Confidence Calibration for Multimodal LLMs: An Empirical Study through Medical VQA

Yuetian Du[1], Yucheng Wang[2], Ming Kong[1], Tian Liang[1], Qiang Long[3], Bingdi Chen[3,4], and Qiang Zhu[1] (✉)

[1] College of Computer Science and Technology, Zhejiang University
[2] School of Computer Science and Technology, Xidian University
[3] Zhihui Medical Technology (Shanghai) Co., Ltd., Shanghai, China
[4] The Institute for Biomedical Engineering & Nano Science, Tongji University
{duyuetian2002, zjukongming, liangtian2022, zhuq}@zju.edu.cn
wangyucheng2004@stu.xidian.edu.cn
alex.long@petctc.com
inanochen@tongji.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) show great potential in medical tasks, but their elicited confidence often misaligns with actual accuracy, potentially leading to misdiagnosis or overlooking correct advice. This study presents the first comprehensive analysis of the relationship between accuracy and confidence in medical MLLMs. It proposes a novel method that combines Multi-Strategy Fusion-Based Interrogation (MS-FBI) with auxiliary expert LLM assessment, aiming to improve confidence calibration in Medical Visual Question Answering (VQA). Experiments demonstrate that our method reduces the Expected Calibration Error (ECE) by an average of 40% across three Medical VQA datasets, significantly enhancing MLLMs' reliability. The findings highlight the importance of domain-specific calibration for MLLMs in healthcare, offering a more trustworthy solution for AI-assisted diagnosis.

Keywords: Medical Visual Question Answering · Confidence Calibration · Multimodal Large Language Models

1 Introduction

Multimodal Large Language Models (MLLMs) [10,14,13,4] have demonstrated significant application potential across various fields, due to their exceptional ability to integrate textual and visual information. However, these models commonly exhibit over-confidence [28,12] in their predictions in practical applications, where there is a notable discrepancy between the confidence assigned to their predictions and their actual accuracy. This issue is particularly pronounced in high-stakes scenarios such as medical diagnosis. For instance, during clinical diagnosis, doctors typically rely on an iterative cycle of hypothesis generation, testing, and validation, integrating multi-source information such as patient history, laboratory results, and imaging data for comprehensive judgment. When

MLLMs are introduced as decision-support tools, the confidence of their predictions must be precisely calibrated to ensure that doctors can effectively utilize the model's recommendations. Respectively, over-confidence in erroneous predictions can lead to misdiagnosis, while under-confidence may cause doctors to overlook correct advice. Therefore, calibrating the confidence of MLLMs in Medical Visual Question Answering (Medical VQA) [7,9,11] tasks to ensure their reliability has become an urgent need in the practical application of AI-assisted diagnosis.

Although significant progress has been made in confidence calibration methods for Large Language Models (LLMs), the application of these methods in multimodal scenarios, especially in the medical field, is still in its infancy. The motivation for this study stems from the increasing prevalence of MLLMs in medical applications and the practical need for these models to provide accurate and reliable confidence estimates. By developing and evaluating confidence calibration methods specifically tailored for medical MLLMs, we aim to enhance the credibility of these models in clinical settings, thereby assisting healthcare professionals in making more accurate diagnostic decisions.

Related Work. In the field of confidence calibration for LLMs, there has been considerable research. [5,27,15] Early methods primarily focused on using LLMs' token-likelihoods [2,8,21,1] as confidence for calibration. However, as such white-box methods are not applicable to closed-source commercial models like ChatGPT, recent research has gradually shifted towards calibrating the verbalized confidence [28,17] of LLMs, which has been empirically shown to achieve better calibration effects [24] compared to token-likelihood confidence. Specific calibration methods include temperature scaling [21,26,1], prompting strategy [28,17,24,22], and reinforcement learning-based methods [8,23]. Additionally, self-consistency methods based on repeated sampling [20] or multi-sampling at different temperatures [28] have also proven effective in calibrating LLMs. Among these, prompting strategy has become the most widely used calibration method due to its plug-and-play nature and good portability. By designing a series of strategy templates, prompting strategy can effectively mitigate MLLMs' over-confidence issue, thus this study also adopts this method. It is noteworthy that in the medical field, limited related research [20,25] has mainly focused on the calibration of LLMs, while multimodal calibration, especially in Medical VQA, has not received sufficient attention. Although calibration strategies for LLMs can be transferred to multimodal scenarios, the effectiveness of these methods in the medical field has not been widely evaluated.

