

Too Many Cooks Spoil the Broth: Auditing Multi-Agent LLM Councils for Medical Reasoning

Mukul Joshi^{1*} Pratik Jalan^{1*} Akhilesh Magotra² Kshitij Jadhav¹

¹Indian Institute of Technology Bombay, Mumbai, India

²Birla Institute of Technology & Science, Pilani, India

Corresponding author: kshitij.jadhav@iitb.ac.in

Abstract

Multi-agent debate among large language models (LLMs) is widely assumed to improve reasoning by letting models correct one another. We audit that assumption for English medical question answering (QA). Five open-weight 4 to 8B medical and generalist LLMs answer 8,402 questions across six benchmarks as a council, measured against the same five prompted individually and a single larger Qwen3-VL-32B, text-only. The council beats every member on five of six datasets, but most of that accuracy is in place before debate runs: 67% of verdicts settle at Round 0, and on five of six those are the more accurate questions. Where debate runs, opinion change concentrates on one anchor; on the hardest benchmark 85% of moves toward it happen while it is wrong. There, agreement is worth less than disagreement: questions all five agree on are correct 6.6% of the time against 14.6% otherwise, tracking a shared bias toward the first option. The council also trails the larger model on four of six by up to 16.2 points, at seven times the model calls for 1.6 times the compute. Agreement among models is a measurement that can be biased, not evidence of correctness. We release the [Council Conversations](#).

1 Introduction

Multi-agent debate among large language models has become a popular recipe for sharper reasoning, with reports of consistent factuality gains on difficult tasks (Du et al., 2023; Liang et al., 2024; Zhang et al., 2024; Chen et al., 2023; Chuang et al., 2024). The recipe has crossed into practice through the LLM Council framework, introduced by Andrej Karpathy. The intuition is wisdom-of-crowds: heterogeneous models will surface complementary

*Mukul Joshi and Pratik Jalan contributed equally to this work and share first authorship.

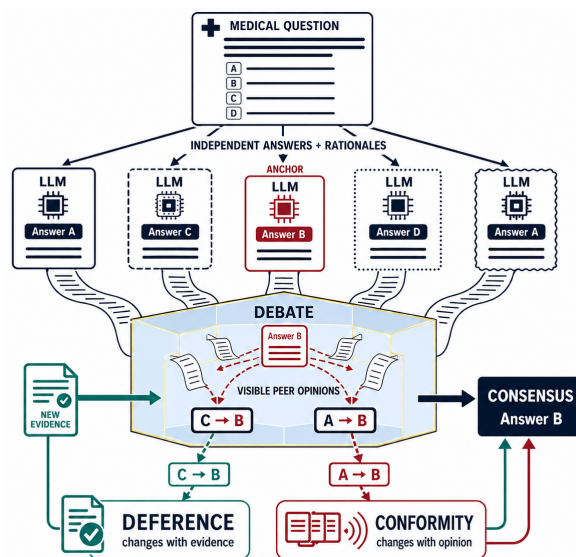


Figure 1: The same consensus verdict can arise through deference to a correct anchor or conformity to an incorrect one. Agreement alone therefore does not reveal how the verdict was formed.

evidence, mutual critique will correct individual errors, and the council’s verdict will be more reliable than any single member.

Two recent threads of work complicate this intuition. (Smit et al., 2024) report that across many tasks, debate gains are matched or exceeded by single-model self-consistency (Wang et al., 2023) or by majority voting over independent samples. Work on sycophancy and conformity in language models (Sharma et al., 2025; Perez et al., 2022; Wei et al., 2024; Zhu et al., 2025; Pitre et al., 2025) shows that models update their answers toward visible group opinion even without new evidence, and (Choi et al., 2026) show that exposing agent identities amplifies these effects in multi-agent settings.

Medical question answering is a natural stress-test. Stakes are real, errors are well characterised (Labkoff et al., 2024), and open-weight medical LLMs in the 4 to 8B range are an attractive de-

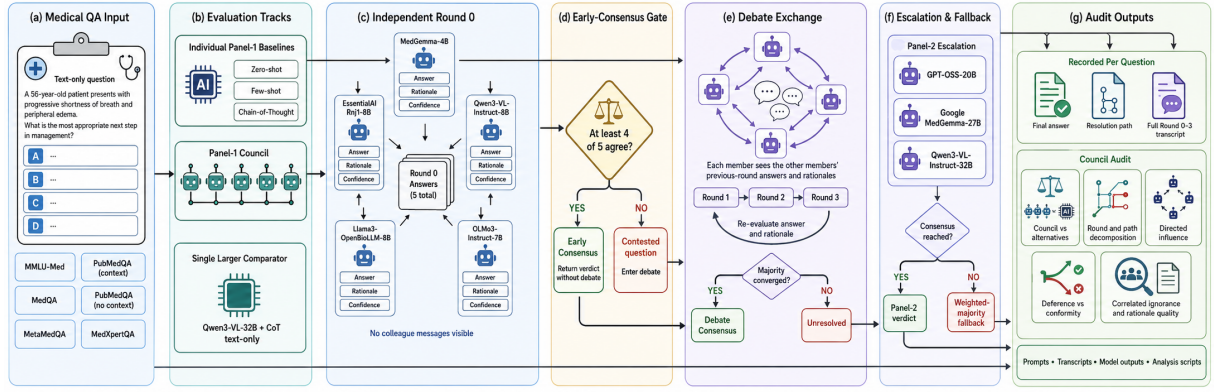


Figure 2: The audited pipeline, from a medical question through the three evaluation tracks and the council’s four resolution routes to the per-question records that drive our analysis.

ployment target precisely because they fit on a single GPU. If the council recipe works anywhere, it should work here.

We do not propose a new debate method. We treat a five-member council as an instrument and ask, on 8,402 questions across six medical benchmarks, the central audit question: does multi-agent debate help, and if so, at which rounds and on which questions? The question decomposes into three parts. Does the council exceed each member’s accuracy under zero-shot prompting, and where do few-shot and chain-of-thought baselines sit? When questions actually enter debate, does per-round accuracy improve from Round 1 to Round 3, or does the debate path under-perform the no-debate path? And when followers change their answers, are they deferring to the most accurate member, or conforming to whichever member happens to anchor the round? Figure 1 summarizes the distinction at the centre of our audit: the same consensus verdict may be produced by a follower moving toward a correct anchor or toward an incorrect but visible anchor. Agreement therefore records an outcome, not the mechanism that produced it.

Our contributions are an audit, not a method. (1) A council-versus-alternatives comparison. We benchmark Panel-1, our five-member small-model council, against every one of its members under zero-shot prompting on six datasets, with few-shot and chain-of-thought as ablations, and against a single larger model with chain-of-thought as the operational alternative (Section 4). (2) A mechanistic audit of what the debate rounds actually do. We decompose the verdict by resolution path to locate where accuracy is produced (Section 5), build a directed influence graph with a deference-versus-

conformity split (Section 6), and characterise the correlated-ignorance regime on MedXpertQA together with the shared first-option preference that accompanies it (Section 7). (3) The released council conversations and a behavioural-quality check. We release the full per-question council conversations, spanning Rounds 0–3 across the six benchmarks, and we score rationale quality with an LLM-as-Judge protocol triangulated against 100 human-rated cases (Section 8). The conversations let future work examine the same per-round dynamics for counterfactual studies, alternative anchor analyses, or new judge protocols. Throughout, we report what our analysis shows; we do not claim that multi-agent debate fails in general.

2 Related Work

LLM-as-judge and chain-of-thought: We use the judge protocol of (Zheng et al., 2023) and (Liu et al., 2023); (Gu et al., 2025) survey the design space. A standing concern for any multi-agent paper is whether debate gains are simply chain-of-thought gains (Wei et al., 2023; Kojima et al., 2023) delivered by another route. We include chain-of-thought as an ablation baseline and note that the council’s per-round prompt itself elicits a structured rationale from every member at every round, so any incremental contribution of the council must lie in the social dynamics rather than in rationale presence.

