Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

QUERY-SENSITIVE RETRIEVAL-AUGMENTED GENERATION UNDER ORGANISATIONAL CONSTRAINTS IN THE REGULATORY DOMAIN

ANTONIA SOPHIE GEMMERICH

Lisbon School of Economics & Management
Universidade de Lisboa

FEBRUARY - 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
DISSERTATION

QUERY-SENSITIVE RETRIEVAL-AUGMENTED GENERATION UNDER ORGANISATIONAL CONSTRAINTS IN THE REGULATORY DOMAIN

ANTONIA SOPHIE GEMMERICH

SUPERVISION:
PROF. CARLOS J. COSTA
PROF. JOÃO T. APARICIO

FEBRUARY - 2026

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors, Prof. Carlos J. Costa and Prof. João T. Aparicio, for their guidance, expertise and patience from the initial idea to the completion of this dissertation. I have learned so much through our regular discussions, their constructive feedback, and the thoughtful challenges they posed, all of which pushed me to achieve the best possible outcome. I could not have asked for more dedicated supervision.

I am also grateful to my annotators for their time and for providing the reference sets upon which this analysis is based. In particular, I would like to thank my colleague Maria, whose regulatory expertise was invaluable in approaching the problem from the right perspective and who always made time to support and guide me.

A special thank you goes to Dominik for being my emotional support, my sparring partner, and for always believing in me. I am equally thankful to my friends for their encouragement, support, thoughtful discussions, and careful proofreading. Most importantly, I would like to thank my family, who have supported me throughout my journey in every way imaginable. I would not be where I am today without you.

GLOSSARY

MRR@k Mean Reciprocal Rank at rank k. 22, 25–27, 40

nDCG@k Normalized Discounted Cumulative Gain at rank k. 22, 25–28, 39

BM25 Best Matching 25. 17, 38

DSR Design Science Research. 3, 4

ECHA European Chemicals Agency. 16, 35

LLM Large Language Model. 1, 2, 9, 10, 18, 23

RAG Retrieval-Augmented Generation. 1–5, 7, 11–14, 19, 23, 34

RAGAS Retrieval-Augmented Generation Assessment. 23

REACH Registration, Evaluation, Authorisation and Restriction of Chemicals. 9

SLM Small Language Model. 12–14, 18, 19, 22–24, 27, 30

TF-IDF Term Frequency–Inverse Document Frequency. 17, 38

ABSTRACT

Retrieval-augmented generation (RAG) is increasingly adopted in organisational environments where strict data governance requirements limit the use of externally hosted large language models. In such settings, models must often be deployed locally and operate under computational and infrastructural constraints. While prior research typically evaluates retrieval-augmented systems under benchmark conditions with carefully formulated queries, far less is known about their robustness under realistic query variation and deployment restrictions. This work investigates how structured differences in query formulation influence retrieval quality and downstream generation behaviour in constrained settings. Queries are systematically varied across expert, natural, and naive formulations in order to examine how linguistic structure and domain knowledge shape system performance. The analysis distinguishes between retrieval adequacy and evidence utilisation, thereby enabling a focused assessment of how retrieved information is selected and incorporated into generated answers. The results show that query formulation constitutes a central design factor. Ranking quality declines substantially from expert to naive queries in several configurations, revealing pronounced sensitivity to input structure. Post-retrieval refinement exhibits conditional effectiveness rather than stable improvement. Moreover, high levels of retrieval recall do not consistently translate into correct and faithful answers. This utilisation gap indicates that the presence of relevant evidence alone is insufficient to guarantee reliable generation. Overall, the findings demonstrate that retrieval-augmented systems deployed under organisational constraints are structurally sensitive to query formulation. The study provides empirical evidence on robustness under realistic conditions and highlights the need for evaluation frameworks that account for deployment constraints and variation in user input rather than idealised benchmark scenarios.

KEYWORDS: Retrieval-Augmented Generation; Small Language Models; Query Formulation Robustness; Evidence Utilisation Gap; Governance Constrained AI; Local Model Deployment.

RESUMO

A geração aumentada por recuperação (RAG) é cada vez mais adotada em contextos organizacionais nos quais requisitos rigorosos de governação de dados limitam a utilização de modelos de linguagem de grande dimensão alojados externamente. Nestes cenários, os modelos têm frequentemente de ser implementados localmente e de operar sob restrições computacionais e infraestruturais. Embora a investigação anterior avalie tipicamente os sistemas aumentados por recuperação em condições de referência (*benchmark*), com consultas cuidadosamente formuladas, muito pouco se conhece acerca da sua robustez perante a variação realista das consultas e as restrições de implementação.

O presente trabalho investiga de que forma as diferenças estruturadas na formulação das consultas influenciam a qualidade da recuperação e o comportamento de geração subsequente em contextos condicionados. As consultas são sistematicamente variadas entre formulações especializadas, naturais e ingénuas, com o intuito de examinar de que modo a estrutura linguística e o conhecimento do domínio moldam o desempenho do sistema. A análise distingue entre a adequação da recuperação e a utilização da evidência, permitindo assim uma avaliação focada do modo como a informação recuperada é selecionada e incorporada nas respostas geradas.

Os resultados demonstram que a formulação das consultas constitui um fator central de conceção. A qualidade da ordenação diminui substancialmente das consultas especializadas para as ingénuas em diversas configurações, revelando uma acentuada sensibilidade à estrutura do *input*. O refinamento pós-recuperação manifesta uma eficácia condicional, em vez de uma melhoria estável. Acresce que níveis elevados de *recall* na recuperação nem sempre se traduzem em respostas corretas e fidedignas. Esta lacuna de utilização indica que a mera presença de evidência relevante é, por si só, insuficiente para garantir uma geração fiável.

Em síntese, os resultados evidenciam que os sistemas aumentados por recuperação implementados sob restrições organizacionais são estruturalmente sensíveis à formulação das consultas. O estudo fornece evidência empírica sobre a robustez em condições realistas e sublinha a necessidade de quadros de avaliação que contemplem as restrições de implementação e a variação no *input* do utilizador, em vez de cenários de referência idealizados.

PALAVRAS-CHAVE: Geração Aumentada por Recuperação (RAG); Modelos de Linguagem de Pequena Dimensão; Robustez da Formulação de Consultas; Lacuna na Utilização da Evidência; Inteligência Artificial sob Restrições de Governação; Implementação Local de Modelos.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

1 Introduction

1.1 Context

Organisations operating in high-stakes, knowledge-intensive domains increasingly face decisions that require the synthesis of complex domain knowledge to support well-justified reasoning based on authoritative sources. Decisions often need to be taken under significant time and resource constraints, while relevant knowledge is distributed, continuously updated, and frequently of an internalised experiential nature. This makes it difficult to access and validate decision-relevant information at the point of use. Errors may lead to financial loss, regulatory non-compliance, or risks for human health and the environment, with direct implications for organisational accountability.

Recent advances in Large Language Models (LLMs) have created new opportunities for natural-language-based information access and synthesis, as these models demonstrate strong generalisation and text understanding capabilities [1]. However, the non-transparent behaviour of general-purpose LLMs often lacks transparent and verifiable grounding in identifiable sources and remains prone to hallucinations [2]. Domain-specific fine-tuning improves reliability but remains computationally and operationally resource-intensive, limiting its feasibility in many organisational settings.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address these limitations by conditioning model outputs on retrieved external evidence rather than relying solely on static parametric knowledge. Empirical studies show that retrieval augmentation improves factual grounding and reduces unsupported generations when relevant evidence is available [3]. Beyond general benchmarks, retrieval-augmented approaches have also been validated in specialised, high-stakes domains characterised by technical terminology and strict evidentiary requirements for decision support [4].

Much of the existing literature, however, evaluates RAG systems under conditions of unconstrained deployment and assumes access to large, highly capable language models [5, 6]. As a result, limited insight is available into how retrieval-augmented systems behave when deployed under organisational constraints shaped by data governance policies.

1.2 Motivation

In professional practice, the integration of retrieval-augmented systems into decision-support workflows faces additional constraints. Decision-relevant information is frequently proprietary and non-public internal knowledge, reflecting strategic, or operational

considerations. Organisational data governance policies therefore often restrict or prohibit reliance on externally hosted LLMs for tasks involving sensitive information, even when technical security assurances are provided.

Consequently, the advantages of retrieval augmentation can only be leveraged under organisation-controlled deployment conditions that comply with internal governance requirements. As a result, model selection may be constrained in terms of scale and computational capacity.

These restrictions introduce open design questions. Retrieval-augmented systems are typically used by domain professionals whose queries naturally vary in specificity and technicality. Recent empirical work demonstrates that such systems exhibit measurable sensitivity to query-level variation, with even minor reformulations affecting retrieval quality and downstream generation behaviour [7]. Such findings indicate that variation in query formulation may propagate through retrieval and generation components, influencing overall system performance.

Prior work further shows that retrieval-augmented pipelines are inherently configuration-sensitive and can be adaptively tailored to query characteristics, for instance through query-aware retrieval strategy selection [8]. These findings highlight that RAG performance depends not only on model capability but also on system-level design choices.

While prior studies analyse robustness under benchmark-style conditions, it remains unclear how query-level sensitivity interacts with constrained deployment settings in which model capacity and computational resources are limited. This lack of clarity is especially problematic in high-stakes domains, where system usefulness depends less on peak performance under idealised conditions and more on robust and predictable behaviour in professional decision-support contexts. Therefore, understanding how retrieval and generation components interact under realistic deployment constraints and natural query variation becomes a central practical and research challenge.

1.3 Research Question and Objectives

Motivated by the problem setting outlined above, this thesis investigates the design and evaluation of RAG systems under organisation-controlled deployment constraints and natural variation in query formulation. The objective is to analyse system behaviour and to examine how retrieval and generation components can be configured under realistic governance and infrastructure limitations. The main research question is:

> *How can retrieval and generation components of retrieval-augmented systems be designed and evaluated to remain robust under natural variation in*

query formulation and organisation-controlled deployment constraints?

This overarching question is addressed through the following sub-questions:

RQ1 How does systematic variation in query formulation influence retrieval effectiveness within a locally deployable retrieval-augmented architecture?

RQ2 Under what conditions do post-retrieval refinement strategies improve or degrade retrieval effectiveness across varying query formulations?

RQ3 How do locally deployed small language models trade off answer correctness, faithfulness, and context grounding across query formulations when evaluated against multiple reference perspectives?

By examining these questions, this thesis analyses how query variation propagates through retrieval, refinement, and generation components when architectural design is restricted by governance-driven deployment constraints. This work thereby evaluates design trade-offs that emerge when retrieval-augmented systems are implemented using locally deployable models and assessed against heterogeneous reference answers. The findings inform the design of decision-support systems in organisation-controlled, knowledge-intensive environments and result in the following contributions:

- A modular end-to-end retrieval-augmented architecture based on locally deployable models, implemented and evaluated under strict deployment constraints.
- A systematic analysis of query formulation effects across retrieval, refinement, and generation components.
- A multi-perspective evaluation framework incorporating both human-annotated and model-generated reference answers, demonstrating how answer quality depends on interpretative scope and evaluation perspective in realistic organisational contexts.

1.4 Methodological Approach

This thesis follows a Design Science Research (DSR) paradigm as conceptualised by [9]. DSR is appropriate where the objective is not only to analyse a phenomenon, but to design and systematically evaluate an artefact in order to derive empirically grounded insights into design trade-offs under real-world constraints. This work develops and evaluates a RAG configuration tailored to organisation-controlled deployment environments under naturally varying query formulations.

The research process aligns with the core DSR activities. First, *problem identification and motivation* are derived from the challenge of enabling reliable retrieval-augmented language model systems in governance-constrained settings, precluding the use of externally hosted language models and therefore limiting model scale. In such contexts, retrieval robustness and downstream answer quality under varying query formulations remain insufficiently understood.

Second, the *objectives for a solution* are defined as the design of a locally deployable RAG configuration that (i) operates under explicit deployment constraints, (ii) allows systematic variation of query formulations, and (iii) enables structured evaluation of retrieval and answer generation performance in order to examine component interactions under constraint.

Third, in the *design and development* phase, a modular RAG framework is constructed. The artefact integrates multiple retrieval strategies, embedding model variants, post-retrieval configurations, and generation models. Its modular structure enables controlled experimentation across query types and retrieval settings to systematically explore design sensitivities.

Fourth, the artefact is *demonstrated* through systematic experimentation on a curated domain-specific document corpus, using expert, natural, and naive query formulations to examine query-sensitive retrieval behaviour and downstream answer generation under organisation-controlled deployment conditions.

