TU WIEN Informatics

Retrieval Augmented Generation: A Multi-Stage Architecture for Verbatim Financial Question Answering

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Tuvshin Selenge, BSc
Matrikelnummer 11815791

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass. Gábor Recski, PhD
Mitwirkung: Ádám Kovács, BSc MSc

Wien, 3. August 2026

Tuvshin Selenge	Gábor Recski

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

Retrieval Augmented Generation: A Multi-Stage Architecture for Verbatim Financial Question Answering

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Tuvshin Selenge, BSc
Registration Number 11815791

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. Gábor Recski, PhD
Assistance: Ádám Kovács, BSc MSc

Vienna, August 3, 2026

Tuvshin Selenge	Gábor Recski

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Tuvshin Selenge, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 3. August 2026

Tuvshin Selenge

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Danksagung

Diese Arbeit wäre ohne die Unterstützung meiner Familie, meiner Freunde, meiner Kollegen und meiner Betreuer nicht möglich gewesen.

Mein besonderer Dank gilt meinen Eltern für ihre stetige Ermutigung und ihren emotionalen Rückhalt während meines gesamten Studiums. Als erstes Mitglied meiner Familie, das eine Diplomarbeit verfasst, bedeuteten mir ihr Vertrauen und ihre Unterstützung insbesondere in herausfordernden Phasen sehr viel.

Darüber hinaus möchte ich meinem Kollegen bei der Bank, Mark Tallai, für seine wertvolle Unterstützung und seinen Beitrag zu dieser Arbeit danken.

Abschließend möchte ich meinem Betreuer, Gábor Recski, sowie meinem Mitbetreuer, Ádám Kovács, meinen aufrichtigen Dank für ihre fachliche Begleitung, ihr konstruktives Feedback und ihre kontinuierliche Unterstützung während der Entstehung dieser Arbeit aussprechen. Ihre Expertise und Ermutigung waren entscheidend dafür, dass ich diese Arbeit erfolgreich abschließen konnte.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

This thesis would not have been possible without the support of my family, friends, colleagues, and supervisors.

I am especially grateful to my parents for their constant encouragement and emotional support throughout my studies. As the first person in my family to write a master's thesis, their belief in me and their support during challenging periods meant a great deal.

I would also like to thank my colleague at the bank, Mark Tallai, for his valuable support and contribution to this thesis.

Finally, I would like to express my sincere gratitude to my supervisor, Gábor Recski, and my co-supervisor, Ádám Kovács, for their guidance, constructive feedback, and continuous support throughout the development of this thesis. Their expertise and encouragement were essential in helping me bring this work to completion.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Finanzielle Due-Diligence-Prozesse erfordern Antworten, die korrekt, nachvollziehbar und unmittelbar durch Quelldokumente belegt sind. Die Identifikation relevanter Informationen in langen und strukturell heterogenen Geschäftsberichten ist jedoch weiterhin eine manuelle und zeitaufwendige Aufgabe, während konventionelle generative Fragebeantwortungssysteme das Risiko bergen, unbelegte oder nur unzureichend prüfbare Antworten zu erzeugen. Diese Arbeit untersucht, ob eine mehrstufige, auf wörtlich übernommenen Textspannen basierende Retrieval-Augmented-Generation-(RAG)-Architektur die Beantwortung von Due-Diligence-Fragebögen (DDQs) im Bankenumfeld unterstützen kann.

Aufbauend auf dem VerbatimRAG-Framework kombiniert die vorgeschlagene Pipeline Query Rewriting, Multi-Query Expansion, hybrides Dense- und Sparse-Retrieval, Cross-Encoder-Reranking sowie eine LLM-basierte Extraktion wörtlicher Textspannen. Extrahierte Textspannen werden nur dann beibehalten, wenn sie mit dem zugrunde liegenden Quelltext abgeglichen und darin verifiziert werden können. Da bislang kein annotierter Datensatz für die Beantwortung von AFME-basierten DDQ-Fragen anhand von Geschäftsberichten existiert, besteht ein weiterer Beitrag dieser Arbeit in der Erstellung eines neuen Evaluationsdatensatzes auf Grundlage der Geschäftsberichte des Jahres 2024 von vier österreichischen Bankinstituten. Der Datensatz umfasst sechzehn Fragen pro Institut sowie manuell annotierte Referenz-Textspannen. Die Annotationen für die Raiffeisen Bank International wurden gemeinsam mit einem Due-Diligence-Praktiker des Instituts erstellt, während die Annotationen für die übrigen drei Banken ausschließlich vom Autor angefertigt wurden.

Die auf dem expertenunterstützt annotierten Datensatz der Raiffeisen Bank International durchgeführte Retrieval-Ablationsstudie zeigt, dass die vollständige mehrstufige Konfiguration die Hit Rate gegenüber der Baseline von 0.583 auf 1.000 erhöht. Unter den untersuchten Rerankern erzielt `BAAI/bge-reranker-v2-m3` die beste Retrieval-Leistung. Bei der Span-Extraktion erreichen `google/gemini-3-flash-preview` und `openai/gpt-5.1` die besten Ergebnisse und erzielen jeweils einen RapidFuzz@75-F1-Wert von 0.690. Die Abstentionsgenauigkeit bei nicht beantwortbaren Fragen variiert erheblich zwischen den Modellen und reicht von 1.000 für `openai/gpt-5.1` bis 0.200 für `google/gemini-2.5-flash-lite`. Die bankübergreifende Evaluation zeigt, dass sich die Retrieval-Architektur gut auf andere Institute übertragen lässt und für alle vier

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Banken eine Hit Rate von mindestens 0.909 erreicht. Die wörtliche Span-Extraktion reagiert hingegen empfindlicher auf institutsspezifische Berichtsstile und Annotationsmerkmale, während die Zuverlässigkeit der Abstention sowohl zwischen Modellen als auch zwischen Instituten erheblich variiert.

Die Ergebnisse legen nahe, dass die vorgeschlagene Architektur die Lokalisierung relevanter Textpassagen in Human-in-the-Loop-Due-Diligence-Prozessen unterstützen und zugleich die Nachvollziehbarkeit der Quellen gewährleisten kann. Verbleibende Extraktionsfehler und ein uneinheitliches Abstentionsverhalten zeigen jedoch, dass das System noch nicht für autonome Entscheidungsprozesse geeignet ist und eine menschliche Prüfung der extrahierten Textspannen weiterhin erforderlich bleibt.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

Financial due diligence requires answers that are accurate, traceable, and directly supported by source documents. However, identifying relevant evidence within long, structurally heterogeneous annual reports remains a manual and time-consuming task, while conventional generative question-answering systems risk producing unsupported or insufficiently auditable responses. This thesis investigates whether a multi-stage, verbatim-grounded retrieval-augmented generation (RAG) architecture can support due diligence questionnaire (DDQ) answering in the banking domain.

Building on the VerbatimRAG framework, the proposed pipeline combines query rewriting, multi-query expansion, hybrid dense and sparse retrieval, cross-encoder reranking, and LLM-based verbatim span extraction. Extracted evidence is retained only if it can be verified against the underlying source text. As no annotated resource exists for AFME-style DDQ answering over annual reports, a further contribution of this thesis is a newly constructed evaluation dataset based on the 2024 annual reports of four Austrian banking institutions. The dataset comprises sixteen questions per institution together with manually annotated gold evidence spans. The Raiffeisen Bank International annotations were created jointly with a due diligence practitioner at the institution, whereas the annotations for the remaining three banks were produced by the author alone.

The retrieval ablation study, conducted on the expert-supported Raiffeisen Bank International dataset, shows that the complete multi-stage configuration improves Hit Rate from 0.583 for the baseline to 1.000. Among the evaluated rerankers, `BAAI/bge-reranker-v2-m3` achieves the strongest retrieval performance. For span extraction, `google/gemini-3-flash-preview` and `openai/gpt-5.1` obtain the best results, both reaching a RapidFuzz@75 F1 score of 0.690. Abstention accuracy on unanswerable questions varies substantially across models, ranging from 1.000 for `openai/gpt-5.1` to 0.200 for `google/gemini-2.5-flash-lite`. Cross-bank evaluation shows that the retrieval architecture generalizes well across institutions, achieving a Hit Rate of at least 0.909 for all four banks. Verbatim extraction remains more sensitive to institution-specific reporting styles and annotation characteristics, while abstention reliability varies substantially across both models and institutions.

These findings suggest that the proposed architecture can support evidence localization within human-in-the-loop due diligence workflows while preserving source traceability. However, residual extraction errors and inconsistent abstention behavior indicate that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the system is not yet suitable for autonomous decision-making and that human review of extracted evidence remains necessary.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem statement . . . 2
1.3 Research questions . . . 2
1.4 Structure of the Thesis . . . 3

2 Background **5**
2.1 Information Retrieval . . . 5
2.2 Large Language Models (LLMs) . . . 9
2.3 Retrieval-Augmented Generation (RAG) . . . 12
2.4 VerbatimRAG . . . 16

3 Use Case **21**
3.1 Overview . . . 21
3.2 Due Diligence as a Financial Question-Answering Task . . . 21
3.3 Dataset Construction . . . 22
3.4 Challenges of Automated Due Diligence Question Answering . . . 28
3.5 Chapter Summary . . . 30

4 Methodology **31**
4.1 Overview . . . 31
4.2 System Architecture . . . 32
4.3 Data and Experimental Setup . . . 39
4.4 Experimental Design . . . 42
4.5 Evaluation Metrics . . . 45

5 Results and Discussion **49**
5.1 Retrieval Ablation Results . . . 49

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.2 Reranker Comparison Results 50
5.3 Extraction Model Comparison Results 51
5.4 Cross-Bank Generalization Results 54
5.5 Discussion 57

6 Conclusion **61**

Overview of Generative AI Tools Used **65**

List of Figures **67**

List of Tables **69**

Bibliography **71**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

Financial due diligence requires precise, auditable, and evidence-based answers grounded in extensive regulatory and corporate disclosure documents. Financial institutions are regularly required to provide verifiable information about their clients, counterparties, and own operations to regulators, supervisory authorities, and business partners. Such due diligence processes are essential for the detection and prevention of money laundering, fraud, and terrorist financing, while also enabling counterparties and investors to assess institutional risk. As a result, banks must respond to due diligence requests with information that is both reliable and traceable to source documentation.

1.1 Motivation

In practice, answering financial due diligence questions remains a highly manual and time-intensive task. Practitioners often need to examine hundreds of pages of annual reports, disclosures, and governance documents in order to identify specific financial figures, ownership structures, governance statements, or risk-related information. This challenge is particularly evident in large and densely structured annual reports, where relevant evidence may be scattered across narrative text, tabular disclosures, and appendices.

The annual reports of Raiffeisen Bank International, Erste Group, BAWAG, and UniCredit Bank Austria for the year 2024 serve as the empirical basis of this study. These documents illustrate the length, density, and structural complexity that characterize financial reporting in practice. These reports were selected because they represent institutions of varying size and reporting scope within the Austrian banking sector, all subject to comparable regulatory disclosure requirements, thereby enabling a controlled cross-institution evaluation.

1.2 Problem statement

The growing capability of large language models has prompted interest in automating document-intensive compliance workflows in the financial sector. However, regulatory and due diligence settings impose a constraint that standard generation-based systems cannot readily satisfy: answers must be directly attributable to verifiable source text, since even minor inaccuracies may carry legal, financial, or compliance consequences Iaroshev et al. (2024).

Retrieval-Augmented Generation (RAG) systems address this challenge by coupling a retrieval component, which searches a document corpus for relevant passages, with a generation component, typically a large language model, that synthesizes an answer conditioned on the retrieved evidence Lewis et al. (2020). This architecture differs from purely parametric LLM-based question answering, in which the model relies solely on knowledge encoded in its parameters during training and cannot verify its output against external, up-to-date source material, making such systems prone to hallucination and difficult to audit in high-stakes settings.

Nevertheless, within RAG pipelines applied to long and heterogeneous financial reports, a persistent "needle in a haystack" challenge remains, as relevant information is often embedded within extensive textual and tabular structures Iaroshev et al. (2024). The retrieval component of standard, single-stage RAG pipelines often fails to surface the most relevant evidence reliably, particularly when the information sought is expressed only indirectly or distributed across multiple sections of a document.

To address these limitations, this thesis builds upon the VerbatimRAG framework Kovács et al. (2025), using its publicly available open-source implementation as the technical foundation for a financial question-answering pipeline. VerbatimRAG is designed to extract and return verbatim evidence spans from retrieved documents rather than generate free-form text. Specifically, a lightweight extractor first identifies which sentences within the retrieved passages are relevant to the query, after which these sentences are inserted directly into a question-specific response template rather than being paraphrased or freely regenerated by the language model. Within this work, the framework is configured and evaluated in a realistic financial due diligence setting in order to assess its suitability for high-stakes reporting tasks.

1.3 Research questions

The objective of this thesis is to conduct a domain-specific evaluation of a VerbatimRAG-based architecture in the context of financial due diligence. While VerbatimRAG provides the technical foundation, the central aim of this work is to assess its effectiveness as a specialized tool for answering due diligence questions over long and complex financial documents. To this end, the study systematically compares an optimized multi-stage pipeline against a minimal VerbatimRAG-style baseline and evaluates the contribution of its individual retrieval and extraction components.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Beyond this system-level evaluation, a central contribution of this thesis is the construction of the evaluation dataset itself. Since no annotated resource exists for answering Association for Financial Markets in Europe (AFME) due diligence questions over annual reports, the reference annotations used in this work were newly created: the Raiffeisen Bank International annotations in collaboration with a due diligence practitioner at the institution, and the annotations for the three remaining banks manually as part of this thesis. The resulting dataset is described in detail in Section 3.3.

The thesis addresses the following research questions:

1. **RQ1: Retrieval Performance** How do the individual retrieval-side components of the adapted VerbatimRAG-based pipeline, including query rewriting, multi-query expansion, and reranking, affect retrieval performance in financial due diligence documents compared to a baseline configuration?
2. **RQ2: Extraction Performance** How accurately does the system identify and extract verbatim evidence spans from dense financial text and tabular layouts, and to what extent do extracted spans align with manually annotated gold references at varying levels of matching strictness?
3. **RQ3: Faithfulness and Abstention** To what extent does the system abstain from returning evidence when no valid answer exists in the document corpus, and how reliably do extracted spans remain verbatim with respect to the underlying source text?

Code Availability The implementation developed for this thesis is publicly available under the MIT License in the accompanying repository Selenge (2026). The repository extends the open-source VerbatimRAG project with the thesis-specific retrieval, evaluation, and financial-domain components described in Chapter 4.

1.4 Structure of the Thesis

The remainder of this thesis is structured as follows:

- **Chapter 2** reviews the theoretical foundations of information retrieval, large language models, retrieval-augmented generation, and the VerbatimRAG framework that forms the methodological foundation of this thesis.
- **Chapter 3** introduces the financial due diligence use case, the selected annual reports, and the evaluation corpus of due diligence questions and annotated gold evidence spans.
- **Chapter 4** presents the methodology, including the multi-stage pipeline architecture and the systematic ablation study covering query rewriting, query variation, and reranking components.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **Chapter 5** presents the results and discussion, providing a quantitative analysis of retrieval, extraction, and faithfulness performance, and discussing the findings with particular emphasis on the reliability and practical suitability of the approach in financial reporting settings.
- **Chapter 6** concludes the thesis by summarizing the key findings and identifying the most reliable RAG configuration within the scope of this study.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Background

In this chapter, the theoretical foundations of Information Retrieval (IR), Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and the VerbatimRAG framework are introduced. The chapter progresses from classical and neural information retrieval, through the capabilities and limitations of LLMs and standard RAG architectures, to the extractive VerbatimRAG framework that forms the methodological foundation of this thesis, laying the conceptual groundwork for the methodology presented in Chapter 4.

2.1 Information Retrieval

2.1.1 Classical Information Retrieval

Information Retrieval (IR) concerns the representation, organization, and search of large collections of predominantly unstructured documents in order to identify information relevant to a user's query or information need Manning et al. (2008).

One of the earliest IR models is the Boolean model, which uses Boolean logic operators to combine query terms and retrieve documents that satisfy specified logical conditions Manning et al. (2008). Another foundational approach is the vector space model, which represents documents and queries as vectors in a term space under the bag-of-words assumption Salton et al. (1975); Zhao et al. (2024). In recent years, research has increasingly focused on Neural Information Retrieval (Neural IR), which applies neural models to learn representations and matching functions for queries and documents. In dense neural retrieval, queries and documents are mapped to continuous vector representations, commonly referred to as embeddings Karpukhin et al. (2020). Unlike traditional sparse representations, these embeddings do not correspond directly to explicit vocabulary dimensions, but instead aim to capture latent semantic properties that support semantic matching Zhao et al. (2024). Dense neural retrieval can therefore improve

retrieval when queries and documents use different surface-level wording by capturing similarity beyond exact term overlap. In an open-domain question-answering setting, Karpukhin et al. (2020) show that a dual-encoder dense retriever trained on a small number of question–passage pairs outperforms a strong BM25 baseline. Lexical retrieval can also mitigate vocabulary mismatch through techniques such as query expansion Manning et al. (2008).

