Ritah Nabawagga

Intent-Aware Retrieval-Augmented Generation for EU Legislation

A Multi-Source Framework for Summarizing Laws, Debates, and News

School of Technology and Innovations
Master of Science in Technology
Master's Programme in Smart Energy

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Ritah Nabawagga
Title of the thesis: Intent-Aware Retrieval-Augmented Generation for EU Legislation : A Multi-Source Framework for Summarizing Laws, Debates, and News
Degree: Master of Science in Technology
Degree Programme: Master's Programme in Smart Energy
Supervisor: Dr. Christos Berberidis
Year: 2026 **Pages:** 87

ABSTRACT:

Retrieval-augmented generation (RAG) pipelines typically employ uniform retrieval strategies across queries, disregarding document-type imbalance in heterogeneous corpora. This thesis investigates whether intent-aware agentic retrieval with source-type weighting improves query-guided summarization over a multi-source EU legislative corpus compared to a standard RAG pipeline. The system classifies each query into one of four intent types (factual, temporal, stance, comparative) and uses this classification to condition source-type retrieval weighting, query decomposition, and summarization prompt selection. These mechanisms are combined with hybrid BM25-dense retrieval using reciprocal rank fusion. A six-level ablation study over 40 benchmark queries evaluates each retrieval component incrementally. The full system achieves attribution of 0.642, outperforming the standard RAG baseline by 15.5%, with modest trade-offs in alignment (−0.025) and key term coverage (−0.028). Hybrid retrieval achieves perfect alignment (1.000) across all 40 queries, making it the strongest single component. The agentic loop increases attribution further (+0.054) but introduces alignment costs on query types where single-pass retrieval is sufficient. An unexpected interaction between hierarchical chunking and inverse-frequency weighting inverts the document-level class distribution at the chunk level. Human evaluation with three raters confirms that automated metrics capture structural summary properties but not substantive adequacy. The findings demonstrate that hybrid retrieval should be the default for multi-source legislative corpora and that component interactions are intent dependent.

KEYWORDS: Retrieval-augmented generation, EU parliamentary legislation, intent-conditioned retrieval, multi-source summarization, agentic retrieval, class imbalance

Contents

Figures

Tables

Abbreviations

Artificial Intelligence Act
AI Act 12
Copyright in the Digital Single Market Directive
Copyright DSM 12
Digital Markets Act
DMA 12
Digital Services Act
DSA 12
Document Understanding Conference
DUC 25
European Union
EU 8
General Data Protection Regulation
GDPR 12
large language models
LLMs 8
Members of the European Parliament
MEP 30
natural language inference
NLI35
natural language processing
NLP 14
Network and Information Security Directive 2 (NIS2 Cybersecurity)
NIS2 Cybersecurity 12
query-focused summarization
QFS 14
Reciprocal Rank Fusion
RRF 18
Retrieval-augmented generation
RAG 8

Use of Artificial Intelligence

During the preparation of this thesis, Claude and Grammarly were used for language-related assistance. They supported grammar correction, style refinement, and clarity improvements throughout the writing process. Claude was additionally used to summarize the author's notes and to provide coding assistance during implementation. The author of the thesis takes full responsibility for the content of the thesis and confirms that all analysis, interpretations, and conclusions were developed by the author.

1 Introduction

This chapter introduces the motivation for the thesis, outlines the research question and hypotheses, and summarizes the main contributions. It also defines the scope and limitations of the work and provides an overview of the thesis structure. Together, these sections establish the conceptual and methodological foundations for the research that follows

1.1 Background and Motivation

Understanding how a piece of European Union(EU) legislation evolves from an initial proposal to an adopted act requires navigating a large and heterogeneous documentary trail. Plenary speeches, legal drafts, amendments, committee reports, and media coverage each capture different aspects of the legislative process, and researchers must determine who supported or opposed provisions, how clauses changed over time, and what obligations the final text imposes. This work relies on extensive manual analysis of diverse sources, which is labor-intensive, error-prone, and difficult to scale (Haag et al., 2025; Xanthaki, 2025). Recent advances in large language models (LLMs) have made it possible to generate coherent, context-aware answers to natural-language questions. Yet purely generative systems remain vulnerable to hallucination, outdated knowledge, and limited traceability. Retrieval-augmented generation (RAG) mitigates these issues by grounding model outputs in retrieved evidence from an external corpus (Gao et al., 2023; Lewis et al., 2021). RAG has demonstrated strong performance in open-domain question answering, scientific literature synthesis, and domain-specific search tasks, and early applications have begun to appear in legal and parliamentary contexts (Arslan et al., 2024). These developments suggest that RAG could support researchers working with EU legislative material by providing targeted access to speeches, proposals, laws, and related documents.

Applying standard RAG pipelines to a multisource legislative corpus, however, introduces challenges that existing work has not addressed jointly. Uniform retrieval strategies assume that all document types are equally informative, even though speeches, legal texts,

and news articles differ sharply in information density, precision, and relevance. Prior work has shown that such uniform retrieval can lead to retrieval drift and suboptimal source selection in mixed-genre corpora (Brown et al., 2025). Second, legislative corpora are structurally imbalanced. In the dataset used in this thesis, plenary speeches outnumber adopted laws by roughly two orders of magnitude; a pattern that reflects institutional practice but poses challenges for dense retrieval. While class-imbalance corrections, such as inverse-frequency weighting, are well established in classification (Cui et al., 2019), their adaptation to retrieval scoring in RAG pipelines has received little systematic attention. Third, standard RAG performs a single retrieval step before generation. This is often insufficient for complex, multi-aspect queries that require synthesizing evidence across multiple documents or source types. Recent work on iterative and agentic RAG introduces mechanisms such as uncertainty-triggered retrieval, self-reflection, and multi-step query refinement (Gao et al., 2023 ;Gupta & Ranjan, 2024). These approaches have not been evaluated in settings where retrieval must navigate both severe class imbalance and heterogeneous document roles, nor have they been conditioned on explicit models of query intent.

1.2 Research Question and Hypotheses

The limitations outlined in Section 1.1 show that existing approaches do not fully accommodate the structure and imbalance of legislative sources or the evidence needs of different query types, which motivates the development of a more adaptive retrieval architecture. The overarching aim of this thesis is to design and evaluate a retrieval-augmented generation pipeline that better supports query-guided summarization of heterogeneous EU parliamentary material.

RQ: **Does intent-aware agentic retrieval with source-type weighting improve query-guided summarization quality across heterogeneous legislative sources compared to a standard RAG pipeline?**

For the purposes of this thesis, a standard RAG pipeline refers to a single-pass dense retriever operating over a multilingual sentence-embedding index using cosine similarity, without any source-type weighting, hybrid retrieval, or iterative refinement. This configuration corresponds to the canonical architecture described in survey work on RAG and serves as the Level-2 baseline in the ablation ladder presented in Chapter 3.

To answer the research question, the thesis decomposes the problem into four hypotheses, each evaluating the incremental contribution of a specific architectural component. Together, these hypotheses trace a progression from simple corpus-aware adjustments to a fully agentic retrieval loop.

H1: **Source-type weighting based on inverse corpus frequency improves summarization quality over unweighted dense retrieval.**

This hypothesis tests whether applying inverse-document frequency weighting to similarity scores at retrieval time can counteract document-type imbalance and improve the quality of retrieved evidence. The comparison is between Level 3 (corpus-frequency-weighted dense retrieval) and Level 2 (unweighted dense retrieval). A positive result would indicate that the extreme speech-to-law imbalance actively harms retrieval and that class-imbalance corrections from classification can be applied to RAG retrieval scoring.

H2: **Intent-conditioned source-type weighting improves summarization quality over corpus-frequency weighting.**

This hypothesis asks whether adjusting the source-type weight distribution based on the query's classified intent provides additional gains beyond a neutral corpus imbalance correction. Factual queries emphasize legal text, stance queries emphasize speeches, and temporal and comparative queries receive intermediate distributions, reflecting the intuition that different source types carry different evidential value depending on the information need. The comparison is between Level 4 (intent-conditioned weights) and Level 3 (corpus-frequency weights).

H3: **Hybrid BM25–dense retrieval improves summarization quality over dense-only retrieval, particularly for factual queries.**
This hypothesis evaluates a hybrid retriever that fuses sparse keyword matching (BM25) with dense semantic retrieval via reciprocal rank fusion. The goal is to better capture exact legal terms, article numbers, and defined concepts that may be missed by dense similarity alone, while retaining the semantic coverage of dense retrieval. The comparison is between Level 5 (hybrid retrieval) and Level 4 (dense-only retrieval), with particular attention to factual queries over legal text.

H4: **Agentic Iterative retrieval with sufficiency checking improves summarization quality over single-pass retrieval, particularly for complex multi-aspect queries.**
This hypothesis tests an agentic retrieval loop that combines LLM-based sufficiency checking, corpus-aware early stopping, and intent-specific query decomposition. The loop is designed to fetch additional evidence only when needed, avoid wasted iterations when the relevant information is structurally absent from the corpus, and decompose complex queries into more targeted sub-queries. The comparison is between Level 6 (agentic iterative retrieval) and Level 5 (single-pass retrieval). In this thesis, the three mechanisms are evaluated jointly as one architectural unit.

The overarching research question is addressed by comparing the full system at Level 6 against the standard RAG baseline at Level 2, while the individual hypotheses identify which components improve quality and under what conditions.

1.3 Contributions

This thesis makes four contributions, three methodological and one infrastructural, aimed at improving retrieval-augmented generation for heterogeneous legislative corpora.

- Intent-conditioned source-type weighting. The thesis introduces a retrieval-scoring approach that adjusts dense similarity scores using both corpus-imbalance

information and query intent, enabling the system to prioritize the document types most relevant to each information need.

- A corpus-aware agentic retrieval loop. An iterative retrieval mechanism is developed that incorporates sufficiency checking, early stopping, and intent-guided query decomposition, allowing the system to acquire additional evidence only when necessary.
- An intent-adaptive hybrid BM25–dense retriever. The system combines sparse and dense retrieval using reciprocal rank fusion, with intent-conditioned source-type weights applied to the dense retrieval scores, improving coverage of legally precise terminology while retaining semantic breadth.
- A reproducible multisource RAG pipeline and evaluation framework. The thesis assembles a curated EU legislative corpus, defines a query intent taxonomy and an evaluation protocol, and provides a modular implementation that supports replication and future research.

1.4 Scope and Limitations

This thesis focuses on retrieval-augmented summarization of EU parliamentary legislative material in English. The corpus contains 1,669 documents from plenary speeches, legislative bills, adopted laws, and news articles, covering six major acts: the General Data Protection Regulation (GDPR), the Artificial Intelligence Act (AI Act), the Digital Services Act (DSA), the Digital Markets Act (DMA), the Copyright in the Digital Single Market Directive (Copyright DSM), and the Network and Information Security Directive 2 (NIS2 Cybersecurity). It includes material discussed up to 2024; later amendments or developments fall outside the system's scope.

Evaluation is based on 40 queries across four intent types. This size balances domain coverage with the computational cost of running seven system configurations, but is not sufficient for statistical significance testing, so results should be interpreted as qualitative patterns.

Summary quality is assessed using five reference-free metrics (coverage, diversity, alignment, attribution, key-term coverage). ROUGE is not used because no gold-standard

summaries exist and producing them would require extensive expert annotation. A small human-rating exercise of 12 summaries provides limited validation.

The system uses a single model, llama-3.3-70b-versatile, for both sufficiency checking and generation, so results reflect the behavior of this model and may not generalize to others. Cross-encoder reranking (bge-reranker-v2-m3) is evaluated only as an optional component due to its computational cost and limited reasoning ability reported in recent legal-retrieval research.

The study excludes committee reports, amendment texts, and other procedural documents, and does not address non-English material or legislative systems outside the EU. The system is a research prototype for evaluating retrieval and summarization methods, not a production tool for legislative analysis.

2 Literature Review

This chapter reviews seven areas of literature that inform the design of the proposed system: RAG, dense and sparse retrieval, hybrid retrieval and cross-encoder reranking, agentic and iterative retrieval, class imbalance correction, multi-source and legal RAG, parliamentary natural language processing (NLP) and legislative corpora, and query-focused summarization (QFS). Each section identifies a specific limitation in the existing work that motivates a corresponding design decision. Section 2.8 draws these threads together and positions the thesis at the intersection of three open problems that prior systems have addressed only in isolation

2.1 Retrieval-Augmented Generation

RAG was introduced by Lewis et al. (2021) as a hybrid architecture combining a pre-trained sequence-to-sequence generator with a non-parametric dense vector index. The core motivation was to address knowledge-intensive tasks where purely parametric models hallucinate or lack current factual grounding, while retaining the flexibility of pre-trained generators. The architecture achieved state-of-the-art performance on open-domain QA benchmarks, with the non-parametric memory updatable by index rebuilding rather than model retraining. Figure 1 illustrates this standard pipeline.

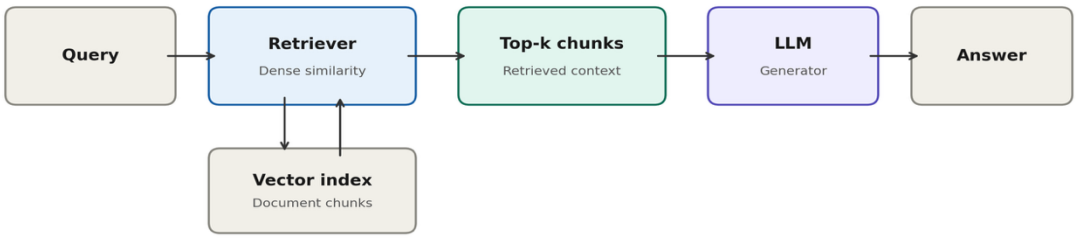

Figure 1. Standard RAG pipeline.

Survey work on RAG situates subsequent developments within a progression from naïve retrieve-read pipelines towards advanced RAG, in which indexing refinements,

query optimization, re-ranking, and mixed sparse–dense retrieval are applied to improve the recall and precision (Gao et al., 2023). These components are examined individually in Section 2.2.

In the legal and parliamentary domains, RAG has been adapted to handle lengthy, structured, domain-specific texts. LexDrafter (Chouhan & Gertz, 2024) grounds LLM-generated term definitions in retrieved fragments from EUR-Lex, demonstrating that RAG can operate on EU legislative material, though it works within a single source type. Jadhav et al. (2025) demonstrate that naive similarity-based retrieval often returns legally irrelevant snippets, and that pipelines incorporating cross-encoder reranking, metadata filtering, and agentic workflows substantially improve the precision-recall trade-off. ConsRAG (Nguyen & Satoh, 2024) introduces aspect-based constraints covering numerical values, definitions, citations, and jurisdictions into both retrieval and generation, combined with an iterative backtracking mechanism that refines retrieval when aspect-level consistency thresholds are not met. CongressRA (Loffredo & Yun, 2025) applies agentic RAG with tool-calling to U.S. Congressional research, while Li et al. (2026) introduce a multi-agent iterative retrieval framework with LLM-based reranking for Chinese statute retrieval called LegalMALR, demonstrating strong gains over single-pass RAG baselines.

Each of these systems works within a homogeneous, single-source corpus. None simultaneously retrieves and synthesizes across source types that differ not only in length and register but also in the kinds of questions they can answer. A plenary speech and an adopted law about the same regulation are not interchangeable evidence: the speech captures political positioning; the law encodes binding obligations. Handling that distinction requires treating source type as a first-class property of retrieval.

2.2 Retrieval Methods for Document Collections

Effective retrieval is the foundation on which the RAG pipeline's quality rests. This section examines the three retrieval paradigms evaluated in this thesis: dense, sparse, and hybrid retrieval, along with cross-encoder reranking as a post-retrieval precision stage.

Each approach addresses a distinct limitation of the others, and their combination is motivated by the specific lexical and semantic demands of EU legislative text.

2.2.1 Dense Retrieval

Dense retrieval represents queries and documents as continuous vectors in a shared embedding space, retrieving the most relevant passages by computing cosine similarity between query and document representations. The approach was established by Karpukhin et al. (2020), who demonstrated that a bi-encoder trained on question–passage pairs substantially outperforms traditional lexical methods in top-k passage recall for open-domain question answering. Subsequent embedding models have built on this foundation through better pre-training objectives and larger training corpora. The E5 model family introduced instruction-tuned embeddings that adapt to different retrieval tasks through natural-language task descriptions, achieving strong performance on the Massive Text Embedding Benchmark. The multilingual variant, intfloat/multilingual-e5-base, used in this work supports over 100 languages, which is particularly relevant given that the EU parliamentary corpus contains speeches originally delivered across 24 official EU languages (Wang et al.,2024).

Dense retrieval excels at capturing semantic similarity between queries and passages that share meaning but differ in vocabulary. However, it has well-documented weaknesses in retrieving passages that contain specific terms, numbers, or identifiers that do not strongly contribute to the overall semantic representation (Karpukhin et al., 2020; Gao et al., 2023). Arabzadeh et al. (2021) argue that no single retrieval approach can effectively serve all query types and propose per-query adaptive selection between retrieval methods. This thesis extends the same principle of per-query adaptation but conditions retrieval on the evidential role of the source type rather than on efficiency, a distinction their work does not address.