The main contributions of this study include the following three aspects:

- **Empirical Study:** We conducted the first comprehensive empirical study on the relationship between the accuracy of MLLMs and their self-assessed confidence in Medical VQA tasks, providing important insights into the calibration needs of multimodal LLMs in high-risk application scenarios.
- **New Calibration Method:** We proposed a new calibration method that combines a **M**ulti-**S**trategy **F**usion-**B**ased **I**nterrogation (MS-FBI) system

with an auxiliary expert LLM assessment framework to better calibrate confidence scores with actual accuracy.

- **Experimental Validation:** We applied this method alongside various LLM baseline methods to a Medical VQA dataset, and the results show that our method significantly outperforms existing technologies, with an average improvement of 40% in Expected Calibration Error (ECE), demonstrating the effectiveness of this method in real-world medical applications.

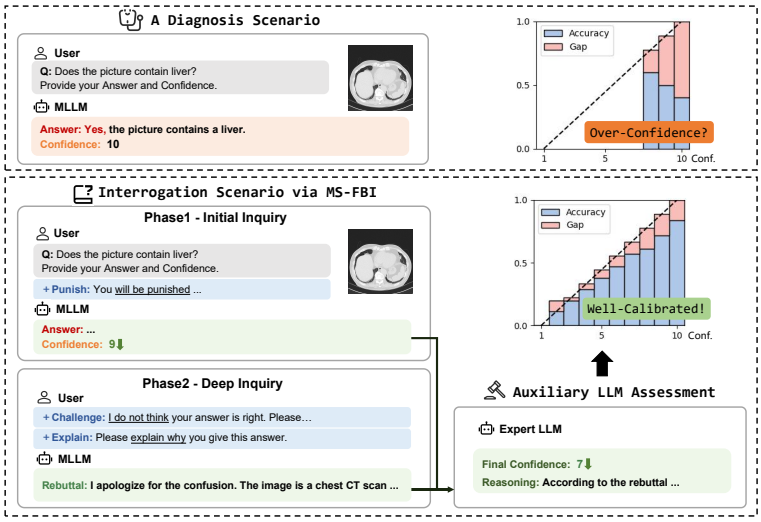

Fig. 1. The MLLM initially overconfidently identifies a liver in a chest CT scan. Through a two-phase interrogation process (MS-FBI), including an initial inquiry and deep inquiry with expert LLM assessment, the model's confidence is adjusted to a well-calibrated level.

2 Method

2.1 Overview

As shown in Figure 1, the proposed method consists of two core components: an interrogation system based on a two-phase multi-strategy fusion approach (MS-FBI, down left), for collecting empirical information; and an auxiliary expert LLM assessment framework (down right), offering final judgement and analysis. This workflow not only achieves calibration between confidence and accuracy but also reveals MLLMs' psychological state under different interrogation scenarios. Through this systematic approach, we obtain auxiliary qualitative reasoning information about the MLLM's behavior, which provides potential opportunities to further enhance question-answering accuracy through the calibration process.

2.2 Multi-Strategy Fusion-Based Interrogation (MS-FBI)

Inspired by the lie detection process in criminal interrogations (FBI, e.g.) [16], this study designs a interrogation system that integrates multiple strategies based on prompt engineering, constructing a dynamically adaptive interrogation scenario, and systematically capturing the deep cognitive information of MLLMs.

Initial Inquiry Phase—Punishment Mechanism. In the first round of interaction, the system requires the MLLM to simultaneously output the answer A and its confidence score C for a Medical VQA question Q. This phase introduces a penalty constraint mechanism ($Punish$) by conditionally embedding the prompt "You will be punished if the answer is wrong but you answer it with high confidence." This strategy aims to curb MLLMs' overconfidence tendency and encourage more cautious and accurate answers. With sampled and comparable MLLM outputs, this design not only simulates the error cost mechanism in real-world decision-making but also effectively evaluates MLLMs' confidence calibration characteristics under pressure.

Deep Inquiry Phase—Dual Verification Strategy. In subsequent interactions, the system adopts a dual-track verification mechanism of $Challenge$ and $Explain$. By embedding prompts behind the context of the previous round of dialogue, the system obtains MLLMs' rebuttal R_{mllm}. Logical challenges expose contradictions in MLLMs' reasoning chain through targeted questioning, forcing it to review and correct its answers. This strategy draws on the core principles of cognitive restructuring. Meanwhile, the explanation reinforcement requires MLLMs to provide a step-by-step interpretation of its answer generation process, revealing potential knowledge gaps or logical flaws. The synergistic application of these two strategies significantly enhances the detection efficacy of deceptive responses. By combining $Challenge$ and $Explain$ strategies, the interrogation process collects more detailed information about MLLMs' reasoning, which is crucial for accurately detecting overconfidence behaviors.

