La confiance de Mistral-7B est-elle justifiée ? Une évaluation en auto-estimation pour les questions biomédicales

Laura Zanella [1] Ambroise Baril[2]
(1) SAS Posos, 44 Rue La Fayette, 75009 Paris, France
(2) Académie de Versailles, 3 Bd de Lesseps, 78017 Versailles, France
laura.zanella-calzada@posos.fr, ambroise-jean-e.baril@ac-versailles.fr

RÉSUMÉ

Évaluer la fiabilité des grands modèles de langage (LLMs) dans des tâches de question-réponse biomédicale est essentiel pour leur déploiement en toute sécurité dans des contextes médicaux. Dans cette étude, nous examinons si Mistral-7B est capable d'estimer avec précision la confiance qu'il accorde à ses propres réponses, en comparant ses scores de similarité auto-attribués à la similarité-cosinus avec des réponses de référence. Nos résultats montrent que Mistral-7B présente une forte tendance à la surconfiance, attribuant systématiquement des scores de similarité élevés, même lorsque la qualité des réponses varie. L'introduction de la génération augmentée par récupération (RAG) améliore la précision des réponses, comme en témoignent les valeurs plus élevées de similarité-cosinus, mais n'améliore pas significativement la calibration de la confiance. Bien que RAG réduise la surconfiance et améliore la corrélation entre les scores de similarité prédits et réels, le modèle continue de surestimer systématiquement la qualité de ses réponses. Ces résultats soulignent la nécessité de mécanismes d'estimation de confiance plus efficaces, afin d'aligner les auto-évaluations du modèle sur la précision réelle de ses réponses. Notre étude montre l'importance d'affiner les techniques de calibration des LLMs pour renforcer leur fiabilité dans les applications biomédicales.

ABSTRACT

Is Mistral-7B's Confidence Justified? Assessing Self-Evaluation in Biomedical QA

Assessing the reliability of LLMs in biomedical QA is crucial for their safe deployment in medical contexts. In this study, we investigate whether Mistral-7B can accurately estimate the confidence of its own responses by comparing its self-assigned similarity scores to the cosine similarity with reference answers. Our results show that Mistral-7B exhibits a strong tendency toward overconfidence, consistently assigning high similarity scores even when response quality varies. The introduction of retrieval-augmented generation (RAG) improves the accuracy of responses, as reflected in higher cosine similarity values, but does not meaningfully enhance confidence calibration. While RAG reduces overconfidence and improves correlation between predicted and actual similarity scores, the model still systematically overestimates answer quality. These findings highlight the need for improved confidence estimation mechanisms to align model self-assessments with actual response accuracy. Our study underscores the importance of refining calibration techniques for LLMs in biomedical applications to enhance their reliability in AI-assisted decision-making.

MOTS-CLÉS: Mistral-7B, Grands modèles de langage (LLM), Calibration des LLM, Auto-évaluation des LLM, Génération augmentée par récupération (RAG)

KEYWORDS: Mistral-7B, LLM calibration, LLM self-assessment, RAG

1 Introduction

Large Language Models (LLMs) are increasingly being used in biomedical applications, where accuracy and reliability are critical. However, a major challenge in deploying these models is their occasional tendency to generate incorrect or misleading information, which can pose significant health risks. Given this challenge, the goal is not to replace human decision-making, but rather to provide a second opinion or decision support. To ensure the safe and effective use of LLMs in such high-stakes domains, it is essential to evaluate not only the quality of their responses but also the reliability of the model's confidence in its own outputs. One possible approach is to have LLMs provide a confidence estimate alongside their responses, allowing users to trust the reliability of the information provided. However, a fundamental question arises: Can we trust the confidence estimates generated by these models? If an LLM is overconfident, users may place excessive trust in incorrect answers, leading to potentially harmful medical decisions. Conversely, if a model systematically underestimates its confidence, it becomes difficult to distinguish between reliable and unreliable responses, limiting its practical usefulness. Simply biasing the model toward more conservative estimates is not a viable solution either, as it would blur the distinction between correct and incorrect answers.