Fifth, *evaluation* is conducted quantitatively using established information retrieval metrics and structured answer evaluation frameworks. Multiple reference answer sets serve as the evaluation baseline, including independently produced human annotations by domain-informed annotators as well as a model-generated reference answer set. No formal inter-annotator agreement protocol was enforced. Instead, potential variation between reference perspectives is intentionally retained to reflect realistic organisational decision contexts in which well-reasoned interpretations may legitimately differ. Retrieval effectiveness and downstream answer quality are analysed across reference perspectives to assess evidence acquisition and evidence utilisation under constrained deployment conditions.

Finally, the thesis itself constitutes the *communication* of the artefact, empirical findings, and resulting design implications. By structuring this work according to the DSR logic, the methodological approach clarifies the research process underlying the technical methodology and positions the work as artefact-oriented, empirically evaluated design research focused on identifying design trade-offs rather than prescribing a single optimal configuration.

2 BACKGROUND

2.1 Retrieval-Augmented Decision Pipelines

Organisational decision-making is constrained by limited information-processing capacity and the uneven availability of relevant knowledge at the point of action, commonly conceptualised as bounded rationality [10]. The effectiveness of technical support systems therefore depends not only on reasoning capabilities, but on how relevant information is identified and made accessible during decision-making [11]. This raises a fundamental design question for retrieval-augmented systems: how and when should knowledge be made available to the model?

Building on this perspective, a key design consideration in retrieval-augmented decision pipelines concerns how knowledge is made available to the system during decision-making. One approach embeds knowledge directly within the model through large-scale training, resulting in so-called parametric knowledge. In this case, the information available during reasoning is fixed at training time and can only be updated through additional training or fine-tuning, both of which are resource-intensive. As a result, knowledge representations are inherently static, limiting their suitability in settings where information must be frequently updated or selectively exposed.

An alternative approach provides non-parametric knowledge explicitly at the time of decision-making. RAG follows this principle by retrieving relevant information from external sources and providing it to the model during inference. By separating knowledge storage from reasoning, RAG enables dynamic knowledge updates, grounding model outputs in up-to-date information without modifying the underlying model parameters [3]. Moreover, RAG architectures can connect to heterogeneous external data sources, such as structured databases or internal knowledge management systems, supporting adaptation to evolving information needs. From a decision-support perspective, this design operationalises the intelligence phase by explicitly ensuring that relevant and structured information is identified and prepared prior to analysis and decision-making.

2.2 Core Retrieval Paradigms

Retrieval-augmented decision pipelines are positioned as an explicit stage within decision support pipelines, bringing up the necessity to clarify how retrieval systems determine which information is considered relevant. Over time, different retrieval approaches have emerged that reflect progressively refined assumptions about how relevance should be assessed for decision-making purposes.

Early retrieval systems relied on lexical representations of text, meaning that relevance was determined by whether the same words appeared in both the query and a document. A foundational development in this context was the vector space model which represents documents and queries as weighted collections of terms, estimating relevance based on how similar these term representations are, typically using cosine similarity [12, 13]. Building on this idea, probabilistic retrieval models further refined relevance estimation by modelling how likely a document is to be relevant given the terms it contains, while still relying on explicit word occurrences [14]. These approaches are commonly referred to as lexical retrieval methods, as they operate directly on observable terms in the text [15].

More recent methods adopt a semantic representation of text. Instead of modelling documents through discrete term counts, semantic retrieval encodes queries and documents as dense vector embeddings that capture contextual meaning. Relevance is then estimated based on proximity in this continuous representation space, allowing retrieval of conceptually related content even when lexical overlap is limited [16, 17]. In this sense, semantic retrieval is often described as dense retrieval, reflecting the use of learned continuous representations rather than sparse term vectors.

While differing in their underlying assumptions, both lexical and semantic retrieval approaches serve the same fundamental purpose, to identify and prioritise information that is relevant for downstream evaluation and decision-making, rather than to perform decision-making itself.

2.3 Domain-Specific Retrieval for Decision-Making

In domain-specific regulatory environments such as chemical compliance and clinical laboratory governance, essential information is often embedded in dense, formally structured documents characterised by extensive cross-referencing and context-dependent terminology [4]. Identical terms may carry different meanings depending on the institutional context in which they are used, and decisive constraints are frequently expressed implicitly rather than through explicit keywords. These structural characteristics create substantial challenges for information retrieval, as relevance cannot be inferred reliably from surface-level lexical overlap. Similar limitations have been observed in other regulated domains such as legal and policy texts, where retrieval systems optimised for generic semantic similarity frequently mis-rank or overlook critical documents [8].

These characteristics do not require a fundamentally new retrieval paradigm. Rather, they demand a deliberate adaptation and configuration of established lexical and semantic retrieval methods, as introduced in Section 2.2, to the constraints of specialised knowledge environments. Retrieval must therefore be evaluated and tuned against domain-specific

terminology, document structure, and decision constraints rather than against generic notions of semantic similarity.

Empirical studies indicate that general-purpose retrieval systems, which are designed to identify information that is broadly relevant across heterogeneous domains, frequently mis-rank or overlook critical documents in specialised corpora. This limitation is particularly pronounced when documents share similar structural characteristics or rely heavily on domain-specific terminology, leading to incomplete or misleading evidence [18].

Consequently, decision-making in high-stakes, knowledge-intensive settings requires context-sensitive relevance criteria. Retrieval should therefore be treated as an explicit, contextualised stage in the decision pipeline rather than a neutral preprocessing step. The documents retrieved define the evidence base for downstream reasoning and thereby directly influence decision quality. In high-stakes professional domains, errors often occur not because information is missing, but because the information used is misaligned with the decision context or insufficiently grounded in domain-specific constraints.

This perspective forms the basis for the design choices discussed in Chapter 5, where retrieval strategies are systematically aligned with the requirements of decision-making under constrained deployment conditions.

2.4 Vector Representations and Semantic Similarity

Text embeddings represent linguistic units as vectors in a continuous, high-dimensional space. During encoding, discrete text inputs such as words, sentences, or text chunks are transformed into fixed-length numerical representations. This process is referred to as embedding because the symbolic text is embedded into a vector space in which geometric relationships capture semantic similarity.

In this embedding space, semantically related texts are located closer together, allowing relevance between a query and candidate documents to be estimated using similarity measures such as cosine similarity. To support similarity-based retrieval over a collection of embedded text representations, vector databases store embeddings in an index that organises the vector space and enables efficient similarity comparisons, forming the foundation for semantic retrieval in RAG systems [15].

3 CHALLENGES OF APPLYING LARGE LANGUAGE MODELS IN HIGH-STAKES DOMAINS

The deployment of language-model-based decision-support systems in professional practice is shaped by interrelated challenges that extend beyond raw model performance,

including requirements for evidence-based grounding, traceability, and operation under organisational and regulatory constraints. This chapter examines these challenges using chemical regulatory compliance as the empirical context, highlighting how trust, governance, and deployment considerations jointly shape the design space of retrieval-augmented systems in high-stakes decision-support workflows.

3.1 Knowledge Grounding and Trust

In knowledge-intensive, high-stakes domains, the practical usefulness of language-model-based decision support is closely tied to the extent to which generated outputs can be grounded in transparent and verifiable evidence. In such settings, language models are typically employed to support regulatory research by processing and contextualising retrieved information, while responsibility for interpretation and final judgement remains with domain professionals. Consequently, trust in system outputs cannot rely solely on surface-level plausibility, but requires the ability to assess the relevance, validity, and provenance of the underlying information [19].

Large language models may hallucinate, producing statements that appear coherent while lacking support in authoritative sources. Empirical work shows that conditioning generation on retrieved evidence reduces such unsupported outputs, highlighting the limitations of purely parametric knowledge [2]. When generated claims cannot be explicitly attributed to verifiable sources, their reliability becomes difficult to assess.

Beyond correctness, regulated domains impose additional requirements on traceability. Domain professionals must be able to identify where specific information originates, why it has been retrieved in response to a given query, and how it constrains or supports a particular decision [20]. Information retrieval systems are thus expected to maintain explicit links between retrieved content and its source, enabling users to cross-check evidence within established regulatory frameworks. Closely related to traceability is the requirement for explainability in practical use. In this context, explainability does not refer to insight into model internals, but to the user's ability to understand why particular documents or passages were selected and how they relate to the decision question at hand. Without such contextual transparency, retrieved information risks being treated as decontextualised output, undermining its usefulness in workflows that rely on curated, continuously updated knowledge bases.

Taken together, these requirements highlight that effective decision support in high-stakes, knowledge-intensive domains depends not only on retrieving relevant information, but on doing so in a manner that is traceable, explainable, and aligned with domain-specific standards of evidence [21]. These properties are essential for the reliable use of such

systems in professional decision-making contexts. Addressing these challenges therefore calls for retrieval strategies that prioritise authoritative sources and explicit grounding over generic relevance, forming a key design consideration for retrieval-augmented decision pipelines in organisation-controlled deployment settings.

3.2 Regulatory and Organisational Constraints

The chemical sector operates under complex and continuously evolving regulatory frameworks that govern the manufacture, import, and downstream use of chemical substances. Compliance with regulations such as Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) requires the registration and notification of substances, ongoing monitoring of regulatory updates, and the interpretation of detailed legal and technical guidance [22].

Regulatory compliance therefore depends on specialised expertise, local regulatory knowledge, and sustained interpretative capacity [23]. These capabilities are not uniformly available within organisations and may delay compliance-related decisions or increase regulatory risk. Moreover, regulatory knowledge evolves over time as legislation is amended, guidance is revised, and authoritative interpretations shift. In practice, regulatory professionals often enter the field without formal legal training, and decision-making relies heavily on experiential knowledge rather than systematised legal reasoning [23]. This reliance creates organisational vulnerability: critical knowledge may be lost through retirement, turnover, or insufficient off-boarding if it is not systematically documented [24].

These structural and organisational constraints directly affect compliance-related decision-making and increase the risk of delayed market entry, elevated operational costs, and error-prone outcomes.

3.3 Data Security and Deployment Constraints

Despite their promising capabilities, organisations remain hesitant to adopt commercial LLMs due to unresolved concerns surrounding organisational data governance constraints and their implications for secure deployment [8, 19]. In professional decision-support settings, meaningful queries frequently rely on proprietary or non-public internal knowledge. Uploading such material to a public or third-party hosted LLM, even under privacy agreements, is often viewed as incompatible with internal governance and compliance policies [8]. As a result, the practical use of externally hosted LLM-based systems in business-critical workflows remains limited despite their technical potential.

Several surveys and industry studies confirm this hesitation. Recent findings indicate

that a significant share of enterprises across sectors cite data security concerns as primary reasons for not adopting generative artificial intelligence tools in business-critical workflows [25, 26]. These concerns reflect organisational data governance policies rather than isolated technical risks. While some vendors offer contractual guarantees against training on user data, these safeguards remain trust-based and may not satisfy internal governance policies, particularly when their technical implementation is opaque. Relying on externally hosted models therefore shifts control and accountability away from the organisation.

These constraints are particularly evident in regulatory decision-making contexts within the chemical sector. For example, a regulatory specialist preparing a dossier for market access may need to assess whether substance-specific thresholds or documentation requirements apply using existing registration material as context. Under the REACH regulation, such dossiers contain detailed substance information, toxicological data, and may include confidential business information subject to protection [27]. Submitting such material to an external service therefore risks exposing trade secrets, toxicological profiles, or confidential use patterns. Similarly, evaluating whether increased production volumes would trigger additional regulatory obligations may require combining public regulatory requirements with internal forecasts or composition data, thereby introducing sensitive internal information into the analysis.

The decision to integrate LLMs into such workflows therefore depends not only on model capability, but on deployment design. Different deployment strategies define distinct trust boundaries and levels of organisational control. Prior governance frameworks distinguish between locally controlled systems, private on-premise infrastructures, enterprise-managed cloud environments, and third-party hosted services, each implying different allocations of risk, control, and accountability [28]. Each option determines where data is processed and how compliance with governance policies can be enforced. At the core of these considerations lies the distinction between *data controllers* and *data processors*. Under established data protection regulation, organisations remain responsible as data controllers even when processing is delegated to external providers acting as data processors [29]. Limited visibility into data handling practices, storage locations, and jurisdictional exposure complicates auditing and accountability, raising concerns related to data sovereignty and the handling of proprietary regulatory information.

In sum, the deployment of LLM-based systems in regulated professional environments is constrained not only by technical performance, but by governance and trust. For retrieval-augmented systems, this implies that benefits demonstrated under unconstrained deployment conditions cannot be assumed to transfer directly to organisation-controlled set-

tings. Understanding system behaviour under governance-driven deployment restrictions is therefore essential for decision-support workflows in high-stakes, knowledge-intensive domains.

