Agri-Query: A Case Study on RAG vs. Long-Context LLMs for Cross-Lingual Technical Question Answering

Anonymous ACL submission

Abstract

We present a case study evaluating large language models (LLMs) with 128K-token context windows on a technical question answering (QA) task. Our benchmark is built on a user manual for an agricultural machine, available in English, French, and German. It simulates a cross-lingual information retrieval scenario where questions are posed in English against all three language versions of the manual. The evaluation focuses on realistic "needle-in-a-haystack" challenges and includes unanswerable questions to test for hallucinations. We compare nine long-context LLMs using direct prompting against three Retrieval-Augmented Generation (RAG) strategies (keyword, semantic, hybrid), with an LLM-as-a-judge for evaluation. Our findings for this specific manual show that Hybrid RAG consistently outperforms direct long-context prompting. Models like Gemini 2.5 Flash and the smaller Qwen 2.5 7B achieve high accuracy (over 85%) across all languages with RAG. This paper contributes a detailed analysis of LLM performance in a specialized industrial domain and an open framework[1] for similar evaluations, highlighting practical trade-offs and challenges.

1 Introduction

Technical user manuals are essential for all equipment. Modern European agricultural machinery is sophisticated with mechatronics and also highly regulated by the European Union for safety. These extensive documents provide comprehensive guidelines covering mechanical, electronic, and agronomical aspects.

Because Europe has many language areas, manufacturers must translate and maintain these manuals in multiple languages. This real-world scenario provides a practical basis for benchmarking the cross-lingual QA capabilities of LLMs. In this paper, we benchmark several state-of-the-art models to assess their QA robustness. For this benchmark, we curated a QA set based on domain expertise, focusing on critical operational and safety information, and including unanswerable questions to test the LLM's ability to avoid hallucinations.

The questions present a needle-in-a-haystack challenge where the answer is typically found in a single location within the user manual. This paper presents a case study comparing RAG approaches against a direct long-context method across various models and languages.

2 Related Work

The ability of LLMs to understand long documents is an active research area. While Retrieval-Augmented Generation (RAG) can outperform Long-Context (LC) models (Yu et al., 2024), the performance is inconsistent across different tasks and datasets (Wang et al., 2024; Li et al., 2025). Newer LC models with large context windows show strong performance without RAG. However, RAG systems are often more resource-efficient and cheaper to maintain than LC systems (Li et al., 2024).
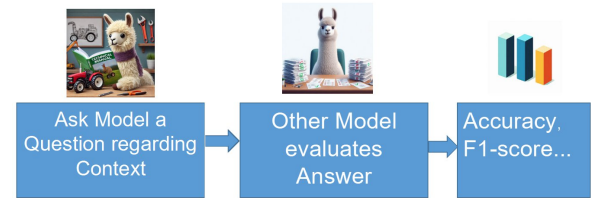

Figure 1: Process of asking questions (Illustrations created by the authors using Microsoft Bing Image Creator powered by DALL-E 3).

3 Materials

3.1 Models / LLM

We tested several openly available LLMs, detailed in Table 1, and the proprietary Gemini 2.5 Flash

---

[1] The code for the framework is available at https://anonymous.4open.science/r/Agri-Query/.

model. All models used a temperature of 0 for deterministic outputs. We used an LLM-as-a-judge framework for automatic evaluation, comparing model outputs to ground-truth answers. Gemma 3 evaluated RAG results, while Gemma 2 evaluated long-context results. We acknowledge that using different, unvalidated judge models is a limitation of this study.

Table 1: Overview of LLMs used in this study

Model	Size	Context	Quantiz.
Qwen 3	8B	128k	Q4_K_M
Qwen 2.5	7B	128k	Q4_K_M
DeepSeek-R1	1.5B	128k	Q4_K_M
DeepSeek-R1	8B	128k	Q4_K_M
Gemma 2 (eval.)	9B	8k	Q4_0
Gemma 3 (eval.)	12B	128k	Q4_K_M
Phi-3 Medium	14B	128k	Q4_0
Llama 3.1	8B	128k	Q4_K_M
Llama 3.2	1B	128k	Q8_0
Llama 3.2	3B	128k	Q4_K_M
Gemini 2.5 Flash	-	1M	-

3.2 User Manual

Our test case is the user manual for the Kverneland Exacta-TLX Geospread GS3, a mechatronic fertilizer spreader. We chose this manual because our familiarity with the machine aided in creating the QA set. We used the official English (KveEN), French (KveFR), and German (KveDE) versions. Each 165-page manual contains approximately 59k tokens and has an identical layout across languages, ensuring consistent page numbering for cross-lingual tests.

4 Methods

This work involved document preparation, QA dataset creation, long-context testing, and RAG system implementation. We converted the PDF manuals to Markdown format using the Docling library (Auer et al., 2024). We developed a small wrapper around the library to enable page-wise conversion. All experiments ran on a single NVIDIA RTX 6000 GPU, totaling approximately 80 GPU hours.

We created a QA test set of 108 questions from our domain expertise, focusing on critical operational and safety information. The dataset is balanced with 54 answerable and 54 unanswerable questions to test for hallucinations. To isolate cross-lingual retrieval capabilities, all questions were posed in English, following benchmarks like XTREME (Hu et al., 2020). Appendix A shows example questions.

Figure 1 illustrates our evaluation process. First, a relevant context is selected. Second, an LLM is prompted with a question about the context. Third, an evaluator LLM assesses the answer's correctness.

4.1 RAG system

We tested three RAG retrieval methods. For all methods, the document was split into chunks of 200 tokens with a 100-token overlap. We used a single embedding model, gte-Qwen2-7B-instruct from Li et al. (2023), for semantic and hybrid retrieval. This model was chosen for its strong performance on the Massive Multilingual Text Embedding Benchmark (MTEB) (Enevoldsen et al., 2025), making it well-suited for our cross-lingual tests. For each question, we retrieved the top three most relevant chunks. These RAG hyperparameters were fixed for all experiments; a sensitivity analysis is a subject for future work. We then provided these chunks and the question to the LLM. The prompts are available in Appendix B.

4.1.1 Keyword-based Retrieval

The BM25 algorithm (Trotman et al., 2014) was used for keyword-based retrieval. This method ranks documents based on the frequency of query terms within them, adjusted for document length and term rarity across the corpus. While efficient, it can fail if queries use synonyms not present in the text.

4.1.2 Semantic Retrieval

This method finds relevant chunks based on semantic meaning. Text is converted into numerical vectors (embeddings). The manual's chunks were vectorized and stored in a local ChromaDB vector database. During a query, the input question is also vectorized, and the database is searched for chunks with the highest cosine similarity to the query vector. Semantic retrieval may sometimes miss important query words. This method is more computationally intensive than keyword retrieval but can find relevant results even if phrasing differs from the document.

4.1.3 Hybrid Retrieval

Hybrid retrieval combines keyword-based (BM25) and semantic retrieval. This approach leverages the strengths of both methods. We perform keyword and semantic searches independently and then merge their ranked results using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to produce a final, more robust ranking. RRF computes a new

score for each retrieved chunk by summing the inverse of its rank from each retrieval list. This method effectively prioritizes chunks that consistently rank high across different search strategies, mitigating the weaknesses of any single method.

4.2 Long-Context Testing

We assessed direct long-context capabilities by providing LLMs with context and a question, without RAG or fine-tuning. For answerable questions, the context included the target page containing the answer. For unanswerable questions, a thematically related page was used. We simulated different context sizes (1k to 59k tokens) by adding surrounding pages as noise while preserving the original document order. This method tests the models' inherent understanding across various context lengths. Appendix B contains the prompts.

5 Results

We evaluated model performance using accuracy, F1 score, precision, recall, and specificity. In our evaluation, a **positive case** corresponds to an answerable question, and a **negative case** corresponds to an unanswerable question. This allows us to assess not only correctness but also the models' ability to avoid hallucination. Appendix C provides the formulas for these metrics.

5.1 Long-Context QA Performance

In our long-context tests without RAG, we provided the LLM with the relevant page plus surrounding pages as noise to reach specific token counts. Figure 2 shows the F1 scores. The Lost in the Middle effect (Liu et al., 2023) was pronounced for smaller models. Larger models like Gemini 2.5 Flash and Phi-3 14B performed better with the full manual context, though some performance degradation was still evident.
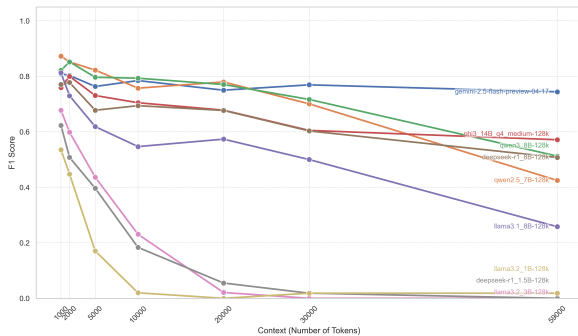
Figure 2: Long-Context QA: F1 score vs. noise: English.

