Applied Computer Science, vol. 20, no. 4, pp. 175–191
doi: https://doi.org/10.35784/acs-2024-46

Submitted: 2024-09-21 | Revised: 2024-11-08 | Accepted: 2024-11-13

Keywords: large language models, retrieval-augmented generation, benchmark, desktop deployment, quantization

Marcin BADUROWICZ [0000-0003-2249-4219]*,
Stanisław P. SKULIMOWSKI [0000-0002-9049-9516]*,
Maciej LASKOWSKI [0009-0006-4255-0686]**

FEASIBILITY OF USING LOW-PARAMETER LOCAL LLMS IN ANSWERING QUESTIONS FROM ENTERPRISE KNOWLEDGE BASE

Abstract

This paper evaluates the feasibility of deploying locally-run Large Language Models (LLMs) for retrieval-augmented question answering (RAG-QA) over internal knowledge bases in small and medium enterprises (SMEs), with a focus on Polish-language datasets. The study benchmarks eight popular open-source and source-available LLMs, including Google's Gemma-9B and Speakleash's Bielik-11B, assessing their performance across closed, open, and detailed question types, with metrics for language quality, factual accuracy, response stability, and processing efficiency. The results highlight that desktop-class LLMs, though limited in factual accuracy (with top scores of 45% and 43% for Gemma and Bielik, respectively), hold promise for early-stage enterprise implementations. Key findings include Bielik's superior performance on open-ended and detailed questions and Gemma's efficiency and reliability in closed-type queries. Distribution analyses revealed variability in model outputs, with Bielik and Gemma showing the most stable response distributions. This research underscores the potential of offline-capable LLMs as cost-effective tools for secure knowledge management in Polish SMEs.

1. INTRODUCTION

Large Language Models (LLMs) have emerged as transformative tools in diverse applications, representing currently one of the fastest evolving components in the field of Machine Learning. Enterprise knowledge management systems can leverage LLMs for question answering over internal databases. Locally deployed LLMs offer a secure alternative to cloud-based solutions as they reduce the risk of exposing personally identifiable information (PII), proprietary information, trade secrets, or other sensitive data to third-party API providers, such as OpenAI. While API providers may have strict

[*] Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Department of Computer Science, m.badurowicz@pollub.pl, s.skulimowski@pollub.pl
[**] Independent researcher, maciej.laskowski@gmail.com

confidentiality policies, legal restrictions may prohibit organizations from transferring data to external servers altogether.

The improvement of information and knowledge management within organizations has been a long-standing objective, yet traditional approaches often struggle due to the extensive volume and structure of enterprise documents, which complicates accessibility, particularly for new employees. Large Language Models, using a Generative AI approach, can interface with local enterprise knowledge bases and provide accurate responses to user queries by employing the Retrieval-Augmented Generation (RAG) technique (Fan et al., 2024) in which relevant document fragments are inserted directly into the LLM context. This approach addresses the issue of limited context lengths in modern LLMs, which often fall short when processing extensive documents.

However, deploying large-scale LLMs presents significant challenges due to substantial hardware requirements, which can impose considerable costs. Organizations may opt for pilot implementations on consumer-grade systems or rely on technical staff to develop custom RAG solutions for internal use, and this is where the quantization process becomes important (Lin et al., 2024). Quantization is used essentially to encode the weights of the models using a smaller data size, using 8, 6, 4 or sometimes only even 2 bits per weight, thus allowing larger models to run on a desktop-class, consumer-grade Graphics Processing Units (GPUs).

This study evaluates eight popular open-source or source-available LLM models, each capable of operating offline operation on desktop-class hardware, by testing their performance in question answering (QA) over a provided documents in popular documentation formats. A unique challenge in this study is the focus on Polish-language datasets for both questions and answers, as LLMs typically perform sub-optimally in less commonly represented languages due to limited training data compared to English, even in case of multilingual models.

2. RELATED WORKS

The potential of Large Language Models in case of knowledge management in the an enterprise is huge (B & Purwar, 2024), especially when such enterprise has knowledge scattered over dozens of internal systems. LLMs and AI in general are already being used (or may be) in text-extraction and classification problems (Cevallos Salas, 2024) in various divisions of enterprises (Bouhsaien & Azmani, 2024; Soni et al., 2023). The next step is to build suitable and robust frameworks for internal documentation chatbots (Soto-Jiménez et al., 2024), technical document analysis (Menon, 2024), client records analysis (Zhu et al., 2024), even with multi-modal processing, and classification (Aydogan-Kilic et al., 2024) and more among other applications. However, a second problem arises - can we trust AI-generated content in professional fields, such as law? (Dahl et al., 2024).

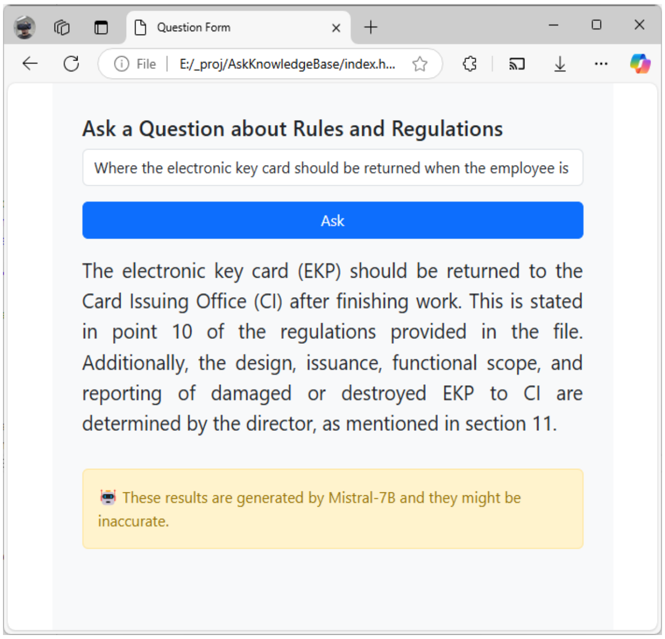
This second issue is not easy to be evaluated (Chen et al., 2021), and the RAG approach to enhance factual accuracy (Li et al., 2024; Zhang et al., 2023) only adds more further layers of difficulty (Ahmed et al., 2024; Gao et al., 2023). The typical benchmarks used for evaluating LLMs in multiple tasks (Bonatti et al., 2024; Tang et al., 2024) are not specifically tuned to such a kind of problem.

In Han et al. (2024), an Aa RAG-QA arena was proposed in which models competed by answering thousands of questions, and there was a framework for assessing human preference in pairs, similar to the previous approach to the same problem (Kamalloo et al., 2024).