2.2.2 Sparse Retrieval and BM25

Sparse retrieval methods represent documents and queries as high-dimensional term-weighted vectors, assigning non-zero values only to vocabulary terms present in the text. BM25, grounded in the probabilistic relevance framework, ranks documents based on term frequency saturation, inverse document frequency, and document length normalization, and remains a consistently strong baseline across retrieval benchmarks, particularly for queries containing specific terms that must be matched exactly (Gao et al.,2023; Robertson & Zaragoza,2009). The BM25F extension further demonstrates that different structural fields within a document carry varying evidential weight, providing an early precedent for field-aware retrieval scoring, conceptually related to the source-type weighting mechanism introduced in this thesis (Robertson & Zaragoza,2009).

The learned sparse retrieval literature has since extended BM25-style term weighting by using neural language models to learn vocabulary-sized impact scores that unify lexical weighting and expansion while remaining compatible with inverted indexes (Brasoveanu et al., 2020; Xu et al., 2025). SPLADE learns sparse lexical and expansion representations via the masked language modeling head of a transformer, implicitly expanding queries and documents to semantically related terms beyond those available in BM25. More recent work replaces encoder-only backbones with decoder-only large language models: Echo-Mistral-SPLADE uses a Mistral-7B causal LLM to achieve state-of-the-art learned sparse retrieval performance on the BEIR benchmark, surpassing all prior SPLADE variants (Doshi et al., 2024), while CSplade scales the approach to 8B-parameter Llama-3 models with Lucene indexes over an order of magnitude smaller than comparable dense indexes (Xu et al., 2025). However, some SPLADE configurations incur substantial per-query latency (Lassance & Clinchant, 2022), and recent multilingual retrieval work further suggests that fixed-vocabulary SPLADE-style models are not yet ideal for cross-lingual settings, with SPLARE outperforming equivalent SPLADE models on multilingual and cross-lingual tasks (Formal et al., 2026).

For the present work, the choice of BM25 over learned sparse alternatives is motivated by two practical considerations. First, BM25 requires no learned representations and no inference-time computation, making it directly deployable without the per-query

latency overhead that some SPLADE configurations incur. Second, BM25 serves a specific architectural purpose in this thesis: as the standalone retrieval baseline at Level 1 of the ablation ladder, providing the simplest possible retrieval lower bound against which each successive component is evaluated, and as the sparse component of the hybrid BM25–dense pipeline at Level 5.

2.2.3 Hybrid Retrieval and Reciprocal Rank Fusion

Hybrid retrieval combines the complementary strengths of sparse and dense methods by fusing their ranked result lists, addressing the documented weakness of each approach in isolation. Sparse retrieval fails when query terms do not appear verbatim in documents, while dense retrieval misses exact lexical matches that are semantically underweighted (Chihaia & Ciobanu, 2025; Karpukhin et al., 2020). The most widely used fusion mechanism is Reciprocal Rank Fusion (RRF), which computes a combined score for each document based on its rank across individual retrieval lists without requiring score normalization between systems. Documents appearing at high ranks across multiple lists receive higher scores, rewarding consensus among retrieval methods while remaining robust to differences in score distributions. Research has shown that RRF consistently outperforms individual retrieval methods and Condorcet-style fusion on TREC collections (Cormack et al., 2009).

Empirical evidence for hybrid retrieval gains is well-established across domains. Lu et al. (2022) show in biomedical literature retrieval that a hybrid of BM25 and a dual encoder consistently outperforms either system individually, and that combining multiple hybrid systems via RRF achieves additional accuracy gains. Sawarkar et al. (2024) confirm in the RAG context that hybrid query strategies consistently outperform single-method retrievers, and that metadata-enriched corpora further amplify the benefits. These findings are particularly relevant to legislative corpora, where legal provisions use precise formal terminology that dense retrieval may underweight, while parliamentary speeches use natural rhetorical language that sparse retrieval may miss.

A more recent development moves beyond static hybrid fusion towards adaptive weighting. Sun (2025) proposes AH-RAG, which dynamically adjusts the relative

contribution of sparse and dense retrieval signals based on query complexity and intent type, demonstrating that dynamic fusion provides measurable gains over fixed-weight hybrid fusion. This thesis applies the same principle where intent-adaptive weights are applied to dense retrieval scores within the hybrid pipeline so that factual queries targeting specific legal provisions receive a stronger sparse signal, while stance and temporal queries receive a stronger semantic signal. The hybrid retrieval component and its intent-adaptive weighting are evaluated as part of the ablation ladder in Chapter 3.

2.2.4 Cross-Encoder Reranking

Cross-encoder reranking operates as a second-stage component in retrieval pipelines, jointly evaluating query–document pairs through full self-attention rather than independently as bi-encoders do (Nogueira & Cho, 2020; Pandit et al., 2026). This joint evaluation enables the model to capture token-level query–document interactions that independent vector embeddings cannot model, yielding more precise relevance scores at the cost of greater computational expense(Yeswanth et al., 2026). Because applying a cross-encoder to an entire corpus is computationally impractical, it is applied only to the top-k candidates returned by a first-stage retriever, a two-stage design that balances retrieval coverage with reranking precision (Busolin et al., 2025). To integrate cross-encoder scores with retrieval signals from earlier stages, a blended scoring approach linearly interpolates cross-encoder outputs with lexical or dense similarity scores, a strategy that has demonstrated gains in both biomedical and domain-specific retrieval settings (Rustamov et al., 2025; Shi et al., 2021). The blend coefficient can be adjusted based on query characteristics, with higher weights assigned to lexical signals for queries where term precision matters and higher weights assigned to semantic signals for queries requiring contextual inference (Stamatis et al., 2024).

For this thesis, bge-reranker-v2-m3 was selected as the cross-encoder reranker. Built on the BGE-M3 backbone, it supports over 100 languages and is trained across diverse retrieval tasks, making it suited to a corpus that combines adopted legal texts, legislative bills, parliamentary speeches, and news articles; document types that differ substantially in register and lexical style (Chen et al., 2024). While LLM-based reranking has shown

promise for complex reasoning tasks, it currently faces practical constraints including high inference latency, sensitivity to input ordering, and a tendency to generate unsupported rationales (Abdallah et al., 2025; Li et al., 2026). Cross-encoder reranking with blended scoring, therefore, represents the appropriate balance between semantic depth, multilingual coverage, and computational tractability for the evaluation conducted in this thesis. The contribution of reranking to overall system performance is reported as a supplementary finding in Chapter 4, evaluated against the full system without reranking.

2.3 Agentic and Iterative Retrieval

Standard RAG, as noted in section 2.1, uses only one retrieval step before generation. This falls short for complex queries that need multi-step reasoning, since a single pass can miss important time periods, overlook some political groups, or only find evidence for one of two entities being compared (Gao et al., 2023; Lewis et al., 2021). Two broad approaches address this limitation: iterative retrieval, which repeats the retrieval step based on a signal from the generation process, and agentic retrieval, which uses an autonomous agent to plan, decompose, and orchestrate retrieval across multiple steps.

FLARE is an iterative system that triggers additional retrieval when the model produces a low-confidence token, using the text generated so far as a new query. Self-RAG extends this by training the model to generate reflection tokens that indicate when to retrieve, whether retrieved passages are relevant, and whether the answer is grounded in evidence. GraphRAG takes a different approach by using knowledge graphs to gather evidence from related entities, supporting queries that require information across multiple connected topics. Adaptive-RAG focuses on how the query's complexity is adjusted, adjusting the number of retrievals as needed. Simple queries end after one round, while complex ones require more rounds, which saves computation without missing important information (Asai et al., 2024; Edge et al., 2024; Jeong et al., 2024; Jiang et al., 2023).

In the legislative domain, CongressRA introduces an agentic RAG system for U.S. Congressional research. This system enables large language models to leverage tools to interact with political knowledge bases, enabling multi-step retrieval and reasoning over

legislative data. LegalMALR uses a multi-agent, iterative retrieval framework for Chinese statutory retrieval. It relies on a planner agent that checks how well the candidate pool is covered in each round and stops when new reformulations are unlikely to yield additional evidence (Li et al., 2026; Loffredo & Yun, 2025). All these iterative and agentic systems demonstrate that multi-step retrieval improves performance on complex queries, and that agent-driven planning adds value when queries require decomposition and coverage assessment across multiple evidence sources. In a multi-source legislative corpus where different document types serve different evidential roles and where query intent determines which sources are most relevant, both properties are needed. This motivates the agentic iterative retrieval loop developed in this thesis, which combines confidence-based termination with intent-specific query decomposition and corpus-aware iteration control.

2.4 Class Imbalance in Retrieval

Class imbalance has been widely studied in supervised classification, where some categories or document types occur much more frequently than others. Cui et al. (2019) defined the effective number of samples and showed that weighting the loss by the inverse of this number, using a square-root variant, improved results on long-tailed datasets compared to simple inverse-frequency weighting. Fernando & Tsokos (2022)extended this with dynamically weighted balanced loss, where class weights depended on both observed frequencies and prediction confidence, helping gradient updates focus on difficult minority-class examples while keeping the model calibrated. In legal text classification, Wang et al. (2022) found severe label imbalance in multi-label medical-dispute cases and proposed an average-sparsity resampling algorithm with ensemble learning, achieving higher recall and F1 than standard resampling across several legal corpora. These studies established that the imbalance must be addressed directly in the model and loss design rather than left to the learning process to resolve.

Applying these ideas to information retrieval means rethinking where imbalance affects the process. In classification, imbalance mainly affects the training loss. In retrieval, it affects the index structure used for ranking. Dai et al. (2024) frame many retrieval biases

as distribution-mismatch problems, where the predicted results differ from a target distribution, and group solutions into a distribution-alignment framework that includes data sampling, rebalancing, and regularization. When some document groups are underrepresented, ranking functions tend to favor majority groups regardless of relevance. Sharma & Bhattarai (2026) note that retrieval from unbalanced corpora can increase existing biases in generated outputs, but do not explore how chunking and document-type frequency together affect retrieval in multi-source settings.

Recent research in retrieval shows that distributional imbalances have a real impact on ranking quality. Ayaou et al. (2026) find that out-of-domain performance is approximately five times worse than in-domain across BM25, dense, and hybrid methods, with dense methods dropping the most, and that hybrid fusion reduces but does not eliminate the problem. Cherumanal et al. (2026) show that standard retrieval methods often under-expose documents linked to minority groups, but score-level re-ranking can rebalance exposure without changing the index. These findings support the principle that adjusting retrieval scores is a practical approach to imbalance correction.

In parliamentary corpora, imbalance comes from how institutions produce documents rather than from label frequencies. Plenary speeches are numerous and short, while adopted laws and formal bills are rare and long. Previous research on domain imbalance (Ayaou et al., 2026), exposure imbalance (Cherumanal et al., 2026), and distribution mismatch (Dai et al., 2024) focuses on either domain or attribute skew but does not address situations in which different document types serve different purposes for different queries. This gap is the reason for the source-type weighting method developed in this thesis, which adjusts dense similarity scores using inverse and intent-based source-type frequencies. This approach extends class-imbalance corrections from loss design and fairness re-ranking to retrieval scoring in a class-imbalanced, multi-source legislative index.

2.5 Multi-Source and Legal RAG

As outlined in Section 2.1, existing legal RAG systems operate almost exclusively on single-source corpora, typically relying on a single homogeneous type of legal material, such as statutes, regulations, or parliamentary research reports. While these systems may

retrieve from multiple documents, they do not integrate heterogeneous evidence types within a unified retrieval pipeline. It is therefore important to distinguish between multi-document legal retrieval, which combines several documents of the same genre, and multi-source retrieval, which integrates fundamentally different kinds of evidence. The literature reviewed in this section shows that while multi-source retrieval has been explored in general-domain settings and multi-document retrieval exists in the legal domain, no prior work combines heterogeneous source types within a legal RAG system.

General-domain research demonstrates why multi-source retrieval requires explicit architectural support. Zhao et al. (2024) show that retrieval sources differ in quality, scale, and knowledge density, and that applying the same retrieval strategy across all sources results in poor source selection. Yu et al. (2025) formalize this as a separate task with OPOR-Bench, a benchmark that requires combining formal news articles and informal social media posts into structured reports with clear source attribution. Their work shows that handling documents from different sources needs special architectural support, not just standard single-source pipelines. Song et al. (2025) address how to manage diverse source types within a single retrieval pipeline by combining keyword, vector, and graph retrieval within a single framework for multi-source data. These systems illustrate the architectural principles needed for multi-source retrieval but do not address the legal domain's structural constraints or evidential requirements.

In contrast, the legal systems in this area operate over multiple legal documents but remain single-source in the sense defined above. Ali et al. (2026) combine several EU regulatory acts using article-based and structure-aware chunking, supported by an LLM-based fact-selection agent. Although this system spans multiple acts, all documents belong to the same source type, and the retrieval pipeline does not incorporate parliamentary debate, legislative proposals, or external reporting. Their finding that human reviewers in compliance settings prefer article-level chunks, even though automated metrics prefer sliding-window chunking, is nonetheless relevant to the chunking decisions made in this thesis.

Thenmozhi et al. (2025) extend this line of work using a hierarchical knowledge graph and agent-driven reasoning over laws, bill histories, and legislative precedents. They

show that fine-tuning for intent identification improves query understanding and that agent-based multi-document retrieval achieves substantially higher semantic similarity scores than traditional keyword methods in complex legal document collections. While this represents a meaningful advance in multi-document legal retrieval, the sources remain homogeneous in genre and evidential role. Goel et al. (2025) introduce a domain-partitioned hybrid RAG that separates legal materials into distinct indexes and routes queries to the appropriate partition. Although this approach acknowledges differences between legal subdomains, it prevents retrieving across multiple source types in a single query, an essential requirement for comparative and temporal legislative questions.

What emerges from this review is a clear separation between systems that handle source heterogeneity in general domains and those that handle multiple documents in legal settings, with no existing system bridging both. To the best of my knowledge, no existing system integrates heterogeneous source types, intent-conditioned retrieval, source-type imbalance correction, and multi-source summarization with explicit attribution within a unified RAG pipeline for EU legislative analysis.

2.6 Parliamentary NLP and Legislative Corpora

Computational analysis of parliamentary text is now a key research area where natural language processing meets political science. This field has produced standardized corpora and analytical methods that are important for the retrieval and summarization tasks in this thesis. The ParlaMint project (Erjavec et al., 2023) compiled uniformly annotated parliamentary corpora from 29 European countries in TEI format with detailed metadata including speaker party, legislative period, and session date. These large, multilingual, and consistently structured corpora demonstrate that high-quality parliamentary data is available at scale for NLP research. European Parliament debates are also available as Linked Open Data, offering metadata-rich resources widely used in machine translation and political text analysis, making the European Parliament one of the most well-documented legislative institutions for computational research (Aggelen et al., 2017).

The EuroParlMin corpus links European Parliament plenary minutes with full session transcripts, creating a structured resource for summarizing parliamentary meetings.

Sood et al. (2026) used EuroParlMin as one of four benchmark corpora in their study of LLM scaling for automatic minuting, finding that EP transcripts consist of long, prepared speeches with few speaker changes, unlike the interactive discussions typical of project or academic meetings. Their findings show that parliamentary texts exhibit unique discourse properties that affect how LLMs process and summarize them.

The primary corpus for this thesis is based on ParlLawSpeech (Schwalbach et al., 2025), which links MEP speeches to the specific legislative proposals and laws they discuss using shared procedure identifiers. This feature makes the corpus useful for temporal and comparative queries. For example, a researcher can track how a regulation changed from proposal to adoption and access both the bill text and related speeches as a linked set. The GDELT (Leetaru, 2013) news archive is also used, providing external coverage of the same legislative procedures and allowing queries about how EU legislation was received outside the parliament. Together, these sources form a genuinely multi-source environment that supports the retrieval and summarization tasks defined in this thesis.

Studies on stance detection in parliamentary texts show that identifying political positions in speeches requires more than just text analysis. Abercrombie & Batista-Navarro (2020) reviewed 61 studies on parliamentary NLP tasks, including sentiment analysis, ideology detection, and position-taking, and found that combining textual and metadata features generally outperforms text-only approaches. Building on this line of work, Kos et al. (2026) demonstrate that transformer-based models can capture ideological patterns in parliamentary debates across 28 countries. Together, these findings support the use of multilingual transformer embeddings for a corpus spanning many EU official languages.

2.7 Query -Focused Summarization

Query-focused summarization (QFS) creates summaries that address a specific information need rather than providing a general overview of the input documents. This task was formalized in the Document Understanding Conference (DUC) evaluation campaigns (Dang, 2005) and reviewed by Nenkova & McKeown (2012). While the literature calls this query-focused summarization, this thesis uses the term query-guided summarization to

highlight that the query influences both retrieval strategy and generation prompt throughout the process, not just as a filter after retrieval. This distinction is important for legislative summarization. For example, a researcher asking which political groups opposed Article 13 of the Copyright Directive needs a system that determines which source types are prioritized, how evidence is weighted, and how the summary is structured.

Early QFS systems often produced summaries that were biased toward the query, focusing too much on matching query terms rather than the actual intent (Katragadda & Varma, 2009). Ya et al. (2020) point out that using only sentence-level attention misses the true intent of the query and leads to mismatches, so they suggest using word-by-word attention and external knowledge to bridge the gap between the query and the document. Reviews show that this problem exists across different methods, from lexical to graph-based to neural approaches, and that modeling intent remains a challenge even for LLM-based QFS systems (Alanzi & Alballaa, 2023; Roy & Kundu, 2024). The need for explicit intent modeling in legislative summarization is therefore grounded in a well-documented limitation of existing approaches

The challenge becomes greater when summarization must draw on heterogeneous sources. Abazari Kia (2021) addresses multi-document summarization from heterogeneous data by jointly learning extractive and abstractive summarization with a temporal structure and combining it with a question-answering framework to focus on relevant evidence. Although this work operates in a corporate rather than legislative context, it demonstrates that heterogeneous, temporally organized document collections require summarization architectures that go beyond standard single-source approaches. Zhang et al. (2025) present a knowledge-intensive QFS system that retrieves relevant documents from a large knowledge base and generates summaries based on both the query and the retrieved context. Liu et al. (2024) introduce QuerySum, a large non-factoid QFS dataset designed to test how well systems model query intent and find that 74% of queries are challenging non-factoid questions. Systems trained on factoid QA datasets often fail when queries require information from multiple perspectives, underscoring the

importance of intent-aware summarization design for queries spanning political positions or cross-regulatory comparisons.