2.1.2 Hybrid Retrieval Methods

Sparse retrieval methods use words or terms as explicit indexing units. Classical sparse methods, such as BM25, rely on exact lexical matching between query and document representations, giving them strong discriminative power, particularly when exact terminology, named entities, or domain-specific expressions are important for retrieval Manning et al. (2008). More recent learned sparse retrieval methods, such as SPLADE, relax this constraint: a pretrained language model predicts weighted term expansions over the vocabulary, allowing a limited degree of semantic matching while retaining a sparse, inverted-index-compatible representation Formal et al. (2021). This distinction is relevant to the present work, since the sparse component of the hybrid retriever described in Section 4.2.3 is a learned encoder of this kind rather than a purely lexical one. Nevertheless, sparse methods of either kind remain comparatively limited in their ability to capture semantic similarity when relevant documents use substantially different wording than the query Guo et al. (2022).

In contrast, dense retrieval methods learn continuous vector representations of queries and documents that capture semantic relationships and enable soft matching beyond exact term overlap. This makes them more robust to lexical variation, but it can also come at the cost of fine-grained lexical precision, as exact wording and other low-level textual features may be less explicitly preserved Guo et al. (2022).

Hybrid retrieval is intended to combine the complementary strengths of sparse and dense retrieval, although its effectiveness depends on the task, corpus, and the chosen score combination method. These approaches generate both sparse and dense representations for queries and documents and then combine the resulting matching signals. In this way, hybrid retrieval aims to balance the lexical precision of sparse retrieval with the semantic generalization capability of dense retrieval Guo et al. (2022). The general architecture of such a hybrid retrieval system is illustrated in Figure 2.1.

A central design choice in hybrid retrieval is how the scores or rankings produced by the individual retrieval methods are combined. Score-based fusion strategies require the underlying scores to be calibrated on a comparable scale, which is often not the case for sparse and dense retrievers. Rank-based fusion methods avoid this problem by operating only on the relative ordering of results. A widely used example is Reciprocal Rank Fusion (RRF) Cormack et al. (2009), which assigns each candidate document d a fused score based on its rank in each retrieval method's result list:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$\mathrm{RRF}(d) = \sum_{r \in R_d} \frac{1}{k_{RRF} + \mathrm{rank}_r(d)} \tag{2.1}$$

where R_d denotes the set of retrieval methods whose ranked result lists contain document d, and k_{RRF} is a constant that dampens the effect of rank differences, particularly among highly ranked documents. Following Cormack et al. (2009), this constant is commonly set to 60. Documents that appear near the top of multiple ranked lists receive higher fused scores, while documents that are retrieved by only one method can still be retained if their rank in that list is sufficiently strong. RRF is attractive in practice because it does not require score normalization and typically requires little or no per-corpus tuning, making it a common default in hybrid retrieval systems.

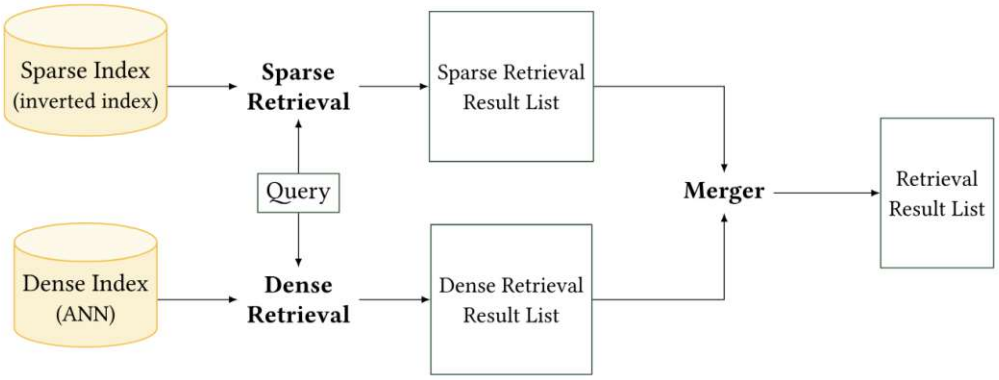

Figure 2.1: The general architecture of hybrid retrieval methods. Source:Guo et al. (2022).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

2.1.3 Retrieval in Financial Documents

Information retrieval in financial documents remains a challenging task. Financial reports are typically characterized by hundreds of pages of dense text, complex tables, figures, and highly domain-specific terminology. These properties make it difficult for traditional retrieval methods to consistently identify the most relevant information.

Recent research has proposed various approaches to address these challenges, though each closes only part of the gap. At the language level, general-purpose models often struggle with specialized financial terminology and reporting conventions; Araci (2019) addresses this with FinBERT, a domain-adapted variant of BERT, showing that even limited domain-specific fine-tuning can significantly improve performance on financial NLP tasks compared to general-purpose models. However, improved language understanding alone does not resolve the retrieval problem: when relevant evidence is distributed across long, heterogeneous documents, the ability to locate the correct passage remains a distinct and persistent challenge.

Other systems target a different axis of the problem: the temporal dynamics of financial information, rather than its specialized language. FinSearch, for instance, introduces an agent-based retrieval framework designed for environments where temporal relevance plays a critical role, such as live market data or evolving financial news Shen et al. (2025). While such approaches demonstrate the value of multi-stage retrieval and reasoning, they are tailored to dynamic corpora rather than the fixed, structured document collections characteristic of annual report analysis. The due diligence setting examined in this thesis therefore poses a different set of challenges, centred on precise evidence localization within long static documents rather than real-time information access.

Existing benchmark datasets reflect a similarly narrow task focus. FinQA Chen et al. (2021) focuses on numerical reasoning over financial reports, while TAT-QA Zhu et al. (2021) extends financial question answering to hybrid tabular and textual evidence. More recent work, such as FinAgentBench Choi et al. (2025), has expanded financial question answering toward retrieval-oriented and agentic settings. Nevertheless, none of these benchmarks directly addresses the compliance- and governance-oriented due diligence questions considered in this thesis. This gap, together with the limited availability of annotated data for financial due diligence, motivates the construction of the purpose-built evaluation dataset described in Chapter 3.

In the annual report setting, these limitations interact in a particularly inconvenient way. Due diligence questions are written in formal compliance language that rarely matches the phrasing of financial disclosures, which disadvantages sparse retrieval; yet annual reports also rely on precise regulatory terminology, rating agency names, capital ratio labels, and entity identifiers, for which exact lexical match matters. Dense encoders trained on general text may collapse the distinction between semantically similar but legally distinct terms in exactly these cases. Domain-adapted encoders such as FinBERT partially close the terminology gap, but neither dense nor sparse retrieval addresses the structural problem: relevant evidence is often a single sentence embedded in a table

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

or a disclosure block that spans several sections, and a retrieval method that ranks passages by overall similarity may return the correct page without surfacing the correct sentence. This combination of vocabulary mismatch and structural dispersion is what makes single-stage retrieval insufficient here, and what motivates the hybrid, multi-stage approach developed in Chapter 4.

To address these limitations, recent work has increasingly explored retrieval-augmented generation (RAG) pipelines for question answering over financial documents, combining retrieval mechanisms with the contextual reasoning capabilities of LLMs Dadopoulos et al. (2025). However, effectively utilizing long and complex contexts remains a significant challenge in practice. More critically for the present setting, standard RAG architectures do not explicitly enforce grounding at the span level, that is, they retrieve passages but do not guarantee that the final answer is drawn verbatim from a specific sentence within those passages.

2.2 Large Language Models (LLMs)

The development of Large Language Models (LLMs) has substantially expanded the capabilities of natural language processing systems, particularly for tasks that require integrating information across long and heterogeneous texts. The key architectural enabler was the Transformer Vaswani et al. (2017), which replaced the sequential processing of recurrent networks with a self-attention mechanism that computes relationships between all input tokens simultaneously. This made training on large corpora computationally tractable and allowed models to be scaled to billions of parameters. Subsequent increases in model size, training data volume, and context window length have since made LLMs applicable to document-intensive tasks such as question answering over financial reports Dadopoulos et al. (2025).

2.2.1 LLMs for Question Answering

Large Language Models can be used for question answering under different learning and inference settings. A common distinction is made between prompting-based approaches and task-specific fine-tuning.

In prompting-based settings, the model is guided through natural language instructions and, optionally, a small number of examples. In *zero-shot prompting*, the model receives only the task instruction or question without any examples. In *one-shot prompting*, a single example is provided, while *few-shot prompting* extends this to several demonstrations, from which the model must infer the task pattern without any parameter updates. A further prompting strategy is *chain-of-thought* prompting, which encourages the model to produce intermediate reasoning steps before arriving at a final answer. This has been shown to improve performance on tasks that require multi-step reasoning, as it helps the model decompose complex problems into smaller and more manageable steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Prompting-based approaches are particularly attractive because they reduce or eliminate the need for task-specific labeled data Brown et al. (2020); Wei et al. (2022).

In contrast, *fine-tuning* involves updating the parameters of a pre-trained model on a supervised dataset for a specific task. This approach has historically been widely used for transformer-based language models and encoders such as BERT, RoBERTa, and domain-adapted variants such as FinBERT. Fine-tuning often yields strong task-specific performance, but it requires labeled training data and may be less robust when transferred to different domains or tasks Brown et al. (2020). The different prompting settings are illustrated in Figure 2.2.

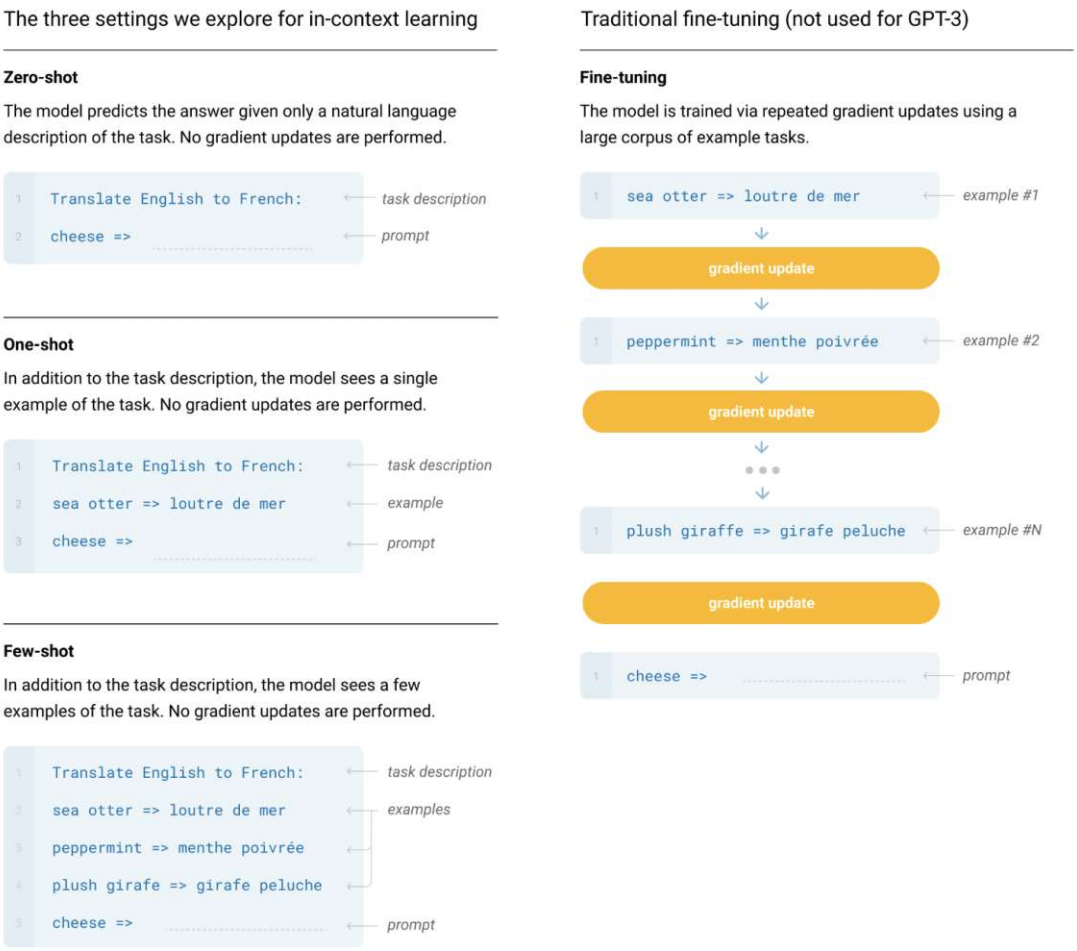

Figure 2.2: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. Source:Brown et al. (2020)

While these approaches have considerably expanded the applicability of LLMs to question answering, they do not eliminate fundamental weaknesses of LLM-based systems. This

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

is particularly relevant in high-stakes domains such as finance, where answers must be precise, source-grounded, and reliable.

2.2.2 General Limitations of Large Language Models

Despite their strong language understanding and generation capabilities, LLMs exhibit several limitations that affect their reliability in document-based question answering.

Lost in the Middle Language models do not necessarily use all parts of a long input context equally effectively. In particular, performance may decline when relevant information is located in the middle of the context, a phenomenon referred to as the *lost in the middle* problem Liu et al. (2024).

Bias LLMs may reproduce or amplify patterns and biases present in their training data Brown et al. (2020). Moreover, their internal decision-making processes are difficult to interpret, which can make it challenging to determine why a particular answer was produced Rudin (2019).

Hallucination A further limitation of generative language models is hallucination, that is, the production of content that is unsupported by or inconsistent with the available evidence. A distinction is commonly made between *intrinsic hallucination*, where generated content contradicts the source, and *extrinsic hallucination*, where generated information cannot be verified from the source Maynez et al. (2020). Empirical evaluations of ChatGPT have also demonstrated that extrinsic hallucinations occur in practice Bang et al. (2023).

2.2.3 Challenges in the Financial Domain

The limitations of LLMs discussed above become particularly consequential in financial applications: documents are characterized by specialized terminology, formal reporting conventions, and institution-specific language, while answers must satisfy strict requirements for factual accuracy and traceability.

Domain-Specific Language Language models pretrained on broad, general-purpose corpora may not adequately represent the terminology and linguistic conventions of specialized domains. Gururangan et al. (2020) show that continued pretraining on domain-specific corpora improves downstream performance across several domains. Similar evidence has been reported in finance: FinBERT, which adapts BERT to financial text, outperforms general-purpose baselines on financial sentiment-analysis tasks Araci (2019). These findings indicate that domain adaptation can improve performance when general-purpose representations do not sufficiently capture specialized financial language.

However, domain adaptation only improves a model's general command of financial language; it does not give the model access to the content of a specific institution's

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

disclosures. Annual reports are long, institution-specific, and not part of any model's pretraining data in their current form, so no degree of domain-adaptive pretraining substitutes for direct access to the source document itself. Grounding an answer in verifiable, institution-specific evidence therefore requires a complementary mechanism, providing the model with the relevant external documents at inference time. This motivates the use of retrieval-augmented approaches, introduced in the following section.

2.3 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) addresses several important limitations of standalone LLMs. In particular, it reduces reliance on knowledge encoded solely in the model weights and enables access to information that lies beyond the model's training cutoff. RAG is especially well suited for document question-answering settings, where answers must be derived from domain-specific or frequently changing sources that were not part of the original training data Lewis et al. (2020); Gao et al. (2023).

A key advantage of RAG is its modularity. RAG workflows can be adapted to different use cases by varying components such as the retrieval mechanism, chunking strategy, and generation model Gao et al. (2023). This flexibility makes RAG particularly suitable for knowledge-intensive tasks in domains such as finance, law, and medicine, where factual accuracy is especially important.

2.3.1 Standard RAG Architecture

A standard RAG pipeline can be decomposed into five sequential stages, from query ingestion through to response generation. Each stage is described below and illustrated in Figure 2.3.

