Human Systems Engineering and Design (IHSED2025), Vol. 198, 2025, 159–169
https://doi.org/10.54941/ahfe1006772

Reliable Information Retrieval With LLMs: Automated Analysis and Comparison of Large PDF Documents

Lara Noe[1], Sebastian Breier[1], and Ruben Nuredini[2]

[1]BettercallPaul GmbH, 70565 Stuttgart, Germany
[2]Heilbronn University of Applied Sciences, 74081 Heilbronn, Germany

ABSTRACT

In high-stakes professional settings like the insurance industry, extracting information from complex PDF documents is critical yet challenging due to document length and technical language. While large language models (LLMs) offer new opportunities for automating document understanding, their usefulness depends on output accuracy, transparency, and verifiability. In critical contexts, users must be able to trace information back to credible sources to reduce risks from hallucinations or misinterpretation. This research explores how LLMs can support reliable, transparent information retrieval (IR) from complex documents. We introduce five IR system variants designed to iteratively improve LLM outputs through better context preservation and fine-grained source attribution. These systems incorporate enhancements such as Markdown-based parsing, retrieval-augmented generation (RAG), and advanced document preprocessing. The final iteration integrates Multi-view Content-aware (MC) indexing, which supports semantically targeted retrieval using keyword, summary, and raw-text views. To evaluate performance, we develop a domain-specific benchmark with curated insurance documents, ground-truth answers, as well as performance metrics for assessment of the answer accuracy, hallucination rate, and source attribution precision. Results show that systems with content-aware chunking or MC-Indexing outperform earlier versions in accuracy and attribution, though with added complexity. Our findings highlight the value of structure-preserving preprocessing, targeted retrieval, and source transparency in developing trustworthy AI tools for document analysis. Future work may explore automated verification loops and user-guided retrieval to improve interpretability and reliability further.

Keywords: Large language models, Information retrieval, Document understanding, Retrieval augmented generation

INTRODUCTION

Large language models (LLMs) have rapidly evolved into popular tools to support users in text-centric tasks, like summarization, question answering, or information retrieval (Bommasani *et al.*, 2022; Joshi *et al.*, 2023). Their capabilities in generating human-like responses and the speed at which they can process vast amounts of information have sparked interest in automating many labor-intensive workflows, in both personal and professional domains.

Among the many emerging application areas for LLMs, one promising domain is the automation of document understanding tasks. Lengthy, structurally complex documents, such as contracts, technical documentation or policies, are often difficult to interpret and compare, even for domain experts (Shaik, 2019). This research was inspired by a concrete example within the insurance industry: to assess their standing on the market and refine their product catalog, insurance companies regularly perform manual comparisons between insurance policies of various competitors. This time-intensive and error-prone task could potentially be streamlined using an LLM-driven information retrieval and comparison system[1].

However, the practical usefulness of any LLM-assisted system is not only determined by the quality of generated responses, but also by the traceability and verifiability of these outputs. It is well-established that language models still struggle with the problem of hallucination, where responses are factually incorrect or fabricated but still sound plausible and match the query at hand (Ji *et al.*, 2023). While retrieval-augmented generation (RAG) frameworks aim to improve factual grounding, they still suffer from fragmented context handling and insufficient source attribution (Nematov *et al.*, 2025). This limitation becomes especially critical in high-stakes environments, like the insurance industry, where incorrect or unverifiable information can lead to operational or legal risks.

This research investigates a central question: How can we enhance the reliability, accuracy, and traceability of LLM-generated answers when working with complex, semantically structured PDF documents? To explore this, we present a series of five system prototypes that incrementally apply advanced document parsing, RAG, and indexing strategies. Each version is designed to mitigate known limitations of LLM-based retrieval, such as context loss, irrelevant retrieval, and opaque source attribution. We perform comparative statistical analysis to assess the performance differences between the different approaches.

ITERATIVE DEVELOPMENT OF ENHANCED IR SYSTEM VARIANTS

To enable automated insurance comparison, a concept for an LLM-driven PDF analysis and comparison tool was developed previous to this research. The idea of this system was to enable users to upload insurance policy PDFs and define a list of comparison questions. Once the answer generation is triggered, the system generates one combined prompt per document, consisting of system instructions, the document's plain text contents, and the entire question list. This prompt is passed to an LLM to produce document-specific answers.