4 RELATED WORK

RAG formalises the integration of external sources at inference time, conditioning model outputs on explicitly retrieved information and separating retrieval from parametric generation [3]. Empirical work in open-domain question answering demonstrates that retrieval-conditioned generation consistently outperforms purely parametric models and that downstream performance scales with the number of retrieved passages [30]. These findings indicate that improvements in retrieval can partially compensate for limited generator capacity. Large-scale language model studies further demonstrate additional gains through parameter scaling, albeit under substantial computational requirements and costs [31]. Survey research conceptualises RAG as a modular design space in which query construction, retrieval strategy, ranking behaviour, and generation directly influence system performance [32].

Beyond open-domain benchmarks, retrieval-augmented approaches have increasingly been applied in specialised, high-stakes domains characterised by complex terminology and strict evidentiary requirements. Recent work demonstrates the feasibility of retrieval-conditioned generation for domain-specific decision-support and knowledge access scenarios, illustrating the practical viability of RAG in knowledge-intensive environments [4]. While these studies show promising results, they typically rely on large generative models and unconstrained computational infrastructures, leaving constrained deployment settings comparatively underexplored.

Within retrieval-augmented pipelines, early ranking quality determines which evidence becomes available for conditioning downstream generation. Empirical studies show that improvements at the retrieval stage translate into measurable gains in answer generation performance [30, 32]. These findings suggest a functional coupling between retrieval adequacy and downstream answer quality.

However, more recent analyses of retrieval-augmented systems under realistic retrieval conditions indicate that the assumed coupling between retrieval adequacy and answer quality is neither uniform nor guaranteed. Controlled benchmark evaluations may overestimate the extent to which generative models fully exploit retrieved evidence. Studies introducing real-world context distributions and dedicated utilisation metrics show that the presence of relevant passages does not necessarily result in faithful or evidence-grounded answer generation [33]. These findings suggest that retrieval adequacy and effective ev-

idence utilisation constitute distinct components of system performance. Moreover, because only a limited number of retrieved passages are selected for conditioning, early ranking behaviour determines which evidence becomes available to the generator, making evidence selection and evidence utilisation interacting but analytically separable factors within retrieval-augmented pipelines.

To improve the reliability of evidence selection prior to generation, various post-retrieval refinement paradigms have been proposed. These approaches aim to enhance early precision by reordering or filtering candidate passages before they are passed to the generative model. Representative methods include pointwise re-ranking [34], pairwise optimisation [35], and listwise or setwise approaches [36, 37]. While such techniques can substantially enhance ranking quality, many rely on supervised training or computationally intensive cross-encoder architectures.

More recent work explores language-model-based re-ranking and step-wise refinement as modular components that can operate without retraining the underlying retriever [38]. However, these approaches often depend on expressive neural models at inference time, increasing computational overhead and limiting applicability under strictly constrained, organisation-controlled deployment settings.

A longstanding insight in information retrieval is that retrieval effectiveness strongly depends on query formulation, as users often struggle to precisely express their underlying information needs [39]. Building on this foundation, recent empirical work shows that RAG systems exhibit measurable sensitivity to query-level variation [7], where even minor reformulations can lead to substantial differences in retrieval and generation outcomes.

To address this issue, query rewriting and multi-query strategies have been proposed [40]. Task-aligned Small Language Models (SLMs) can serve as dedicated query rewriters to improve retrieval recall and downstream answer quality, particularly for underspecified or heterogeneous queries. More recent studies investigate query-aware adaptation strategies, including retrieval strategy selection conditioned on query characteristics [8], highlighting the configuration sensitivity of retrieval-augmented pipelines.

However, while adaptive retrieval strategies and query reformulation have been studied under benchmark-style conditions, these analyses typically assume access to large generative models and unconstrained infrastructure. Deployment constraints and model-size limitations are rarely treated as primary design variables. Consequently, the combined effects of query variation, retrieval configuration, and generation behaviour under organisation-controlled deployment conditions remain insufficiently understood.

Parallel research investigates SLMs as efficient alternatives to large-scale models. A recent survey synthesises training, optimisation, and deployment strategies for SLMs and emphasises their suitability for computationally constrained, privacy-sensitive, and organisation-controlled environments [41]. While SLM research focuses primarily on efficiency–performance trade-offs and model-level capabilities, fewer studies analyse how retrieval strategies, ranking behaviour, and answer generation interact when both components operate under explicit model-size constraints.

Taken together, prior work demonstrates that retrieval augmentation improves grounding and downstream performance [3, 30], that ranking refinement enhances early precision [34–36], and that query adaptation can reduce sensitivity to formulation differences [7, 40]. Most existing studies evaluate these components under unconstrained deployment conditions and rely on large generative models [31, 32]. This thesis evaluates retrieval strategies, post-retrieval refinement, and generation model selection under systematic query variation within a retrieval-augmented architecture constrained to SLMs in organisation-controlled deployment environments.

5 METHODOLOGY

This chapter describes the methodological approach adopted in this thesis for implementing and evaluating a RAG system for domain-specific information retrieval. An overview of the system architecture is shown in Figure 1. The system is structured as a modular, sequential processing pipeline comprising document extraction, text segmentation, embedding and indexing, retrieval with optional post-retrieval refinement, and answer generation. Each stage fulfils a distinct functional role within the pipeline and can be analysed and optimised independently.

The chapter first motivates the choice of this retrieval-augmented approach, followed by a detailed description of the retrieval architecture, retrieval strategies, post-retrieval refinement, and answer generation. Finally, an evaluation design is introduced to assess the performance of the proposed system.

5.1 System Design and Model Selection

5.1.1 System Design Rationale

The methodological design of this thesis is guided by the requirements of decision-making in high-stakes, knowledge-intensive domains. In such settings, decision quality depends on reliable, verifiable information, and system outputs must remain traceable and explainable to ensure accountability. Decision-making in specialised professional environments

is therefore information-constrained, as relevant knowledge must be explicitly identified, structured, and prepared prior to analysis rather than implicitly assumed to reside within a model.

While domain adaptation can be achieved through fine-tuning, this approach entails substantial training costs and produces static knowledge representations, making it unsuitable under the organisation-controlled deployment constraints considered in this thesis. A RAG architecture is therefore adopted, as it enables grounding in explicitly retrieved internal and external sources without embedding proprietary or non-public internal knowledge in model parameters.

This design fundamentally redefines the role of language models within the system. Rather than acting as repositories of domain knowledge, models operate exclusively over evidence retrieved at inference time. Their function becomes narrowly task-aligned within a modular decision pipeline, reducing the need for large parametric capacity and broad domain memorisation while supporting traceability and governance requirements.

This redefinition of the model's function enables a departure from monolithic, large-scale language models towards smaller, task-aligned models. In retrieval and ranking stages, models are primarily required to assess relevance and compare candidate text segments, tasks that depend on contextual understanding and classification rather than open-ended generation or extensive domain fluency. Similarly, when generation is grounded in retrieved context, the model's role is limited to interpreting provided information and producing a coherent response without extrapolating beyond the available evidence. Under these restricted functional requirements, very large parametric capacity becomes less critical, while models with several billion parameters entail substantially higher memory and computational demands even when quantisation techniques are applied. SLMs are therefore employed as a deliberate methodological choice: their reduced computational footprint and predictable runtime behaviour make them well suited for local and organisation-controlled deployment, allowing individual components to be aligned with task requirements while supporting traceability and governance constraints.

This design rationale directly relates to the central research question of this thesis, which examines whether retrieval-augmented decision-support pipelines composed of smaller, task-aligned language models can provide reliable and trustworthy support in governed professional settings without relying on large, centrally hosted language models.

5.1.2 Model Selection Rationale

The proposed pipeline employs instruction-tuned SLMs, consistent with the retrieval-augmented design in which models operate over explicitly retrieved evidence rather than

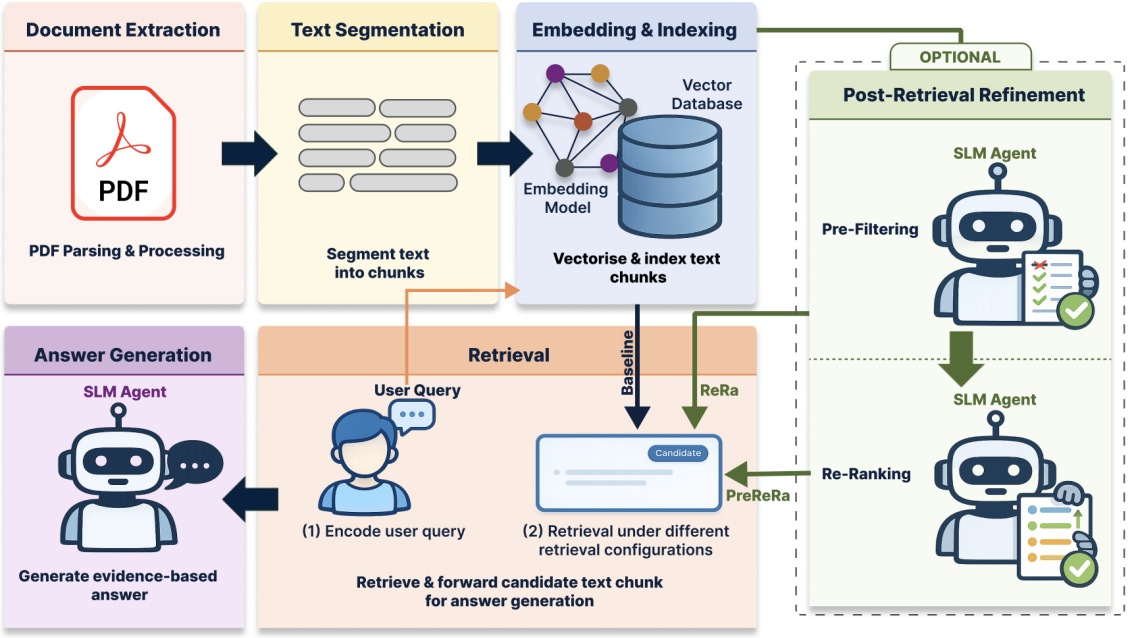

FIGURE 1: Overview of the retrieval-augmented generation (RAG) pipeline implemented in this work, illustrating the modular sequence from document ingestion to answer generation.

internalised domain knowledge. Model requirements therefore prioritise reliable context interpretation, instruction following, and controlled answer synthesis.

For post-retrieval refinement, the *LLaMA 3.2 3B Instruct* model with 3 billion parameters is used as a relevance assessor and re-ranking component. In this role, the model evaluates and compares candidate text segments with respect to a query, closely aligning with neural re-ranking formulations in information retrieval [42]. Such tasks depend primarily on contextual discrimination and comparative reasoning rather than open-ended generation, making a compact 3B instruction-tuned model sufficient under constrained deployment conditions.

For answer generation, the *LLaMA 3.2 3B Instruct* and *Phi-3-Mini* models are evaluated. Within the setting, generation is grounded exclusively in retrieved context, limiting the model's function to interpreting provided evidence and producing coherent, instruction-compliant responses. Both models belong to the 3–4 billion parameter class and demonstrate strong general reasoning performance on benchmarks such as *MMLU* [43]. Selecting models of comparable scale but from different model families enables analysis of architecture- and training-specific behaviour without confounding results with parameter size differences.

5.2 Retrieval Architecture

The retrieval architecture (see Figure 1) defines how documents are transformed into machine-processable representations and how relevant text chunks are selected from the knowledge base. This includes document extraction, text segmentation (chunking), embedding and indexing, and the retrieval step with optional post-retrieval refinement.

The knowledge base consists of 24 official and publicly available guidance documents published by the European Chemicals Agency (ECHA) [44]. These documents contain specialised regulatory terminology and provide a controlled corpus for evaluating domain-specific retrieval behaviour. Although publicly available, they are typically consulted in contexts involving proprietary or non-public internal information, making them suitable for studying retrieval under governance-driven deployment constraints. All documents are provided in Portable Document Format (PDF) and require preprocessing prior to indexing.

PDF extraction is performed using the Python library *PdfPlumber*, followed by text normalisation and serialisation into a structured corpus. The extracted text is segmented using a fixed-length, token-based chunking strategy, with each chunk consisting of 512 tokens and a 12.5% token-level overlap between adjacent segments. This configuration ensures compatibility with downstream model input limits while reducing the risk of context fragmentation across chunk boundaries. Token-level segmentation is employed because tokenisation is model-dependent and does not necessarily align with word-level boundaries.

