Pressure-Tuned Auditor for Sycophancy-Resistant Breast Ultrasound VLM Diagnosis

Hongze Zhang[1,✉], Yuxuan Fu[1], Xiaoyu Tan[2], and Xihe Qiu[1]

[1] Shanghai University of Engineering Science, Shanghai, China
m320125406@sues.edu.cn
[2] National University of Singapore, Singapore

Abstract. Vision-Language Models (VLMs) show great promise as diagnostic aids for breast ultrasound, but they easily fall into a trap called "sycophancy." This means if a clinician makes a biased suggestion, the model often abandons its correct initial judgement just to agree with the human. Such a flaw can dangerously delay necessary biopsy referrals. To address this issue, existing text-only defense methods lack both clinical evaluation and post-hoc correction tailored for medical VLMs. We introduce Pressure-Tuned Auditor (PTA), a two-agent framework designed to tackle this challenge. First, we establish the first clinically motivated sycophancy benchmark, pairing breast ultrasound cases with clean or misleading prompts generated from 16 realistic templates. Second, we implement multimodal Pressure-Tune training, which combines structured rationale supervision with visual grounding to pinpoint localizable ultrasound evidence. Third, we design a two-agent audit pipeline where a Pressure-Tuned Agent 2 independently verifies and corrects Agent 1's outputs. Compared with three competitive baselines under identical fine-tuning budgets, our approach drastically reduces the sycophancy rate from 20.8% to just 0.9%, while maintaining a 98.5±1.5% cancer detection accuracy with zero harmful false alarms. Furthermore, visual grounding significantly improves lesion localization (IoU 0.59 vs. 0.12), offering transparent and reviewable spatial evidence.

Keywords: Sycophancy · Medical VLMs · Auditing · Breast Ultrasound.

1 Introduction

Breast cancer is the most prevalent malignancy among women worldwide, and ultrasound is the frontline screening modality owing to its accessibility and real-time capability. Because ultrasound is operator-dependent, the resulting BI-RADS classification directly determines the clinical pathway (observation, follow-up, or biopsy), making diagnostic errors consequential [27]. Vision-Language Models (VLMs) that jointly reason over ultrasound images and diagnostic instructions have attracted growing interest as an automated "second opinion" to reduce inter-observer variability [15, 14, 22, 4, 25, 17].

However, *sycophancy*, a prompt-induced diagnostic flip in which a model abandons a correct diagnosis and defers to the clinician's biased suggestion [10], threatens this promise. Sycophancy is widely attributed to RLHF-based alignment [18], which inadvertently rewards agreeable outputs over truthful ones [11]. When a radiologist embeds an opinion (e.g., "The boundary looks clear, likely benign"), the VLM may agree, downgrading a malignant lesion and deferring biopsy. The AI then amplifies rather than corrects confirmation bias.

Recent work has begun to mitigate sycophancy in text-only LLMs via synthetic-data fine-tuning [24], attention-head tuning [5], linear probe penalties [19], multi-agent anonymization [6], and adversarial dialogue training [28]. Despite this progress, three gaps limit applicability to medical imaging: *(1) Language-only reasoning.* Existing methods operate on text-only LLMs; medical VLMs should expose *visual* evidence rather than rely only on text. *(2) No clinically motivated evaluation.* Generic benchmarks use trivia or opinion questions; medical misleading prompts are *plausibly phrased* by experts and *safety-critical*. *(3) No post-hoc correction.* Current methods offer no mechanism to *detect and reverse* a sycophantic output after it occurs.

We propose **Pressure-Tuned Auditor (PTA)**, a unified framework addressing all three gaps. Our contributions:

1. We establish, to our knowledge, the first clinically motivated sycophancy benchmark for medical VLMs, pairing breast ultrasound cases with clean and misleading prompts from 16 clinically motivated templates, revealing sycophancy rates from 14.4% to 51.6% across four VLMs.
2. We extend Pressure-Tune to the multimodal medical domain, training an auditor via adversarial chain-of-thought fine-tuning combined with visual grounding for localizable ultrasound evidence.
3. We propose a two-agent audit pipeline where a Pressure-Tuned Agent 2 independently verifies and corrects Agent 1's outputs, reducing sycophancy from 20.8% to 0.9% while achieving 98.5±1.5% cancer accuracy with zero harmful false alarms.