However, the specifics of small and medium-sized enterprises are different. Although it is possible to build QA systems (Zhou et al., 2022) and such systems are being successfully prepared, there may be two additional needs: working completely privately (or even offline) and on desktops if there is no available budget to prepare a common server machine for multiple users. Especially small and medium businesses are likely to evaluate only a RAG QA approach for their internal documentation and knowledge bases using smaller LLMs. In this paper, the authors are concentrating on a question answering approach using a RAG method running on desktop-class devices, and – more importantly – in a relatively unpopular language. The results may be used to measure usefulness for internal and early evaluation of such methods in the small and medium-sized enterprises.

3. THE BENCHMARK METHODOLOGY

The authors chose to use Microsoft's kernel-memory, an open-source RAG solution. The pipeline was based on an in-memory concept, where all embedding data was generated from scratch after every run of the tool, to achieve completely new context. Kernel-memory version 0.71 was used in the experiments.



Ask a Question about Rules and Regulations

Where the electronic key card should be returned when the employee is

The electronic key card (EKP) should be returned to the Card Issuing Office (CI) after finishing work. This is stated in point 10 of the regulations provided in the file. Additionally, the design, issuance, functional scope, and reporting of damaged or destroyed EKP to CI are determined by the director, as mentioned in section 11.

These results are generated by Mistral-7B and they might be inaccurate.

Fig. 1. The question answering application

The authors base the experiment on a prototype question answering system prepared for internal deployment in the enterprise, presented in Figure 1.

During the introduction of the application there was a need for benchmarking the models used for question answering and thus the second, more automated application was prepared. Both the benchmarking system and the production system were using the same kernel-memory version based on LlamaSharp wrapper for a llama.cpp project for loading model files. LlamaSharp version used was 11b84eb4, and models were provided in the native llama.cpp's GGUF format. The pipeline of the experiment is presented in the Figure 2, below.

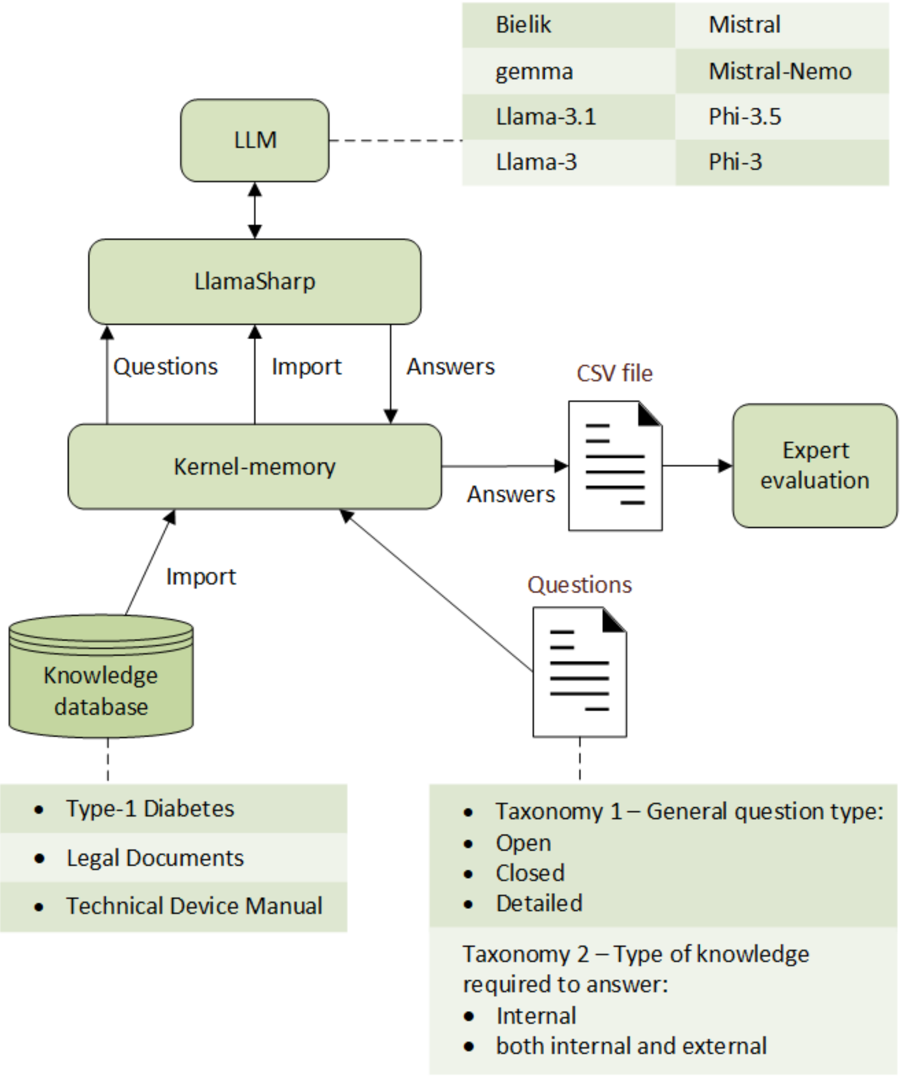


Fig. 2. The benchmark pipeline

3.1. The knowledge database

The knowledge database was segmented into three distinct datasets:

- The first dataset encompassed information related to the treatment of Type-1 diabetes, a well-documented topic. For this dataset, the model could draw upon both their internal (training-based) knowledge and information contained within the provided documents. This dataset was built from approximately 14.6 KB of text data provided in the form of internal documents exported to the PDF format,
- The second dataset consisted of legal documents in PDF format, with pure text-only sizes of 124 KB and 135 KB, respectively, containing complex legal terminology. As these documents were proprietary to the company, the models lacked prior exposure to the exact data. Thus, all references needed to be accurately cited from the documents themselves,
- The third dataset involved a user manual for a technical device, also in PDF format, totaling 5.5 MB, including images, with 92 KB of extractable text. While general operating procedures within this domain may have been part of the model's prior training, the specific details pertained exclusively to this device model. The files were imported directly into the pipeline without any alterations. None of the files were altered in any way; they were directly imported into the pipeline, from which text extraction and tokenization were subsequently performed.

3.2. Large Language Models included in the benchmark

Eight large language models (LLMs) were selected for benchmarking. Every model used was an 8-bit quantization variant (Q8_0), which is a high value, not very effective as for VRAM usage, yet minimizing hallucinations. All models were of the "instruct" type, optimized for instruction-following tasks, and the RAG (Retrieval-Augmented Generation) pipeline was configured accordingly.

The following set of models presented in Table 1 were chosen to be included in the experiments.

Tab. 1. Models to be benchmarked

Name	Size [number of parameters]	License
speakleash/Bielik-11B-v2.2-Instruct	11.2B	Apache 2 and custom terms
google/gemma-2-9b	9.24B	Custom terms
meta-llama/Llama-3.1-8B	8.03B	Custom terms
meta-llama/Meta-Llama-3-8B	8.03B	Custom terms
mistralai/Mistral-7B-Instruct-v0.2	7.24B	Apache 2.0
mistralai/Mistral-Nemo-Instruct-2407	12.2B	Apache 2.0
microsoft/Phi-3.5-mini-instruct	3.82B	MIT
microsoft/Phi-3-mini-4k-instruct	3.82B	MIT

The models were chosen based on their availability (licensing), popularity, size and rankings: (SpeakLeash | Spichlerz, n.d.; Ociepa, 2023), including the lowest number of "hallucination" (Vectara, 2024), which is a critical case in case of responding based on the known knowledge.