5.2 RAG Performance

Next, RAG performance was tested with the same QA set and prompt structure. Figure 3 shows the F1 score. Hybrid retrieval consistently achieved the highest accuracy and F1 scores. Gemini 2.5 Flash had the highest accuracy, while naively inserting the full manual yielded significantly worse results. This indicates RAG, especially Hybrid retrieval, is better for needle-in-a-haystack tasks and suits technical documents like our agricultural manuals. Notably,smaller models like Llama 3.2 3B and Qwen 2.5 7B also achieved high performance (Accuracy > 0.85, Specificity = 0.815), demonstrating that RAG can enable effective results on resource-constrained hardware.

Table 2 shows more detailed results for Hybrid RAG. Precision and recall are very similar across models. However, specificity is much lower for smaller models. This indicates smaller models are more likely to produce false positives (i.e., hallucinate answers to unanswerable questions), while larger models are more likely to produce false negatives (i.e., fail to find an existing answer).
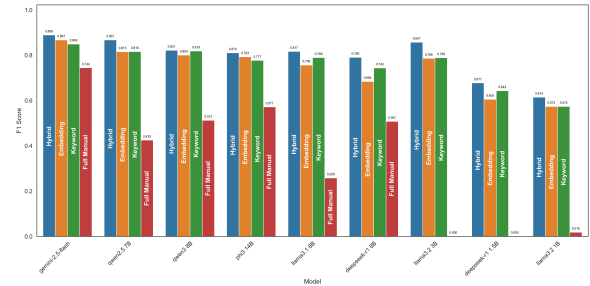
Figure 3: F1 comparison for English language across RAG retrieval and Full Manual (59k tokens).

Table 2: Performance on English Manual using Hybrid retrieval

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.880**	**0.889**	0.825	**0.963**	0.796
Qwen 2.5 7B	0.861	0.867	**0.831**	0.907	**0.815**
Qwen3 8B	0.815	0.821	0.793	0.852	0.778
Phi3 14B	0.796	0.810	0.758	0.870	0.722
Llama3.1 8B	0.796	0.817	0.742	0.907	0.685
Deepseek-R1 8B	0.759	0.790	0.700	0.907	0.611
Llama3.2 3B	0.852	0.857	0.828	0.889	0.815
Deepseek-R1 1.5B	0.630	0.677	0.600	0.778	0.481
Llama3.2 1B	0.500	0.614	0.500	0.796	0.204

5.3 Cross-lingual Performance using Hybrid RAG

Lastly, we evaluated the models' cross-lingual information retrieval capabilities. As described, this

setup involves posing questions in English against non-English documents (French and German) to assess the system's ability to bridge this language gap. Figure 4 shows accuracy, and Figure 5 shows the F1 score across English (EN), French (FR), and German (DE) for Hybrid RAG. For most models, performance on French or German was comparable to English, demonstrating that hybrid RAG with a strong multilingual embedding model offers robust cross-lingual retrieval.
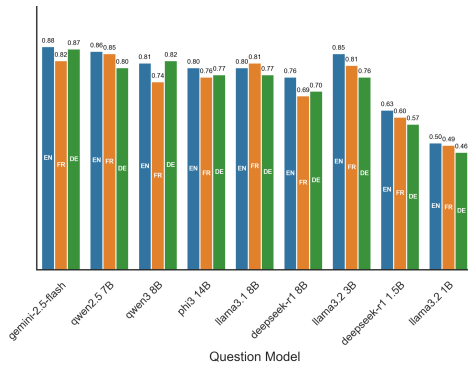
Figure 4: Accuracy comparison across different languages.
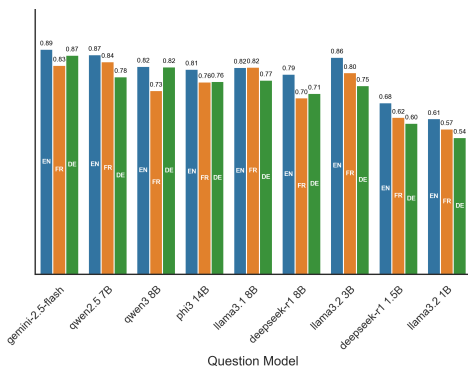
Figure 5: F1 score comparison across different languages.

6 Discussion

Our results offer a case study on applying LLMs to a real-world technical QA task, highlighting key points about RAG and long-context models in a specialized, multilingual domain.

6.1 RAG vs. Long-Context

For the tested agricultural manual, Hybrid RAG consistently outperformed the direct long-context approach. This was true even when comparing smaller models using RAG against larger models using the full context. The pronounced "Lost in the Middle" effect in our long-context tests (Figure 2) underscores the ongoing challenges large context models face in reliably locating specific facts within long, noisy inputs. Our results suggest that for applications requiring high-precision factual retrieval from dense technical documents, a well-configured RAG system remains a more robust choice.

Cross-Lingual Capabilities: The hybrid RAG approach demonstrated strong cross-lingual performance. High-performing models like Gemini 2.5 Flash and Qwen 2.5 7B maintained high accuracy when querying in English against French and German manuals. This indicates that the combination of a powerful multilingual embedding model and a capable LLM can effectively bridge language gaps for information retrieval tasks.

Failure Modes: A qualitative review revealed two primary failure modes: retrieval failure and hallucination. Retrieval failure occurs when the RAG system does not retrieve the correct context, a problem less frequent with the hybrid approach. Hallucination was more common, especially for unanswerable questions, with smaller models showing a higher tendency to invent answers (lower specificity in Table 2).

Potential Risks While LLMs can enhance access to information, they also pose risks of misinformation and over-reliance on AI-generated content. Users must exercise caution, especially with unanswerable questions, as models may generate plausible but incorrect answers.

Future Research: Future work should expand the benchmark to other technical domains to test generalizability. It could also explore more complex queries requiring information synthesis and test queries in the document's native language. A sensitivity analysis of RAG hyperparameters (e.g., chunk size, embedding model) is needed.

7 Conclusions

We present a framework for benchmarking LLMs on RAG and long-context tasks. Our findings show that for the agricultural manual tested, RAG is highly effective, enabling even small models to achieve strong results. The Lost in the Middle effect highlights that context window length is a critical factor for long-context models. Finally, our results demonstrate that robust cross-lingual performance is achievable with Hybrid RAG paired with capable LLMs.

Limitations

This study has several limitations.

Scope and Generalizability: The benchmark uses a single agricultural manual. Findings may not generalize to other domains or document types.

Dataset: The QA dataset is limited to 108 questions curated with domain expertise. A larger, more diverse dataset would provide greater statistical power. We did not perform statistical significance testing.

Evaluation Methodology: Our evaluation uses an LLM-as-a-judge framework not validated against human annotators, which may introduce bias. Using different judge models for RAG (Gemma 3) and long-context (Gemma 2) experiments is a confounding variable that complicates direct comparison.

Experimental Design:

- **RAG Hyperparameters:** The RAG configuration was fixed (chunk size: 200, overlap: 100, top-k: 3) and used a single embedding model. The reported superiority of Hybrid RAG may be configuration-specific, and a hyperparameter sweep could yield different results.

- **Cross-Lingual Task:** Questions were posed only in English. This setup tests cross-lingual information retrieval but does not fully represent a scenario where a native speaker would query the document in their own language.

- **Question Complexity:** The questions primarily target factual, localized information. The benchmark does not assess the models' ability to synthesize information across multiple sections, reason about complex procedures, or interpret tables and figures.

Reproducibility: Only a single test run was conducted for each experiment, so we do not report variance in the results.

Acknowledgements

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling technical report. *Preprint*, arXiv:2408.09869.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

KveDE. 2024. *Betriebsanleitung - Exacta-TLX-GEOSPREAD GS3*. Kverneland Group Nieuw-Vennep B.V., Hoofdweg 1278, NL-2153 LR Nieuw-Vennep, The Netherlands. Reference Number: A148818240-1; Language: German.

KveEN. 2021. *User Manual - Exacta-TLX-GEOSPREAD GS3*. Kverneland Group Nieuw-Vennep B.V., Hoofdweg 1278, NL-2153 LR Nieuw-Vennep, The Netherlands. Reference Number: A148703540-2; Language: English; Applicable from software version V: GS3 x.x., machine number 1001, serial number VN407.

KveFR. 2021. *Manuel d'utilisation - Exacta-TLX-GEOSPREAD GS3*. Kverneland Group Nieuw-Vennep B.V., Hoofdweg 1278, NL-2153 LR Nieuw-Vennep, The Netherlands. Reference Number: A148703640-2; Language: French.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025. Lara: Benchmarking retrieval-augmented generation and long-context llms – no silver bullet for lc or rag routing. *Preprint*, arXiv:2502.09977.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

5

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *Preprint*, arXiv:2407.16833.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Preprint*, arXiv:2406.17419.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024. In defense of rag in the era of long-context language models. *Preprint*, arXiv:2409.01666.

