MedBench: Deliberative Evaluation of Medical Language Models

Pratik Jalan[*,1] **Mukul Joshi**[*,1] **Akhilesh Magotra**[2] **Kshitij Jadhav**[1]
[1]Indian Institute of Technology Bombay, Mumbai, India
[2]Birla Institute of Technology & Science, Pilani, India
{30006531,30004808,kshitij.jadhav}@iitb.ac.in akhilesh.magotra@gmail.com

Abstract

We introduce **MedBench**, a benchmark for evaluating medical language models as deliberating agents rather than isolated predictors. MedBench evaluates eight models (4B-32B) on 19,625 questions from six medical QA datasets using Consensus-Aware Model Panel (CAMP), a two-tier protocol in which five 4B–8B models answer independently, revise after observing peer reasoning, and escalate persistent disagreements to larger 20B-32B models. Compared with zero-shot, few-shot, and chain-of-thought baselines, CAMP shows that deliberation is not uniformly accuracy-improving, but reveals interaction-driven behaviors hidden by single-model evaluation. On PubMedQA without external context, the 4B-8B panel outperforms the evaluated 20B-32B individual zero-shot models (54.1% vs. 33.9%), and achieves the best evaluated result with context (75.7%), suggesting that structured interaction can sometimes complement scale. Across five datasets, initial inter-model agreement is positively associated with correctness and serves as a useful difficulty signal. However, on MedXpertQA, unanimous agreement yields only 6.6% accuracy despite 14.4% overall accuracy, suggesting correlated ignorance, where shared biases make consensus misleading. Error analysis shows that most failures are debate-insufficient cases, where incorrect majorities persist despite interaction (93-97%), while debate-harmful cases account for 3-7%. MedBench positions deliberative evaluation as a complement to accuracy-centric benchmarking, measuring when model interaction corrects errors, reinforces shared mistakes, or signals the need for stronger evidence and human review.

1 Introduction

The dominant paradigm in medical AI benchmarking is simple: ask a model a question, record whether it answers correctly, and report a single accuracy number. Considerable progress has followed from scaling (Nori et al., 2023; Sellergren et al., 2026; Singhal et al., 2025) and prompting (Wei et al., 2022; Brown et al., 2020). Yet this paradigm has a structural blind spot: it cannot detect failure modes that only emerge when models are placed in relation to each other, including shared biases, correlated misconceptions, and the qualitative difference between "every model is wrong because the question is hard" and "every model agrees because they all share the same hallucination."

Medicine itself provides a precedent for deliberative evaluation. Physicians facing genuine diagnostic uncertainty consult colleagues, revise initial impressions in light of contradicting reasoning, and escalate to specialists when the collective remains uncertain. Benchmarks, by contrast, ask language models the same question in isolation. The result is that two qualitatively different failure modes are collapsed into a single metric: cases where models are wrong but could be corrected through interaction, and cases where models are wrong in a way that collective deliberation would only reinforce the incorrect reasoning. We operationalize this intuition in **MedBench**, a benchmark that evaluates medical language models not as isolated oracles but as deliberating agents. MedBench specifies both a collection of evaluation datasets spanning the clinical difficulty gradient and the CAMP protocol as its assessment mechanism; the two are designed together so that the protocol's deliberative structure is what makes the benchmark's diagnostic findings possible.

The CAMP (Consensus-Aware Model Panel) is an evaluation protocol, not a deployment proposal. In this protocol, models iteratively revise their answers after observing peer reasoning under a fixed update protocol (Du et al., 2023; Liang et al., 2023). Unlike self-consistency (Wang et al., 2023) and ensemble voting, which aggregate independent pre-

[*]Equal contribution.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 981–991

dictions, CAMP enables iterative interaction between models, making the influence structure observable and measurable. For each question, the protocol generates an inter-model agreement trajectory, per-model position-change sequences, and a resolution pathway, collectively constituting a richer characterization of model capability than accuracy alone. We treat deliberation as an observable process that reveals latent error structure in language models: interaction converts hidden error correlations into measurable signals.

Why this matters in medicine: In high-stakes domains such as clinical decision support, understanding when model confidence is reliable versus when it reflects shared bias is not merely an academic concern. A finding of unanimous model agreement on a specialist question is equally consistent with collective expertise and with collective hallucination; standard benchmarks cannot distinguish the two. MedBench provides formal tools to make this distinction, including the agreement ratio as an uncertainty proxy and the correlated ignorance diagnosis for cases where consensus signals failure rather than reliability.

The evaluation gap: Medical QA benchmarks span a wide difficulty gradient, from MMLU-style questions tractable for 4B models to ten-option specialist questions that remain difficult for the evaluated model sizes (Jin et al., 2020; Zuo et al., 2025; Hendrycks et al., 2021). Standard evaluation conflates confident correct reasoning with lucky guessing, and cannot distinguish "all models disagree because the question is hard" from "all models agree because they share the same bias." MedBench is designed to make this distinction explicit.

Contributions:

1. **MedBench benchmark.** To our knowledge, MedBench is among the first benchmarks to evaluate medical language models under a deliberative protocol across six datasets and eight models (19,625 questions), extending medical NLP evaluation beyond isolated single-model accuracy.

2. **Deliberation can complement scale in a scoped setting.** A panel of 4–8B models outperforms the individual 20–32B models evaluated here on PubMedQA without context and obtains the strongest result among our evaluated settings on PubMedQA with context (75.68%), suggesting that deliberation can sometimes complement scale.