While Llama-family (Meta Llama, 2024a; Meta Llama, 2024b) of models are not officially described as supporting Polish language, the support is working quite well.

The Mistral family (Jiang et al., 2023; Mistral AI_, 2024a; Mistral AI_, 2024b) officially supports Polish language.

Phi-family (Microsoft, 2024a; Microsoft, 2024b) were chosen because of their lowest size, and the language quality was expected to be poor.

The Bielik (SpeakLeash | Spichlerz, 2024) was specifically chosen as a officially Polish-language supporting model. All the models included in the benchmark have a license of either Apache 2 or MIT or, if there are custom terms, allowing commercial usage in small and medium companies (SMEs) – in case of Gemma and Bielik the limitations are described in the prohibited usage policy and reflect illegal or malicious activities, as for Llama there is a limit of 700 million monthly active users.

3.3. Question sets and taxonomies

Twenty questions have been asked for each of the knowledge datasets. The answers were checked by the human judges and separate grades were used. The first one, the factual quality of the answer, ranged from -1 to 2 points, where 0 was awarded if there was no answer or the resulting string contained 'INFO NOT FOUND', as requested in the question, and negative values were awarded if the answer was factually incorrect. One or two positive points were given if the answer was partial (1 point) or full (2 points).

The "language" score was either 0 or -1, where the negative values were given by the referees, if the output was incoherent or incomprehensible, or if there were words which do not exist in Polish language. Finally, the "English" negative score was given if the answer was in English, not adjusting to the prompt, which required the model to answer in the same language as the question.

To sum up, it was possible to get a max of 120 points (3 domains, 20 questions, 2 points per question) for the factual grade, -60 for language grade and -60 for (in)adherence to the Polish language.

The questions have been divided into two taxonomies:

1. Taxonomy 1 – If the question is *open* (O_), *closed* (C_) or *detailed* (D_).

- The notion of "open" question means that the answer for that question was is expected to be in the form of a sentence or a paragraph.
- The notion of "closed" question means that answer is expected to be limited to one where it should be only one to a few words long.
- The "detailed" question means where the answer should be similar to answer to a "closed" question, but with a small additional comment.

2. Taxonomy 2 – If the question requires only internal knowledge extracted from the provided documents – *internal* (_I) –, or the model requires internal knowledge and may use its own knowledge from the training process – *both* (_B).

3.4. Benchmark environment details

The experiments were performed on an AMD Ryzen 9 7900X-based desktop PC with 64 GB of RAM and AMD Radeon 7900XT GPU with 20 GB of VRAM memory, allowing the models to be loaded fully into the graphics card memory, and still being the desktop-class device suitable for developers or power users of the enterprise, with its price around $1000.

The experiment scripts were importing the documents for a given dataset and then automatically asking questions and logging both the answers and their answer time to CSV files. The exported CSV files were later sent to the domain experts for evaluation and annotation.

4. RESULTS AND DISCUSSION

The total sum for the factual answer grade for each model (with names abbreviated names) is presented in Figure 3, below. As mentioned earlier, the highest possible value of factual grade was 120 points, thus meaning that the best models, Google's Gemma-9B and Speakleash's Bielik only got 45% and 43% of maximum points, respectively. But just the total sum of factual grades (blue bars on chart) is not extensive enough to finally tell if the model is good enough for the desired usage. Furthermore, it can be seen that an increase in model size (number of parameters) does not always mean an increase in model performance.

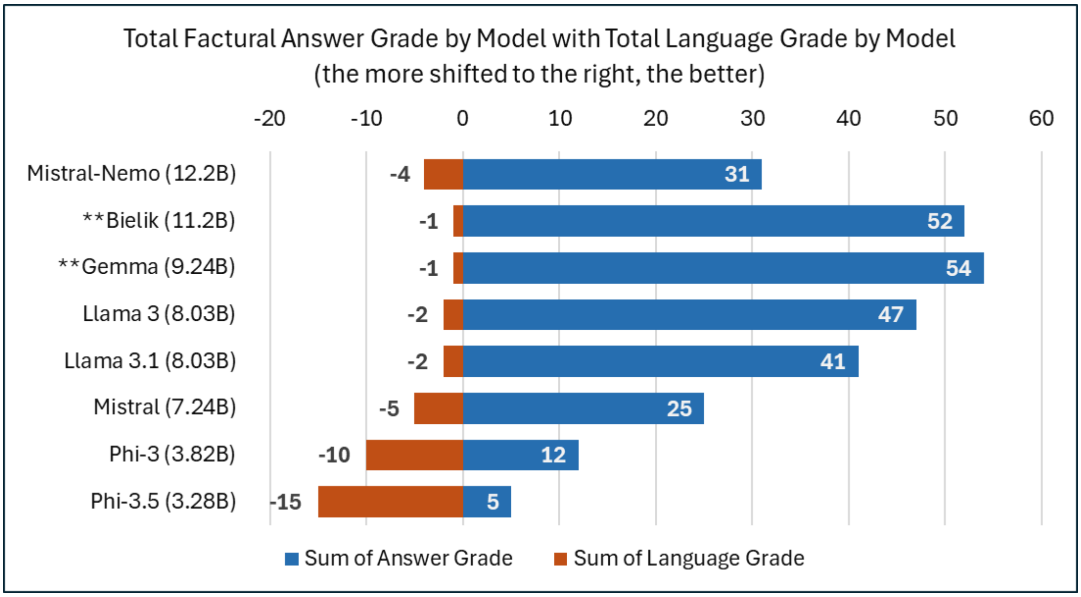


Fig. 3. Total Answer Grade by Model with Total Language Grade by Model (sum of the "Language quality" and "English" negative grades by model). The maximum possible grade was 120. LLMs sorted by number of parameters. Best scores labeled with stars

The quality of the language quality (also analyzed by the domain experts) was a big significant factor in calculating the final score – as expected, the Phi-3-mini and Phi-3.5-mini, being the smallest (in terms of number of parameters) hugely lacked in quality of the Polish language, mostly in inflections of verbs and nouns. The Mistral-7B, was preferring tended to answer in English, ignoring the prompt asking the model to answer in the same language as the original question language. The quality of the language: summed "English" negative score and "Language quality" negative scores of the answers are presented in Figure 3 as red bars. For the best models in the "Factual" grade, Bielik and Gemma, the score is both just -1, meaning only one answer (0.8%) was using wrong language.

Two additional questions were:

– whether the models perform better with certain types of questions,
– if there are models more effective in working with only the data provided from the documents in the knowledge base.