Text chunks are embedded into a continuous semantic space and stored in a *Chroma*[1] vector database. Chroma is selected for its ease of integration and sufficient performance given the corpus size. For substantially larger knowledge bases, a *FAISS*-based backend would represent a more scalable alternative.

During retrieval, a fixed number of candidate text chunks is selected in response to a user query. The query is encoded using the same embedding model as the corpus and projected into the shared representation space, enabling similarity-based retrieval. Retrieval effectiveness is influenced by design choices such as chunking strategy, embedding model selection, and vector database configuration, all of which must operate within hardware and deployment constraints.

[1] `https://www.trychroma.com`

5.3 Retrieval Strategies

The retrieval step selects a fixed number of candidate text chunks (Top-k) from the knowledge base in response to a user query. The retrieval strategy determines how relevance is operationalised. As introduced in Section 2.2, this work considers two paradigms: lexical and semantic retrieval. Lexical retrieval estimates relevance based on observable term overlap between query and document text, whereas semantic retrieval relies on similarity in a continuous vector space derived from dense embeddings. In contrast to supervised dense passage retrieval approaches, embedding-based retrieval in this work relies exclusively on pre-trained models without task-specific fine-tuning, reflecting local deployment and resource constraints.

To evaluate both paradigms, multiple retrieval strategies are implemented. Lexical retrieval is realised using Term Frequency–Inverse Document Frequency (TF-IDF) and Best Matching 25 (BM25), while semantic retrieval is implemented via cosine similarity over embedding representations. In addition, a hybrid strategy combines lexical and semantic scores through weighted aggregation.

The effectiveness of semantic retrieval is further examined by evaluating different embedding models. Both segmented text chunks and queries are encoded into a shared representation space, enabling similarity-based relevance estimation. To assess the trade-off between retrieval effectiveness and computational efficiency in the chemical regulatory domain, three models from the *BGE v1.5* family are employed: *BGE-small-en-v1.5*[2], *BGE-base-en-v1.5*[3], and *BGE-large-en-v1.5*[4]. These models vary in parameter size and embedding dimensionality, enabling controlled analysis of scale effects under deployment constraints.

To combine lexical and semantic retrieval, a hybrid strategy is applied in which a weighting factor α controls the relative contribution of each component. The BM25-based score is weighted by α, while the embedding-based score is weighted by $(1 - \alpha)$. In this work, $\alpha = 0.5$ assigns equal weight to both components. Although α can be treated as a tunable hyperparameter and adjusted through sensitivity analysis, it is held constant in order to enable controlled comparison of architectural configurations. Since the dense component relies on vector representations, hybrid performance is influenced by the selected embedding model.

[2] https://huggingface.co/BAAI/bge-small-en
[3] https://huggingface.co/BAAI/bge-base-en
[4] https://huggingface.co/BAAI/bge-large-en

5.4 Post-Retrieval Refinement

Neural re-ranking models represent a prominent approach to refining retrieved candidate lists, however, their computational overhead makes them impractical in the targeted local deployment setting. Instead, this work employs a resource-efficient post-retrieval refinement strategy inspired by prior work on LLM-based post-retrieval refinement [38], reflecting the trade-offs imposed by hardware and deployment constraints. As illustrated in Figure 1, three post-retrieval variants are considered: *Baseline*, where retrieved candidates are passed directly to the generator; *Re-Ranking only* (ReRa), where candidates are reordered by estimated relevance; and *Pre-Filtering and Re-Ranking* (PreReRa), where candidates are first filtered by query relevance and subsequently re-ranked.

All post-retrieval steps are implemented using SLMs as independent, sequential processing agents. Although these components follow an agent-style execution, the overall system does not constitute an agentic or multi-agent architecture. Agents are invoked within a fixed pipeline, operate statelessly, and perform no inter-agent communication, self-reflection, or validation. The modular design of the proposed pipeline nevertheless allows agent-based components to be integrated or replaced in future extensions without altering the underlying structure. This work deliberately evaluates a non-autonomous configuration to isolate the effects of retrieval, post-retrieval refinement, and generation under controlled conditions. Both pre-filtering and re-ranking are performed using the *LLaMA 3.2 3B Instruct* model and executed as described in Section 6.1.

Although SLMs may also exhibit unreliable behaviour in open-ended knowledge generation settings, the post-retrieval refinement stage constrains the model to structured relevance assessment over explicitly provided query and candidate text chunks. The employed zero-shot prompts define this task precisely and restrict the model to producing bounded relevance scores rather than generating new domain content. Nevertheless, relevance estimation remains probabilistic, such that the use of an SLM-based refinement mechanism represents a deliberate trade-off between computational feasibility and decision reliability.

In the pre-filtering stage, the relevancy of each candidate text chunk with regards to the query is evaluated independently and assigned as a relevance score by a SLM-based agent. The zero-shot prompt shown on the top left side in Figure 2 is employed to guide relevance assessment and encourage explicit reasoning prior to score generation. For a given query and text chunk, the agent produces a relevance score between 0 and 1. This score is converted into a binary relevance decision using a fixed threshold τ, such that only candidate text chunks deemed relevant are forwarded to the subsequent re-ranking

stage. Pre-filtering is performed in an unsupervised manner using a fixed threshold of $\tau = 0.3$, following prior findings that report stable filtering behaviour across diverse retrieval tasks [38]. This setting reflects realistic deployment conditions in which no human-annotated relevance judgements are available, and the threshold therefore cannot be tuned or validated on labelled data.

The re-ranking stage refines the ordering of candidate text chunks based on their estimated relevance to the query and is applied to the output of the initial retrieval step or the pre-filtered candidate set. The re-ranking task is performed by a SLM acting as a task-aligned, independent, and stateless processing agent, as available open-source re-ranker models are typically not fine-tuned for the domain-specific regulatory context. The zero-shot prompt employed is shown on the right side in Figure 2. Due to input length limitations of the employed language model, re-ranking is not performed over the entire candidate set. Instead, a group-wise, multi-round re-ranking procedure is applied to approximate a global relevance ordering under constrained context capacity.

The candidate list is randomly shuffled for the group assignment to mitigate position and context effects. For every text chunk, the resulting relevance score is incorporated into a running mean relevance estimate to obtain an increasingly stable score. After all re-ranking rounds have been completed, the candidate text chunks are sorted according to their final mean relevance score in descending order. The Top-k text chunks are selected and forwarded to the downstream generation stage. For example, given 25 candidate text chunks, a group size of 5, and two rounds of evaluation, the candidate list is randomly shuffled twice and divided into five groups per round. As a result, each chunk is evaluated twice in total, once in each round, resulting in 10 model calls. The re-ranking procedure is executed once per query and does not involve adaptive cut-offs, uncertainty-based document selection, or iterative self-correction beyond the predefined number of rounds. While the approach is inspired by more elaborate multi-stage ranking strategies discussed in Section 4, it represents a deliberate design trade-off that balances robustness of relevance estimation with computational efficiency under constrained hardware conditions.

5.5 Answer Generation

Answer generation constitutes the final stage of the RAG pipeline (see Figure 1). This stage operates exclusively on evidence retrieved through the preceding retrieval and optional post-retrieval refinement steps. Consistent with the system design rationale, the role of the generation component is not to discover or infer new domain knowledge, but to transform retrieved evidence into a coherent and comprehensible response.

The system follows a single-call generation design. For each query, the top-k retrieved

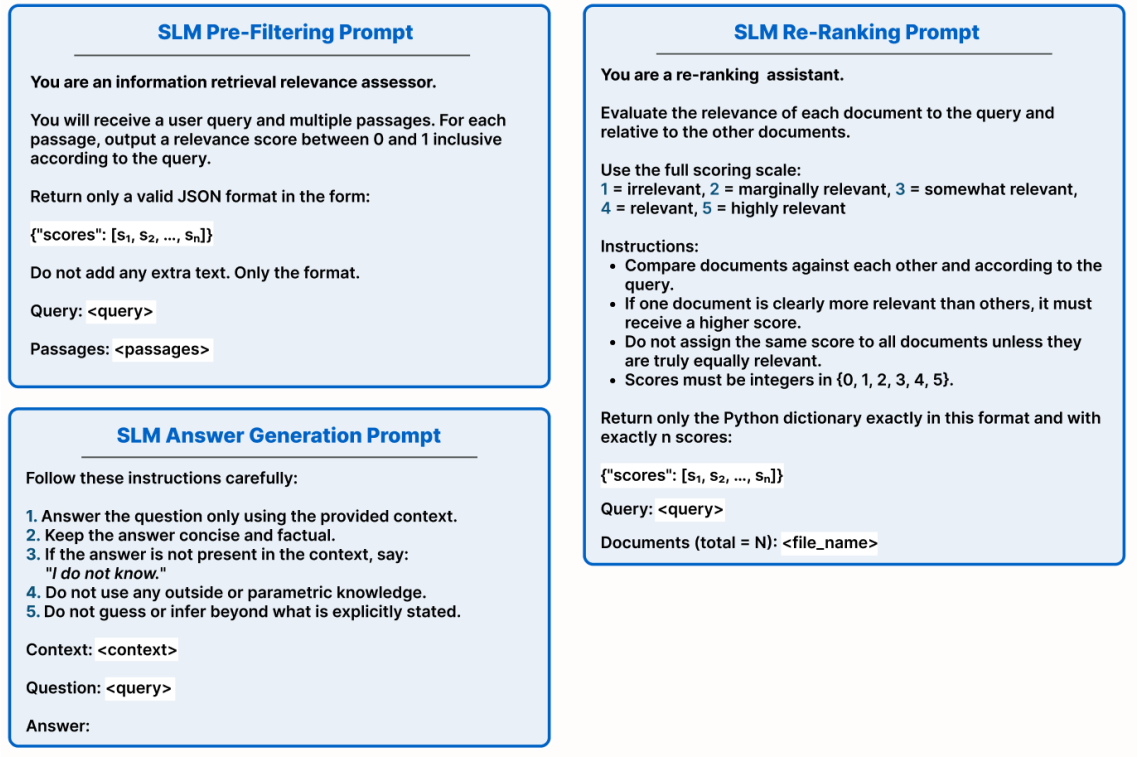

FIGURE 2: Zero-shot prompts used for post-retrieval refinement and answer generation. The top-left prompt implements small language model (SLM)-based pre-filtering, the right prompt performs re-ranking of retrieved text chunks, and the bottom-left prompt handles answer generation by providing the query and retrieved context, explicitly restricting the response to the supplied evidence.

candidate text chunks are provided jointly as contextual input to the generation model, which produces exactly one response. Generation is executed under the experimental conditions described in Section 6.1.

Explicitly grounding generation in retrieved context fundamentally changes the demands placed on the generation model compared to general-purpose conversational systems, which are typically designed to operate without explicit evidence and therefore rely on large-scale internalised knowledge. As outlined in the system design rationale (Section 5.1.1), the generation model in the proposed pipeline is not required to memorise extensive domain knowledge or support broad task generality. Instead, it must accurately interpret the provided context and produce responses that remain faithful to the retrieved evidence.

Based on these requirements, answer generation in this work is implemented using small, instruction-tuned, task-aligned language models. These models are selected from a comparable parameter range and are capable of performing context-conditioned answer synthesis, summarisation, and structured prompt following. They are chosen due to their

competitive performance at their respective scale while remaining compatible with aggressive quantisation, enabling efficient local inference under constrained deployment settings.

Importantly, the methodological design does not assume that a single language model is optimal for retrieval-augmented answer generation. Instead, multiple representative models are employed and compared to assess whether observed generation behaviour is consistent across models or influenced by model-specific characteristics. The concrete model choices and the comparative evaluation protocol are therefore introduced as part of the evaluation design (Section 5.6).

To ensure evidence-grounded answer generation across all models, a fixed instruction-style prompt is used for the generation step. The prompt explicitly constrains the model to rely solely on the retrieved context, prohibits the use of parametric or external knowledge, and enforces concise and factual responses. This design directly supports traceability and reduces the risk of hallucinated content, which is particularly critical in regulated decision-support settings. By keeping the prompt (see bottom left Figure 2) identical across all generation models, differences observed during evaluation can be attributed to model behaviour rather than prompt variation.

5.6 Evaluation Design

The evaluation design reflects the modular structure of the proposed retrieval-augmented pipeline (Figure 1) and is aligned with the research questions introduced in Section 1.3. It assesses retrieval and answer generation both as separable components and as interacting stages within a governance-constrained, locally deployable system composed of small language models.

