AlignVQA: Debate-Driven Multi-Agent Calibration for Vision Language Models

Ayush Pandey [1], **Jai Bardhan** [1] , **Ramya Hebbalaguppe**[1]
[1]TCS Research
ayush.p4@tcs.com, jai.bardhan90@gmail.com, ramya.hebbalaguppe@tcs.com,

Abstract

In the context of Visual Question Answering (VQA) and Agentic AI, calibration refers to how closely an AI system's confidence in its answers reflects their actual correctness. While modern VQA systems, powered by advanced vision-language models (VLMs), are increasingly used in high-stakes domains like medical diagnostics and autonomous navigation due to their improved accuracy, the reliability of their confidence estimates remains under-examined. Particularly, these systems often produce overconfident responses. To address this, we introduce *AlignVQA*, a debate-based multi-agent framework, in which diverse specialized VLM – each following distinct prompting strategies – generate candidate answers and then engage in two-stage interaction: generalist agents critique, refine and aggregate these proposals. Furthermore, we introduce a novel differentiable calibration-aware loss function called *AlignCal* designed to fine-tune the specialized agents by minimizing an upper bound on the calibration error. This objective explicitly improves the fidelity of each agent's confidence estimates. Empirical results across multiple benchmark VQA datasets substantiate the efficacy of our approach, demonstrating substantial reductions in calibration discrepancies.

Introduction

Visual Question Answering (VQA) is a foundational task in multimodal artificial intelligence that requires models to jointly process visual content and natural language to generate accurate answers to open-ended questions about images. First introduced to connect vision and language for goal-oriented reasoning (Antol et al. 2015), VQA has evolved into a benchmark for evaluating systems' abilities in compositional reasoning, visual grounding, and language understanding (Agrawal et al. 2016).

Agentic architectures for VQA: Recent advancements in VQA have embraced agentic architectures, where multiple interacting agents collaboratively solve complex visual reasoning tasks. For instance, Jiang et al. (Jiang et al. 2024) introduced a zero-shot multi-agent system with specialized experts coordinated adaptively. Hu et al. (Hu et al. 2024) proposed a team of LLM-based agents with tool access, whose outputs are aggregated via voting. Wang et al. (Wang et al. 2023a) designed explainable agents with dedicated roles (Responder, Seeker, Integrator) that operate in a top-down reasoning loop.

Need for Calibration in VQA: Due to its practical relevance, VQA is increasingly being deployed in high-stake real-world domains such as medical diagnosis (Lin et al. 2023; Zhou et al. 2023; Canepa, Singh, and Sowmya 2023), autonomous navigation (Qian et al. 2024; Sima et al. 2025; Marcu et al. 2024; Atakishiyev et al. 2023), and assistive technologies for the visually impaired (Gurari et al. 2018; Chanana et al. 2017). In these settings, it is not only essential for VQA systems to be accurate, but also to be calibrated. A model is said to be calibrated if it's confidence matches the probablity of occurance (Guo et al. 2017). A calibrated model knows when to trust their predictions.

Calibration in SOTA VQA architectures: Several recent works have attempted to improve VQA calibration. Whitehead et al. (Whitehead et al. 2022) proposed a selective answering strategy where the model abstains when it is unsure. Mozaffari et al.'s GLEN framework (Mozaffari, Sapkota, and Yu 2025) introduced a combination of model simplification and focal loss to enhance calibration. IVON by Wieczorek et al. (Wieczorek et al. 2025) leveraged Bayesian variational fine-tuning to capture model uncertainty through posterior weight distributions.

Humans rarely make decisions in isolation—opinions evolve through discussion, critique, and consensus. Inspired by this; we introduce *AlignVQA*, a calibration method for MCQ VQA that draws on human-like deliberation through a structured multi-agent debate. Specialized agents produce initial answers, while generalist agents critique, revise, and update both predictions and confidences via iterative argument exchange; final decisions are aggregated by a confidence-aware mechanism to enhance robustness and calibration. Complementing this framework, we propose *AlignCal*, a differentiable calibration-aware loss that serves as a surrogate for minimizing an upper bound on miscalibration, jointly optimizing correctness and confidence.