Multi-agent debate: (Du et al., 2023) introduced multi-agent debate as a remedy for hallucination, with consistent factuality gains across tasks; (Liang et al., 2024) added divergent-thinking rounds and (Chen et al., 2023; Park et al., 2023) explored persona variants. (Chuang et al., 2024) compare col-

Dataset	Questions	Format	Options	Agreement Ratio	Panel Accuracy
MMLU-Med	1,089	Multiple-choice	4	0.684	74.4%
PubMedQA with context	995	Yes/No/Maybe with abstract	3	0.778	75.7%
MedQA	1,500	USMLE multiple-choice	4	0.606	58.3%
PubMedQA without context	995	Yes/No/Maybe	3	0.773	54.1%
MetaMedQA	1,373	Multiple-choice	4	0.615	49.0%
MedXpertQA	2,450	Expert multiple-choice	10	0.522	14.4%

Table 1: The six benchmarks audited in this study, totalling 8,402 questions. Agreement Ratio and Panel Accuracy are defined in Section 3.

lective intelligence in human and LLM-based agent crowds. The re-evaluation by (Smit et al., 2024) casts doubt on the universality of these gains: debate is matched or exceeded by single-model self-consistency (Wang et al., 2023) or by independent-sample majority voting on many tasks. Our work sits in this re-evaluation thread, in medical reasoning, with five heterogeneous open-weight models and an end-to-end audit of internal state rather than a black-box accuracy comparison.

Conformity and sycophancy: (Sharma et al., 2025) characterise sycophantic behaviour, (Perez et al., 2022) report self-evaluations of agreement bias, and (Wei et al., 2024; Jalan et al., 2026; Banerjee et al., 2026) reduce sycophancy with simple synthetic data. In multi-agent settings, (Zhu et al., 2025) measure Asch-style conformity in language models, (Baltaji et al., 2024) document conformity, confabulation, and impersonation in multi-agent collaboration, (Pitre et al., 2025) treat sycophancy as the central bottleneck of debate quality, and (Choi et al., 2026) show that exposing agent identities amplifies it. Identities are exposed throughout our council by design.

3 Setup

Council: The Panel-1 council is fixed throughout the paper and comprises five small open-weight medical or generalist large language models, each in the 4 to 8 billion parameter range: MedGemma-4B (Sellersgren et al., 2026), Rnj1-8B, Qwen3-VL-8B (Bai et al., 2025), OpenBioLLM-8B, and OLMo3-7B (Olmo et al., 2026). We use “Panel-1” and “council” interchangeably in the rest of the paper to refer to this same five-model group, with “council” preferred in narrative prose and “Panel-1” preferred in figures, tables, and architecture references. Where axis space is tight, figures abbreviate the five members as MedG, Rnj1, Qwen, Bio and OLMo respectively. At Round 0, every member

produces an independent answer and rationale. If four of five members agree, an early-consensus verdict is returned; otherwise, up to three debate rounds (Round 1 to Round 3) follow, in which each member sees the previous-round answers and rationales of the other four. In the debate rounds the council returns a verdict once a majority of members converge on the same answer, and the agreement ratio of that final round is reported as the declared ratio. Cases that still lack consensus escalate to Panel-2, a three-model larger-scale panel comprising GPT-OSS-20B, MedGemma-27B, and Qwen3-VL-32B (each in the 20 to 32 billion parameter range); where no rule is satisfied even there, the verdict is a weighted majority vote over the members’ final answers. Escalation is rare, at most 5.47% of questions on MedXpertQA and none at all on MetaMedQA, and Appendix J reports its accuracy contribution. Figure 2 shows the pipeline end to end, from the question through the three evaluation tracks to the per-question records the analysis draws on. Hugging Face identifiers for every model are given in Appendix A, Table 3.

We assembled the council around a deployable configuration rather than a leaderboard: open weights, 4B to 8B parameters so repeated multi-round inference fits on one GPU, and a mix of medical specialists (MedGemma-4B, OpenBioLLM-8B) with instruction-tuned generalists (Qwen3-VL-8B, Rnj1-8B, OLMo3-7B) so members do not share a training lineage. Models outside that band, not open-weight, impractical for repeated inference, or duplicating a family already represented were excluded. Four-of-five is strict enough that Round 0 does not absorb contested questions while staying reachable with five heterogeneous members; debate is capped at three rounds. Vote entropy is close to flat after Round 1 on every dataset (Appendix E), so we see no continued convergence within the three rounds we run; behaviour beyond

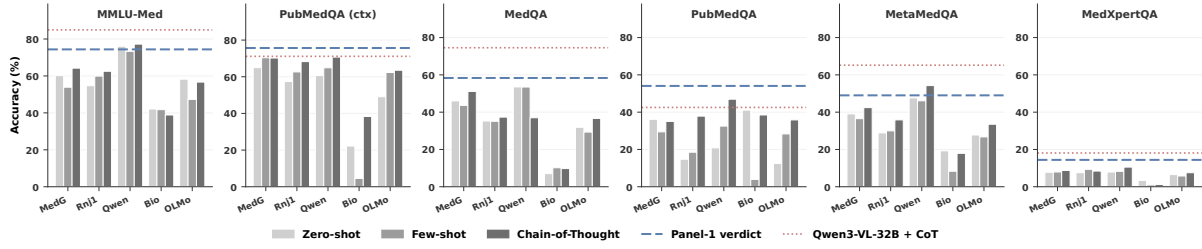


Figure 3: Per-dataset accuracy of every Panel-1 member at zero-shot, few-shot and chain-of-thought, against the council verdict (dashed blue) and a single Qwen3-VL-32B with chain-of-thought (dotted red).

Round 3 is untested.

Datasets: We audit the council on six medical question-answering benchmarks that together total 8,402 questions and span clinical knowledge, biomedical reading comprehension, and expert specialist reasoning: MMLU-Med (Hendrycks et al., 2021), MedQA (Jin et al., 2020), MetaMedQA (Griot et al., 2025), PubMedQA with and without an abstract context (Jin et al., 2019), and MedXpertQA (Zuo et al., 2025). Question counts, formats, the mean Round-0 agreement ratio across council members, and the council’s final accuracy are reported in Table 1. The Round-0 agreement ratio of a question is the size of the largest bloc of members choosing the same option, divided by five; four members agreeing scores 0.8. Unparseable responses contribute no vote and are counted in the denominator.

Baselines: Our primary baseline is zero-shot prompting of each Panel-1 member, run by us on the same 8,402 questions and benchmarked against the council verdict on every dataset (the council is compared against the strongest individual member at zero-shot per dataset). Few-shot (with in-context examples) and chain-of-thought (with an explicit step-by-step prompt) baselines are also produced by us and reported as an ablation in Appendix I. The retained baseline records are aggregate accuracies rather than per-question predictions, so council-versus-baseline contrasts are reported as accuracy differences without significance tests. Comparisons internal to the council use the per-question records; since the Early-Consensus and Debate-Consensus paths partition the questions into disjoint subsets, we compare them with a two-proportion test rather than a paired one.

Cross-check and cost: Every metric we report in this paper is re-derived from the per-question records produced by our council run, by a sec-

ond analysis pass written independently of the one that produced them; on a one-decimal cross-check the two derivations agree on panel-final accuracy for all six datasets. A council pass averages $5 \times (1 + \bar{r}) \approx 7.45$ model calls per question, with $\bar{r} \approx 0.49$ the mean number of debate rounds, against one call for a single Qwen3-VL-32B with chain-of-thought. Weighting each call by the parameter count of the model serving it, the council costs $(1 + \bar{r}) \times 35B \approx 52$ B-parameter-passes against 32, a factor of about 1.6. The gap between the two ratios is the point: roughly seven times the calls for 1.6 times the parameter-weighted compute, so what a deployer pays for is orchestration rather than throughput. Both are first-order proxies, and neither models batching, latency, memory placement, or the parallel execution a multi-model council permits.