Results are presented in Figures 4, 6 and 7, where it is shown shows that Mistral and its derivative, Bielik, got the highest grades for the open questions, with Bielik being also the top in the case of detailed questions. Llama3 and Gemma were best for the closed type of questions, where a single response was preferred.

It is worth noting that the quality of the provided answers (even within one questions' category) was not the same for all models. Figure 5 presents the distribution of the types of answers provided by the models to Closed Questions, presented generally in Figure 4. The fewest answers (0 points) were provided by Mistral-Nemo and the most answers were provided by Mistral, although its quality was the worst. A small share of partially correct answers (1 point) is a common feature of all the distributions and the clear division can be noticed between very good (2 points) and very bad answers (-1 point). The Mistral and Phi-3.5 also presented the highest number of factually wrong answers (-1 point), which means they may be not the suitable solutions for the QA.

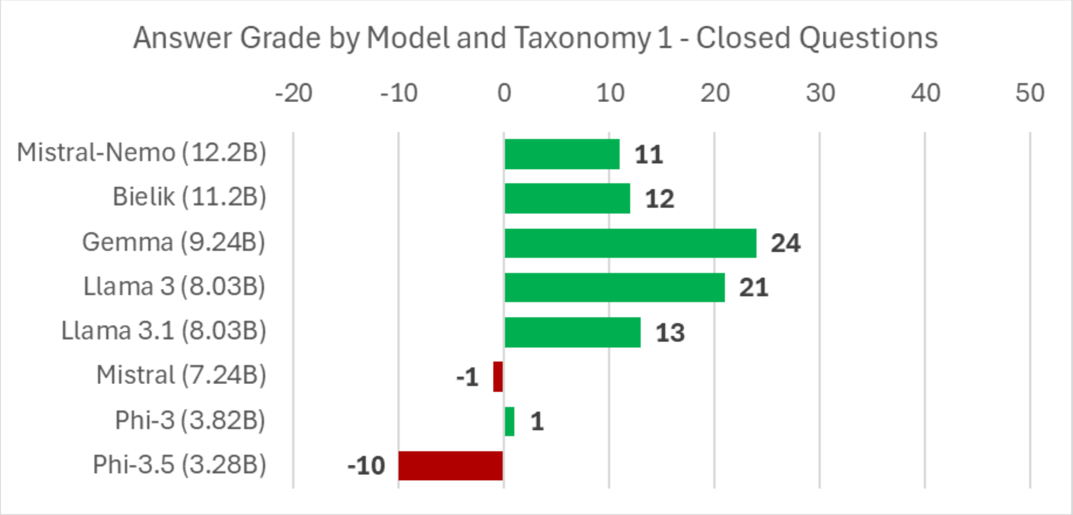


Fig. 4. Answer grade for each model for Closed Questions (C_). The maximum possible grade was 50. Sorted by LLMs size [number of parameters]

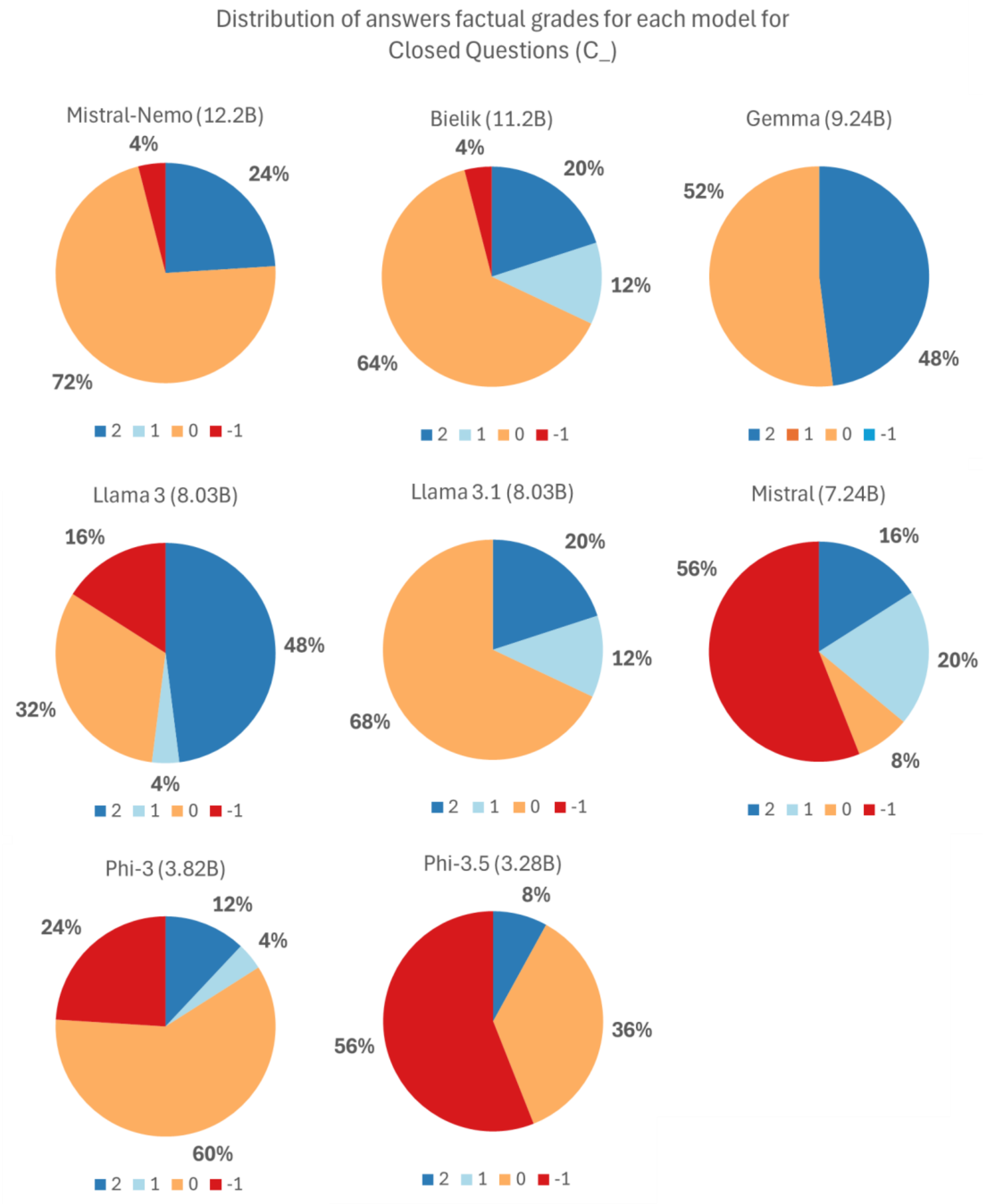


Fig. 5. Distribution of answers factual grades for each model for Closed Questions (C_)

Analysis of the Histogram analysis of the results of individual LLMs allowed the authors to determine the structure of the responses generated (Table 2). For the majority of LLMs, the kurtosis (Fisher standardized) exhibits a negative value, indicating that the distribution is platykurtic and devoid of any tendency for outliers. The grades of the answers were concentrated in groups in which the mode was 0.