Related Works

The common calibration techniques used in classification tasks include:(1.) **Train-time Calibration methods** aim to improve confidence estimates during the training phase by modifying the loss function. These methods generally

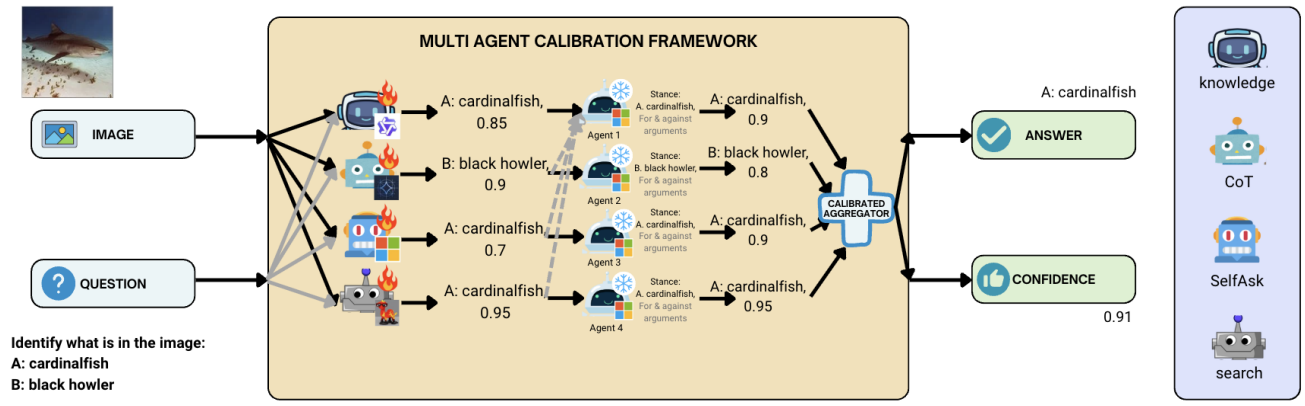


Figure 1: **AlignVQA Multi-Agent Calibration Model.** Given an input image and question, the model first queries a set of specialized agents—*Chain-of-Thought* (Wei et al. 2022), *Search-Augmented*, *SelfAsk* (Press et al. 2022), and *GENREAD Knowledge-based* (Yu et al. 2022) models—each fine-tuned for calibration using our custom proposed loss *AlignCal* . These agents independently produce answer classes (e.g., A: cardinalfish, B: black howler). In the second stage, a group of general agents is instantiated, with each agent probabilistically initialized to a specific answer class based on the distribution of predictions from the specialized agents.

smooth confidence scores in a sample-agnostic manner– applying regularization uniformly across samples. (2.) **Post-hoc Calibration** Post-hoc calibration methods adjust a fully trained model's confidence scores using a separate hold-out set.

Multi-Agent Calibration in LLM: Collaborative Calibration (Yang et al. 2024) was introduced as a multi-agent deliberation framework where agents share their predictions, confidence estimates, and the reasoning steps to engage in a simulated group dialogue. **Calibration in VQA** Utilizing a popular strategy of using consistency among samples to estimate confidence, Eisenschlos et al. (Eisenschlos et al. 2024) introduced a method for improving the reliability of visual question answering (VQA) models.

Multi-Agent approaches in VQA tasks Wang et al. (Wang et al. 2023b) proposed a multi-agent architecture for VQA that draws inspiration from the human process of top-down reasoning—where individuals leverage prior knowledge and contextual cues to infer new information (e.g., predicting rain from observing cloudy skies). To our knowledge, no prior work leverages multi-agent methods for calibration in VQA.

Proposed Methodology

Agent Ensemble and Stance Generation

Given an input image-question pair (for example, "Identify the type of fish in the picture" (c.f. Fig. 1), we first generate candidate answers from a diverse set of *specalized expert agents*. Each agent is created with a different VLM backbone – Qwen2.5-VL-3B-Instruct (Bai et al. 2025), Llava-Onevision(Li et al. 2024), Gemma 3 4B (Team et al. 2025) and Phi-4-multimodal-instruct (Abouelenin et al. 2025) – and a distinct prompting strategy to encourage diverse reasoning. [1] Specifically, we employ: Chain-of-Thought prompting (Wei et al. 2022) for multi-hop reasoning, Self-Ask prompting (Press et al. 2022) for recursive problem decomposition, Search-style prompting startegy to incorporate external retrieval cues and the GENREAD style prompting (Yu et al. 2022) for structured comprehension.