- **User prompt:** The process begins with a user query, which defines the information need and serves as the basis for retrieval.
- **Query representation:** The query is transformed into a representation suited to the retrieval mechanism in use. This may take the form of a dense vector embedding for similarity-based (semantic) search, a sparse lexical representation for term-based matching, or both in hybrid retrieval settings.
- **Information retrieval:** Based on this representation, the system retrieves the most relevant document chunks from the indexed corpus. The retrieval mechanism is not fixed by the RAG paradigm itself and can range from sparse lexical matching to dense embedding-based similarity search, or a hybrid combination of both.
- **Context augmentation:** The top-k retrieved chunks are provided to the generation model as additional context, commonly by concatenating them with the query or prompt.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **Response generation:** The LLM then generates an answer conditioned on both the original query and the retrieved context, ideally producing a response that is more accurate, relevant, and grounded in the source documents Lewis et al. (2020); Gao et al. (2023).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Vector Database

RAG systems that rely on dense or hybrid retrieval require an efficient way to store and query embeddings, the numerical vector representations of text or other data objects. These embeddings make it possible to compare semantic similarity between a user query and indexed document chunks Gao et al. (2023). Sparse retrieval, by contrast, relies on the inverted index described in Section 2.1.2 rather than a vector-based structure. The concepts introduced here therefore apply specifically to the dense-retrieval branch of the hybrid architecture shown in Figure 2.1.

At the implementation level, dense retrieval may use either a standalone vector-search library or a vector database. A vector database combines vector indexing and similarity search with data-management functions such as persistence, metadata storage, filtering, and updates Wang et al. (2021); Databricks (2023).

In RAG systems that employ a vector database, document embeddings can be stored together with their corresponding text and metadata and retrieved based on their similarity to a query embedding.

Chunking Strategies in RAG

Before documents can be indexed for retrieval, they are typically divided into smaller text segments, commonly referred to as chunks. Depending on the retrieval architecture, these chunks may be represented using sparse lexical features, dense embeddings, or both. The choice of chunking strategy is an important design decision in RAG systems, as it can strongly influence retrieval quality. In particular, chunking affects whether relevant information is captured within a single passage, how much contextual information is preserved, and how precisely the retrieval component can identify useful evidence.

Several chunking strategies have been proposed in the literature:

- **Fixed-size chunking:** One of the most common approaches is to divide text into segments of a predefined size, often measured in characters or tokens. This method is usually controlled by two main hyperparameters: chunk size and chunk overlap. While fixed-size chunking is simple and efficient, it may split semantically coherent content across chunk boundaries and can be less suitable for highly structured documents Brådland et al. (2025); Gao et al. (2023).
- **Recursive character chunking:** This strategy recursively splits text according to a hierarchy of separators, such as paragraphs, sentences, or whitespace, until all resulting chunks satisfy a predefined maximum size. Compared with purely fixed-size chunking, this approach often preserves textual structure more effectively Brådland et al. (2025); Gao et al. (2023).
- **Semantic chunking:** Semantic chunking aims to create segments that are coherent in meaning rather than determined only by length or surface-level delimiters.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Typically, the process begins with smaller units such as sentences, which are then grouped into larger chunks based on semantic similarity Brådland et al. (2025). Related approaches include proposition-based retrieval units, which index self-contained atomic factual statements rather than fixed-size spans Chen et al. (2024), and late chunking, which embeds the full document before segmenting it so that chunk-level embeddings retain document-level context Günther et al. (2024). This can improve retrieval quality, although it is usually more computationally expensive than simpler chunking methods.

Each of these strategies involves trade-offs. Smaller chunks may improve retrieval precision by isolating relevant information more clearly, but they can also omit important surrounding context. Larger chunks preserve more context, but may reduce retrieval specificity by including more irrelevant information. Consequently, the most suitable chunking strategy depends on the characteristics of the documents and the requirements of the specific use case; for annual reports, whose tables, headings, and disclosure blocks carry structural meaning that plain length-based splitting would discard, this motivates the structure-aware chunking approach adopted in Section 4.3.1.

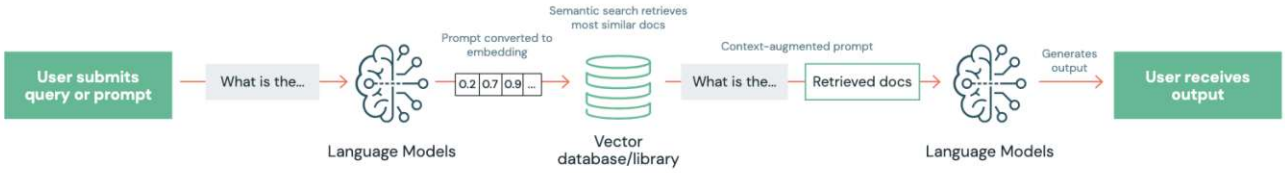

Figure 2.3: Overview of the retrieval-augmented generation (RAG) workflow. Source: Databricks (2023)

2.3.2 Limitations of Standard RAG

Despite its advantages, standard RAG still exhibits important limitations. One concern relates to *bias and fairness*. Survey literature identifies several ways in which biases in the source corpus, retrieval mechanism, or generation model may influence RAG outputs Gupta et al. (2024). Examining this issue empirically, Wu et al. (2025) evaluate fairness across multiple RAG configurations and find that fairness concerns persist across both retrieval and generation components, while improvements in answer utility do not necessarily translate into fairer outcomes.

Another important limitation concerns *retrieval quality and relevance*. The effectiveness of a RAG system depends heavily on whether the retrieval component returns passages that contain useful and sufficiently complete evidence. Retrieval failures may instead expose the generator to irrelevant or insufficient context, negatively affecting downstream performance Gupta et al. (2024). In an ablation study, Shi et al. (2024) show that ensembling randomly sampled documents worsens language-modeling performance, whereas ensembling relevant retrieved documents improves it. In their experimental setting, these

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

results show that the benefit of retrieval augmentation depends on the relevance of the supplied documents.

A further challenge concerns *evidence utilization and faithfulness*. Even when relevant information has been retrieved, the generation model may fail to use it faithfully or may supplement it with unsupported information. Based on a manually annotated corpus of RAG-generated responses, Niu et al. (2024) show that retrieval-augmented models can still produce claims that are unsupported by or contradict the retrieved context. Similarly, Kovács et al. (2025) observe that even when evidence sentences are inserted verbatim, generated answer templates may introduce subtle inaccuracies or misleading framing.

Retrieving a relevant passage is, however, not equivalent to identifying the specific evidence span that answers the question. In financial due diligence, where each response must be traceable to an exact source, this distinction is critical: a system may retrieve the correct page yet still produce a paraphrased, incomplete, or unsupported answer. This limitation motivates approaches that constrain the factual content of the response to evidence copied directly from the source. The following section introduces VerbatimRAG, the extractive framework that forms the methodological foundation of this thesis.

2.4 VerbatimRAG

2.4.1 Motivation and Design Principle

VerbatimRAG is an extractive retrieval-augmented question-answering framework designed for settings in which answers must remain directly traceable to their source documents. In contrast to many conventional generative RAG systems, in which an LLM synthesizes or paraphrases a final answer from retrieved context, VerbatimRAG follows an extractive approach that constrains the answer-bearing content to verbatim text spans selected from the retrieved documents Kovács et al. (2025). Its central design principle can therefore be summarized as *extract rather than generate*.

The distinction has direct practical consequences. Providing an LLM with relevant context does not guarantee that its generated response will reproduce the source faithfully. The model may omit qualifications, alter numerical values, introduce unsupported statements, or combine evidence in a misleading manner. VerbatimRAG addresses this generation-side risk by separating evidence identification from response formatting: the system first determines which parts of the retrieved text answer the question and subsequently constructs the response from those source passages (Kovács, 2025).

2.4.2 Original Verbatim Pipeline

The original VerbatimRAG approach was introduced by Kovács et al. (2025) for evidence-based question answering over electronic health records in the ArchEHR-QA 2025 shared task. The proposed pipeline consists of two principal stages.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

First, a sentence-level extractor identifies the sentences within the retrieved medical notes that are relevant to the question. The original system supports both zero-shot LLM-based extraction and supervised extraction using a fine-tuned Clinical ModernBERT classifier, a biomedical variant of ModernBERT adapted for long clinical narratives Kovács et al. (2025). This stage transforms a retrieved passage into a smaller collection of answer-bearing evidence sentences.

Second, an LLM constructs a question-specific answer template containing placeholders. These placeholders are subsequently filled with the extracted sentences without paraphrasing them. The factual evidence in the resulting response is therefore reproduced verbatim, while the template provides a question-appropriate response structure and source references Kovács et al. (2025).

This architecture separates two operations that are combined in conventional generative RAG: determining which evidence is relevant and expressing the final answer. The language model may assist with evidence selection and formatting, but the factual content presented to the user is constrained to extracted source text.

2.4.3 Framework Architecture and Implementation

The subsequent open-source implementation generalizes the original pipeline beyond its medical application and separates the system into two independently configurable layers Kovács (2025).

The first layer, referred to as the *VerbatimIndex* (Kovács, 2025), is responsible for document ingestion, parsing, chunking, indexing, and retrieval. The retrieval component is not restricted to one representation method: depending on the configuration, documents can be retrieved using sparse, dense, or hybrid representations. The indexing layer can also be replaced by an existing retrieval system, allowing the verbatim extraction component to be added to an established RAG pipeline.

The second layer, referred to as the *Verbatim Core* (Kovács, 2025), receives the user query and the retrieved chunks and identifies the source spans that answer the question. Span extraction may be performed using a prompted LLM or a fine-tuned extractive model (Kovács, 2025); Recski et al. (2026) additionally evaluate a non-generative semantic highlighting mechanism as a further extraction alternative. Extracted spans are then verified against the retrieved source text before they are used in the final response (Kovács, 2025). This verification step ensures that retained evidence is present in the source rather than newly generated by the extraction model.

The framework additionally provides different response-template strategies Kovács (2025). A static template uses a predetermined response format for all questions. A contextual template is generated dynamically according to the question and available evidence. Question-specific templates select among predefined formats based on the type of information requested. These alternatives represent a trade-off between natural response formatting and strict control: dynamic templates provide greater flexibility, whereas

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

static templates reduce the possibility that generated framing introduces unsupported meaning.

2.4.4 Recent Developments and Extraction Models

More recent work applies VerbatimRAG to scientific literature through the ACL-Verbatim system Recski et al. (2026). This study indexes more than 114,000 research papers from the ACL Anthology and directly maps user queries to evidence spans in retrieved sections of research papers. In addition to an end-to-end retrieval system, the work introduces a manually annotated benchmark for evaluating the extraction stage independently from retrieval.

Of the 100 query-chunk pairs in this benchmark, roughly half (47) were judged to contain answerable evidence, together yielding 78 gold spans; the remaining pairs were judged irrelevant to their paired query. The study compares prompted LLM extractors, semantic highlighting and context-pruning systems, and compact token-classification models trained using synthetically generated silver supervision Recski et al. (2026).

Among all evaluated systems, the best performance is reached by a 150-million-parameter ModernBERT token classifier fine-tuned on the silver data, obtaining a word-level F1 of 53.63; the best LLM-based extractor, GLM-5, trails at 48.71. The compact classifier provides higher precision and more frequently abstains on irrelevant chunks. In contrast, the LLM-based extractors generally achieve higher recall and evidence coverage, but also extract more false-positive spans from irrelevant passages Recski et al. (2026). These findings show that the extraction component should be treated as a separate modelling problem and that extractor selection involves a trade-off between evidence coverage, precision, computational cost, and abstention behaviour.

2.4.5 Limitations and Relevance to This Thesis

Verbatim extraction improves provenance but does not guarantee that an answer is correct in every relevant sense. A system may copy a span exactly while selecting evidence that is irrelevant, incomplete, outdated, or insufficient for answering the question. Verbatim verification establishes that the output occurs in the source text, but not that the selected text correctly satisfies the user's information need. Retrieval quality and extraction relevance therefore remain critical sources of error.

The response-template stage introduces an additional distinction. When a static template presents extracted spans directly, the scope for generated distortion is limited. When an LLM dynamically generates the surrounding template, however, the framing of otherwise verbatim evidence may still introduce subtle inaccuracies or unsupported implications Kovács et al. (2025). Furthermore, a strictly extractive approach is less suitable for questions requiring interpretation, aggregation, or synthesis across multiple sources, because the complete answer may not occur verbatim in any individual passage (Kovács, 2025).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

These limitations define the point of departure for the present thesis. This work applies the VerbatimRAG principle to financial due diligence question answering, where exact source attribution is required, but where evidence must be identified within long and structurally heterogeneous annual reports. The framework is extended with query rewriting, multi-query expansion, hybrid sparse-dense retrieval, cross-encoder reranking, and alternative span extractors. The experiments in Chapter 4 evaluate not only whether relevant document chunks are retrieved, but also whether the system identifies the correct verbatim spans and abstains when the available documents contain no supported answer.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Use Case

3.1 Overview

Building on the problem motivation introduced in Chapter 1, this chapter presents the concrete use case addressed in this thesis: answering standardized due diligence questionnaires (DDQs) issued by the Association for Financial Markets in Europe (AFME), using evidence drawn exclusively from the relevant institution's publicly available annual report. The chapter first positions this task within the broader landscape of due diligence practices, before describing the dataset in detail, including the underlying annual reports, the question set, and the reference answers. It then discusses the key challenges that arise in this setting.

3.2 Due Diligence as a Financial Question-Answering Task

Due diligence questionnaires (DDQs) are standardized information requests used across the financial industry to assess a counterparty's operational, legal, regulatory, and risk-related standing, typically during onboarding or periodic review. Depending on the issuing body and purpose, DDQ practices differ substantially: examples include mergers-and-acquisitions due diligence, credit risk assessment, AML/KYC counterparty screening, and vendor or regulatory due diligence, each relying on different document types, stakeholders, and workflows. This thesis does not address DDQs in general. It focuses specifically on the questionnaire issued by AFME, described in detail in Section 3.3, and treats it as a fixed set of natural-language questions that must be answered using evidence from a single source per institution: its publicly available annual report.

Answering such questionnaires is a time-consuming and labor-intensive process that requires extensive search for relevant information within large document collections. The questions vary in format, ranging from short factual information requests to more

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

open-ended questions that require the identification of specific disclosures in the source documents.

As outlined in the introduction, these tasks are often repetitive and operational in nature. However, they still require substantial domain knowledge as well as a broader understanding of the institution and its disclosures, particularly for questions whose interpretation depends on institution-specific disclosure practices.

3.3 Dataset Construction

The dataset used in this thesis consists of three main components. First, the source documents comprise annual reports of selected Austrian banks. Second, the question set combines twelve due diligence questions drawn from questionnaire material published by the Association for Financial Markets in Europe (AFME), an industry association representing wholesale financial market participants in Europe Association for Financial Markets in Europe (nd), with four questions authored for this thesis to enable the evaluation of abstention behavior. The selection procedure is described in Section 3.3.2. Third, the answer annotations define the reference evidence spans used for evaluation. These annotations did not exist prior to this work. Their creation constitutes one of the main contributions of this thesis.

3.3.1 Source Documents

The dataset is constructed from four publicly available annual reports of Austrian banks for the year 2024: Raiffeisen Bank International, Erste Group, BAWAG Group, and UniCredit Bank Austria. These documents were selected because they are directly relevant to the financial due diligence setting considered in this thesis, provide sufficiently detailed public disclosures, and exhibit a broadly comparable reporting structure that supports cross-institution evaluation.

As described in Section 3.3.2, the annotation procedure differs between Raiffeisen Bank International and the remaining banks. This difference may introduce some variability in annotation quality and is therefore considered when interpreting the evaluation results.

The dataset is not intended to be statistically representative of the entire banking sector. Rather, it is designed as a realistic and controlled evaluation corpus for assessing retrieval and extraction performance on long, complex, and heterogeneous financial reports. Since all source documents are publicly available, the experimental setup remains transparent and reproducible.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Dataset Statistics

The annual reports were processed and segmented into text chunks for indexing in the vector database. The exact parameters and methodology used for this segmentation (such as chunk size and overlap) are detailed in Section 4.3.1.

Table 3.1 summarizes the key characteristics of the resulting dataset.

Table 3.1: Summary statistics of the annual reports included in the dataset

Bank	Pages	Number of Chunks
Raiffeisen Bank International	632	2244
UniCredit Bank Austria	615	2199
BAWAG Group	363	1414
Erste Group	494	2011

3.3.2 Annotation Procedure

The annotations were created in two phases: a collaborative phase for Raiffeisen Bank International (RBI), carried out jointly by a due diligence practitioner at the institution and the author, and a subsequent single-annotator phase in which the author annotated the remaining three banks.

Question Selection The evaluation benchmark is based on the Association for Financial Markets in Europe (AFME) Due Diligence Questionnaire (DDQ). In practical due diligence, the full questionnaire covers information that may be distributed across different document types, including public disclosures as well as internal policies, audit reports, and other non-public documentation. Since this thesis specifically investigates automated question answering over publicly available annual reports, a due diligence practitioner at RBI selected twelve AFME questions for which the RBI annual report provides sufficient supporting evidence. These questions form the AFME-based core question set used throughout the evaluation. For RBI, all twelve are answerable from the annual report. For the other institutions, answerability depends on the disclosures contained in the respective report. The resulting task remains a realistic RAG use case because, even when the requested information is present, identifying the relevant evidence within annual reports spanning several hundred pages and containing heterogeneous textual and tabular structures constitutes a non-trivial retrieval problem.