Evaluating QFS without reference summaries is a known challenge. Traditional metrics like ROUGE (Lin, 2004) need human-written reference summaries, which are costly to create for domain-specific queries over varied collections. Roy & Kundu (2024) point out the limits of reference-based evaluation in settings with multiple sources and intents and suggest that reference-free metrics focused on specific quality aspects provide better feedback when gold references are unavailable. The evaluation framework in this thesis follows this approach as described in Chapter 3.

2.8 Research Gap and Positioning

The seven areas reviewed above highlight three main gaps that existing systems have addressed individually but not in combination.

Gap 1: Query-agnostic retrieval in multi-source corpora*.* Standard RAG uses the same retrieval strategy for all query types. Zhao et al. (2024) and domain-partitioned methods (Goel et al., 2025) showed that different sources offer different levels of informational value, and Sun (2025) found that intent-conditioned fusion weights perform better than static ones. What the existing work does not provide is a method that simultaneously adapts retrieval scoring to both the query's intent and the distribution of source types in the corpus. This gap motivates Hypotheses H1 and H2.

Gap 2: Document-type class imbalance in RAG retrieval. Inverse-frequency weighting is well established as a correction for imbalanced classification (Cui et al., 2019; Fernando & Tsokos, 2022), but its application to retrieval scoring introduces complications that the classification literature does not address, particularly when the chunking strategy alters the class distribution at the index level. Dai et al. (2024) describe retrieval biases as distribution-mismatch problems. Ayaou et al. (2026) show that dense retrieval performance drops under distributional skew, and Sharma & Bhattarai (2026) note that retrieving from unbalanced corpora can amplify bias in generated outputs. None of these studies examines how document-type frequency and chunking strategy interact in multi-

source retrieval settings, which is the specific problem examined through Hypotheses H1 and H2.

Gap 3: Agentic retrieval over class-imbalanced multi-source corpora. As established in Section 2.3, iterative and agentic retrieval systems demonstrate genuine gains for complex queries, and Thenmozhi et al. (2025) show that intent identification improves query understanding in multi-document legal retrieval. To the best of my knowledge, no existing system combines corpus-aware iteration control and intent-specific query decomposition in a setting with severe document-type class imbalance. This gap motivates Hypothesis H4.

This thesis aims to address all three gaps simultaneously using a single integrated architecture and to evaluate each part through a controlled six-level ablation study. The closest systems, reviewed in Sections 2.1, 2.3, and 2.5, each handle one or two dimensions: agentic retrieval without source-type awareness, legal RAG without heterogeneous sources, or multi-source retrieval without intent conditioning. To the best of my knowledge, no prior system combines intent-aware retrieval, source-type imbalance correction, and agentic iterative retrieval within a unified framework and evaluates them on a multi-source EU parliamentary corpus for query-guided summarization. Chapter 3 describes how the system is designed to address these three gaps.

3 Methodology

This chapter explains the methods used to design, build, and evaluate the intent-aware agentic retrieval-augmented generation (RAG) system for summarizing EU parliamentary legislation. Section 3.1 outlines the research design and explains why a comparative experimental approach was chosen. Section 3.2 describes how the multi-source legislative corpus was built, covering the sources of data, the selection of domains, the handling of class imbalance, and the chunking method used. Section 3.3 explains the system architecture, describing each stage and its importance, considering previous research limitations. Section 3.4 introduces the query taxonomy that guides intent-based retrieval. Section 3.5 lays out the evaluation framework, including the metrics, the design of benchmark queries, baseline definitions, and the ablation ladder. Section 3.6 provides details on the experimental setup to ensure the work can be reproduced.

3.1 Research Design Overview

The study investigates whether intent-aware agentic retrieval with source-type weighting improves query-guided summarization of EU parliamentary legislative material compared with a RAG pipeline. The central question is whether adapting retrieval to the user's intent leads to higher-quality summaries across five dimensions: coverage, source diversity, intent alignment, attribution density, and key-term recall.

To isolate the contribution of each system component, the study employs a within-system ablation design. A single configurable pipeline is evaluated across multiple configurations, ranging from a sparse retrieval baseline to the full agentic system. This design allows attributing performance differences to specific mechanisms, such as intent classification or source-type weighting, while keeping the corpus, embeddings, and evaluation queries constant across all conditions.

The evaluation uses a 40-query benchmark constructed specifically for this thesis. The benchmark includes 10 queries per intent type and spans six EU legislative domains, ensuring balanced coverage of both intent categories and policy areas. Because no gold-standard summaries exist for multi-source EU legislative queries, the study relies on

reference-free quantitative metrics to assess summarization quality. These metrics are complemented by a human validation step, in which three independent raters evaluate a subset of 12 system outputs using a three-point scale for relevance, attribution accuracy, and intent alignment. Rater scores provide an external check on the reference-free results.

3.2 Dataset Construction

This section explains how the multi-source EU legislative corpus was built for the evaluation. It details the source corpora, the selection of legislative domains, and the class imbalance that influenced some design choices. It also describes the chunking method used to prepare documents for retrieval and the unified index that enables cross-source retrieval within a single embedding space.

3.2.1 Source Corpora

The corpus brings together four different types of sources, each offering a unique perspective on legislative questions. Plenary speeches show the political positions and arguments of Members of the European Parliament (MEP). Legislative bills are the original proposals being considered, and adopted laws are the final texts that have legal force. News articles add current journalistic context and show how legislative debates are presented to the public. The parliamentary documents are drawn from the ParlLawSpeech dataset (Schwalbach et al., 2025), which links MEP plenary speeches to the corresponding legislative proposals and adopted laws through shared procedure identifiers. The European Parliament publishes plenary transcripts in English, and ParlLawSpeech preserves these transcripts alongside the associated legislative texts. This pre-linked structure supports cross-source retrieval across all 24 official EU languages without requiring manual document alignment. News articles were collected from the GDELT Global Knowledge Graph (Leetaru, 2013) and filtered by legislative keywords and date ranges to align with the parliamentary materials.The final corpus comprises 1,669 documents distributed as shown in Table 1.

Table 1. Corpus composition by source type.

Source Type	Documents	Languages	Provenance
Plenary speeches	1,562	24 (translated to EN)	ParlLawSpeech
Legislative bills	16	English	ParlLawSpeech
Adopted laws	11	English	ParlLawSpeech
News articles	80	English	GDELT
Total	**1,669**		

3.2.2 Domain Selection

Six legislative domains were chosen as case studies to capture variation in three key areas. First, they represent different policy fields, such as data protection, platform regulation, intellectual property, and cybersecurity. This helps prevent the system from focusing too narrowly on one type of regulation. Second, the cases span a range of time periods, from the GDPR (proposed in 2012, adopted in 2016) to the AI Act (proposed in 2021, adopted in 2024), allowing for testing questions about changes over time. Third, the cases differ in the level of political controversy. The Copyright Directive and AI Act sparked intense debate in parliament, while NIS2 faced less disagreement. This provides a mix of evidence with varying levels of debate and polarization. The six domains are the GDPR, the AI Act, the DSA, the DMA, the Copyright DSM, and the NIS2 Cybersecurity.

3.2.3 Class Imbalance

One major feature of the corpus is a strong class imbalance at the document level. There are about 142 speeches per law and 98 per bill. This is not due to how the data was collected, but rather shows that the European Parliament creates much more deliberative content than legislative text. This imbalance creates a real challenge for retrieval. Dense vector similarity search tends to return documents that span most of the

embedding space. As a result, in a typical RAG pipeline, a factual question about a legal obligation might return mostly speeches that mention the topic but do not include the exact legal text.

This imbalance led to two main design choices, explained in Section 3.3. First, source-type weighting at the retrieval stage (Stage 2) changes similarity scores to help balance the less common source types. The weighting method is based on class imbalance research, particularly the square-root inverse frequency rule (Cui et al., 2019), which yields a law-to-speech weight ratio of about 12:1 for factual queries. Second, intent-conditioned weight matrices adjust the degree of upweighting for each source type based on the query type. Section 3.3.2 explains this design further.

3.2.4 Chunking Strategy

Legislative texts are much longer than speeches, with bills averaging 220,000 characters and laws about 252,000 characters. To manage this, all documents are divided into retrievable chunks before embedding. The system splits documents at meaningful structural points rather than at arbitrary character counts.

Laws and bills are divided at article boundaries, using structural markers like Article, Section, Chapter, and Recital headings. Each article remains a complete, self-contained unit. Speeches are split by paragraphs, and short paragraphs are combined until they reach about 1,500 characters. News articles are split into groups of sentences, each around 800 characters.

Hierarchical chunking results in a very different distribution of chunks compared to that of whole documents. For example, the 11 laws in the dataset create 2,064 chunks, or about 188 chunks per law. The 16 bills result in 2,566 chunks. In contrast, most speeches become a single chunk, totaling 1,562 speech chunks, and the 80 news articles each form one chunk.

Figure 2 shows the chunk-level distribution by source type and its inversion relative to the document-level composition.

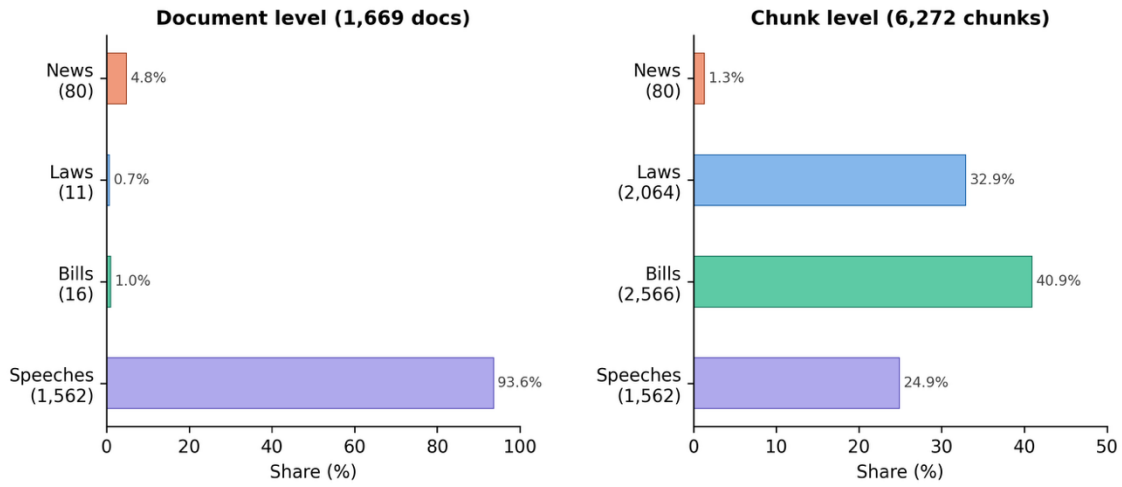

Figure 2. Document Vs Chunk distribution.

This distribution matters because the weighting formula in the corpus-frequency ablation baseline (Section 3.3.2) uses chunk counts instead of document counts. Laws and bills are uncommon as whole documents, but they make up most of the chunks since hierarchical chunking breaks each legal document into many separate articles. Chapter 5 looks at how this chunking method affects frequency-based weighting.

3.2.5 Unified Index Design

All four source types are indexed together in a single FAISS vector space (Johnson et al., 2021) rather than using separate indexes for each type. There are three main reasons for this. First, since all source types are natural-language text embedded by the same model (intfloat/multilingual-e5-base), their cosine similarity scores can be directly compared in the shared space. Using one index avoids the need to normalize scores that would come from separate indexes with different distributions. Second, differences between source types are handled by the intent-conditioned weight matrix described in Section 3.3.2. This method adjusts retrieval scores by source type, so there is no need to separate the indexes unless the source types use completely different embedding spaces, as in image–text cross-modal retrieval (D'Asaro et al., 2024). In this case, the main difference between speeches and laws is due to the makeup of the dataset, not how they are represented. Third, the unified index makes it easier to use the agentic retrieval loop

(Section 3.3.3). When the query decomposer creates sub-questions for different source types, each one searches the full index. This way, evidence from any source type can be found without needing to decide in advance where to look.

3.3 System Architecture

The system works as a pipeline, with each query passing through four main stages in order. There is also an optional fifth stage, cross-encoder reranking, which can be used if GPU resources are available. The design is based on the idea that different legislative queries need different retrieval methods. By adjusting how the system retrieves information based on the query's intent, it can create better summaries than a system that uses the same method for every query.

Stage 1 classifies each query into one of four intent types. This classification then informs three downstream stages. In Stage 2, it selects the appropriate row of the source-type weight matrix for retrieval scoring, prioritizing law chunks for factual queries and speech chunks for stance queries. In Stage 3, it determines the agentic loop's decomposition strategy, splitting temporal queries into time-period sub-queries and comparative queries into dimension-specific sub-queries. In Stage 4, it selects the summarization prompt template, providing a citation-heavy legal format for factual queries and a party-position breakdown for stance queries. The intent signal shapes evidence retrieval, the expansion of evidence, and the structure of the summary. Figure 3 below illustrates the pipeline architecture.

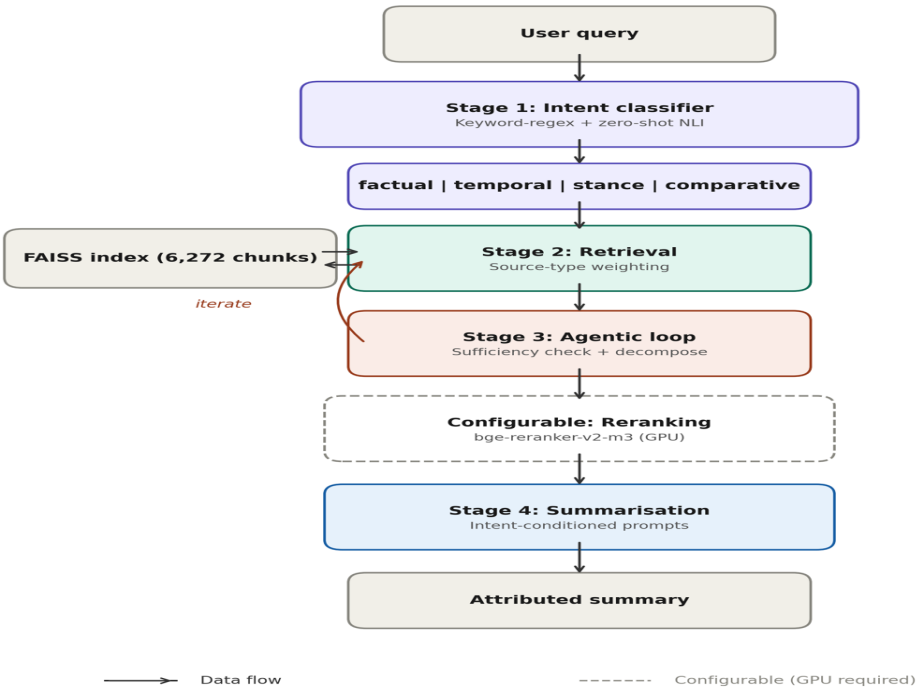

Figure 3. System Architecture.

3.3.1 Stage 1: Intent Classification

Purpose. Standard RAG pipelines handle all queries the same way, retrieving the top-k most semantically similar chunks, regardless of the information the user is looking for. For example, a question about specific legal obligations in the GDPR needs different evidence, mainly legislative text, compared to a question about which political groups opposed the AI Act, which relies more on plenary speeches. In Stage 1, each incoming query is classified into one of four intent types, allowing later stages to adjust their approach as needed.

Design. The classifier uses a hybrid approach that combines keyword-based pattern matching with zero-shot natural language inference (NLI). The keyword layer uses handcrafted regular expressions to identify clear intent signals. For example, queries with phrases like "evolution of" or "over time" are marked as temporal, while those with "compare" or "vs." are marked as comparative. If the keyword layer does not find a match, the query goes to a zero-shot NLI model (facebook/bart-large-mnli) (Yin et al.,

2019). This model checks the query against four natural-language hypothesis templates and selects the intent with the highest entailment score.

Justification. The hybrid design was selected for two main reasons. First, it does not require labeled training data. The keyword rules are derived from analyzing legislative query patterns, and the NLI model operates in a zero-shot setting. This matters because there is no large dataset of EU parliamentary queries labeled with intent. Second, the keyword layer provides high-precision shortcuts for clear cases, reducing unnecessary NLI inference calls and improving speed. The classifier reached 97.5% accuracy on the 40-query benchmark set, with only one misclassification on a query that sits between factual and comparative categories

3.3.2 Stage 2: Source-Type Weighting

Purpose. Even when intent is classified correctly, dense retrieval on an imbalanced dataset tends to miss minority source types. Stage 2 addresses this by using intent-based class weights on retrieval scores, increasing the presence of minority sources based on how underrepresented they are.