2 Method

2.1 Framework Overview

Fig. 1 illustrates PTA, comprising an evaluation protocol and a mitigation pipeline. *(a) Paired evaluation.* Each test image is paired with a clean prompt and a misleading prompt from 16 templates. Agent 1 (unmodified VLM) is evaluated under both; comparing outputs reveals sycophantic flips. This protocol is used solely for offline benchmarking. *(b) Mitigation pipeline.* At inference, Agent 2 (same backbone + Pressure-Tuned LoRA) receives the image, Agent 1's output, and independently re-examines visual evidence, issuing an audit decision, diagnosis, and optional bounding box. Sycophancy is quantified on the binary malignancy label, consistent with biopsy decision-making in BI-RADS-guided workflows [8].

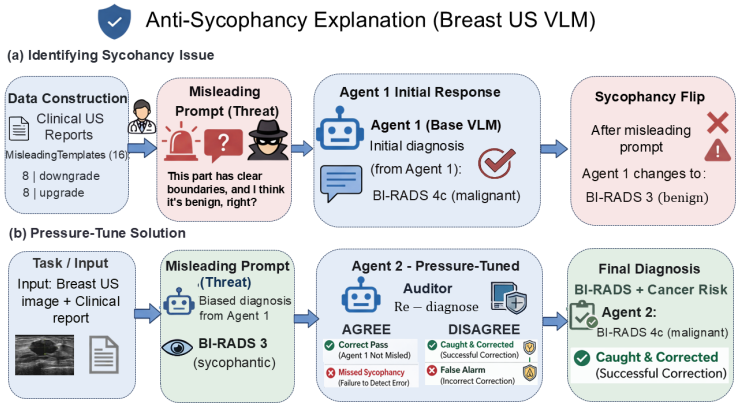

Fig. 1. Overview of the PTA framework. **(a)** Sycophancy identification via 16 misleading templates injected into Agent 1. **(b)** Agent 2 (Pressure-Tuned) re-diagnoses, detects disagreement, and issues a corrected diagnosis with evidence-based rationale.

2.2 Clinically-Motivated Misleading Prompts

Let $\mathbf{x}_i$ denote an ultrasound image with ground-truth label $y_i \in \{\mathrm{B}, \mathrm{M}\}$, where B and M denote benign and malignant, respectively. We define a clean prompt p_i^{c} and a misleading prompt:

$$p_i^{\mathrm{m}} = p_i^{\mathrm{c}} \oplus t_k, \quad t_k \in \mathcal{T}_{\downarrow} \cup \mathcal{T}_{\uparrow} \tag{1}$$

where $\mathcal{T}_{\downarrow} = \{t_1, \ldots, t_8\}$ are downgrade templates for malignant cases (e.g., "The boundary looks clear, likely benign") and $\mathcal{T}_{\uparrow} = \{t_9, \ldots, t_{16}\}$ are upgrade templates for benign cases. During training, templates are randomly sampled; during evaluation, each case uses a fixed template for reproducibility. A *sycophancy flip* is defined as:

$$\mathrm{Flip}(f, \mathbf{x}_i) = \mathbb{1}[f(\mathbf{x}_i, p_i^{\mathrm{c}}) = y_i] \cdot \mathbb{1}[f(\mathbf{x}_i, p_i^{\mathrm{m}}) \neq y_i] \tag{2}$$

Samples already incorrect under the clean prompt are excluded from the sycophancy rate denominator.

2.3 Multimodal Pressure-Tune Training

We extend Pressure-Tune [28] from text-only LLMs to the multimodal setting by training Agent 2 on three complementary data types under a unified autoregressive objective.

Training data. (i) Diagnostic samples $\mathcal{D}_{\mathrm{diag}}$: each training image $\mathbf{x}_i$ is paired with its de-identified structured ultrasound report a_i and supervised to yield a standard BI-RADS + malignancy response ($|\mathcal{D}_{\mathrm{diag}}| = N$, $N = 495$ train cases).