6

A Appendix: QA Test Dataset

A.1 Answerable question samples

- Q: What torque (in Nm) should be applied to the vane lock nuts? A: 50 Nm

- Q: What is the required grease level in mm below the filler opening for spreading disc gearboxes after the machine has stood still? A: 35 mm

- Q: Where is the main switch button to turn the control box on or off located? A: The main switch button is located on the upper left in the red extension.

- Q: How to enable fine application for dosing low application rate? A: Move the fine application handle to the fine dosing position on both sides.

- Q: Where is the RS 232 connector located? A: At the back of the control box.

- Q: How often should the agitator axle seal be replaced? A: Every season and after every 100 operational hours.

- Q: From which machine point is the spreading height measured to the ground or the crop? A: Measured from the bottom of the vanes.

- Q: What materials are required to perform the tray test? A: A measuring tape or ruler, a spirit level, 7 troughs, 7 graduated tubes, a funnel, a notebook, pen, calculator, this manual, and the software's instruction manual.

- Q: How many parts does the distribution meter have? A: Seven.

- Q: Should the parking brake of the tractor be engaged before connecting the machine? A: Yes.

- Q: How long should the main switch button be pressed to turn the control box on or off? A: At least 1 second.

- Q: What is the overlap percentage for the full field spreading pattern? A: 100% overlap.

- Q: What determines the machine's working width? A: Spreading disc RPM

- Q: When shortening a coupling shaft, how far must profiled tubes at least overlap in mm? A: 150mm.

A.2 Unanswerable question samples

- Q: How much extra diesel does the tractor consume to use the Exacta-TLX GEOSPREAD? A: Not found in context

- Q: Is one-sided boundary spreading suitable for small gardens? A: Not found in context

- Q: Can IsoMatch Tellus be connected to an external mouse? A: Not found in context

- Q: Can the linkage pin for the tractor be made out of aluminium? A: Not found in context

- Q: What is the maximum height the fertilizer flies when spreading without GEOCONTROL headland? A: Not found in context

- Q: What happens if the machine grease nipples are never lubricated? A: Not found in context

- Q: What kind of protective safety gloves are needed for cleaning fertiliser remnants from the Exacta-TLX GEOSPREAD before welding? A: Not found in context

- Q: What specific 'grease' type is recommended for 'profiled tubes' of the coupling shaft? A: Not found in context

- Q: What is the minimum 'baud rate' for 'RS 232 connection'? A: Not found in context

- Q: What is the recommended tire pressure 'range' for transport mode? A: Not found in context

- Q: What is the drain rate in liters per minute of 'drain kit' for hopper emptying? A: Not found in context

- Q: How many hours of continuous operation can the IsoMatch Tellus operate before flattening a typical tractor battery if left switched on with the engine off? A: Not found in context

- Q: What specific paint should be used to paint any damaged paintwork at the end of the season the machine in preparation for winter storage? A: Not found in context

- Q: Can the IsoMatch universal ISOBUS terminal be used to check the weather? A: Not found in context

B Appendix: Prompts

B.1 Question prompts

```
<purpose> Extract a precise, concise answer to the question from the given context. Adhere strictly to the instructions.
Base your answer on the context. </purpose>
    <instructions>
        <instruction> Read the entire context carefully </instruction>
        <instruction> Focus ONLY on the specific information related to the question </instruction>
        <instruction> Provide an extremely precise answer </instruction>
        <instruction> Match the expected answer format exactly </instruction>
        <instruction> If unsure, respond with "Unknown" or "Not found in context" </instruction>
        <instruction> Answer in English </instruction>
    </instructions>

    <context>
      {context}
    </context>
    <question>
      {question}
    </question>
    <answer>
      [Carefully extract the EXACT information that directly answers the question, keeping it as brief and precise as
possible]
    </answer>
```

B.2 Evaluation prompt

```
<purpose> ANSWER COMPARISON TASK. Do ANSWER_ONE and ANSWER_TWO convey the same informa-
tion regarding the QUESTION? Adhere strictly to the INSTRUCTIONS. Base your ANSWER on the CONTEXT.
</purpose>
    <INSTRUCTIONS>
        <instruction> - Respond 'yes' if ANSWER_ONE and the ANSWER_TWO convey the SAME TECHNICAL
MEANING </instruction>
        <instruction> - Consider 'yes' if differences are INSIGNIFICANT to the core technical content
</instruction>
        <instruction> - Respond 'no' ONLY if there are MEANINGFUL differences that alter the technical understand-
ing </instruction>
        <instruction> - Assess the SUBSTANCE of the information, not surface-level variations </instruction>
        <instruction> - Answer ONLY with yes or no </instruction>
        <instruction> - Don't provide additional information </instruction>
    </INSTRUCTIONS>

    <CONTEXT>
        <QUESTION>
      {question}
    </QUESTION>
        <ANSWER_ONE>
      {model_answer}
    </ANSWER_ONE>
        <ANSWER_TWO>
      {expected_answer}
    </ANSWER_TWO>
    </CONTEXT>

    <ANSWER>
      (yes/no)
    </ANSWER>
```

C   Appendix: Detailed Results

We used the following metrics to evaluate the performance of the models. For a given question, the outcome is classified into one of four categories based on whether the question is answerable and whether the model's response is correct. A positive case is an answerable question, and a negative case is an unanswerable question.

- **TP (True Positive):** The model correctly answers an answerable question.

- **TN (True Negative):** The model correctly identifies an unanswerable question (e.g., by responding "Not found in context").

- **FP (False Positive):** The model provides an incorrect answer to an unanswerable question (hallucination).

- **FN (False Negative):** The model fails to answer an answerable question correctly.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$F_1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
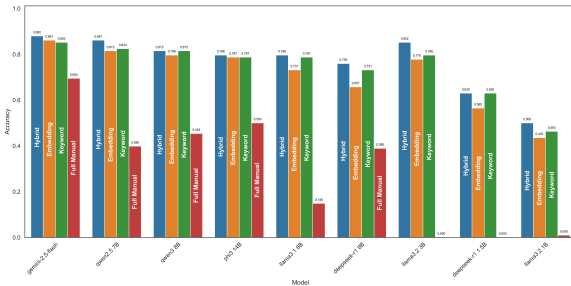
Figure 6: Accuracy comparison for English language across RAG retrieval and Full Manual (59k tokens).

C.1   Keyword RAG Performance

This section includes tables detailing the performance metrics for models utilizing the Keyword RAG retrieval algorithm across various languages.

Table 3: Performance of English Keyword Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.852**	**0.849**	**0.865**	0.833	**0.870**
Qwen 2.5 7B	0.824	0.816	0.857	0.778	**0.870**
Qwen3 8B	0.815	0.818	0.804	**0.833**	0.796
Phi3 14B	0.787	0.777	0.816	0.741	0.833
Llama3.1 8B	0.787	0.789	0.782	0.796	0.778
Deepseek-R1 8B	0.731	0.743	0.712	0.778	0.685
Llama3.2 3B	0.796	0.788	0.820	0.759	0.833
Deepseek-R1 1.5B	0.630	0.643	0.621	0.667	0.593
Llama3.2 1B	0.463	0.574	0.476	0.722	0.204

Table 4: Performance of French Keyword Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.583**	0.328	0.846	0.204	0.963
Qwen 2.5 7B	0.556	0.273	0.750	0.167	0.944
Qwen3 8B	0.556	0.294	0.714	0.185	0.926
Phi3 14B	0.546	0.310	0.647	0.204	0.889
Llama3.1 8B	**0.583**	**0.348**	0.800	**0.222**	0.944
Deepseek-R1 8B	0.491	0.267	0.476	0.185	0.796
Llama3.2 3B	0.565	0.230	**1.000**	0.130	**1.000**
Deepseek-R1 1.5B	0.389	0.233	0.312	0.185	0.593
Llama3.2 1B	0.213	0.206	0.208	0.204	0.222

Table 5: Performance of German Keyword Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	0.528	0.190	0.667	0.111	0.944
Qwen 2.5 7B	**0.537**	0.138	**1.000**	0.074	**1.000**
Qwen3 8B	**0.537**	0.194	0.750	0.111	0.963
Phi3 14B	0.528	0.164	0.714	0.093	0.963
Llama3.1 8B	0.519	0.161	0.625	0.093	0.944
Deepseek-R1 8B	0.472	0.174	0.400	0.111	0.833
Llama3.2 3B	0.509	0.102	0.600	0.056	0.963
Deepseek-R1 1.5B	0.306	0.096	0.138	0.074	0.537
Llama3.2 1B	0.398	**0.198**	0.296	**0.148**	0.648

9

C.2 Embedding RAG Performance

This section includes tables detailing the performance metrics for models utilizing the Embedding RAG retrieval algorithm across various languages.