Each expert agent i (independently) produces an output $v_i = (\hat{y}_i, p_i)$, where $\hat{y}_i$ is the answer string and p_i is its sequence probability. We infer the confidence of the sequence through the geometric mean of probabilities of next-tokens generated. This serves as the initial confidence estimation for a candidate answer of a particular agent. We merge lexically different but semantically equivalent answers into K unique stances $\{s_1, \ldots, s_K\}$ $(K \leq N)$ using a GPT-3.5 judge (Tian et al. 2023). For each stance s_k we define,

$$\mathcal{I}_k = \{\, i : \hat{y}_i = s_k \,\}, \qquad f_k = |\mathcal{I}_k|, \qquad \bar{c}_k = \frac{1}{f_k} \sum_{i \in \mathcal{I}_k} c_i,$$

where I_k is the index set, f_k is frequency of stance s_k and c_i is the sequence probability of answer i.

Illustrated Failure Case. As illustrated in Fig. 1, in the first stage, three agents each give the answer *cardinal fish* with confidence scores of 0.85, 0.70, and 0.95, producing an average confidence of 0.83 for that stance, while a fourth agent answers *black howler* with confidence 0.90. A majority confidence based system would adopt *black howler* as the consensus after the first stage, despite it being incorrect and lacking group support. Stage 2, is designed to revisit and refine such consensus through deliberation and counter argumentation.

[1] This agentic framework is model agnostic, any set of VLM backbones can be substituted in place of those used here.

Architecture	VQARad Dataset							ScienceQA Dataset						
	↑Acc.	↑F_1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE	↑Acc.	↑F_1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE
Agentic Framework	65.70%	0.540	0.554	0.544	0.133	0.146	0.820	72.80%	0.340	0.346	0.328	0.265	0.270	0.438
Post-Hoc Calibration														
Agentic + TS	65.70%	0.540	0.554	0.544	0.114	0.117	0.765	72.80%	0.340	0.346	0.328	0.255	0.268	0.421
Agentic + DC	65.70%	0.540	0.554	0.554	0.097	**0.041**	**0.113**	72.80%	0.340	0.346	0.328	–	–	–
Train-Time Calib.														
Agentic + FL	**68.50%**	0.571	0.542	0.605	0.116	0.073	0.393	74.40%	0.424	0.480	0.381	0.142	0.180	0.678
Agentic + LS	67.70%	**0.650**	**0.652**	**0.650**	0.175	0.183	0.543	75.20%	0.467	0.532	**0.424**	0.186	0.186	0.916
Proposed Method														
Agentic+*AlignCal* +FL	68.20%	0.548	0.517	0.583	**0.095**	0.098	0.267	**76.10%**	**0.472**	**0.540**	0.418	**0.110**	**0.055**	**0.331**

Table 1: Comprehensive comparison of calibration strategies across VQARad and ScienceQA datasets. Bold values indicate best performance for each metric within each dataset, underlined values indicate second-best performance. The proposed method (Agentic + *AlignCal* + FL) demonstrates superior calibration performance with competitive accuracy across both datasets. It is not possible to perform Dirchelet Calibration in the case of ScienceQA Dataset due to the unavailability of the probabilities of other options.

Group Debate with Rationale and Feedback

The second stage introduces a set of **generalist deliberation agents** (no specialized prompting) whose role is to critically examine, defend and revise the candidate stances produced in Stage 1, forming a structured debate ensemble. To maintain the prior group consensus while allowing contrarian exploration, each generalist agent j is assigned a stance s_j by sampling proportionally to the frequencies f_k, i.e., $\Pr(s_j = s_k) \propto f_k$. This maintains a soft bias towards majority supported views while still allowing minority stances to be reconsidered.

Each agent then argues for its stance by exploring diverse reasoning and developing rationales for defending it. Each reasoning path is unique and develops an ensemble of rationales for a particular stance. Agents then provide ratings and feedback to each rationale in terms of logical consistency, factuality, clarity and conciseness. Specifically, Chain-of-Verification style prompting (Dhuliawala et al. 2024) is used to check the factuality by generating underlying premises or assumptions. These premises are then further checked with a search augmented agent to identify unfactual statements in the feedback.