To enable a controlled evaluation of model abstention, the dataset incorporates two types of unanswerable cases. First, four additional questions were authored for this thesis to form a common negative subset. These questions concern plausible information needs for which no supporting disclosure is present in any of the four evaluated annual reports, allowing abstention behaviour to be compared across institutions using the same questions. Second, naturally occurring unanswerability is retained: because disclosure practices

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

differ between institutions, some of the twelve AFME questions that are answerable for RBI cannot be answered from the annual reports of the other banks. These cases are likewise treated as unanswerable during evaluation.

Span Annotation for RBI For each of the twelve answerable questions, the gold evidence spans were identified in joint working sessions between the due diligence practitioner and the author. Both searched the annual report for passages answering the question, and each answer-bearing passage was recorded verbatim in a shared annotation file. A question may be answered by a single span or by multiple spans located in different parts of the report, for instance when the same disclosure appears in both a summary table and the narrative text. In such cases, all identified passages were recorded as gold spans. The four unanswerable questions were assigned an empty gold span set. Span relevance and boundaries were settled by discussion during the sessions, with the practitioner's judgment decisive in cases of doubt.

Derivation of Retrieval Labels The span annotations directly provide the extraction labels described in Section 4.3.2. The retrieval labels were derived from them after corpus indexing: each gold span was located in the chunked corpus, and the identifiers of the chunks containing the span text were recorded as the expected retrieval targets. Every annotated span was contained within a single indexed chunk. The retrieval labels are therefore a deterministic function of the span annotations and the chunking configuration: they were assigned by string matching against the stored chunk text and do not depend on the output of any retrieval configuration. Retrieval runs during development were used to assist in locating candidate chunks, but every label was verified directly against the stored chunk text.

Annotation for the Remaining Banks After completion of the RBI dataset, the annotations for BAWAG Group, Erste Group, and UniCredit Bank Austria were created by a single annotator, following the span format and annotation conventions established in the RBI sessions. Because all four institutions are subject to comparable EU and Austrian banking disclosure requirements, their annual reports address similar categories of information, such as credit ratings, capital adequacy, and governance disclosures. The RBI gold spans therefore served as a reference for what constitutes an answer-bearing disclosure in the remaining reports. For example, the question asking for the contracting entity is answered by institution-identifying passages such as "Raiffeisen Bank International AG" for RBI, "UniCredit Bank Austria AG" for UniCredit, and "Erste Group Bank AG" for Erste. Similarly, the Standard & Poor's rating question is answered by comparable rating disclosures across institutions, for example "Standard & Poor's A-" for RBI and "Standard & Poor's ... A-" for UniCredit. Although the exact wording and location of these disclosures differ between reports, the underlying information categories and the type of evidence required to answer the questions are comparable. Further examples are shown in Table 3.3. Questions for which no supporting disclosure could be found were marked as unanswerable for that institution. No second

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

annotator or expert review was available for these banks. This limitation is discussed in Section 5.5.2.

3.3.3 Question and Answer Set

The evaluation uses sixteen due diligence questions applied consistently to all four banks: twelve drawn from the AFME questionnaire and four authored for this thesis to enable the evaluation of abstention behavior. The questions vary in structure and difficulty, ranging from short factual requests, such as the name of the contracting party, to questions whose answers require longer or multi-part disclosures. Because the same question set is evaluated against different annual reports, answerability is institution-specific: the four authored questions are unanswerable for every bank, while the twelve AFME questions are answerable or not depending on what the individual report discloses. For RBI, all twelve AFME questions are answerable, whereas for the other institutions between nine and eleven are.

For each bank-question pair, reference answer spans were annotated according to the procedure described in Section 3.3.2. These spans serve as the *gold spans* for the evaluation, that is, the reference evidence that the system is expected to retrieve and extract from the source document. If no valid supporting evidence could be identified in the corresponding annual report, the gold-span set was left empty and the question was classified as unanswerable.

Table 3.2: Question and annotation statistics per institution. The same sixteen questions are posed to every bank. Answerability is determined by the individual annual report, except for the four authored questions, which are unanswerable by construction. Span lengths are measured in words after whitespace normalization.

	Questions		Gold spans		Span length (words)		
Bank	Answ.	Unansw.	Total	per answ. q.	Avg.	Med.	Max.
Raiffeisen Bank International	12	4	35	2.92	32.5	22.0	293
BAWAG Group	9	7	26	2.89	28.4	23.0	63
Erste Group	11	5	21	1.91	34.2	20.0	187
UniCredit Bank Austria	11	5	30	2.73	38.4	32.5	131

Table 3.2 summarizes the resulting question and annotation statistics. Although the same sixteen questions are posed to every bank, the number of answerable questions ranges from nine for BAWAG Group to twelve for RBI, largely reflecting differences in the information disclosed by the institutions. Annotation density also varies. The RBI, BAWAG, and UniCredit annotations average close to three gold spans per answerable question, often because relevant information appears in both summary tables and narrative disclosures, whereas the Erste annotations average fewer than two spans per answerable question.

The span-length distributions also differ across institutions. The BAWAG gold spans are comparatively short and uniform, with a maximum length of 63 words, whereas the RBI

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and Erste annotations each contain a small number of substantially longer spans, reaching 293 and 187 words respectively. These differences are considered when interpreting the cross-bank results in Section 5.4, since annotation density and span length may influence retrieval coverage and strict span-matching scores. The different numbers of unanswerable questions should likewise be considered when comparing abstention accuracy, particularly because the estimates are based on relatively small denominators.

To illustrate the questions and the corresponding annotation format, two representative AFME questions are presented below, followed by selected gold spans for each institution in Table 3.3. The AFME questionnaire used in this study is in English, and although the selected banks publish annual reports in both German and English, only the English-language versions were used for dataset construction and evaluation. Consequently, all questions and source documents were processed in English, and no machine translation was required.

Example Questions

1. Please advise the name of the contracting party providing custody and/or client money services in your jurisdiction and responding to this questionnaire. If applicable, please also advise the name of the local delegate if different from the contracting entity.

2. Please state the credit ratings of the contracting party for long-term credit rating (Standard & Poor's).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 3.3: Selected gold answer spans for two AFME due diligence questions

Bank	**Q.**	**Illustrative Gold Span(s)**
Raiffeisen Bank International	Q1	"Raiffeisen Bank International AG"; "RBI AG is the central institution of the Raiffeisen Banking Group Austria (RBG)."
	Q2	"Standard & Poor's A-"; "S&P long-term rating A-"
UniCredit Bank Austria	Q1	"UniCredit Bank Austria AG, Rothschildplatz 1, 1020 Vienna, Austria ... is a universal bank ..."
	Q2	"Standard & Poor's ... A-"
BAWAG Group	Q1	"BAWAG Group AG is the parent company of BAWAG Group ..."
	Q2	No gold span annotated, as no answer is available in the document.
Erste Group	Q1	"Erste Group Bank AG ..."
	Q2	"Standard & Poor's ... confirming the A+/A-1 rating"

The abbreviated spans in Table 3.3 provide a compact cross-institutional overview but abstract away the document layout. In financial reporting, isolated values or short spans may require structural context for their correct interpretation, such as table headers or disclosure metadata. To make this relationship explicit, two RBI examples are shown below in their original report context.

Q1 - Raiffeisen Bank International (Annual Report 2024, p. 430)

Source context: Consolidated financial statements, ESEF information.

Name of reporting entity	**Raiffeisen Bank International AG**
Legal form of entity	AG
Address of entity's registered office	Am Stadtpark 9, 1030 Vienna
Country of incorporation	Austria

The surrounding fields explicitly identify Raiffeisen Bank International AG as the reporting entity and provide additional information about its legal form and registered location. This context supports the entity name selected as answer-bearing evidence for Q1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Q2 - Raiffeisen Bank International (Annual Report 2024, p. 23)

Source context: Section "RBI rating".

Rating	Moody's Investors Service	Standard & Poor's
Long-term rating	A1	**A-**

The row identifies the rating as long-term, while the column identifies the rating agency as Standard & Poor's. Their intersection therefore establishes **A-** as the value requested in Q2, illustrating why the surrounding table structure is required to interpret the short gold span.

The examples also illustrate two properties of the dataset. First, a single question may be associated with multiple valid evidence spans when the same information is disclosed in different parts of an annual report. Second, answerability is institution-specific: for example, the Standard & Poor's rating question is answerable for RBI, UniCredit, and Erste, whereas no corresponding evidence was identified in the BAWAG annual report.

3.4 Challenges of Automated Due Diligence Question Answering

Building on the general limitations of Large Language Models and Retrieval-Augmented Generation discussed in Chapter 2, the financial due diligence setting introduces several additional task-specific challenges. These challenges arise from the structure of financial documents, the nature of the questions, and the strict requirements for accuracy and traceability.

Complex Document Structure Financial annual reports are long, heterogeneous, and non-linearly structured documents that combine narrative disclosures, complex tabular layouts, and document-internal cross-references. These document characteristics present distinct retrieval challenges in the evaluation dataset:

- **Tabular coordinate dependencies:** An extracted value may derive its meaning from its surrounding row and column headers. For example, in the RBI report, the long-term credit rating is represented by the isolated value "A-", which is ambiguous without the intersection of the row header (*Long-term rating*) and column header (*Standard & Poor's*).
- **Document-internal cross-references:** Information may be linked through pointers to other sections. For instance, the section describing RBI's ownership structure directs the reader to Note (29), *"Equity and non-controlling interests"*, for further disclosures on the ownership structure.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **Dispersed evidence across distant sections:** Relevant evidence for a single due diligence item may occur in different parts of the same report. For the question on regulatory stress testing, valid evidence occurs in separate sections discussing climate stress testing, normative stress scenarios, and liquidity stress testing.

Such structural heterogeneity makes it difficult for conventional single-step retrieval to reliably localize relevant evidence. This motivates the use of structure-aware document representations as well as multi-stage retrieval and reranking to improve evidence localization across both tabular and dispersed textual content.

Domain-Specific Language and Terminology Financial documents and due diligence questions employ specialized terminology, reporting conventions, and institution-specific phrasing. The same underlying concept may therefore be expressed differently across institutions or even within the same report. For example, the AFME questionnaire asks for the "contracting party", whereas the RBI annual report identifies the relevant institution using the field "Name of reporting entity". Credit-rating information also appears in different forms: RBI reports the Standard & Poor's rating in a structured table, while Erste states the corresponding rating in narrative text. Such variation limits the effectiveness of purely lexical retrieval and motivates query rewriting and query expansion to bridge the gap between questionnaire wording and document terminology.

Handling Unanswerable Questions Some questions cannot be answered from the available annual report and must therefore be treated as unanswerable. For example, the Standard & Poor's long-term rating question is answerable for RBI, UniCredit, and Erste, whereas no corresponding evidence was identified in the BAWAG annual report. In such cases, forcing the system to return an answer may result in unsupported content. The system must therefore be able to recognize when no valid evidence is available and abstain from answering. This is particularly important in financial due diligence, where an unsupported compliance statement or credit rating could misrepresent the institution's risk profile.

The Limits of Full Automation While retrieval-side techniques can substantially improve evidence localization, the outputs of automated systems cannot be treated as final due diligence judgments without expert review. Financial due diligence is governed by strict requirements concerning regulatory compliance, auditability, and legal liability. Because LLM-based systems are probabilistic and cannot assume legal accountability, their outputs must remain verifiable and subject to human validation in high-stakes environments. Even with advanced RAG architectures, the cost of a hallucinated or misinterpreted answer can be high.

These challenges directly correspond to the core evaluation dimensions of this thesis: retrieval performance, extraction accuracy, and the ability to abstain from unsupported answers. Addressing them requires a system design that prioritizes precise evidence identification and strict grounding in source documents.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Accordingly, the approach proposed in this thesis is designed strictly as an assistive "human-in-the-loop" system. By emphasizing verifiable, evidence-based answering rather than autonomous decision-making, the system aims to accelerate the manual review process while ensuring that final accountability remains with the human expert.

3.5 Chapter Summary

This chapter detailed the concrete use case of financial due diligence that forms the empirical basis of this thesis. It introduced the dataset, comprising the annual reports of four Austrian banks, a fixed set of sixteen due diligence questions, and the corresponding gold-span annotations, created with expert involvement for RBI and by a single annotator for the remaining institutions. It then discussed the task-specific challenges of this setting, ranging from complex document structure and domain-specific terminology to the handling of unanswerable questions. These challenges underline the limits of full automation and motivate the assistive, human-in-the-loop design adopted in this thesis.

Building on this foundation, the following chapter will introduce the methodology and system architecture designed to address these challenges. Specifically, it will detail the techniques employed by the proposed framework to navigate complex financial texts and extract verbatim evidence spans, ultimately aiming to accelerate and support the manual due diligence process.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Methodology

4.1 Overview

This chapter describes the methodology developed to investigate how a grounded, verbatim answering approach can be applied to due diligence questionnaire (DDQ) answering in a banking context. To address the research questions outlined in Chapter 1, the proposed pipeline is evaluated across three dimensions: retrieval quality, verbatim span extraction quality, and abstention behavior for unanswerable questions. The chapter is organized as follows. Section 4.2 introduces the overall system architecture and its components. Section 4.3 describes the corpus, indexing procedure, and gold standard annotation process. Section 4.4 presents the experimental design, including ablation studies for each pipeline stage and a cross-bank generalization study. Finally, Section 4.5 defines the evaluation metrics used throughout the experiments.

Implementation Provenance The `LocalMilvusStore` and `VerbatimIndex` abstractions are inherited from the upstream VerbatimRAG project and were not implemented as part of this thesis. The thesis-specific extensions include the local index-connection configuration, bank-to-report scope resolution, query rewriting, multi-query expansion, candidate deduplication, the integration of BGE-based cross-encoder reranking, the financial-domain benchmark datasets, evaluation metric implementations, and the experimental analysis pipeline. These components are contained primarily in the `custom/` directory of the accompanying open-source repository Selenge (2026). The repository is distributed under the MIT License, and all thesis-specific code described in this work is released as part of this repository.

4.2 System Architecture

The proposed system builds upon the VerbatimRAG framework introduced by KR Labs, discussed in Section 2.4, which was originally developed for evidence-grounded question answering in the medical domain. Given its emphasis on verbatim span extraction and answer traceability, VerbatimRAG serves as a natural foundation for the DDQ answering task, where responses must remain directly grounded in source documents. This section describes how the original framework is adapted and extended into a baseline configuration suitable for the financial domain, followed by a detailed discussion of each pipeline component: query processing, hybrid retrieval, reranking, and verbatim extraction.

4.2.1 Baseline Configuration

The baseline pipeline consists of the hybrid Milvus index described in Section 4.3.1, the raw user question, and the unmodified VerbatimRAG extractor. It serves as the control condition for every experiment in Section 4.4, so that the contribution of each additional component (query rewriting, multi-query expansion, and cross-encoder reranking) can be measured as an isolated delta on top of the same retrieval and extraction back-end.

Concretely, given a DDQ question q targeting bank b, the baseline performs the following steps:

1. **Query processing.** The question q is forwarded unchanged to the retriever.

2. **Retrieval.** Hybrid dense and sparse retrieval is run over the Milvus index. The dense and sparse rankings are fused using weighted Reciprocal Rank Fusion (RRF). The RRF smoothing constant is set to $k_{\text{RRF}} = 60$, corresponding to the Milvus default. This value was retained without separate optimization. The fused ranking is truncated to the top $k = 5$ chunks, which are passed to the subsequent pipeline stages. This retrieval depth was fixed throughout the experiments and was not tuned per configuration.

   Because the index contains chunks from all four banks, the search is restricted at query time to the bank that the question concerns using a Milvus metadata filter derived from the bank profile. In the implemented system, this filter is resolved from source-identifying metadata such as the indexed PDF filename. Without this scope restriction, a question about one bank could be answered using chunks from another institution.

3. **Reranking.** None is applied in the baseline configuration.

4. **Verbatim extraction.** The retrieved chunks are passed to the VerbatimRAG extractor, which selects up to $N_{\text{spans}} = 5$ grounded spans. The extractor uses Gemini 3 Flash-Preview (via OpenRouter) by default. The same model is used in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

every configuration so that retrieval-side effects can be isolated, and the comparison of alternative extraction LLMs is reported separately in Section 4.4.3.

A summary of the baseline settings is given in Table 4.1.

Table 4.1: Baseline pipeline configuration. Components marked "fixed" are held constant across the retrieval-side ablation studies in Section 4.4, while components marked "varied" are the ones manipulated by an ablation experiment.