Strategy Combination. The overall prompt design references existing templates from relevant literature [17]. In practical application scenarios, the system provides an expandable strategy combination space: the activation state of the punishment mechanism constitutes a binary choice, while at least one of the $Challenge$ and $Explain$ strategies must be activated, resulting in six combination modes (2×3). This modular design enables multi-dimensional exploration of the MLLMs' cognitive boundaries [8,12] through strategy combinations and can be innovatively adjusted according to specific research needs.

2.3 Auxiliary Expert LLM Assessment

In this section, we template the previously collected information (Q, A, C, R_{mllm}) and input it into an expert LLM (llama3-instruct-8B, in our practice) for providing calibrated evaluation of the MLLMs' responses. The expert LLM's output includes the reassessed confidence C_r adjusted based on the former interrogation process, as well as the reasoning R_{exp} for C_r that helps better understand the MLLMs' behavior. This information can be used to comprehensively

evaluate the MLLMs' performance and identify areas where MLLMs may be prone to errors or inconsistencies. By comparing C and C_r, MLLMs' introspective ability can be quantified, while the R_{exp} generated by the expert LLM reveals the MLLMs' decision bias patterns, providing interpretable insights for model optimization.

3 Experiments and Results

3.1 Experimental Setup

Datasets. We extensively evaluate our proposed method on three public medical VQA datasets:

- **Med-VQA** [7]: contains over 4,000 medical images and 50,000 question-answer pairs.
- **VQA-RAD** [9]: includes 315 radiological images annotated by clinicians and 3,515 question-answer pairs.
- **SLAKE** [11]: a semantically annotated knowledge-enhanced medical VQA dataset with 642 images and 14,000 bilingual question-answer pairs.

Ultimately, we select 1,179 closed-ended questions from the test sets of the three datasets as a benchmark for evaluating MLLMs' capabilities and self-awareness.

MLLM Backbones. We test three MLLMs to evaluate their performance on medical VQA tasks under various calibration methods (including ours):

- **LLaVA-1.5-Med-Mistral-7b** [10]: a multimodal large model specifically designed for the biomedical field, demonstrates excellent cross-modal understanding abilities.
- **LLaVA-NeXT-Mistral-7b** [14]: an upgraded version of LLaVA-1.5 [13], shows significant improvements in visual dialogue and reasoning capabilities.
- **Molmo-7b** [4]: a general-purpose MLLM developed by Ai2, and its smaller parameter model outperforms models with 10 times more parameters.
- **MedVLM-R1** [18]: a recent medical reasoning model that leverages reinforcement learning, specifically the GRPO algorithm, achieves strong results in medical VQA.

Metrics. We evaluate MLLM calibration using two core metrics. Expected Calibration Error **(ECE)** [6] quantifies the deviation between model confidence and predictive accuracy by partitioning confidence scores into M equally spaced bins B_m $(m = 1, \ldots, M)$. It calculates the weighted difference between average accuracy and confidence within each bin. Due to its intuitiveness, it is also widely regarded as the primary metric for assessing the calibration degree of models. The metric is formulated as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{1}$$

Area Under Receiver Operating Characteristic Curve **(AUROC)** [3] assesses confidence scores' diagnostic capability in distinguishing correct predictions via ROC curve integration. The metric is formulated as:

$$\text{AUROC} = \frac{1}{|C^+||C^-|} \sum_{C_i \in C^+} \sum_{C_j \in C^-} \mathbb{I}(C_i > C_j) \tag{2}$$

where C_i and C_j denote confidence scores for correct and incorrect predictions respectively, and $\mathbb{I}(\cdot)$ is an indicator function returning 1 when $C_i > C_j$ and 0 otherwise.

Table 1. Performance comparison of different calibration methods across MLLMs and medical VQA datasets (Values are converted to percentages, and the optimal and suboptimal results in each column are highlighted in bold and underlined, respectively. The same applies to Table 2.)