Qwen3-VL-32B is used in text-only mode throughout, with no image input. It serves as an open-weight 32B reference from the same family as one council member, not as a multimodal comparator.

4 Does the Council Beat Its Members? Primary Baseline Comparison

We take the first audit question here: does the Panel-1 council beat each of its members on each dataset, and where do the few-shot and chain-of-thought baselines sit? Figure 3 gives the per-dataset comparison, with every council member shown under three prompting modes against the council verdict (dashed blue) and a single Qwen3-VL-32B with chain-of-thought (dotted red) as the operational alternative. The primary contrast is the light-grey zero-shot bars against the dashed blue line; the comparisons below take the visible patterns in turn.

Against its own members: The council verdict exceeds each Panel-1 member’s individual zero-shot accuracy on five of six datasets. The exception

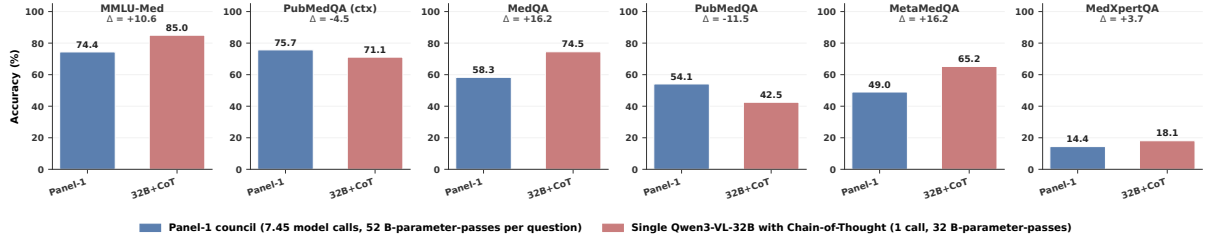


Figure 4: Panel-1 council against a single Qwen3-VL-32B with chain-of-thought, per dataset, with the (single – panel) difference in percentage points.

is MMLU-Med, where Qwen3-VL-8B at 76.1% slightly exceeds the council at 74.4%. Elsewhere the margin ranges from 1.3 points on MetaMedQA (Qwen3-VL-8B at 47.7% versus the council at 49.0%) to 13.0 points on PubMedQA without context (OpenBioLLM-8B at 41.1% versus the council at 54.1%). On those five the council does deliver a lift over every individual zero-shot member.

Prompting mode: Within Panel-1, few-shot and chain-of-thought help different members on different datasets, and prompting mode can move a member a long way: Qwen3-VL-8B falls from 53.5% zero-shot to 37.0% with chain-of-thought on MedQA, while MedGemma-4B gains 5.1 points there. The ordering does not change. Only one few-shot or chain-of-thought baseline overtakes the council, Qwen3-VL-8B with chain-of-thought on MMLU-Med (77.2% against 74.4%), the dataset where that member already led at zero-shot. The council’s lift therefore cannot be attributed to a missing rationale signal at the individual level. Table 5 reports the breakdown for two representative datasets.

Against a single larger model: A single Qwen3-VL-32B with chain-of-thought, the dotted red line in Figure 3, leads on four of the six datasets: by 16.2 points on MedQA (74.5% against 58.3%) and by the same margin on MetaMedQA (65.2% against 49.0%), by 10.6 on MMLU-Med and 3.7 on MedXpertQA. The council leads on both PubMedQA variants, by 4.5 points with the abstract context supplied and 11.5 without it. These are the two Yes/No/Maybe tasks in the suite and also the two with the highest Round-0 agreement (Table 1); we report the association without claiming a mechanism.

Cost: The single larger model is more accurate on four of the six benchmarks and needs roughly one-seventh of the model calls (Section 3), so for this council, this comparator and this benchmark suite,

a deployer who can afford either configuration has a reason to prefer the single model; Figure 4 pairs accuracy against cost per dataset; its differences are computed on unrounded accuracies and so sit a decimal away from the difference of the rounded bar labels in two places. The two exceptions matter: on both PubMedQA variants the council is the more accurate option, so the preference is not uniform even within our own suite, and we do not extend it beyond it.

5 Where Does the Council’s Lift Come From? Per-Round Decomposition

The headline that “the council beats its members” invites the question: **when** does the council do its work? The five-member council reaches a verdict through one of four routes: Early Consensus at Round 0 (four-of-five agreement before any debate fires), Debate Consensus across one or more rounds (Round 1 to Round 3), Panel-2 Escalation on unresolved cases, and a Weighted-Majority fallback when even Panel-2 disagrees. The audit question is whether the debate steps (Round 1 to Round 3) are the source of the lift, or whether the lift is already booked at Round 0.

Before debate: Across the 8,402 questions, 67.3% are decided by Early Consensus at Round 0. The Early-Consensus share is 89.9% on PubMedQA with context, 78.2% on MMLU-Med, and 47.0% on MedXpertQA. The debate mechanism is invoked on the remainder. We are not arguing that this in itself is a problem: easy questions should resolve without debate. But it does mean that two-thirds of the council’s reported accuracy is, by construction, an Early-Consensus number with no contribution from the debate scaffolding. Appendix B gives the per-dataset share and path-conditional accuracy of every route.

Path accuracy: Table 2 reports the path-conditional accuracies. On five of six datasets, the

Dataset	Panel Final	Early Consensus	Debate Consensus	Panel-2 Escalation
MMLU-Med	74.4%	76.2%	67.7%	80.0% (n = 5)
PubMedQA with context	75.7%	78.2%	54.3%	33.3% (n = 6)
MedQA	58.3%	60.0%	53.9%	66.7% (n = 39)
PubMedQA without context	54.1%	55.8%	45.6%	16.7% (n = 6)
MetaMedQA	49.0%	51.4%	44.3%	n/a (n = 0)
MedXpertQA	14.4%	12.5%	15.3%	23.1% (n = 134)

Table 2: Per-dataset accuracy of the Panel-1 council broken down by resolution route, with the escalation count for each route in parentheses.

questions that resolve at Round 0 (Early Consensus) end up with higher accuracy than the questions that pass through the debate rounds: +8.5 points on MMLU-Med, +24.0 on PubMedQA with context, +6.2 on MedQA, +10.2 on PubMedQA without context, and +7.2 on MetaMedQA.* All five gaps are significant at $p < 0.05$ under a two-proportion test, and the largest, on PubMedQA with context, at $p < 10^{-6}$. Appendix C plots the two path accuracies against each other per dataset. MedXpertQA is the one dataset where the relationship inverts (−2.8 points): both paths are near the 10% random baseline, and Section 7 treats it as a separate case.

Per round: A natural follow-up question: among the questions that do enter debate, does accuracy improve from Round 1 to Round 3? On the debate subset of each dataset the per-round Panel-1 accuracy is flat. MedQA moves from 53.5% (Round 1) to 53.8% (Round 2) to 54.0% (Round 3); MetaMedQA from 44.1 to 44.3 to 44.4; MMLU-Med from 67.5 to 67.7 to 67.7; PubMedQA with context from 54.1 to 54.2 to 54.3; PubMedQA without context from 45.4 to 45.6 to 45.6; and MedXpertQA from 14.8 to 15.0 to 15.3. Gains from Round 1 to Round 3 are within roughly one percentage point on every dataset. These later-round figures rest on shrinking subsets, since most questions that enter debate converge in Round 1 and leave.† The Round 1 already-converged majority is not subsequently rescued by Round 2 or Round 3. The per-round vote-entropy trajectory (Appendix E, Figure 10) corroborates this: entropy collapses early on the easier datasets and stays near

1.4 bits through Round 3 on MedXpertQA.