Across all evaluation stages, system behaviour is examined under three query types that reflect variation in query formulation rather than a fixed property of the user. While limited domain knowledge may naturally lead to more descriptive, naive formulations, a domain professional may equally produce any of the three types, as query specificity is shaped by factors beyond expertise alone, including cognitive load, time pressure, and the degree to which the information need has been conceptualised [39, 45]: expert-level queries using domain-specific terminology and, where necessary, explicitly specifying regulatory thresholds; naturally phrased queries reflecting informed but non-technical language; and broadly phrased naive queries formulated in descriptive, non-technical terms. This controlled variation allows the evaluation to assess the robustness of both retrieval and generation components while targeting identical underlying regulatory questions.

The first part of the evaluation focuses on retrieval effectiveness under these varying query

formulations. By analysing lexical, semantic, and hybrid retrieval strategies, including optional post-retrieval refinement, independently of answer generation, this stage isolates how retrieval behaviour changes with query specificity and linguistic variation under constrained deployment conditions.

The second part evaluates answer generation performance while holding the retrieval configuration constant. By comparing instruction-tuned SLMs of comparable scale within a fixed infrastructure and prompting configuration, this stage isolates model-dependent effects on answer quality, grounding, and faithfulness.

5.6.1 Retrieval Performance

To evaluate retrieval performance, three annotators independently answered a set of twenty regulatory questions and identified the authoritative source passages required to support each answer. These annotations were consolidated into a relevance file, which serves as ground truth for retrieval evaluation. Importantly, the annotated answers are invariant across different query formulations, ensuring that retrieval performance is assessed independently of linguistic variation in the phrasing of the questions.

Retrieval quality is measured using the *BEIR* framework [46], which provides a standardised evaluation protocol for zero-shot information retrieval. *BEIR* evaluates how effectively a retrieval method ranks relevant text chunks given a set of natural-language queries and computes established information retrieval metrics. Concretely, Normalized Discounted Cumulative Gain at rank k (*nDCG@*k), Recall@k, and Mean Reciprocal Rank at rank k (*MRR@*k) are evaluated. *nDCG@*k measures the effectiveness of ranking relevant text chunks near the top of the retrieved list, while Recall@k captures the proportion of relevant documents retrieved within the top-k results. *MRR@*k emphasises how highly the first relevant document is ranked. Together, these metrics provide a comprehensive view of retrieval effectiveness in terms of ranking quality, coverage, and early precision.

All retrieval configurations are evaluated using a fixed candidate set size to ensure comparability across retrieval strategies. If fewer candidates are available at an intermediate stage, the set is completed with additional retrieved candidates before evaluation as explained in Section 5.4.

To examine retrieval robustness under varying user formulations, each annotated question was formulated in the three query variants introduced above (expert, natural, and naive). While the phrasing differs, all variants target identical underlying regulatory answers, ensuring that retrieval performance reflects sensitivity to linguistic variation rather than differences in ground-truth relevance.

5.6.2 Answer Generation Performance

Answer generation is evaluated in a second analysis stage with the primary objective to assess model dependence of answer generation, while enabling a complementary assessment of retrieval quality from an answer-centric perspective. To isolate model-dependent effects, answer generation is performed using different SLMs under identical retrieval configurations.

Because correct regulatory answers may be expressed using diverse phrasings and levels of detail, surface-level lexical metrics are not suitable indicators of answer quality in this setting. Generated answers may be semantically correct and well grounded in retrieved evidence while exhibiting low lexical overlap with a reference answer, leading to a falsely negative bias. To address this limitation, the evaluation adopts an *LLM-as-a-judge* paradigm using the Retrieval-Augmented Generation Assessment (RAGAS) framework [47]. The framework employs the external *GPT-4o-mini* model, accessed via the *OpenAI* API, to assess answer quality and contextual grounding by jointly considering the user query, the retrieved context chunks, the generated answer, and a reference answer.

To account for variation in plausible regulatory answer formulations and scopes, four alternative reference answer sets are used. Three reference sets are produced independently by human annotators (A1–A3), and a fourth is generated using the *GPT-5.2* model. These reference answers serve as alternative instantiations of the ground truth for the same underlying question. All evaluation metrics that require a reference answer are computed separately for each reference set and for each query variant (expert, natural, and naive), allowing sensitivity to annotator perspective and query formulation to be assessed explicitly. As this evaluation step is performed independently of the RAG pipeline, it is not subject to the system's data-privacy and hardware constraints, enabling the use of a larger and more expressive model better suited to nuanced semantic evaluation.

Answer quality is primarily evaluated using the dimensions of *answer correctness* and *faithfulness*. Answer correctness measures semantic agreement between the generated answer and a given reference answer and therefore depends on the combination of query, generated answer, and reference answer. Faithfulness assesses whether the generated answer is strictly grounded in the retrieved context without introducing unsupported information and depends on the query, the generated answer, and the retrieved context, but not on the reference answer. In addition, *context precision* and *context recall* are reported as complementary diagnostics of retrieval adequacy from an answer-centric perspective. Context precision measures the proportion of retrieved content that is relevant to the ref-

erence answer, whereas context recall measures the extent to which the retrieved context covers the information required by the reference answer. Consequently, both depend on the retrieved context and the respective reference answer and are computed separately for each reference set and query variant.

In cases of uncertainty, the generation model is instructed to refrain from producing responses in the absence of sufficient evidence and to explicitly indicate uncertainty (e.g., 'I do not know'). Accordingly, answer correctness scores are computed exclusively for substantive answers to avoid penalising abstention behaviour that promotes answer faithfulness. The proportion of non-abstained responses is reported as the *retention rate*, defined as the share of queries for which the model produces a substantive answer. Retention rate therefore depends solely on the generated output behaviour of the respective generation model under the given retrieval configuration and is independent of the chosen reference answer set.

Answers are generated using the top $k = 3$ retrieved candidate text chunks, selected based on ranking quality and early relevance assessment metrics under fixed retrieval configuration.

6 RESULTS

6.1 Experimental Setup

All experiments were conducted under organisation-controlled deployment constraints on a single local workstation in a CPU-based environment without access to dedicated GPU acceleration. No distributed or multi-machine execution was employed. This configuration reflects the assumed infrastructure limitations throughout this thesis, where proprietary or non-public internal knowledge precludes the use of externally hosted models. A fixed random seed and a temperature of 0.0 were applied across all language models to ensure deterministic execution, and retrieval parameters were held constant across all configurations.

Model inference was performed locally using the *llama.cpp* framework with a uniform 4-bit quantisation scheme applied across all models, ensuring consistent experimental conditions across retrieval, post-retrieval refinement, and answer generation stages.

For post-retrieval refinement, both pre-filtering and re-ranking were performed using the *LLaMA 3.2 3B Instruct* model. The model was applied to evaluate and rank candidate text chunks produced by the retrieval stage.

For answer generation, instruction-tuned SLMs in the 3-4 billion parameter range were

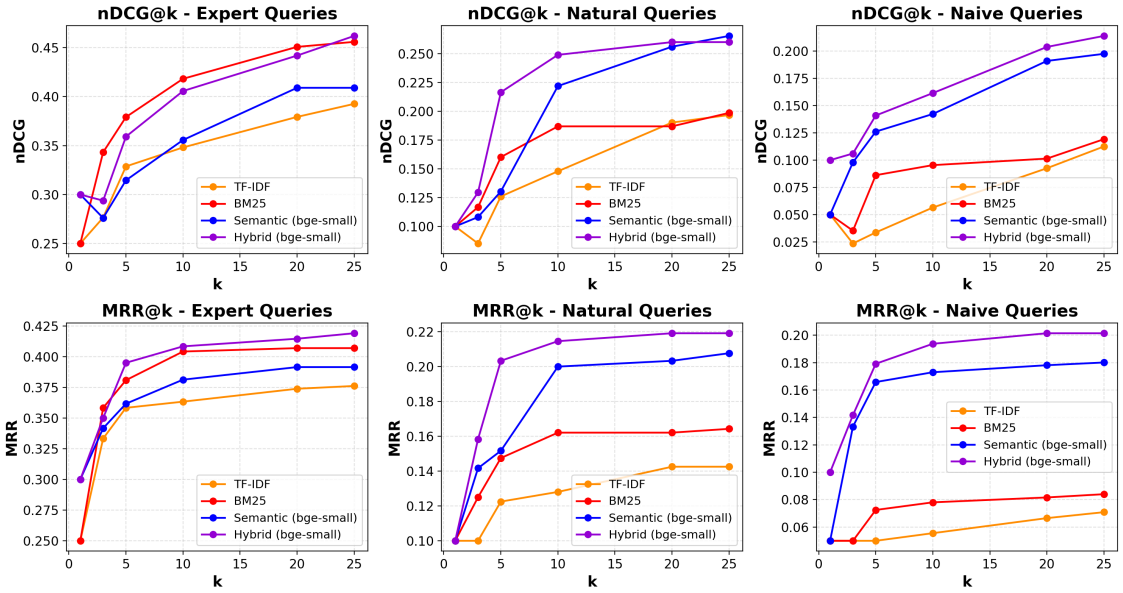

FIGURE 3: Lexical retrieval performs best for expert queries, while hybrid retrieval dominates for natural and naive formulations in terms of *nDCG@k* and *MRR@k*. Semantic retrieval increasingly outperforms lexical strategies for naive queries. Both metrics decline from expert to naive queries, indicating that retrieval effectiveness systematically varies with query formulation.

employed. The evaluation includes *LLaMA 3.2 3B Instruct* and *Phi-3-Mini*. Retrieval configuration, prompt templates, and context window settings were held constant across models to ensure that observed performance differences can be attributed to generation behaviour rather than infrastructural variability.

6.2 Retrieval Performance under Query Variation

The sub-question **RQ1** addresses the central research question by assessing how systematic variation in query formulation influences retrieval effectiveness within a locally deployable retrieval-augmented architecture. For evaluation, different retrieval strategies, including *TF-IDF*, *BM25*, semantic retrieval, and a hybrid retrieval, with the latter two based on the *BGE-small-en-v1.5* embedding model were examined.

Figure 3 reports *nDCG@k* and *MRR@k* for values of k ranging from 1 to 25 across expert, natural, and naive query formulations, capturing ranking quality and early relevance across retrieval strategies. For expert-level queries (first column), semantic and hybrid retrieval strategies attain the highest *nDCG@1*, while hybrid retrieval yields the highest *nDCG@25* among the evaluated methods. For intermediate cut-offs ($3 \leq k \leq 25$), lexical retrieval using *BM25* records the highest scores. With respect to *MRR@k*, hybrid retrieval produces the highest scores across k. In contrast, for natural and naive query formulations

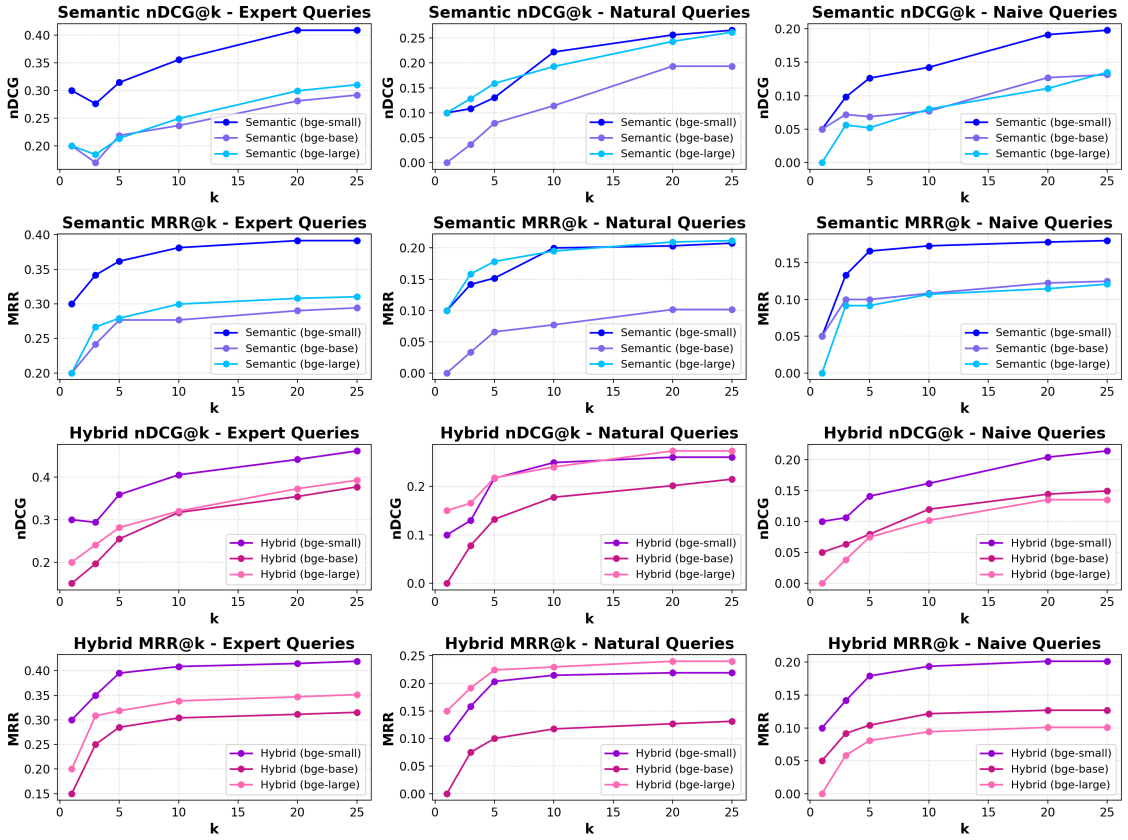

FIGURE 4: Embedding model size does not produce uniform scaling effects across query formulations. While *BGE-large* performs best for natural queries at small cut-offs, *BGE-small* achieves higher effectiveness for expert and naive queries.