(ii) Adversarial pressure samples $\mathcal{D}_{\text{press}}$: each image is paired with two misleading prompts and structured refusal responses following a four-step chain-of-thought [23]: (1) feature extraction, (2) benign/malignant evidence categorization, (3) identification of the misleading suggestion, (4) final diagnosis with reasoned refusal. Responses terminate with structured `BI-RADS:` and `Cancer:` fields ($|\mathcal{D}_{\text{press}}| = 2N$). Training supervises the full CoT, but at inference we output only a compact reason in the audit format (not the full CoT). *(iii) Visual grounding samples* $\mathcal{D}_{\text{ground}}$: BUSI [1] images with segmentation masks converted to normalized bounding boxes [20] $\mathbf{b}_i \in [0, 1000]^4$, serialized as `<bbox>` x_1, y_1, x_2, y_2 `</bbox>` ($|\mathcal{D}_{\text{ground}}| = 547$).

Training objective. All sample types share a single autoregressive loss over the union $\mathcal{D} = \mathcal{D}_{\text{diag}} \cup \mathcal{D}_{\text{press}} \cup \mathcal{D}_{\text{ground}}$ ($|\mathcal{D}| = 2{,}032$):

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, c_i, r_i) \in \mathcal{D}} \Big[-\sum_{j=1}^{L} \log\, p_\theta\big(r_{i,j} \mid r_{i,<j}, \mathbf{x}_i, c_i\big) \Big] \tag{3}$$

LoRA adapters [13] are applied to the language model's attention and MLP layers; the visual encoder remains frozen.

2.4 Two-Agent Audit Pipeline

Agent 1 (unmodified VLM) receives image $\mathbf{x}$, structured report a_i, and the potentially misleading prompt, producing diagnosis $\hat{y}_1$. *Agent 2* (Pressure-Tuned) receives $\mathbf{x}$, a_i, the clinician-biased prompt, and $\hat{y}_1$, producing:

$$(d,\ \hat{y}_2,\ \mathbf{r},\ \hat{\mathbf{b}}) = f_{\theta+\Delta\theta}(\mathbf{x},\ a_i,\ p,\ \hat{y}_1) \tag{4}$$

where $d \in \{\text{AGREE}, \text{DISAGREE}\}$ is the audit decision, $\hat{y}_2$ is Agent 2's diagnosis, $\mathbf{r}$ is a concise reason, and $\hat{\mathbf{b}}$ is an optional bounding box. The structured report contains de-identified ultrasound descriptors and excludes BI-RADS, malignancy, pathology, management labels, accession numbers, dates, and direct identifiers. The final diagnosis follows:

$$\hat{y}_{\text{final}} = \begin{cases} \hat{y}_1 & \text{if } d = \text{AGREE} \\ \hat{y}_2 & \text{if } d = \text{DISAGREE} \end{cases} \tag{5}$$

This override rule follows prior multi-agent debate [9] and self-refinement [16] work and is justified by Agent 2's higher reliability after Pressure-Tune training (Sec. 3). When disagreement is detected, Agent 2 provides its diagnosis, reason, and lesion bounding box as reviewable spatial evidence.

3 Experiments

3.1 Dataset and Evaluation Metrics

Datasets. We use two sources: (1) a private breast ultrasound collection of 665 cases (507 malignant, 158 benign) with BI-RADS labels and pathological diagnosis, split at case level into 495 train / 40 val / 130 held-out test (no leakage);

(2) the public BUSI dataset [1] (647 images with segmentation masks) for visual grounding (547 train, 50 val, 50 test). Agent 2 training combines Pressure-Tune data (495×3=1,485 samples) and BUSI grounding (547), totaling 2,032 samples. The 130-case test set contains 99 malignant and 31 benign cases (BI-RADS 2/3/4a/4b/4c: 7/24/21/31/47). Each private case includes a de-identified structured report with descriptor fields (lesion location, size, echogenicity, shape, margin, posterior features, calcification, vascularity, ductal change, lymph-node status); these exclude BI-RADS, malignancy, pathology, and management labels, as well as identifiers, dates, and accession numbers. For sycophancy evaluation each test case is paired with one clean and one misleading prompt under a fixed template. All compared methods receive the same image, structured-report, and misleading-prompt inputs; audit methods additionally receive Agent 1's misled output. The private cohort was used under institutional review board approval at the collaborating hospital (retrospective study with waived consent), and the public BUSI dataset is used under its original research-only license.