Design. First, each retrieval candidate is scored using cosine similarity between the query embedding and each chunk embedding. These scores are then adjusted by multiplying them by a weight from a 4×4 matrix W, which is organized by intent and source type as shown in equation (1).

$$\text{adjusted\_score}(q, c) = \text{cosine\_sim}(q, c) \times W[\text{intent}(q), \text{source\_type}(c)] \quad (1)$$

The weight matrix shows that different types of questions need different sources. For factual questions, laws get the highest weight (0.60) because they provide the most reliable answers. Speeches get the lowest weight (0.05) since they mostly share opinions about laws. For stance questions, the pattern is reversed. Speeches have the highest weight (0.65) because political positions are usually found in parliamentary debates, while laws have the lowest (0.05). Table 2 displays the full weight matrix.

Table 2. Intent-conditioned source-type weight matrix W[intent, source_type].

Intent	**Law**	**Bill**	**Speech**	**News**
Factual	0.60	0.25	0.05	0.10
Temporal	0.15	0.25	0.50	0.10
Stance	0.05	0.10	0.65	0.20
Comparative	0.35	0.30	0.25	0.10

Derivation. The law weight for factual queries is based on the square-root inverse frequency method, which is often used in class-imbalanced classification (Cui et al., 2019). At the document level, there are 142 speeches for every law, so the raw inverse frequency weight for laws is 142. Applying the square root gives about 12. After normalizing across source types and rounding, this results in a factual-law weight of 0.60. The other weights are set based on domain knowledge. For temporal queries, speeches are prioritized because they show the order of debates, and bills are included because they precede final laws. Stance queries focus on speeches and news, as these reflect political positions. Comparative queries use both laws and bills to allow for provision-level comparison. The weight matrix was set before evaluation and was not adjusted using the benchmark queries.

Ablation baseline. To see if intent conditioning offers benefits beyond correcting for corpus imbalance, the evaluation uses a corpus-frequency weighting baseline. In this setup, all intents are given equal weights, calculated as the inverse chunk frequency of each source type and normalized to sum to 1.0. This method follows the standard inverse-frequency correction used in research on class imbalance (Cui et al., 2019; Fernando & Tsokos, 2022). If the intent-conditioned weights (Table 2) perform better than the corpus-frequency weights, it supports the value of the domain-specific reasoning in the matrix. If both perform the same, any improvement is due solely to weighting, not to intent conditioning. This comparison tests H2 as discussed in Chapter 4.

3.3.3 Stage 3: Agentic Retrieval Loop

Purpose. Standard RAG pipelines have a known limitation: single-pass retrieval (Gao et al., 2023). Complex legislative queries often need evidence from several sub-topics. For example, a comparative question about the DSA and DMA might require content moderation rules from the DSA and gatekeeper obligations from the DMA, which probably will not be found in one retrieval pass. Stage 3 solves this by using an iterative loop that breaks down complex queries, gathers more evidence, and stops when enough context has been collected. Figure 4 illustrates the agentic retrieval loop.

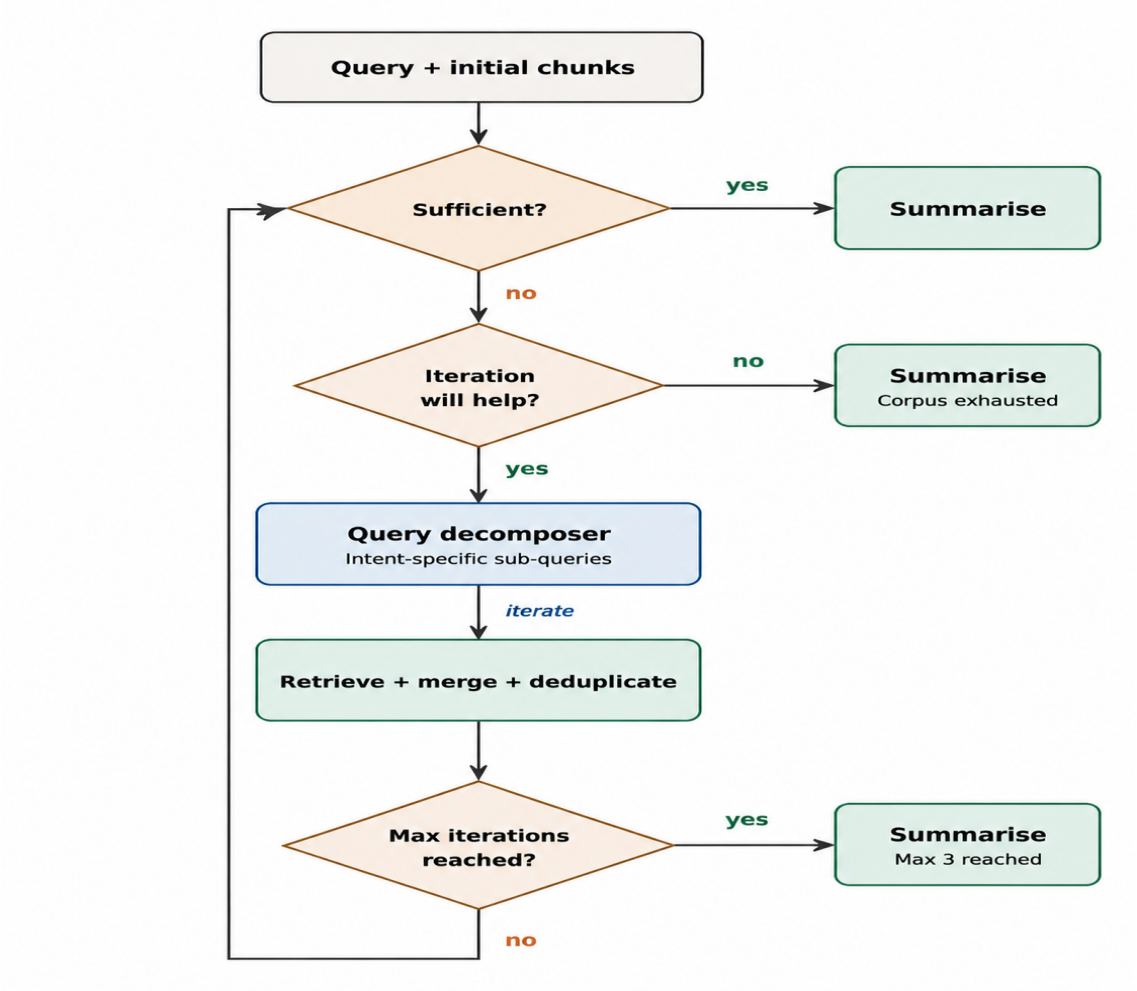

Figure 4. Agentic Loop.

Design. The agentic loop runs for up to three iterations. This limit helps balance the amount of evidence gathered with API rate limits and inference speed, and aligns with

the iteration limits used in similar systems (Jiang et al., 2023; Loffredo & Yun, 2025). Each iteration follows these steps.

First, a sufficiency checker (llama-3.3-70b-versatile via the Groq API) reviews the retrieved chunks to determine whether they provide sufficient evidence to answer the query. The prompt asks the LLM to check if the main parts of the question can be answered using the sources found, without needing every detail. The checker provides a structured response with a sufficiency judgment, a confidence score, a list of evidence gaps, and an ITERATION_WILL_HELP flag indicating whether another round of retrieval could fill those gaps based on the known corpus. If the confidence score is above 0.65, the loop stops early.

Second, if the sufficiency checker finds that the context is insufficient and another round could help, an intent-specific query decomposer creates subqueries to address the gaps. For example, temporal queries are split into time-period subqueries, stance queries are divided by political group, and comparative queries are broken down into subqueries that compare regulations on specific dimensions at once. Factual queries are made broader by removing article-specific limits. Third, these sub-queries go through Stages 1 and 2, and the new results are combined with the existing context, removing any duplicates by chunk identifier.

Corpus-aware early stopping works through the ITERATION_WILL_HELP field. If the LLM finds that the needed information is missing from the corpus, for example, when a query asks about events after the corpus coverage period, it stops the loop instead of making more retrieval calls that would only bring in unrelated material. This idea builds on active retrieval methods such as FLARE (Jiang et al., 2023) and Self-RAG (Asai et al., 2024), as well as the coverage-based stopping rule from LegalMALR (Li et al., 2026). The main change here is that the system checks both the query intent and the corpus structure to decide if enough information has been found, not just the retrieval results.

3.3.4 Stage 4: Intent-Conditioned Summarization

Purpose. The last main stage creates a structured summary with clear sources from the retrieved information. Using a single generic summarization prompt would not meet the

distinct needs of each query type. For example, factual queries need precise legal summaries with many citations, while stance queries need a breakdown by political group. Stage 4 solves this by sending the context to one of four prompts, each designed for a specific intent.

Design. The summarization model used is llama-3.3-70b-versatile, accessed via the Groq API with a temperature of 0.3 and a maximum output length of 1,000 tokens. Four different prompts guide the model to produce the right output for each intent. For factual queries, the model must cite specific articles and provisions from laws. For temporal queries, it gives a timeline showing what changed at each legislative stage. For stance queries, it lists at least three political groups' positions, including MEP names and examples of their speeches. For comparative queries, it uses a four-part structure to show shared scope, unique provisions for each act, and a summary of key differences. The full prompt templates for all four intent types are provided in Appendix 3.

Justification*.* Llama 3.3 70B was chosen for two main reasons. First, it performs well on legal and legislative reasoning tasks and is available through a free-tier API, making it easy to reproduce results without extra costs. Second, its 70B parameter size allows it to follow detailed instructions for the structured outputs needed by each prompt. Smaller models did not do this reliably in early tests.

The intent-conditioned summarization prompts are applied in all system configurations from Level 2 onwards in the ablation ladder. This means that all systems benefit from intent-aware generation. The ablation ladder, therefore, tests the contribution of intent-aware retrieval mechanisms, specifically source-type weighting, hybrid fusion, and agentic iteration, while holding the generation prompts constant. Any differences between ablation levels are attributable to changes in retrieval, not in generation.

3.3.5 Configurable Stage: Cross-Encoder Reranking

Purpose*.* Dense bi-encoder retrieval works efficiently, but it is only approximate. Since query and chunk embeddings are computed separately, they miss detailed token-level interactions. Cross-encoder reranking solves this by using a more accurate, though costlier, relevance model to reorder the top candidates before summarization.

Design. The cross-encoder reranking step uses BAAI/bge-reranker-v2-m3 (Chen et al., 2024), a multilingual reranker with 568 million parameters based on XLM-RoBERTa. This model was chosen instead of the commonly used ms-marco-MiniLM-L-6-v2 (Nogueira & Cho, 2020) for two main reasons. First, its multilingual training data, which includes the MIRACL multilingual retrieval benchmark, is a better fit for the translated EU parliamentary corpus. Second, early experiments showed that ms-marco-MiniLM reduced alignment on stance and comparative queries, while bge-reranker-v2-m3 kept alignment across all intent types.

Justification for configurable status. This stage is configured as a part of the system, not a core pipeline stage, for two reasons. First, cross-encoder inference needs GPU resources that are not available on the main development machine, so it must run on Google Colab with T4 or A100 instances. Second, recent research on legal statute retrieval (Li et al., 2026) found that lightweight cross-encoder rerankers lack sufficient reasoning ability to assess statutory applicability, whereas LLM-based reranking performs better. The main ablation ladder (Levels 1 through 6) is tested without cross-encoder reranking. Reranking is tested separately on top of the full system (Level 6), and its results are reported in Chapter 4.

3.4 Query Taxonomy and Intent Types

The system identifies four types of intent, based on an analysis of the questions that researchers, policymakers, and journalists often ask about EU legislative processes. This classification covers all possible information needs in the context of EU parliamentary work, with each type clearly separated from the others.

Factual queries focus on the specific rules, requirements, penalties, or definitions found in legislative texts. Answers to these questions usually come from laws and bills. For example: "What are the penalty provisions of the GDPR?"

Temporal queries examine how a legislative proposal changed over time, what differences exist between the first draft and the final version, or how political debates shifted across different sessions. Answers come from bills, laws, and speeches arranged by date. For example: "How did the AI Act evolve from proposal to adoption?"

Stance queries are about the political views or positions of MEPs, political groups, or member states on a specific legislative issue. Answers usually come from speeches and news articles. For example: "What was the EPP Group's position on the Digital Markets Act?"

Comparative queries request a clear comparison between two or more legislative acts, or between the draft and final versions of one act. Answers are based on laws and bills, with speeches providing additional context for why the differences exist. For example: "How do the DSA and DMA differ in their regulatory approach?"

3.5 Evaluation Framework

This section defines the evaluation framework used to assess the six ablation configurations. It covers the five automated metrics, the benchmark query set, the baseline systems, and the ablation ladder.

3.5.1 Evaluation metrics

The system is assessed with five reference-free metrics, each focusing on a different aspect of summarization quality. Reference-free evaluation is used because there are no gold-standard summaries for multi-source EU parliamentary queries. Creating these references would require domain experts to read and summarize hundreds of legislative documents for each query, which is not practical for this study.

Coverage measures the proportion of content words from the top retrieved chunks that are included in the summary. It is calculated as the fraction of unique content words, not counting stop words, from the top five retrieved chunks that appear in the summary. Higher coverage means the summary uses more of the available evidence.

Source Diversity measures the number of distinct source types in the retrieved chunks. It is calculated by counting the number of distinct source types (law, bill, speech, news) present in the retrieved context, then dividing by four. A score of 1.0 means all four source types are present.

Intent Alignment checks if the summary's language matches the expected style for its intent type. This is done by looking for certain linguistic markers. For example, factual summaries should include words like "article", "requires", or "defines", while stance summaries should have words like "supported", "opposed", or party names. If at least three expected markers are found in the summary, it gets a score of 1.0.

Attribution measures how many identifiable source references from the retrieved chunks are mentioned in the summary. This is done by collecting domain names, party affiliations, and source-type labels from the retrieved chunks and checking if they appear in the summary. Higher attribution means the summary cites its sources more completely.

Key-Term Recall measures how many predefined domain-specific key terms for each query appear in the summary. Each benchmark query has a list of expected key terms, such as "erasure", "right", "personal data", and "controller" for a GDPR query. Key-term recall is the fraction of these terms found in the summary.

Alongside these automated metrics, a human validation exercise was carried out. Three independent raters reviewed 12 summaries from four system configurations (BM25, No-weights, Hybrid, Full) and rated them on a three-point scale for relevance, attribution accuracy, and intent alignment. Details about the human validation design, results, and how they relate to automated scores are given in Chapter 4.

3.5.2 Benchmark queries

The evaluation set includes 40 benchmark queries, with 10 for each intent type, spread across six legislative domains as listed in Appendix 1. These queries meet three main criteria. First, each one has a clear expected answer that can be checked against the legislative record to assess factual accuracy. Second, the queries cover all six domains to avoid overfitting to any single area. Third, each query is complex enough to need evidence from different types of sources, which tests the system's ability to retrieve information from multiple places. Each query also comes with set key terms and a domain filter for focused evaluation.

3.5.3 Baseline Systems and ablation ladder

To isolate the contribution of each system component, six configurations are evaluated in an ablation ladder, ordered from simplest to most complex. Each configuration adds exactly one component to the previous level, and each transition corresponds to one of the four hypotheses defined in Section 1.2. Table 3 defines the six ablation configurations and their corresponding hypotheses, and Figure 5 illustrates the progressive addition of components.

Table 3. Ablation ladder. Each level adds one component, and each transition **tests one hypothesis**

#	Configuration	Description	Tests
1	BM25	Sparse keyword retrieval (BM25), no embeddings	Baseline
2	No-weights	Dense retrieval (cosine similarity), no source weighting. Standard RAG baseline.	Baseline
3	Corpus-freq	Dense + corpus-frequency source weighting (inverse chunk frequency, same weights for all intents)	H1
4	Single-pass	Dense + intent-conditioned source weighting (weights vary by classified intent)	H2
5	Hybrid	BM25 + dense RRF fusion + intent-conditioned weights	H3
6	Full agentic	Hybrid + iterative agentic retrieval loop (sufficiency checking, early stopping, query decomposition)	H4

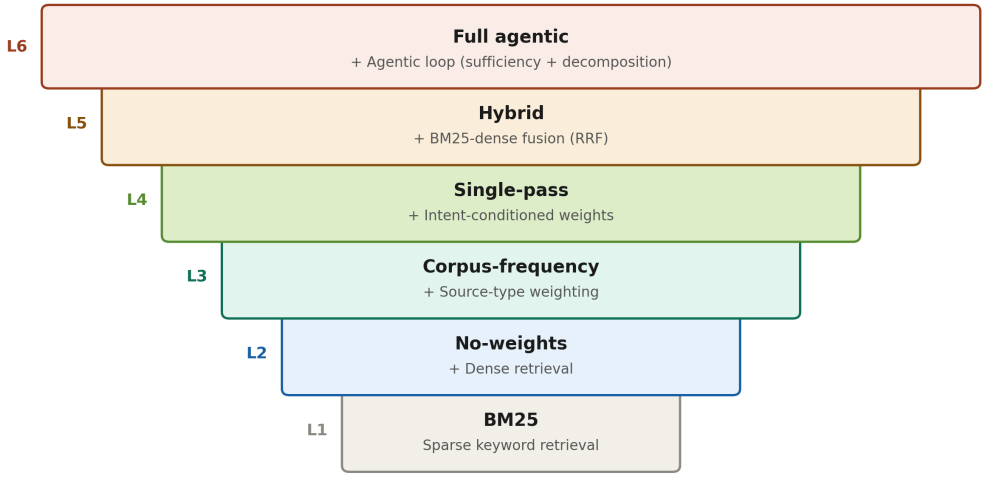

Figure 5. Ablation Ladder.