Stage	Setting	Status
Query processing	identity (raw q)	varied
Retrieval mode	hybrid dense + sparse	fixed
Fusion	weighted RRF, $k_{\text{RRF}} = 60$	fixed
Top-k chunks	$k = 5$	fixed
Reranker	none (baseline)	varied
Extractor	VerbatimRAG	fixed
Extractor LLM	Gemini 3 Flash-Preview (OpenRouter)	fixed*
Max extracted spans	$N_{\text{spans}} = 5$	fixed

* The extractor LLM is fixed for the retrieval-side ablation studies and varied only in the extraction model comparison (Section 4.4.3).

4.2.2 Query Processing

Query processing is a critical stage in the proposed pipeline because DDQ questions are typically written in a formal, compliance-oriented style that may differ substantially from the terminology used in annual reports. For example, one AFME question asks for the "contracting party providing custody and/or client money services", whereas the corresponding RBI disclosure identifies Raiffeisen Bank International AG under the field "Name of reporting entity". Although both expressions refer to the institution whose identity is required, they share little lexical overlap.

When such questions are passed directly to the retriever, this mismatch between question wording and document terminology may reduce the retrieval score of otherwise relevant chunks and push them out of the top-k candidate set. To reduce this vocabulary mismatch, the system applies two sequential query-processing steps: query rewriting and multi-query expansion, each described below.

Query Rewriting

The original question q is first rewritten into a simpler, retrieval-oriented formulation $\tilde{q}$. The goal is to preserve the semantic intent of the question while stripping procedural or compliance-oriented phrasing that is unlikely to appear verbatim in annual report text. For example, verbose requests such as *"Please provide documentation or evidence to confirm whether..."* are transformed into concise factual questions that more closely resemble statements found in financial disclosures.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The rewriter is also responsible for resolving entity references: if the question refers to the contracting party, organisation, or uses a bank-specific ticker or acronym, e.g. `RBI`, these are expanded to the full institution name. This reduces ambiguity during retrieval and ensures that bank-specific terminology is handled consistently across questions.

Rewriting is performed by an LLM prompted with the system instruction shown in Figure 4.1. The prompt is bank-aware: the bank name and short name are injected at runtime, allowing the same rewriter to be applied across all four banks in the corpus without modification. Although this introduces an additional LLM call and therefore additional latency and computational cost, the rewriting step is not expected to be executed repeatedly for every retrieval request in a live system. Since the DDQ questions are standardized, rewritten queries for a given question and institution can be generated once and cached for subsequent reuse.

```
You are an expert in Due Diligence / finance questions. You rewrite
Due Diligence / finance questions into simple questions suitable for
searching the {bank_name} Annual Report.

Rules:
- Output ONLY questions.
- No bullet points, no numbering, no prefixes, no commentary.
- Translate abbreviations into their full form.
- If the question refers to the contracting party or organisation,
  treat it as referring to {bank_name}.
- When the text uses a short ticker or acronym (e.g. {bank_short}),
  treat it as referring to {bank_name}.
- Grasp the meaning of the question and rewrite it simply.

Example:
  Input:   Please provide documentation or evidence to confirm
           whether your service infrastructure undergoes independent
           third-party audits, such as SOC2 Type II or ISO 27001,
           on an annual basis."
  Output:  Does your service undergo independent third-party
           security audits every year?
```

Figure 4.1: System prompt used for query rewriting. Bank name and abbreviation are injected at runtime.

Multi-Query Expansion

The rewritten query $\tilde{q}$ is subsequently provided to an LLM that is prompted to generate at most three alternative search queries. This prompt-level limit was selected to broaden retrieval while controlling the number of additional retrieval calls and reranking candidates. It was not separately optimized.

The generated variants are intended to capture complementary lexical and semantic formulations of the same information need, including synonyms, domain-specific abbre-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

viations, and bank-specific terminology. This is particularly important in the financial domain, where the same concept may be expressed differently across annual reports. For example, a question referencing *S&P* may need to match passages containing the agency's full name, *Standard & Poor's*. Similarly, a general question about systemic importance may need to match regulatory terminology such as *O-SII buffer*, *systemic risk buffer*, or *capital buffer for systemically important institutions*.

The motivation for multi-query expansion is that a single query formulation may capture only one lexical and semantic representation of the underlying information need. For sparse retrieval, passages using different terminology may receive weaker sparse-retrieval scores, although SPLADE partially mitigates this problem through learned term expansion. For dense retrieval, a single query embedding may not align equally well with all semantically relevant formulations in the document. Submitting several alternative formulations therefore diversifies both the lexical and semantic retrieval signals and increases the likelihood that at least one query matches the terminology used in the relevant report passage.

Query generation is performed using the prompt shown in Figure 4.2. The model is instructed to produce short, keyword-oriented queries in a structured JSON object. This format facilitates automatic parsing and avoids unnecessary conversational content in the generated output.

```
You generate search queries to retrieve relevant chunks from a bank
annual report for: {bank_name} ({bank_short}).

Return JSON only with this schema:
{"queries": ["q1", "q2", "q3"]}

Rules:
- Make 3 queries max.
- Do NOT start queries with: Confirm / Please / Advise.
- Use short keyword-like phrases that would appear in annual reports.
- Include 1 exact-name query with the bank name or short name
  when helpful.
- Include variants with synonyms or acronyms (e.g. O-SII, G-SIB,
  systemic risk buffer, Moody's credit rating).

Question: {question}
```

Figure 4.2: Prompt used for multi-query expansion. The model is instructed to return a structured JSON object containing up to three search queries.

Representative Expansion Examples Table 4.2 illustrates two cases in which alternative query formulations can bridge the terminology used in an AFME question and the terminology appearing in an annual report. The variants shown are illustrative formulations consistent with the generation prompt rather than logged outputs from a particular experimental run.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.2: Representative multi-query expansions and corresponding annual-report terminology

Information need	**Individual generated query strings**	**Relevant report terminology and benefit**
RBI long-term S&P rating	q_1: `RBI Standard & Poor's long-term rating` q_2: `Raiffeisen Bank International S&P credit rating` q_3: `RBI rating agency long-term rating`	The report refers to *Standard & Poor's* and contains a rating table with a *Long-term rating* row. Query q_1 expands *S&P* to *Standard & Poor's*, thereby increasing lexical correspondence with the source terminology.
RBI systemic importance	q_1: `RBI systemically important institution` q_2: `RBI O-SII buffer` q_3: `Raiffeisen systemic risk buffer`	The relevant disclosure refers to a *capital buffer for systemically important banks (O-SII buffer)* and a *systemic risk buffer*. Queries q_2 and q_3 introduce regulatory terminology that does not occur explicitly in the original AFME question.

The generated response is parsed as JSON, after which exact duplicate query strings are removed. If generation fails, produces an empty response, or yields output that cannot be parsed, the rewritten query $\tilde{q}$ is used as a fallback. Failure of the expansion stage therefore does not prevent retrieval from being performed.

Each resulting query variant is submitted independently to the hybrid retrieval index. In the experimental configuration, up to 20 chunks are retrieved per query before reranking. This first-stage retrieval depth was fixed to provide the reranker with a broader candidate pool than the final top-5 context while keeping the reranking volume bounded. It was not separately optimized. When the model follows the three-query instruction, the initial candidate pool may contain up to 60 chunks before duplicate chunks are removed. The deduplicated candidates are subsequently scored by the cross-encoder reranker, which retains the five highest-ranked chunks. Multi-query expansion thus prioritizes recall during first-stage retrieval, while deduplication and reranking control the additional noise and computational cost introduced by the broader search.

4.2.3 Hybrid Retrieval

After query processing, each generated query variant is submitted to a hybrid retrieval component. The purpose of this stage is to identify candidate annual report chunks that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

are relevant either because they are semantically similar to the query or because they share important financial terminology with it. This is important for DDQ questions because some answers require conceptual matching, whereas others depend on exact terms such as regulatory acronyms, credit rating agency names, risk categories, or institution-specific labels. The retrieval index is implemented using Milvus and stores both dense and sparse representations for each document chunk. Dense embeddings are produced using a sentence-transformer model and capture semantic similarity between the query and the report text. In parallel, sparse representations are produced using a SPLADE-style learned sparse encoder, which assigns weights to lexical terms and expanded term features. The sparse component therefore helps retrieve chunks containing exact or near-exact financial terminology, while the dense component can still retrieve relevant chunks when the wording differs from the query.

For example, an illustrative expanded query such as *RBI Standard & Poor's long-term rating* shares several exact terms with the relevant RBI passage, which states that the bank is rated by *Standard & Poor's* and contains a table row labelled *Long-term rating* with the value *A-*. This provides a strong sparse-retrieval signal because the bank name, rating agency, and rating horizon occur explicitly in both the query and the passage.

By contrast, the question *Are you a branch of an entity incorporated in another country?* can be answered by the Erste Group passage *Erste Group Bank AG is a stock corporation established under Austrian law.* Although the question and passage have limited surface-level overlap, they express the same underlying legal-status relationship: the passage indicates that Erste Group Bank AG is an Austrian legal entity rather than a foreign branch. Dense retrieval can therefore help identify this passage through semantic similarity even when the exact wording of the question is absent.

For a processed query q_i, the system computes a dense query vector $\mathbf{d}_{q_i}$ and a sparse query vector $\mathbf{s}_{q_i}$. Milvus is then queried separately over the dense and sparse vector fields. The dense search returns chunks ranked by vector similarity, while the sparse search returns chunks ranked by lexical-neural term overlap. These two ranked lists are merged using weighted reciprocal rank fusion (RRF). For a candidate chunk c, the hybrid score is computed as:

$$\text{score}(c) = \sum_{r \in R} w_r \cdot \frac{1}{k_{\text{RRF}} + \text{rank}_r(c)}$$

where R is the set of retrieval methods, here dense and sparse retrieval, w_r is the weight assigned to retrieval method r, and $\text{rank}_r(c)$ is the rank of chunk c in that method's result list. Chunks that appear highly ranked in both dense and sparse retrieval receive a higher fused score, while chunks found by only one method can still be retained if they are strongly ranked by that method.

In the implemented system, dense and sparse retrieval are weighted equally. This choice reflects the complementary role of the two signals: dense retrieval addresses semantic

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

mismatch between DDQ questions and annual report prose, whereas sparse retrieval preserves sensitivity to exact financial and regulatory terms. The fused list is truncated to the top candidate chunks for each query variant.

Because the query expansion stage may generate up to three alternative queries, the hybrid retrieval process is repeated for each query variant. The resulting candidate lists are then merged and deduplicated using chunk metadata, including the source file and chunk index. This produces a broader candidate pool while avoiding repeated presentation of the same annual report passage. In the implementation, each query variant retrieves up to 20 candidates before the merged candidate pool is passed to the reranking stage.

4.2.4 Reranker

Hybrid retrieval is designed to surface a broad candidate set: by combining semantic and lexical signals across multiple query variants, it prioritizes coverage over precision. However, the resulting scores reflect retrieval-level similarity rather than a direct assessment of whether a passage actually answers the question. A chunk may rank highly because it shares domain terminology with the query, or because it discusses the same general topic, without containing the specific evidence needed. To address this, a second-stage cross-encoder reranker is applied to re-score the merged candidate pool before final selection.

The reranker is a cross-encoder model that receives the rewritten query $\tilde{q}$ and each candidate chunk as a paired input. Unlike bi-encoder retrieval, where query and document representations are computed independently, the cross-encoder jointly encodes the query–chunk pair and produces a direct relevance score. This allows the model to capture fine-grained relationships, such as whether a chunk actually answers the question rather than merely containing related terminology. Formally, for each candidate chunk c_j, the reranker computes:

$$r_j = f_\theta(\tilde{q}, c_j)$$

where $\tilde{q}$ is the rewritten query and f_θ is the cross-encoder reranking model. Candidate chunks are sorted by their reranking score in descending order, and only the top five are passed to the verbatim extraction stage. The two stages play different roles: the hybrid retriever casts a wide net over the annual report to gather as many potentially relevant chunks as possible, while the reranker carefully selects the passages most likely to contain a direct answer to the question.

4.2.5 Verbatim Extraction and Answer Generation

After reranking, the system uses the top-ranked chunks to construct a grounded answer. Rather than generating an answer directly from the retrieved text, the pipeline first extracts the exact passages that are relevant to the original DDQ question. This step is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

necessary because the final response should not only be informative, but also traceable to specific wording in the annual report.

The extraction stage receives the rewritten retrieval query and the top five reranked chunks as input. In the experiments reported in Table 4.1, the span extractor uses Gemini 3 Flash-Preview. The extractor identifies sentences or paragraphs that explicitly answer the question and is constrained to copy these passages verbatim, preserving the original wording, punctuation, and capitalization from the annual report. Across the candidate set, up to $N_{\text{spans}} = 5$ verbatim spans are retained as the evidence used in the final answer. If a retrieved chunk does not contain relevant information, no span is extracted from that chunk.

To ensure that the extracted text is genuinely verbatim, the extracted spans are verified against the source chunk text. Only spans that can be matched back to the retrieved annual report text are retained. The implementation also allows for whitespace differences, which is useful because text extracted from PDF annual reports often contains irregular line breaks, tabs, or spacing artifacts. This verification step reduces the risk that the LLM contributes unsupported or paraphrased output instead of evidence copied from the source document.

The verified spans are then passed to the answer generation component. Instead of allowing the LLM to freely summarize the full retrieved chunks, the system uses the extracted spans as the factual basis for the response. A template manager constructs a question-appropriate response structure with placeholders for the extracted facts. In the implemented contextual mode, this template is generated dynamically from the question and the available spans, and uses placeholders such as `[FACT_1]`, `[FACT_2]`, and `[CITATION_REFS]`. These placeholders are then filled with the verified verbatim spans and citation references. This keeps the generated answer concise while preserving the original source wording.

Finally, the response builder links the extracted spans back to the retrieved documents by creating highlights and citations. Each retained span is associated with its source chunk, allowing the final answer to be inspected together with the supporting annual report passage. This design is particularly important in a due diligence setting, where users need to verify not only the answer itself but also the documentary evidence behind it.

4.3 Data and Experimental Setup

Section 4.3.1 explains how the annual report PDFs are parsed, chunked, and stored in the vector index. Section 4.3.2 describes the two forms of gold annotation used to evaluate retrieval and extraction.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.3.1 Corpus and Indexing

Before indexing, the annual report PDFs are parsed into a structured document representation. This step is important because annual reports are not plain continuous text documents: they contain hierarchical section headings, tables, captions, footnotes, lists, and disclosure blocks. If such documents are chunked only by a fixed token limit or by a simple sliding window with overlap, important context may be split across chunk boundaries. For example, a table may be separated from its surrounding disclosure text, or a paragraph may lose the section heading that explains whether it refers to capital ratios, credit ratings, risk management, or governance.

For this reason, the corpus-indexing pipeline implemented for this thesis uses Docling as the underlying document-processing library for structure-aware parsing and chunking. Each PDF is first processed using Docling's document converter with table-structure recognition enabled. The resulting document representation is then segmented using Docling's existing `HybridChunker`. For chunk-length measurement, the chunker uses the Hugging Face tokenizer associated with `sentence-transformers/all-MiniLM-L6-v2`, configured with a maximum chunk length of 512 tokens. The document-conversion and chunking algorithms are provided by Docling, while their configuration and integration into the financial-document indexing pipeline form part of the thesis-specific implementation released in the accompanying repository.

The same model checkpoint, `sentence-transformers/all-MiniLM-L6-v2`, is used separately as the dense encoder. Unlike the tokenizer used for chunk-length measurement, the complete SentenceTransformer model transforms each chunk into a 384-dimensional dense vector. It was selected as a fixed configuration because it provides a practical trade-off between retrieval quality and computational efficiency and can run locally without specialised hardware. The embedding model was not separately optimised. Instead, the evaluation focuses on the multi-stage retrieval and extraction architecture.

This choice is supported by the hybrid system design: terminology-sensitive matching is strengthened by the SPLADE sparse retriever, while final candidate ordering is performed by a cross-encoder reranker. The dense encoder therefore does not carry the complete retrieval burden alone.

Instead of splitting the report at arbitrary character or token offsets, the chunker operates over Docling's document representation and is therefore designed to better preserve document elements such as headings, lists, and tables. Peer merging is enabled, so neighbouring elements that belong to the same local document context can be combined into a more coherent chunk rather than being indexed as many small fragments. This is useful for annual reports, where a short heading, caption, or list item may only become meaningful when kept together with nearby text.

Tables are handled explicitly during this process. The index creation pipeline uses a Markdown table serializer so that tables are stored in a readable textual form. This is particularly important for annual reports, where relevant evidence often appears in tabular disclosures, for example in rating summaries, capital figures, or risk-related

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

breakdowns. Serializing tables as Markdown preserves row and column structure more clearly than plain text extraction, which supports both retrieval and later span verification against the stored chunk text.