Reconciliation: The overall Panel-1 accuracy in Table 1 is the share-weighted combination of the three subset columns in Table 2. Early Consensus carries roughly two-thirds of the questions, though only 47.0% on MedXpertQA, and is the most accurate path on five of six datasets, so it dominates the average. The Debate Consensus figures quoted above (54.0% at Round 3 on MedQA) are conditional on a smaller, harder subset and sit below the dataset’s Panel Final value (58.3%): the headline accuracy is high because Early Consensus carries both the largest share and the highest accuracy, not because the debate rounds are.

Interpretation: The debate scaffolding does contribute, but modestly, and mainly on questions where the knowledge is already present in at least one member. Where it is absent in most of them, more rounds yield similar entropy and similar accuracy: debate does not synthesise knowledge the council does not have. For this council on this dataset family, the debate steps do less work than the accuracy suggests.

6 Influence Structure: Deference or Conformity?

We compute directed influence as follows: for each question and each round transition $r-1 \rightarrow r$, we count an event $i \rightarrow j$ whenever model j changed its answer in round r to match model i ’s round- $(r-1)$ answer, given that i and j disagreed at round $r-1$. Aggregated across all six datasets, Qwen3-VL-8B generates 4,274 outgoing influence events to the other four members, compared with 3,173 for MedGemma-4B, 2,233 for Rnjl-8B, 1,480 for OLMo3-7B, and 1,315 for OpenBioLLM-8B. Aggregated over the six datasets the largest dyadic channel is Qwen3-VL-8B $\rightarrow$ Rnjl-8B (368, 973, and 338 events on MedQA, MedXpertQA, and

*Gaps are computed on unrounded path-conditional accuracies and may differ by a decimal from the rounded values in Table 2.

†Questions still unresolved at Round 3: 264 on MedXpertQA, 69 on MedQA and on MetaMedQA, 19 on PubMedQA without context, 11 on MMLU-Med, 9 on PubMedQA with context (Figure 10). On the last three the Round-3 value is indicative rather than precise.

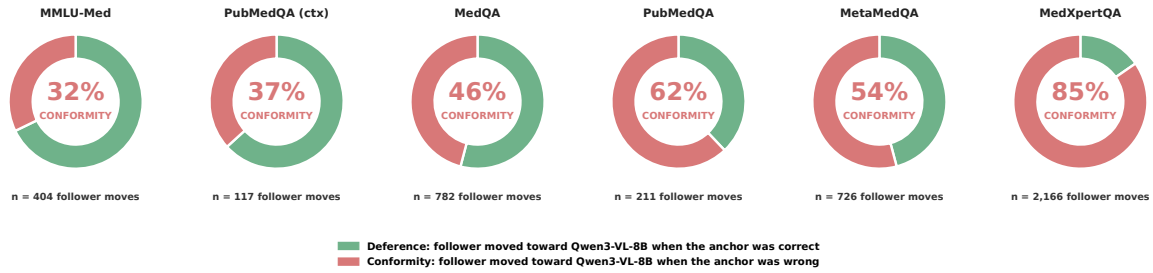


Figure 5: Follower position-changes toward Qwen3-VL-8B split by whether the anchor was correct (green) or wrong (red). The centre number is the wrong-anchor share; the count below is the total moves.

MetaMedQA respectively), though it is not the largest dyad on either PubMedQA variant. The counts in this section are raw event totals taken from the per-question records; the per-dataset 5×5 matrices in Appendix D (Figure 9) show the same events with each panel normalised to 100%.

A natural counter-reading is that Qwen3-VL-8B is the strongest member at zero-shot (Figure 3; Table 5), so its high outgoing influence may be rational deference: the other four correctly updating toward the most accurate member. To separate deference from conformity, we split every follower change toward Qwen3-VL-8B by whether the anchor was itself correct at the prior round.

Deference and conformity: Figure 5 shows the split. On the easier datasets (MMLU-Med, PubMedQA with context), the conformity share is around one-third: most follower moves toward the anchor happen when the anchor is correct, the rational-deference pattern. On the harder datasets the share climbs, though not monotonically, and reaches 85% on MedXpertQA. Of every 100 follower moves toward Qwen3-VL-8B on MedXpertQA, 85 occur in cases where Qwen3-VL-8B itself was wrong. Follower moves toward the anchor are, on this dataset, largely insensitive to whether the anchor was right, which is not what rational deference would predict. The evidence is observational: we count answer changes in the transcripts and split them by anchor correctness, and we do not re-run the council with colleague messages removed.

Self-correction and corruption: Appendix G (Figure 12) reports per-model self-correction (wrong at Round 0, right at the final answer) against corruption (right to wrong). On the easier datasets the ratio is roughly 2.5:1 in favour of self-correction; on MedXpertQA it collapses to 1.16:1, near parity. Appendix L (Figure 14) follows one

question through all three rounds and shows both patterns in a single case: one member updates on new evidence, another holds an incorrect position throughout.

7 Correlated Ignorance on MedXpertQA

On every dataset except MedXpertQA, panel-final accuracy is non-decreasing in the Round-0 agreement ratio: conditional on all five members agreeing at Round 0, accuracy ranges from 61.2% on PubMedQA without context to 92.1% on MMLU-Med (Figure 6). On MedXpertQA the relationship inverts. Of the 61 questions where all five members agreed at Round 0, the panel was correct on only 6.6% (4 of 61). On the 2,389 questions where they did not all agree, the panel was correct on 14.6%.

Control: A sceptical reader might ask whether the inversion tracks agreement specifically or simply marks the questions on which these models are confidently wrong. We can separate the two only partially. On the 108 questions where all five members report a confidence of at least 0.85 at Round 0, regardless of which answer each chose, the panel verdict is correct on 11.1%: below the 14.4% dataset baseline, so confidence alone already selects harder questions, and numerically above the 6.6% recorded when the five agree on one option. At 61 and 108 questions the two are not distinguishable (two-sided Fisher exact, $p = 0.42$), so we do not claim agreement contributes a drop beyond confidence. Both conditions are council-level; a single-model equivalent would need per-model confidence-conditioned accuracies our aggregate records do not retain.

Mechanism: Behind the MedXpertQA inversion is a bias the five members hold in common. The ten options A to J are close to uniform in the ground truth (rarest 222, most common 263), yet the council returns option “A” 558 times, more than twice

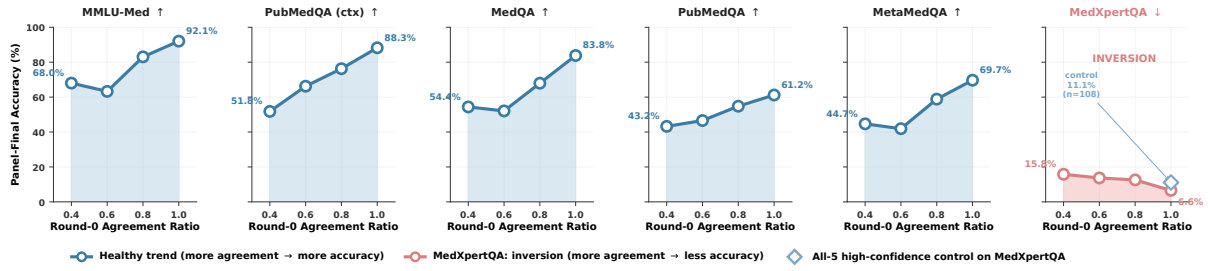


Figure 6: Round-0 agreement ratio against panel-final accuracy, per dataset. The diamond on the MedXpertQA panel is the high-confidence control.

as often as any other option and 2.3 times the 241 questions on which A is in fact correct (Appendix F, Figure 11; binomial test $p < 10^{-30}$). A first-option preference is well documented in individual instruction-tuned models (Zheng et al., 2024). What our transcripts add is that the preference is still visible in the council’s final answers. Because this count pools questions settled at Round 0 with questions that went through debate, it does not on its own show whether exchange weakens the bias. This is the concrete form correlated ignorance takes here, and why agreement and accuracy come apart on this dataset. When members share a heuristic, their agreeing tells you the heuristic fired, not that the evidence supports the answer, and a consensus rule that reads agreement as reliability cannot tell the two apart. We have not run the causal test: re-running MedXpertQA with option positions shuffled would show whether the same content is still preferred once it no longer appears first. What we report is the association under the benchmark’s native ordering.