While an initial proof-of-concept prototype of this system demonstrated that LLMs can extract useful information from insurance policy documents, it also revealed critical limitations of the system: lack of traceability,

[1]These insights are based on internal discussions with industry practitioners who are directly involved in conducting comparative analyses of insurance products.

undetectable hallucinations and limited options for source attribution, due to the loss of structural information within the plain text document representation. To address these challenges, we designed five successive prototype variants that build on the same foundational workflow as the initial concept. All systems use an identical system prompt and rely on the LLaMA 3.3 Instruct (70B) model (Meta Llama, 2024), selected for its robust instruction adherence and strong performance in structured question answering tasks (Fourrier *et al.*, 2024). Table 1 provides an overview of key differences between the developed system variants.

Table 1: Description of five iterative development steps of an LLM-driven IR system for automated PDF analysis and comparison.

Prototype	Context-Handling	Description
P0: Baseline	Full document text as plain text	Extracts plain text from PDF; includes entire content in prompt. The LLM is instructed to generate a short answer, supporting quote, page number and list of all parent headers of quoted passage. A heuristic hallucination flagging mechanism compares answers against source text to flag potentially unreliable answer components.
P1: Markdown Parsing	Full document text in Markdown format	Preservation of structural and contextual information like headings and page breaks through OCR-based parsing to Markdown format.
P2: RAG	Fixed-length character-based text-splitting; retrieval of three most relevant sections	Integration of retrieval augmented generation approach (Lewis *et al.*, 2021); Splits document into fixed-size segments embedded using the jina-embeddings-v2-base-de model (Mohr *et al.*, 2024). Uses hybrid retrieval (dense FAISS and sparse BM25) and re-ranking to fetch the three most relevant sections per query, improving focus and reducing context overload.
P3: Content-Aware Splitting	Markdown header-based splitting into the original document sections	Splits document into contextually coherent sections using Markdown headings to avoid arbitrary fixed-length splits and better preserve document structure (Dong *et al.*, 2024).

Continued

Table 1: Continued

Prototype	Context-Handling	Description
P4: MC-Indexing	Markdown header-based splitting and Multi-view Content-aware (MC) indexing based on Markdown sections as described by Dong et al. (2024)	Builds on P3 by generating two additional "views" per section: keyword lists and summaries, created via LLaMA 3.1 Instruct (8B) (Meta, 2024). These, along with raw text, are embedded and stored. A multi-vector retriever fetches relevant segments from any view, but raw text is passed to the LLM for answer generation.

EXPERIMENTAL DESIGN

To enable an objective comparison between the five previously described variants of the IR system, we conducted an evaluation on a custom benchmark for insurance comparison. The following sections describe the methodology used to assess the performance of each prototype version in terms of answer quality, source attribution accuracy, and hallucination rates.

Evaluation Benchmark

The evaluation in this study is based on a purpose-built benchmark dataset, specifically designed to support the comparison of insurance policy documents in German language. The benchmark covers three distinct insurance topics for comparison: dental, life, and animal health insurance. It includes a total of 15 documents (5 per domain) and 60 user-centered questions (20 per domain). This results in 300 unique topic-question-document combinations, each annotated with structured ground truth data: the expected short answer, a supporting quote, and detailed source attribution, including the page number, a hierarchical parent header list and the directly relevant "perfect" header.

The benchmark reflects real-world use cases and domain-specific challenges: documents vary in structure, terminology, and formality, and some questions are only answerable in a subset of documents. This introduces realistic variability in difficulty and supports fine-grained evaluation of answer accuracy, contextual grounding, and source traceability.

Evaluation Metrics

To assess the performance of the different IR system variants, we use a set of benchmark-specific evaluation metrics. Together, these metrics support a multi-faceted evaluation of each system's ability to extract, justify, and correctly attribute relevant information from PDF documents.

Answer Metrics. For each user question, the system generates two types of answers: (1) a concise short answer that directly addresses the question and (2) a direct quote from the source document that supports the short answer and provides contextual evidence. To evaluate the accuracy and faithfulness of these outputs, the benchmark defines two answer-level metrics.