Model	Method	med-vqa		vqa-rad		slake		Avg.	
		ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑
Llava-1.5-med-7B	Vanilla	47.71	48.46	39.44	50.63	45.70	49.36	44.28	49.48
	Punish	46.68	48.47	38.31	51.30	42.43	49.55	42.47	49.77
	Top-K	43.85	50.00	42.41	50.00	41.69	50.00	42.65	50.00
	Ours	**27.89**	**55.70**	**21.77**	**56.80**	**29.00**	**52.70**	**26.22**	**55.07**
Llava-NeXT-7B	Vanilla	25.94	57.01	35.16	54.56	37.00	**57.48**	32.70	56.35
	Punish	25.58	58.28	36.37	55.85	37.04	56.65	33.00	56.93
	Top-K	27.80	47.43	33.69	54.03	36.88	53.25	32.79	51.57
	Ours	**13.80**	**58.44**	**20.90**	**56.88**	**19.52**	55.73	**18.07**	**57.02**
Molmo-7B	Vanilla	25.08	51.45	30.28	49.87	26.87	**54.45**	27.68	51.92
	Punish	23.54	51.69	32.47	46.98	29.35	50.19	30.91	49.62
	Top-K	39.28	45.46	31.04	**56.47**	48.71	39.48	39.68	47.14
	Ours	**14.52**	**62.20**	**24.52**	52.39	**20.22**	51.25	**19.75**	**55.28**
MedVLM-R1	Vanilla	20.82	49.15	25.59	54.95	30.46	44.81	25.62	49.64
	Punish	24.58	49.71	27.63	57.59	33.44	42.38	28.55	49.89
	Top-K	13.24	62.51	28.91	**59.29**	16.76	**60.59**	19.64	**60.80**
	Ours	**12.52**	**63.32**	**16.85**	59.03	**14.06**	55.62	**14.48**	59.32

Baselines. We test the performance of three LLM confidence calibration methods:

- **Vanilla** [28]: directly extracts the model's confidence in the verbal answer. Its advantage is that it provides a basic confidence indicator without additional computation.
- **Punish** [17]: adds a prompt "You will be punished if the answer is wrong but you answer it with high confidence" to encourage the model to be cautious in answering.
- **Top-K** [24]: requires the model to provide k best guesses $\{G_1, \ldots, G_k\}$ with corresponding probabilities $\{P_1, \ldots, P_k\}$, eliciting verbalized likelihoods. For fair comparison across methods, we apply computational normalization by

scaling the maximum probability to confidence through linear transformation:

$$\text{Confidence} = \max\{P_1, \ldots, P_k\} \times 10. \tag{3}$$

3.2 Main Results

As shown in Table 1, we evaluate four confidence calibration methods for the medical VQA task through comparative experiments. Notably, our proposed method outperforms the baseline methods across all the medical VQA datasets and MLLMs, reducing the average ECE of LLaVA-1.5-med-7B from 44.28% to 26.22% (a 40.8% reduction) and increasing AUROC from 49.48% to 55.07%. This demonstrates significant improvement in both calibration and the model's ability to distinguish between correct and incorrect answers. Moreover, our findings show that domain-specific characteristics significantly affect calibration performance. General-domain MLLMs, like Molmo-7B, exhibit low ECE across calibration methods, with its vanilla method achieving an average ECE of 27.68%. In contrast, medical-domain MLLMs, such as LLaVA-1.5-med-7B, show persistent overconfidence, with an average ECE of 44.28%, much higher than general-domain MLLMs. This suggests that supervised fine-tuning (SFT) on domain-specific data may impair confidence calibration [19].

Table 2. Performance comparison of different strategy combinations.

Method	Llava-1.5-med-7B			Llava-NeXT-7B			Molmo-7B		
	med-vqa	vqa-rad	slake	med-vqa	vqa-rad	slake	med-vqa	vqa-rad	slake
Ch.	32.36	21.52	<u>27.47</u>	14.04	**19.21**	<u>20.31</u>	<u>17.00</u>	<u>25.16</u>	<u>22.09</u>
Exp.	34.46	21.53	35.93	16.92	26.49	26.15	27.72	28.03	29.57
Ch.+Exp.	30.51	<u>20.82</u>	30.74	14.42	21.04	22.19	21.80	25.95	25.96
Pu+Exp.	34.00	22.29	29.72	14.89	25.32	24.33	26.58	27.06	27.07
Pu.+Ch.	**27.89**	21.77	29.00	**13.80**	<u>20.90</u>	**19.52**	**14.52**	**24.52**	**20.22**
Pu+Ch+Exp.	<u>29.28</u>	**17.90**	**25.77**	<u>13.88</u>	22.37	21.54	20.52	25.45	23.12

3.3 Ablation Study

Ablation experiments reveal that strategy combinations significantly impact the calibration performance of medical visual question answering models, measured by ECE. The *Punish* only approach (Pu.) was excluded from the table as it omitted the model's rebuttal phase, making the expert model's calibrated confidence scores unreliable. The *Punish* and *Challenge* (Pu.+Ch.) combination proved most effective on the Llava-1.5-med-7B model, with an ECE of 27.89%, and also demonstrated optimal performance in cross-model scenarios, reducing Molmo-7B's ECE by 14.6%. However, strategy effectiveness is not tied to complexity, as combining all three methods yielded suboptimal results.

