Queryable Shared Reference Repository

Atyab Hakeem, Kishan Sathish Babu, Naga Kushal Ageeru, Pranav Kanth Anbarasan
https://github.com/hakeematyab/Queryable-Shared-Reference-Repository

1. Summary

1.1 Problem Description

Research groups must manage an ever-growing volume of scientific literature. While reference managers allow storage and basic retrieval, they lack intelligent, context-aware querying that integrates both paper content and metadata. Large Language Models (LLMs) can enhance search and synthesis but raise privacy concerns and introduce risks of hallucination and inconsistent accuracy. This project aims to address these limitations by developing a on-device, shared, queryable repository of scientific papers that enables natural language queries and minimizes fabricated outputs through careful design and evaluation.

1.2 Dataset

The dataset currently consists of 300 curated scientific papers in PDF form, along with supplementary webpages and bibliographic files (e.g., .bib). Both full text (abstracts, methods, results, figures, tables) and metadata (title, authors, journal, year, DOI, citations) are extracted and structured for flexible querying. Early processing includes text cleaning, formatting corrections, and semantic chunking to optimize retrieval. While we currently have 300 papers, the dataset may scale to around 3,000**–**10,000 papers.

1.3 Related work:

Semantic Scholar, developed by the Allen Institute for AI, indexes over 200 million academic papers using machine learning and citation-informed embeddings to provide AI-enhanced search and research recommendations. CORD-19 aggregated 1M+ coronavirus papers through Semantic Scholar with structured full-text parses, facilitating numerous COVID-19 text mining systems. However, both systems' cloud-based architectures pose privacy risks for sensitive research data and lack integrated hallucination detection mechanisms to ensure factual accuracy. Our work addresses these limitations by developing a secure, on-premises RAG system with query-type-specific hallucination detection, specifically designed for research groups managing proprietary scientific literature.

2. Methods

2.1 Data Processing

The data processing pipeline standardizes diverse research documents into clean, metadata-linked text suitable for querying and analysis. PDFs are parsed using PyMuPDF for direct text extraction from page objects, while BeautifulSoup collects visible text from HTML sources, and python-docx is used for DOCX files. Unreadable or restricted files are automatically skipped to maintain data integrity. All processed text files are stored in a uniform structure with a one-to-one mapping to their raw documents. Metadata tagging is performed by parsing .ris files (a standard citation format used by reference managers) with rispy to extract title, authors, year, and DOI. Titles are cleaned and deduplicated, and filenames are fuzzy-matched to metadata using RapidFuzz at a ≥ 70% similarity threshold.

We additionally experimented with small language models (SLMs) for document transcription and structured text generation. IBM's Granite-Docling model, comprising approximately 0.3 billion parameters, was employed for general-purpose image-to-text transcription. It is designed to handle multiple document formats and types, offering robust capabilities for converting visual data into text. Additionally, Nougat (Neural Optical Understanding for Academic Documents) was utilized to process research-oriented PDFs. Nougat follows a vision encoder-decoder architecture, combining a Swin Transformer as the visual encoder with an mBART-based text decoder. This configuration allows the model to interpret PDF pages as images and produce structured markdown or LaTeX output. The base model contains around 250 million parameters, while the larger variant scales to approximately 350 million, striking a balance between computational efficiency and high accuracy on complex scientific documents.

2.2 Synthetic Data Generation

Synthetic evaluation data was generated using Llama 3.3 70B (Meta, 2024) at half precision. The model generated questions along with the exact excerpt to answer the corresponding questions. Similarly, to evaluate the hallucination detection system, a synthetic dataset was generated with a question, a true answer, and a wrong answer. To make the evaluation challenging, the model was instructed to generate wrong answers that resembled the correct answer.

2.3 Embedding Models

Embedding models convert text into dense vector representations that capture the semantics of the text. Similarity search can be performed on these embeddings to retrieve similar text. Through surveying the embedding models with the best reported performance in the scientific literature domain, five models were shortlisted: SPECTER (Cohan et al., 2020), MiniLM-L6 (Wang et al., 2020), MPNet (Song et al., 2020), SciNCL (Ostendorff et al., 2022), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2020).