To investigate this issue, we assess the ability of Mistral-7B[1] to provide reliable confidence estimates. Specifically, we analyze whether its self-reported confidence scores align with actual answer quality by measuring the cosine similarity. Additionally, we explore the potential of Retrieval-Augmented Generation (RAG) to improve answer quality and mitigate overconfidence in LLMs. RAG enhances a model's responses by retrieving relevant external information before generating an answer, helping to ground its outputs in factual data. This method has been shown to reduce hallucinations where the model generates false information by integrating evidence-based information (Yuan *et al.*, 2024; Ng *et al.*, 2025). Given this, we hypothesize that RAG could also improve confidence calibration by reducing the model's reliance on uncertain internal knowledge. In our setup, the model is presented with additional information retrieved through semantic search, without being explicitly told that it comes from an external source or is directly related to the answer. By analyzing the impact of this implicit exposure to high-quality information, we assess whether this approach leads to improved answer accuracy and better-aligned confidence estimates.

We aim to explore whether LLMs can provide reliable confidence estimates in biomedical question-answering (QA) and how RAG influences their accuracy. Our main contributions are listed above.

Evaluation of Mistral-7B's Self-Assessment: We assess the ability of the model to estimate the quality of its answers by comparing its self-assigned similarity scores with the cosine similarity between its responses and the gold standard answers.

Impact of Context-Enriched Prompting: We evaluate the model in two settings: (1) a baseline setup where the model is prompted to answer questions and provide similarity scores, and (2) a context-enriched setup where relevant information retrieved via RAG is incorporated into the prompt to support the model's responses.

Analysis of Confidence and Accuracy: We show that while RAG improves answer quality, it also increases the model's confidence in its predictions, sometimes leading to an overestimation of similarity scores.

[1] https://huggingface.co/mistralai/Mistral-7B-v0.1

2 Related works

Confidence Estimation and Calibration in LLMs

The estimation of confidence in LLMs for QA has been widely studied, with a focus on quantifying uncertainty and improving reliability. A well-calibrated model should not only provide high discrimination, i.e., assigning higher confidence to correct responses than incorrect ones, but also exhibit proper calibration, meaning that its confidence scores accurately reflect the true probability of correctness (Chen & Mueller, 2024; Gawlikowski *et al.*, 2023; Abdar *et al.*, 2021).

Recent work has shown that LLMs often struggle with calibration, typically displaying overconfidence, particularly in complex domains like biomedicine. For example, models can assign confidence levels of 90% to responses that are only correct 70–80% of the time, conducting to misleading trust in their output. This issue is critical in sensitive applications, such as medical decision making, where overconfidence can result in unsafe recommendations. Studies have explored various uncertainty quantification techniques, including token-level probabilities, confidence elicitation, and sample consistency measures (Savage *et al.*, 2024). Other recent analyses of LLM confidence metrics highlight that while these models can exhibit strong discrimination, their calibration remains unreliable, particularly in automated evaluation tasks (Stureborg *et al.*, 2024).

A 2023 survey (Geng *et al.*, 2023) offers a comprehensive overview of early efforts in confidence estimation and calibration for LLMs. It identifies key challenges specific to language generation—such as the vast and semantically variable output space—and organizes existing approaches across generative and classification settings. While more recent studies have built on this foundation, the survey remains a key reference for understanding the conceptual and technical landscape of the field.

In parallel, the TruthfulQA benchmark (Lin *et al.*, 2021) addresses the issue of factual correctness from a different angle, by evaluating whether LLMs generate truthful answers to questions that often elicit misconceptions. The study finds that larger models are more prone to imitative falsehoods—plausible but incorrect answers learned from human text—exposing a counterintuitive "inverse scaling" effect. These findings highlight that increasing model size alone is not sufficient for improving truthfulness, and emphasize the need for evaluation frameworks that go beyond surface-level fluency.

A related direction explores whether models can introspect about their own knowledge by predicting their own probability of providing a correct answer ($\mathbb{P}(\text{TRUE})$), or estimating in advance whether they are likely to know the answer to a question ($\mathbb{P}(\text{IK})$, IK stands for "I know"). Recent work shows that larger models exhibit strong calibration on diverse question formats, and that self-evaluation improves when models generate multiple answer candidates before scoring their plausibility. Furthermore, models trained to estimate $\mathbb{P}(\text{IK})$ can anticipate whether they are likely to answer a question correctly—generalizing partially across tasks and improving when contextual information (e.g., hints or source documents) is available. These techniques suggest a promising route for developing more honest and self-aware systems, going beyond mere imitation of training data (Kadavath *et al.*, 2022).