Furthermore, the experiments show cross-model heterogeneity: the single *Challenge* (Ch.) strategy was superior for the Llava-NeXT-7B model on the

vqa-rad dataset, while Pu.+Ch. performed best on the slake dataset for multiple models. This variation stems from pre-training differences, where overconfident models like Llava-Med benefit from multi-strategy intervention, while more capable models like Molmo-7B perform better with simpler strategies.

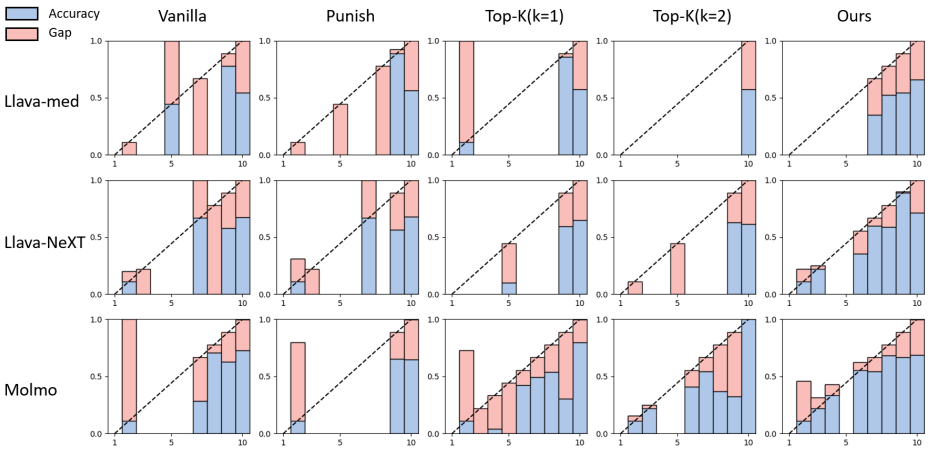

Fig. 2. Visualization analysis of the confidence (x-axis) vs. accuracy (y-axis) calibration comparison across different baselines (including ours), with all datasets aggregated.

3.4 Visualization and Analysis

Figure 2 demonstrates a comparison of the calibration effects between different calibration methods across three datasets, with confidence (x-axis) and accuracy (y-axis). The visual analysis reveals significant differences in calibration characteristics under various models (Llava-med, Llava-NeXT, and Molmo) for each calibration method. Specifically, the Llava-med model exhibits a clear calibration underperformance under the Vanilla method, showing a significant discrepancy between confidence and accuracy. In contrast, our MS-FBI method substantially reduces the gap between confidence and accuracy, showing superior calibration performance. Similar trends are observed in the Llava-NeXT and Molmo models, confirming the method's advantage in cross-model generalization. Overall, the method proposed in this paper demonstrates the best calibration effects across all models and datasets, effectively enhancing the consistency between model confidence and accuracy.

4 Discussion and Conclusion

This study is the first to combine multi-strategy fusion-based interrogation with expert LLM evaluation framework to explore the calibration of medical MLLMs,

achieving significant progress. The experiments show that the proposed method effectively improves the alignment between model confidence and prediction accuracy, and reveals MLLM's decision-making patterns in complex medical scenarios through a multi-round interrogation mechanism. Ablation experiments confirm the contributions of optimization strategies and expert LLMs to the calibration effect, providing support for enhancing model reliability.