2.4 Text Chunking

Chunking is a process of segmenting large documents into smaller, semantically coherent units for efficient retrieval and processing. Since embedding models have limited context windows and LLMs exhibit reduced performance when processing excessive irrelevant information, effective chunking strategies must balance preserving contextual completeness with maintaining retrieval precision. Recursive character splitting hierarchically divides the text using a prioritized sequence of separators (newlines, paragraphs, section breaks), splitting the text into semantically coherent sections. This approach is efficient and preserves naturally existing boundaries. Semantic chunking uses the similarity between the embeddings of the text to merge similar text or identify breakpoints. Although this approach groups text that is semantically similar, it is resource-heavy.

Smith and Troynikov (2024) found that recursive character text splitting with a chunk size of 200 and without any chunk overlap offered the best tradeoff between latency and quality. Subburaj et al. (2025) demonstrated that medium-sized chunks of 512-1,024 tokens performed best for complex analytical queries in financial documents. Cluster semantic chunking, on the other hand, yielded the highest evaluation metrics, albeit at the expense of greater latency and resource consumption. Due to practical constraints, including time and resources, our search was limited to these strategies. The shortlisted embedding models were evaluated against the synthetic question-excerpt dataset across the aforementioned chunking strategies. Hit rate and MRR (Mean Reciprocal Rank) were used to assess the chunking strategies.

2.5 Hallucination

The definition of hallucination varies based on the project scope–hence, we define hallucination on three different types of queries: retrieval-based, fact-based, and brainstorming. For retrieval-based queries, any generated output that is not grounded in a retrieved document is a hallucination. Outputs with low confidence score or high perplexity (model internal metrics) will be classified as hallucinated for fact-based queries. It is important to note that it doesn't necessarily mean the output is false; it simply means the model has not seen the information beforehand, which in many cases can be a proxy for untrue statements. For the same reason, we propose an unshackled mode that would not provide warnings for hallucination for brainstorming queries. This allows unbridled idea exploration, as novel ideas would likely have low confidence.

MiniCheck (Tang et al., 2024) and (HHEM-2.1-Open; Bao et al., 2024) models were shortlisted and benchmarked. To make information less prone to interpretation and more calibrated to our experimental findings, we propose a three-tiered actionable hallucination reporting system based on our experimental results (Figure A.3, A.4).

2.6 Evaluation Metrics

To evaluate the quality of chunking, two primary metrics are used: Hit Rate and Mean Reciprocal Rank (MRR). Hit Rate@K measures the quality of the retrieval of documents by computing the proportion of queries for which the retrieved K documents contain the relevant document. MRR evaluates the ranking quality by computing the mean of the reciprocals of the ranks of the relevant document. A document is considered relevant if there is a certain percentage of overlap between the excerpt (ground truth) and the retrieved document.

Hit Rate = $(1/n) \sum_{i=1}^{n} \mathbb{1}(\text{recall}_i > \tau)$ and **MRR** = $(1/n) \sum_{i=1}^{n} (1/\text{rank}_i)$

where $\text{recall}_i = |D_i \cap G_i| / |G_i|$; n = number of queries; D_i = retrieved documents for query i; G_i = ground truth documents for query i; τ = threshold; $\mathbb{1}(\cdot)$ = indicator function; rank_i = rank of the first relevant document; relevant document is defined as $\text{recall}_i > \tau$.

3. Results

3.1 Exploratory Data Analysis

The dataset comprises computational imaging biology papers primarily in PDF format, with HTML and DOCX variants. Metadata (DOI, publication) enables initial filtering. Table A.1 shows the text data statistics. Figure A.1 visualizes that the jargon used in computational imaging biology is most prevalent. Figure A.2 shows a highly right-skewed document size distribution with most documents being around a thousand words.

An initial examination of relevant models showed the token-to-word ratio (~1.5) indicates minimal word-splitting by Byte Pair Encoding, meaning technical jargon exists in the model vocabulary. Therefore, models should perform well without fine-tuning, allowing us to focus on optimizing data transformation for retrieval.