3. **Agreement ratio as an uncertainty proxy.** Across five datasets, the fraction of Panel 1 models initially agreeing is a strong predictor of final accuracy, serving as a practical uncertainty signal. On MedXpertQA the relationship inverts completely, providing a diagnostic criterion for correlated ignorance, a failure mode where unanimous collective confidence actively misleads.

4. **Debate error taxonomy.** Debate failures partition into two primary debate-related modes: debate-insufficient (93–97%, incorrect majority persists despite interaction) and debate-harmful (3–7%, a correct minority model is overturned by peer influence), with Panel 2 insufficiency tracked separately for escalated cases. These error types are invisible to accuracy-only evaluation and suggest different intervention strategies.

2 Related Work

Medical LLM benchmarking: Foundational medical QA benchmarks include MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), and MMLU (Hendrycks et al., 2021). Performance has risen sharply with scale (Nori et al., 2023) and domain fine-tuning (Sellergren et al., 2026). MedMCQA (Pal et al., 2022) expanded evaluation to Indian medical licensing questions. MedXpertQA (Zuo et al., 2025) recently introduced ten-option specialist questions designed to resist saturation; our analysis suggests that it reveals a qualitatively different failure mode. Most benchmark evaluations still treat models in isolation; MedBench instead uses a deliberative protocol to study interaction dynamics across this landscape.

Multi-agent deliberation: (Du et al., 2023) showed that having multiple LLM instances debate their responses over multiple rounds improves factual validity and mathematical reasoning. (Liang et al., 2023) introduced structured multi-agent debate to encourage divergent thinking and found that forcing models to take opposing positions reduces groupthink. (Chan et al., 2023) applied multi-agent debate for chatbot evaluation. CAMEL (Li et al., 2023) explored role-playing agents for collaborative task completion. In clinical reasoning, (Misaghi et al., 2026) study deliberative multi-agent

councils for ophthalmology vignettes. Our work differs from these lines in three ways: (1) we use deliberation as an evaluation protocol applied to a multi-dataset medical QA benchmark suite rather than as a generation strategy for a single domain; (2) we introduce a two-tier hierarchical structure with explicit escalation; and (3) we measure diagnostic failure modes, including agreement-ratio inversion, debate-induced corruption, and escalation pathways, alongside performance gains.

Self-consistency and ensemble methods: Self-consistency (Wang et al., 2023) samples diverse reasoning paths and takes the majority vote without explicit argument exchange. MedBench's debate rounds enable explicit evidence sharing, allowing us to track social influence events (how often one model switches another's answer) and to separate productive self-correction from harmful corruption. This separation is central to our analysis of when deliberation helps and when it fails.

Confidence and calibration: (Xiong et al., 2024) systematically evaluated confidence elicitation in LLMs and found that expressed confidence is poorly calibrated under distribution shift. Our error analysis corroborates this at scale: models report 0.706 to 0.815 average confidence on incorrect answers, approaching their confidence on correct ones, and the observed high confidence on incorrect answers prevents confidence from functioning as a reliable escalation signal in our setting.

Disagreement as signal: (Kuai et al., 2026) showed that language models trained on overlapping corpora exhibit systematic behavioral entanglement that undermines the independence assumptions required for reliable ensemble voting. MedBench reframes inter-model disagreement not as noise to suppress but as an informative signal: high disagreement reliably indicates question difficulty, while unexpectedly unanimous agreement exposes correlated failure modes. This shift from disagreement-as-error to disagreement-as-diagnostic is central to MedBench's contribution as an evaluation framework.

3 Benchmark Datasets

MedBench evaluates models on six medical QA datasets designed to span a difficulty gradient from standard medical curricula to expert specialist knowledge. Table 1 provides an overview.

Dataset	N	Options	Domain
MedQA	12,723	4	USMLE clinical
MedXpertQA	2,450	10	Expert specialist
MetaMedQA	1,373	4	Multilingual clinical
MMLU-Med	1,089	4	6 med. subjects
PubMedQA	995	3	Biomed. literature
PubMedQA+C	995	3	Biomed. + abstract
Total	**19,625**		

Table 1: Composition of the MedBench benchmark suite, including dataset size, number of answer options, and clinical or biomedical domain coverage.

MedQA (Jin et al., 2020) contains four-option questions in USMLE Step 1/2/3 style, testing clinical reasoning required for US medical licensure. **MedXpertQA** (Zuo et al., 2025) presents ten-option expert-level questions spanning treatment, diagnosis, and basic science. With a random baseline of 10%, this dataset is especially challenging for the evaluated models. **MetaMedQA** aggregates multilingual clinical MCQs testing robustness across question styles and language registers. **MMLU-Med** (Hendrycks et al., 2021) draws six medical subcategories from the Massive Multitask Language Understanding benchmark. **PubMedQA** and **PubMedQA+C** (Jin et al., 2019) ask three-class (yes/no/maybe) questions about biomedical research; the paired design isolates retrieval from comprehension by providing or withholding the source abstract.

Together, these datasets span a controlled difficulty gradient, enabling us to study how deliberation behaves under regimes where the evaluated models have comparatively stronger performance (MMLU-Med, MedQA), partial evidence or mixed robustness (MetaMedQA, PubMedQA+C), and expert-level difficulty or absence of provided context (MedXpertQA, PubMedQA without context).