RAG and Its Impact on Model Calibration

RAG has been proposed as a way to improve response accuracy by integrating external knowledge into model prompts. Previous work suggests that RAG can reduce hallucinations and improve factual grounding, but its impact on confidence calibration remains inconclusive (Li *et al.*, 2024; Yarie *et al.*, 2024; Song *et al.*, 2024). While some analyses indicate a better alignment between model confidence and correctness when retrieved context is included, others report persistent overconfidence and limited

adjustment of certainty based on the relevance or quality of the retrieved content (Savage *et al.*, 2024).

Recent findings suggest that hallucination in LLMs is inevitable, as these models inherently generate plausible but incorrect outputs, even with improved training and retrieval mechanisms (Xu *et al.*, 2024). This further complicates the challenge of trustworthiness in AI-driven applications. Similarly, a recent study on belief alignment in LLMs emphasizes that users should be cautious when interpreting model confidence, as LLMs can exhibit confidence biases that do not always correlate with factual accuracy (Yadkori *et al.*, 2024).

3 Assessing Mistral-7B confidence Estimation

3.1 Preliminaries

We assess Mistral-7B's ability to self-evaluate its confidence in a biomedical QA task. Specifically, we evaluate the model's confidence calibration by assessing: (1) its ability to generate answers based on internal knowledge or retrieved context, (2) its self-assigned similarity score, estimating alignment with the correct response (which remains unseen), and (3) the correlation between these scores and cosine similarity to the gold answer. A well-calibrated model should align its predicted similarity with actual response quality. Overconfident models inflate scores for incorrect answers, while underconfident models undervalue correct ones. Our goal is to analyze these patterns and assess Mistral-7B's reliability in self-assessment. Additionally, we examine whether RAG mitigates overconfidence by enriching the model's knowledge at inference time. Specifically, we explore whether Mistral-7B adjusts its confidence when given additional context, ideally increasing scores when the retrieved information is helpful and moderating them when uncertain.

3.2 Semantic-enriched prompt

We follow a RAG-based approach to add semantic information to the prompt, where Mistral-7B receives additional context retrieved via semantic search. Specifically, we include in the prompt the gold answer and the most similar response from the dataset, without explicit confirmation of relevance. Since RAG reduces hallucinations by grounding responses in factual data, we hypothesize that it may also enhance confidence calibration. By incorporating retrieved knowledge, we evaluate whether Mistral-7B's self-assigned similarity scores better reflect actual response quality.

3.3 Similarity Comparison Evaluation

To assess the reliability of Mistral-7B self-assigned similarity scores, we compare them against an objective measure of similarity between the model's answers and the gold-standard responses.

3.3.1 Objective Similarity Measure

We compute the cosine similarity between the embeddings of the generated answer and the gold answer, which provides a numerical estimate of how closely the two responses align. Cosine similarity

is defined in Equation 1,

$$cos\_sim(x, y) = \frac{x \cdot y}{\|x\|_2 \times \|y\|_2} \tag{1}$$

where x and y are the vector representations of the generated and gold-standard answers, and with $x \cdot y$ the dot product of x and y. We refer to the cosine similarity estimated by Mistral-7B as "predicted similarity", while the actual computed cosine similarity serves as the reference measure.

3.3.2 Equivalence Testing with TOST

To determine whether Mistral-7B's self-assigned similarity scores are statistically equivalent to the actual cosine similarity, we employ the Two One-Sided Test (TOST) (Jones, 1952) for equivalence testing. The test evaluates whether the difference between the two similarity scores falls within an acceptable equivalence range. We set a tolerated error threshold of $\varepsilon = 0.3$, based on prior work (Liu *et al.*, 2024; Savage *et al.*, 2024), where a similarity above 0.7 is commonly considered sufficient for correct answers.

We compute a 90 % confidence interval ($\alpha = 0.90$) for the true mean (the limit with probability 1 of the empirical mean when the number of experiments tends to infinity) difference between the predicted similarity and the cosine similarity. If this confidence interval falls entirely within the equivalence range $[-0.3, 0.3]$, the test concludes that the predicted similarity scores are statistically indistinguishable from the true cosine similarity.

Mathematically, the test is conclusive if $I_\alpha \subseteq [-\varepsilon, +\varepsilon] = [-0.3, 0.3]$. If the test is conclusive, the difference between Mistral-7B's estimated similarity and the actual similarity is considered negligible, indicating that the model's confidence scores are well-calibrated.

3.3.3 Evaluating Discrimination and Calibration

We assess the quality of Mistral-7B's self-assigned similarity scores using two metrics:

Discrimination: the model's ability to distinguish correct answers from incorrect ones, measured by the correlation coefficient between the predicted similarity and the actual cosine similarity. A higher correlation indicates that the model reliably differentiates good answers from poor ones.