Although the generalizability of the method has been preliminarily validated, more refined calibration strategies may be needed in the case of general LLMs. The study also highlights key future research directions, such as domain-specific calibration for medical models, unification of expert evaluation standards, and achieving efficient calibration without increasing computational costs. Addressing these issues will advance MLLM calibration technologies and provide more reliable intelligent tools for clinical decision support systems.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064 and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ahuja, K., Sitaram, S., Dandapat, S., et al.: On the calibration of massively multilingual language models. arXiv preprint (2022), https://arxiv.org/abs/2210.12265, arXiv:2210.12265
2. Andrey, M., Mark, G.: Uncertainty estimation in autoregressive structured prediction. arXiv preprint (2020), https://arxiv.org/abs/2002.07650, arXiv:2002.07650
3. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145–1159 (1997)
4. Deitke, M., Clark, C., Lee, S., et al.: MOLMO and PIXMO: Open weights and open data for state-of-the-art multimodal models. arXiv preprint (2024), https://arxiv.org/abs/2409.17146, arXiv:2409.17146
5. Geng, J., Cai, F., Wang, Y., et al.: A survey of confidence estimation and calibration in large language models. arXiv preprint (2023), https://arxiv.org/abs/2311.08298, arXiv:2311.08298
6. Guo, C., Pleiss, G., Sun, Y., et al.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
7. Hasan, S.A., Ling, Y., Farri, O., et al.: Overview of ImageCLEF 2018 medical domain visual question answering task. In: Proceedings of CLEF Working Notes (2018)
8. Kadavath, S., Conerly, T., Askell, A., et al.: Language models (mostly) know what they know. arXiv preprint (2022), https://arxiv.org/abs/2207.05221, arXiv:2207.05221
9. Lau, J., Gayen, S., Ben Abacha, A., et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(180251) (2018)
10. Li, C., Wong, C., Zhang, S., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems. vol. 36 (2024)

11. Liu, B., Zhan, L.M., Xu, L., et al.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: IEEE 18th International Symposium on Biomedical Imaging. pp. 1650–1654 (2021)
12. Liu, G., Wang, X., Yuan, L., et al.: Examining LLMs' uncertainty expression towards questions outside parametric knowledge. arXiv preprint (2023), https://arxiv.org/abs/2311.09731, arXiv:2311.09731
13. Liu, H., Li, C., Li, Y., et al.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
14. Liu, H., Li, C., Li, Y., et al.: LLaVA-Next (2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
15. Liu, Y., Yao, Y., Ton, J.F., et al.: Trustworthy llms: a survey and guideline for evaluating large language models' alignment. arXiv preprint (2023), https://arxiv.org/abs/2308.05374, arXiv:2308.05374
16. Manea, T.: Lie detection during the interview and interrogation process: A psychosocial criminal approach. Balkan Social Science Review **17**, 41–55 (2021)
17. Ni, S., Bi, K., Guo, J., et al.: When do LLMs need retrieval augmentation? Mitigating LLMs' overconfidence helps retrieval augmentation. arXiv preprint (2024), https://arxiv.org/abs/2402.11457, arXiv:2402.11457
18. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Cheng, O., Rueckert, D.: MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. arXiv preprint (2025), https://arxiv.org/abs/2502.19634, arXiv:2502.19634
19. Ren, Y., Sutherland, D.J.: Learning dynamics of llm finetuning. In: International Conference on Learning Representations (2025), iCLR
20. Savage, T., Wang, J., Gallo, R., et al.: Large language model uncertainty measurement and calibration for medical diagnosis and treatment. medRxiv (2024)
21. Si, C., Zhao, C., Min, S., et al.: Re-examining calibration: The case of question answering. arXiv preprint (2022), https://arxiv.org/abs/2205.12507, arXiv:2205.12507
22. Steyvers, M., Tejeda, H., Kumar, A., et al.: What large language models know and what people think they know. Nature Machine Intelligence pp. 1–11 (2025)
23. Tao, S., Yao, L., Ding, H., et al.: When to trust llms: Aligning confidence with response quality. arXiv preprint (2024), https://arxiv.org/abs/2404.17287, arXiv:2404.17287
24. Tian, K., Mitchell, E., Zhou, A., et al.: Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint (2023), https://arxiv.org/abs/2305.14975, arXiv:2305.14975
25. Wada, A., Akashi, T., Shih, G., et al.: Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. Diagnostics **14**, 1541 (2024)
26. Wang, C., Szarvas, G., Balazs, G., et al.: Calibrating verbalized probabilities for large language models. arXiv preprint (2024), https://arxiv.org/abs/2410.06707, arXiv:2410.06707
27. Wen, B., Yao, J., Feng, S., et al.: Know your limits: A survey of abstention in large language models. arXiv preprint (2024), https://arxiv.org/abs/2407.18418, arXiv:2407.18418
28. Xiong, M., Hu, Z., Lu, X., et al.: Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv preprint (2023), https://arxiv.org/abs/2306.13063, arXiv:2306.13063