4 The MedBench Evaluation Protocol

The CAMP routes each question through a hierarchical two-tier deliberation protocol. Figure 1 illustrates the overall architecture.

4.1 Formal Protocol

We formalize deliberation as an iterative update process where models revise their predictions based on peer responses. This allows us to quantify not only outcomes but also interaction dynamics such as influence and leadership. The protocol can be inter-

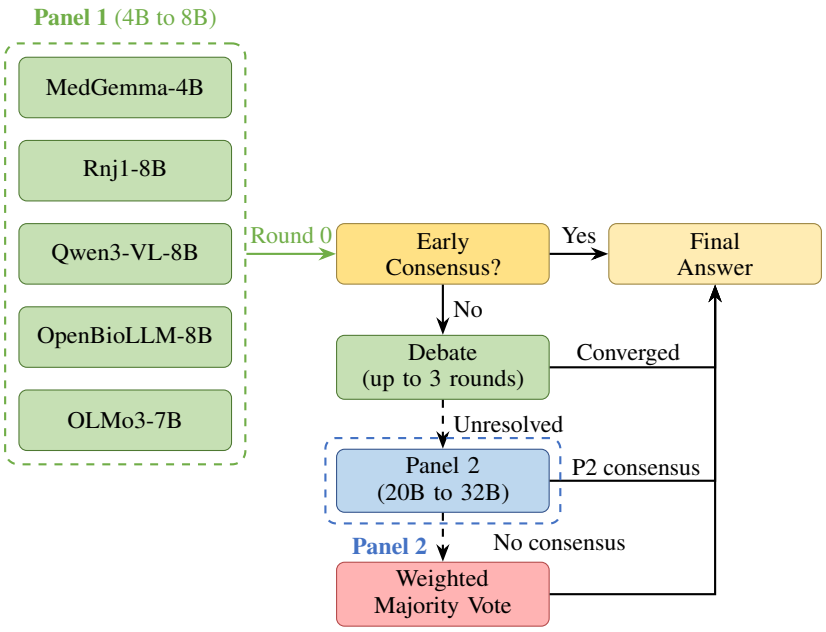

Figure 1: CAMP evaluation architecture. Panel 1 models first reason independently (Round 0). Questions where at least 4 of 5 agree are resolved immediately via Early Consensus. Those that fail are debated for up to three rounds. Persistent Panel 1 disagreement triggers escalation to Panel 2 specialist models (20B to 32B); a confidence-weighted majority vote (WMV) fallback fires only if Panel 2 also fails to converge.

preted as a discrete-time opinion dynamics system over model predictions (Cisneros-Velarde, 2024), where each model's state is updated by exposure to peer reasoning rather than by direct parameter modification.

Let $\mathcal{Q}$ be a question with answer set $\mathcal{A}$, and let $\mathcal{M}_1 = \{m_1, \ldots, m_5\}$ denote the five Panel 1 models.

Independent reasoning (Round 0): Each model m_i receives a role-structured prompt ("You are a board-certified physician on a diagnostic panel...") requiring a JSON response containing a selected option $a_i^{(0)} \in \mathcal{A}$, a scalar confidence $c_i^{(0)} \in [0, 1]$, and a free-text clinical reasoning chain $r_i^{(0)}$.

Early consensus: After Round 0, the agreement ratio (AR) is computed as:

$$\mathrm{AR}^{(t)} = \frac{\max_{a \in \mathcal{A}} \left| \{ i : a_i^{(t)} = a \} \right|}{|\mathcal{M}_1|} \tag{1}$$

If $\mathrm{AR}^{(0)} \geq 0.8$ (at least 4 of 5 models agree), the majority answer becomes the Panel 1 verdict without debate.

Structured debate: When $\mathrm{AR}^{(0)} < 0.8$, each model m_i receives the full set of peer responses from the previous round and updates its answer:

$$\left(a_i^{(t)},\, c_i^{(t)},\, r_i^{(t)}\right) = m_i\Big(\mathcal{Q},\ \big\{(a_j^{(t-1)}, r_j^{(t-1)})\big\}_{j \neq i}\Big) \tag{2}$$

Debate terminates when $\mathrm{AR}^{(t)} \geq 0.8$ or $t = R = 3$.

Influence and leadership: We define two descriptive interaction metrics for this study, using opinion-dynamics work as conceptual background rather than adopting a standard clinical metric. The influence of model m_i on m_j across all debate rounds is:

$$\mathrm{Inf}_{i \to j} = \sum_{t=1}^{R} \mathbf{1}\left[a_j^{(t)} \neq a_j^{(t-1)} \ \wedge \ a_j^{(t)} = a_i^{(t-1)} \right] \tag{3}$$

The leader-follower index for model m_i is:

$$\mathrm{LF}_i = \frac{\sum_{j \neq i} \mathrm{Inf}_{i \to j} - \sum_{j \neq i} \mathrm{Inf}_{j \to i}}{\sum_{j \neq i} \left(\mathrm{Inf}_{i \to j} + \mathrm{Inf}_{j \to i} \right) + \epsilon} \tag{4}$$

with $\epsilon = 1$ to prevent division by zero. $\mathrm{LF}_i > 0$ indicates a net debate leader; $\mathrm{LF}_i < 0$ indicates a net follower.