The Phi-3 model shows an almost mesokurtic distribution. This is further supported by the low value of the standard deviation. This may suggest more certainty about the quality of the responses generated by this model.

The Mistral Nemo model exhibits a leptokurtic distribution, with the highest concentration of results occurring at 0. The high positive skewness value indicates that this model has relatively few higher grades. This is also confirmed by the low mean value.

Taking into account all the examined characteristics of the histograms into account, Bielik and Gemma achieved the most favorable outcomes, sustaining a high mean result, stability, and robust grouping of results. The worst results are shown by Mistral and Phi-3.5 with a negative average rating and large value fluctuation

Figures 8 and 9 present Gemma's highest total score in the answers based on the knowledge base, with Llama3 being the second. Bielik, however, got the highest results for the answers both including the knowledge base and the model's knowledge. Phi-3.5 demonstrated that answers based purely on document analysis were false in many cases, further disqualifying this model for RAG-QA usage.

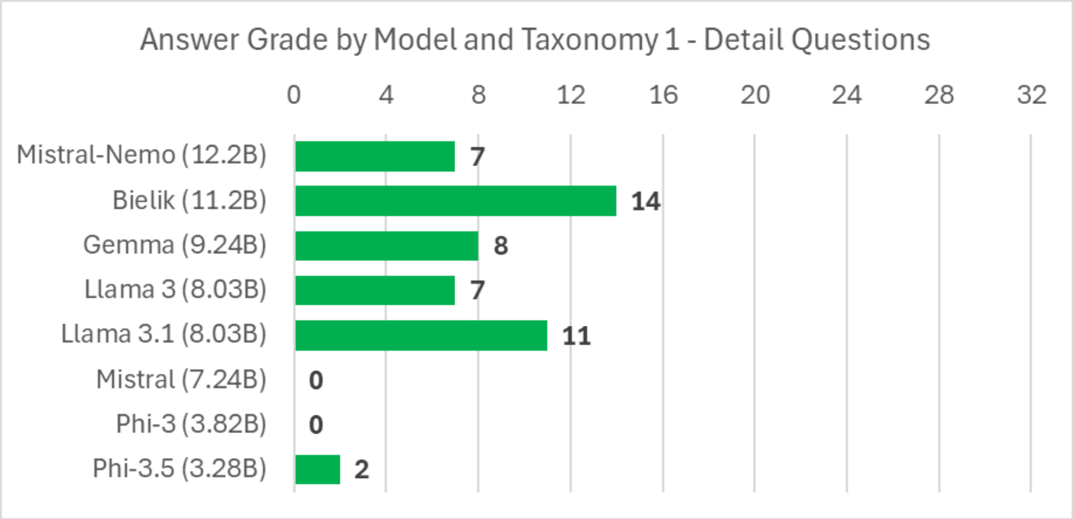


Fig. 6. Answer grade for each model for Detail Questions (D_). The maximum possible grade was 32. Sorted by LLMs size [number of parameters]

Results show that Bielik and Gemma models were the best ones, with a final result of 51 points out of 120 possible (42.5%).

An attempt was also made to evaluate the LLMs taking into account the results in the question categories. The results were as follows:

- answer quality with category weight – The sum of the scores in a given category was normalized to the maximum possible score in a given category. Then the value was multiplied by the category weight, for better visualization experience. The categories CI, OI and DI were assigned a weight of 1.5, while CB, OB and DB were assigned a weight of 1.0. The greater the value the better.
- average response time – For this purpose, the average response time for questions in each category was determined. The lower the time the better.

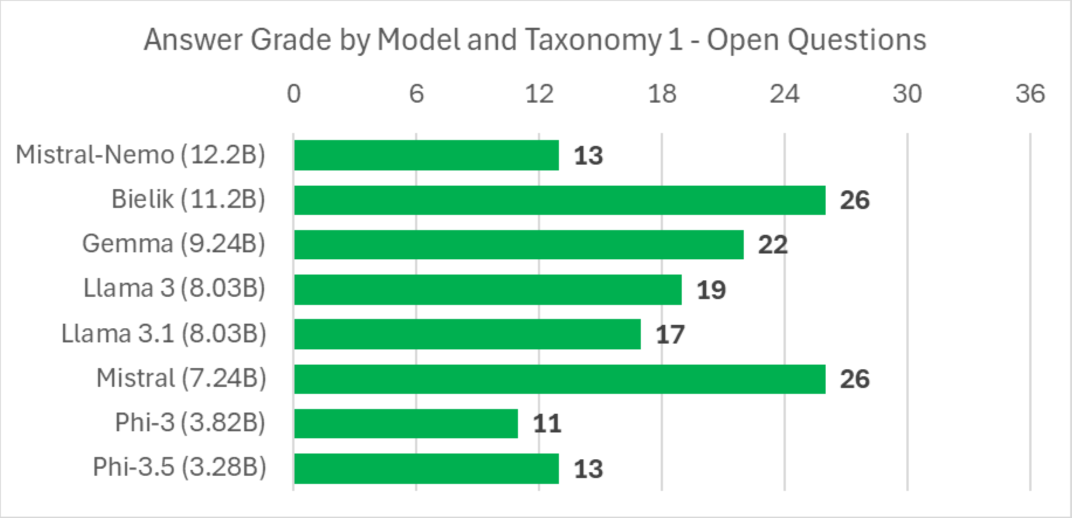


Fig. 7. Answer grade for each model for Open Questions (O_). The maximum possible grade was 36. Sorted by LLMs size [number of parameters]