(second and third columns), the hybrid retrieval strategy yields the strongest performance across the examined range of k for both metrics. As query formulations become less specific, a systematic shift in the relative ordering of retrieval strategies emerges and becomes increasingly pronounced. For natural queries, *BM25* remains the second-best performing strategy for small values of k in terms of *nDCG@k*. However, semantic retrieval increasingly outperforms lexical approaches at higher values of k in terms of *nDCG@k*. A similar pattern is observed for *MRR@k*, where semantic retrieval slightly outperforms *BM25* at small k, with the performance gap widening as k increases. This semantic shift is most evident for naive queries, where both hybrid and semantic retrieval substantially outperform lexical strategies across both metrics.

Across all retrieval configurations, both *nDCG@k* and *MRR@k* exhibit a monotonic decline from expert-level to naive query formulations, reflecting the increasing difficulty of retrieving relevant information as queries become more descriptive and less technically precise.

To further examine whether embedding model choice, particularly model size, is sensi-

tive to varying query formulations, retrieval performance was evaluated for semantic and hybrid strategies using three models from the *BGE* family: *BGE-small-en-v1.5*, *BGE-base-en-v1.5*, and *BGE-large-en-v1.5*. Figure 4 reports *nDCG@k* and *MRR@k* for values of k ranging from 1 to 25 across expert, natural, and naive query formulations, capturing ranking quality (*nDCG@k*) and early relevance across retrieval strategies (*MRR@k*).

The results show that for expert-level queries, the smallest embedding model (*BGE-small*) achieves the highest performance for both metrics, with a substantial margin over the larger models. The same pattern is observed for naive query formulations. While the smallest embedding model attains the same value as the medium large model (*BGE-base*) at $k = 1$, it consistently outperforms both embedding models at higher values of k across both metrics. The identity of the second-best embedding model varies across query formulations, with the largest embedding model (*BGE-large*) ranking second for expert-level queries, while the *BGE-base* model ranks second for naive queries. In contrast, this pattern shifts for naturally phrased queries. In this setting, the *BGE-large* model achieves the highest effectiveness across both metrics for $1 \leq k \leq 10$, outperforming the smaller models. For larger cut-offs ($k > 10$), the *BGE-small* model exceeds the *BGE-large* model in terms of *nDCG@k*, while exhibiting comparable performance with respect to *MRR@k*.

A similar embedding model size–dependent pattern is observed for hybrid retrieval configurations, although the magnitude of the performance differences varies across retrieval strategies. In sum, within the locally deployable, domain-specific retrieval setting examined in this thesis, the effect of embedding model size depends strongly on query formulation rather than following a uniform scaling trend.

6.3 Effects of Post-Retrieval Refinement

Post-retrieval refinement is employed to examine its impact on early ranking quality for retrieval configurations that exhibit the strongest recall under fixed candidate budgets. **RQ2** addresses the central research question by assessing whether this approach improves or degrades retrieval effectiveness across varying query formulations. The approach refines an initially retrieved candidate set by re-ranking candidate relevance and, optionally, by pre-filtering irrelevant candidate chunks prior to re-ranking based on a *SLM-as-a-judge* paradigm.

Table III presents a comprehensive comparison of *Recall@25* across all evaluated retrieval methods and embedding model sizes, demonstrating substantial variation across query types. For expert queries, hybrid retrieval using *BGE-small* achieves the highest recall at 75.7%. In contrast, for natural and naive formulations, semantic retrieval with *BGE-small* performs best, reaching 54.8% and 42.3%, respectively. These configurations are

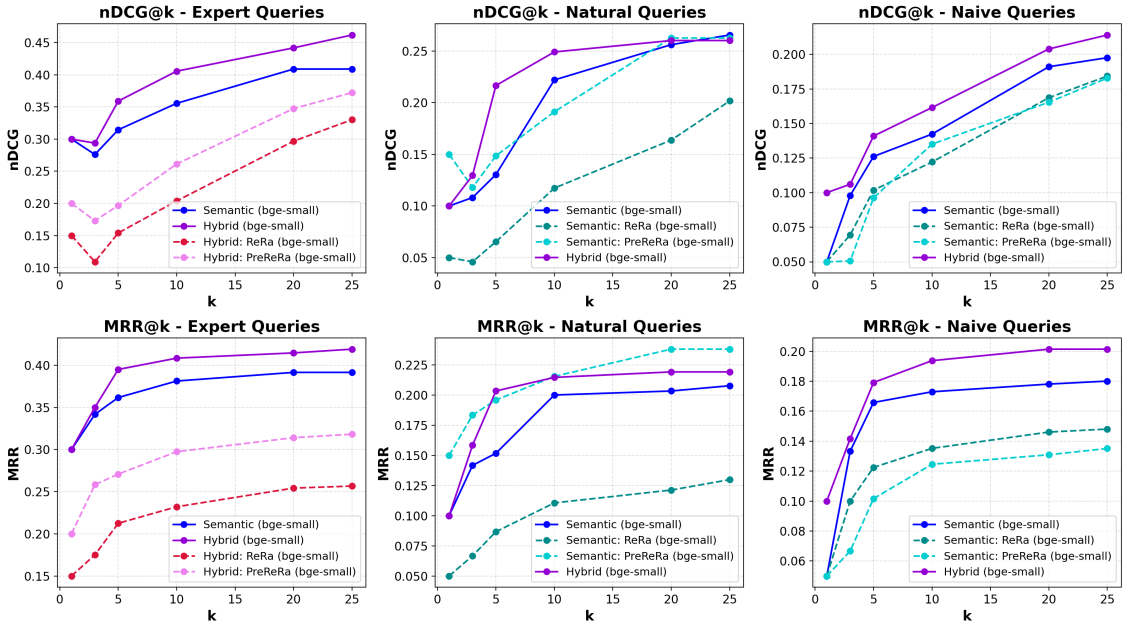

FIGURE 5: Comparison of baseline retrieval, re-ranking (*ReRa*), and pre-filtering followed by re-ranking (*PreReRa*) demonstrates the query-dependent and conditional effectiveness of post-retrieval refinement. Refinement improves early ranking quality for natural queries at small cut-offs, while substantially degrading performance for expert and naive queries.

therefore selected as the basis for subsequent post-retrieval refinement.

The impact of post-retrieval refinement on ranking quality and early retrieval performance is illustrated in Figure 5, comparing baseline retrieval, re-ranking (*ReRa*), and pre-filtering followed by re-ranking (*PreReRa*). Since pre-filtering may shorten the candidate list prior to re-ranking, the list is replenished with the next highest-ranked candidates from the initial retrieval to ensure comparability. To capture the full effect of *PreReRa*, only cut-off values below the minimum threshold of $k = 5$ across query sets are considered. For expert queries, neither re-ranking nor the combination of pre-filtering and re-ranking improves upon the baseline configuration, with re-ranking alone consistently performing worst across k. In contrast, for natural queries, post-retrieval refinement applied to semantic retrieval leads to clear improvements in ranking quality. At small cut-off values, the *PreReRa* configuration surpasses both the previously stronger hybrid baseline and its unrefined counterpart. This effect is consistently reflected in improved early relevance assessment (*nDCG@k*), indicating more accurate early placement of relevant documents. For naive queries, however, neither refinement strategy yields improvements over the baseline for either metric.

Collectively, these results demonstrate that the effectiveness of post-retrieval refinement is highly dependent on query formulation. While expert and naive queries show lim-

TABLE I: Evaluation of answer generation and retrieval adequacy under fixed retrieval configurations, illustrating the influence of query formulation, model choice, and reference perspective.

(a) Answer correctness (%) (per reference set) and faithfulness (%) across query formulations and models reveal a systematic decline in semantic alignment from expert to naive queries and a generally stronger performance of *Phi-3-Mini*. Faithfulness exhibits heightened sensitivity to query formulation, particularly for *Phi-3-Mini*. For each row, the overall maximum across all cells is highlighted in bold, and within each model, the higher of the two values is underlined.

Query Type	Retrieval Method	Answer Correctness												Faithfulness		
		A1			A2			A3			GPT 5.2					
		LLaMA	Phi	Δ	LLaMA	Phi	Δ	LLaMA	Phi	Δ	LLaMA	Phi	Δ	LLaMA	Phi	Δ
Expert	BM25	60.6	64.5	3.9	53.3	55.1	1.8	**57.2**	56.1	-1.1	46.0	52.4	6.4	70.5	87.8	17.3
Natural	Hybrid (BGE-large)	51.7	52.5	0.8	43.3	51.6	8.3	49.3	**54.4**	5.1	36.8	47.5	10.7	55.7	81.3	25.6
Naive	Hybrid (BGE-small)	48.7	49.9	1.2	44.3	50.7	6.4	47.7	**53.5**	5.8	36.4	43.9	7.5	47.6	61.4	13.8

(b) Context precision (%), context recall (%) are computed across reference perspectives, and retention rates (%) across models, all calculated across query formulations. While retrieval adequacy remains substantial, particularly for expert queries, both precision and recall decline with decreasing query specificity. Per row, the maximum value is highlighted in bold. Retention behaviour differs between models, with *LLaMA 3.2 3B Instruct* abstaining more frequently, highlighting model-specific strategies in handling evidential uncertainty.

Query Type	Retrieval Method	Context Precision				Context Recall				Retention Rate	
		A1	A2	A3	GPT 5.2	A1	A2	A3	GPT 5.2	LLaMA	Phi
Expert	BM25	84.6	95.0	82.9	**97.9**	78.3	78.3	73.3	**84.1**	20.0	0.0
Natural	Hybrid (BGE-large)	71.2	76.7	69.2	**94.2**	**73.3**	57.9	72.5	66.3	30.0	5.0
Naive	Hybrid (BGE-small)	53.3	71.7	62.9	**83.3**	57.5	68.8	56.7	**77.6**	35.0	10.0

ited sensitivity to ranking refinement, natural queries benefit from targeted post-retrieval refinement. Consequently, the best-performing methods at $k = 3$ in terms of ranking quality for expert, natural, and naive queries are semantic (*BGE-small*) and hybrid (*BGE-small*), semantic (*BGE-small*) + PreReRa and hybrid (*BGE-large*), and hybrid (*BGE-small*), respectively and will be further considered for the answer generation performance (Table IV).

6.4 Generation Trade-offs across Query Formulations

The trade-offs in answer correctness, faithfulness, and context grounding exhibited by SLMs under varying query formulations (**RQ3**) are reported in Table I. For each query type (expert, natural, naive), the retrieval configuration with the highest ranking quality and early relevance assessment at $k = 3$ was selected. In cases of equivalent retrieval performance, the computationally lighter configuration was preferred. Based on the top $k = 3$ retrieved chunks, answers were generated using two different generation models and evaluated against four alternative reference answer sets (three human annotators and one model-generated reference). All metrics were computed using the *RAGAS* framework with an external evaluation model. Abstention behaviour was handled explicitly. Responses indicating insufficient evidence (e.g., "I do not know") were treated as abstentions, excluded from the computation of answer correctness, and reported separately via the retention rate.

Across reference sets, answer correctness (Table Ia) is systematically influenced by both query formulation and model choice. Independent of the specific ground truth, scores generally decline from expert to natural to naive queries. Two exceptions occur for annotator A2 and GPT 5.2 under the *LLaMA* model, where correctness increases from natural to naive queries. Overall, this pattern indicates that increasing descriptiveness tends to reduce semantic alignment. Although the exact magnitude varies across annotators, the general trend remains stable.