8 Behavioural Quality of Rationales

The accuracy and structural findings of Sections 4 through 7 rest on the assumption that the council’s rationales themselves are reasonable. We test this assumption with an LLM-as-Judge protocol following (Zheng et al., 2023) and (Liu et al., 2023), scoring each rationale on seven dimensions. Four are analysed here: **Faithfulness** (1–5; does the selected option follow from the written rationale, that is, answer-rationale consistency rather than fidelity to the model’s internal computation), **Groundedness** (1–5; is each claim supported by the question stem), **Helpfulness** (1–5; is the rationale actionable for a clinician), and **Sycophancy rate** (a proxy: the fraction of round-to-round position changes that move toward the visible majority without introducing a new factual claim). Clinical correctness, Calibrated

uncertainty and Hallucination count were also collected but are not analysed here. The primary judge is Qwen3-32B (Yang et al., 2025) (text-only); a secondary GPT-OSS-20B judge scores a 20% overlap subsample. Two human raters independently score faithfulness and sycophancy on 100 cases drawn from the same subsample. The values reported here come from a 1,500-question subsample allocated equally across the six datasets, 250 questions each, under a fixed random seed.

The picture is stable across rounds and datasets. On a 1 to 5 scale, Faithfulness sits between 3.45 and 4.02 across the five members, Groundedness between 3.36 and 3.95, and Helpfulness between 3.20 and 3.86, with Qwen3-VL-8B highest and OpenBioLLM-8B lowest. The Sycophancy Rate rises with round number and is highest in the two members that change position most readily (Rnj1-8B at +0.32 and OLMo3-7B at +0.30 from Round 0 to Round 3), and lowest in the anchor Qwen3-VL-8B at +0.10. The full per-round, per-model judge plot (Figure 13) and per-model summary (Table 4) are in Appendix H. This matches Section 6: followers adopt the anchor’s position in later rounds, and the judge marks that as sycophancy when the change brings no new factual claim.

On 100 cases scored by two raters, inter-rater Cohen κ on sycophancy is 0.47 and Spearman ρ on faithfulness 0.61; judge-to-human agreement is $\kappa = 0.37$ and $\rho = 0.64$ respectively. The judge labels sycophancy more conservatively than the human raters.

9 Discussion

Does three-round multi-agent debate help the council reach the correct answer in medical reasoning, and if so by how much? Our audit answer is softly negative. The council does beat each member’s zero-shot accuracy on five of six datasets, and 67.3% of questions are settled at Round 0 before

any exchange. Among the questions that do enter debate, path-conditional accuracy moves by at most half a point across the rounds we report. Because Early-Consensus and Debate-Consensus are disjoint, difficulty-selected subsets, and because we report no all-question Round-0 majority-vote baseline, this contrast describes where accuracy sits rather than isolating what the debate rounds contribute. Held against a single Qwen3-VL-32B with chain-of-thought prompting, the same council trails on four of six datasets by up to 16.2 points, and leads on the two PubMedQA variants (Figure 4).

Two design implications follow. The first is about composition, and it is not the lesson we expected: lineage diversity is not sufficient. Our five members come from five different lineages, yet Section 7 shows them agreeing on the first option and being wrong together, with agreement giving no warning. Treating consensus as a measurement that can itself be biased matters more here than adding members. The second is about routing. A fixed four-of-five threshold spends the same budget on every question while two-thirds never reach debate; letting it respond to difficulty, or gating the council on a reference model’s confidence, would move that budget somewhere useful. We evaluated neither.

10 Conclusion

We audited a five-member council of open-weight medical and generalist LLMs on 8,402 questions across six benchmarks. It beats each member’s zero-shot accuracy on five of six datasets, and most verdicts are settled at Round 0, where accuracy is higher than on the questions that enter debate. These are different question subsets, so we do not read that contrast as a causal estimate of what the debate rounds add. The verdict tracks one anchor, and the harder the dataset, the more of those moves happen while it is wrong. On the hardest, the direction reverses: the questions the council agrees on at Round 0 are the ones it most often gets wrong, and what it converges on is disproportionately the first option.

Limitations

Our evidence base is narrower than the questions it is asked to settle. The zero-shot, few-shot and chain-of-thought baselines are our own runs, but we retained only their aggregate-accuracy summaries and not the per-question predictions, so ev-

ery council-versus-baseline contrast in Section 4 is a difference in aggregate accuracy, reported without paired tests or per-question intervals. A cleaner comparison would have been a Round-0 majority vote computed over all 8,402 questions, including the third that the council routes into debate: because the Early-Consensus path is by construction the subset on which members already agree, such a baseline would separate ordinary ensembling from debate more sharply than our path decomposition can. We report the decomposition instead because that baseline was not part of the evidence reviewed for this paper. The judge results in Section 8 rest on a narrower base still, a 1,500-question subsample allocated equally across the six datasets and scored by a single primary judge (Qwen3-32B) with a 20% secondary-judge overlap (GPT-OSS-20B); we used the secondary judge as a consistency check during scoring and report no inter-judge agreement statistic, which would require several judges scored on a common subsample and a broader pool of human raters. The accuracy and structural claims of Sections 4 through 7 are computed on all 8,402 questions and do not depend on the judge. We also did not run the counterfactual that would settle the sycophancy question: re-inferring every model at every round with the colleague messages removed roughly doubles the compute of our council run and was outside our budget, leaving the deference-versus-conformity decomposition of Section 6 as the closest observational analogue available to us.

Several confounds we cannot exclude bear on how those numbers should be read. All six benchmarks are public research datasets, so we cannot rule out that some questions appeared in the pre-training data of one or more council members, and absolute accuracies should be read with that in mind. Our central claims rest on comparisons made within a single run over the same instances, but that does not make the resolution paths comparable here: Early Consensus and Debate Consensus hold different questions, and contamination need not reach every model or stage equally, so it could shift both absolute accuracy and which questions settle at Round 0. A second gap is positional. We did not re-run MedXpertQA with the option positions shuffled, so the first-option preference reported in Section 7 is an association under the benchmark’s native ordering rather than a demonstrated positional cause. A third is that the council aggregates votes rather than probabilities: members report a self-assessed confidence that we use

only for the control in Section 7, and self-reported confidence is not a calibrated probability (Kadavath et al., 2022). Aggregating calibrated option probabilities and measuring the calibration of members and of the council as a whole (Guo et al., 2017) would test whether the failures we document are aggregation failures or calibration failures; any such calibration would need held-out data kept strictly separate from the evaluation questions.

Finally, we do not claim that multi-agent debate fails in general. The audit is bounded to one prompt template used by every member at every round, the four-of-five consensus threshold, the Panel-1 composition, a three-round debate cap, and the medical question-answering domain; Appendix K sets out in one place what the audit varies and what it holds fixed. Council size is among the fixed axes, so we cannot say whether three members would aggregate as well as five or would simply lean harder on the anchor, and a Round-0 subpanel analysis would not settle it, because removing a member changes the messages every other member sees in later rounds. We compare against one larger model, where a text-only Qwen3-32B, rather than a vision-language model run without images, would isolate the effect of vision-language pretraining. We did not test closed-weight councils, and whether five or more rounds would close the early-versus-debate gap is unknown. Nor can we speak to debate among larger models: Panel-2 brings 20B to 32B models to bear, but only as an escalation stage on 190 questions, debating no more than the questions it inherits; it is not a large-model council. Our one operational claim, that a deployer might prefer a single Qwen3-VL-32B with chain-of-thought to this five-member council, is bounded accordingly, and does not hold uniformly even within our own suite, where the council is the more accurate option on both PubMedQA variants. The cost comparison behind it is a first-order proxy over model calls and parameter-weighted passes that does not model batching, latency, memory placement, or the parallelism a multi-model council allows.