Panel 2 escalation: Questions where Panel 1 fails to reach consensus after $R = 3$ rounds are forwarded to Panel 2 ($\mathcal{M}_2 = \{$GPT-OSS-20B, MedGemma-27B, Qwen3-VL-32B$\}$). Each Panel

Model	Size	Panel	Type
MedGemma	4B	P1	Medical
Rnj1	8B	P1	General
Qwen3-VL	8B	P1	Multimodal
OpenBioLLM	8B	P1	Biomedical
OLMo3	7B	P1	General
GPT-OSS	20B	P2	General
MedGemma	27B	P2	Medical
Qwen3-VL	32B	P2	Multimodal

Table 2: Model configuration for the two-panel CAMP evaluation setup. Panel 1 contains the primary models and handles 94.5–99.5% of questions, while Panel 2 is invoked only under persistent Panel 1 disagreement.

Model	Panel	Source
MedGemma-4B	P1	HF
Rnj1-8B	P1	HF
Qwen3-VL-8B	P1	HF
OpenBioLLM-8B	P1	HF
OLMo3-7B	P1	HF
GPT-OSS-20B	P2	HF
MedGemma-27B	P2	HF
Qwen3-VL-32B	P2	HF

Table 3: Public Hugging Face sources for the eight models evaluated in MedBench.

2 model first reasons independently (blind to its peers' Panel 2 responses), and then participates in up to two additional debate rounds if Panel 2 initial agreement falls below 2/3. When Panel 2 also fails, confidence-weighted majority vote across all eight models provides the final answer.

Design rationale: The five-model Panel 1 is intended to balance model diversity, local deployability, and a clear majority threshold. The 4-of-5 early-consensus rule resolves questions only when the small-model panel shows strong agreement, while 3–2 splits remain eligible for debate. The three-round limit allows models to revise after seeing peer reasoning while bounding cost and avoiding open-ended deliberation; we do not claim this is the optimal number of rounds. Panel 2 is reserved for persistent Panel 1 disagreement, and weighted majority voting is used only as a final fallback when both tiers fail to converge.

4.2 Model Configuration

Table 2 lists all eight models. Panel 1 comprises five open-weight models spanning 4B to 8B parameters, each deployable locally. Panel 2 comprises three larger models spanning 20B to 32B. This separation between tiers helps us compare deliberation with parameter scale while keeping larger-model conclusions cautious because Panel 2 is invoked only for a small escalated subset. The public source table lists model identities for reproducibility (Sellergren et al., 2026; Vaswani et al., 2025; Bai et al., 2025; OpenBioLLM, 2024; Olmo et al., 2026; OpenAI et al., 2025).

5 Experimental Setup

Baselines: For each of the eight models we run: (1) zero-shot with a minimal prompt; (2) five-shot with domain-matched examples; and (3) chain-of-thought (CoT) prompting (Wei et al., 2022; Brown et al., 2020). In the CoT setting, models are prompted to reason step by step before committing to a final answer. All three protocols form the individual model baseline suite.

Metrics: Primary metric is accuracy. Secondary metrics include: debate uplift (CAMP's Round 0 majority accuracy), escalation rate, self-correction rate (fraction of position changes toward the correct answer during debate), and corruption rate (fraction away from the correct answer). We apply McNemar's two-tailed test for pairwise comparisons, with $p < 0.001$ considered significant given dataset sample sizes.

Implementation: All Panel 1 models run locally on A100 GPUs. Panel 2 models are accessed via API endpoints. All models are evaluated under identical prompting and output constraints to ensure comparability across protocols. All models use a consistent JSON response format; temperature is set to 0 for all Round 0 responses and to 0.7 for debate rounds to allow opinion updating.

6 Individual Model Results

Table 4 reports accuracy for all eight models under all three baseline protocols.

Table 4 also shows dataset-specific empirical patterns among Panel 1 zero-shot models. Qwen3-VL-8B has the highest Panel 1 zero-shot accuracy on four of six datasets, especially the four exam-style MCQ datasets: MedQA, MedXpertQA, MetaMedQA, and MMLU-Med. OpenBioLLM-8B has the highest Panel 1 zero-shot accuracy on PubMedQA without context but is weak on several MCQ exam datasets, suggesting that biomedical pretraining alone does not ensure robust option-selection behavior in this evaluation. MedGemma-