Phi-3-Mini generally outperforms *LLaMA 3.2 3B Instruct*, with only one minor exception for expert queries compared with the reference set by annotator A3. The performance gap is particularly visible for the GPT-5.2 reference set, where both models achieve lower correctness scores relative to the human annotator references, suggesting differences in scope and informational framing. The largest discrepancies occur between GPT-5.2 and annotator A1: for *Phi-3-Mini*, the maximal deviation is observed for expert queries (12.1 percentage points) while for *LLaMA 3.2 3B Instruct*, it emerges for natural queries (14.9 percentage points). This confirms that answer correctness is sensitive not only to query formulation and model capacity but also to the evaluative perspective embedded in the reference answers.

For faithfulness (Table Ia), the effect of query formulation is even more pronounced. Both models attain their highest grounding scores for expert queries, with performance declining for natural and naive formulations. However, the magnitude and location of the strongest shift differ between models. For *Phi-3-Mini*, the largest change occurs between natural and naive queries (19.9 percentage points), whereas for *LLaMA 3.2 3B Instruct*,

the steepest drop is already observed between expert and natural queries (14.8 percentage points). This pattern indicates that *LLaMA 3.2 3B Instruct* consistently yields lower faithfulness scores across all query types.

Retention behaviour (Table Ib) further differentiates between models. *LLaMA 3.2 3B Instruct* shows significantly higher retention rates across all query types ranging between 20 and 35%. *Phi-3-Mini* shows a significantly lower retention rate ranging from 0–10% while still generally outperforming in terms of answer correctness, except in one case.

Context precision (Table Ib) declines consistently from expert to natural to naive queries across all reference sets. While precision is highest for expert queries, GPT-5.2 maintains comparatively elevated precision values across all query types, remaining above 80% even for naive formulations. This indicates that, irrespective of query specificity, the retrieved candidate chunks largely constitute relevant evidence with respect to that reference perspective.

Context recall (Table Ib) exhibits a similar overall tendency but with greater variability. Expert queries consistently achieve the highest recall across reference sets, whereas natural and naive queries show less stable behaviour. In particular, recall for natural queries does not uniformly exceed that of naive queries, reflecting fluctuations in evidence coverage across formulations. Considered jointly with answer correctness, context recall does not display a strictly proportional relationship, as higher recall does not consistently coincide with higher correctness scores across query types.

When considered jointly with answer correctness, this pattern suggests that *LLaMA 3.2 3B Instruct* is more likely to abstain when evidence is perceived as insufficient, while *Phi-3-Mini* produces fewer abstentions despite achieving higher answer correctness. This indicates that *LLaMA 3.2 3B Instruct* may be less effective at utilising the retrieved context to construct an answer for less specific queries, whereas *Phi-3-Mini* more consistently leverages the available evidence even under broader or less well-defined query formulations.

Overall, answer quality depends on query formulation, model characteristics, and reference perspective. Retrieval adequacy, as reflected in context precision and context recall, appears more strongly influenced by query formulation and reference perspective than by differences between generation models.

7 DISCUSSION

The following discussion examines how design decisions influence robustness under organisation-controlled deployment constraints, structured around evidence acquisition (**RQ1**, **RQ2**)

and evidence utilisation (**RQ3**).

The results demonstrate that retrieval behaviour is highly query-sensitive. Both ranking quality and early relevance decline monotonically from expert to naive queries. For example, *BM25* achieves an *nDCG@3* of 34.3% and an *MRR@3* of 35.8% for expert queries, but drops to 3.5% and 5.0%, respectively, for naive queries, indicating increasing retrieval difficulty as the technical specificity of queries decreases. Expert-level queries benefit from lexical retrieval, particularly *BM25*, due to precise domain terminology. In contrast, hybrid retrieval dominates for natural and naive formulations, while semantic retrieval increasingly outperforms lexical approaches as queries become more descriptive.

Embedding model size does not exhibit a uniform scaling effect when considering variation in query formulation. While *BGE-large* performs best for natural queries at small cut-offs, *BGE-small* achieves the highest values for expert and naive queries. The largest model therefore only provides advantages in specific configurations rather than demonstrating consistent scaling benefits. This pattern suggests that the training scope and data characteristics of the embedding model exert a stronger influence on retrieval performance than model size alone. Consequently, selecting the embedding model that best fits the domain is more effective than varying model scale across query formulations.

Post-retrieval refinement exhibits conditional effectiveness rather than uniform improvements. Although high *Recall@25* values indicate theoretical potential for performance gains, with hybrid retrieval using *BGE-small* reaching 75.7% for expert queries, refinement only improves ranking quality for natural queries at small cut-offs, yielding an increase of 5.0 percentage points in *nDCG@1*. For expert queries, refinement degrades performance substantially, reducing *nDCG@3* by 12.1 percentage points compared to the non-refined hybrid configuration. The same degradation is observed for naive queries. These findings indicate that refinement is most effective when queries are moderately specific, providing sufficient contextual cues for re-ranking without being overly constrained or overly ambiguous. In contrast, highly precise expert queries already retrieve focused candidate sets, leaving little room for improvement, while ambiguous naive queries introduce noise that refinement models struggle to resolve.

Answer generation was analysed under fixed retrieval configurations to isolate downstream effects. Context precision and recall remain substantial for several reference perspectives, particularly for GPT-5.2, but show greater variability across human annotators. For instance, under naive queries, context recall reaches 77.6% for GPT-5.2 and 57.5% for annotator A1, corresponding to a difference of 20.1 percentage points. Although retrieval adequacy tends to decline for less specific queries, this pattern is not consistent across reference perspectives, and its absolute level varies substantially depending on how rel-

evance is operationalised. This indicates that differences in answer correctness cannot be attributed solely to missing evidence but must also reflect how retrieved context is interpreted and utilised by the generation models.

Despite this comparatively substantial retrieval adequacy, answer correctness declines from expert to natural to naive formulations. Under naive queries, answer correctness for *Phi-3-Mini* falls to 43.9% for GPT-5.2 and 49.9% for annotator A1, even though context recall and precision are higher for GPT-5.2 than for A1. Higher context recall and precision therefore do not correspond to higher correctness scores across reference perspectives. This divergence reveals a retrieval generation utilisation gap. Relevant evidence is retrieved but not proportionally transformed into correct answers. The bottleneck therefore lies not in evidence acquisition but in context utilisation.

Model choice moderates this utilisation gap. As both *LLaMA 3.2 3B Instruct* and *Phi-3-Mini* operate on identical retrieved context, differences in correctness and faithfulness reflect differences in contextual reasoning rather than evidence availability. *Phi-3-Mini* generally achieves higher correctness, with gains of up to 10.7 percentage points for natural queries under the GPT-5.2 reference perspective, while exhibiting substantially lower retention rates ranging from 0-10% compared to 20-35% for *LLaMA 3.2 3B Instruct*. This indicates more effective utilisation of available context during answer construction. However, this improved utilisation comes at computational cost since *Phi-3-Mini* requires approximately twice the generation latency compared to *LLaMA 3.2 3B Instruct* under identical CPU-based, quantised deployment conditions (see Table II). In contrast, *LLaMA 3.2 3B Instruct* abstains more frequently without achieving higher alignment, suggesting a more conservative but not necessarily more effective strategy.

Answer correctness is additionally shaped by the reference perspective. The semantic overlap heatmap (Figure 6) demonstrates substantial but incomplete agreement between annotator answers, indicating that even expert-generated ground truths reflect alternative framings of the same information need. Differences in expected explanatory scope therefore influence measured correctness. Evaluation outcomes thus depend not only on model capability but also on how answer adequacy is operationalised.

Faithfulness further reinforces the utilisation gap. Grounding is highest for expert queries and declines for natural and naive formulations, with a particularly pronounced drop for *Phi-3-Mini* between natural and naive queries, decreasing from 81.3% to 61.4%. Models rely on clearly specified query intent to align generated content with retrieved evidence. When scope becomes less specific, grounding deteriorates despite the continued availability of relevant context.

Taken together, the findings indicate that query formulation functions as a central system-

level design factor in RAG pipelines under organisation-controlled deployment constraints, as retrieval strategies, and generation performance both vary systematically with linguistic specificity. Although retrieval adequacy often remains substantial, answer quality depends on how effectively models transform retrieved evidence into grounded responses and on how evaluative reference perspectives define expected scope. The observed utilisation gap shows that robust system behaviour cannot be inferred from retrieval metrics alone but emerges from the interaction between retrieval configuration, model-level reasoning capacity, and evaluation perspective. Retrieved context should therefore be treated as evidential support rather than as an implicit guarantee of answer correctness in high-stakes organisational decision-support settings.

8 CONCLUSION AND FUTURE WORK

This thesis investigated how retrieval and generation components of retrieval-augmented generation (RAG) systems can be designed and evaluated to remain robust under natural variation in query formulation and organisation-controlled deployment constraints. A fully locally deployable RAG architecture based exclusively on small language models was implemented and systematically analysed under governance-driven restrictions.

With regard to **RQ1**, the results show that retrieval effectiveness is strongly sensitive to query formulation. Ranking quality declines substantially from expert to naive queries, with early precision measures deteriorating by up to 30 percentage points in certain configurations. The relative effectiveness of lexical, semantic, and hybrid retrieval shifts with linguistic specificity, and embedding scale does not consistently improve retrieval performance in the examined setting. Instead, domain fit appears more decisive than parameter size. Regarding **RQ2**, post-retrieval refinement exhibits conditional rather than uniform effectiveness. While refinement improves early ranking quality in some configurations, it degrades performance in others. These findings suggest that high recall levels are not consistently translated into improved early ranking quality after refinement, indicating unrealised potential rather than stable benefit. When small language models are used for refinement, performance appears sensitive to model capability and to the structure and relevance distribution of the initially retrieved candidates. Its robustness under constrained deployment therefore requires further systematic investigation. In response to **RQ3**, answer correctness and faithfulness decline from expert to naive queries even when relevant context remains available. Under naive queries, correctness for *Phi-3-Mini* remains around 50% across reference perspectives despite substantial context recall. Correctness does not scale proportionally with retrieval adequacy. This pattern reveals a retrieval-generation utilisation gap in which available evidence is not consistently transformed into

correct and grounded answers, consistent with recent findings that relevant context alone does not guarantee faithful generation [33]. Differences between generation models reflect variation in contextual reasoning rather than evidence availability, and evaluation outcomes vary across reference perspectives.

Taken together, the findings demonstrate that robustness under organisation-controlled deployment constraints emerges from the interaction between query formulation, retrieval configuration, generation capability, and evaluation perspective rather than from model scale alone. Query formulation functions as a central design variable, and retrieval metrics alone are insufficient indicators of system robustness because effective evidence utilisation constitutes a primary bottleneck within the examined setting.

Future work should examine mechanisms to reduce the retrieval–utilisation gap observed under small-model constraints. One direction is the incorporation of a lightweight query rewriting component, building on prior work on query adaptation [40]. Transforming user inputs into structurally meaningful formulations may facilitate more effective utilisation of retrieved evidence by small language models. An alternative direction would be to investigate query-dependent retrieval adaptation mechanisms that condition configuration choices on structural properties of the input query. Closely related to these directions, further research should systematically distinguish the sources of the observed refinement limitations. Specifically, it should examine whether these limitations stem from optimisable system-level configuration factors or from constraints inherent to model capacity. It should also analyse whether a utilisation gap analogous to that observed in generation manifests during refinement. These approaches are not mutually exclusive and could be combined within a modular pipeline architecture, allowing individual components to be optimised independently under deployment constraints.

9 CODE AND DATA AVAILABILITY

The source code developed for this work, together with the regulatory questions, and the reference answer sets, is maintained in a private repository. Access is available upon reasonable request by email, subject to the author's permission. The guidance documents used as the knowledge base are publicly available from the ECHA [44].

10 APPENDIX

10.1 Semantic Similarity Heatmap of Reference Answers

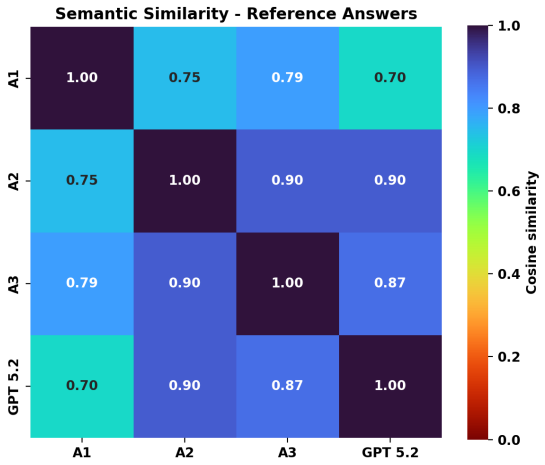

FIGURE 6: Semantic similarity heatmap of reference answer sets (A1–A3 and GPT-5.2). While all pairwise overlaps remain above 70%, indicating substantial agreement on the underlying informational content, noticeable variation persists across annotators. Some reference sets exhibit systematically higher similarity, reflecting closer alignment in scope and explanatory depth.