Table 6: Performance of English Embedding Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.861**	**0.867**	**0.831**	**0.907**	**0.815**
Qwen 2.5 7B	0.815	0.815	0.815	0.815	**0.815**
Qwen3 8B	0.796	0.800	0.786	0.815	0.778
Phi3 14B	0.787	0.793	0.772	0.815	0.759
Llama3.1 8B	0.731	0.756	0.692	0.833	0.630
Deepseek-R1 8B	0.657	0.684	0.635	0.741	0.574
Llama3.2 3B	0.778	0.786	0.759	0.815	0.741
Deepseek-R1 1.5B	0.565	0.605	0.554	0.667	0.463
Llama3.2 1B	0.435	0.573	0.461	0.759	0.111

Table 7: Performance of French Embedding Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.806**	**0.784**	**0.884**	**0.704**	**0.907**
Qwen 2.5 7B	0.741	0.714	0.795	0.648	0.833
Qwen3 8B	0.657	0.626	0.689	0.574	0.741
Phi3 14B	0.648	0.642	0.654	0.630	0.667
Llama3.1 8B	0.704	0.704	0.704	**0.704**	0.704
Deepseek-R1 8B	0.537	0.545	0.536	0.556	0.519
Llama3.2 3B	0.648	0.548	0.767	0.426	0.870
Deepseek-R1 1.5B	0.380	0.385	0.382	0.389	0.370
Llama3.2 1B	0.324	0.425	0.370	0.500	0.148

Table 8: Performance of German Embedding Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.824**	**0.822**	**0.830**	**0.815**	**0.833**
Qwen 2.5 7B	0.759	0.740	0.804	0.685	**0.833**
Qwen3 8B	0.694	0.692	0.698	0.685	0.704
Phi3 14B	0.722	0.732	0.707	0.759	0.685
Llama3.1 8B	0.722	0.732	0.707	0.759	0.685
Deepseek-R1 8B	0.537	0.583	0.530	0.648	0.426
Llama3.2 3B	0.713	0.674	0.780	0.593	**0.833**
Deepseek-R1 1.5B	0.389	0.431	0.403	0.463	0.315
Llama3.2 1B	0.565	0.561	0.566	0.556	0.574

C.3 Hybrid RAG Performance

This section includes tables detailing the performance metrics for models utilizing the Hybrid RAG retrieval algorithm across various languages.

Table 9: Performance of English Hybrid Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.880**	**0.889**	0.825	**0.963**	0.796
Qwen 2.5 7B	0.861	0.867	**0.831**	0.907	**0.815**
Qwen3 8B	0.815	0.821	0.793	0.852	0.778
Phi3 14B	0.796	0.810	0.758	0.870	0.722
Llama3.1 8B	0.796	0.817	0.742	0.907	0.685
Deepseek-R1 8B	0.759	0.790	0.700	0.907	0.611
Llama3.2 3B	0.852	0.857	0.828	0.889	**0.815**
Deepseek-R1 1.5B	0.630	0.677	0.600	0.778	0.481
Llama3.2 1B	0.500	0.614	0.500	0.796	0.204

Table 10: Performance of French Hybrid Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	0.824	0.826	0.818	**0.833**	0.815
Qwen 2.5 7B	**0.852**	**0.840**	**0.913**	0.778	**0.926**
Qwen3 8B	0.741	0.725	0.771	0.685	0.796
Phi3 14B	0.759	0.759	0.759	0.759	0.759
Llama3.1 8B	0.815	0.818	0.804	**0.833**	0.796
Deepseek-R1 8B	0.685	0.696	0.672	0.722	0.648
Llama3.2 3B	0.806	0.796	0.837	0.759	0.852
Deepseek-R1 1.5B	0.602	0.619	0.593	0.648	0.556
Llama3.2 1B	0.491	0.574	0.493	0.685	0.296

Table 11: Performance of German Hybrid Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.870**	**0.865**	**0.900**	**0.833**	**0.907**
Qwen 2.5 7B	0.796	0.780	0.848	0.722	0.870
Qwen3 8B	0.824	0.819	0.843	0.796	0.852
Phi3 14B	0.769	0.762	0.784	0.741	0.796
Llama3.1 8B	0.769	0.766	0.774	0.759	0.778
Deepseek-R1 8B	0.704	0.714	0.690	0.741	0.667
Llama3.2 3B	0.759	0.745	0.792	0.704	0.815
Deepseek-R1 1.5B	0.574	0.596	0.567	0.630	0.519
Llama3.2 1B	0.463	0.540	0.472	0.630	0.296

C.4 Full Manual Performance (Long-Context @ approx. 59k Tokens)

This section presents performance metrics for models under the "Full Manual" configuration, corresponding to Long-Context evaluations with a context of approximately 59,000 tokens (entire document).

Table 12: Performance of English Full Manual Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.694**	**0.744**	**0.640**	**0.889**	**0.500**
Qwen 2.5 7B	0.398	0.425	0.407	0.444	0.352
Qwen3 8B	0.454	0.512	0.463	0.574	0.333
Phi3 14B	0.500	0.571	0.500	0.667	0.333
Llama3.1 8B	0.148	0.258	0.229	0.296	0.000
Deepseek-R1 8B	0.389	0.507	0.425	0.630	0.148
Llama3.2 3B	0.000	0.000	0.000	0.000	0.000
Deepseek-R1 1.5B	0.000	0.000	0.000	0.000	0.000
Llama3.2 1B	0.009	0.018	0.018	0.019	0.000

Table 13: Performance of French Full Manual Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.704**	**0.754**	0.645	**0.907**	0.500
Qwen 2.5 7B	0.620	0.549	**0.676**	0.463	**0.778**
Qwen3 8B	0.398	0.414	0.404	0.426	0.370
Phi3 14B	0.537	0.528	0.538	0.519	0.556
Llama3.1 8B	0.148	0.193	0.183	0.204	0.093
Deepseek-R1 8B	0.259	0.310	0.290	0.333	0.185
Llama3.2 3B	0.019	0.036	0.036	0.037	0.000
Deepseek-R1 1.5B	0.009	0.018	0.018	0.019	0.000
Llama3.2 1B	0.000	0.000	0.000	0.000	0.000

Table 14: Performance of German Full Manual Models

Metric LLM	Acc.	F1	Prec.	Rec.	Spec.
Gemini 2.5 Flash	**0.685**	**0.738**	**0.632**	**0.889**	0.481
Qwen 2.5 7B	0.546	0.380	0.600	0.278	**0.815**
Qwen3 8B	0.352	0.364	0.357	0.370	0.333
Phi3 14B	0.519	0.480	0.522	0.444	0.593
Llama3.1 8B	0.194	0.269	0.246	0.296	0.093
Deepseek-R1 8B	0.231	0.303	0.277	0.333	0.130
Llama3.2 3B	0.037	0.071	0.069	0.074	0.000
Deepseek-R1 1.5B	0.056	0.105	0.100	0.111	0.000
Llama3.2 1B	0.019	0.036	0.036	0.037	0.000

Performance Heatmaps: Multilingual Algorithm vs. Model

Accuracy Score (Multi-Language, Sorted by Algo): Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7b-128k	qwen3_8b-128k	phi3_14b_q4_medium-128k	llama3.1_8b-128k	deepseek-r1_8b-128k	llama3.2_3b-128k	deepseek-r1_1.5b-128k	llama3.2_1b-128k
english_hybrid	0.880	0.861	0.815	0.796	0.796	0.759	0.852	0.630	0.500
german_hybrid	0.870	0.796	0.824	0.769	0.769	0.704	0.759	0.574	0.463
french_hybrid	0.824	0.852	0.741	0.759	0.815	0.685	0.806	0.602	0.491
english_embedding	0.861	0.815	0.796	0.787	0.731	0.657	0.778	0.565	0.435
german_embedding	0.824	0.759	0.694	0.722	0.722	0.537	0.537	0.389	0.565
french_embedding	0.806	0.741	0.657	0.648	0.704	0.537	0.648	0.380	0.324
english_keyword	0.852	0.824	0.815	0.787	0.787	0.731	0.796	0.630	0.463
german_keyword	0.528	0.537	0.537	0.528	0.519	0.472	0.509	0.306	0.398
french_keyword	0.583	0.556	0.556	0.546	0.583	0.491	0.565	0.389	0.213
english_in-Context_1000	0.769	0.870	0.815	0.759	0.778	0.778	0.741	0.639	0.574
german_in-Context_1000	0.722	0.833	0.787	0.741	0.722	0.620	0.778	0.435	0.287
french_in-Context_1000	0.704	0.833	0.713	0.704	0.750	0.630	0.769	0.398	0.333
english_in-Context_10000	0.741	0.769	0.778	0.713	0.417	0.657	0.259	0.176	0.083
german_in-Context_10000	0.704	0.667	0.694	0.676	0.435	0.444	0.093	0.056	0.213
french_in-Context_10000	0.685	0.694	0.704	0.685	0.407	0.491	0.046	0.056	0.167
english_in-Context_30000	0.722	0.676	0.685	0.565	0.389	0.537	0.009	0.009	0.009
german_in-Context_30000	0.694	0.611	0.648	0.593	0.306	0.426	0.009	0.019	0.009
french_in-Context_30000	0.694	0.667	0.620	0.657	0.287	0.361	0.037	0.009	0.000
english_in-Context_59000	0.694	0.398	0.454	0.500	0.148	0.389	0.000	0.000	0.009
german_in-Context_59000	0.685	0.546	0.352	0.519	0.194	0.231	0.037	0.056	0.019
french_in-Context_59000	0.704	0.620	0.398	0.537	0.148	0.259	0.019	0.009	0.000

Figure 7: Accuracy score heatmap comparing multilingual algorithms and models.