3.2 Embedding Models & Chunking Strategies

Figures 1 and 2 provide quality metric scores - Hit Rate and MRR, across embedding models and chunking strategies. MiniLM consistently obtains the highest scores across the chunking strategies, followed by MPNET. Strategies 2 and 4 (Table 1) obtain the highest scores across the models. In addition, MiniLM also has the lowest parameter count (Figure 3) and query time (Figure 4). Figure 5 shows the normalized, weighted scores computed across the quality metrics and runtime performance metrics (parameter count and search time). Quality metrics were weighted heavily at 85%, with parameter count and search time at 10% and 5% respectively, reflecting the system's high fidelity requirements. MiniLM with strategy 4 achieved the highest combined score.

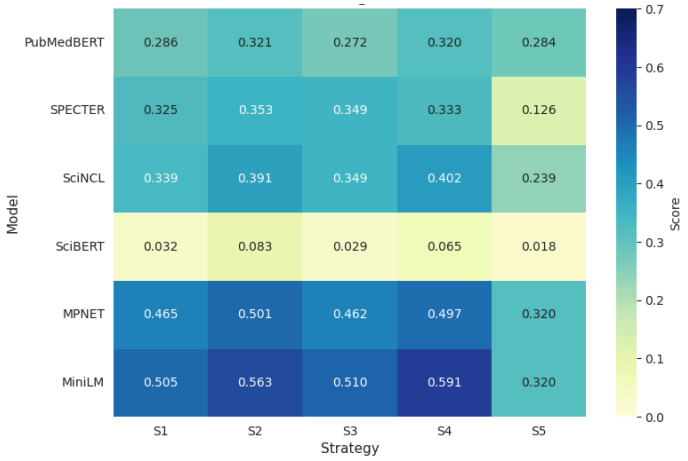

Figure 1: Heat map of Hit Rate across embedding models and chunking strategies with parsed text data

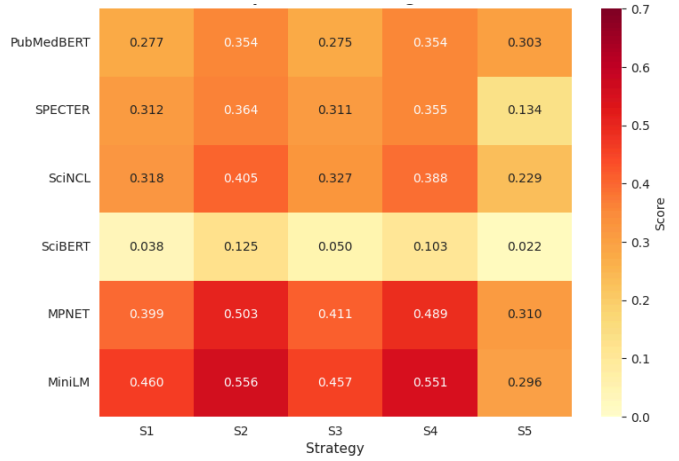

Figure 2: Heat map of MRR across embedding models and chunking strategies with parsed text data