Protocol	Model	Accuracy (%)					
		MedQA	MXQA	MMed	MMLU	PMed	PMed+C
Zero-Shot	MedGemma-4B	46.01	7.77	39.08	60.17	36.15	**64.96**
	Rnj1-8B	35.26	7.54	28.86	54.72	14.74	57.37
	Qwen3-VL-8B	**53.53**	**7.89**	**47.66**	**76.10**	20.83	60.62
	OpenBioLLM-8B	7.00	3.39	19.26	42.09	**41.11**	22.21
	OLMo3-7B	31.88	6.51	27.75	58.24	12.51	49.09
	GPT-OSS-20B	60.79	19.59	56.90	75.18	33.88	68.27
	MedGemma-27B	57.92	11.72	55.35	77.13	28.99	63.21
	Qwen3-VL-32B	66.19	14.85	60.14	81.85	22.80	56.44
Few-Shot	MedGemma-4B	43.63	7.91	36.51	53.90	29.44	**70.35**
	Rnj1-8B	35.09	**9.33**	29.98	59.92	18.51	62.64
	Qwen3-VL-8B	**53.47**	8.26	**46.09**	**73.31**	**32.53**	64.90
	OpenBioLLM-8B	10.18	0.95	8.32	41.73	3.92	4.55
	OLMo3-7B	29.32	5.77	26.74	47.31	28.33	62.30
	GPT-OSS-20B	61.88	18.30	65.18	71.70	40.94	70.03
	MedGemma-27B	62.38	13.47	55.03	75.68	38.53	67.53
	Qwen3-VL-32B	66.77	14.17	64.59	81.63	32.95	62.61
CoT	MedGemma-4B	**51.11**	8.75	42.44	64.22	35.04	70.21
	Rnj1-8B	37.34	8.41	35.86	62.52	37.88	68.25
	Qwen3-VL-8B	36.99	**10.54**	**54.24**	**77.23**	**46.94**	**70.72**
	OpenBioLLM-8B	9.74	1.18	17.92	38.85	38.49	38.34
	OLMo3-7B	36.64	7.52	33.52	56.67	35.84	63.55
	GPT-OSS-20B	26.89	14.78	62.02	84.54	28.51	52.49
	MedGemma-27B	**79.52**	**17.46**	**66.12**	78.66	47.44	72.42
	Qwen3-VL-32B	74.55	18.11	65.19	**84.95**	42.53	71.13

Table 4: Individual model accuracy (%) under three evaluation protocols across six benchmarks. **Bold** marks the best Panel 1 result per dataset per protocol. Highest marks the overall best per dataset. MXQA = MedXpertQA; MMed = MetaMedQA; PMed = PubMedQA; PMed+C = PubMedQA+C.

4B has the highest Panel 1 zero-shot accuracy on PubMedQA+C, suggesting benefit when biomedical context is provided. These observations describe patterns in this benchmark only and should not be read as universal claims about model capability.

CoT produces divergent effects: it lifts MedGemma-27B from 57.92% to 79.52% on MedQA ($p < 0.001$), establishing the strongest single-model result in our benchmark, while degrading OpenBioLLM-8B on multiple datasets. CoT prompting may conflict with some models' answer-format behavior, amplifying format errors rather than improving reasoning quality. Few-shot prompting severely degrades OpenBioLLM-8B on PubMedQA (41.11% zero-shot to 3.92% few-shot), consistent with negative transfer from examples whose format does not align with the model's response tendencies.

These results highlight that individual model performance is highly sensitive to prompting strategy and training alignment. A 22-point swing from CoT alone, and near-total failure under few-shot for one model, reinforces the need for evaluation frameworks that move beyond single-model accuracy to surface how models respond to external signals.

7 Panel-Based Evaluation

7.1 Overall CAMP Accuracy

Table 5 compares CAMP accuracy against the best individual model under each baseline protocol. The CAMP uses Panel 1 alone for 94.5 to 99.5% of questions, invoking Panel 2 only for 0.5 to 5.5% of escalated cases. CAMP outperforms the best individual Panel 1 zero-shot model on five of six datasets and remains close on MMLU-Med. Comparisons to CoT should be interpreted as comparisons across prompting settings rather than as direct zero-shot baselines.

The headline result is PubMedQA without context: CAMP achieves 54.07%, surpassing the best Panel 2 zero-shot model by 20.2 percentage points, the best zero-shot result by 12.96 points, and the best CoT result (MedGemma-27B, 47.44%) by 6.6 percentage points ($p < 0.001$). This result is notable because PubMedQA without context requires models to answer biomedical research questions

Dataset	CAMP Accuracy
MedQA	58.33
MedXpertQA	14.41
MetaMedQA	49.02
MMLU-Med	74.38
PubMedQA	54.07
PubMedQA+C	75.68

Table 5: CAMP accuracy across benchmarks. Individual model baselines are reported in Table 4.

Pathway	N	%	Acc.	Trigger
EC (Panel 1)	5,656	67.3	53.66%	AR ≥ 0.8
DC (Panel 1)	2,523	30.0	35.47%	1–3 rounds
P2C (Panel 2)	160	1.9	37.50%	P2 agree
WMV (fallback)	63	0.7	19.05%	P2 disagree

Table 6: CAMP resolution pathways for 8,402 questions with complete routing metadata. EC denotes Early Consensus, DC denotes Debate Consensus, P2C denotes Panel 2 Consensus, and WMV denotes Weighted Majority Vote. Percentages use the 8,402-question routing-metadata denominator.

without the source abstract, a regime where larger models might be expected to help. In this evaluated setting, CAMP's panel-based aggregation among five 4–8B models produces better results than any individual model and prompting setting we tested.

We hypothesize that panel aggregation can combine partially overlapping signals across models, enabling the collective elimination of incorrect options even when no single model has complete evidence. The three-class structure of PubMedQA (yes/no/maybe) may be amenable to this mechanism because the "maybe" class can reflect conflicting or insufficient evidence. This interpretation is suggestive rather than conclusive and requires broader validation.

7.2 Resolution Pathway Analysis

Table 6 reports the distribution and accuracy of CAMP resolution pathways for the 8,402 questions for which complete routing metadata were available. Percentages in this table are computed over this 8,402-question routing-metadata denominator.