Each hypothesis is tested by comparing two consecutive levels. H1 (Level 3 vs 2) checks if using corpus-frequency weighting is better than unweighted retrieval. H2 (Level 4 vs 3) looks at whether intent-conditioned weights outperform corpus-frequency weights. H3 (Level 5 vs 4) examines if hybrid BM25-dense retrieval is better than dense-only retrieval. H4 (Level 6 vs 5) tests if the agentic loop improves results compared to single-pass retrieval. To answer the main research question, we compare the full system (Level 6) with the standard RAG baseline (Level 2).

3.6 Experimental Setup

This section describes the technical setup for running the evaluation, covering the models and infrastructure, the development environment, and the evaluation protocol.

3.6.1 Models and Infrastructure

All embeddings are generated using intfloat/multilingual-e5-base (Wang et al., 2024), a multilingual sentence-embedding model that creates 768-dimensional vectors. This model was selected because the ParlLawSpeech corpus includes speeches originally given in 24 EU languages. Although English translations are used throughout, the

multilingual embedding space better captures legislative terms that may retain traces of the source language phrasing. The embedding index is stored in a FAISS IndexFlatIP (Johnson et al., 2021) with normalized vectors, equivalent to cosine similarity, enabling exact nearest-neighbor search across the 6,272 chunk vectors.

Summarization and sufficiency checking are performed via the Groq API, which provides low-latency inference for open-weight large language models. The summarization model (llama-3.3-70b-versatile) is used with a temperature of 0.3 to balance factual precision with natural language fluency. The sufficiency checker uses the same model to ensure consistent assessment quality across the pipeline. A 15-second sleep interval is enforced between consecutive API calls to respect rate limits during evaluation runs.

Cross-encoder reranking uses BAAI/bge-reranker-v2-m3 (Chen et al., 2024), a multilingual reranker with 568 million parameters based on XLM-RoBERTa, trained on multilingual datasets, including MIRACL. As the reranker requires GPU resources, it is executed on Google Colab using T4 instances.

3.6.2 Development Environment

Primary development was conducted in VS Code on an Apple M-series machine (16GB RAM). GPU-intensive tasks, specifically cross-encoder reranking and full ablation runs, are executed on Google Colab with T4 or A100 GPU instances. The codebase is maintained in a public GitHub repository[1] with reproducibility documentation, including a Colab setup guide that specifies the exact steps to reproduce each evaluation session from a cloned repository and pre-built FAISS index files stored on Google Drive.

3.6.3 Evaluation Protocol

Each of the six ablation configurations uses the same set of 40 benchmark queries. In every session, queries ran one after another with a 15-second pause between them. The system behaves predictably during retrieval and reranking, since cosine similarity and

[1] https://github.com/Ritahrouse/eu-parliament-rag

cross-encoder scores stay the same when using the same index. The summarization step introduces a small amount of randomness due to the LLM's temperature setting (0.3), but this value remains the same across sessions. All evaluation outputs were saved, such as retrieved chunks, generated summaries, metric scores, and raw API responses in structured JSON and CSV files for later analysis. Every configuration uses the hierarchical chunking index. Results are reported as the mean and standard deviation across the 40 queries to show both the average performance and the extent of variation within each intent.

4 Results and Analysis

This chapter presents the quantitative evaluation results across the six-level ablation ladder and the supplementary cross-encoder reranking configuration. Section 4.1 reports the overall results. Section 4.2 presents the hypothesis tests. Section 4.3 provides a per-intent analysis. Section 4.4 examines the behavior of the agentic retrieval loop. Section 4.5 reports the supplementary reranking results. Section 4.6 presents the human validation exercise. All results are based on the N=40 benchmark (10 queries per intent type) evaluated across six EU legislative domains. Results are reported as mean ± standard deviation to characterize both central tendency and within-intent variability. Given the sample size (N=40, n=10 per intent), these comparisons are interpreted as directional patterns rather than statistically significant differences.

4.1 Overall results

Table 4 presents the overall performance of each system configuration across the five evaluation metrics, and Figure 6 visualizes this performance.

Table 4. Overall evaluation results (mean ± SD, N=40). Bold values indicate the best performance for each metric across the six main ablation levels.

System	Coverage	Diversity	Alignment	Attribution	Key Terms
L1: BM25	.178±.076	**.719±.116**	.983±.105	.443±.158	.650±.211
L2: No-wt	.216±.085	.519±.207	**1.00±.000**	.556±.214	**.667±.233**
L3: Corp-freq	.210±.081	.519±.222	.958±.172	.559±.245	.651±.233
L4: Single	.218±.086	.500±.196	.983±.074	.547±.207	.657±.253
L5: Hybrid	**.223±.094**	.450±.213	**1.00±.000**	.588±.222	.627±.244
L6: Full	.218±.087	.469±.213	.975±.117	**.642±.222**	.639±.239

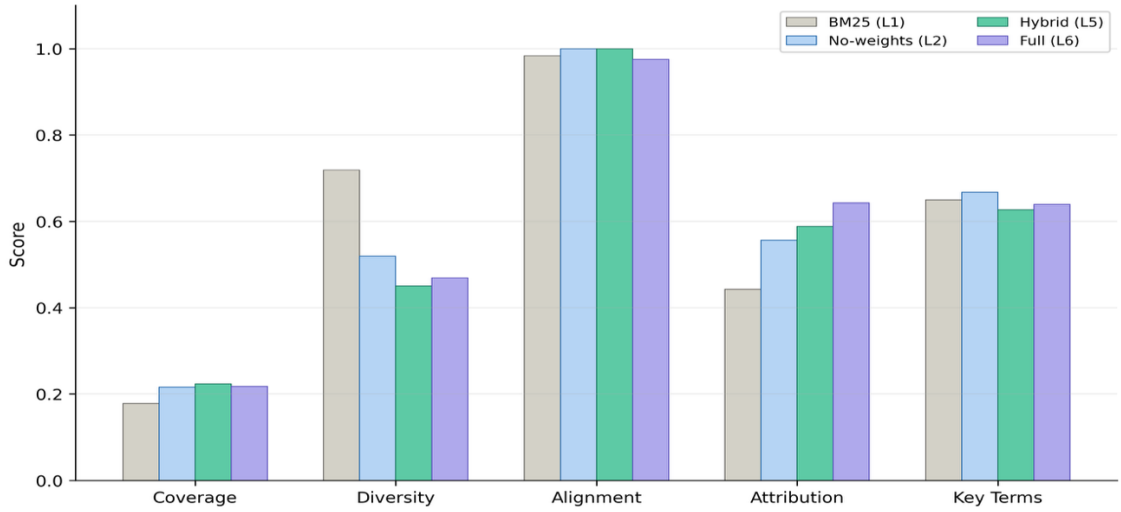

Figure 6. Overall Results.

Two configurations stand out. The Hybrid system (Level 5) achieves perfect alignment (1.000 ± 0.000), the highest coverage (0.223 ± 0.094), and strong attribution (0.588 ± 0.222). A standard deviation of zero on alignment indicates that every query at this level produces a summary whose structure matches the intended intent. The Full system (Level 6) achieves the highest attribution (0.642 ± 0.222), representing a 9.2% increase over Hybrid and a 15.5% increase over the No-weights baseline. The No-weights configuration (Level 2) also achieves perfect alignment and the highest key term coverage (0.667 ± 0.233), establishing a strong reference point for the hypothesis comparisons.

BM25 (Level 1) produces the highest diversity (0.719 ± 0.116) with the lowest variability across queries, consistent with the broader lexical reach of sparse keyword matching. Standard deviations across all systems and metrics range from 0.07 to 0.25, reflecting the inherent difficulty variation across query types and domains.

4.2 Hypothesis Tests

Each hypothesis isolates the contribution of one architectural component by comparing two adjacent levels of the ablation ladder. Table 5 reports the delta between the test and baseline systems.

Table 5. Hypothesis test results (Δ = test − baseline). ↑ improvement > 0.005, ↓ degradation < −0.005, → negligible.

H	Comparison	Cov.	Div.	Align.	Attr.	KT
H1	L3 vs L2	−.006↓	.000→	−.042↓	+.003→	−.017↓
H2	L4 vs L3	+.008↑	−.019↓	+.025↑	−.012↓	+.006↑
H3	L5 vs L4	+.005→	−.050↓	+.017↑	+.041↑	−.030↓
H4	L6 vs L5	−.005↓	+.019↑	−.025↓	+.054↑	+.012↑
RQ	L6 vs L2	+.001→	−.050↓	−.025↓	+.086↑	−.028↓

4.2.1 H1: Corpus-Frequency Weighting (L3 vs L2)

Verdict: Not supported. Three of five metrics decline when corpus-frequency weights are applied: coverage (−0.006), alignment (−0.042), and key terms (−0.017). Attribution increases by 0.003, and diversity is unchanged. The alignment decreased from 1.000 ± 0.000 to 0.958 ± 0.172, the largest single-component degradation observed for this metric. The increased standard deviation indicates that the degradation affects some queries more than others.

The corpus-frequency weight distribution assigns a news weight of 0.891, reflecting the small news sample (80 articles) rather than genuine institutional rarity. This configuration applies the same weight distribution to all query types regardless of intent. In the ablation ladder, corpus-frequency weighting functions serve as a negative control: they isolate the effect of naive inverse-frequency correction when sample sizes are determined by study design rather than by the underlying population distribution.

4.2.2 H2: Intent-Conditioned Weighting(L4 vs L3)

Verdict: Partially supported. Intent-conditioned weighting increases coverage (+0.008), alignment (+0.025), and key terms (+0.006) as compared to corpus-frequency weighting. Attribution decreases by 0.012, and diversity decreases by 0.019. The alignment shift from 0.958 ± 0.172 to 0.983 ± 0.074 is accompanied by a reduction in standard deviation,

indicating more consistent alignment across queries when weights are conditioned on intent.

4.2.3 H3: Hybrid Retrieval (L5 vs L4)

Verdict: Supported for attribution and alignment. Hybrid BM25–dense retrieval produces the largest attribution increase of any single component in the ablation ladder (+0.041) and restores perfect alignment (1.000 ± 0.000). Coverage increases by 0.005. Key term coverage decreases by 0.030, and diversity decreases by 0.050.

The attribution gains concentrate on temporal queries (+0.111) and comparative queries (+0.051), where BM25 retrieves date-specific legal provisions and cross-regulation terms that dense retrieval alone does not surface. The diversity reduction occurs because the combination of two retrieval signals that both favor the most relevant chunks narrows the range of source types in the retrieval window. These gains apply to all types of intent, not just factual queries as we first expected. This wider improvement goes beyond our original hypothesis and shows why the Hybrid configuration is the strongest overall, not just for factual tasks. It is the only system that achieves perfect alignment and the highest coverage.

4.2.4 H4: Agentic Retrieval Loop (L6 vs L5)

Verdict: Partially supported. The agentic loop produces the highest attribution of any system configuration (0.642 ± 0.222), an increase of 0.054 over Hybrid. Diversity increases by 0.019 and key terms increase by 0.012. These gains come at the cost of alignment, which decreases by 0.025 from 1.000 to 0.975 ± 0.117, and coverage, which decreases by 0.005.

The attribution improvement is distributed across intents: factual attribution increases from 0.602 to 0.752 (+0.150), temporal attribution increases from 0.456 to 0.506 (+0.050), and stance attribution increases from 0.707 to 0.720. The alignment decrease concentrates on factual queries, where alignment drops from 1.000 to 0.900 ± 0.225.

Temporal, stance, and comparative queries each maintain alignment at or near 1.000. The agentic loop's iteration behavior is reported in Section 4.4.

4.2.5 Research Question: Full System vs Standard RAG (L6 vs L2)

Verdict: The full system improves attribution over standard RAG but comes with trade-offs in other metrics. Attribution increases by 0.086, from 0.556 ± 0.214 to 0.642 ± 0.222, representing a 15.5% relative improvement. Coverage is unchanged (+0.001). Diversity decreases by 0.050, alignment decreases by 0.025, and key terms decrease by 0.028.
The attribution gain indicates that the combination of intent-conditioned weighting, hybrid retrieval, and agentic iteration produces summaries with more traceable claims and cited sources than the standard RAG baseline. The aggregate result is that the full system offers a different quality profile from standard RAG, stronger evidence grounding at a modest cost to structural consistency, rather than a uniform improvement across all dimensions.

4.3 Per-Intent Analysis

The aggregate metrics in Section 4.1 average across four intent types with distinct evidential requirements. This section reports per-intent results for the three configurations that represent the main comparison points: No-weights (Level 2), Hybrid (Level 5), and Full (Level 6).

4.3.1 Factual Queries

Factual queries present a trade-off between alignment and attribution as shown in Table 6. Both No-weights and Hybrid maintain perfect factual alignment (1.000 ± 0.000), while the Full system achieves the highest factual attribution (0.752 ± 0.195) at the cost of reduced alignment (0.900 ± 0.225). The elevated standard deviation on Full system alignment indicates that a subset of factual queries is affected rather than all of them. No-weights achieves the highest factual coverage (0.278 ± 0.047) with the lowest variability. The Full system's diversity (0.650 ± 0.211) is the highest among the three configurations,

indicating that iterative retrieval introduces additional source types into the factual context window.

Table 6. Factual query results (mean ± SD, n=10).

System	Coverage	Diversity	Alignment	Attribution	Key Terms
No-weights	**.278±.047**	.550±.158	**1.00±.000**	.582±.207	.713±.267
Hybrid	.266±.109	.625±.212	**1.00±.000**	.602±.199	.655±.259
Full	.247±.132	**.650±.211**	.900±.225	**.752±.195**	**.702±.238**

4.3.2 Temporal Queries

All three systems maintain perfect temporal alignment (1.000 ± 0.000) as noted in Table 7. Attribution increases progressively across the three configurations: 0.362 for No-weights, 0.456 for Hybrid, and 0.506 for Full. The Full system also produces the lowest attribution standard deviation (0.111), indicating more consistent source grounding across temporal queries. This progressive improvement suggests that each pipeline component contributes to acquiring temporally distributed evidence. No-weights achieves the highest temporal diversity (0.625 ± 0.132), a pattern that reverses for Hybrid and Full as these systems concentrate retrieval on the most chronologically relevant chunks.

Table 7. Temporal query results (mean ± SD, n=10).

System	Coverage	Diversity	Alignment	Attribution	Key Terms
No-weights	.214±.064	**.625±.132**	**1.00±.000**	.362±.121	.515±.259
Hybrid	.197±.058	.325±.121	**1.00±.000**	.456±.126	.470±.230
Full	.196±.039	.375±.132	**1.00±.000**	**.506±.111**	.450±.183

4.3.3 Stance Queries

All three systems achieve perfect stance alignment (1.000 ± 0.000). No-weights produces the highest stance attribution (0.748 ± 0.178) and key terms (0.733 ± 0.159). Both Hybrid and Full produce diversity of 0.250 ± 0.000, indicating that these systems consistently retrieve only parliamentary speeches for stance queries. While substantively appropriate given that political positions are expressed in speeches, this ceiling means that neither BM25 fusion nor the agentic loop contributes additional source types for this intent. The elevated attribution standard deviation for Hybrid and Full (0.294 for both) indicates greater variability in source citation quality compared to No-weights (0.178). See Table 8.

Table 8. Stance query results (mean ± SD, n=10).

System	Coverage	Diversity	Alignment	Attribution	Key Terms
No-weights	.175±.093	**.350±.211**	**1.00±.000**	**.748±.178**	**.733±.159**
Hybrid	.156±.055	.250±.000	**1.00±.000**	.707±.294	.650±.235
Full	.158±.059	.250±.000	**1.00±.000**	.720±.294	.633±.220

4.3.4 Comparative Queries

Comparative queries require evidence from multiple regulations for systematic comparison. All three systems achieve perfect comparative alignment (1.000 ± 0.000) as recorded in Table 9. Hybrid produces the highest coverage (0.275 ± 0.096). Both Hybrid and Full achieve diversity of 0.600 ± 0.129, indicating three of four source types are consistently retrieved. The Full system achieves the highest comparative key terms (0.770 ± 0.214) and attribution (0.589 ± 0.185), both marginally above Hybrid. The low coverage standard deviation for Full (0.045) compared to Hybrid (0.096) and No-weights (0.099) indicates more consistent coverage across comparative queries.

Table 9. Comparative query results (mean ± SD, n=10).

System	Coverage	Diversity	Alignment	Attribution	Key Terms
No-weights	.199±.099	.550±.230	**1.00±.000**	.532±.157	.707±.193
Hybrid	**.275±.096**	**.600±.129**	**1.00±.000**	.587±.189	.734±.205
Full	.270±.045	**.600±.129**	**1.00±.000**	**.589±.185**	**.770±.214**

4.4 Agentic Loop Behavior

The agentic retrieval loop iterates up to a maximum of three retrieval passes per query. At each iteration, a sufficiency checker assesses whether the accumulated evidence is adequate to address the query. If the confidence threshold (0.65) is met, the loop terminates early. Table 10 reports the iteration statistics.

Table 10. Agentic loop iteration statistics for the Full system (N=40).