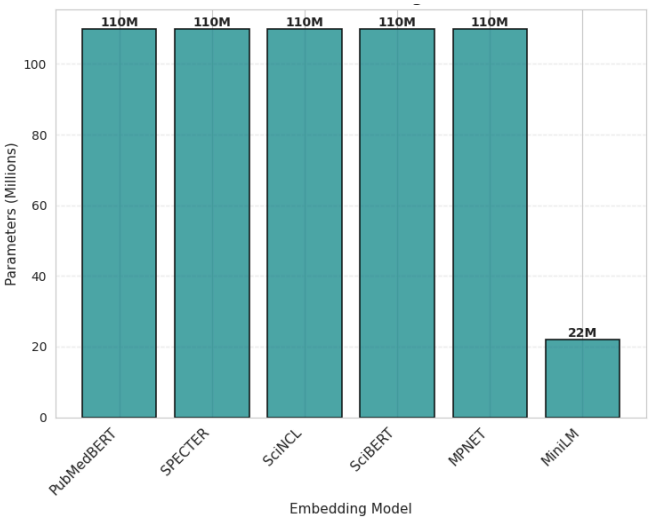

Figure 3: Comparison of the parameter count across embedding models

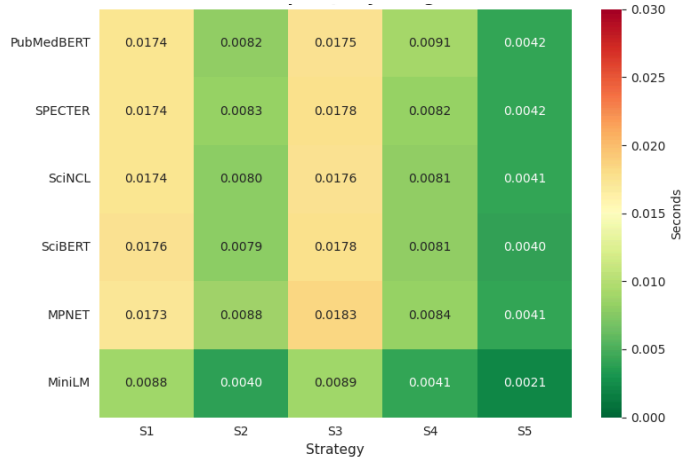

Figure 4: Heat map of MRR across embedding models and chunking strategies with parsed text data

.

Strategy	Type	Chunk Size	Chunk Overlap
S1	Recursive	256	0
S2	Recursive	512	0
S3	Recursive	256	25
S4	Recursive	512	50
S5	Semantic	**Breakpoint** : Percentile	**Threshold**: 90

Table 1: Chunking strategy definition

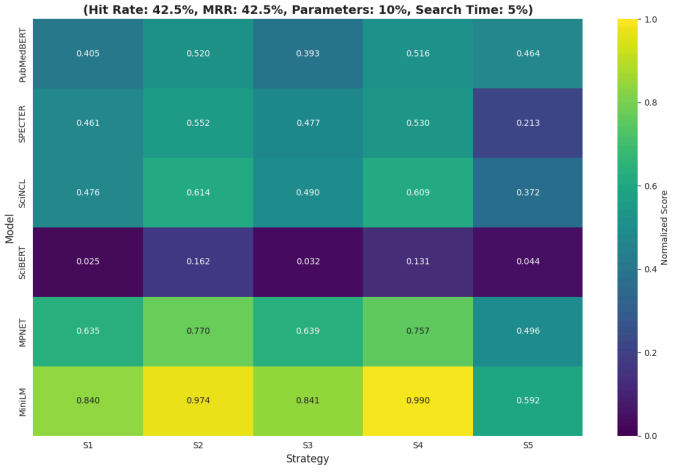

Figure 5: Heat map of combined scores across embedding models and chunking strategies with parsed text data

3.3 Hallucination Detection

Bespoke Flan-T5 achieved the highest F1 score (Figure 6), followed by Bespoke MiniCheck. However, Bespoke RoBERTa had the highest F1 score per billion parameters (Figure 7). To achieve the optimal tradeoff between precision and recall, grid search over varying thresholds was performed (Figure A.3). This was especially important as recall dropped precipitously past the optimal

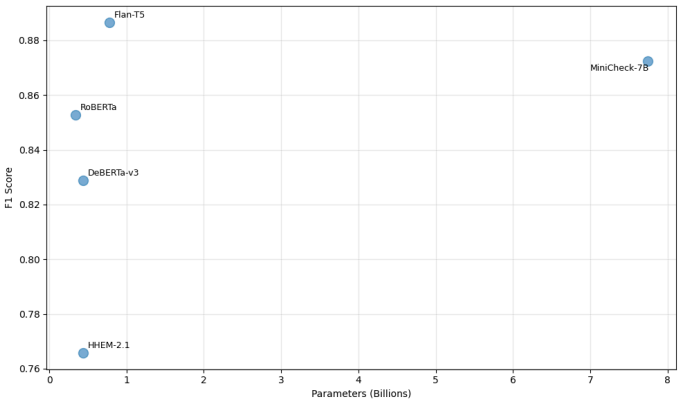

Figure 6: Plot of F1 score vs the parameter count across hallucination detection models