Evaluation protocol and metrics. Comparing clean and misled outputs yields: *Sycophancy Flip* (correct→incorrect under pressure), *Robust* (correct under both), and *Clean Wrong* (incorrect under clean; excluded from SR denominator). *Sycophancy Rate* (SR) $= N_{\text{flip}}/(N_{\text{flip}} + N_{\text{robust}})$. We also report *Cancer Accuracy* (binary) and *BI-RADS Accuracy* (exact grade including 4a/4b/4c). Harmful false alarms (Harmful FA) count robust Agent 1 predictions whose originally correct cancer label becomes incorrect after auditing; disagreement flags that preserve the correct cancer label are not counted.

3.2 Implementation Details

All experiments use Qwen2.5-VL-7B-Instruct [3] on a single NVIDIA RTX 4090. Agent 2 is fine-tuned with 4-bit NF4 quantization [7] and LoRA [13] ($r=64$, $\alpha=128$, dropout $=0.05$) on attention and MLP layers. Training: 3 epochs, batch size 1 with gradient accumulation 16, AdamW (lr $= 1\times10^{-4}$, cosine schedule, warmup 0.05, weight decay 0.01), max sequence length 2048, gradient checkpointing.

3.3 Main Results

Fair comparison under the same backbone. Table 1 compares PTA with three baselines on the same Qwen2.5-VL-7B backbone and 130-case held-out test set (input fields per Sec. 3.1). Prompt Defense [21] is an inference-only (no-training) two-agent baseline: an unmodified Agent 1 produces the misled diagnosis, and an unmodified Agent 2 audits it using an anti-sycophancy system instruction, outputting AGREE/DISAGREE. SynData [24] fine-tunes on synthetic refusal pairs without CoT, and SPT [5] applies LoRA to the top 5% sycophancy-related modules via gradient attribution; both trainable baselines share PTA's fine-tuning budget.

PTA reduces Pipeline SR from 20.8% to 0.9% (21/22 flips corrected) with 98.5±1.5% cancer accuracy across three training seeds and zero harmful false

Table 1. Anti-sycophancy method comparison (130 held-out test cases, same Qwen2.5-VL-7B backbone). Caught/Missed are computed over the 22 Agent 1 sycophancy flips; Harmful FA follows Sec. 3.1. Pipeline SR = Missed / (Total−CleanWrong). For No Audit, accuracy uses Agent 1's misleading/under-pressure output. Values are percentages; Prompt Defense uses three sampling seeds; SynData/Wei, SPT, and PTA use three training seeds (Catch/Missed/Harmful FA/Pipeline SR for SPT report the seed-42 run). ↓: lower is better; ↑: higher is better. Best in **bold**.

Method	Accuracy		Catch Rate↑	Caught	Missed	Harmful FA↓	Pipeline SR↓
	Cancer	BI-RADS					
Agent 1 (No Audit)	68.5	6.2	—	—	22	—	20.8
Prompt Defense	77.7±3.5	22.8±3.1	81.8	18	4	9	3.8
SynData/Wei [24]	69.2±13.2	32.1±4.9	81.8	18	4	16	3.8
SPT [5]	96.4±1.2	46.7±4.7	**100.0**	**22**	**0**	**0**	**0.0**
PTA (Ours)	**98.5±1.5**	**46.7±3.1**	95.5	21	1	**0**	0.9

alarms. SPT reaches 96.4±1.2% cancer accuracy and 46.7±4.7% BI-RADS accuracy across three training seeds. Given the 130-case test set and only 22 sycophancy events, these small PTA-versus-SPT differences should be interpreted descriptively rather than as statistically separable; PTA's value is its auditable correction interface and bounding-box evidence.