Intent	Mean iters	Max iters	Mean chunks	Sufficient
Factual	1.60	3	5.4	7/10
Temporal	1.80	3	14.6	5/10
Stance	1.40	3	14.6	9/10
Comparative	1.60	3	7.1	7/10
Overall	**1.60**	**3**	**10.4**	**28/40**

The sufficiency checker terminates the loop early for 28 of 40 queries (70%). Stance queries have the highest early-termination rate (9/10), followed by factual and comparative (7/10 each) and temporal (5/10). The mean iteration count across all queries is 1.60, ranging from 1.40 for stance to 1.80 for temporal.

Mean chunk counts vary by intent. Temporal and stance queries accumulate the most chunks on average (14.6 each), while factual queries accumulate the fewest (5.4).

Comparative queries retrieve 7.1 chunks on average. The overall mean of 10.4 chunks across 40 queries is lower than the 12-chunk maximum retrieved by single-pass systems.

4.5 Supplementary: Cross-Encoder Reranking

As recorded in Table 11, reranking increases coverage (+0.021) and key terms (+0.021) while leaving attribution unchanged and reducing alignment by 0.017. The coverage improvement indicates that the cross-encoder promotes more content-relevant chunks to higher positions in the retrieval window. Attribution remains at 0.642 for both configurations, suggesting that reranking reorders chunks without changing which sources are represented in the context window. The alignment decrease is concentrated in factual queries (0.900 to 0.867) and comparative queries (1.000 to 0.967). Temporal and stance alignment remain at 1.000. Factual attribution increases from 0.752 to 0.793, the largest per-intent attribution gain from reranking.

Table 11. Full system with and without cross-encoder reranking (mean ± SD, N=40).

Configuration	Coverage	Diversity	Alignment	Attribution	Key Terms
L6: Full	.218±.087	.469±.213	.975±.117	.642±.222	.639±.239
L6 + Rerank	**.239±.083**	.463±.208	.958±.135	.642±.202	**.660±.211**
Δ	+.021	−.006	−.017	.000	+.021

4.6 Human Validation

Out of 36 ratings for each dimension (12 summaries × 3 raters), the mean human scores were 2.44 ± 0.68 for relevance, 2.50 ± 0.73 for attribution, and 2.58 ± 0.60 for alignment. Most scores fell between 2 and 3 as shown in Table 12, indicating that raters usually found the summaries at least partially adequate across all three areas.

Table 12. Human validation ratings (mean of 3 raters, 1-3 scale) alongside automated metric scores for each summary.

ID	System	Intent	H-Rel	H-Attr	H-Align	A-Cov	A-Attr	A-Align	A-KT
S01	Hybrid	Factual	3.00	2.67	3.00	0.392	1.000	1.000	1.000
S02	Full	Stance	3.00	2.33	2.67	0.221	0.500	1.000	0.800
S03	Hybrid	Compara-tive	2.33	3.00	2.67	0.278	0.500	1.000	0.875
S04	BM25	Temporal	2.67	2.67	2.33	0.191	0.333	1.000	0.200
S05	Full	Stance	2.33	1.00	2.67	0.245	1.000	1.000	0.800
S06	Full	Compara-tive	1.33	2.67	2.00	0.343	0.400	1.000	0.800
S07	BM25	Factual	2.33	2.67	2.33	0.265	0.333	0.333	0.600
S08	No-weights	Temporal	2.67	2.67	2.67	0.152	0.500	1.000	0.500
S09	Full	Temporal	3.00	2.67	2.67	0.126	0.500	1.000	0.800
S10	No-weights	Stance	2.67	1.67	2.33	0.044	0.333	1.000	0.400
S11	Hybrid	Compara-tive	2.33	3.00	2.67	0.209	1.000	1.000	1.000
S12	Full	Factual	1.67	3.00	3.00	0.086	1.000	1.000	0.40

Inter-rater agreement was measured using Fleiss' kappa, a statistic that quantifies the degree to which multiple raters assign consistent scores beyond what would be expected by chance alone. A kappa of 1.0 means perfect agreement, 0 means agreement is only by chance, and negative values mean less agreement than expected by chance. We also used Spearman's rank correlation to see if summaries rated highly by humans were also rated highly by automated metrics. A positive rho indicates that the two rankings align in the same direction. Relevance yielded a kappa of 0.120 (slight agreement), attribution yielded 0.150 (slight agreement), and alignment yielded −0.178 (below chance

agreement). These low kappa values indicate substantial rater disagreement, particularly for alignment, where raters inconsistently interpreted intent-appropriate structure. This level of agreement is consistent with the inherent difficulty of evaluating domain-specific legal summaries on a compressed three-point scale. It is comparable to inter-annotator agreement levels reported in meeting summarization evaluation(Sood et al., 2026).

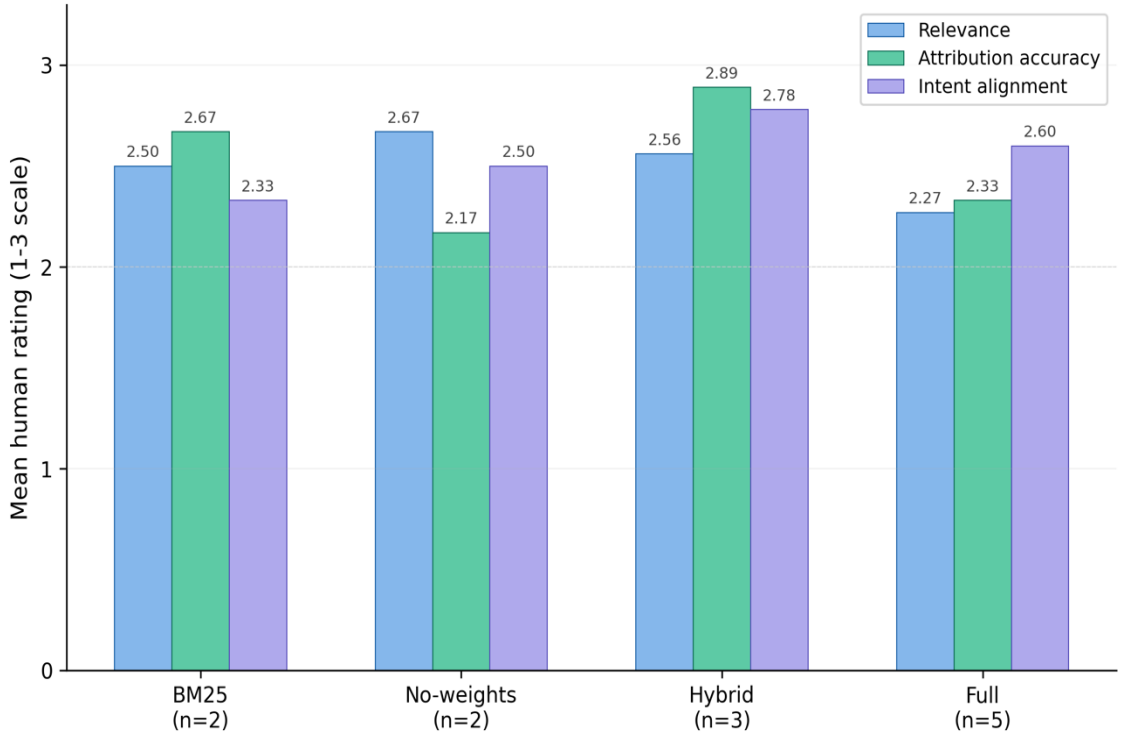

Figure 7. Mean human ratings by system (averaged across summaries and raters).

Figure 7 shows the mean human ratings by system configuration. The Hybrid system achieved the highest mean attribution (2.89) and alignment (2.78) from human raters, which is consistent with automated evaluations indicating the highest alignment and strong attribution for this system. In contrast, the Full system obtained the lowest mean relevance (2.27), primarily due to low ratings on S06 (comparative, mean 1.33) and S12 (factual, mean 1.67). These two summaries exemplify instances in which the system generated well-structured, well-cited output that failed to substantively address the query, a pattern further analyzed in the Discussion.

Spearman's rank correlations between human ratings and automated scores were calculated for the 12 summaries. Human and automated attribution showed a positive but

non-significant correlation (rho = 0.269, p = 0.397). Similarly, human and automated alignment showed a positive but non-significant correlation (rho = 0.330, p = 0.295). None of the correlations reached statistical significance, which is expected given the small sample size (n = 12) and the limited statistical power of rank correlations at this scale. The positive correlations for both attribution and alignment suggest partial correspondence between human and automated assessments; however, the sample size is insufficient to confirm this relationship.

4.7 Summary of Findings

The key findings from the N=40 evaluation are summarized below.

- Hybrid retrieval is the most consistent component. The Hybrid configuration (Level 5) achieves perfect alignment across all 40 queries, the highest coverage (0.223), and strong attribution (0.588). No other system achieves this combination.
- The agentic loop produces the highest attribution of any configuration. The Full system (Level 6) achieves an attribution of 0.642, a 15.5% increase over the standard RAG baseline and a 9.2% increase over Hybrid. This gain is distributed across factual (+0.150 over Hybrid), temporal (+0.050), and stance (+0.013) intents.
- Intent-conditioned weighting produces targeted but modest gains**.** It recovers alignment degraded by corpus-frequency weights and improves coverage but does not clearly outperform the standard RAG baseline on aggregate metrics.
- The full system offers a different quality profile rather than a uniform improvement. Compared to standard RAG, the Full system trades diversity (−0.050) and alignment (−0.025) for substantially better attribution (+0.086).
- Cross-encoder reranking provides complementary gains in coverage and key terms. Adding bge-reranker-v2-m3 increases coverage and key terms by 0.021 each without affecting attribution. Alignment decreases by 0.017.

5 Discussion

This chapter interprets the findings from Chapter 4 and places them within the context of existing literature. It starts in Section 5.1 by evaluating the research question and each hypothesis. Next, Section 5.2 examines the comparative query challenge in detail. Section 5.3 shows concrete system outputs to illustrate these points. The discussion then proceeds to Section 5.4, which evaluates the validity of the automated metrics used in this study. Finally, Section 5.5 concludes the chapter by identifying the limitations inherent in this work.

5.1 H1: The Failure of Naive Corpus-Frequency Weighting

This section explains the results for each hypothesis, using the per-intent analysis and agentic loop behavior from Chapter 4 to clarify the patterns seen. At the end, the research question is answered by bringing together the findings from each hypothesis.

5.1.1 H1: The Failure of Naive Corpus-Frequency Weighting

The corpus-frequency weighting scheme increases retrieval scores for source types with fewer chunks in the index, following the inverse-frequency correction principle established by Cui et al. (2019). The rationale is that when one source type dominates the index, a retrieval system is more likely to return documents from that source type because they are abundant in the embedding space, rather than their relevance. Upweighting minority source types is intended to mitigate this bias. However, empirical results do not support this hypothesis. Corpus-frequency weighting reduces alignment by 0.042 and key term retrieval by 0.017 compared to unweighted retrieval.

The problem originates in the interaction between hierarchical chunking and the weighting formula. At the document level, speeches outnumber laws and bills (1,562 speeches versus 11 laws and 16 bills). However, the weighting formula operates at the chunk level, and hierarchical chunking splits each law and bill into individual articles and sections. The 11 laws in the corpus yield 2,064 chunks under hierarchical chunking,

averaging approximately 188 per law document. As a result, the chunk-level distribution is inverted. Bills account for 2,566 chunks (40.9%), laws for 2,064 chunks (32.9%), speeches for 1,562 chunks (24.9%), and news for just 80 chunks (1.3%). Laws and bills, which are rare at the document level, become the majority at the chunk level.

The inverse-frequency formula accurately represents this inverted distribution. It assigns the highest weight to news (0.891) due to its minimal chunk count, while laws and bills receive the lowest weights (0.035 and 0.028) because they have the most chunks. The intended outcome was to upweight scarce legal text, increasing its presence in retrieval results. In practice, the effect is the opposite: the formula downweighs legal text and significantly upweights news commentary. As a result, queries using corpus-frequency weighting retrieve a disproportionate number of news chunks, which typically contain general media commentary rather than the specific legal content required by most queries.

This finding indicates that applying inverse-frequency corrections at the chunk level can yield counterproductive results when the chunking strategy alters the class distribution. Within the ablation ladder, corpus-frequency weighting functions as a negative control, isolating the effects of applying inverse-frequency correction without considering the interaction between document structure and chunking in shaping the retrievable index

5.1.2 The Value and Limits of Intent Conditioning

Intent-conditioned weighting increases alignment from 0.958 to 0.983 and enhances both coverage and key term retrieval compared to the corpus-frequency baseline. These results suggest that conditioning the weight distribution on query intent yields a more effective correction than applying uniform weights. By assigning greater law weights to factual queries and greater speech weights to stance queries, the system adjusts its retrieval scoring to meet the evidential requirements of each query type.

However, the H2 comparison is affected by the limitations of the pathological corpus-frequency baseline. When evaluated against the Level 2 standard RAG baseline, intent conditioning yields smaller and more inconsistent effects. This limitation does not undermine the validity of the intent-conditioning mechanism. The per-intent analysis in

Chapter 4 demonstrates that different source types are appropriately prioritized for distinct intents. Nevertheless, this issue restricts the strength of the conclusions that can be drawn from the aggregate comparison. Additionally, the current design changes two variables simultaneously between L3 and L4. The base weight derivation shifts from chunk-level inverse frequency to document-level inverse frequency, and intent conditioning is introduced at the same time. A cleaner test would hold the base frequency distribution constant and vary only the presence or absence of intent-specific adjustments, so that the independent contribution of intent conditioning can be measured without confounding.

It is important to note that intent conditioning in this thesis operates at two levels. The summarization prompts (Stage 4) are conditioned on intent in all system configurations from Level 2 onwards, and the perfect alignment achieved by L2 indicates that intent-aware generation is effective on its own. The H2 comparison isolates intent conditioning at the retrieval scoring level only. The modest aggregate effect of intent-conditioned retrieval weighting does not mean that intent awareness is unimportant to the system. It means that the generation prompts already capture most of the intent-specific benefit, and that adding intent-specific retrieval weights provides incremental rather than transformative improvement.

5.1.3 Hybrid Retrieval as the Strongest Single Component

Hybrid retrieval produces the largest single-component attribution gain (+0.041 over dense-only retrieval) and achieves perfect alignment across all 40 queries. This finding is consistent with the hybrid retrieval literature. Cormack et al. (2009) demonstrate that reciprocal rank fusion consistently outperforms individual retrieval methods, and Lu et al. (2022) confirm that hybrid retrieval gains are robust across domains where documents vary in vocabulary and structure.

The specific mechanism is interpretable. BM25 surfaces exact article numbers, defined legal terms, and date references that dense retrieval underweights because they do not contribute strongly to the overall semantic representation. For temporal queries, BM25 retrieves provisions dated to specific legislative sessions. For comparative queries, it

retrieves provisions from both compared regulations by matching their names as keywords. These are precisely the cases where dense retrieval's semantic smoothing produces retrieval drift, returning passages that are topically related but lack the specific terms the query requires.

The key term decrease (−0.030) that accompanies hybrid retrieval warrants discussion. BM25 promotes chunks containing exact query terms, which may displace semantically relevant chunks that use different vocabulary but contain domain-specific terminology that the evaluation's key term list expects. This suggests a tension between lexical precision, which BM25 provides, and terminological coverage, which the key term metric measures. The practical implication is that hybrid retrieval optimizes for retrieval precision and source grounding at the expense of vocabulary breadth.

5.1.4 The Agentic Loop as an Attribution Amplifier

The agentic loop yields the highest attribution among all system configurations (0.642), surpassing Hybrid by 0.054 and the standard RAG baseline by 0.086. This improvement is observed across all four intent types, with the most substantial increase occurring in factual queries (+0.150 over Hybrid). The observed diversity gain (+0.019) indicates that iterative retrieval uncovers additional source types that are not accessed in a single retrieval pass.

The alignment cost (−0.025) primarily affects factual queries, with alignment decreasing from 1.000 to 0.900. This trend indicates that additional chunks introduced by subsequent iterations reduce the structural coherence of factual summaries. When the initial retrieval pass already identifies the relevant legal provision, further iterations introduce tangentially related content, such as speeches discussing the provision and news articles providing context. The summarizer integrates this supplementary material in ways that deviate from the expected factual format.

The sufficiency checker is intended to prevent unnecessary iterations by terminating the loop when the accumulated evidence is deemed adequate. In this evaluation, it terminates 70% of queries after a single iteration, with stance queries exhibiting the highest early-termination rate (9 out of 10). This outcome is appropriate, as stance queries do

not benefit from additional iterations, as demonstrated by the per-intent results. However, the checker does not prevent the degradation of factual alignment described previously. This limitation arises because the sufficiency assessment determines only whether more relevant content exists in the corpus, which is often the case for factual queries. For any given legal provision, there are typically related speeches, news articles, and bill texts. While the checker accurately identifies this material as retrievable, it does not assess whether incorporating it would enhance the summary. The initial retrieval already provides the relevant legal text, and subsequent passes introduce contextual material that diminishes the factual structure. This reveals a gap in the sufficiency assessment: the checker evaluates the existence of additional evidence but does not consider whether its inclusion would improve summary quality for the specific intent type.
The calibration of the sufficiency prompt constitutes a key design parameter that directly influences system performance. The prompt instructs the language model to determine whether the main aspects of the question can be addressed using the retrieved sources, with a confidence threshold set at 0.65. A more permissive prompt would preserve the number of iterations and maintain alignment, but at the expense of attribution. Conversely, a stricter prompt would increase iterations and enhance attribution but reduce alignment and increase computational costs. The threshold selected in this implementation corresponds to a specific point in this trade-off space. The sensitivity of system performance to prompt calibration is itself a significant finding for the broader design of agentic RAG systems. This trade-off is a well-recognized challenge in iterative RAG, as observed by Gao et al. (2023), that multi-step retrieval can introduce noise when the initial retrieval is already sufficient.