Calibration: how well the predicted similarity scores reflect actual similarity, measured using the mean squared error (MSE) between the predicted similarity and the cosine similarity. A lower MSE indicates better calibration, meaning the model's confidence estimates closely match reality.

By jointly analyzing equivalence testing, discrimination, and calibration, we look for providing a comprehensive evaluation of Mistral-7B's ability to assess its own confidence accurately.

4 Experiments

4.1 Dataset

We used MedQuAD [2] for the development of our experiments, a dataset that comprises 47,457 medical question-answer pairs sourced from 12 National Institutes of Health (NIH) websites. It contains 37 question types, such as Treatment, Diagnosis, among other medical entities. For our approach, we randomly selected a subset of 1k questions to be asked to the model.

4.2 Experimental Setup

We conducted our experiments under two different settings:

Baseline: the model was prompted to both answer the question and provide a similarity score, relying solely on its internal knowledge without external context.

Context-Enriched (RAG-Augmented): The prompt was supplemented with two additional context sentences retrieved from the dataset using semantic search. This setup allowed us to assess the impact of additional context on both answer quality and the model's confidence in its predictions.

We used the following prompt to describe the model the tasks to accomplish:

From the MedQuad MedicalQA Dataset: Given the medical question and question type, provide an accurate answer.
1. Provide the most accurate answer based solely on your medical knowledge.
2. Estimate a similarity score (0 to 1) representing how closely you believe your answer aligns with the real correct answer (which you do NOT have access to). This is a measure of your confidence in your answer's accuracy and completeness.
IMPORTANT: Only provide the answer and the estimated similarity score. Do NOT include any sections about the correct answer or additional commentary.
- Question Type:
- Question:

To retrieve relevant context in the RAG-Augmented setting, we used the *all-mpnet-base-v2*[3] model for sentence embedding, considering only the answer texts. For each query, we retrieved both the gold answer and the most similar answer based on cosine similarity. We added the answer texts in the prompt next to the *Question Type* line, replacing the description of the type of question, and without specifying that we were adding context related to the question.

For inference, we used the Mistral-7B model with 4-bit quantization to optimize memory usage and inference speed. Specifically, the model was loaded using the *BitsAndBytesConfig* with the following settings: *load_in_4bit*=True, *bnb_4bit_use_double_quant*=True, *bnb_4bit_quant_type*="nf4", and *bnb_4bit_compute_dtype*=torch.bfloat16. Outputs were generated using greedy decoding with *max_new_tokens*=1000 and a *repetition_penalty* of 1.15 to reduce redundancy. The model's response was post-processed using regular expressions to extract the generated answer and the associated similarity score.

[2] https://huggingface.co/datasets/keivalya/MedQuad-MedicalQnADataset
[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

4.3 Main results

Figure 1 presents the histograms of predicted and cosine similarity scores obtained, both with and without RAG. The baseline model's predicted scores (top-left) are highly concentrated around 0.8–0.9, suggesting consistent overconfidence. In contrast, the actual cosine similarity scores (top-right) are more widely distributed, reflecting varying response quality. With RAG, predicted similarity (bottom-left) remains skewed toward high values, with even more responses above 0.9, while cosine similarity (bottom-right) shifts higher, indicating improved answer accuracy. However, Mistral-7B's confidence calibration remains inadequate, as it fails to adjust its predicted similarity scores to account for the improvements from retrieved knowledge.