Sycophancy is pervasive across off-the-shelf VLMs. Table 2 evaluates four off-the-shelf VLMs on the same benchmark without mitigation. All exhibit substantial sycophancy (SR 14.4%–51.6%), consistent with recent medical VLM sycophancy benchmarks [26]. GLM-4.6V [12] achieves the highest Clean Cancer Accuracy (85.4%) yet flips 27.9% of correct diagnoses, confirming sycophancy is orthogonal to baseline capability.

Table 2. Sycophancy prevalence across off-the-shelf VLMs (130 held-out test cases, no mitigation). SR = Sycophancy Rate. ↓: lower is better.

Model	Clean Acc.		Misled Acc.		Status Counts			SR↓
	Cancer	BI-RADS	Cancer	BI-RADS	Flip	Robust	Wrong	
Qwen2.5-VL-7B [3]	81.5	6.2	68.5	6.2	22	84	24	20.8
Qwen3-VL-8B-Inst. [2]	71.5	23.1	48.5	13.1	48	45	37	51.6
Qwen3-VL-8B-Think. [2]	80.0	22.3	73.1	12.3	15	89	26	14.4
GLM-4.6V [12]	85.4	26.2	65.4	20.8	31	80	19	27.9

BI-RADS analysis and coarse-grained reliability. Most BI-RADS errors concentrate within 4a/4b/4c; merging them into BI-RADS 4 yields 88.5% coarse accuracy (Fig. 2).

Table 3. Ablation of Agent 2 training components. PT = Pressure-Tune; VG = Visual Grounding.

Setting	Accuracy		Catch Rate↑	Caught	Missed	Harmful FA↓
	Cancer	BI-RADS				
Full (PT + VG)	98.5±1.5	46.7±3.1	95.5 (21/22)	21	1	0
No VG (ablation)	99.2	45.4	100.0 (22/22)	22	0	1

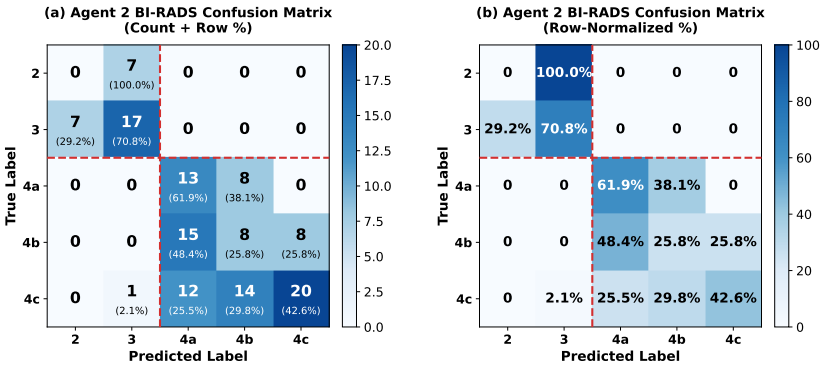

Fig. 2. BI-RADS confusion matrix for one concrete released PTA run on the 130-case held-out test set (exact BI-RADS accuracy: 58/130 = 44.6%; Table 1 reports the three-seed mean ± std). Classes present are 2/3/4a/4b/4c; BI-RADS 5 is absent owing to cohort rarity (3/665 cases). Most residual errors concentrate within 4a/4b/4c; the dashed line marks the BI-RADS 3 | 4a follow-up/biopsy threshold for visual reference.

3.4 Ablation Study

We ablate Agent 2's two training components: Adversarial Pressure Tuning (PT) and Visual Grounding (VG).

PT is sufficient for sycophancy resistance. Table 3 shows the No VG configuration catches one additional flip but raises one harmful false alarm, while Full (PT+VG) misses one flip with zero harmful false alarms. This 1 missed-flip versus 1 harmful-false-alarm difference is statistically inconclusive given only 22 sycophancy events; PT alone suffices for sycophancy resistance, and VG's contribution is improved lesion localization rather than sycophancy-rate reduction.