For each generated chunk, the indexing script stores the chunk text together with metadata extracted from the Docling representation. This metadata includes the source PDF filename, page number where available, chunk index, detected headings, chunk type, and an approximate token count estimated from the chunk text. The chunk type distinguishes between text, table, list, figure, and formula-like content when such information is available. In the present experiments, this field is not used as a hard retrieval filter or reranking feature. Instead, it is stored for traceability, diagnostics, and analysis of the indexed corpus.

The resulting chunks are stored in a local Milvus vector index. For each chunk, both dense and sparse representations are computed and stored. Dense embeddings are produced with the model described above, while sparse representations are produced with `opensearch-neural-sparse-encoding-doc-v2-distill`. Dense embeddings support semantic similarity search, whereas sparse representations support lexical and terminology-sensitive matching. At query time, metadata filters restrict retrieval to the annual report of the bank relevant to the question, preventing evidence from one institution's report from being retrieved for a question about another.

4.3.2 Gold Standard Annotation

The evaluation uses two complementary forms of gold annotation: retrieval labels and extraction labels. These annotations correspond to different stages of the pipeline and are therefore used for different metrics. The same set of 16 questions is applied to each of the four banks, however, whether a question is answerable depends on the individual annual report, so the unanswerable subset differs per institution.

Retrieval labels. For retrieval evaluation, each answerable question is associated with one or more expected chunk identifiers. These identifiers specify which indexed chunks contain the evidence required to answer the question. They are used to evaluate whether the retrieval and reranking stages successfully surface the relevant evidence in the top-ranked candidate set, and form the basis for Hit Rate, Recall@K, and Mean Reciprocal Rank (MRR), as defined in Section 4.5.

Extraction labels. For extraction evaluation, each question is associated with one or more gold answer spans representing the exact evidence text that the system is expected to extract from the annual report. If no valid evidence exists in the report, the question is treated as unanswerable by leaving the gold span set empty. This makes it possible to evaluate both successful evidence extraction for answerable questions and abstention behavior for questions where the annual report does not contain a supported answer.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

As described in Section 3.3, the Raiffeisen Bank International annotations were created jointly with a domain expert as part of this work, while the annotations for BAWAG Group, Erste Group, and UniCredit Bank Austria were created manually without expert involvement.

4.4 Experimental Design

The experiments are organized as controlled ablations over the DDQ answering pipeline. Each experiment keeps the indexed corpus, chunking strategy, embedding models, hybrid retrieval setup, and evaluation annotations fixed, while varying one component of the pipeline at a time. The first set of experiments is conducted on the Raiffeisen Bank International (RBI) annotations, which serve as the primary development and ablation set. The selected configuration is then evaluated across the remaining banks to assess cross-bank generalization.

4.4.1 Retrieval Ablation Study

The retrieval ablation study examines the incremental contribution of each retrieval-side component by adding components to a fixed retrieval backbone in a controlled sequence. All five configurations share the same indexed corpus, chunking strategy, embedding models, and gold annotations, ensuring that any observed difference in performance reflects the contribution of the varied component alone. This experiment is conducted exclusively on the expert-annotated RBI dataset, which serves as the primary development set. The remaining banks rely on manually created annotations and are reserved for the cross-bank generalization study in Section 4.4.4. All configurations are evaluated using Hit Rate, Recall@K, and Mean Reciprocal Rank (MRR), as defined in Section 4.5.1.

The five configurations evaluated in this ablation are summarized in Table 4.3.

Table 4.3: Retrieval ablation configurations. Each condition isolates one retrieval-side variation or combination, while keeping all other pipeline stages fixed.

Configuration	**Query Processing**	**Reranker**
Baseline	Raw DDQ question	None
Baseline + Reranker	Raw DDQ question	Cross-encoder
Baseline + Rewriting + Reranker	Rewritten query	Cross-encoder
Baseline + Multi-Query + Reranker	Expanded queries	Cross-encoder
Full Pipeline	Rewritten + expanded queries	Cross-encoder

The ordering of conditions reflects the architectural dependency between pipeline components. The baseline establishes a lower bound using raw DDQ questions with hybrid dense and sparse retrieval and no reranking. Cross-encoder reranking is introduced first, as it operates directly on the candidate set returned by the retriever without modifying the query, making it the most straightforward component to evaluate in isolation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Query rewriting is then added to assess whether reformulating the DDQ question into a retrieval-oriented phrasing improves the quality of the initial candidate set. Multi-query expansion is evaluated next to determine whether generating lexically varied query variants further broadens coverage of the indexed corpus. Finally, the full pipeline combines all components, allowing us to assess whether the joint application of query processing and reranking produces effects beyond what each component contributes individually.

4.4.2 Reranker Comparison

The retrieval ablation study in Section 4.4.1 establishes whether reranking improves retrieval quality, but does not determine which reranking model is most suitable for the DDQ setting. A separate reranker comparison is therefore conducted, in which all pipeline components are held fixed at the full pipeline configuration, and only the cross-encoder model used in the reranking stage is varied. The following four models are evaluated:

- `cross-encoder/ms-marco-MiniLM-L-6-v2`
- `cross-encoder/ms-marco-MiniLM-L-12-v2`
- `BAAI/bge-reranker-base`
- `BAAI/bge-reranker-v2-m3`

These models were selected to cover both widely used MS MARCO cross-encoders and BGE rerankers, while also comparing smaller and larger variants under the same candidate generation setting. The reranking stage receives the merged and deduplicated candidate pool produced by hybrid retrieval, scores each query and chunk pair, and passes the top five chunks to evaluation using the same metrics and annotations as Section 4.4.1.

4.4.3 Extraction Model Comparison

After selecting the retrieval-side configuration and reranker, the next experiment evaluates the span extraction stage. While the previous experiments measure whether the correct evidence chunks are retrieved, this experiment measures whether the system can identify the exact answer-bearing spans within those chunks.

All retrieval-side components are held fixed using the best-performing configuration from the retrieval ablation and reranker comparison. For each question, the system retrieves and reranks the top five chunks, after which different extraction methods are applied to the same candidate evidence. This ensures that differences in performance are attributable to the extraction method rather than to retrieval quality.

Multiple extraction methods are compared. The first group consists of VerbatimRAG LLM-based span extractors, where the extractor is prompted to copy only evidence spans

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

that directly answer the DDQ question while preserving the original wording of the annual report.

The following models are evaluated in this setting:

- `google/gemini-3-flash-preview`
- `google/gemini-2.5-flash-lite`
- `openai/gpt-5.1`
- `openai/gpt-4.1-mini`
- `moonshotai/kimi-k2-0905`
- `meta-llama/llama-4-scout`

These models provide a comparison across different model families and capability levels under the same retrieval and reranking conditions. In addition, a semantic span extraction approach using `zilliz/semantic-highlight-bilingual-v1` is included as a non-generative baseline. This model identifies semantically relevant passages from the retrieved chunks as a more constrained span-selection mechanism, making it possible to assess whether LLM-based extraction provides a measurable advantage for DDQ evidence selection.

Because DDQ answering requires traceability, all methods are evaluated on their ability to return evidence that matches the gold-annotated spans rather than on free-form answer quality, using the extraction and abstention metrics defined in Section 4.5.2.

4.4.4 Cross-Bank Generalization

The retrieval ablation and reranker comparison are conducted exclusively on the RBI dataset, which provides expert-annotated gold labels and serves as the primary development set. To assess whether the selected configuration generalizes beyond this dataset, the best-performing pipeline is subsequently evaluated on the three remaining banks: BAWAG Group, Erste Group, and UniCredit Bank Austria.

All pipeline components are held fixed at the configuration selected from the RBI experiments. No further tuning is performed on the remaining banks, ensuring that the evaluation reflects out-of-development-set generalization rather than bank-specific optimization. The same set of 16 questions is applied to each institution, with retrieval restricted to the corresponding annual report via the Milvus metadata filter described in Section 4.3.1. Results are reported using the same retrieval and extraction metrics as the preceding experiments.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.5 Evaluation Metrics

4.5.1 Retrieval Metrics

Retrieval quality is evaluated by comparing the ranked list of retrieved chunks against the gold chunk identifiers associated with each question. For a question q, let G_q denote the set of gold chunks that contain evidence needed to answer the question, and let $R_q^K = (r_1, r_2, \ldots, r_K)$ denote the top-K chunks returned by the retrieval or reranking stage. Questions without a valid gold evidence chunk are excluded from retrieval metric computation and are evaluated separately through the abstention metrics. In all experiments, $K = 5$, matching the number of chunks passed to the span extraction stage.

Hit Rate. Hit Rate measures whether at least one relevant evidence chunk appears in the top-K retrieved results. For a single question, the hit indicator is defined as:

$$\text{Hit@K}(q) = \begin{cases} 1, & \text{if } R_q^K \cap G_q \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

The final Hit Rate is the mean of this indicator over all evaluated questions:

$$\text{HitRate@K} = \frac{1}{|Q|} \sum_{q \in Q} \text{Hit@K}(q).$$

Recall@K. Recall@K measures the proportion of all gold evidence chunks that are retrieved in the top-K list. This is important because some DDQ questions require evidence from more than one chunk. The reported Recall@K is averaged across all evaluated questions:

$$\text{Recall@K} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|R_q^K \cap G_q|}{|G_q|}.$$

Mean Reciprocal Rank. Mean Reciprocal Rank at K evaluates how early the first relevant evidence chunk appears in the ranked list. Let rank_q be the rank position of the first retrieved chunk in R_q^K that also belongs to G_q. If no gold chunk is retrieved in the top-K list, the reciprocal rank is set to zero:

$$\text{RR@K}(q) = \begin{cases} \frac{1}{\text{rank}_q}, & \text{if } R_q^K \cap G_q \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

MRR@K is then computed as the mean reciprocal rank over all evaluated questions:

$$\text{MRR@K} = \frac{1}{|Q|} \sum_{q \in Q} \text{RR@K}(q).$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Together, these three metrics capture complementary aspects of retrieval quality. Hit Rate measures whether any supporting evidence is found, Recall@K measures coverage of all annotated evidence chunks, and MRR@K measures how highly the first relevant chunk is ranked.

4.5.2 Extraction Metrics

Extraction quality is evaluated by comparing the spans produced by the system against the gold answer spans annotated for each question. For a question q, let $S_q = \{g_1, \ldots, g_m\}$ denote the set of gold spans and let $P_q = \{p_1, \ldots, p_n\}$ denote the set of predicted spans extracted from the retrieved chunks. The goal is not to evaluate free-form answer generation, but to measure whether the system identifies the same evidence passages as the gold annotation.

Because extracted spans may differ slightly from the gold spans due to whitespace, punctuation, or boundary differences, extraction is evaluated with multiple complementary similarity measures.

RapidFuzz Span Matching. Fuzzy lexical similarity between every predicted span p_i and gold span g_j is computed using `fuzz.WRatio` from the `rapidfuzz` package. Before scoring, spans are lowercased, Markdown formatting is removed, tables are flattened, punctuation is discarded, and whitespace is normalized. The pairwise score is

$$s_{ij}^{\mathrm{RF}} = \frac{\mathrm{WRatio}\left(N(p_i), N(g_j)\right)}{100},$$

where $N(\cdot)$ denotes text normalization and $s_{ij}^{\mathrm{RF}} \in [0, 1]$.

`WRatio` is a composite lexical similarity measure based on normalized string-edit similarity and token-based variants of this comparison. It considers the basic sequence similarity between the two strings as well as token-based comparisons that are less sensitive to differences in token order or the presence of additional tokens. When the input strings differ substantially in length, partial variants are additionally considered in order to identify the best matching region of the longer string. Rather than averaging these component scores, `WRatio` returns the highest appropriately weighted score. Token-based comparisons are slightly downweighted, and partial comparisons receive an additional scaling penalty.

This scoring behavior makes the resulting span evaluation comparatively tolerant of over-extraction. A prediction that contains the complete gold evidence together with additional surrounding text may still receive a high `WRatio` score. At the same time, lexical overlap can also remain high when a short answer-bearing value is missing from an otherwise similar span. Consequently, a high `WRatio` score does not by itself guarantee that the predicted span is concise or contains every materially relevant value. These limitations are examined qualitatively in Section 5.3.1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

RapidFuzz scores are evaluated at thresholds $t \in \{0.50, 0.75, 0.90, 0.95\}$, corresponding to increasingly strict matching requirements. For each threshold, scores below t are set to zero, whereas scores meeting the threshold retain their original continuous value:

$$s_{ij}^{(t)} = \begin{cases} s_{ij}^{\mathrm{RF}}, & s_{ij}^{\mathrm{RF}} \geq t, \\ 0, & s_{ij}^{\mathrm{RF}} < t. \end{cases}$$

Thus, RF@75, for example, denotes span evaluation with a minimum pairwise similarity of 0.75. It is not a retrieval-rank cutoff.

ROUGE-L. ROUGE-L Lin (2004) is computed using `RougeScorer` from the `rouge-score` package. It measures the longest common subsequence between a predicted and gold span. Unlike a substring, the tokens in a subsequence do not need to be contiguous, but their relative order must be preserved. The pairwise ROUGE-L F1 value is used as the similarity score for one-to-one span matching.

BERTScore. BERTScore Zhang et al. (2020) is computed using `BERTScorer(lang="en")` from the `bert-score` package. It compares contextual token embeddings and therefore captures semantic correspondence beyond exact surface-form overlap. Pairwise BERTScore precision and recall are first combined into pairwise F1, and this F1 value forms the similarity matrix used for one-to-one matching.

One-to-One Matching Procedure. When multiple predicted and gold spans are available, every predicted span is first compared with every gold span to construct a pairwise similarity matrix. For RapidFuzz, this is the threshold-gated matrix defined above; for ROUGE-L and BERTScore, it contains the corresponding continuous pairwise F1 values.

The Hungarian algorithm, implemented using `linear_sum_assignment` from `scipy.optimize`, selects the one-to-one assignment that maximizes total similarity. Consequently, each predicted span can match at most one gold span, and each gold span can be matched at most once. This prevents several similar predictions from receiving credit for the same annotated span.

Let M denote the sum of the optimally assigned similarities, n the number of predicted spans, and m the number of gold spans. Soft set-level precision and recall are calculated as

$$P = \frac{M}{n}, \qquad R = \frac{M}{m},$$

and F1 is their harmonic mean:

$$F1 = \frac{2PR}{P+R}.$$

These are graded, similarity-weighted measures rather than conventional binary true-positive counts. Unmatched predicted spans increase the precision denominator and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

therefore penalize over-generation, while unmatched gold spans increase the recall denominator and penalize missing evidence. If both span sets are empty, the question receives a precision, recall, and F1 score of 1; if only one set is empty, all three scores are 0. Final results are obtained by macro-averaging the per-question precision, recall, and F1 scores.

Together, RapidFuzz, ROUGE-L, and BERTScore capture complementary aspects of span quality: fuzzy lexical correspondence, ordered token overlap, and contextual semantic similarity. The shared one-to-one assignment framework ensures that all three measures consistently penalize duplicate predictions, additional predicted spans, and missing gold evidence. This set-level penalty is distinct from over-extraction within a single predicted span, to which `WRatio` can remain comparatively tolerant.

4.5.3 Abstention and Faithfulness Metrics

Questions with no gold evidence span are treated as unanswerable. For these questions, the correct behavior is to return no extracted span. Abstention accuracy is computed as the proportion of unanswerable questions for which the system produces an empty span set:

$$\text{AbstentionAccuracy} = \frac{\#\text{correct abstentions}}{\#\text{unanswerable questions}}.$$

Faithfulness is enforced procedurally through verbatim span verification: extracted spans are retained only if they can be matched back to the retrieved source chunk. However, verbatim verification alone does not determine whether a verified span actually answers the question. The quantitative evaluation of unsupported-answer avoidance is therefore captured by abstention accuracy, which measures whether the extractor returns no span for questions where no valid gold evidence exists in the annual report.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Results and Discussion

This chapter presents and discusses the empirical results of the proposed DDQ answering pipeline, following the structure of the pipeline itself. First, the retrieval ablation study examines how query rewriting, multi-query expansion, and reranking affect the ability to retrieve annotated evidence chunks. Second, different reranking models are compared under the selected retrieval configuration. Third, the span extraction stage is evaluated by comparing LLM-based extractors with a semantic highlighting baseline, followed by a qualitative analysis of the extracted spans. Finally, the selected configuration is applied to the remaining banks to assess cross-bank generalization. Faithfulness (RQ3) is not evaluated as a standalone experiment. The corresponding abstention-accuracy results are reported alongside the cross-bank experiments and drawn together in the cross-experiment synthesis that closes the chapter, together with a discussion of limitations and future work.