F1 Score (Multi-Language, Sorted by Algo): Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7b-128k	qwen3_8b-128k	phi3_14b_q4_medium-128k	llama3.1_8b-128k	deepseek-r1_8b-128k	llama3.2_3b-128k	deepseek-r1_1.5b-128k	llama3.2_1b-128k
english_hybrid	0.889	0.867	0.821	0.810	0.817	0.790	0.857	0.677	0.614
german_hybrid	0.865	0.780	0.819	0.762	0.766	0.714	0.745	0.596	0.540
french_hybrid	0.826	0.840	0.725	0.759	0.818	0.696	0.796	0.619	0.574
english_embedding	0.867	0.815	0.800	0.793	0.756	0.684	0.786	0.605	0.573
german_embedding	0.822	0.740	0.692	0.732	0.732	0.583	0.574	0.431	0.581
french_embedding	0.784	0.714	0.626	0.642	0.704	0.545	0.548	0.385	0.425
english_keyword	0.849	0.816	0.818	0.777	0.789	0.743	0.788	0.643	0.574
german_keyword	0.190	0.138	0.194	0.164	0.161	0.174	0.102	0.096	0.198
french_keyword	0.328	0.273	0.294	0.310	0.348	0.267	0.230	0.233	0.206
english_in-Context_1000	0.809	0.873	0.821	0.759	0.812	0.770	0.878	0.623	0.535
german_in-Context_1000	0.766	0.827	0.789	0.741	0.750	0.643	0.769	0.419	0.421
french_in-Context_1000	0.750	0.824	0.710	0.729	0.769	0.649	0.747	0.393	0.471
english_in-Context_10000	0.785	0.757	0.793	0.705	0.547	0.694	0.231	0.183	0.020
german_in-Context_10000	0.746	0.600	0.701	0.648	0.555	0.474	0.075	0.105	0.190
french_in-Context_10000	0.734	0.637	0.719	0.667	0.529	0.513	0.072	0.106	0.237
english_in-Context_30000	0.769	0.701	0.717	0.605	0.500	0.603	0.000	0.018	0.018
german_in-Context_30000	0.740	0.488	0.672	0.577	0.380	0.446	0.018	0.036	0.018
french_in-Context_30000	0.740	0.591	0.643	0.661	0.374	0.378	0.017	0.018	0.000
english_in-Context_59000	0.744	0.425	0.512	0.571	0.258	0.507	0.000	0.000	0.018
german_in-Context_59000	0.738	0.380	0.364	0.480	0.269	0.303	0.071	0.105	0.036
french_in-Context_59000	0.754	0.549	0.414	0.528	0.193	0.310	0.036	0.018	0.000

Figure 8: F1 score heatmap comparing multilingual algorithms and models.

General Questions Success Rate (General Questions, Sorted by Algo): Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7b-128k	qwen3_8b-128k	phi3_14b_q4_medium-128k	llama3.1_8b-128k	deepseek-r1_8b-128k	llama3.2_3b-128k	deepseek-r1_1.5b-128k	llama3.2_1b-128k
english_hybrid	0.963	0.907	0.852	0.870	0.907	0.907	0.889	0.778	0.796
german_hybrid	0.833	0.722	0.796	0.741	0.759	0.741	0.704	0.630	0.630
french_hybrid	0.833	0.778	0.685	0.759	0.833	0.722	0.759	0.648	0.685
english_embedding	0.907	0.815	0.815	0.815	0.833	0.741	0.815	0.667	0.759
german_embedding	0.815	0.685	0.685	0.759	0.759	0.648	0.593	0.463	0.556
french_embedding	0.704	0.648	0.574	0.830	0.704	0.556	0.426	0.389	0.500
english_keyword	0.833	0.778	0.833	0.741	0.796	0.778	0.759	0.667	0.722
german_keyword	0.111	0.074	0.111	0.093	0.093	0.111	0.056	0.074	0.148
french_keyword	0.204	0.167	0.185	0.204	0.222	0.185	0.130	0.185	0.204
english_zeroshot_noise_1000	0.981	0.889	0.852	0.759	0.963	0.870	0.759	0.704	0.704
german_zeroshot_noise_1000	0.907	0.796	0.796	0.741	0.833	0.685	0.741	0.407	0.519
french_zeroshot_noise_1000	0.889	0.778	0.704	0.796	0.833	0.685	0.685	0.389	0.593
english_zeroshot_noise_10000	0.944	0.722	0.852	0.685	0.704	0.778	0.222	0.185	0.019
german_zeroshot_noise_10000	0.870	0.500	0.722	0.593	0.704	0.500	0.074	0.056	0.185
french_zeroshot_noise_10000	0.870	0.537	0.759	0.630	0.667	0.537	0.074	0.111	0.259
english_zeroshot_noise_30000	0.926	0.759	0.796	0.667	0.611	0.704	0.000	0.019	0.019
german_zeroshot_noise_30000	0.870	0.370	0.722	0.556	0.426	0.463	0.019	0.037	0.019
french_zeroshot_noise_30000	0.870	0.481	0.685	0.667	0.426	0.389	0.074	0.019	0.000
english_zeroshot_noise_59000	0.889	0.444	0.574	0.667	0.296	0.630	0.000	0.000	0.019
german_zeroshot_noise_59000	0.889	0.278	0.370	0.444	0.296	0.333	0.074	0.111	0.037
french_zeroshot_noise_59000	0.907	0.463	0.426	0.519	0.204	0.333	0.037	0.019	0.000

Figure 9: Success heatmap for general questions: multilingual algorithms vs. models.

Precision Score (Multi-Language, Sorted by Algo): Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7b-128k	qwen3_8b-128k	phi3_14b_q4_medium-128k	llama3.1_8b-128k	deepseek-r1_8b-128k	llama3.2_3b-128k	deepseek-r1_1.5b-128k	llama3.2_1b-128k
english_hybrid	0.825	0.831	0.793	0.758	0.742	0.700	0.828	0.600	0.500
german_hybrid	0.900	0.848	0.843	0.784	0.774	0.690	0.792	0.567	0.472
french_hybrid	0.818	0.913	0.771	0.759	0.804	0.672	0.837	0.593	0.493
english_embedding	0.831	0.815	0.786	0.772	0.692	0.635	0.759	0.554	0.461
german_embedding	0.830	0.804	0.698	0.707	0.707	0.530	0.780	0.403	0.566
french_embedding	0.884	0.795	0.689	0.654	0.704	0.536	0.767	0.382	0.370
english_keyword	0.865	0.857	0.804	0.816	0.782	0.712	0.820	0.621	0.476
german_keyword	0.667	1.000	0.750	0.714	0.625	0.400	0.600	0.138	0.296
french_keyword	0.846	0.750	0.714	0.647	0.800	0.476	1.000	0.312	0.208
english_in-Context_1000	0.688	0.857	0.793	0.759	0.703	0.891	0.612	0.559	0.432
german_in-Context_1000	0.682	0.860	0.782	0.741	0.682	0.607	0.800	0.431	0.354
french_in-Context_1000	0.649	0.875	0.717	0.672	0.714	0.617	0.822	0.396	0.390
english_in-Context_10000	0.671	0.796	0.742	0.725	0.447	0.627	0.240	0.182	0.021
german_in-Context_10000	0.653	0.750	0.684	0.711	0.458	0.450	0.077	0.056	0.196
french_in-Context_10000	0.635	0.784	0.698	0.708	0.439	0.492	0.070	0.100	0.219
english_in-Context_30000	0.658	0.651	0.652	0.654	0.423	0.528	0.000	0.018	0.018
german_in-Context_30000	0.644	0.714	0.629	0.600	0.343	0.431	0.018	0.036	0.018
french_in-Context_30000	0.644	0.765	0.607	0.665	0.333	0.368	0.069	0.018	0.000
english_in-Context_59000	0.640	0.407	0.463	0.500	0.229	0.425	0.000	0.000	0.018
german_in-Context_59000	0.632	0.630	0.357	0.522	0.246	0.277	0.069	0.100	0.036
french_in-Context_59000	0.645	0.676	0.404	0.538	0.183	0.290	0.036	0.018	0.000

Figure 10: Precision score heatmap comparing multilingual algorithms and models.