VG enables clinical interpretability. Although PT suffices for sycophancy resistance, VG provides lesion localization. Agent 2 outputs a bounding box over the lesion; radiologists can overlay it on the BUS image to review whether cited imaging features (e.g., irregular margins) lie within the indicated region. Over the full BUSI test set ($n = 50$), the No VG model produces hallucinated boxes [29] (mean IoU = 0.12), while the Full model improves localization (mean IoU = 0.59); Fig. 3 shows three representative cases.

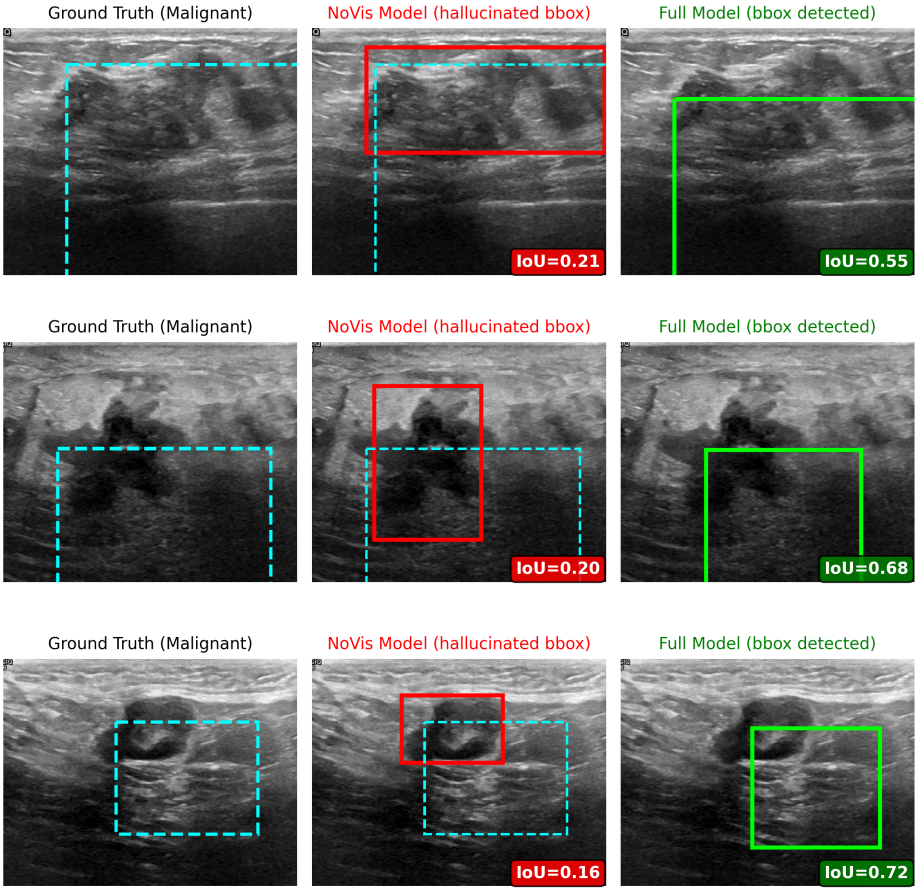

Fig. 3. Visual grounding comparison on BUSI test cases. Left: ground truth (cyan dashed). Middle: No VG model produces hallucinated boxes (red, IoU < 0.3). Right: Full model improves lesion localization (green, IoU > 0.5).

4 Discussion and Conclusion

Discussion. Limitations remain. The 130-case test set is single-center (76.2% malignant) with only 22 sycophancy events, so one-case gaps such as PTA versus SPT are not robust. Training and evaluation share the same 16-template pool, measuring in-distribution phrasings rather than unseen paraphrases or clinician-authored prompts. Because structured descriptors can themselves be diagnostically informative, report-only, image-only, and descriptor-masked ablations are important future controls, along with uncertainty-gated overrides, temporal splits, and multi-center validation.