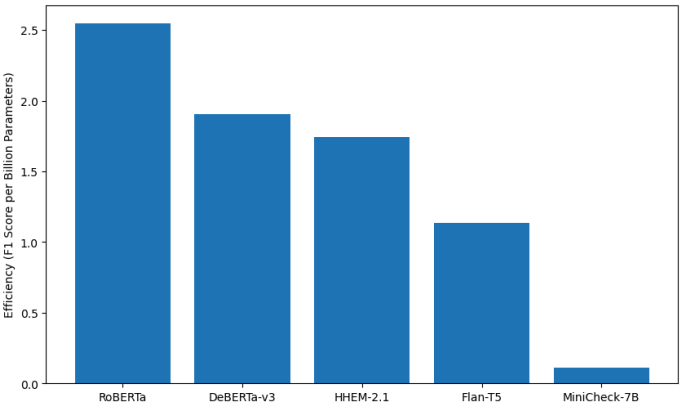

Figure 7: F1 Score per billion parameters comparison across hallucination detection models

threshold. Based on these experimental results, we propose a three-tiered actionable hallucination reporting system. This would not only prevent scores from being misconstrued but also ensure that the reports are calibrated to our experimental findings. Qualitative analysis of the results on Bespoke RoBERTa revealed that the model is prone to false negatives when the inputs are numerical and prone to false positives when the hallucinated content is similar to the truthful content. Additionally, the model struggles when the text is lengthy and with irregular formatting.

4. Discussion

The prevalence of chunking strategies 2 and 4 across embedding models indicates that larger text segments with greater context completeness are favored. An MRR of ~55% indicates that relevant queries are ranked between 1st and 2nd. A hit rate of ~60% indicates that approximately 60% of queries successfully fetched relevant content. This could either mean that the system accuracy is suboptimal or that the evaluation dataset is overly stringent. Examining the synthetic data revealed that a sizeable number of questions were not self-contained. However, a tradeoff must be made to ensure queries are both plausible and sufficiently challenging by providing adequate keywords to resemble deployment conditions, which is challenging to achieve with synthetic data generation. Therefore, a more accurate evaluation would only be possible with data generated by lab members based on their actual search patterns. Additionally, filters could be applied to narrow the search space and further improve performance.

The study also involved an experimental fine-tuning attempt on Nougat using a small custom training set of ten research papers; however, the results did not yield any significant improvement in performance or output quality.

The limitations of the hallucination detection system are due to the way large language models inherently process and store information. Potential solutions include an additional numerical features layer, fine-tuning, small chunk sizes, simpler queries, and stringent processing. However, this represents a growing area of research, and we acknowledge this limitation while proceeding with the current framework

5. Statement of Contributions

Atyab Hakeem designed the evaluation methodology, conducted comprehensive surveys and benchmarking of embedding models, chunking strategies, and hallucination detection models, and developed the synthetic dataset generation pipeline for evaluation. He defined three trajectories of query-based hallucination detection, analyzed the experimental results, and proposed the three-tiered reporting strategy.

Kishan worked on using language models to extract data and tried training them to make the extraction more accurate and reliable.

Kushal experimented with various retrieval strategies and evaluated open-source vector databases to identify one that met our project requirements. He also contributed to developing a data pipeline that handled data loading, embedding generation, and storage in the selected vector database.

Pranav built the data preprocessing pipeline, extracting and standardizing text from PDFs, HTML, and DOCX files, and linking them with cleaned metadata parsed from .ris files using rispy and RapidFuzz for matching. The final structured dataset enabled seamless downstream retrieval and analysis.