Early consensus (EC) is the dominant pathway among questions with complete routing metadata, resolving 5,656 of 8,402 questions (67.3%), and is also the most accurate pathway (53.66%). The 18.19-point accuracy gap between EC (53.66%) and debate consensus (DC, 35.47%) should be interpreted primarily as difficulty triage rather than as evidence that debate itself is harmful: EC questions are those where Panel 1 models already show strong initial agreement, whereas DC questions enter debate precisely because Panel 1 is initially split. Within this harder contested group, debate provides a modest uplift of 4–5 percentage points over the Round 0 majority baseline.

Escalation pathways are rare in the routing-metadata analysis. Panel 2 consensus (P2C) accounts for 160 questions (1.9%), while the weighted-majority fallback is used for only 63 questions (0.7%). The low accuracy of WMV cases (19.05%) suggests that questions unresolved by both Panel 1 and Panel 2 may represent a collective knowledge boundary for the evaluated models. In a deployment-oriented setting, such fallback cases would be natural candidates for human review rather than automatic resolution.

8 Diagnostic Analysis

8.1 Agreement Ratio as a Difficulty Signal

One of the informative signals in the CAMP framework is the initial agreement ratio (AR, Equation 1). Figure 2 shows accuracy stratified by AR bin across all six datasets. For five of six datasets, the relationship between agreement, accuracy is generally positive. MedXpertQA is the exception: unanimous agreement (AR = 1.0) predicts only 6.6% accuracy, less than half the dataset mean of 14.4%. This inversion is associated with a systematic positional bias (Ko et al., 2020). Of 2,450 MedXpertQA, 22.8% of final answers select option A (558 total), while A is correct only 9.8% of the time (241 instances), a 2.3× over-selection ratio. This pattern may reflect shared training or evaluation artifacts, common instruction-tuning conventions, option-position bias, or correlated benchmark exposure; without training-data disclosure, we cannot isolate the source. We refer to the resulting behavior as correlated ignorance: not the stochastic uncertainty that deliberation can resolve, but a systematic shared error that debate may reinforce. The practical implication for benchmark is that agreement ratio is a useful dataset-level diagnostic. When a benchmark exhibits the normal positive AR-accuracy relationship, it can help calibrate escalation thresholds. When agreement and accuracy diverge, the benchmark exposes a qualitative failure boundary where deliberation may be counterproductive. The key takeaway is that agreement ratio is a reliable uncertainty proxy only under

Model	Stub.	LF-Idx	Total Inf.	Role
Qwen3-VL-8B	0.913	+0.306	4,274	Leader
OpenBioLLM-8B	0.892	+0.028	1,315	Outlier
OLMo3-7B	0.832	-0.006	1,480	Neutral
MedGemma-4B	0.806	+0.111	3,173	Active
Rnj1-8B	0.737	-0.012	2,233	Follower

Table 7: Per-model debate behaviour metrics averaged across all datasets. Stubbornness denotes the fraction of debate rounds in which a model maintains its position, LF-Idx denotes the leader–follower index from Equation 4, and Total Inf. reports cumulative outgoing influence events.

sufficient error independence (Kuai et al., 2026); when models share systematic biases, unanimity can signal failure rather than confidence.

8.2 Debate Dynamics and Model Roles

Position changes decay rapidly across debate rounds: Round 1 accounts for 67.7 to 81.8% of all position changes, Round 2 for 14.8 to 23.3%, and Round 3 for only 3.3 to 11.2%. This suggests that setting $R = 3$ rounds is adequate for the protocol studied here, although we do not claim optimality without ablations. The pattern also provides no evidence that simply increasing R would resolve MedXpertQA-class failures, where the dominant issue appears to be correlated error rather than unresolved disagreement. Table 7 characterises per-model debate behaviour. Qwen3-VL-8B has the highest outgoing influence count, exerting 4,274 cumulative influence events, 35% more than the next-highest model (MedGemma-4B, 3,173). The single strongest dyadic influence channel is Qwen3-VL-8B $\rightarrow$ Rnj1-8B (368 events on MedQA alone), consistent with Rnj1-8B's role as the panel's primary swing voter (lowest stubbornness: 0.737; most negative LF-index: -0.012). The combination of Rnj1-8B's flexibility and Qwen3-VL-8B's anchoring behaviour is associated with much of the panel's observed self-correction.

The paired PubMedQA experiment isolates the value of providing the source abstract. Accuracy rises from 54.07% on PubMedQA without context to 75.68% on PubMedQA+C, a 21.6 percentage point gain associated with context availability. Debate adds only 0.80 points on PubMedQA+C and 0.00 on PubMedQA without context. In this paired setting, access to the relevant evidence appears more important than adding further deliberation. For system builders, this suggests that retrieval or context provision should be evaluated before more elaborate debate protocols.

8.3 Error Decomposition

We categorise debate errors into two primary modes and analyze Panel 2 insufficiency separately for escalated cases. **Type 1 (debate insufficient):** the dominant failure (93 to 97% of all errors), where the panel converges on the wrong answer regardless of debate, often because models share the same misconception. **Type 2 (debate harmful):** a model initially correct was persuaded to switch away during debate (3 to 7% of errors). This corruption rate is relatively consistent across datasets, suggesting that harmful influence is not confined to a single benchmark. Separately, Panel 2 can remain insufficient after escalation, as on MedXpertQA where 76.87% of escalated cases remain wrong after Panel 2 processing. A cross-cutting finding is overconfidence on errors: models report 0.706 to 0.815 average confidence when they are wrong, approaching their confidence on correct answers. This confidence pattern limits the utility of confidence-weighted voting and suggests that confidence-gated escalation would require stronger calibration methods. Figure 3 visualizes the error-mode pattern and the wrong-answer confidence values.