Ethical Considerations

The questions used in this audit are drawn from publicly released medical question-answering benchmarks intended for research evaluation, not from patient records. No identifiable patient data appears in any council transcript. Our negative findings should not be read as endorsement of any single

model for clinical use. The safe interpretation is that the failure modes we identify (correlated ignorance, anchored agreement) are properties that any council-based clinical decision support pipeline should be tested for explicitly before deployment, consistent with the recommendations of (Labkoff et al., 2024). We release the full per-question council conversations at [Council Conversations](#). AI assistants were used for writing support and sentence polishing.

Acknowledgement

We gratefully acknowledge the Koita Centre for Digital Health, IIT Bombay, IIT Bombay Trust Lab, the Department of Computer Science and Engineering, IIT Bombay, and BITS Pilani 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.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang 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.
- Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. 2024. [Persona inconstancy in multi-agent llm collaboration: Conformity, confabulation, and impersonation](#). *Preprint*, arXiv:2405.03862.
- Shayantan Banerjee, Pratik Jalan, Rudrapratap Chinnara, Chinmayee Kalle, Pramod P Wangikar, and Kshitij Jadhav. 2026. Ontology-guided harmonization enables unified discovery of public metabolomics studies within and across repositories. *bioRxiv*, pages 2026–07.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. [Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors](#). *Preprint*, arXiv:2308.10848.
- Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2026. [When identity skews debate: Anonymization for bias-reduced multi-agent reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14284–14311, San Diego, California, United States. Association for Computational Linguistics.

- Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. [The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents](#). *Preprint*, arXiv:2311.09665.
- 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.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). *Preprint*, arXiv:1706.04599.
- 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.
- Pratik Jalan, Mukul Joshi, Akhilesh Magotra, and Kshiti Sharad Jadhav. 2026. Medbench: Deliberative evaluation of medical language models. In *BioNLP 2026*, pages 981–991.
- 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](#). *Preprint*, arXiv:1909.06146.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#). *Preprint*, arXiv:2205.11916.
- Steven Labkoff, Bilikis Oladimeji, Joseph Kannry, Anthony Solomonides, Russell Leftwich, Eileen Koski, Amanda L Joseph, Monica Lopez-Gonzalez, Lee A Fleisher, Kimberly Nolen, and 1 others. 2024. Toward a responsible future: recommendations for ai-enabled clinical decision support. *Journal of the American Medical Informatics Association*, 31(11):2730–2739.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). *Preprint*, arXiv:2305.19118.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#). *Preprint*, arXiv:2303.16634.
- 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.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). *Preprint*, arXiv:2304.03442.
- Ethan Perez, Sam Ringer, Kamilė Lukošiuotė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2022. [Discovering language model behaviors with model-written evaluations](#). *Preprint*, arXiv:2212.09251.
- Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. [CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda

Zhang, and Ethan Perez. 2025. [Towards understanding sycophancy in language models](#). *Preprint*, arXiv:2310.13548.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. 2024. [Should we be going mad? a look at multi-agent debate strategies for llms](#). *Preprint*, arXiv:2311.17371.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. [Simple synthetic data reduces sycophancy in large language models](#). *Preprint*, arXiv:2308.03958.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. [Exploring collaboration mechanisms for llm agents: A social psychology view](#). *Preprint*, arXiv:2310.02124.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). *Preprint*, arXiv:2309.03882.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. [Conformity in large language models](#). *Preprint*, arXiv:2410.12428.

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](#). *Preprint*, arXiv:2501.18362.

A Model Identifiers

The exact checkpoints used in the audit are listed in Table 3. The judge and the comparator described in

Section 4 were run text-only; the latter corresponds to the relevant Panel-2 entry. All model weights are publicly downloadable, although some require acceptance of the publisher’s access terms.

Role	Checkpoint	Link
Panel-1	google/medgemma-4b-it	HF
Panel-1	EssentialAI/rnj-1-instruct	HF
Panel-1	Qwen/Qwen3-VL-8B-Instruct	HF
Panel-1	aaditya/Llama3-OpenBioLLM-8B	HF
Panel-1	allenai/OLMo-3-7B-Instruct	HF
Panel-2	openai/gpt-oss-20b	HF
Panel-2	google/medgemma-27b-it	HF
Panel-2	Qwen/Qwen3-VL-32B-Instruct	HF
Judge	Qwen/Qwen3-32B	HF

Table 3: Models checkpoints used in the audit.

B Resolution-Path Geography

Figure 7 reports, per dataset, both the share of council decisions resolved by each path and the path-conditional accuracy of that path. Aggregated across the 8,402 questions, 67.3% of council decisions are reached by Early Consensus at Round 0 and 30.0% by Debate Consensus across Rounds 1 to 3. The remaining 2.3% escalate to Panel-2 (190 questions), of which a small residue falls through to the Weighted-Majority fallback.

C Early- versus Debate-Consensus Accuracy (Dot Plot)

Figure 8 pairs Early-Consensus accuracy and Debate-Consensus accuracy per dataset and shows the Early-minus-Debate gap in percentage points on the right margin. On five of six datasets the gap is positive and between 6.2 and 24.0 points. MedXpertQA is the outlier with a small negative gap, both paths near the random baseline.

D Influence Heatmaps

Figure 9 reports the per-pair directed influence matrices per dataset, normalised so each panel sums to 100%. The Qwen3-VL-8B row is the highest-magnitude row across all six datasets. The Qwen3-VL-8B → Rnj1-8B channel is the largest dyad on MedQA, MetaMedQA and MedXpertQA, and is tied for largest on MMLU-Med; on PubMedQA with context the largest is MedGemma-4B → Rnj1-8B, and on PubMedQA without context it is Qwen3-VL-8B → MedGemma-4B.

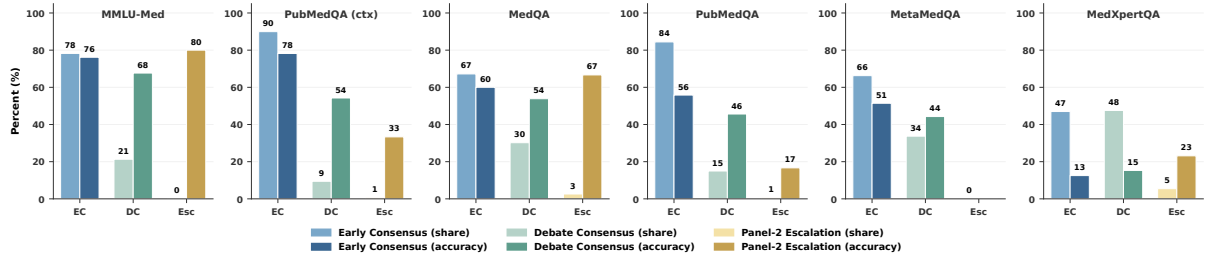


Figure 7: Per-dataset share of council decisions following each resolution route, and the path-conditional accuracy of that route.

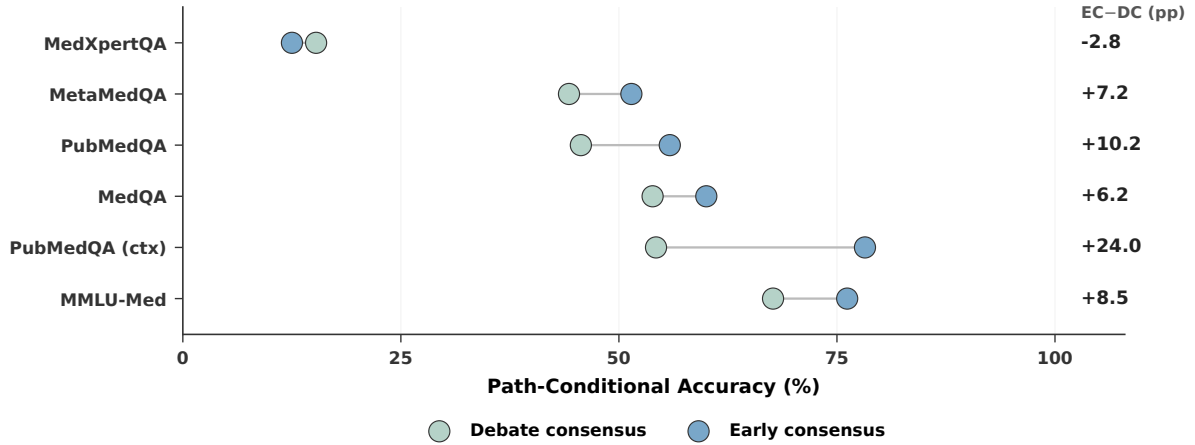


Figure 8: Early-Consensus accuracy (blue) against Debate-Consensus accuracy (green) per dataset, with the gap in percentage points on the right margin.