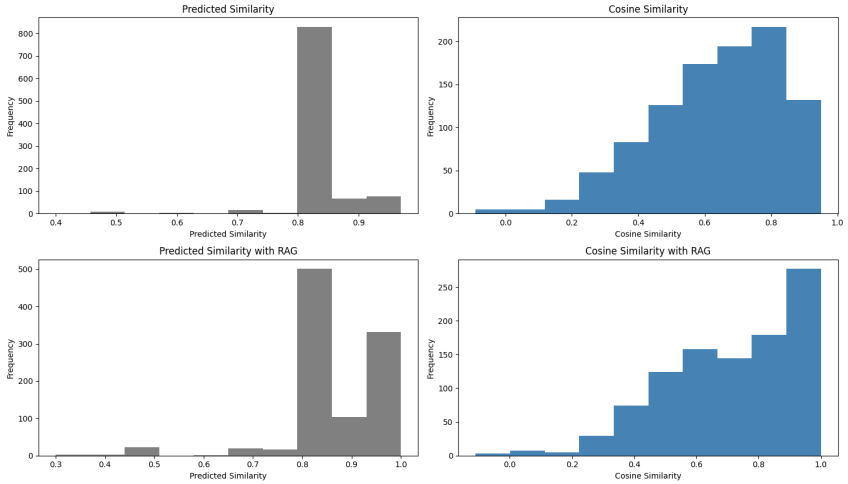

Figure 1: (Left side) Histograms of the similarities predicted by the model. (Right side) Histograms of the cosine similarities obtained between the predicted answers and the gold data.

Figures 2 and 3 show the TOST results for equivalence between Mistral-7B's predicted and actual similarity scores. Without RAG, the mean difference is -0.198 (90% confidence interval: $[-0.206, -0.191]$), while with RAG, it improves to -0.137 ($[-0.144, -0.129]$). In both cases, confidence intervals fall within the equivalence bounds $[-0.3, 0.3]$, indicating that the discrepancy is not statistically significant. Notably, the use of RAG leads to a smaller mean difference, suggesting a modest improvement in alignment between Mistral-7B's confidence estimates and actual response quality. However, the persistent negative bias indicates that the model still systematically overestimates its answer quality.

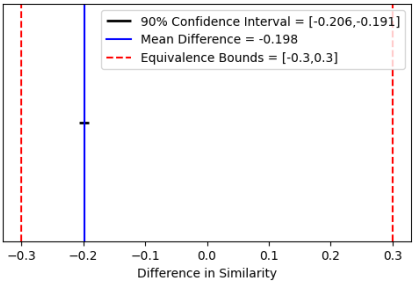

Figure 2: TOST between cosine similarities and predicted similarities.

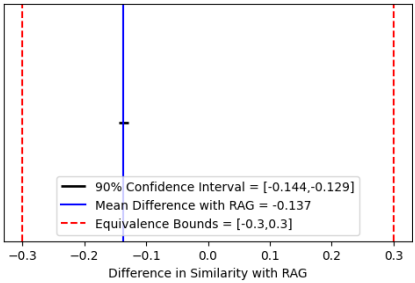

Figure 3: TOST between cosine similarities and predicted similarities with RAG.

Figure 4 illustrates the relationship between Mistral-7B's predicted similarity and the actual cosine similarity, highlighting areas of overconfidence (orange), underconfidence (blue), and acceptable confidence (gray). The left plot corresponds to the baseline model, while the right plot presents the results with RAG. The dashed black line represents the ideal scenario where predicted similarity perfectly matches cosine similarity, while the shaded regions indicate deviations beyond the predefined threshold of ± 0.3. Without RAG, the model exhibits a weak correlation ($r = 0.182$) between predicted and actual similarity, with a strong tendency towards overconfidence, as seen in the large cluster of points in the overconfidence region. Many predictions are concentrated near the upper bound (0.8–1.0), even when the actual cosine similarity is lower, suggesting that Mistral-7B systematically overestimates the quality of its responses. With RAG, the correlation improves ($r = 0.39$), indicating a stronger alignment between predicted and actual similarity. Additionally, the number of overconfident predictions decreases, and more points fall within the acceptable confidence region, showing better calibration. However, underconfidence remains an issue, as reflected by the small number of points in the underconfidence area.

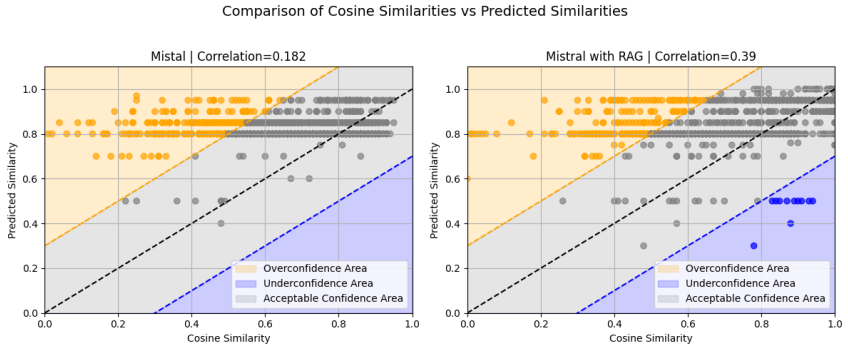

Figure 4: The threshold for overconfidence and underconfidence is fixed at +0.3 and -0.3

Overall, while RAG enhances correlation and reduces overconfidence, it does not fully resolve the calibration issue. The model still lacks a precise confidence estimation mechanism, and further adjustments, such as explicit confidence calibration techniques, may be necessary to improve reliability in practical medical applications.