5.1 Retrieval Ablation Results

The retrieval ablation study evaluates how each retrieval-side component affects the system's ability to surface the correct evidence chunks. The baseline in this experiment is not a conventional generative RAG system. Rather, it follows the VerbatimRAG principle of evidence-grounded answering: the system retrieves annual report chunks and subsequently extracts verbatim evidence spans from them. However, in the baseline configuration, the DDQ question is passed directly to the hybrid retriever without query rewriting, multi-query expansion, or reranking. It therefore represents a minimal VerbatimRAG-style configuration against which the additional retrieval components can be measured.

As shown in Table 5.1, the baseline achieves a Hit Rate of 0.583, Recall@K of 0.368, and MRR of 0.542. This indicates that the hybrid dense–sparse retriever is able to retrieve relevant evidence for some questions, but its performance remains limited when formal

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

DDQ questions are used unchanged as search queries. The relatively low Hit Rate shows that in a substantial number of cases no annotated evidence chunk appears in the top-five results. Similarly, the Recall@K score suggests that the baseline often retrieves only part of the required evidence when multiple gold chunks are available. This confirms the need for additional retrieval-side improvements before verbatim span extraction is applied.

Adding a cross-encoder reranker improves Hit Rate from 0.583 to 0.667 and MRR from 0.542 to 0.625, showing that reranking can help place relevant chunks higher when they are present in the candidate set. However, Recall@K decreases from 0.368 to 0.292. Since the final evaluation is computed over the top five chunks after reranking, this suggests that the reranker moved some gold chunks below the top-five cutoff. Reranking alone therefore improves ranking precision in some cases, but does not necessarily improve evidence coverage.

The strongest improvements are observed when query processing is introduced. Query rewriting combined with reranking increases Hit Rate to 0.833 and Recall@K to 0.478, demonstrating that simplifying DDQ questions into retrieval-oriented formulations improves the initial candidate set. Multi-query expansion with reranking achieves the same Hit Rate of 0.833 but a higher Recall@K of 0.542 and MRR of 0.708, indicating that lexical query variation is particularly useful for retrieving a broader set of relevant chunks.

The full pipeline, combining query rewriting, multi-query expansion, and reranking, achieves the best overall retrieval performance, with a Hit Rate of 1.000, Recall@K of 0.596, and MRR of 0.704. This means that at least one gold evidence chunk is retrieved for every evaluated RBI question. Although the MRR is slightly lower than in the multi-query-only configuration with reranking, the full pipeline provides the strongest coverage and reliability. It is therefore selected as the retrieval configuration for the subsequent reranker, extraction, and cross-bank experiments.

Table 5.1: Retrieval ablation results on the expert-annotated RBI dataset. Retrieval scores are reported as percentages.

Strategy	**Hit Rate (%)**	**Recall@K (%)**	**MRR (%)**	**LLM Calls**
Baseline	58.3	36.8	54.2	0
Baseline + Reranker	66.7	29.2	62.5	0
Rewriting + Reranker	83.3	47.8	56.9	1/query
Multi-Query + Reranker	83.3	54.2	**70.8**	1/query
Full Pipeline	**100.0**	**59.6**	70.4	2/query

LLM call counts refer to query-processing calls only and exclude the span extraction step, which is held fixed across all configurations.

5.2 Reranker Comparison Results

The retrieval ablation study uses `BAAI/bge-reranker-v2-m3` as the provisional default reranker. The separate reranker comparison reported in this section validates

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

whether this default choice is appropriate under the full pipeline configuration. In this experiment, query rewriting, multi-query expansion, hybrid retrieval, and candidate deduplication are held fixed, so that differences in performance can be attributed to the reranker alone.

Table 5.2 shows that `BAAI/bge-reranker-v2-m3` performs best across all three retrieval metrics. It achieves a Hit Rate of 1.000, Recall@K of 0.603, and MRR of 0.718. This means that the model retrieves at least one annotated evidence chunk for every evaluated RBI question and also provides the strongest overall evidence coverage.

The two MS MARCO cross-encoders achieve the same Hit Rate of 0.833, but lower Recall@K and MRR. This indicates that they often retrieve at least one relevant evidence chunk, but they retrieve fewer of the annotated gold chunks overall and tend to rank the first relevant chunk lower. `BAAI/bge-reranker-base` performs worst among the evaluated models, with a Hit Rate of 0.667 and MRR of 0.514.

These results confirm `BAAI/bge-reranker-v2-m3` as the most suitable reranker for this setting and validate its use as the default throughout the ablation study.

Table 5.2: Reranker comparison results on the expert-annotated RBI dataset.

Reranker	**Hit Rate (%)**	**Recall@K (%)**	**MRR (%)**
`cross-encoder/ms-marco-MiniLM-L-6-v2`	83.3	45.7	60.4
`cross-encoder/ms-marco-MiniLM-L-12-v2`	83.3	45.3	58.3
`BAAI/bge-reranker-v2-m3`	**100.0**	**60.3**	**71.8**
`BAAI/bge-reranker-base`	66.7	40.8	51.4

5.3 Extraction Model Comparison Results

After fixing the retrieval configuration and reranker, the extraction model comparison evaluates how well different span extractors identify the annotated evidence spans within the retrieved chunks. Unlike the previous experiments, which focus on whether the correct chunks are retrieved, this experiment measures the quality of the extracted evidence itself. The LLM-based extractors are evaluated using the Verbatim-RAG span extraction prompt, while the semantic highlighting baseline uses `zilliz/semantic-highlight-bilingual-v1` in sentence-level mode.

Table 5.3 reports the extraction results. For the LLM-based extractors, the values are averaged over five runs. The strongest overall extraction performance is achieved by `google/gemini-3-flash-preview` and `openai/gpt-5.1`. Both models obtain the highest RapidFuzz@75 F1 score of 0.690 and similar RapidFuzz@50 F1 scores of 0.722 and 0.700 respectively, indicating strong near-match performance. Under the stricter RapidFuzz@95 criterion, `google/gemini-3-flash-preview` performs slightly better (0.457 vs. 0.419). In terms of abstention, `openai/gpt-5.1` achieves both the highest BERTScore F1 of 0.709 and a perfect abstention accuracy of 1.000

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

across all five runs, making it the most reliable model for unanswerable questions. `google/gemini-3-flash-preview` also achieves strong abstention accuracy of 0.900.

`openai/gpt-4.1-mini` performs competitively on span matching, achieving the highest RapidFuzz@50 F1 among the remaining models, but its scores fall below the two best-performing models at stricter thresholds, and its abstention accuracy of 0.500 indicates that it frequently fails to abstain on unanswerable questions. `moonshotai/kimi-k2-0905` shows a similar abstention accuracy of 0.900 but weaker span matching overall. The remaining LLMs show weaker extraction performance, particularly under stricter fuzzy matching thresholds, and `google/gemini-2.5-flash-lite` performs poorly on abstention with an accuracy of only 0.200.

The semantic highlighting baseline performs below all LLM-based extractors on the main span matching metrics. Its RapidFuzz@75 F1 score of 0.399 and RapidFuzz@90 F1 score of 0.157 are substantially lower than the best LLM-based extractors. Notably, however, its RapidFuzz@50 F1 of 0.456 is closer to the weaker LLMs, suggesting that the semantic highlighter can identify broadly relevant regions but struggles to align its span boundaries with the gold annotations at stricter thresholds. Its abstention accuracy of 0.750 is comparatively strong among the evaluated models, though it does not fully abstain in all unanswerable cases.

Table 5.3: Extraction model comparison results. LLM results are averaged over five runs.

Extractor	**RapidFuzz F1 (%)**				**ROUGE-L F1 (%)**	**BERTScore F1 (%)**	**Unans. Acc. (%)**
	@50	**@75**	**@90**	**@95**			
`Gemini 3 Flash`	**72.2**	**69.0**	**53.9**	**45.7**	**54.1**	70.5	90.0
`GPT-5.1`	70.0	**69.0**	53.6	41.9	53.1	**70.9**	**100.0**
`GPT-4.1-mini`	63.4	61.0	43.4	34.7	44.9	64.6	50.0
`Kimi K2`	54.8	54.8	44.7	41.7	45.5	55.7	90.0
`Gemini 2.5 Flash Lite`	47.6	48.5	28.9	23.6	30.5	52.2	20.0
`Llama 4 Scout`	45.4	44.2	31.6	28.2	32.8	46.5	45.0
`Semantic Highlight`	45.6	39.9	15.7	11.5	20.2	47.6	75.0

5.3.1 Qualitative Analysis of Extracted Spans

The aggregate scores reported in Table 5.3, however, do not on their own indicate whether a predicted span would be useful in a due diligence review. This section therefore examines predicted–gold span pairs produced by `google/gemini-3-flash-preview` for several RBI questions under the full pipeline configuration. The examples illustrate different types of extraction results but do not indicate how frequently each type occurred.

An extraction is considered practically useful when it contains the requested fact, does not substitute a different disclosure item or conflicting value, and can be verified directly from the extracted text. For consistency with the quantitative evaluation, `WRatio` remains the primary similarity measure. `token_set_ratio` is used only as a supplementary

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

diagnostic because it ignores token order and may assign a maximal score when one span's token set is contained in the other. The RF@K values in Table 5.3 and Table 5.4 remain based on `WRatio`.

Exact and complete matches. A score of 100 generally corresponds to directly usable evidence. For the question on whether RBI is a branch of a foreign-incorporated company, the extractor reproduces the gold spans *"Country of incorporation | Austria"* and *"Domicile of entity | Austria"* exactly. Exact extraction also occurs for longer narrative evidence, such as the ownership disclosure beginning *"RBI has no majority shareholder. The eight regional Raiffeisen banks are core shareholders ..."*. These examples show that exact extraction is achievable for both tabular and narrative content.

Correct but over-inclusive matches. For the Standard & Poor's rating question, the extractor returns the complete ratings table, whereas the gold span contains only *"Long-term rating A-"*. The pair receives a `WRatio` score of 90.0. Although the prediction contains substantial additional information, the requested rating is present and identifiable, so the span remains useful. Similarly, for the registered-address question, the narrative extraction *"Its address is Am Stadtpark 9, 1030 Vienna."* differs from the tabular gold span but preserves the complete requested address. These cases show that imperfect span boundaries do not necessarily indicate an incorrect answer. Whether over-inclusive extraction is disadvantageous depends on the specific case: additional surrounding information may be unnecessary, but it can also provide useful context for interpreting or verifying the requested fact.

Topically related but non-answering spans. In the inspected examples, lower scores correspond to passages that discuss the correct topic but omit the requested fact. For example, the extraction

> *"Standard & Poor's last reaffirmed its rating in May 2024, citing the continued good business performance and large capital buffers that could absorb risks from the Russian business."*

receives a `WRatio` score of 58.1 against *"Standard & Poor's A-"*. The passage concerns the correct rating agency but does not state the rating grade. Comparable errors include returning *"Legal form of entity | AG"* for a question asking for the entity name. These spans are relevant to the same disclosure section but are not usable answers.

High overlap with a missing material value. The supplementary `token_set_ratio` analysis reveals an additional failure mode. A prediction may reproduce the descriptive part of a gold span while omitting the short value on which the answer depends. For example, the Standard & Poor's commentary above receives a `token_set_ratio` score of 93.3 against *"Standard & Poor's A-"*, despite omitting the rating grade. Likewise,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

"RBI continues to be rated by Moody's Investors Service and Standard & Poor's" receives a token-set score of 94.1 against a gold span containing the Moody's A1 rating, although the value A1 is absent. The same behaviour appears under the reported measure. The Moody's span above receives a `WRatio` score of 86.7 against the same gold annotation and therefore passes RF@50 and RF@75 in Table 5.3, despite omitting the value A1. This failure mode is consequently not an artifact of the supplementary diagnostic, but a property of the reported scores: lexical overlap alone cannot determine whether a short, financially material value is present.

Implications for threshold interpretation. Within the inspected `WRatio` examples, RF@50 and RF@75 appear to be lenient indicators that may include broadly relevant but non-answering spans. The useful over-inclusive ratings-table extraction receives a score of 90.0, whereas the passage that discusses Standard & Poor's without stating the requested rating receives only 58.1. Among the evaluated thresholds, RF@90 therefore appears to provide the most defensible indicator of answer-bearing overlap in the inspected sample, while RF@95 provides a stricter proxy for near-exact agreement with the annotated span.

These interpretations remain empirical rather than universal. No lexical threshold alone guarantees practical correctness, particularly for questions whose answer depends on a single rating, figure, date, country, percentage, or other short identifier. The automatic scores should therefore be interpreted together with qualitative inspection of whether the requested value itself is present and correct.

5.4 Cross-Bank Generalization Results

The final experiment evaluates whether the selected pipeline generalizes beyond the RBI development setting. The configuration selected from the previous experiments is kept fixed: query rewriting, multi-query expansion, hybrid dense–sparse retrieval, the selected BGE reranker, and the `google/gemini-3-flash-preview` span extractor. No bank-specific tuning is performed. Retrieval is restricted to the corresponding annual report using the Milvus metadata filter.

Table 5.4 reports the results for RBI, BAWAG Group, Erste Group, and UniCredit Bank Austria. The pipeline generalizes well on Hit Rate, reaching 1.000 for RBI, BAWAG, and UniCredit, and 0.909 for Erste. This means that at least one annotated evidence chunk is retrieved for nearly every evaluated question across all institutions. Recall@K shows greater variation: UniCredit achieves the highest value of 0.932, followed by Erste (0.909), BAWAG (0.778), and RBI (0.579). The lower Recall@K for RBI and BAWAG indicates that despite high hit rates, some gold evidence chunks fall outside the top-five candidate set, particularly when multiple chunks are required to answer a single question. These results suggest that the retrieval architecture transfers well across institutions, though evidence coverage remains sensitive to document structure and the number of required gold chunks.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The extraction results show greater variation across banks. BAWAG achieves the strongest extraction performance, with RF@75-F1 of 0.781, RF@95-F1 of 0.571, ROUGE-L F1 of 0.658, and BERTScore F1 of 0.808. This suggests that the retrieved evidence chunks for BAWAG contain spans that align closely with the gold annotations and can be extracted reliably. RBI performs second best on extraction, with RF@75-F1 of 0.634 and BERTScore F1 of 0.665, outperforming UniCredit, which achieves RF@75-F1 of 0.588 and BERTScore F1 of 0.626. Erste has the weakest extraction results overall, with RF@75-F1 of 0.480 and BERTScore F1 of 0.481. Notably, its RF@50-F1 and RF@75-F1 scores are identical at 0.480, indicating that even lenient fuzzy matching does not improve coverage, which points to a systematic mismatch between extracted spans and gold annotations rather than a span boundary alignment issue.

The difference between retrieval and extraction performance is important. Although the system retrieves at least one relevant chunk for almost every question, extracting the exact annotated span remains more sensitive to document wording, annotation style, and span boundary decisions. This is especially visible for Erste, where retrieval performance is high but extraction metrics are substantially lower. The cross-bank experiment therefore suggests that the retrieval architecture is robust across institutions, while verbatim span extraction remains the more bank-sensitive stage of the pipeline.

Part of this variation reflects measurable properties of the four datasets rather than the system alone (Table 3.2). RBI carries the densest annotation, with 2.92 gold spans per answerable question, which makes full evidence coverage within the top-five candidate set more difficult and is consistent with its lower Recall@K despite a perfect Hit Rate. BAWAG has the shortest and most uniform gold spans (median 23 words, maximum 63), which may contribute to its stronger strict-threshold scores, whereas RBI and UniCredit contain much longer spans, reaching 293 and 131 words respectively. Erste combines strong retrieval with the weakest extraction scores and the sparsest annotation, with 21 gold spans across eleven answerable questions, so annotation density should be considered when interpreting its results. More generally, the BAWAG, Erste, and UniCredit annotations were produced by a single annotator without an independent completeness check. Although all identified valid passages were recorded, exhaustive coverage of every possible gold span cannot be guaranteed. A valid prediction could therefore in principle correspond to an unannotated passage and receive a low similarity score. This possibility was not systematically quantified and should be considered when interpreting the cross-bank extraction metrics.

Abstention performance varies considerably across institutions. BAWAG achieves the highest unanswerable accuracy of 0.857 (6 of 7 unanswerable questions), followed by RBI at 0.750 (3 of 4) and UniCredit at 0.600 (3 of 5). Erste obtains the lowest score of 0.400 (2 of 5), meaning the extractor returned at least one span for the majority of questions carrying no gold evidence annotation. Given these denominators, a single question shifts the accuracy by 0.14 to 0.25, so the cross-bank differences should be read as indicative rather than precise. Two explanations are possible and cannot be fully disentangled. The system may genuinely struggle to recognize the absence of relevant evidence in Erste's

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

report, for instance, because topically related but non-answering passages mislead the extractor. Alternatively, since the Erste unanswerable labels were produced by a single annotator without quality control, some questions marked as unanswerable may in fact be answerable, making the annotation rather than the system the source of the discrepancy. In either case, abstention reliability cannot be treated as a stable property of the pipeline across institutions and warrants further investigation.