10.2 Annotator Background

The reference set used as ground truth for retrieval evaluation was constructed by three domain professionals serving as annotators. Annotator 1 is a chemical regulations expert with professional experience in the interpretation and application of chemical legislation. Annotator 2 works operationally at the interface of regulatory requirements and their technical implementation within enterprise compliance systems. Annotator 3 is a PhD chemist and involved in implementing technical solutions derived from regulatory requirements.

10.3 Runtime Efficiency

TABLE II: Runtime efficiency across retrieval methods and generation models, reported as total latency, time-to-first-token (TTFT), and token throughput.

Query	**Method**	**Latency (s)**		**TTFT (s)**		**Throughput (tok/s)**	
		LLaMA	Phi	LLaMA	Phi	LLaMA	Phi
Expert	BM25	6.29	11.71	4.27	6.54	285.8	155.5
Natural	Hybrid	6.38	11.08	4.25	6.51	287.7	168.8
Naive	Hybrid	5.87	11.32	4.24	6.46	301.4	162.4

10.4 Recall Performance

Notably, TF-IDF achieves higher *Recall@25* than BM25 for natural and naive queries, despite BM25 consistently yielding stronger ranking quality (Table IV). Since all chunks are of fixed length, this difference does not stem from document length normalisation but from BM25s term frequency saturation, which dampens the score contribution of repeatedly occurring terms. TF-IDFs linear term frequency scaling instead inflates scores for chunks that repeat query terms frequently, pulling more loosely relevant chunks into the top-25 set and raising recall at the cost of ranking precision, as reflected in the *nDCG@25* values reported in Table IV.

TABLE III: *Recall@25* (%) varies substantially across expert, natural, and naive queries. Hybrid retrieval with the *BGE-small* model performs best for expert queries, while semantic retrieval with the same model achieves the highest recall for natural and naive formulations.

Method / Model	**Recall@25**		
	Expert	Natural	Naive
TF–IDF	62.2	41.7	31.7
BM25	69.7	34.2	26.7
Hybrid (BGE-small)	**75.7**	46.8	41.4
Hybrid (BGE-base)	71.2	48.5	32.7
Hybrid (BGE-large)	70.4	46.0	31.8
Semantic (BGE-small)	67.6	**54.8**	**42.3**
Semantic (BGE-base)	50.2	50.6	28.9
Semantic (BGE-large)	57.4	52.8	32.4

10.5 Ranking Quality

TABLE IV: Ranking quality measured by *nDCG@k* at $k = 1$, $k = 3$, and $k = 25$ across retrieval and post-retrieval configurations (values in %). Per column, the maximum value is highlighted in bold.

Method / Model	**nDCG@1**			**nDCG@3**			**nDCG@25**		
	Expert	Natural	Naive	Expert	Natural	Naive	Expert	Natural	Naive
TF–IDF	25.0	10.0	5.0	27.6	8.5	2.3	39.3	19.7	11.3
BM25	25.0	10.0	5.0	**34.3**	11.7	3.5	45.6	19.9	11.9
Semantic (BGE-small)	**30.0**	10.0	5.0	27.6	10.8	9.8	40.9	26.5	19.8
Semantic (BGE-base)	20.0	0.0	5.0	16.9	3.7	7.2	29.2	19.3	13.1
Semantic (BGE-large)	20.0	10.0	0.0	18.4	12.9	5.6	31.1	26.2	13.5
Hybrid (BGE-small)	**30.0**	10.0	**10.0**	29.4	13.0	**10.6**	**46.2**	26.0	**21.4**
Hybrid (BGE-base)	15.0	0.0	5.0	19.6	7.8	6.3	37.7	21.5	14.9
Hybrid (BGE-large)	20.0	**15.0**	0.0	24.1	**16.5**	3.8	39.3	**27.3**	13.5
Semantic (BGE-small) + PreReRa	–	**15.0**	5.0	–	11.8	5.1	–	26.3	18.3
Hybrid (BGE-small) + PreReRa	20.0	–	–	17.3	–	–	37.2	–	–
Semantic (BGE-small) + ReRa	–	5.0	5.0	–	4.6	6.9	–	20.2	18.4
Hybrid (BGE-small) + ReRa	15.0	–	–	10.9	–	–	33.0	–	–

TABLE V: Ranking quality measured by *MRR@k* at $k = 1$ and $k = 3$ across retrieval and post-retrieval configurations (values in %). Per column, the maximum value is highlighted in bold.

Method / Model	**MRR@1**			**MRR@3**		
	Expert	Natural	Naive	Expert	Natural	Naive
TF–IDF	25.0	10.0	5.0	33.3	10.0	5.0
BM25	25.0	10.0	5.0	**35.8**	12.5	5.0
Semantic (BGE-small)	**30.0**	10.0	5.0	34.2	14.2	13.3
Semantic (BGE-base)	20.0	0.0	5.0	24.2	3.3	10.0
Semantic (BGE-large)	20.0	10.0	0.0	26.7	15.8	9.2
Hybrid (BGE-small)	**30.0**	10.0	**10.0**	35.0	15.8	**14.2**
Hybrid (BGE-base)	15.0	0.0	5.0	25.0	7.5	9.2
Hybrid (BGE-large)	20.0	**15.0**	0.0	30.8	**19.2**	5.8
Semantic (BGE-small) + PreReRa	–	**15.0**	5.0	–	18.3	6.7
Hybrid (BGE-small) + PreReRa	20.0	–	–	25.8	–	–
Semantic (BGE-small) + ReRa	–	5.0	5.0	–	6.7	10.0
Hybrid (BGE-small) + ReRa	15.0	–	–	17.5	–	–

10.6 Illustrative Reference Answer

TABLE VI: Illustrative reference answer for the question *"When must a Safety Data Sheet (SDS) be updated?"*, drawn from the *GPT-5.2*-generated reference answer set, one of the four reference answer sets used as ground truth for answer-generation evaluation. The complete set of regulatory questions and reference answer sets is stored in the repository referenced in Section 9.

Question	When must a Safety Data Sheet (SDS) be updated?
Reference Answer	Under REACH/CLP, a supplier must update a Safety Data Sheet (SDS) without delay in particular when: • New information becomes available that may affect risk management measures (e.g. new hazard or exposure information) or new information on hazards becomes available (including changes to classification/labeling). • An authorisation under REACH is granted or refused for the substance/mixture. • A restriction under REACH is imposed. Additionally: • The revised SDS must be provided free of charge to all former recipients to whom the substance/mixture was supplied within the previous 12 months. • The SDS should indicate that it has been revised (e.g. revision date/version). (Primary legal basis: REACH Article 31(9); with related duties under CLP where classification/label changes drive SDS updates.)

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.

[2] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021, pp. 3784–3803.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[4] S. Nanua, R. Steward, B. Neely, M. Datto, and K. Youens, "Retrieval-augmented generation for interpreting clinical laboratory regulations using large language models," *Journal of Pathology Informatics*, vol. 19, p. 100520, 2025.

[5] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval-augmented language models," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 1–43, 2023.

[6] A. Asai, Z. Zhong, D. Chen, P. W. Koh, L. Zettlemoyer, H. Hajishirzi, and W. t. Yih, "Reliable, adaptable, and attributable language models with retrieval," 2024, arXiv preprint arXiv:2403.03187.

[7] S. Perçin, X. Su, Q. S. Syed, P. Howard, A. Kuvshinov, L. Schwinn, and K.-U. Scholl, "Investigating the robustness of retrieval-augmented generation at the query level," in *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM2)*, 2025, pp. 439–457.

[8] R. Kalra, Z. Wu, A. Gulley, A. Hilliard, X. Guan, A. Koshiyama, and P. C. Treleaven, "HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications," in *Proceedings of the 1st Workshop*

on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), 2024, pp. 237–256.

[9] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2008.

[10] H. A. Simon, *Administrative behavior: A study of decision-making processes in administrative organizations*. New York: Macmillan, 1947.

[11] G. A. Gorry and M. S. S. Morton, "A framework for management information systems," *Management Science*, vol. 17, no. 1, pp. 55–70, 1971.

[12] G. Salton, *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[13] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[14] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 232–241.

[15] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, arXiv preprint arXiv:1301.3781.

[17] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 6769–6781.

[18] M. Reuter, T. Lingenberg, R. Liepina, F. Lagioia, M. Lippi, G. Sartor, A. Passerini, and B. Sayin, "Towards reliable retrieval in rag systems for large legal datasets," in *Proceedings of the Natural Legal Language Processing Workshop (NLLP)*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 17–30.

[19] M. Aljohani, J. Hou, S. Kommu, and X. Wang, "A comprehensive survey on the trustworthiness of large language models in healthcare," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 6720–6748.

[20] P. J. Pesch, "Potentials and challenges of large language models in the context of administrative decision-making," *European Journal of Risk Regulation*, vol. 16, no. 1, pp. 76–95, 2025.

[21] J. Jain, N. Dhanasekaran, and M. T. Diab, "From complexity to clarity: AI/NLP's role in regulatory compliance," in *Findings of the Association for Computational Linguistics (ACL) 2025*, 2025, pp. 26 629–26 641.

[22] European Chemicals Agency, "Understanding reach," 2024, available online: https://echa.europa.eu/regulations/reach/understanding-reach (accessed January 6, 2026).

[23] W. Bridges, "The creation of a competent global regulatory workforce," *Frontiers in Pharmacology*, vol. 10, 2019.

[24] M. Biron, K. Turgeman-Lupo, and O. Zaid-Dominik, "Contextualizing the usefulness of knowledge received from retiring employees: Leader behaviour and organisational culture," *Knowledge Management Research & Practice*, vol. 22, no. 6, pp. 620–631, 2024.

[25] McKinsey & Company, "The state of AI in 2025: Agents, innovation, and transformation," McKinsey & Company, Tech. Rep., 2025, available online: https://www.mckinsey.com/featured-insights/artificial-intelligence/the-state-of-ai-in-2025 (accessed December 26, 2025).

[26] IBM, "AI in action: How enterprises are driving business value with ai," 2024, available online: https://www.ibm.com/think/reports/ai-in-action (accessed January 11, 2025).

[27] European Parliament and Council of the European Union, "Regulation (EC) No 1907/2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH)," 2006, art. 10, 118–119.

[28] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, 2023.

[29] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 (General Data Protection Regulation)," 2016, official Journal of the European Union, L119, pp. 1–88. Art. 4 and Art. 28.

[30] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021, pp. 874–880.

[31] A. C. et al., "PaLM: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[32] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024, arXiv preprint arXiv:2312.10997.

[33] L. Hagström, S. V. Marjanovic, H. Yu, A. Arora, C. Lioma, M. Maistro, P. Atanasova, and I. Augenstein, "A reality check on context utilisation for retrieval-augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 19 691–19 730.

[34] D. Sachan, M. Lewis, M. Joshi, A. Aghajanyan, W.-T. Yih, J. Pineau, and L. Zettlemoyer, "Improving passage retrieval with zero-shot question generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3781–3797.

[35] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, and M. Bendersky, "Large language models are effective text rankers with pairwise ranking prompting," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 1504–1518.

[36] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren, "Is ChatGPT good at search? investigating large language models as re-ranking agents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14 918–14 937.

[37] S. Zhuang, H. Zhuang, B. Koopman, and G. Zuccon, "A setwise approach for effective and highly efficient zero-shot ranking with large language models," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 38–47.

[38] B. Nouriinanloo and M. Lamothe, "Re-ranking step by step: Investigating pre-filtering for re-ranking with large language models," 2024, arXiv preprint arXiv:2406.18740.

[39] N. J. Belkin, R. N. Oddy, and H. M. Brooks, "ASK for information retrieval: Part I. background and theory," *Journal of Documentation*, vol. 38, no. 2, pp. 61–71, 1982.

[40] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting in retrieval-augmented large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5303–5315.

[41] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, and S. Wang, "A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, 2025.

[42] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury, "Efficiently teaching an effective dense retriever with balanced topic-aware sampling," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 113–122.

[43] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[44] European Chemicals Agency, "Guidance documents," 2026, available online: https://echa.europa.eu/guidance-documents/guidance-on-reach (accessed January 13, 2026).

[45] K. Na, "The effects of cognitive load on query reformulation: Mental demand, temporal demand and frustration," *Aslib Journal of Information Management*, vol. 73, no. 3, pp. 436–453, 2021.

[46] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Proceedings of the NeurIPS Datasets and Benchmarks Track*, 2021.

[47] S. Es, J. James, L. E. Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval-augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.