Recall Score (Multi-Language, Sorted by Algo): Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7b-128k	qwen3_8b-128k	phi3_14b_q4_medium-128k	llama3.1_8b-128k	deepseek-r1_8b-128k	llama3.2_3b-128k	deepseek-r1_1.5b-128k	llama3.2_1b-128k
english_hybrid	0.963	0.907	0.852	0.870	0.907	0.907	0.889	0.778	0.796
german_hybrid	0.833	0.722	0.796	0.741	0.759	0.741	0.704	0.630	0.630
french_hybrid	0.833	0.778	0.685	0.759	0.833	0.722	0.759	0.648	0.685
english_embedding	0.907	0.815	0.815	0.815	0.833	0.741	0.815	0.667	0.759
german_embedding	0.815	0.685	0.685	0.759	0.759	0.648	0.593	0.463	0.556
french_embedding	0.704	0.648	0.574	0.830	0.704	0.556	0.426	0.389	0.500
english_keyword	0.833	0.778	0.833	0.741	0.796	0.778	0.759	0.667	0.722
german_keyword	0.111	0.074	0.111	0.093	0.093	0.111	0.056	0.074	0.148
french_keyword	0.204	0.167	0.185	0.204	0.222	0.185	0.130	0.185	0.204
english_in-Context_1000	0.981	0.889	0.852	0.759	0.963	0.870	0.759	0.704	0.704
german_in-Context_1000	0.907	0.796	0.796	0.741	0.833	0.685	0.741	0.407	0.519
french_in-Context_1000	0.889	0.778	0.704	0.796	0.833	0.685	0.685	0.389	0.593
english_in-Context_10000	0.944	0.722	0.852	0.685	0.704	0.778	0.222	0.185	0.019
german_in-Context_10000	0.870	0.500	0.722	0.593	0.704	0.500	0.074	0.056	0.185
french_in-Context_10000	0.870	0.537	0.759	0.630	0.667	0.537	0.074	0.111	0.259
english_in-Context_30000	0.926	0.759	0.796	0.667	0.611	0.704	0.000	0.019	0.019
german_in-Context_30000	0.870	0.370	0.722	0.556	0.426	0.463	0.019	0.037	0.019
french_in-Context_30000	0.870	0.481	0.685	0.667	0.426	0.389	0.074	0.019	0.000
english_in-Context_59000	0.889	0.444	0.574	0.667	0.296	0.630	0.000	0.000	0.019
german_in-Context_59000	0.889	0.278	0.370	0.444	0.296	0.333	0.074	0.111	0.037
french_in-Context_59000	0.907	0.463	0.426	0.519	0.204	0.333	0.037	0.019	0.000

Figure 11: Recall score heatmap comparing multilingual algorithms and models.

Specificity Score (Multi-Language, Sorted by Algo):
Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7B-128k	qwen3_8B-128k	phi3_14B_q4_medium-128k	llama3.1_8B-128k	deepseek-r1_8B-128k	llama3_3B-128k	deepseek-r1_1.5B-128k	llama3.2_1B-128k
english_hybrid	0.796	0.815	0.778	0.722	0.685	0.611	0.815	0.481	0.204
german_hybrid	0.907	0.870	0.852	0.796	0.667	0.667	0.815	0.519	0.296
french_hybrid	0.815	0.926	0.796	0.759	0.796	0.648	0.852	0.556	0.296
english_embedding	0.815	0.815	0.778	0.759	0.630	0.574	0.741	0.463	0.111
german_embedding	0.833	0.833	0.704	0.685	0.685	0.426	0.833	0.315	0.574
french_embedding	0.907	0.833	0.741	0.667	0.704	0.519	0.870	0.593	0.204
english_keyword	0.870	0.870	0.796	0.833	0.778	0.685	0.833	0.593	0.648
german_keyword	0.944	1.000	0.963	0.963	0.944	0.833	0.963	0.537	0.648
french_keyword	0.963	0.944	0.926	0.889	0.944	0.796	1.000	0.593	0.222
english_in-Context_1000	0.556	0.852	0.778	0.759	0.593	0.611	0.519	0.444	0.074
german_in-Context_1000	0.537	0.870	0.778	0.741	0.611	0.556	0.815	0.463	0.056
french_in-Context_1000	0.519	0.889	0.722	0.611	0.667	0.574	0.852	0.407	0.074
english_in-Context_10000	0.537	0.815	0.704	0.741	0.130	0.537	0.296	0.167	0.148
german_in-Context_10000	0.537	0.833	0.667	0.759	0.167	0.389	0.111	0.056	0.241
french_in-Context_10000	0.500	0.852	0.648	0.741	0.148	0.444	0.019	0.000	0.074
english_in-Context_30000	0.519	0.593	0.574	0.463	0.167	0.370	0.000	0.000	0.000
german_in-Context_30000	0.519	0.852	0.574	0.630	0.185	0.389	0.000	0.000	0.000
french_in-Context_30000	0.519	0.852	0.556	0.648	0.148	0.333	0.000	0.000	0.000
english_in-Context_59000	0.500	0.352	0.333	0.333	0.000	0.148	0.000	0.000	0.000
german_in-Context_59000	0.481	0.815	0.333	0.593	0.093	0.130	0.000	0.000	0.000
french_in-Context_59000	0.500	0.778	0.370	0.556	0.093	0.185	0.000	0.000	0.000

Question Model

Figure 12: Specificity score heatmap comparing multilingual algorithms and models.

Unanswerable Questions Success Rate (Unanswerable Questions, Sorted by Algo):
Algorithm/Language vs LLM Model (RAG: C=200/O=100, ZeroShot: Noise [1000, 10000, 30000, 59000])

Plot Index	gemini-2.5-flash-preview-04-17	qwen2.5_7B-128k	qwen3_8B-128k	phi3_14B_q4_medium-128k	llama3.1_8B-128k	deepseek-r1_8B-128k	llama3_3B-128k	deepseek-r1_1.5B-128k	llama3.2_1B-128k
english_hybrid	0.796	0.815	0.778	0.722	0.685	0.611	0.815	0.481	0.204
german_hybrid	0.907	0.870	0.852	0.796	0.778	0.667	0.815	0.519	0.296
french_hybrid	0.815	0.926	0.796	0.796	0.796	0.648	0.852	0.556	0.296
english_embedding	0.815	0.815	0.778	0.759	0.630	0.574	0.741	0.463	0.111
german_embedding	0.833	0.833	0.704	0.685	0.685	0.426	0.833	0.315	0.574
french_embedding	0.907	0.833	0.704	0.667	0.704	0.519	0.870	0.370	0.148
english_keyword	0.870	0.870	0.796	0.833	0.778	0.685	0.833	0.593	0.648
german_keyword	0.944	1.000	0.963	0.963	0.944	0.833	0.963	0.537	0.648
french_keyword	0.963	0.944	0.926	0.889	0.944	0.796	1.000	0.593	0.222
english_zeroshot_noise_1000	0.556	0.852	0.778	0.759	0.593	0.811	0.519	0.444	0.074
german_zeroshot_noise_1000	0.537	0.870	0.778	0.741	0.611	0.556	0.815	0.463	0.056
french_zeroshot_noise_1000	0.519	0.889	0.722	0.611	0.667	0.574	0.852	0.407	0.074
english_zeroshot_noise_10000	0.537	0.815	0.704	0.741	0.130	0.537	0.296	0.167	0.148
german_zeroshot_noise_10000	0.537	0.833	0.667	0.759	0.167	0.389	0.111	0.056	0.241
french_zeroshot_noise_10000	0.500	0.852	0.648	0.741	0.148	0.444	0.019	0.000	0.074
english_zeroshot_noise_30000	0.519	0.593	0.574	0.463	0.167	0.370	0.000	0.000	0.000
german_zeroshot_noise_30000	0.519	0.852	0.574	0.630	0.185	0.389	0.000	0.000	0.000
french_zeroshot_noise_30000	0.519	0.852	0.556	0.648	0.148	0.333	0.000	0.000	0.000
english_zeroshot_noise_59000	0.500	0.352	0.333	0.333	0.000	0.148	0.000	0.000	0.000
1_zeroshot_noise_59000	0.481	0.815	0.333	0.593	0.093	0.130	0.000	0.000	0.000
french_zeroshot_noise_59000	0.500	0.778	0.370	0.556	0.093	0.185	0.000	0.000	0.000

Question Model

Figure 13: Success heatmap for unanswerable questions: multilingual algorithms vs. models.

Long-Context QA Performance vs. Noise

Figure 14: Long-Context QA accuracy vs. noise: English.

Figure 15: Long-Context QA F1 score vs. noise: English.