Table 5.4: Cross-bank generalization results using the selected pipeline configuration. All scores are reported as percentages.

Bank	**Retrieval**			**RapidFuzz F1**				**Additional Metrics**		
	Hit	**Recall@K**	**MRR**	**@50**	**@75**	**@90**	**@95**	**ROUGE-L**	**BERT**	**Abst.**
RBI	**100.0**	57.9	71.1	65.6	63.4	47.0	42.1	50.1	66.5	75.0
BAWAG	**100.0**	77.8	**85.6**	**78.9**	**78.1**	**64.0**	**57.1**	**65.8**	**80.8**	**85.7**
Erste	90.9	90.9	79.5	48.0	48.0	40.8	37.5	41.0	48.1	40.0
UniCredit	**100.0**	**93.2**	84.8	59.5	58.8	49.9	43.2	49.5	62.6	60.0

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.5 Discussion

5.5.1 Cross-experiment synthesis

As already indicated by the cross-bank results (Section 5.4), the central finding across all four experiments is an asymmetry between retrieval and extraction: the retrieval architecture generalizes robustly across institutions and configurations, while verbatim span extraction remains the more sensitive and variable stage of the pipeline.

On the retrieval side, the ablation study demonstrates that formal DDQ questions are poorly suited as direct search queries. The baseline hybrid retriever achieves a Hit Rate of only 0.583, confirming that the lexical and semantic gap between DDQ wording and annual report language is substantial. Query rewriting and multi-query expansion each contribute meaningfully to closing this gap, and their combination with cross-encoder reranking produces the strongest and most reliable retrieval configuration, achieving a perfect Hit Rate of 1.000 on the RBI development set. The LLM call overhead introduced by query processing is modest in the present DDQ setting, with two additional calls per question, and is reported explicitly in Table 5.1, where it can be weighed against the retrieval improvements obtained. In high-throughput production systems such as search engines or chatbots, however, additional LLM calls per query can introduce substantial latency and computational overhead.

The reranker comparison confirms that model choice has a measurable effect even when retrieval inputs are held fixed. The performance gap between `BAAI/bge-reranker-v2-m3` and the MS MARCO cross-encoders suggests that training data domain matters: the BGE model's multilingual and broader domain coverage appears better suited to the heterogeneous language of financial disclosures than models trained specifically on web passage retrieval.

Extraction performance is generally strong for the top-performing models but degrades noticeably under stricter matching thresholds. The drop from RF@50 to RF@95 across all extractors suggests that models more frequently produce broadly overlapping evidence than near-exact matches. The qualitative analysis shows that the decline at stricter thresholds reflects several failure modes, including over-inclusive extraction, alternative but valid formulations, imprecise span boundaries, and passages that discuss the correct topic while omitting a short material value. The decline therefore cannot be interpreted solely as a span-boundary problem.

Abstention behavior varies substantially between models and correlates with model capability. `openai/gpt-5.1` achieves perfect abstention accuracy across all five runs, while lighter models such as `google/gemini-2.5-flash-lite` and `meta-llama/llama-4-scout` fail to abstain reliably. The two top-performing models, `google/gemini-3-flash-preview` and `openai/gpt-5.1`, achieve nearly identical span matching scores RF@75-F1 of 0.690 for both, with a difference of 0.038 at RF@95 making it difficult to declare a clear winner on extraction quality alone. `google/gemini-3-flash-preview` is therefore selected for the cross-bank experiments on

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

practical grounds: it offers lower inference latency and substantially reduced API cost while producing comparable span extraction quality. In a production setting where abstention reliability is the primary concern and cost is less constrained, `openai/gpt-5.1` would be the more appropriate choice given its perfect abstention accuracy.

The cross-bank results confirm that the retrieval and extraction asymmetry identified on the RBI development set is not dataset specific. Hit Rate reaches 1.000 for three of the four banks, and Recall@K is consistently high, indicating that the retrieval architecture transfers well to different annual report structures and institution-specific language. Extraction performance varies more strongly, with Erste Group showing substantially weaker scores despite strong retrieval. The identical RF@50 and RF@75 scores for Erste (0.480) are particularly telling. Normally, a lenient threshold produces higher scores than a stricter one, because more spans qualify as a partial match. When both thresholds yield the same score, it indicates that the extracted spans do not even loosely resemble the gold annotations, suggesting that the system selected different passages rather than mis-cut the correct ones. This pattern underlines that verbatim extraction remains the bottleneck stage of the pipeline and is more sensitive to institution-specific document characteristics than retrieval.

5.5.2 Limitations

Several limitations constrain the scope and generalizability of the results presented in this thesis.

The first limitation concerns data availability and task scope. Bank-specific due diligence data is difficult to obtain due to confidentiality constraints and disclosure limitations in the DACH banking context. The experiments therefore rely on AFME-style questions that can be answered from publicly available annual reports. While this makes the evaluation feasible and reproducible, it covers only a subset of real DDQ workflows. In practice, many DDQ questions require evidence from internal policies, audit reports, regulatory filings, or security documentation that are not publicly accessible. Multi-source and non-public evidence settings are outside the scope of this thesis.

The second limitation concerns annotation quality and scale. Only the RBI annotations were produced collaboratively with a due diligence practitioner. The BAWAG, Erste, and UniCredit datasets were annotated by a single annotator, which introduces the risk of systematic annotation bias. Furthermore, the relatively small number of questions per bank means that the reported metrics should be treated as indicative rather than statistically robust estimates of generalization performance.

Finally, the evaluation focuses exclusively on annual-report-grounded verbatim evidence extraction. The benchmark does not contain questions whose correct answer requires synthesis across multiple documents or controlled paraphrasing of evidence. Some questions require evidence from multiple parts of the same annual report, but these can still be represented as multiple verbatim spans and therefore do not require the system to construct a new factual statement. Questions that require information to be

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

combined, derived, or reformulated, for example by comparing figures across documents or calculating a value from several disclosures, remain outside the scope of the benchmark. Such cases represent an important limitation of the VerbatimRAG approach and an important practical requirement for real DDQ workflows.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.5.3 Future Work

Several directions for future research follow:

- **Multi-source and agentic RAG.** Real DDQ workflows require evidence from heterogeneous sources, including annual reports, security policies, audit reports, regulatory filings, and internal documentation. An agentic architecture that coordinates source selection, retrieval, extraction, and answer assembly across these document types would be a natural extension of the pipeline proposed here and would bring the system closer to production-level DDQ answering.
- **Domain-specific reranking.** The reranker comparison shows that model training domain has a measurable effect on retrieval quality. The performance gap between `BAAI/bge-reranker-v2-m3` and the MS MARCO cross-encoders suggests that rerankers trained or fine-tuned on financial documents, regulatory disclosures, or DDQ-style relevance judgments could improve retrieval robustness further, particularly for institutions with less standard reporting language.
- **Fine-tuned span extraction.** LLM-based extraction performs well under lenient matching criteria but degrades at stricter thresholds, indicating that exact span selection remains difficult. The ACL-Verbatim results provide a direct example of the potential of dedicated extraction models: a fine-tuned ModernBERT token classifier outperformed the evaluated LLM-based extractors on their evidence-extraction benchmark Recski et al. (2026). Training or fine-tuning a comparable extraction model on financial-domain evidence, with annotations that explicitly define span boundaries, could therefore improve RF@90 and RF@95 performance and reduce unsupported extractions on unanswerable questions.
- **Multimodal document understanding.** Annual reports frequently present quantitative and qualitative information in tables, charts, and figures that are not captured by plain-text chunking. Integrating multimodal document understanding components could improve evidence coverage for questions that require information embedded in non-textual report elements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Conclusion

This thesis investigated the applicability of a grounded, verbatim answering approach to financial due diligence question answering over long and heterogeneous annual reports. Building on the VerbatimRAG framework originally developed for the medical domain, a multi-stage pipeline was designed, adapted, and evaluated in the context of compliance-oriented DDQ answering across four Austrian banking institutions. In addition to the system evaluation, the thesis contributes a newly constructed dataset comprising DDQ questions and manually annotated evidence spans for four Austrian banks, including expert-supported annotations for RBI. The study addressed three research questions concerning retrieval performance, verbatim span extraction quality, and the system's ability to abstain from answering when no valid evidence is present in the source documents.

Retrieval performance (RQ1). The retrieval ablation study demonstrates that raw DDQ questions are poorly suited as direct search queries. The baseline hybrid retriever achieves a Hit Rate of only 0.583 on the expert-annotated RBI dataset, confirming that the lexical and semantic gap between formal compliance language and annual report prose is substantial. Introducing cross-encoder reranking improves the ranking of relevant candidates, while query rewriting and multi-query expansion further increase the likelihood that suitable evidence enters the candidate set. The full pipeline, combining all three components, achieves the strongest overall retrieval performance and a Hit Rate of 1.000 on RBI, at the cost of two additional LLM calls per query. These results indicate that the strongest performance arises from the combination of query processing and second-stage reranking, rather than from any single component in isolation.

Extraction performance (RQ2). Among the evaluated span extractors, `google/gemini-3-flash-preview` and `openai/gpt-5.1` achieve the strongest performance, with identical RapidFuzz@75 F1 scores of 0.690 and near-identical ROUGE-L and BERTScore results. However, extraction quality degrades noticeably at stricter

matching thresholds across all models. The qualitative analysis shows that this decline reflects not only imprecise span boundaries, but also over-inclusive extraction, alternative valid evidence passages, and topically related spans that omit a requested material value. The non-generative semantic highlighting baseline performs substantially below the strongest LLM-based extractors, suggesting that the verbatim extraction task benefits from the contextual reasoning capabilities of large generative models. Overall, the models often identify broadly relevant evidence, but are less consistent in selecting complete and precisely bounded answer-bearing spans. Verbatim span extraction therefore remains the principal performance bottleneck in the evaluated pipeline.

Faithfulness and abstention (RQ3). Faithfulness at the textual level is enforced through verbatim span verification: only extracted spans that can be matched back to the retrieved source chunks are retained. This establishes that the returned evidence occurs in the underlying annual report, although it does not by itself guarantee that the selected passage is relevant, complete, or sufficient to answer the question. Abstention behavior varies substantially across models and institutions. On the RBI dataset, `openai/gpt-5.1` achieves an abstention accuracy of 1.000 across all five evaluation runs, compared with only 0.200 for `google/gemini-2.5-flash-lite` and 0.450 for `meta-llama/llama-4-scout`. In the cross-bank evaluation using `google/gemini-3-flash-preview`, abstention accuracy ranges from 0.857 for BAWAG to 0.400 for Erste. These results show that abstention reliability depends strongly on both the selected extractor and the institution and cannot be assumed to transfer consistently across model families or document collections.

Cross-bank generalization. The cross-bank experiment reveals a clear asymmetry between the two main pipeline stages. The retrieval architecture transfers robustly across institutions, achieving a Hit Rate of 1.000 for three of the four banks. Verbatim span extraction is considerably more sensitive to institution-specific reporting style, annotation granularity, and span boundary conventions, as reflected in the substantially lower extraction scores for Erste Group despite strong retrieval. This asymmetry constitutes the central empirical finding of this thesis: within the evaluated four-bank setting, the fixed retrieval configuration transfers effectively across institutions, whereas verbatim extraction remains the more institution-sensitive performance bottleneck.

Practical implications. The results support the use of the proposed pipeline as an assistive tool within a human-in-the-loop due diligence workflow. By grounding responses in verbatim evidence spans with explicit source citations, the system has the potential to reduce the effort required for evidence localization while supporting auditability. However, the residual extraction errors and imperfect abstention behavior observed across institutions confirm that the system is not suitable for autonomous decision-making. Final accountability must remain with the human reviewer, and extracted spans should be treated as evidence candidates rather than fully validated answers.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In summary, this thesis demonstrates that a multi-stage, verbatim-grounded RAG architecture can achieve reliable evidence retrieval across diverse annual reports, while also identifying span extraction and abstention calibration as the primary remaining challenges for production-level financial DDQ answering.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

- **Claude** (Anthropic, Claude Sonnet 5, 2026): Used as a writing aid for improving and refining text passages, as well as for writing and debugging code used in the experimental evaluation. All outputs were reviewed, edited, and validated by the author.
- **Cursor** (Anysphere, 2026): Used as an AI-assisted code editor for developing, refining, and debugging code used in the experimental evaluation. All generated code and suggested modifications were reviewed and validated by the author.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

2.1 The general architecture of hybrid retrieval methods. Source:Guo et al. (2022). 7
2.2 Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. Source:Brown et al. (2020) 10
2.3 Overview of the retrieval-augmented generation (RAG) workflow. Source: Databricks (2023) 15

4.1 System prompt used for query rewriting. Bank name and abbreviation are injected at runtime. 34
4.2 Prompt used for multi-query expansion. The model is instructed to return a structured JSON object containing up to three search queries. 35

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

3.1 Summary statistics of the annual reports included in the dataset 23
3.2 Question and annotation statistics per institution. The same sixteen questions are posed to every bank. Answerability is determined by the individual annual report, except for the four authored questions, which are unanswerable by construction. Span lengths are measured in words after whitespace normalization. 25
3.3 Selected gold answer spans for two AFME due diligence questions 27

4.1 Baseline pipeline configuration. Components marked "fixed" are held constant across the retrieval-side ablation studies in Section 4.4, while components marked "varied" are the ones manipulated by an ablation experiment. . . 33
4.2 Representative multi-query expansions and corresponding annual-report terminology 36
4.3 Retrieval ablation configurations. Each condition isolates one retrieval-side variation or combination, while keeping all other pipeline stages fixed. . . 42

5.1 Retrieval ablation results on the expert-annotated RBI dataset. Retrieval scores are reported as percentages. 50
5.2 Reranker comparison results on the expert-annotated RBI dataset. 51
5.3 Extraction model comparison results. LLM results are averaged over five runs. 52
5.4 Cross-bank generalization results using the selected pipeline configuration. All scores are reported as percentages. 56

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

Association for Financial Markets in Europe (n.d.). Introducing AFME.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 675–718.

Brådland, H., Goodwin, M., Andersen, P.-A., Nossum, A. S., and Gupta, A. (2025). A new HOPE: Domain-agnostic automatic evaluation of text chunking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 170–179.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, T., Wang, H., Chen, S., Yu, W., Ma, K., Zhao, X., Zhang, H., and Yu, D. (2024). Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15159–15177.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., and Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Choi, C., Kwon, J., Lopez-Lira, A., et al. (2025). Finagentbench: A benchmark dataset for agentic retrieval in financial question answering. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25)*, pages 632–637.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pages 758–759.

Dadopoulos, M., Ladas, A., Moschidis, S., and Negkakis, I. (2025). Metadata-driven retrieval-augmented generation for financial question answering. *arXiv preprint arXiv:2510.24402*.

Databricks (2023). The big book of MLOps. Databricks E-book, 2nd Edition.

Formal, T., Piwowarski, B., and Clinchant, S. (2021). SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Günther, M., Mohr, I., Williams, D. J., Wang, B., and Xiao, H. (2024). Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*.

Guo, J., Cai, Y., Fan, Y., et al. (2022). Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems*, 40(4):66:1–66:42.

Gupta, S., Ranjan, R., and Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Iaroshev, I., Pillai, R., Vaglietti, L., and Hanne, T. (2024). Evaluating retrieval-augmented generation models for financial report question answering. *Applied Sciences*, 14(20):9318.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Kovács, Á. (2025). Build hallucination-free rag with verbatim. Hugging Face Blog. Published November 18, 2025.

Kovács, Á., Schmitt, P., and Recski, G. (2025). Kr labs at archehr-qa 2025: A verbatim approach for evidence-based question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, N. F. et al. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

Recski, G., Tóth, S., Verdha, N., Boros, I., and Kovács, Á. (2026). ACL-Verbatim: Hallucination-free question answering for research. arXiv preprint arXiv:2605.21102.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Selenge, T. (2026). verbatim-r: Thesis-specific extensions to verbatimrag. `https://github.com/TuvshinSelenge/verbatim-r`. Open-source software, MIT License.

Shen, Y., Zhang, J., Chen, F., Yan, K., and Li, H. (2025). Finsearch: A temporal-aware search agent framework for real-time financial information retrieval with large language models. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, pages 10–17. Association for Computing Machinery.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. (2024). REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., Jiang, R., Wei, Y., and Xie, C. (2021). Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)*, pages 2614–2627.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wu, X., Li, S., Wu, H.-T., Tao, Z., and Fang, Y. (2025). Does RAG introduce unfairness in LLMs? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10021–10036.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTscore: Evaluating text generation with BERT.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):89:1–89:60.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., and Chua, T.-S. (2021). Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287. Association for Computational Linguistics.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.