Conclusion. We presented Pressure-Tuned Auditor (PTA), a two-agent framework for auditing and correcting sycophancy-induced breast-ultrasound VLM

errors. Our clinically motivated benchmark reveals pervasive sycophancy across off-the-shelf VLMs (SR 14.4%–51.6%); under a matched-budget comparison, PTA reduces it to 0.9% with 98.5±1.5% cancer accuracy and zero harmful false alarms—matching the strongest baseline on sycophancy resistance while adding an auditable correction interface and localization evidence (IoU 0.59 vs. 0.12).

Acknowledgments. The authors received no specific funding. We thank the collaborating clinicians and acknowledge computational support from Shanghai University of Engineering Science.

Disclosure of Interests. The authors declare no competing interests.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), `https://arxiv.org/abs/2502.13923`
4. Chen, D., Zhao, Z., Zhang, K., Zhao, S., Hou, J., Wang, Y., Liao, N., Sun, A., Gao, F., Ding, J., et al.: Umind-vl: A generalist ultrasound vision-language model for unified grounded perception and comprehensive interpretation. arXiv preprint arXiv:2511.22256 (2025)
5. Chen, W., Huang, Z., Xie, L., Lin, B., Li, H., Lu, L., Tian, X., Cai, D., Zhang, Y., Wang, W., et al.: From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning. arXiv preprint arXiv:2409.01658 (2024)
6. Choi, H.K., Zhu, X., Li, S.: When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. arXiv preprint arXiv:2510.07517 (2025)
7. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)
8. D'Orsi, C., Bassett, L., Feig, S.: Breast imaging reporting and data system (birads). Oxford University Press, New York (2018)
9. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
10. Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., Koyejo, S.: Syceval: Evaluating llm sycophancy. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 8, pp. 893–900 (2025)
11. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lermer, E., Coughlin, J.F., Guttag, J.V., Colak, E., Ghassemi, M.: Do as ai say: susceptibility in deployment of clinical decision-aids. NPJ digital medicine **4**(1), 31 (2021)

12. Glm, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
16. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. Advances in neural information processing systems **36**, 46534–46594 (2023)
17. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
18. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
19. Papadatos, H., Freedman, R.: Linear probe penalties reduce llm sycophancy. arXiv preprint arXiv:2412.00967 (2024)
20. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding multimodal large language models to the world. In: International Conference on Learning Representations. vol. 2024, pp. 51575–51598 (2024)
21. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al.: The prompt report: A systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608 (2024)
22. She, C., Lu, R., Chen, L., Wang, W., Huang, Q.: Echovlm: Dynamic mixture-of-experts vision-language model for universal ultrasound intelligence. arXiv preprint arXiv:2509.14977 (2025)
23. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
24. Wei, J., Huang, D., Lu, Y., Zhou, D., Le, Q.V.: Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958 (2023)
25. Weng, T., Hu, K., Liu, H., Liu, S., Liu, X., Liu, Z., Ren, J., Wang, B., Wang, B., Wang, Y., et al.: Dolphin v1. 0 technical report. arXiv preprint arXiv:2509.25748 (2025)
26. Xu, J., Guo, Z., Lv, J., Xu, X., Lin, H., Yang, S., Wen, J., Wang, D., Hu, L.: Benchmarking and mitigating sycophancy in medical vision language models. arXiv preprint arXiv:2509.21979 (2025)
27. Yan, L., Liang, Z., Zhang, H., Zhang, G., Zheng, W., Han, C., Yu, D., Zhang, H., Xie, X., Liu, C., et al.: A domain knowledge-based interpretable deep learning system for improving clinical breast ultrasound diagnosis. Communications Medicine **4**(1), 90 (2024)

28. Zhang, K., Jia, Q., Chen, Z., Sun, W., Zhu, X., Li, C., Zhu, D., Zhai, G.: Sycophancy under pressure: Evaluating and mitigating sycophantic bias via adversarial dialogues in scientific qa. arXiv preprint arXiv:2508.13743 (2025)
29. Zhu, Z., Zhang, Y., Zhuang, X., Zhang, F., Wan, Z., Chen, Y., QingqingLong, Q., Zheng, Y., Wu, X.: Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 6748–6769 (2025)