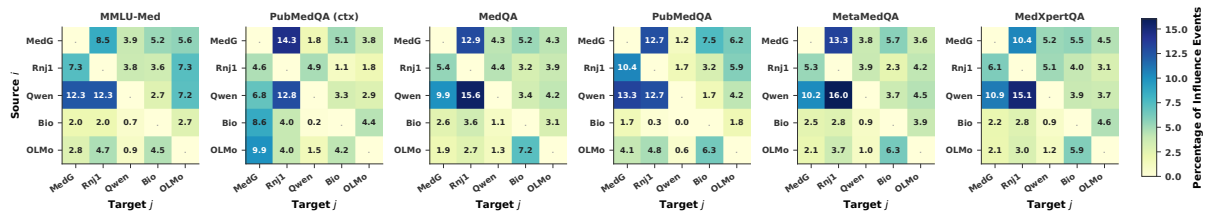


Figure 9: Per-pair directed influence $i \rightarrow j$ across all rounds, normalised to the percentage of all influence events per dataset. Members are abbreviated as in Section 3.

E Vote-Entropy Trajectory

Figure 10 tracks the mean vote entropy of the council across rounds. On the easier datasets entropy collapses almost immediately: most questions leave the debate after Round 1, and the count on each marker shows how few remain. On MedXpertQA it stays near 1.4 bits through Round 3, and the questions that survive to the final round are no closer to agreement than those that entered at Round 1. This is the trajectory Section 5 draws on when it reports that per-round accuracy is flat, and it is also why the Round-3 figures for MMLU-Med, PubMedQA with context and PubMedQA without context rest on fewer than twenty questions each.

F MedXpertQA Option-Selection Bias

Figure 11 compares how often the council returns each of the ten MedXpertQA options against how often that option is correct. The ground truth is close to uniform, from 222 to 263 questions per option, so no option is intrinsically favoured by the benchmark. The council nonetheless returns option “A” 558 times against the 241 questions on which A is correct, more than twice as often as any other option; a binomial test against the ground-truth rate gives $p < 10^{-30}$. Section 7 reads this as the mechanism behind the agreement inversion on this dataset.

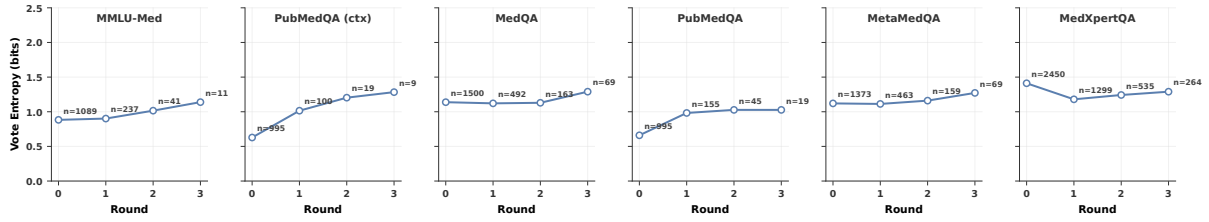


Figure 10: Mean vote entropy across rounds per dataset; n on each marker is the number of questions still unresolved at that round.

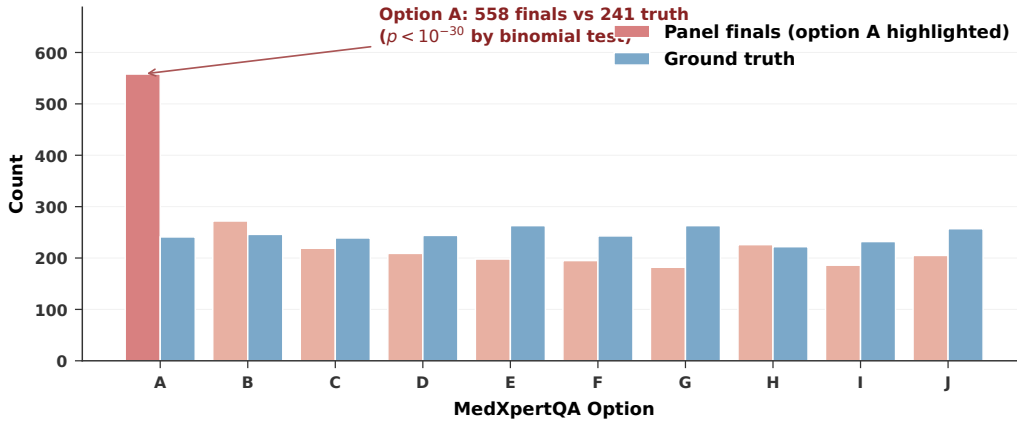


Figure 11: MedXpertQA option-selection histogram, council finals against ground truth.

G Self-Correction and Corruption

Figure 12 reports per-model Self-Correction (a member changes from wrong at Round 0 to right at the final answer) and Corruption (right at Round 0 to wrong at the final answer) per dataset. The Self-Correction-to-Corruption ratio is roughly 2.5:1 on the easier datasets and falls to 1.16:1 on MedXpertQA.

H Judge-Based Behavioural Quality

Figure 13 shows the four judge metrics per round for each council member, and Table 4 gives the per-model means. Faithfulness, Groundedness and Helpfulness stay roughly flat from Round 0 to Round 3, so the rationales do not visibly degrade as the debate proceeds. The sycophancy rate behaves differently: it rises with round number for every member, most steeply for Rnj1-8B and OLMo3-7B, the two members that change position most readily, and least for the anchor Qwen3. Section 8 reads this alongside the influence analysis of Section 6.

I Few-Shot and Chain-of-Thought Ablation

Table 5 reports per-model accuracies under each prompting mode side by side with the Panel-1 coun-

cil verdict for two representative datasets. Few-shot and chain-of-thought help different members on different datasets, and the spread across modes for a single member reaches 16.5 points (Qwen3-VL-8B on MedQA, 53.5% zero-shot against 37.0% with chain-of-thought).

J Panel-2 Ablation

Panel-2 is the three-model escalation panel (GPT-OSS-20B, MedGemma-27B, Qwen3-VL-32B). Escalation is rare: no questions on MetaMedQA, 0.46% on MMLU-Med, and at most 5.47% on MedXpertQA, for 190 escalated questions in total across the six datasets. On the two datasets with enough escalated questions to read, Panel-2 accuracy is above the dataset baseline (66.7% on MedQA with $n = 39$, 23.1% on MedXpertQA with $n = 134$; Table 2), but the absolute contribution to overall accuracy is small, and the remaining four datasets escalate too few questions to support any estimate. Panel-2 is an escalation stage, not a large-model debate: it sees only the questions Panel-1 could not resolve, and it is not run on the full benchmark.