Each general agent then receives a pair of arguments one sampled from the set of supporting arguments and one sampled from one of the opposing sides to form a debate pair. This mirrors two-sided deliberation paradigms in multi-agent reasoning and debate systems. Based on these arguments, each agent produces a final answer that incorporates the provided opposing argument, supporting argument and its previously assigned stance. Thus, the final answer is given by $y'_j = f_j(s_j, \bar{c}_j, a_p, a_n)$, where a_p, a_n are the supporting and opposing arguments with ratings and feedback, and $s_j, \bar{c}_j$ is the initial stance with its associated confidence assigned to agent j. We also record the sequence probability of each agent's final response y'_j, which serves as the refined confidence score $\mathrm{Conf}(y'_j)$.

After collecting refined outputs $\{(y'_j, \mathrm{Conf}(y'_j))\}_{j=1}^{M}$ from M generalist agents, we aggregate in two steps. First, for each stance s_k define the agent index set $\mathcal{I}'_k = \{\, j \mid y'_j = s_k \,\}$, set $f'_k = |\mathcal{I}'_k|$, and compute the mean refined confidence

$$\widehat{c}_k \;=\; \frac{1}{|\mathcal{I}'_k|} \sum_{j \in \mathcal{I}'_k} \mathrm{Conf}(y'_j).$$

The final answer is selected by majority vote:

$$s^* \;=\; \arg \max_{k \in \{1,\ldots,K\}} f'_k. \tag{1}$$

The final confidence is the mean confidence of the agents supporting the chosen stance. This aggregation yields a better indication of prediction, by weighing different arguments through deliberation. To further improve calibration, we introduce the following loss.

Calibration Aware Finetuning

Our system can benefit from better calibrated VLMs. Therefore, we introduce a novel surrogate loss function that directly minimizes a tight upper bound on miscalibration during training, thereby avoiding the pitfalls of post-hoc fixes. Classical metrics like ECE average confidence–accuracy gaps over broad bins so they often fail to estimate the reliability of a single test example. UBCE(Zhong et al. 2025) was proposed to overcomes this and capture per-instance miscalibration by averaging absolute gaps between confidence and correctness. We minimize a differentiable plug-in surrogate of UBCE, directly shrinking these gaps and thereby reducing ECE and improving MCE.

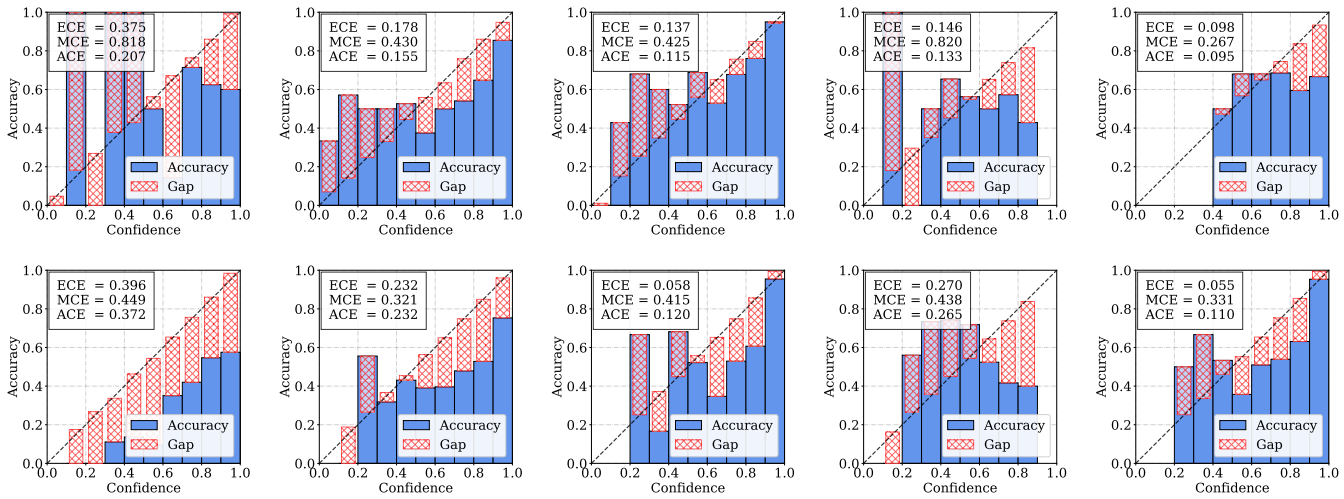
Our loss function *AlignCal*: Formally, given softmax outputs $\mathbf{p} = (p_1, \ldots, p_K)$ with logits z_i, true label y, top predicted confidence $p_{\max} = \max_i p_i$ and predicted ground truth class probability p_y, we define the soft-calibration loss:

$$\mathcal{L}_{\textit{AlignCal}}\,(p_y, p_{\max}) = p_y(1 - p_{\max}) + (1 - p_y)p_{\max} \tag{2}$$

The full training objective therfore becomes:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{FL}} + \lambda \mathcal{L}_{\textit{AlignCal}}\,,$$

where λ is a tuneable hyperparameter and $\mathcal{L}_{\text{FL}}$ is the focal loss (Mukhoti et al. 2020)



(a) Gemma 3 4B (b) FL (c) *AlignCal* + FL (d) Agentic Framework (e) Agentic + *AlignCal*

Figure 2: Reliability plots of the datasets, VQARad (top) and ScienceQA (bottom). 2a shows the calibration from base Gemma model. 2b shows plot on FL finetuned Gemma 3 4B model. 2c shows the plot on FL + *AlignCal* finetuned Gemma 3 4B model. 2d shows the plot obtained from Agentic Framework. 2e shows the plot obtained from Agentic framework where agents are finetuned with *AlignCal* + FL.

Dataset and Evaluation

Datasets and Models: We took 2 publicly available datasets, ScienceQA (Lu et al. 2022) and medical dataset VQARad (Lau et al. 2018). ScienceQA consists of 21,208 multimodal multiple choice questions with diversE science topics and annotations of their answers with corresponding lectures and explanations. VQA-RAD is manually constructed dataset in radiology where answers about images are naturally created and validated by clinicians. VQA-RAD dataset contains 3,515 total visual questions. We only consider Yes/No type of questions from this dataset.

Evaluation: We evaluate calibration with ECE (Guo et al. 2017), ACE, and MCE. To visualize miscalibration, we include reliability diagrams. For task performance, we additionally report Accuracy, F1-score, Precision, and Recall. Maximum Calibration Error (MCE) is the largest absolute difference between predicted confidence and empirical accuracy across all confidence bins. Adaptive Calibration Error (ACE) splits the sorted predictions into bins each containing an equal number of examples, then computes the mean absolute gap between empirical accuracy and average confidence across those bins.

Experiment and Results

Agentic Framework We report here the Agentic framework results. From Fig. 2, we can show that agentic VLM debate leads to better calibration and more reliable responses. On ScienceQA dataset ECE decreases from 0.396 to 0.270, while MCE and ACE also show consistent reductions from 0.449 to 0.438 and 0.372 to 0.265, respectively. In the VQARad dataset, ECE decreases from 0.375 to 0.146, ACE also shows consistent reduction 0.207 to 0.133.

***AlignCal* Results.** Fig. 2 reports the calibration improvements achieved by our loss function *AlignCal* on two VQA benchmarks. On the VQARad dataset, ECE decreases from 0.178 to 0.137, and ACE decreases from 0.155 to 0.115. Similarly, on the ScienceQA dataset, ECE is reduced from 0.232 to 0.058, while ACE falls from 0.232 to 0.120. We also compared our results with other training calibration methods focal loss (Lin et al. 2017) and label smoothing (Szegedy et al. 2016). For LS, we use $\alpha = 0.1$ and for FL, we use $\gamma = 2$.