5 Conclusion

Our study evaluates Mistral-7B's ability to assess the quality of its own answers by comparing its self-assigned similarity scores to an objective measure. Our findings reveal that RAG enhances the quality of the model's responses. However, the model tends to be overconfident in its predictions, with this effect increasing when additional context from RAG is provided. Despite this overconfidence, there is a meaningful correlation between the model's predicted similarity scores and cosine similarity, suggesting that its confidence scores follow a consistent pattern rather than being assigned randomly. These insights contribute to understanding Mistral-7B's self-assessment capabilities and highlight areas for improving confidence calibration.

It is important to note, however, that these results are specific to the prompt used and to the Mistral-7B model. The scope of our experiments was limited due to resource constraints, which means further investigation is needed to generalize these findings across other prompts, models, or tasks.

Future Work

Future work will address several key areas to enhance our study. First, we plan to incorporate multiple similarity measures and additional evaluation metrics to provide a more comprehensive assessment of the model's confidence. Second, we will extend our experiments beyond a single model by exploring other architectures, particularly those pre-trained in the biomedical domain, to evaluate generalizability. Third, we aim to expand our analysis to multiple biomedical tasks, allowing for broader insights into self-assessment capabilities. Finally, we will investigate the impact of domain-specific training by evaluating fine-tuned versions of Mistral-7B to determine whether such adaptations improve confidence calibration.

References

ABDAR M., POURPANAH F., HUSSAIN S., REZAZADEGAN D., LIU L., GHAVAMZADEH M., FIEGUTH P., CAO X., KHOSRAVI A., ACHARYA U. R. *et al.* (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, **76**, 243–297.

CHEN J. & MUELLER J. (2024). Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 5186–5200.

GAWLIKOWSKI J., TASSI C. R. N., ALI M., LEE J., HUMT M., FENG J., KRUSPE A., TRIEBEL R., JUNG P., ROSCHER R. *et al.* (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, **56**(Suppl 1), 1513–1589.

GENG J., CAI F., WANG Y., KOEPPL H., NAKOV P. & GUREVYCH I. (2023). A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*.

JONES L. V. (1952). Test of hypotheses: one-sided vs. two-sided alternatives. *Psychological Bulletin*, **49**(1), 43.

KADAVATH S., CONERLY T., ASKELL A., HENIGHAN T., DRAIN D., PEREZ E., SCHIEFER N., HATFIELD-DODDS Z., DASSARMA N., TRAN-JOHNSON E. *et al.* (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

LI Z., XIONG J., YE F., ZHENG C., WU X., LU J., WAN Z., LIANG X., LI C., SUN Z. *et al.* (2024). Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. *arXiv preprint arXiv:2410.02719*.

LIN S., HILTON J. & EVANS O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

LIU L., PAN Y., LI X. & CHEN G. (2024). Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

NG K. K. Y., MATSUBA I. & ZHANG P. C. (2025). Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, **2**(1), AIra2400380.

SAVAGE T., WANG J., GALLO R., BOUKIL A., PATEL V., AHMAD SAFAVI-NAINI S. A., SOROUSH A. & CHEN J. H. (2024). Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv*, p. 2024–06.

SONG M., SIM S. H., BHARDWAJ R., CHIEU H. L., MAJUMDER N. & PORIA S. (2024). Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *arXiv preprint arXiv:2409.11242*.

STUREBORG R., ALIKANIOTIS D. & SUHARA Y. (2024). Characterizing the confidence of large language model-based automatic evaluation metrics. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 76–89.

XU Z., JAIN S. & KANKANHALLI M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

YADKORI Y. A., KUZBORSKIJ I., GYÖRGY A. & SZEPESVÁRI C. (2024). To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

YARIE L., SORIANO D., KACZMAREK L., WILKINSON B. & VASQUEZ E. (2024). Mitigating token-level uncertainty in retrieval-augmented large language models. *Authorea Preprints*.

YUAN Y., LIU C., YUAN J., SUN G., LI S. & ZHANG M. (2024). A hybrid rag system with comprehensive enhancement on complex reasoning. *arXiv preprint arXiv:2408.05141*.