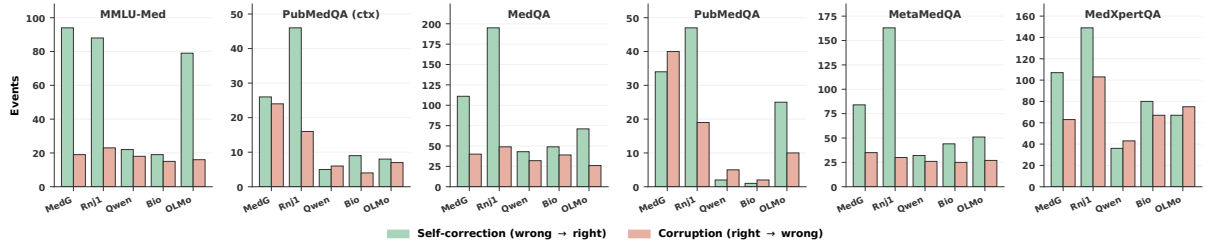


Figure 12: Self-correction (green) against corruption (red), per model per dataset. Members are abbreviated as in Section 3.

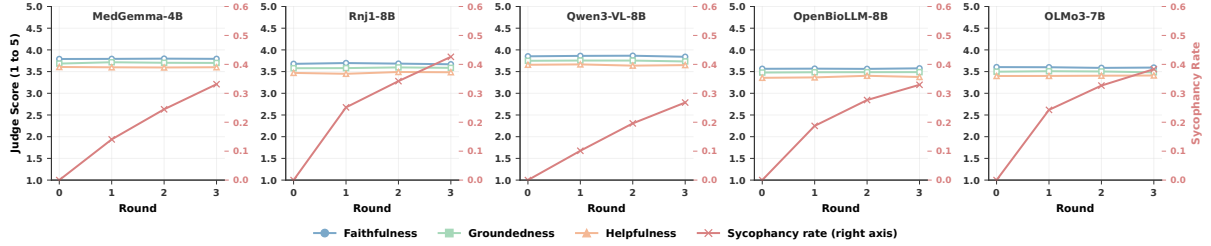


Figure 13: Judge-based behavioural quality per round, per model. Left axis: Faithfulness, Groundedness, Helpfulness on a 1 to 5 scale. Right axis: sycophancy-proxy rate.

Model	Faithfulness	Groundedness	Helpfulness	Sycophancy Rate Δ (R0 to R3)
MedGemma-4B	3.91	3.83	3.74	+0.16
Rnj1-8B	3.69	3.61	3.49	+0.32
Qwen3-VL-8B	4.02	3.95	3.86	+0.10
OpenBioLLM-8B	3.45	3.36	3.20	+0.21
OLMo3-7B	3.49	3.40	3.25	+0.30

Table 4: Per-model means for the three line metrics in Figure 13, averaged over rounds and datasets, with the change in sycophancy-proxy rate from Round 0 to Round 3.

Dataset	Model	ZS	FS	CoT	Panel
MedQA	MedGemma-4B	46.0	43.6	51.1	58.3
	Rnj1-8B	35.3	35.1	37.3	
	Qwen3-VL-8B	53.5	53.5	37.0	
	OpenBioLLM-8B	7.0	10.2	9.7	
	OLMo3-7B	31.9	29.3	36.6	
MMLU-Med	MedGemma-4B	60.2	53.9	64.2	74.4
	Rnj1-8B	54.7	59.9	62.5	
	Qwen3-VL-8B	76.1	73.3	77.2	
	OpenBioLLM-8B	42.1	41.7	38.8	
	OLMo3-7B	58.2	47.3	56.7	

Table 5: Per-model accuracies (%) for two representative datasets under zero-shot (ZS), few-shot (FS) and chain-of-thought (CoT). Panel is the Panel-1 council verdict.

K Prompt and Model Scope

This audit varies prompting mode, benchmark, Panel-1 member, model scale and debate round, while holding fixed the debate prompt, consen-

sus rules, three-round cap, visible member identities, greedy decoding and native option ordering. Visible identities may strengthen conformity (Choi et al., 2026), and we did not test alternative prompts, thresholds or identity settings. Exact model checkpoints, roles and source links are listed in Appendix A. We also release the council conversations and prompts for research use.

L A Three-Round Debate Transcript

Figure 14 shows an illustrative MedQA case in which all three debate rounds fired. We display the three members with a parseable rationale at every round; OLMo3-7B and OpenBioLLM-8B contributed extractable answers but no rationale that could be quoted, so they remain in the five-member vote tallies without displayed turns. Rationales are quoted verbatim and truncated to approximately 280 characters per turn.

MedQA question 1081 ground truth: B (Cystathionine synthase deficiency) council final: B (correct)

Case. A 3-year-old boy is seen in clinic. He was born at home without perinatal care. He was apparently normal at birth, but later developed failure to thrive and developmental delay. He also has a history of cataracts. His older brother had a myocardial infarction at the age of 18 and is rather lanky and tall in appearance. Laboratory testing of his urine showed an increase in the level of an amino acid. What is the most likely mechanism responsible for this boy's pathology?

Options. (A) Hereditary defect of renal amino acid transporter; (B) Cystathionine synthase deficiency; (C) Decreased phenylalanine hydroxylase; (D) Deficiency of homogentisic acid oxidase; (E) Inability to degrade branched chain amino acids.

Round 0: independent answers (no colleague messages visible).

MedGemma-4B selects (C). Confidence 0.95. "Failure to thrive, developmental delay, cataracts, and an increased urinary amino acid level strongly suggest phenylketonuria. Phenylketonuria is caused by a deficiency in phenylalanine hydroxylase."

Rnj1-8B selects (B). Confidence 0.90. "Cataracts and developmental delay suggest a metabolic disorder. The elevated urinary amino acid points to a defect in amino acid metabolism. Cystathionine synthase deficiency leads to elevated homocysteine, which can cause developmental delay and cataracts; the brother's early myocardial infarction fits homocystinuria's vascular risk."

Qwen3-VL-8B selects (E). Confidence 0.95. "Failure to thrive, developmental delay, cataracts, and family history of early myocardial infarction all point to a metabolic disorder; the elevated amino acid in urine suggests a defect in branched-chain amino acid metabolism (maple syrup urine disease)."

Round 0 Vote. A: 2, B: 1, C: 1, E: 1. Agreement Ratio = 0.4. **No early consensus; council enters debate.**

Round 1: each member sees the other four members' Round 0 answers and rationales.

MedGemma-4B selects (C). Confidence 0.95. "While Panellists B and E present compelling arguments, the key distinguishing feature is the increased urinary amino acid level. This is most characteristic of phenylketonuria."

Rnj1-8B selects (B). Confidence 0.90. "I reaffirm cystathionine synthase deficiency. The elevated amino acid in urine, combined with cataracts and developmental delay, fits homocystinuria; the brother's early myocardial infarction is a key cardiovascular tell."

Qwen3-VL-8B selects (B). Confidence 0.95. **Position change from (E) to (B).** "Panellist B's argument is strengthened by the family history of early myocardial infarction, which aligns with homocystinuria's known cardiovascular risk. Cataracts and developmental delay are classic features. My position updates to B."

Round 2: continued debate.

MedGemma-4B selects (C). Confidence 0.95. "B's argument for homocystinuria is compelling, but the urinary amino acid is more characteristic of phenylketonuria."

Rnj1-8B selects (B). Confidence 0.95. "Family history of myocardial infarction at 18 plus cataracts plus developmental delay is the classic homocystinuria triad."

Qwen3-VL-8B selects (B). Confidence 0.95. "My position is unchanged. The combination of features, especially the early myocardial infarction, is classic for homocystinuria. Phenylketonuria does not cause cataracts."

Round 3: final round.

MedGemma-4B selects (C). Confidence 0.95. "I maintain C. Phenylketonuria is most consistent with the urinary amino acid finding."

Rnj1-8B selects (B). Confidence 0.95. "Homocystinuria remains the most likely diagnosis."

Qwen3-VL-8B selects (B). Confidence 0.95. "Position unchanged."

Round 3 Vote. B: 3, C: 1, A: 1 (OLMo3-7B unchanged on A). Agreement Ratio = 0.6. **Council final: B (correct), by majority agreement in the debate rounds.**

Figure 14: One council transcript (MedQA question 1081) in which all three debate rounds fired. Discussed in Appendix L.