:39:42	02:15:11
270	0.415	28.8364	96.53	0.8799	02:15:18	06:45:07
271	0.3469	18.1324	94.49	0.631	08:39:42	02:16:34
272	0.4132	28.8364	96.53	0.8799	02:15:18	02:18:49
273	0.4068	25.1685	96.03	0.8514	04:41:44	02:20:27
274	0.3686	11.7719	91.51	0.79	09:59:54	02:19:55
275	0.413	23.364	95.72	0.8525	01:37:24	02:15:39
276	0.3804	17.7911	94.38	0.6944	10:08:14	02:18:57
277	0.3041	19.3864	94.84	0.6111	08:39:42	02:14:34
278	0.3804	37.6828	97.35	0.6612	01:37:24	02:19:58
279	0.3295	17.7914	94.38	0.6555	10:08:14	02:23:03
280	0.3686	41.0639	97.56	0.6573	04:41:44	02:12:46
281	0.3701	60.3568	98.34	0.6413	02:15:18	02:14:11
282	0.2999	15.1941	93.42	0.6743	09:59:54	02:17:33
283	0.3352	1.8148	44.9	0.7036	08:39:42	02:13:24
284	0.4082	1.3962	28.38	0.9766	01:37:24	02:12:09
285	0.3834	1.8405	45.67	0.8115	10:08:14	02:13:39
286	0.4062	1.5108	33.81	0.9741	04:41:44	02:15:35
287	0.4077	1.6338	38.79	0.9804	02:15:18	02:12:54
288	0.4048	1.2374	19.19	0.9765	09:59:54	02:17:25
289	0.301	21.0446	95.25	0.6085	14:52:53	02:21:10
290	0.359	20.1821	95.05	0.6467	16:38:42	02:14:59
291	0.2454	19.0241	94.74	0.5833	22:11:05	02:22:39

292	0.2881	19.6424	94.91	0.6026	13:26:38	02:23:11
293	0.3813	18.3498	94.55	0.7845	14:52:53	02:15:12
294	0.3906	16.4252	93.91	0.7973	16:38:42	02:15:39
295	0.3694	16.1402	93.8	0.7665	22:11:05	02:15:44
296	0.3815	16.3605	93.89	0.777	13:26:38	02:16:56
297	0.304	21.0446	95.25	0.6085	14:52:53	02:26:00
298	0.358	20.1821	95.05	0.6467	16:38:42	02:17:08
299	0.2494	19.0241	94.74	0.5833	22:11:05	02:22:59
300	0.2935	19.6424	94.91	0.6026	13:26:38	02:24:28
301	0.3661	22.4165	95.54	0.7272	16:38:42	02:16:06
302	0.3304	30.3265	96.7	0.7082	14:52:53	02:18:05
303	0.3026	21.7534	95.4	0.6865	22:11:05	02:21:01
304	0.3291	22.348	95.53	0.6999	13:26:38	02:18:55
305	0.3348	36.6689	97.27	0.5918	16:38:42	02:16:22
306	0.2624	69.8917	98.57	0.5572	14:52:53	02:16:32
307	0.2643	36.7661	97.28	0.5552	13:26:38	02:18:33
308	0.2156	36.5554	97.26	0.5382	22:11:05	02:19:38
309	0.3382	4.5799	78.17	0.7528	14:52:53	02:18:39
310	0.3029					02:22:24
311	0.3781	12.8827	92.24	0.8407	00:54:55	02:20:26
312	0.2831	26.894	96.28	0.8793	02:15:18	00:50:28
313	0.4031	6.6439	84.95	0.8593	01:37:24	02:17:07
314	0.411	35.3472	97.17	0.8105	04:41:44	02:15:35
315	0.4131	33.5587	97.02	0.8075	01:37:24	02:12:09
316	0.4142	42.1988	97.63	0.8386	02:15:18	02:12:54

317	0.3871	22.2815	95.51	0.6621	10:08:14	02:13:39
318	0.4713	28.8364	96.53	0.8799	02:15:18	09:33:51
319	0.4022	2.0872	52.09	0.7878	09:42:43	02:19:37