Debate with Calibrated Agents From Table 1, we see that in the Agentic debate framework when the agents are finetuned with *AlignCal* loss, it leads to better calibration. ECE reduces from 0.375 to 0.098, MCE from 0.818 to 0.267 and ACE from 0.207 to 0.095 on VQARad dataset. On ScienceQA dataset ECE reduces from 0.396 to 0.055, ACE reduces from 0.372 to 0.110 and MCE from 0.449 to 0.331. The significant reduction in both across the VQARad and ScienceQA datasets—relative to using fine-tuning or debate in isolation—indicates that the calibrated-agents debate yields substantially more reliable confidence estimates and, more trustworthy answers. The supplementary material includes additional results and ablation studies.

Conclusion

In this work, we presented AlignVQA, a novel approach to improving confidence calibration in Visual Question Answering (VQA) through a multi-agent debate framework. By incorporating a multi-agent debate process and calibration-aware loss, AlignVQA yields more reliable answers with better-calibrated confidence scores. Future work will include reducing its inference overhead (e.g., via selective agent activation) and applying it to risk-aware applications.

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2016. VQA: Visual Question Answering. arXiv:1505.00468.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Atakishiyev, S.; Salameh, M.; Babiker, H.; and Goebel, R. 2023. Explaining Autonomous Driving Actions with Visual Question Answering. arXiv:2307.10408.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Canepa, L.; Singh, S.; and Sowmya, A. 2023. Visual Question Answering in the Medical Domain. arXiv:2309.11080.

Chanana, P.; Paul, R.; Balakrishnan, M.; and Rao, P. 2017. Assistive technology solutions for aiding travel of pedestrians with visual impairment. *Journal of rehabilitation and assistive technologies engineering*, 4: 2055668317725993.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578. Bangkok, Thailand: Association for Computational Linguistics.

Eisenschlos, J. M.; Maina, H.; Ivetta, G.; and Benotti, L. 2024. Selectively Answering Visual Questions. *arXiv preprint arXiv:2406.00980*.

Grunewalder, S. 2018. Plug-in Estimators for Conditional Expectations and Probabilities. In Storkey, A.; and Perez-Cruz, F., eds., *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 1513–1521. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. arXiv:1802.08218.

Hu, Z.; Yang, P.; Li, B.; and Wang, Z. 2024. Multi-agents based on large language models for knowledge-based visual question answering. *arXiv preprint arXiv:2412.18351*.

Jiang, B.; Zhuang, Z.; Shivakumar, S. S.; Roth, D.; and Taylor, C. J. 2024. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; and Ge, Z. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143: 102611.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Marcu, A.-M.; Chen, L.; Hünermann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; Arani, E.; and Sinavski, O. 2024. LingoQA: Visual Question Answering for Autonomous Driving. arXiv:2312.14115.

Mozaffari, M.; Sapkota, H.; and Yu, Q. 2025. GLEN: Generalized Focal Loss Ensemble of Low-Rank Networks for Calibrated Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19563–19571.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33: 15288–15299.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. arXiv:2305.14836.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2025. DriveLM: Driving with Graph Visual Question Answering. arXiv:2312.14150.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Wang, Z.; Wan, W.; Lao, Q.; Chen, R.; Lang, M.; Wang, K.; and Lin, L. 2023a. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

Wang, Z.; Wan, W.; Lao, Q.; Chen, R.; Lang, M.; Wang, K.; and Lin, L. 2023b. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Whitehead, S.; Petryk, S.; Shakib, V.; Gonzalez, J.; Darrell, T.; Rohrbach, A.; and Rohrbach, M. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, 148–166. Springer.

Wieczorek, T. J.; Daun, N.; Khan, M. E.; and Rohrbach, M. 2025. Variational Visual Question Answering. *arXiv preprint arXiv:2505.09591*.

Yang, R.; Rajagopal, D.; Hayati, S. A.; Hu, B.; and Kang, D. 2024. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Yu, W.; Iter, D.; Wang, S.; Xu, Y.; Ju, M.; Sanyal, S.; Zhu, C.; Zeng, M.; and Jiang, M. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Zhong, M.; Wang, G.; Chuang, Y.-N.; and Zou, N. 2025. Quantized Can Still Be Calibrated: A Unified Framework to Calibration in Quantized Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30503–30517. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhou, Y.; Mei, J.; Yu, Y.; and Syeda-Mahmood, T. 2023. Medical visual question answering using joint self-supervised learning. arXiv:2302.13069.