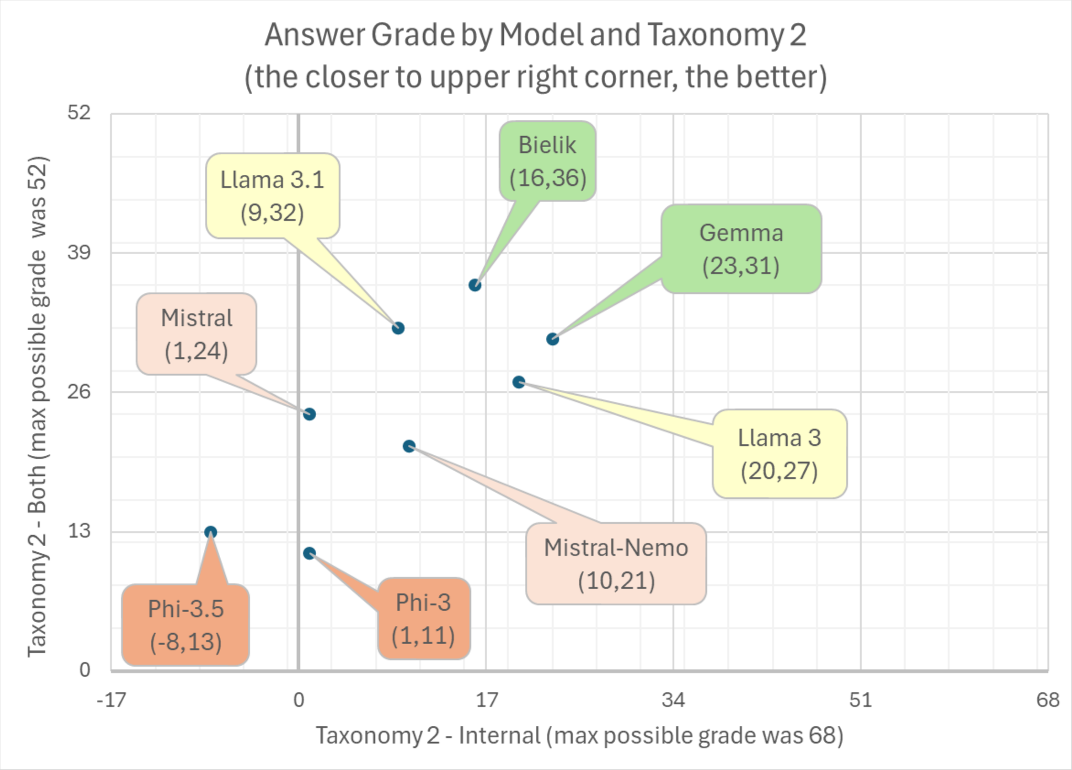


Fig. 8. Answer grade for each model by the type of knowledge used (Taxonomy 2). Scores gradient: orange-low score, green- high score

The time of documents import and processing the documents was also measured and Bielik was the slowest one in total – Gemma needed only 75% of the time (Figure 9).

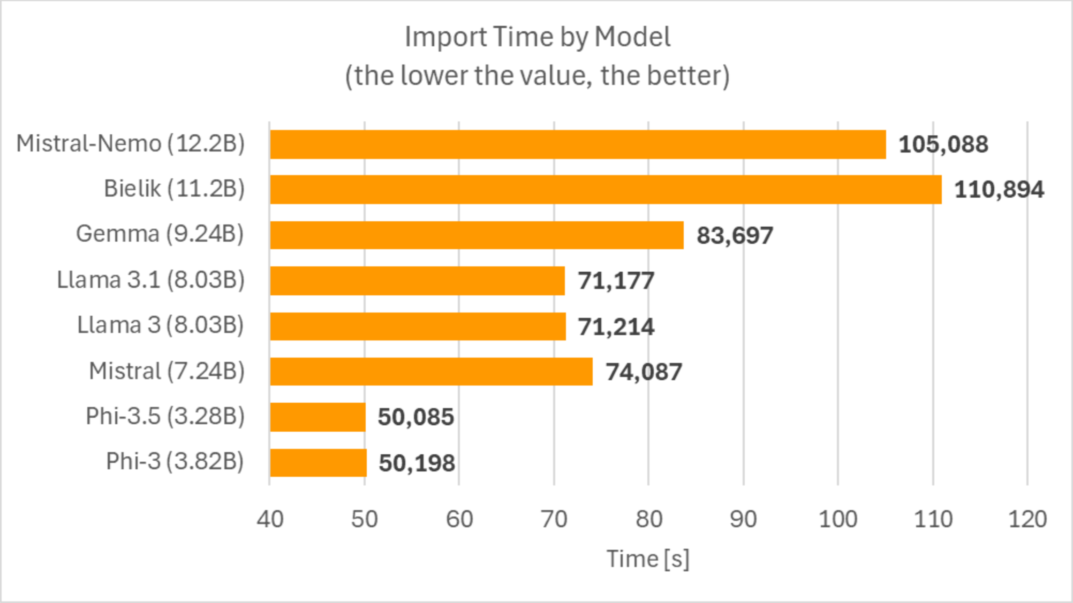


Fig. 9. Import Time [s] by Model. Sorted by LLMs size [number of parameters]

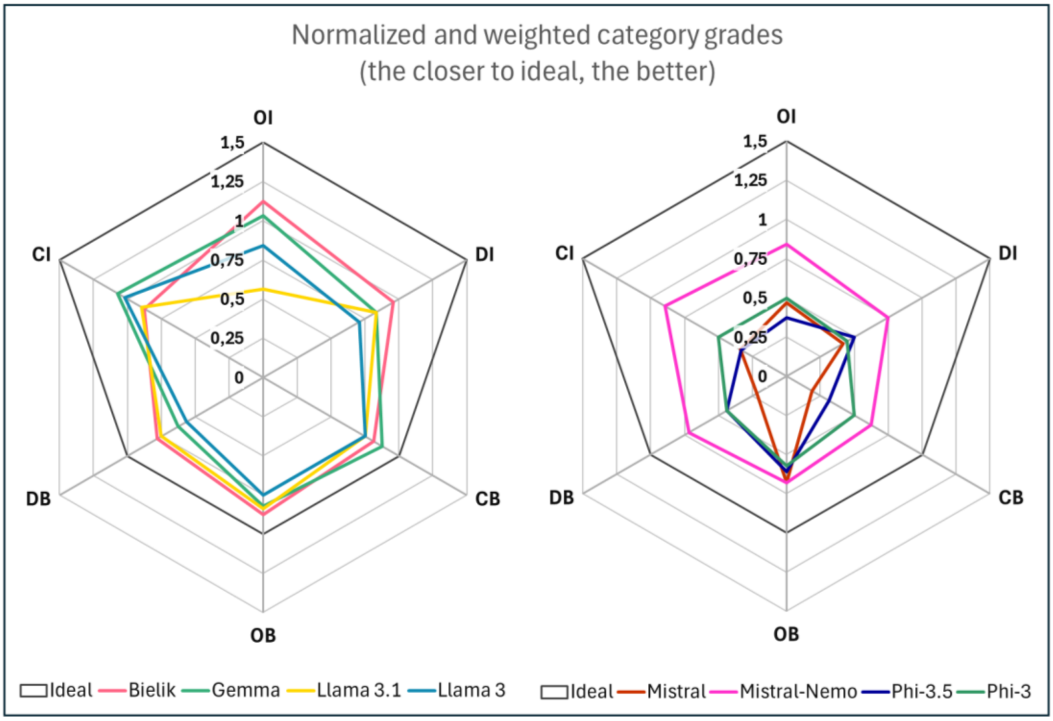


Fig. 10. Radar chart of the normalized and weighted grades obtained in individual categories by each of the examined LLM. Values increase from the center outwards. Divided into two sub-diagrams for greater readability

The analysis results were recorded in cross-tabulation form and visualized as radar diagrams (Figures 10. and 11.). This visualization method allows to check the quality of the answers and the response time (the larger the figure the better).

Most models show a balanced answers quality with a small bias towards the CI, OI and DI categories (Figure 10.) what can be determined by bringing the graph closer to the ideal shape. This is especially true for Bielik and Gemma. The only exception is Llama 3.1, whose only outlier result is the lower values in the OI category.