8.4 Conservative Evidence on Design Choices

The current results provide limited, observational evidence about the CAMP design choices, but they should not be interpreted as full ablations. The 4-of-5 early-consensus threshold separates a high-agreement subset with higher accuracy from lower-agreement questions that require debate, while the MedXpertQA inversion shows why agreement must be treated as a diagnostic signal rather than a universal confidence guarantee. The observed decay in position changes across rounds supports the choice to cap debate at three rounds for this evaluation, but it does not prove that three rounds is optimal. Similarly, the low Panel 2 activation rate keeps the protocol compute-efficient, but the small escalated sample means conclusions about larger models must remain cautious.

Taken together, these findings support a conservative interpretation: CAMP is useful not because it always improves accuracy, but because it exposes when agreement, disagreement, escalation, and confidence behave differently across datasets. This diagnostic role is especially important in medical

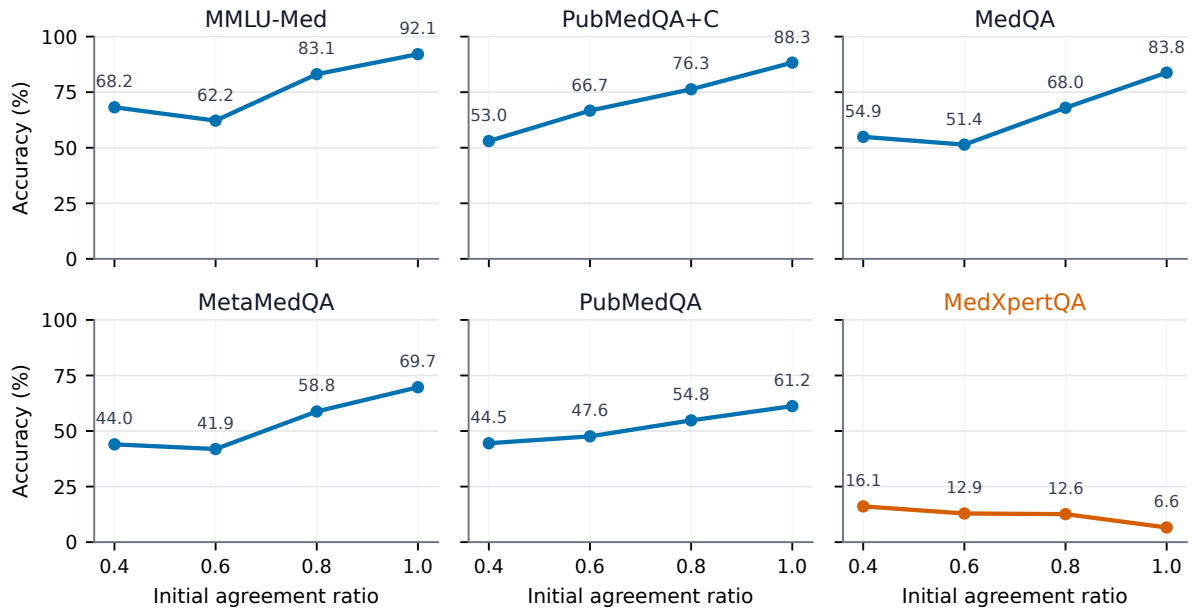

Figure 2: Accuracy conditioned on initial agreement-ratio bin. Five datasets show a generally positive relationship between agreement and accuracy, while MedXpertQA inverts this pattern, indicating correlated ignorance in this evaluation.

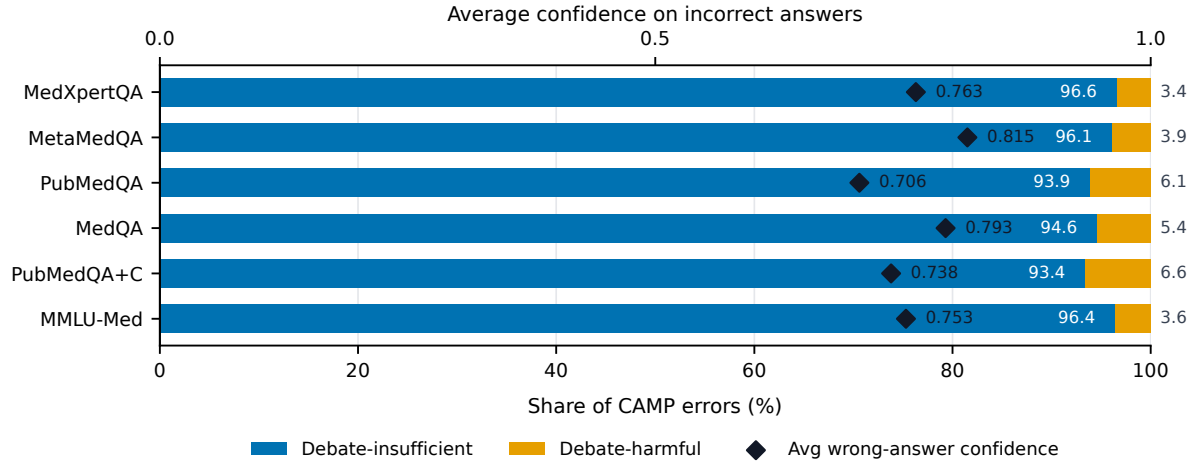

Figure 3: CAMP error decomposition across datasets. Debate-insufficient errors dominate on all datasets, while debate-harmful errors remain smaller but non-negligible; diamonds show the average confidence assigned to incorrect answers.