5.1.5 The Research Question

The research question investigates whether intent-aware agentic retrieval with source-type weighting enhances query-guided summarization quality compared to a standard RAG pipeline. The full system demonstrates the highest attribution among all evaluated configurations (0.642), surpassing the standard RAG baseline by 15.5%, offering a distinct quality profile that prioritizes evidence grounding over structural consistency. This

represents a substantial improvement. Attribution assesses the extent to which the summary's claims are traceable to cited sources, a critical quality dimension for systems designed to support legislative research. Researchers relying on the system's output must verify claims against the underlying documents, and higher attribution facilitates this verification process. This improvement incurs a modest reduction in alignment (−0.025) and key term coverage (−0.028), while overall coverage remains essentially unchanged (+0.001).

The ablation structure reveals how each component contributes to this overall profile. Hybrid retrieval yields the most consistent improvements, achieving perfect alignment, the highest coverage, and strong attribution, making it the most robust general-purpose configuration. The agentic loop further increases attribution but introduces an alignment trade-off, particularly affecting factual queries. Intent-conditioned weighting directs retrieval toward appropriate source types for each query intent, yet does not yield clear aggregate gains over the unweighted baseline. The system achieves intent-appropriate summarization by combining intent-conditioned generation prompts with intent-specific retrieval mechanisms. The generation prompts are the primary driver of alignment, producing intent-appropriate output structure from Level 2 onwards. The retrieval mechanisms (source-type weighting, hybrid fusion, and agentic iteration) shape which evidence reaches the summarizer, and the ablation demonstrates that these mechanisms contribute differently across query types. Hybrid retrieval provides broad benefits across all intents. The agentic loop provides targeted benefits for temporal queries while introducing costs for factual queries. Intent-conditioned retrieval weighting shapes the source-type composition of the retrieval window without producing clear aggregate gains. Together, these findings confirm the thesis's central premise that different legislative queries require different retrieval strategies, while revealing that the strongest single intervention is the combination of hybrid retrieval with intent-conditioned generation rather than intent conditioning at every pipeline stage.

The recommended general-purpose configuration for multi-source legislative retrieval is Hybrid (Level 5). The agentic loop should be applied selectively in scenarios where evidence grounding is prioritized over structural consistency, especially for temporal

queries that require evidence from multiple time periods. The optimal configuration depends on the specific quality dimension prioritized. For alignment and consistency, Hybrid is preferable. For attribution and source traceability, the full system with the agentic loop offers measurable improvements.

5.2 Validity of Automated Evaluation Metrics

The human validation exercise provides an opportunity to assess how closely these automated metrics align with human judgment. The results reveal both partial correspondence and systematic discrepancies.

The positive Spearman correlations between human and automated scores for attribution (rho = +0.269) and alignment (rho = +0.330) suggest that the automated metrics capture aspects similar to what humans look for. For example, when the automated alignment score is low (S07, auto alignment 0.333), the human alignment score is also below average (2.33). When automated attribution is low (S04, auto attribution 0.333), the human attribution score is moderate (2.67). Still, these correlations are not statistically significant given only 12 samples, and the relationship does not always hold, especially when automated metrics miss important details.

The clearest example of a mismatch is S05, a stance query about AI regulation from the Full system. The automated attribution score is 1.000 because the summary lists source-type labels (S&D, PPE, Greens/EFA) and MEP names found in the retrieved text. However, all three human raters gave attribution a score of 1 out of 3, the lowest possible. They explained that "none of these claims were backed with citations." The summary names the MEPs and describes their views, but it does not include citation brackets linking each claim to a specific speech document. The automated metric only checks if source identifiers are present in the text. Human raters look for claims that are clearly tied to verifiable sources. These are similar, but not identical, and the automated metric cannot distinguish between simply mentioning a source and formally citing it.

A second discrepancy is found in S12, a factual question about prohibited AI practices. The automated attribution and alignment scores are both 1.000 because the summary cites Article 5 of the AI Act with a formal citation bracket and uses factual structural

markers. However, the human relevance score is 1.67. One rater noted that the response "acknowledges there are prohibitions but fails to mention them." This observation is confirmed by the summary itself, which explicitly states that "the specific prohibited AI practices are not fully listed in the provided source." The summarizer correctly identified the relevant legal provision and correctly recognized that the retrieved chunks did not contain the full article text. The failure is in retrieval, not generation. The chunks that reached the summarizer referenced Article 5 but did not include the enumerated list of prohibited practices. The automated metrics score this output highly because the formal properties are correct. The citation is present, the structure is factual, and the article number is accurate. But the substantive information the query requested is absent, and the automated metrics cannot detect this.

A third discrepancy is seen in S06, a comparative question about the differences between NIS and NIS2. The automated alignment score is 1.000 because the summary includes the four-part comparative structure. However, the human relevance score is 1.33, the lowest among all 12 summaries. All three raters noted that the summary discusses similarities and individual provisions of each directive but does not highlight the differences between them. This pattern reflects a limitation of the comparative prompt template, which applies the same four-part structure to all comparative queries regardless of what they specifically ask. The template begins with a shared scope, which encourages the LLM to discuss similarities before differences. For a query that explicitly asks about differences, this structure works against the information need. A more adaptive template that adjusts the section ordering based on the specific comparative question would address this problem. The automated alignment metric detects the structural markers but cannot assess whether the structure serves the query.

These three cases reveal a consistent limitation. Automated metrics focus on the formal aspects of the summary, including source identifiers, structural markers, and domain-specific terms. However, they do not verify whether the content within these structures addresses the query. A summary might achieve perfect automated scores by citing the correct regulation, following the expected structure, and using relevant terms, yet still fail to provide the requested information. This gap between meeting formal

requirements and delivering substantive answers is not unique to this thesis. It is a broader issue in the reference-free evaluation of generative systems, where automated metrics are necessary but insufficient to judge output quality. This gap has two sources. First, the summarization prompts define the expected output structure but do not require the model to verify that the requested information is present in its response. A prompt that additionally instructs the model to check whether it has enumerated the specific items the query asks for would reduce cases like S12, where the model acknowledges a gap but does not attempt to fill it. Second, the automated metrics are designed to measure prompt compliance rather than answer completeness. They check whether the structural markers and source identifiers required by the prompts are present in the output. Improving both the prompts to include content verification instructions and the metrics to assess whether the query's specific information need is addressed would narrow this gap. Both are identified as future work.

The low inter-rater agreement (Fleiss' kappa 0.046 overall) weakens the conclusions that can be made from the human validation. Disagreement is influenced by the limited three-point scale, the specialized nature of EU legislative content, and the inherent difficulty of assessing legal summary quality without domain-specific training. A stronger validation would involve domain experts, a five-point scale with clear rubrics, and a larger set of summaries. These improvements are planned for future work.

Even with these limitations, the human validation achieves its main goal. It finds specific problems that automated metrics miss, such as citations without proper attribution, summaries that follow structure but lack substance, and comparisons that are not actually made. These results help explain how to interpret the automated results in Chapter 4. High automated scores indicate that the system produces summaries that are well-structured, well-cited, and use the appropriate terms. However, they do not ensure that the summaries always meet the user's information needs.

5.3 Limitations

The study has several limitations arising from the evaluation process itself, beyond the scope outlined in Section 1.4.

- Inter-rater agreement was low (Fleiss' kappa 0.046) as reported in Section 4.6, limiting the strength of conclusions that can be drawn from the human validation exercise. The exercise nonetheless reveals that the automated metrics capture formal properties of summary quality but not substantive adequacy, as demonstrated by the failure modes identified in Section 5.2. High automated scores are, therefore, necessary but not sufficient indicators of summary quality.
- The sufficiency checker's behavior depends on the formulation of the prompt and the confidence threshold. The current calibration (balanced prompt requesting assessment of whether "main aspects" can be addressed, threshold 0.65) produces early termination for 70% of queries. A different prompt formulation would produce different iteration patterns and potentially different system performance. The sensitivity of system performance to this single design parameter was not anticipated before the evaluation and represents a finding about the design space of agentic RAG systems.
- The interaction between hierarchical chunking and corpus-frequency weighting was discovered during analysis rather than anticipated in the system design. Hierarchical chunking splits each law into individual articles, producing approximately 188 chunks per law document. This inverts the document-level class distribution at the chunk level, making laws and bills the majority class. The inverse-frequency formula then assigns them the lowest weights, which is the opposite of the intended correction. This interaction would affect any system applying frequency-based weighting to a corpus with heterogeneous document lengths and should be tested for explicitly during system design.

.

6 Conclusion and Future Work

This thesis investigated whether an intent-aware agentic retrieval system with source-type weighting can improve query-guided summarization over a heterogeneous EU legislative corpus compared with a standard RAG pipeline. Using a six-level ablation study on 40 benchmark queries across four intent types and six legislative domains, the results showed that the system's components contribute in different ways rather than producing a single uniformly superior configuration.

The full system achieved the strongest attribution, producing summaries with more traceable claims and clearer grounding in legislative sources. This benefit, however, came with small reductions in alignment and key-term coverage, indicating that improved evidence traceability does not automatically yield better structural fit or terminology preservation. Hybrid retrieval offered the most balanced performance: it achieved perfect alignment across all queries, maintained the highest coverage, and still delivered strong attribution. These findings suggest that hybrid retrieval should serve as the default configuration for multi-source legislative corpora, while the agentic loop is most valuable when additional evidence gathering justifies a modest trade-off in summary consistency.

Beyond the performance results, the thesis identified several broader design lessons for retrieval-augmented generation in heterogeneous corpora. The interaction between chunking and frequency-based weighting shows that retrieval corrections must account for how documents are segmented, not only how often they appear in the corpus. The behavior of the sufficiency checker demonstrates that prompt formulation is not a minor implementation detail but a mechanism that directly shapes when and how the agentic loop retrieves additional evidence. The comparison between automated metrics and human judgements further shows that while formal metrics capture structural and citation-related properties, they do not fully measure whether a summary substantively answers the query.

Several directions for future work follow from these findings. A larger, more evenly distributed benchmark would support stronger statistical testing and more reliable intent-specific comparisons. LLM-based reranking is a promising next step, as it could improve

both attribution and alignment by selecting chunks more precisely according to the query's evidential needs. Future research should also explore selective use of the agentic loop, triggering iteration only for query types that genuinely benefit from it. The sufficiency checker warrants more systematic study by varying prompt wording and confidence thresholds to understand how these settings influence retrieval depth and summary quality. A more robust evaluation could combine automated metrics with domain-expert review or hybrid scoring methods that better capture substantive adequacy.

References

Abazari Kia, M. (2021). Automated Multi-document Text Summarization from Heterogeneous Data Sources. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12657 LNCS*, 667–671. https://doi.org/10.1007/978-3-030-72240-1_78

Abdallah, A., Piryani, B., Mozafari, J., Ali, M., & Jatowt, A. (2025). *How Good are LLM-based Rerankers? An Empirical Analysis of State-of-the-Art Reranking Models*. https://github.com/DataScienceUIBK/

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270. https://doi.org/10.1007/s42001-019-00060-w

Alanzi, E., & Alballaa, S. (n.d.). Query-Focused Multi-document Summarization Survey. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Number 6). Retrieved www.ijacsa.thesai.org

Ali, Șener, Oprea, S. V., & Bâra, A. (2026). Engineering Trustworthy Retrieval-Augmented Generation for EU Electricity Market Regulation. *Electronics (Switzerland)*, *15*(4). https://doi.org/10.3390/electronics15040749

Arabzadeh, N., Yan, X., & Clarke, C. L. A. (2021). *Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection*. http://arxiv.org/abs/2109.10739

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia Computer Science*, *246*(C), 3781–3790. https://doi.org/10.1016/j.procs.2024.09.178

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). *SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION*. https://selfrag.github.io/.

Ayaou, I., Cavallucci, D., & Chibane, H. (2026a). DAPFAM: A Domain-Aware Family-level Dataset to benchmark cross domain patent retrieval. *Array*, *29*. https://doi.org/10.1016/j.array.2026.100720

Ayaou, I., Cavallucci, D., & Chibane, H. (2026b). DAPFAM: A Domain-Aware Family-level Dataset to benchmark cross domain patent retrieval. *Array*, *29*. https://doi.org/10.1016/j.array.2026.100720

Brasoveanu, A., Moodie, M., & Agrawal, R. (2020). Textual evidence for the perfunctoriness of independent medical reviews. *CEUR Workshop Proceedings*, *2657*, 1–9. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Brown, A., Roman, M., & Devereux, B. (2025). A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. In *Big Data and Cognitive Computing* (Vol. 9, Number 12). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/bdcc9120320

Busolin, F., Lucchese, C., Nardini, F. M., Orlando, S., Perego, R., Trani, S., & Veneri, A. (2025). Efficient Re-ranking with Cross-encoders via Early Exit. *SIGIR 2025 - Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2534–2544. https://doi.org/10.1145/3726302.3729962

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. https://github.com/FlagOpen/FlagEmbedding.

Cherumanal, S. P., Scholer, F., & Spina, D. (2026). Diversification and Fairness in Search: Two Sides of the Same Coin? *ACM Transactions on Information Systems*, *44*(2), 1–31. https://doi.org/10.1145/3774320

Chihaia, T., & Ciobanu, R. I. (2025). Keyword vs Semantic Search for Retrieval-Augmented Generation: A Survey. *Proceedings - 2025 25th International Conference on Control Systems and Computer Science, CSCS 2025*, 169–174. https://doi.org/10.1109/CSCS66924.2025.00033

Chouhan, A., & Gertz, M. (2024). *LexDrafter: Terminology Drafting for Legislative Documents using Retrieval Augmented Generation*. https://lawdroid.com/,

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). *Class-Balanced Loss Based on Effective Number of Samples*.

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 6437–6447. https://doi.org/10.1145/3637528.3671458

Dang, H. T. (2005). *Overview of DUC 2005*. http://www.isi.edu/

D'Asaro, F., De Luca, S., Bongiovanni, L., Rizzo, G., Papadopoulos, S., Schinas, M., & Koutlis, C. (2024). Zero-Shot Content-Based Crossmodal Recommendation System. *Expert Systems with Applications*, *258*. https://doi.org/10.1016/j.eswa.2024.125108

Doshi, M., Kumar, V., Murthy, R., P, V., & Sen, J. (2024). *Mistral-SPLADE: LLMs for better Learned Sparse Retrieval*. http://arxiv.org/abs/2408.11119

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Fernando, K. R. M., & Tsokos, C. P. (2022). Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2940–2951. https://doi.org/10.1109/TNNLS.2020.3047335

Formal, T., Louis, M., Déjean, H., & Clinchant, S. (2026). *Naver Labs Europe @ WSDM CUP | Multilingual Retrieval*. http://arxiv.org/abs/2602.20986

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. http://arxiv.org/abs/2312.10997

Goel, R., Kumar, S. P., Agrawal, A., Poddar, D., Narang, P., & Kumar, D. (2025). *Domain-Partitioned Hybrid RAG for Legal Reasoning: Toward Modular and Explainable Legal AI for India*. http://arxiv.org/abs/2602.23371

Gordon V. Cormack, Charles L A Clarke, & Stefan Buettcher. (2009). *SIGIR '09 : the 32nd international ACM SIGIR conference on research and development in information*

retrieval, Boston, MA, USA, 19-23.07.2009. A.C.M. https://doi.org/https://doi.org/10.1145/1571941.1572114

Gupta, S., & Ranjan, R. (n.d.). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions*.

Haag, M., Hurka, S., & Kaplaner, C. (2025). Policy complexity and implementation performance in the European Union. *Regulation and Governance*, *19*(3), 656–674. https://doi.org/10.1111/rego.12580

Jadhav, C. D., Liu, C., & Zhao, J. (2025). *Legal Retrieval Augmented Generation with Structured Retrieval and Iterative Refinement*. 1–9. https://doi.org/10.1109/pst65910.2025.11268884

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). *Active Retrieval Augmented Generation*. https://github.com/

Johnson, J., Douze, M., & Jegou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. http://arxiv.org/abs/2004.04906

Katragadda, R., & Varma, V. (2009). Query-Focused Summaries or Query-Biased Summaries ? In *ACL and AFNLP*.

Kos, B., Tuković, L., Vlainić, E., & Babac, M. B. (2026). Unveiling ideological extremes in parliamentary debates using transformer-based language models. *Journal of Information Technology & Politics*, 1–21. https://doi.org/10.1080/19331681.2026.2616668

Lassance, C., & Clinchant, S. (2022). An Efficiency Study for SPLADE Models. *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2220–2226. https://doi.org/10.1145/3477495.3531833

Leetaru, K. S. P. (2013, April). GDELT: Global Data on Events, Language, and Tone, 1979-2012. *International Studies Association Annual Conference*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. http://arxiv.org/abs/2005.11401

Li, Y., Xie, M., Kang, G., Gong, Z., Wu, G., & Yang, M. (2026). *LegalMALR:Multi-Agent Query Understanding and LLM-Based Reranking for Chinese Statute Retrieval*. http://arxiv.org/abs/2601.17692

Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*.