Four categories of possible short answers are defined in the benchmark, and different checks are used to assess the correctness of generated short answers based on the expected answer category. Binary, numerical, and time span answers are evaluated using rule-based comparison techniques, including normalization and pattern matching. Textual answers are assessed using cosine similarity between text embeddings created using the jina-embeddings-v2-base-de model (Mohr *et al.*, 2024). A similarity threshold of 0.5 determines correctness (Crocetti, 2015).

The faithfulness of the quoted passage is assessed using the ROUGE-L score (Lin, 2004). This metric measures how closely a generated quote matches the expected reference quote through longest common subsequence (LCS) matching. The final score is the ROUGE-L F1 measure, which balances precision and recall, allowing for minor variations in the length of quoted passages while still ensuring close alignment with the expected quote.

Source Attribution Metrics. For each generated quote, the newly developed system versions P0-P4 return associated source attribution metadata, including the page number and a list of related parent headers that help localize the referenced passage within the source document.

The benchmark defines four complementary metrics to assess the accuracy of this data compared to the expected values: (1) Page Found assesses the correctness of the generated page number; (2) Perfect Header Found checks whether the expected "perfect header" can be found within the generated header list; (3) Header Intersection Rate reflects the percentage of accurately returned headers in the generated header list; and (4) Location Score is a composite metric represented by the average across the above three components to assess the overall fidelity of the source attribution.

Hallucination Rates. This metric group is not benchmark-defined but makes use of the hallucination flagging mechanism implemented by each prototype variant. Each answer generated by the system contains a set of four hallucination flags, which were used to evaluate the hallucination rates across the different prototype versions. The mechanism sets four Boolean flags for each answer element: (1) "Quote" checks whether the quote can be found within the original text; (2) "Headers" raises a flag if any of the generated headers does not appear in the source; (3) "Page" determines whether the quoted passage appears on the page referenced by the generated page number; and (4) "Irrelevant Headers" assesses whether any of the generated headers are relevant to the quoted passage, i.e. can be found on the same page or one prior.

Experimental Procedure

To account for the non-deterministic nature of LLMs, each system variant was executed 40 times on the benchmark. Each run constitutes a complete pass through the benchmark, meaning that the system generates one answer for every topic-question-document combination defined by the task. This results in 300 individual answers per run. After completing all 40 runs for a

system, each answer is compared to the expected output values and assessed using the defined metrics. The metric scores are then averaged across all answers within one run, yielding a single mean score per metric. Overall, this process produces 40 independent data points per system, which serve as the basis for downstream statistical analysis.

Given the sample size per group (n = 40), the Central Limit Theorem justifies the assumption of normality, supporting the use of parametric tests (Ross, 2014). For each comparison, we perform two-tailed independent t-tests with a significance level of $\alpha = 0.05$. A Welch's t-test was used for comparisons where Levene's test indicated unequal variances; otherwise, Student's t-test was used. To indicate not only statistical significance but also the magnitude of observed effects, we report Cohen's d effect sizes alongside p-values.

RESULTS

This chapter presents the empirical findings based on 40 independent runs for each system prototype. The results are organized around three key dimensions: answer quality, source attribution, and hallucination rates. Trends are visualized in Figures 1–3, while Table 2 presents statistical comparisons across successive versions.

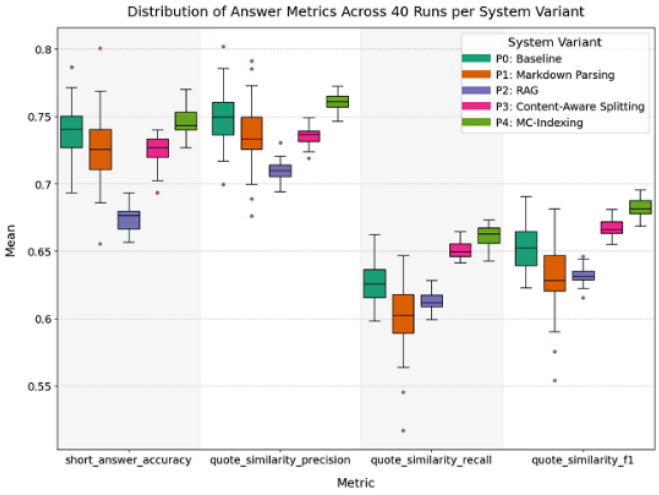

Figure 1: Boxplots describing the distribution of answer accuracy across 40 independent runs of each prototype system on the benchmark.