QA, where a high-accuracy aggregate can obscure whether the system is correcting errors, reinforcing shared mistakes, or deferring to larger models on too few cases to draw broad conclusions.

9 Conclusion

MedBench shows that evaluating medical language models only as isolated predictors misses important interaction-level behavior. Through CAMP, deliberation among smaller models improves performance in selected settings, particularly PubMedQA, but does not consistently replace scale or resolve expert-level gaps. More importantly, the protocol exposes when consensus is informative and when it is misleading: agreement tracks difficulty across most datasets, while MedXpertQA reveals correlated ignorance. The observed debate-insufficient failures, harmful persuasion, and over-confidence on wrong answers suggest that deliberative evaluation should complement accuracy-based benchmarking, with future work on panel design, calibration, model diversity, and human oversight.

Limitations

Knowledge ceiling: MedXpertQA demonstrates that deliberation does not reliably overcome expert-level gaps for the evaluated 4B to 8B models. At 14.4% accuracy (barely above 10% random), this dataset marks a difficult boundary for the current panel. Our results do not show evidence that simply increasing debate rounds would resolve this failure mode.

Correlated training data: Panel 1 models may share pretraining, instruction-tuning, evaluation, or formatting artifacts, which could amplify correlated failure modes such as the positional bias observed on MedXpertQA. Because training-data disclosure is incomplete, we cannot isolate the source of this correlation. A panel with more diverse knowledge sources might exhibit different deliberation dynamics.

Small Panel 2 sample sizes: Panel 2 handles only 0.5 to 5.5% of questions per dataset (4 to 134 questions). Panel 2 statistics (particularly for PubMedQA where only 6 questions escalated) should be interpreted cautiously.

No ablations on panel design: We did not ablate panel size (3 vs. 5 models), debate round limit (1 vs. 2 vs. 3), or escalation threshold. These design choices likely affect both accuracy and the diagnostic value of agreement ratio as a signal.

Confidence calibration: The observed overconfidence on incorrect answers prevents confidence from serving as a useful escalation gate in this study. Calibration-aware prompting or post-hoc calibration strategies are needed before confidence can reliably inform routing decisions.

Ethical Considerations

Clinical deployment risk: MedBench is an evaluation benchmark; the CAMP is an evaluation protocol. The 41.7% error rate on MedQA and 85.6% on MedXpertQA make clear that neither the individual models nor the panel are ready for autonomous clinical deployment without extensive validation, regulatory approval, and mandatory human oversight.

Dataset biases: MedQA and MMLU-Med reflect US-centric medical education. MetaMedQA provides some multilingual coverage, but demographic and geographic biases in underlying training data may produce unequal performance across patient populations.

Overconfidence risk: Models report high average confidence (0.706 to 0.815) on incorrect answers. Any downstream clinical decision support tool must prominently communicate this overconfidence risk to end users.

Acknowledgements

We gratefully acknowledge the Koita Centre for Digital Health, IIT Bombay, and IIT Bombay Trust Lab for their support, resources, and funding that enabled this work. The authors received funding for this project from IIT Bombay Trust Lab. We are sincerely grateful for their generous support. We also thank Mayank Kumar for participating in helpful discussions.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Pedro Cisneros-Velarde. 2024. On the principles behind opinion dynamics in multi-agent systems of large language models. *Preprint*, arXiv:2406.15492.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2567–2577. ArXiv:1909.06146.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. *Preprint*, arXiv:2004.14602. ArXiv:2004.14602.

Chenchen Kuai, Jiwan Jiang, Zihao Zhu, Hao Wang, Keshu Wu, Zihao Li, Yunlong Zhang, Chenxi Liu, Zhengzhong Tu, Zhiwen Fan, and Yang Zhou. 2026. How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles. *Preprint*, arXiv:2604.07650. ArXiv:2604.07650.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36. ArXiv:2303.17760.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2305.19118.

Ehsan Misaghi, Sean T Berkowitz, Bing Yu Chen, Qingyu Chen, Renaud Duval, Pearse A Keane, Danny A Mammo, Ariel Yuhan Ong, Mertcan Sevgi, Sumit Sharma, Sunil K Srivastava, Yih Chung Tham, and Fares Antaki. 2026. Deliberative multi-agent large language models improve clinical reasoning in ophthalmology. *Preprint*, arXiv:2603.21447.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *Preprint*, arXiv:2303.13375. ArXiv:2303.13375.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenBioLLM. 2024. Llama3-OpenBioLLM-8B model card. Hugging Face model card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. ArXiv:2203.14371.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Ashish Vaswani, Mike Callahan, Adarsh Chaluvaraju, Aleksa Gordić, Devaansh Gupta, Yash Jain, Divya Mansingka, Philip Monk, Khoi Nguyen, Mohit Parmar, Michael Pust, Tim Romanski, Peter Rushton, Ali Shehper, Divya Shivaprasad, Somanshu Singla, Kurt Smith, Saurabh Srivastava, Anil Thomas, and 4 others. 2025. Rnj-1-Instruct. Instruction-tuned model release.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*. ArXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35. ArXiv:2201.11903.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations*. ArXiv:2306.13063.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*. ArXiv:2501.18362.