Liu, Y., Wang, Z., & Yuan, R. (2024). *QuerySum: A Multi-Document Query-Focused Summarization Dataset Augmented with Similar Query Clusters*. www.aaai.org

Loffredo, J. R., & Yun, S. (2025). *Agent-Enhanced Large Language Models for Researching Political Institutions*. https://doi.org/10.1561/113.00000125

Lovely Yeswanth, P., Satya Krishna, N., Lavanya, P., Sriya, P., & Yogeshvar Reddy, K. (2026). *Modified Cross Encoder for Two-Stage Passage Ranking* (pp. 41–53). https://doi.org/10.1007/978-981-96-8898-2_4

Lu, J., Ma, J., & Hall, K. (2022). *Zero-shot Hybrid Retrieval and Reranking Models for Biomedical Literature*. https://pubmed.ncbi.nlm.nih.gov/

Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining Text Data* (Vol. 9781461432234, pp. 43–76). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_3

Nguyen, H. T., & Satoh, K. (2024). ConsRAG: Minimize LLM Hallucinations in the Legal Domain. *Frontiers in Artificial Intelligence and Applications*, *395*, 327–332. https://doi.org/10.3233/FAIA241263

Nogueira, R., & Cho, K. (2020). *Passage Re-ranking with BERT*. http://arxiv.org/abs/1901.04085

Pandit, T., Mahendru, S., Raval, M., & Upadhyay, D. (2026). *The Evolution of Reranking Models in Information Retrieval: From Heuristic Methods to Large Language Models* (pp. 56–67). https://doi.org/10.1007/978-981-95-4788-3_6

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Roy, P., & Kundu, S. (2024). Review on Query-focused Multi-document Summarization (QMDS) with Comparative Analysis. *ACM Computing Surveys*, *56*(1). https://doi.org/10.1145/3597299

Rustamov, Z., Gasimzade, M., & Rustamov, S. (2025). *Enhancing Retrieval and Re-ranking in RAG: A Case Study on Tax Law*. 1–7. https://doi.org/10.1109/aict67988.2025.11268677

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). *Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers*. https://doi.org/10.1109/MIPR62202.2024.00031

Schwalbach, J., Hetzer, L., Proksch, S.-O., Rauh, C., & Sebők, M. (2025). *ParlLawSpeech Codebook*. https://parllawspeech.org/

Sharma, P., & Bhattarai, S. (2026). A Review on Retrieval-Augmented Generation: Architectures, Research Challenges, and Emerging Frontiers. *Journal of Future Artificial Intelligence and Technologies*, *2*(4), 616–628. https://doi.org/10.62411/faith.3048-3719-297

Shi, M.-X., Pan, T.-H., Huang, H.-H., & Chen, H.-H. (n.d.). *Hybrid Re-ranking for Biomedical Information Retrieval at the TREC 2021 Clinical Trials Track*. Retrieved http://www.trec-cds.org/2021.html

Song, Y., Yan, L., Qin, L., Wang, G., Huang, X., Hu, L., & Liu, W. (2025). URAG:Unified Retrieval-Augmented Generation. *ICCIP 2024 - 2024 the 10th International Conference on Communication and Information Processing*, 660–667. https://doi.org/10.1145/3708657.3708763

Sood, A., Singh, M., Gardiner, B., & Condell, J. (2026). MeetMulti-X: A benchmark analysis of scaling and prompting large language models on automatic minuting. *Expert Systems with Applications*, *303*. https://doi.org/10.1016/j.eswa.2025.130428

Stamatis, V., Salampasis, M., & Diamantaras, K. (2024). A novel re-ranking architecture for patent search. *World Patent Information*, *78*. https://doi.org/10.1016/j.wpi.2024.102282

Sun, Y. (2025). *Adaptive Hybrid Retrieval-Augmented Generation with Context-Aware Confidence Control*. 891–895. https://doi.org/10.1109/eit67313.2025.11231898

Thenmozhi, D., AG, M. V., & M, N. S. (2025). Hierarchical knowledge graph based legal information retrieval from multiple documents using intelligent agents. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-025-09491-5

Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The debates of the European Parliament as Linked Open Data. *Semantic Web*, *8*(2), 271–281. https://doi.org/10.3233/SW-160227

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report*. http://arxiv.org/abs/2402.05672

Wang, T., Weng, T., Ji, J., Zhong, M., & Zhang, B. (2022). A Text Multi-label Classification Scheme Based on Resampling and Ensemble Learning. *Communications in Computer and Information Science*, *1587 CCIS*, 67–80. https://doi.org/10.1007/978-3-031-06761-7_6

Xanthaki, H. (2025). *Legislative complexity and monitoring the application of EU law*. https://www.europarl.europa.eu/committees/en/supporting-analyses/sa-highlights

Xu, Z., Feng, A., Tian, Y., Ding, H., & Cheong, L. L. (2025). *CSPLADE: Learned Sparse Retrieval with Causal Language Models*. http://arxiv.org/abs/2504.10816

Ya, J., Liu, T., & Guo, L. (2020). A Compare-Aggregate Model with External Knowledge for Query-Focused Summarization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12343 LNCS*, 68–83. https://doi.org/10.1007/978-3-030-62008-0_5

Yin, W., Hay, J., & Roth, D. (2019). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach*. https://cogcomp.seas.upenn.edu/page/

Yu, J., Xu, Y., Li, H., Li, J., Feng, Y., Zhu, L., Shen, H., & Shi, L. (2025). *OPOR-Bench: Evaluating Large Language Models on Online Public Opinion Report Generation*. http://arxiv.org/abs/2512.01896

Zhang, W., Huang, J. H., Vakulenko, S., Xu, Y., Rajapakse, T., & Kanoulas, E. (2025). Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models. *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) , *15319 LNCS*, 89–104. https://doi.org/10.1007/978-3-031-78495-8_6

Zhao, W. X., Liu, J., Ren, R., & Wen, J. R. (2024). Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Transactions on Information Systems*, *42*(4). https://doi.org/10.1145/3637870

Appendices

Appendix 1. Benchmark Queries

The following 40 benchmark queries were designed by the author to cover all four intent types and six EU legislative domains. Each set of 10 queries targets one intent type. Domain assignments for comparative queries reflect the primary acts being compared.

1.1 Factual Queries

ID	Domain	Query
F01	GDPR	What does Article 17 of the GDPR say about the right to erasure?
F02	DMA	What is the definition of a gatekeeper under the Digital Markets Act?
F03	AI Act	What obligations does the AI Act place on high-risk AI systems?
F04	DSA	What does the DSA require from very large online platforms?
F05	GDPR	What penalties does the GDPR impose for non-compliance?
F06	Copyright DSM	What exceptions does the Copyright DSM Directive provide for text and data mining?
F07	NIS2	What cybersecurity risk management obligations does the NIS2 Directive impose on essential entities?
F08	AI Act	What does the AI Act define as prohibited artificial intelligence practices?
F09	DMA	What transparency obligations does the DMA impose on gatekeepers regarding advertising?
F10	DSA	What rules does the DSA establish for recommender systems used by online platforms?

1.2 Temporal Queries

ID	Domain	Query
T01	GDPR	How did the GDPR proposal change between 2012 and 2016?
T02	NIS2	How has the EU approach to cybersecurity evolved since 2013?
T03	Copyright DSM	What shifted in the Copyright Directive debate after 2017?
T04	AI Act	How did the AI Act evolve from the 2021 Commission proposal to the final text?

T05	DSA	How did parliamentary positions on the DSA change during negotiations?
T06	DMA	How did the debate on gatekeeper regulation evolve before the DMA was adopted?
T07	GDPR	How did EU data protection rules change from the 1995 Directive to the GDPR?
T08	AI Act	How did the scope of the AI Act expand during the legislative process?
T09	DSA	How did the enforcement provisions of the DSA develop during negotiations?
T10	Copyright DSM	How did the Copyright Directive's provisions on press publishers evolve from proposal to adoption?

1.3 Stance Queries

ID	Domain	Query
S01	Copyright DSM	Which political groups opposed Article 13 of the Copyright Directive?
S02	AI Act	Who supported stronger AI regulation in the European Parliament?
S03	DMA	What was the EPP position on the Digital Markets Act?
S04	GDPR	Which MEPs or parties opposed the GDPR in the European Parliament?
S05	NIS2	How did the Greens group approach NIS2 cybersecurity legislation?
S06	DSA	What was the S&D group's position on the Digital Services Act?
S07	AI Act	Which MEPs argued the AI Act was too restrictive for innovation?
S08	DMA	How did the ECR group approach the Digital Markets Act?
S09	GDPR	What positions did MEPs take on the right to be forgotten during the GDPR debate?
S10	NIS2	Which political groups called for stronger cybersecurity measures during the NIS2 debate?

1.4 Comparative Queries

ID	Domain	Query
C01	DSA / DMA	How does the DSA differ from the DMA in scope and obligations?
C02	AI Act	Compare the Commission proposal for the AI Act with the final adopted text.

C03	GDPR	How do MEP speeches on GDPR compare to what the final law actually says?
C04	NIS2	What are the differences between the NIS and NIS2 directives?
C05	Copyright DSM	How does the Copyright DSM Directive compare to previous EU copyright law?
C06	GDPR / DSA	How do the penalty frameworks of the GDPR and the DSA compare?
C07	DMA / DSA	How does the DMA's approach to interoperability compare to the DSA's approach to transparency?
C08	GDPR / AI Act	Compare the GDPR's consent requirements with the AI Act's transparency obligations.
C09	Copyright DSM / DSA	What are the differences between the Copyright DSM Directive's platform obligations and those in the DSA?
C10	NIS2 / GDPR	How does the NIS2 Directive's incident reporting compare to the GDPR's data breach notification?

Appendix 2. Sample System outputs

Three representative system outputs are provided below. Each is referenced in the Discussion (Chapter 5) to illustrate the relationship between automated metrics and summary quality. Automated scores are reported for coverage (Cov), alignment (Align), attribution (Attr), and key terms (KT).

2.1 Strong Stance Output(S05)

Query*:* Who supported stronger AI regulation in the European Parliament?

System: Full (L6) | Intent: Stance | Cov: 0.245 | Align: 1.000 | Attr: 1.000 | KT: 0.800

```
Based on the provided speeches, the positions of the major political
groups on stronger AI regulation in the European Parliament are:

1. S&D (Socialists and Democrats): supported — The S&D Group's shadow
rapporteur, Miapetra Kumpula-Natri, stated that the AI Act is an
important step towards building a society that respects fundamental
rights and individual freedoms, and that it creates a framework for
ethical and trustworthy AI development and deployment.

2. PPE (European People's Party): supported — Deirdre Clune from the
PPE Group welcomed the final adoption of the Artificial Intelligence
Act, stating that it provides a robust regulatory framework, promotes
trust and transparency in AI systems, and upholds fundamental rights
in the digital age.

3. Greens/EFA (The Greens/European Free Alliance): supported — Ser-
gey Lagodinsky from the Greens/EFA Group stated that the AI Act is
a timely attempt to control future developments in AI, and that
while it is weaker than desired, it is better than no regulation at
all and pioneers work in many areas.

Note: The positions of other groups, such as the ECR (European
Conservatives and Reformists), Renew Europe, and The Left, are un-
clear due to insufficient evidence from the provided speeches.
```

2.2 Factual Alignment Degradation (S07)

Query: What penalties does the GDPR impose for non-compliance?

System: BM25 (L1) | Intent: Factual | Cov: 0.265 | Align: 0.333 | Attr: 0.333 | KT: 0.600

Note: The system produced a four-part comparative structure instead of a factual response. BM25 retrieved both the GDPR proposal and a Council Decision, and the LLM interpreted the dual-document context as a comparative query.

```
. SHARED SCOPE: The GDPR imposes penalties for non-compliance with
its provisions, aiming to protect the rights of data subjects and
ensure the proper processing of personal data [BILL | GDPR | 2012-
01-25]. The regulatory objective is to enforce compliance with the
GDPR's requirements, such as providing mechanisms for data subject
requests and responding promptly to these requests [BILL | GDPR |
2012-01-25].

2. FIRST ACT ONLY: Article 79 of the GDPR covers the imposition of
administrative sanctions, including fines up to 250,000 EUR or 0.5%
of an enterprise's annual worldwide turnover for intentionally or
negligently failing to provide mechanisms for data subject requests
[BILL | GDPR | 2012-01-25]. Additionally, it imposes fines up to
500,000 EUR or 1% of annual worldwide turnover for intentionally or
negligently providing incomplete information to data subjects or
failing to comply with the right to be forgotten [BILL | GDPR |
2012-01-25].

3. SECOND ACT ONLY: The second act, a proposal for a Council Decision,
does not specifically address penalties for non-compliance with the
GDPR, but focuses on the adoption of a decision by the EEA Joint
Committee to ensure the homogeneity of the Internal Market [BILL |
GDPR | 2018-05-02].

4. MOST SIGNIFICANT DISTINCTION: The most significant distinction
is that Article 79 of the GDPR specifically outlines the administra-
tive sanctions for non-compliance, whereas the Council Decision does
not address penalties at all, instead focusing on the implementation
and adoption of the GDPR in the EEA [BILL | GDPR | 2012-01-25] [BILL
| GDPR | 2018-05-02].
```

3.3 Strong Comparative Output (S11)

Query: How does the DSA differ from the DMA in scope and obligations?

System: Hybrid (L5) | Intent: Comparative | Cov: 0.209 | Align: 1.000 | Attr: 1.000 | KT: 1.000

```
. SHARED SCOPE: Both the Digital Services Act (DSA) and the Digital
Markets Act (DMA) address the regulation of digital services in the
European Union, with a common objective of promoting fair competition,
transparency, and accountability in the digital market [LAW | DSA |
Regulation (EU) 2022/2065 | 2022-10-19]. They share the goal of
protecting users and promoting a safe online environment [LAW | DSA
| 2022-10-19].
```

2. FIRST ACT ONLY (DSA): The DSA covers the regulation of intermediary services, including mere conduit, caching, and hosting, with specific obligations such as the exemption from liability for neutral intermediaries (Article 19, Recital 19) [LAW | DSA | 2022-10-19]. Additionally, the DSA introduces a supervisory fee for competent authorities to cover the costs of fulfilling their tasks (Recital 112) [LAW | DSA | 2022-10-19].

3. SECOND ACT ONLY (DMA): The DMA focuses on regulating gatekeepers, which are large online platforms with significant market power, and imposes specific obligations such as data sharing and interoperability requirements [BILL | DMA]. The DMA also introduces a set of dos and don'ts for gatekeepers, including prohibitions on self-preferencing and requirements for transparency in online advertising [BILL | DMA].

4. MOST SIGNIFICANT DISTINCTION: The most significant distinction is the scope of their regulatory objectives, with the DSA focusing on intermediary services and the DMA targeting large online platforms with significant market power (gatekeepers) [LAW | DSA | 2022-10-19]. This difference results in distinct enforcement mechanisms, with the DSA relying on Digital Services Coordinators and the DMA on the European Commission's enforcement powers [LAW | DSA | 2022-10-19].

Appendix 3. Summarization Prompt Templates

The following four prompt templates are used in Stage 4 (intent-conditioned summarization). Each is selected based on the classified intent of the query. All prompts instruct the model to cite sources using bracketed labels matching the retrieved chunk headers. The model used is llama-3.3-70b-versatile with temperature 0.3 and a maximum output of 1,000 tokens.

3.1 Factual prompt

```
You are an EU legislative expert. Answer the question using only the
provided sources. Structure your answer as follows:
1. State what the relevant article or law requires or defines.
2. Provide the specific legal obligation, right, or definition.
3. Note any conditions, exceptions, or thresholds.

Use precise legal language: 'Article X states...', 'The law re-
quires...', 'This provision defines...', 'According to...', 'Under
the regulation...'. Always name the specific article number, section,
or law you are drawing from.

CITATION RULE: Every sentence making a factual claim MUST end with
a bracketed source label matching the source headers provided, e.g.
[LAW | GDPR], [BILL | AI_Act], [SPEECH | Copyright_DSM].
```

3.2 Temporal prompt

```
You are an EU legislative historian. In 4-6 sentences explain how
the topic evolved over time. Explicitly contrast early positions
with later ones. Mention specific changes, not just general state-
ments.
```

CITATION RULE: Every sentence making a factual claim MUST end with a bracketed source label matching the source headers provided, e.g. [LAW | GDPR], [BILL | AI_Act], [SPEECH | Copyright_DSM].

3.3 Stance prompt

```
You are an EU parliamentary analyst. Based on the provided speeches,
identify the positions of ALL major political groups on this topic.
You MUST cover at least three different groups if evidence exists.
For each group write:

[Group]: [supported/opposed/mixed] — [specific reason from speeches].
```

```
If a group's position is unclear from the sources, write: [Group]:
unclear — insufficient evidence. Always name the specific MEP or
speech where possible.
```

3.4 Comparative prompt

```
You are an EU policy analyst comparing two legislative acts. Struc-
ture your answer in four parts:
1. SHARED SCOPE: In 1-2 sentences, state what both acts address and
their common regulatory objectives.
2. FIRST ACT ONLY: What the first act covers that the second does
not — with specific obligations and article references.
3. SECOND ACT ONLY: What the second act covers that the first does
not — with specific obligations and article references.
4. MOST SIGNIFICANT DISTINCTION: The single most important diffe-
rence in scope, enforcement, or legal effect.

Prioritise differences over similarities — shared scope should be
brief context, not the main content. Reference the specific law or
article for every claim. If comparing a proposal with a final text,
focus on what was ADDED, REMOVED, or CHANGED between versions.

CITATION RULE: Every sentence making a factual claim MUST end with
a bracketed source label matching the source headers provided, e.g.
[LAW | GDPR], [BILL | AI_Act], [SPEECH | Copyright_DSM].
```