References

1. Lewis, P., et al. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks.* NeurIPS.
2. Thakur, N., et al. (2021). *BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models.* NeurIPS.
3. Gao, L., et al. (2022). *Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline.* ACL.
4. RAGAS (2023). *Faithfulness evaluation for RAG systems.* https://github.com/explodinggradients/ragas
5. NEU Magi Cluster Documentation.
6. Wang, L. L., Lo, K., Chandrasekhar, Y., et al. (2020). CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706*.

7. Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *Proceedings of ACL 2020*.
8. Roberts, K., et al. (2020). TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9), 1431-1436.
9. Voorhees, E., et al. (2020). TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).
10. Meta AI. (2024). Llama 3.3: Multilingual large language model (Version 70B) [Computer software]. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3
11. NVIDIA. (2024). Llama 3.3 70B Instruct FP8 [Quantized model]. Hugging Face. https://huggingface.co/nvidia/Llama-3.3-70B-Instruct-FP8
12. Smith, B., & Troynikov, A. (2024). Evaluating chunking strategies for retrieval (Technical Report). Chroma. https://research.trychroma.com/evaluating-chunking
13. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3615-3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371
14. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2270-2282). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.207
15. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.
16. Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., & Rehm, G. (2022). Neighborhood contrastive learning for scientific document representations with citation embeddings. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 11670-11688). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.802
17. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv preprint arXiv:2002.10957
18. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In Proceedings of NeurIPS 2020
19. Subburaj, A., Dua, K., Shah, V., Baumann, P., Poon, H., & Xu, F. (2025). Finding the best chunking strategy for accurate AI responses. NVIDIA Technical Blog. https://developer.nvidia.com/blog/finding-the-best-chunking-strategy-for-accurate-ai-responses/
20. Tang, L., Laban, P., & Durrett, G. (2024). MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8818-8847). Association for Computational Linguistics.
21. Bao, F., Li, M., Luo, R., & Mendelevitch, O. (2024). HHEM-2.1-Open: Hughes Hallucination Evaluation Model. Hugging Face. https://doi.org/10.57967/hf/3240

Appendix

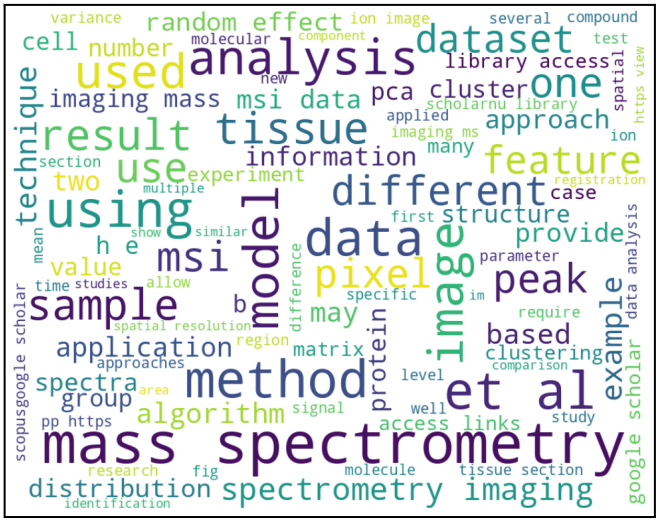

Figure -1: World Cloud Visualization

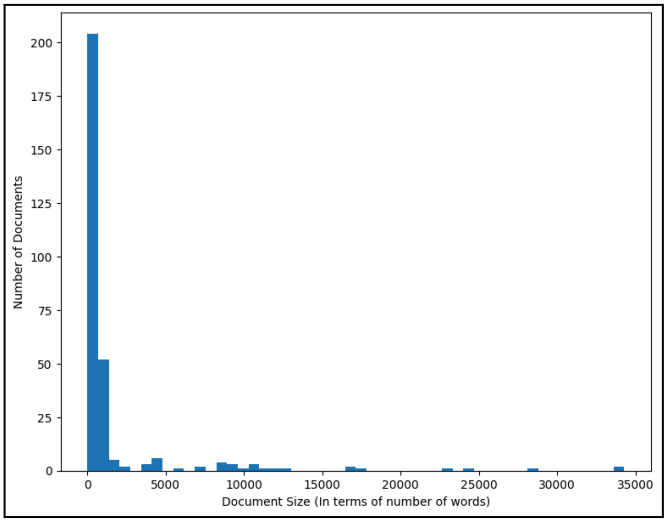

Figure -2: Document Length Distribution

Number of Documents	Average Number of Characters	Average Number of Words	Vocabulary Size	Average Number of Tables	Average Number of Figures
297	9,543	1,782	39,144	1	7