Most models show a bias toward faster response generation in the CB, DI, and CI categories (Figure 11.). The greatest category bias represents Bielik – it clearly shifts towards all _I categories. Also, Bielik and Mistral Models represent the worst respond time, with lowest polygons area and lowest perimeter.

Llama 3 shows the least bias towards any category (low standard deviation of its vertices distance to origin point). This model exhibits the greatest balance in time required for answer across all categories, in terms of the small distance from centroid to the coordinate system origin, long perimeter, and great polygon area.

The analysis of the histograms of individual LLMs for the overall score is presented in Table 2. An internal similarity in the overall set of values is indicated by the standard deviation. All LLMs except one exhibit similar deviation values. Mistral exhibits a slightly different value, but looking at the mean value, it can be concluded that this distinction is negative. All LLMs have a positive skewness, which means that all of them tend to produce low values (which was presented earlier, for example, in Figure 5). The kurtosis of the majority of LLMs exhibits negative values, indicating platykurtic distributions - low values without a distinct distinction. The exception is Mistral-Nano, whose leptokurtic distribution suggests a significant concentration of low values.

Tab. 2. Results summary for tested LLMs. The grade relates to the overall score, without division into groups. Alphabetic order by model name

Model	Mean	Median	Mode	Standard deviation	Skewness	Kurtosis (Fisher)
Bielik	0.850	1	0	0.90967	0.61186	-1.62348
Gemma	0.850	0	0	0.96307	0.97446	-1.22421
Llama 3	0.583	0	0	1.05343	0.46234	-0.81415
Llama 3.1	0.650	0	0	0.96307	1.37088	1.03831
Mistral	-0.633	0	-2	1.30341	0.61464	-1.02685
Mistral Nemo	0.466	0	0	0.90308	1.98222	3.95122
Phi-3	0	0	0	0.85635	0.92303	-0.03783
Phi-3.5	-0.25	0	0	0.99373	0.25853	-1.53482

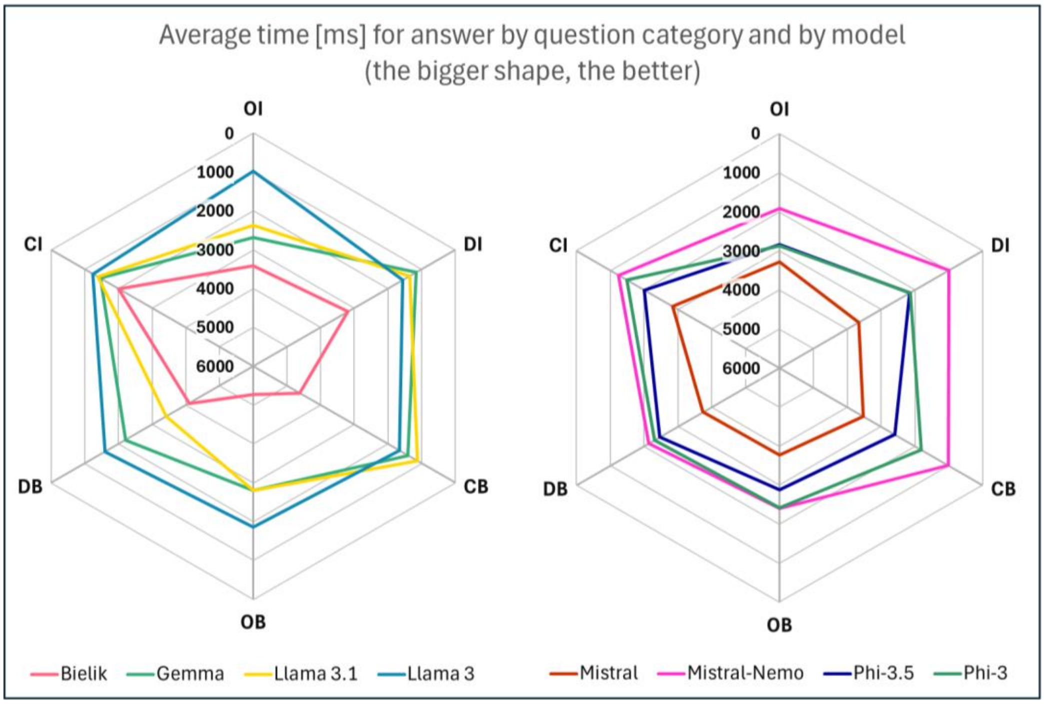


Fig. 11. Radar chart of average times for answer in each category by each of the LLMs studied. Values increase from the outside inwards. Time in milliseconds. Divided into two sub-diagrams for greater readability

5. CONCLUSIONS AND FUTURE WORKS

The authors successfully established a comprehensive benchmark for evaluating Large Language Models (LLMs) capable of running on desktop-class hardware for retrieval-augmented question answering (RAG-QA). The study focuses on Polish-language knowledge bases, typically stored in PDF documents. The results indicate that these locally-deployable models show potential for pilot implementations, though the highest factual accuracy scores remain moderate, as Google's Gemma-9B and Speakleash's Bielik-11B achieved the top factual accuracy scores, with 45% and 43% of the maximum points, respectively. Even with such values of accuracy pilot deployment may be valuable within small and medium enterprises (SMEs), for searching through internal databases of PDF documents.

The findings highlight a marked advantage of newer and larger models in language quality (with Bielik and Gemma standing out as the most reliable performers). It must be noted, that Bielik excelled in open-ended and detailed questions, while Gemma performed best on closed-type queries. Distribution analysis using kurtosis and skewness measures revealed that Bielik and Gemma offered stable, grouped responses, whereas Mistral and Phi-3.5 exhibited high variability, indicating inconsistent outputs. Regarding domain-specific knowledge retrieval, Gemma scored highest on responses based solely on the knowledge base, while Bielik excelled in combining external knowledge with the retrieved content.

However, Phi-3.5 faced challenges in RAG-specific tasks, signaling limitations in its design for such applications.

In terms of processing efficiency, Bielik was the slowest model, while Gemma processed documents 25% faster. Despite these limitations, both models are positioned as strong contenders in the benchmark, showcasing their feasibility for secure and cost-effective deployment in SMEs.

Future work will address three key areas:

- VRAM Optimization: Advanced quantization methods such as Q4_K_M or imatrix-based IQ4_XS could significantly reduce memory requirements while maintaining model quality, improving accessibility and lowering deployment costs for SMEs – in this paper, the authors were working with Q8 class-optimization, while 4-bit quantization can run on even smaller hardware, allowing cheaper deployment in smaller enterprises.
- Baseline Comparison: Testing against state-of-the-art models like GPT-4o or larger variants of the Llama and Mistral families will establish clearer performance benchmarks and provide insights into the scalability of the proposed approach.
- Energy Efficiency: A detailed evaluation of the energy consumption of locally-deployed models will be conducted to assess their alignment with "Green AI" principles, balancing cost savings and sustainability goals.