Answer Quality

As illustrated in Figure 1, median scores for answer-level metrics generally increase across prototype versions, while variability decreases. Early prototypes (P0 and P1) display higher variability and lower quote metric performance. Later prototypes, particularly P4, achieve the highest median scores and lowest variability across all answer quality metrics. These results indicate a trend toward more consistent and higher-quality answer generation

in later system iterations. This trend can also be observed in the results of the two-sample t-tests presented in Table 2: Initial significant declines in both short answer and quote metrics are followed by significant improvements for later prototype stages.

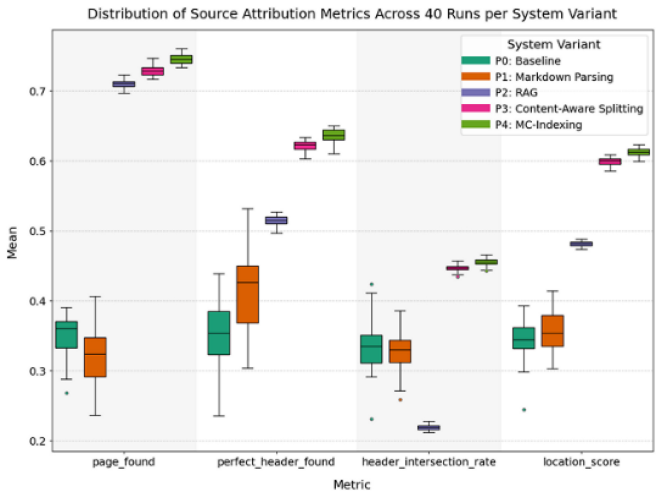

Figure 2: Boxplots describing the distribution of source attribution metrics across 40 independent runs of each prototype system on the benchmark.

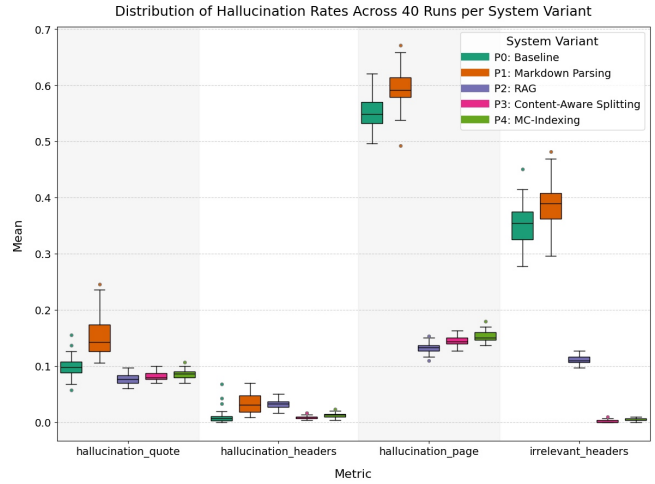

Figure 3: Boxplots describing the distribution of hallucination rates across 40 independent runs of each prototype system on the benchmark.

Source Attribution

For all source attribution metrics, median values increase, and variability decreases in later system versions, as illustrated in Figure 3. Early prototypes show lower and more variable scores, while P3 and P4 demonstrate higher medians and narrower interquartile ranges. As presented in Table 2, the composite location score significantly improves with every prototype

iteration. Overall, the results show that later systems provide more accurate and consistent source attribution.

Hallucination Rates

Hallucination rates, as measured by the percentage of flagged responses in each category, decrease across successive system versions. The earlier prototype versions P0 and P1 show elevated median hallucination rates across most metrics. P2 reduces some hallucination types, and P3 and P4 exhibit the lowest rates and least variability across all hallucination categories. These patterns indicate that later system versions are associated with fewer hallucinations and greater consistency in output. As detailed in Table 2, hallucination rates fluctuated across system versions, with significant reductions from P1 to P2, but notable increases in most metrics between P0 to P1 and again from P3 to P4.

Table 2: Pairwise t-test results between successive prototype versions. Each cell shows the *p*-value (significance of change) and Cohen's *d* (effect size), with (+) indicating a significant improvement, (–) a significant decline, and (ns) a non-significant change. Welch's or Student's t-tests were used based on variance equality.