Figure 16: Long-Context QA precision vs. noise: English.
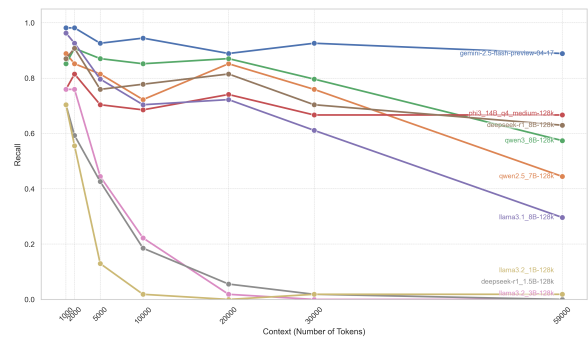
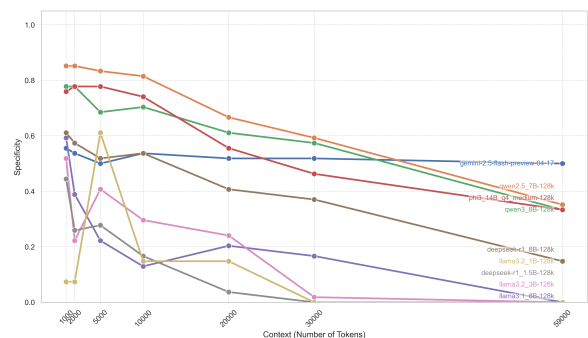

Figure 17: Long-Context QA recall vs. noise: English.

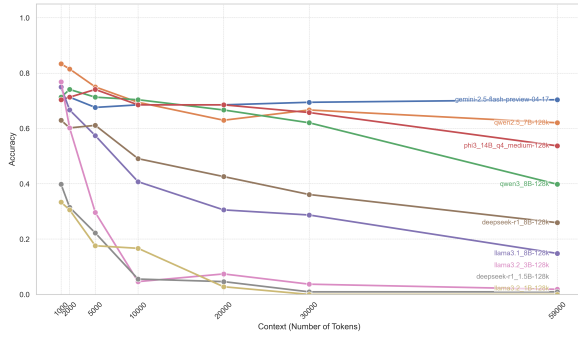Figure 18: Long-Context QA specificity vs. noise: English.

12

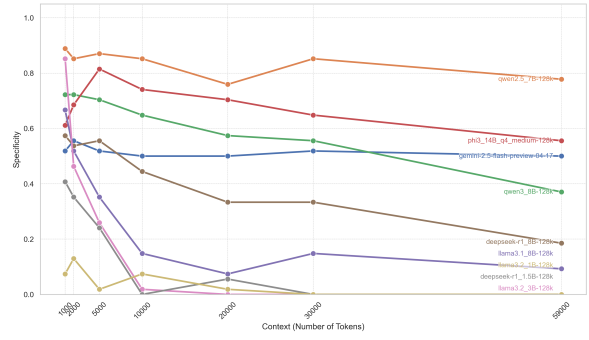Figure 19: Long-Context QA accuracy vs. noise: French.

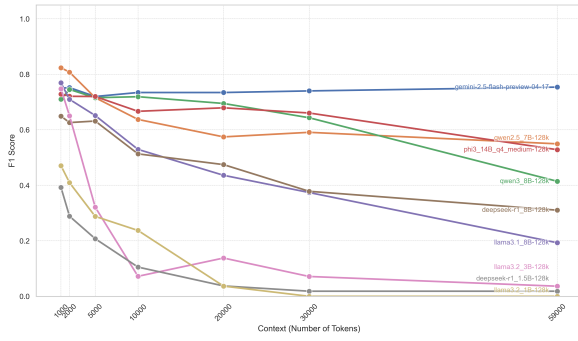Figure 20: Long-Context QA F1 score vs. noise: French.
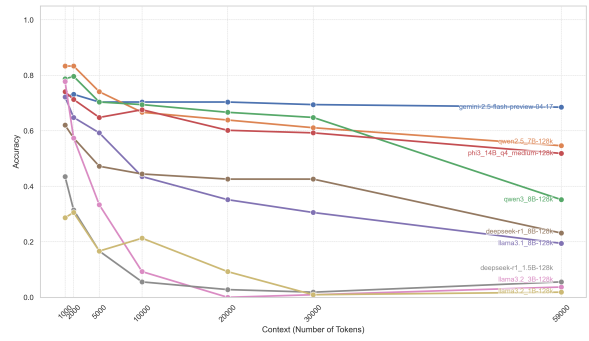

Figure 21: Long-Context QA precision vs. noise: French.

Figure 22: Long-Context QA recall vs. noise: French.

Figure 23: Long-Context QA specificity vs. noise: French.
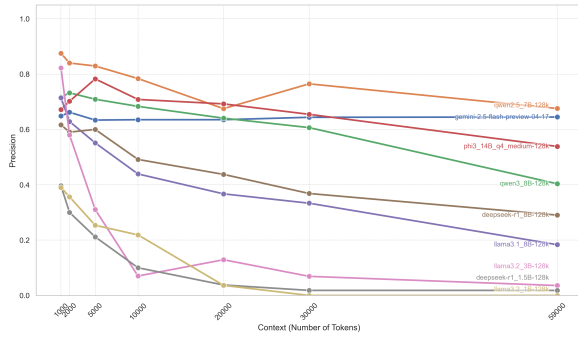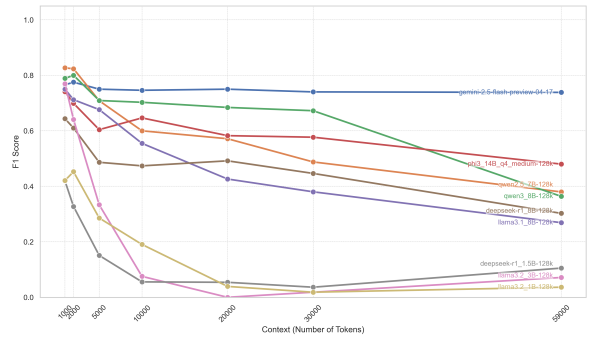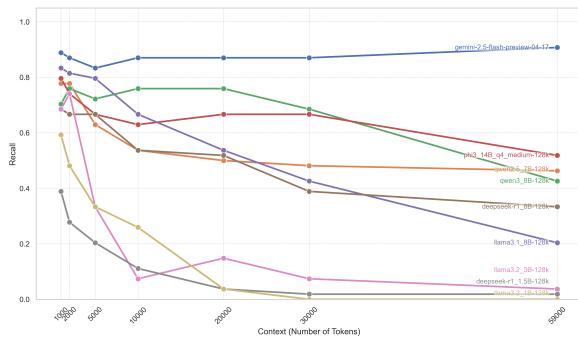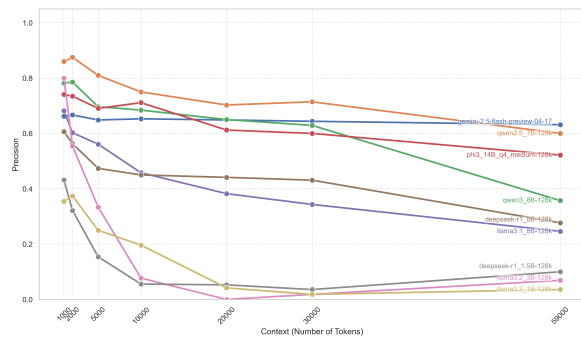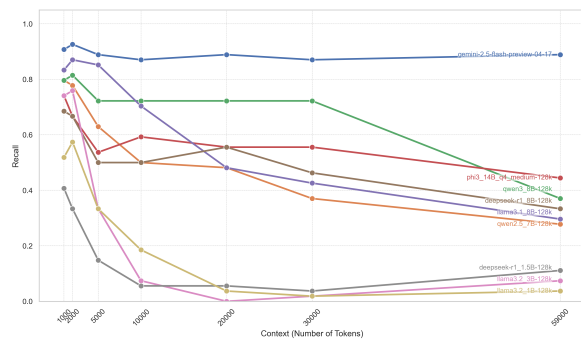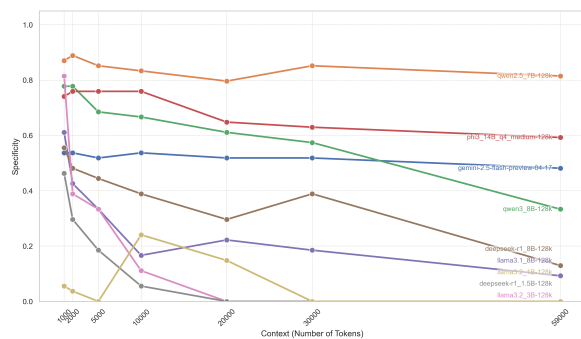

Figure 24: Long-Context QA accuracy vs. noise: German.