Authors would like to suggest practical recommendations for SMEs include prioritizing scenarios that demand secure knowledge management and offline document retrieval, where the models' strengths in language quality and knowledge integration can be maximally leveraged.

Author Contributions

M. Badurowicz - initial concept and experiments
S.P. Skulimowski - data analysis and visualizations
M. Laskowski - language and factual correctness

Conflicts of Interest

The authors declare there were no conflicts of interest.

REFERENCES

Ahmed, T., Bird, C., Devanbu, P., & Chakraborty, S. (2024). Studying LLM performance on closed- and open-source data. *ArXiv, abs/2402.15100*. https://doi.org/10.48550/arXiv.2402.15100

Aydogan-Kilic, D., Kilic, D. K., & Nielsen, I. E. (2024). Examination of summarized medical records for ICD code classification Via BERT. *Applied Computer Science*, *20*(2), 60-74. https://doi.org/10.35784/acs-2024-16

B, G., & Purwar, A. (2024). Evaluating the efficacy of open-source LLMs in enterprise-specific RAG systems: A comparative study of performance and scalability. *ArXiv, abs/2406.11424*. https://doi.org/10.48550/arXiv.2406.11424

Bonatti, R., Zhao, D., Bonacci, F., Dupont, D., Abdali, S., Li, Y., Lu, Y., Wagle, J., Koishida, K., Bucker, A., Jang, L., & Hui, Z. (2024). Windows agent arena: Evaluating multi-modal OS agents at scale. *ArXiv, abs/2409.08264*. https://doi.org/10.48550/arXiv.2409.08264

Bouhsaien, L., & Azmani, A. (2024). The potential of Artificial Intelligence in human resource management. *Applied Computer Science*, *20*(3), 153-170. https://doi.org/10.35784/acs-2024-34

Cevallos Salas, F. A. (2024). Digital news classification and punctuaction using Machine Learning and text mining techniques. *Applied Computer Science*, *20*(2), 24-42. https://doi.org/10.35784/acs-2024-14

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., … Zaremba, W. (2021). Evaluating Large Language Models trained on code. *ArXiv, abs/2107.03374*. https://doi.org/10.48550/arXiv.2107.03374

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in Large Language Models. *Journal of Legal Analysis*, *16*(1), 64-93. https://doi.org/10.1093/jla/laae003

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T. S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented Large Language Models. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '24)* (pp. 6491-6501). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671470

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for Large Language Models: A survey. *ArXiv, abs/2312.1099*. https://doi.org/10.48550/arXiv.2312.1099

Han, R., Zhang, Y., Qi, P., Xu, Y., Wang, J., Liu, L., Wang, W. Y., Min, B., & Castelli, V. (2024). RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. *ArXiv, abs/2407.13998*. https://doi.org/10.48550/arXiv.2407.13998

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. Le, Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. *ArXiv, abs/2310.06825*. https://doi.org/10.48550/arXiv.2310.06825

Kamalloo, E., Upadhyay, S., & Lin, J. (2024). Towards robust QA evaluation via open LLMs. *47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)* (pp. 2811-2816). Association for Computing Machinery. https://doi.org/10.1145/3626772.3657675

Li, J., Yuan, Y., & Zhang, Z. (2024). Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *ArXiv, abs/2403.10446*. https://doi.org/10.48550/arXiv.2403.10446

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. *ArXiv, abs/2306.00978*. https://doi.org/10.48550/arXiv.2306.00978

Menon, K. (2024). *Utilizing open-source AI to navigate and interpret technical documents : Leveraging RAG models for enhanced analysis and solutions in product documentation*. http://www.theseus.fi/handle/10024/858250

Meta Llama. (2024a, July 23). *meta-llama/Llama-3.1-8B*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/meta-llama/Llama-3.1-8B

Meta Llama. (2024b, April 24). *meta-llama/Meta-Llama-3-8B*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/meta-llama/Meta-Llama-3-8B

Microsoft. (2024a, September 18). *microsoft/Phi-3.5-mini-instruct*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/microsoft/Phi-3.5-mini-instruct

Microsoft. (2024b, September 20). *microsoft/Phi-3-mini-4k-instruct*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Mistral AI_. (2024a, September 27). *mistralai/Mistral-7B-Instruct-v0.2*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Mistral AI_. (2024b, November 6). *mistralai/Mistral-Nemo-Instruct-2407*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

Ociepa, K. (2023). *PoLEJ - Polish Open LLM Leaderboard*. Azurro. Retrieved October 30, 2024 from https://polej.azurro.pl/

Soni, S., Datta, S., & Roberts, K. (2023). quEHRy: A question answering system to query electronic health records. *Journal of the American Medical Informatics Association*, *30*(6), 1091-1102. https://doi.org/10.1093/JAMIA/OCAD050

Soto-Jiménez, F., Martínez-Velásquez, M., Chicaiza, J., Vinueza-Naranjo, P., & Bouayad-Agha, N. (2024). RAG-based question-answering systems for closed-domains: Development of a prototype for the pollution domain. In K. Arai (Ed.), *Intelligent Systems and Applications* (Vol. 1065, pp. 573-589). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-66329-1_37

SpeakLeash | Spichlerz. (2024, October 26). *speakleash/Bielik-11B-v2.2-Instruct*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/speakleash/Bielik-11B-v2.2-Instruct

SpeakLeash | Spichlerz. (n.d.). *Open PL LLM Leaderboard - a Hugging Face Space*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/spaces/speakleash/open_pl_llm_leaderboard

Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Lin, C., Li, W., Mahmood, M. F. F. Bin, Feng, H., Zhao, Z., Wang, Y., Liu, Y., Liu, H., Bai, X., & Huang, C. (2024). MTVQA: Benchmarking multilingual text-centric visual question answering. *ArXiv, abs/2405.11985*. https://doi.org/10.48550/arXiv.2405.11985

Vectara. (2024, December 11) *Hallucination Evaluation Leaderboard - a Hugging Face Space*. Hugging Face. Retrieved October 30, 2024 from https://huggingface.co/spaces/vectara/Hallucination-evaluation-leaderboard

Zhang, Y., Khalifa, M., Logeswaran, L., Lee, M., Lee, H., & Wang, L. (2023). Merging generated and retrieved knowledge for open-domain QA. *2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (EMNLP 2023)* (pp. 4710-4728). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.286

Zhou, Y. Q., Liu, X. J., & Dong, Y. (2022). Build a robust QA system with transformer-based mixture of experts. *ArXiv, abs/2204.09598*. https://doi.org/10.48550/arXiv.2204.09598

Zhu, Y., Ren, C., Xie, S., Liu, S., Ji, H., Wang, Z., Sun, T., He, L., Li, Z., Zhu, X., & Pan, C. (2024). REALM: RAG-driven enhancement of multimodal electronic health records analysis via Large Language Models. *ArXiv, abs/2402.07016*. https://doi.org/10.48550/arXiv.2402.07016