	P0 → P1	P1 → P2	P2 → P3	P3 → P4
Short Answer Accuracy	.02 (–), 1.0	<.01 (–), 2.69	<.01 (+), – 4.98	<.01 (+), – 1.92
Quote Similarity (ROUGE-L F1)	<.01 (–), .97	–.75 (ns)	<.01 (+), 5.41	<.01 (+), 2.37
Location Score Hallucination	p = .11 (ns)	<.01 (+), 6.29	<.01 (+), 26.15	<.01 (+), 2.29
Quote	<.01 (–), – 1.94	<.01 (+), 3.10	<.01 (–), –.77	.06 (ns)
Page	<.01 (–), – 1.39	<.01 (+), 17.88	<.01 (–), 1.70	<.01 (–), –.81
Headers	<.01 (–), – 1.56	–.95 (ns)	<.01 (+), –4.09	<.01 (–), –.68
Irrelevant Headers	<.01 (–), –.99	<.01 (+), 9.32	<.01 (+), 20.25	<.01 (–), – 1.18

DISCUSSION

The results of this study reveal distinct performance characteristics among the evaluated LLM-based document IR systems. In terms of answer performance, the baseline system P0 already shows strong capabilities at abstracting information from large context sizes and quoting relevant passages. Notably, the traditional RAG approach underperforms all other systems, supporting the idea that arbitrary context splitting can severely reduce retrieval and answer accuracy. P4 restored short answer accuracy to levels comparable to P0, which confirms that enhancements in retrieval and structural augmentation do not come at the expense of short answer quality.

Quotation accuracy showed a similar trend, with P4 outperforming all earlier system versions. P1 produced the highest rate of quote hallucinations, likely due to formatting inconsistencies and Markdown parsing artifacts. The added noise from including entire document contents appeared to impair

instruction following for P0 and P1, leading to increased hallucination rates. By contrast, RAG-based systems with targeted, concise prompts (P2–P4) largely overcame these issues. Despite recent advances enabling large context windows, our results show that including the entire document as context can degrade answer performance.

The performance across source attribution metrics improved consistently from P0 through P4. Adding parent headers and page numbers as explicit metadata, along with refining section-level retrieval made it easier for the LLM to accurately locate and attribute information within the documents. The most reliable results were achieved by versions P3 and P4, which both used markdown header-based section splitting and enriched the context with relevant metadata. These systems almost entirely eliminated header-based hallucinations, with mean rates dropping as low as 0.002 and 0.005. This demonstrates that providing clear structural cues and comprehensive metadata is highly effective for minimizing attribution errors.

The retrieval strategy emerged as a central factor in overall system performance. P2, while the first to implement RAG, underperformed due to its fixed-length chunking approach, which often fragmented the logical structure of source documents. P3 and P4, incorporating structure-aware indexing and document-defined sectioning, substantially improved both source attribution precision and answer accuracy. These findings highlight the importance of aligning retrieval mechanisms with document structure to support trustworthy and accurate information retrieval from documents.

While these results provide valuable insights, several limitations should be acknowledged to further contextualize the findings. Firstly, the evaluation benchmark was custom-developed for this study and is limited to 15 German-language insurance documents. Although this benchmark was designed to reflect real-world use cases, its narrow focus may limit the generalizability of the results to other insurance products, document formats, or languages. Broader validation will be necessary to confirm that the proposed systems are applicable across other domains. In addition, the limited dataset size may not capture the full spectrum of document complexity or user questions encountered in production environments. As a result, the statistical power of the comparisons may be limited.

Furthermore, the statistical comparisons were conducted using parametric tests, supported by the Central Limit Theorem given the number of samples per system. However, based on this assumption, no formal checks for normality were conducted beyond Levene's test for equal variances. If the underlying data distributions are highly skewed, the results of these tests may be less robust.

Moreover, the hallucination detection method used in this study relies on heuristics, such as exact string matches and page number alignment, to identify unsupported content. While effective for certain hallucination types, this approach does not capture all forms of factual inconsistency, and the approach has not yet been empirically validated. This limitation should be considered when interpreting hallucination-related results, since they are only intended as indicators and may produce false positives. Looking ahead, the proposed set of overlap-based metrics using ROUGE and string comparisons could be further complemented with more advanced evaluation metrics,

which align better with human judgement. The integration of model-graded approaches, like G-Eval (Liu *et al.*, 2023), could offer more nuanced insights into semantic similarities, domain-specific correctness, and faithfulness even when lexical overlap is low.