Figure 25: Long-Context QA F1 score vs. noise: German.

Figure 26: Long-Context QA precision vs. noise: German.

Figure 27: Long-Context QA recall vs. noise: German.

Figure 28: Long-Context QA specificity vs. noise: German.

Model Performance: Precision by Language and Retrieval Method
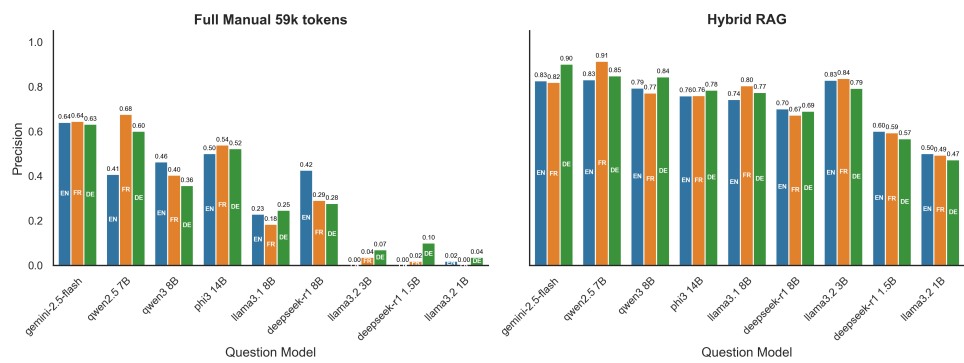
Figure 29: Precision comparison of different models.

Model Performance: Recall by Language and Retrieval Method
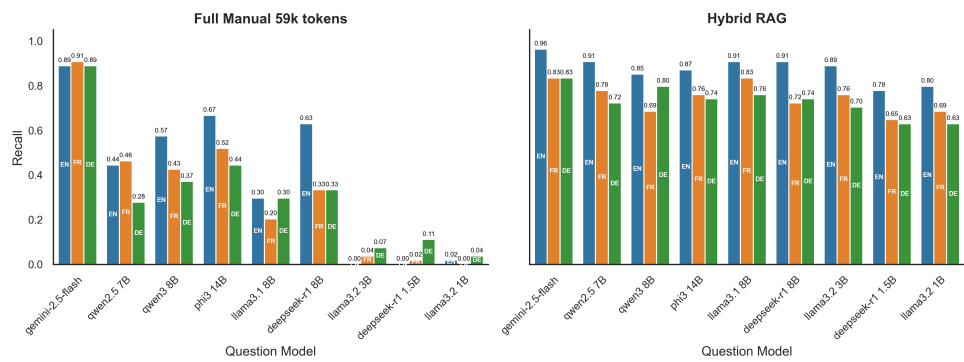
Figure 30: Recall comparison of different models.

Model Performance: Specificity by Language and Retrieval Method
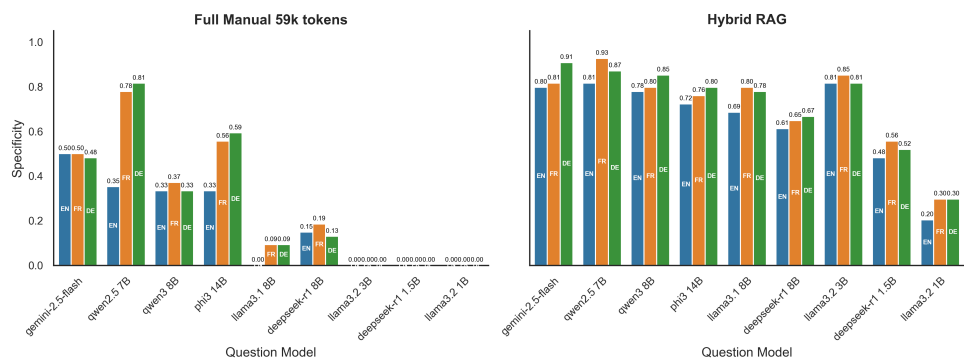
Figure 31: Specificity comparison of different models.

D Appendix: Framework Usage Guide

D.1 Overview

The benchmarking framework[2] is designed to evaluate Large Language Models (LLMs) on technical document understanding. It comprises two main projects: one for Long-Context (LC) testing, often referred to as "Zeroshot" testing in the codebase (located in the `ZeroShot/` directory), and another for Retrieval Augmented Generation (RAG) testing (located in the `RAG/` directory). Both projects share a common goal: to assess how well LLMs can answer questions based on a provided manual. This appendix provides instructions on data preparation and evaluation execution using both frameworks.

D.2 Manual Preparation

The framework primarily ingests manuals in plain text format, often with each page as a separate entry or segment. Manuals in PDF format must be converted to text. The ZeroShot project includes a utility script, `docling_page_wise_pdf_converter.py` (located in `ZeroShot/docling_page_wise_pdf_converter/`), for this purpose. Executing the main script in the ZeroShot project (`ZeroShot/main.py`) will automatically attempt to download and convert PDF manuals specified in its configuration, saving them as `.txt` files. This converter can also be used to prepare text files for the RAG framework.

D.3 Question Dataset Creation

Evaluation of LLMs on a new manual requires a corresponding question-answer dataset. This dataset must be a JSON file containing a list of question objects. Each object must include an `"id"`, `"question"`, `"expected_answer"`, and `"target_page"` (the page number in the manual where the answer can be found, or a relevant page for unanswerable questions). For unanswerable questions, the `"expected_answer"` should typically be `"Not found in context"` or a similar designated phrase.

Custom question datasets, for example `my_manual_questions.json`, are placed inside the `ZeroShot/question_datasets/` folder for the Long-Context framework. For the RAG framework, the dataset is placed in the `RAG/question_datasets/` folder. An example structure for a question entry is shown below:

```
// ZeroShot/question_datasets/my_manual_questions.json
// or RAG/question_datasets/my_manual_questions.json

[
  {
    "question": "How many dosing openings are closed during fine application?",
    "answer": "Two of three dosing openings are closed.",
    "page": 24
  },
  {
    "question": "Some question?",
    "answer": "Some answer",
    "page": 99
  }
]
```

D.4 Long-Context (Zeroshot) Testing Framework

The Long-Context testing framework, found in the `ZeroShot/` directory, evaluates an LLM's ability to answer questions when provided with the entire document or large sections of it. This method is also referred to as Zeroshot testing within the project because it tests the model's direct inference capabilities without retrieval augmentation specific to the query. Usage of this framework requires configuration of the `ZeroShot/config.json` file. This file is used to specify the LLM models, paths to the question datasets, the path or URL to the manual, and other parameters such as noise levels.

---

[2]The code for our framework is available at https://anonymous.4open.science/r/Agri-Query/.

The `main.py` script in the `ZeroShot/` directory is the entry point for running tests. It is executed from
the command line, specifying arguments such as the model, context type, and noise levels. Detailed
instructions and configuration options are available in the `ZeroShot/README.md` file. An example of
relevant parts to update in `ZeroShot/config.json` for a new manual and dataset:

```
// ZeroShot/config.json
{
  "llm_models": { /* ... define models ... */ },
  "evaluator_model": "gemma2:latest",
  "prompt_paths": { /* ... */ },
  "question_dataset_paths": [
    "question_datasets/my_manual_questions.json", // Add new dataset here
    /* ... other existing datasets */
  ],
  // Update document_path or ensure documents_to_test in main.py includes the manual:
  "document_path": "https://yourdomain.com/path/to/your/manual.pdf", // Example for auto download
  /* ... other configurations */
}
```

An example command to run `ZeroShot/main.py` from within the `ZeroShot/` directory:

```
# From the ZeroShot directory
python main.py --models your_chosen_model --mode all --noise_levels 1000 5000 59000
```

D.5 Retrieval Augmented Generation (RAG) Testing Framework

The RAG testing framework, located in the `RAG/` directory, evaluates LLMs by first retrieving relevant
document chunks using various strategies (keyword, semantic, hybrid) and then providing these chunks
along with the question to the LLM. Configuration for the RAG framework, including LLM models, em-
bedding models, and dataset paths, is primarily managed through its configuration files (e.g., `config.ini`
or JSON configurations) and command-line arguments for its main evaluation scripts. Manuals must
be prepared (e.g., converted to TXT using the `docling_page_wise_pdf_converter.py` script from the
ZeroShot project and placed in a directory such as `RAG/manuals/`). The corresponding question dataset
must be placed in the `RAG/question_datasets/` folder.

The RAG pipeline can be tested with a single question using the `ask_question_demo.ipynb` script,
which is typically found within the `RAG/` directory. This script facilitates inputting a question and
specifying the document to observe the retrieved context and the LLM's answer, which is helpful for
debugging and exploration before running full-scale evaluations.

For comprehensive evaluations using various RAG strategies and LLMs, the `RAG/README.md` file
provides detailed setup, data preparation (including document processing and vector store creation), and
execution instructions for its main evaluation scripts.