Table -1: Text Statistics

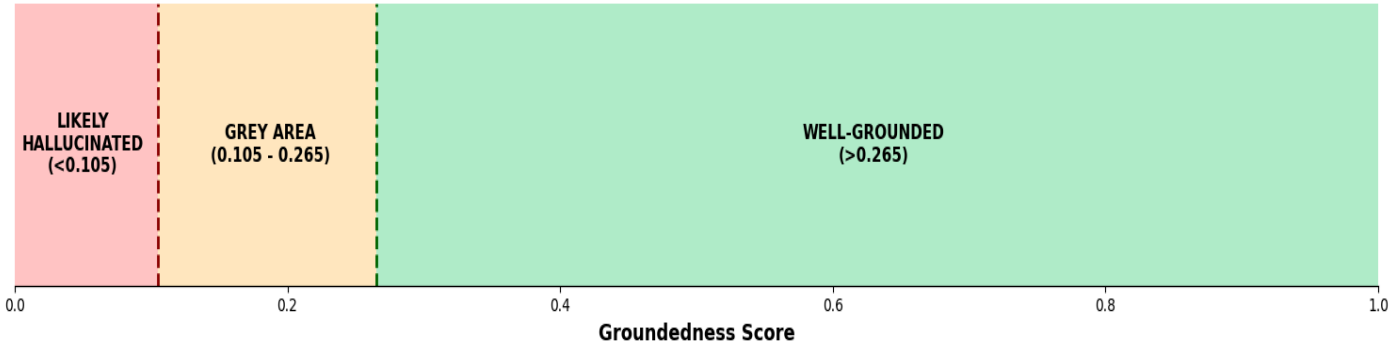

Figure 3: Hallucination detection and reporting decision framework

Question	What is the correlation between any two adjacent time points in the first case of the repeated measures ANOVA model?
Excerpt	In the first case of the repeated measures ANOVA model, we assumed ρ = 0…
Claim	0.7
Answer	0.7

Table 2: Example of false negatives (RoBERTa)

Question	What could be removed by aligning each spectrum to the mean spectrum and re-calibrating the m/z positions via the internal calibrants?
Excerpt	These m/z shifts could be removed by aligning each spectrum to the mean spectrum…
Claim	baseline shifts
Answer	m/z shifts

Table 3: Example of false positives (RoBERTa)

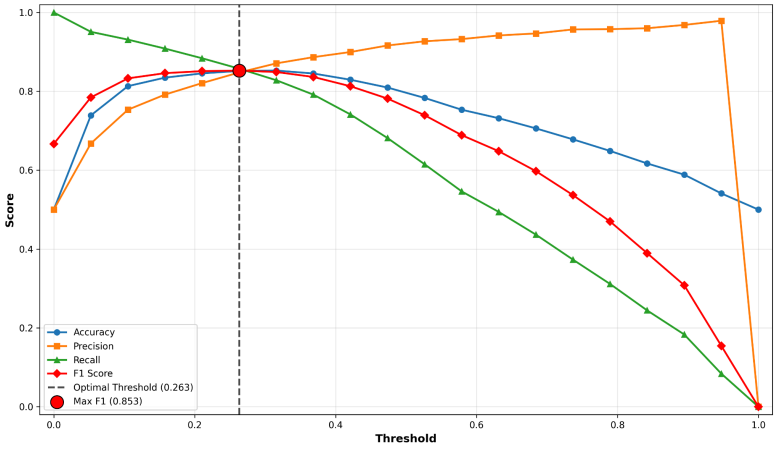

Figure 4: Confidence threshold gridsearch across accuracy, precision, recall, and f1 score on RoBERTa model