Despite its limitations, this research provided valuable insights into the performance of the successive prototype versions developed to build upon the original concept for an automated analysis and comparison system for insurance documents. Based on these findings, the next logical step is to integrate the most promising techniques, like header-based splitting and MC-Indexing, into a complete system with a user interface (UI). Developing a dedicated UI would not only support pilot deployment within the insurance sector but also enable the collection of direct feedback from end users.

A more advanced UI could incorporate the generated source attributions and hallucination flags into interactive verification features that allow users to review and validate answers directly within their workflow. Besides the added transparency for users, these feature integrations could enable a systematic evaluation of their impact on user trust, perceived reliability, and overall usability, particularly in high-stakes environments.

In parallel, this work did not assess computational complexity, such as indexing cost, required memory size, or end-to-end latency. Since these factors will be crucial for sizing and running on-prem production deployments of the system, conducting a formal evaluation of relevant complexity measures across prototype variants should be a critical focus for future work.

To further refine system performance, active learning strategies could be explored. For example, by integrating user corrections into a continuous feedback loop, the system could incrementally improve its accuracy and robustness over time. Future work should also explore the use of other state-of-the-art open-source models or the fine-tuning of these models on domain-specific insurance documents to further enhance answer quality and source attribution accuracy.

CONCLUSION

This study demonstrates that advanced context-handling techniques can improve the performance of information retrieval systems not just in answer accuracy but also in document-level traceability, which is critical for risk-sensitive domains like insurance. The evaluation framework and insights presented here offer a solid foundation for future research into more accountable and transparent document AI systems, with a strong focus on enhancing user trust and reliability.

ACKNOWLEDGMENT

This research builds upon the findings of my master's thesis, developed in collaboration with BettercallPaul in Stuttgart and the Heilbronn University of Applied Sciences. I am deeply grateful for the opportunity to explore this

topic and for the invaluable technical guidance and professional support provided by all supervisors involved throughout the course of this project.

REFERENCES

Bommasani, R. *et al.* (2022) 'On the Opportunities and Risks of Foundation Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2108.07258.

Crocetti, G. (2015) 'Textual Spatial Cosine Similarity'. arXiv. Available at: https://doi.org/10.48550/arXiv.1505.03934.

Dong, K. *et al.* (2024) 'Multi-view Content-aware Indexing for Long Document Retrieval'. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.15103.

Fourrier, C. *et al.* (2024) 'Open LLM Leaderboard v2'. Hugging Face. Available at: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Ji, Z. *et al.* (2023) 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys*, 55(12), pp. 1–38. Available at: https://doi.org/10.1145/3571730.

Joshi, I. *et al.* (2023) "'With Great Power Comes Great Responsibility!": Student and Instructor Perspectives on the influence of LLMs on Undergraduate Engineering Education'. arXiv. Available at: https://doi.org/10.48550/arXiv.2309.10694.

Lewis, P. *et al.* (2021) 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.11401.

Lin, C.-Y. (2004) 'ROUGE: A Package for Automatic Evaluation of summaries', in *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*, p. 10.

Liu, Y. *et al.* (2023) 'G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment'. arXiv. Available at: https://doi.org/10.48550/arXiv.2303.16634.

Meta (2024) *Llama 3.1 | Model Cards and Prompt formats*. Available at: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/ (Accessed: 15 July 2025).

Meta Llama (2024) *meta-llama/Llama-3.3–70B-Instruct*. Available at: https://huggingface.co/meta-llama/Llama-3.3–70B-Instruct (Accessed: 15 July 2025).

Mohr, I. *et al.* (2024) 'Multi-Task Contrastive Learning for 8192-Token Bilingual Text Embeddings'. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.17016.

Nematov, I. *et al.* (2025) 'Source Attribution in Retrieval-Augmented Generation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2507.04480.

Ross, S. M. (2014) 'Chapter 6 - Distributions of Sampling Statistics', in S. M. Ross (ed.) *Introduction to Probability and Statistics for Engineers and Scientists (Fifth Edition)*. Boston: Academic Press, pp. 207–233. Available at: https://doi.org/10.1016/B978-0-12-394811-3.50006-X.