

The Consistency Illusion: How Multi-Agent Debate Hides Reasoning Misalignment

Xiaoyang Wang¹, Christopher C. Yang¹

¹Drexel University

xw388@drexel.edu, chris.yang@drexel.edu

Abstract

Multi-agent LLM systems for medical question answering treat consensus as a reliability signal: if multiple agents agree on an answer, it is presumed trustworthy. However, answer-level consensus does not entail reasoning-level alignment. We introduce **CARA** (Cross-Agent Reasoning Alignment), a family of automated metrics that measure whether agents who agree on an answer also agree on the reasoning. Applying CARA to a standard debate system on two medical QA benchmarks (MedQA-USMLE and MedThink-Bench), we identify the **consistency illusion**: a failure mode where debate reduces detectable contradictions between agents while simultaneously decreasing the semantic similarity of their reasoning chains; agents appear to agree more but reason less consistently. To improve this misalignment, we propose the **Grounded Debate Protocol** (GDP), a prompt-level intervention that requires agents to commit to named medical facts and take explicit stances on other agents' claims. GDP produces large, consistent alignment improvements (Cohen's $d = +1.43$ to $+1.99$) across two datasets and two backbone models, without adding LLM calls or modifying system architecture. Our results motivate cross-agent reasoning alignment as a quantity to audit alongside accuracy in safety-critical domains.

1 Introduction

Multi-agent LLM systems are increasingly used for medical question answering. Frameworks such as MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), and MCC (Sun et al., 2026) deploy several LLM instances that debate, consult, and vote to answer clinical questions, and report accuracy gains over single-agent baselines on benchmarks such as MedQA. Because the dominant evaluation pipeline primarily reports macroscopic answer accuracy under majority vote, the underlying *trust assumption* is straightforward: if multiple

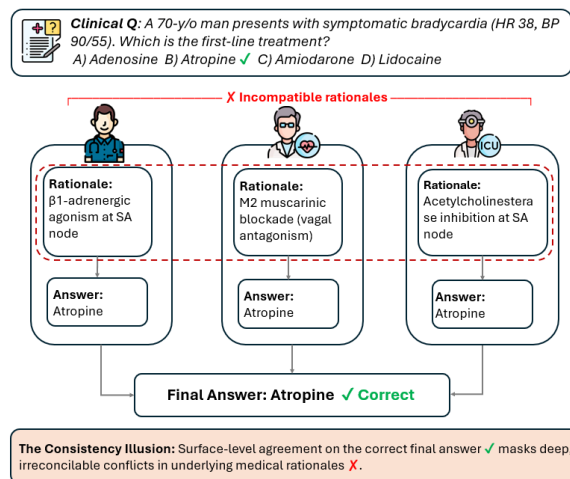


Figure 1: The **consistency illusion**: three agents independently agree on the clinically correct answer (atropine for symptomatic bradycardia), yet their rationales invoke three mutually exclusive pharmacological targets. Empirical 2D visualization on D2 (MedThink-Bench) appears in Appendix N.

agents independently converge on the same answer, that answer is reliable.

We argue this trust assumption is incomplete: answer-level consensus does not entail *reasoning-level* alignment. Figure 1 illustrates the gap on a high-stakes case: three agents independently agree that atropine is the first-line treatment for symptomatic bradycardia, yet their underlying rationales invoke three mutually exclusive pharmacological targets (β_1 -adrenergic agonism, M_2 -muscarinic blockade, and acetylcholinesterase inhibition). The agents reach the correct answer through medically incompatible reasoning; we term this pattern the **consistency illusion**. This paper therefore measures **cross-agent reasoning alignment**: whether agents that converge on an answer also align in the reasoning that produced it. This question is orthogonal to both macroscopic answer accuracy and individual, single-agent reasoning faithfulness.

This phenomenon has gone undetected because existing evaluations measure something else. Medical multi-agent systems report accuracy and answer agreement, while reasoning-faithfulness methods were built to audit a *single* agent’s trace rather than alignment *across* agents; to our knowledge, no prior method measures whether agents that agree on an answer also agree on the reasoning behind it. Recent theoretical critiques of multi-agent deliberation (Choi et al., 2025; Zhu et al., 2025) characterize debate dynamics as a voting-driven martingale rather than genuine synthesis. They leave open a complementary question central to safety-critical settings: when agents *do* agree, does the reasoning behind that agreement hold up?

To answer this question, we introduce **CARA** (Cross-Agent Reasoning Alignment), a family of automated metrics that decompose each agent’s response into reasoning steps and score pairwise alignment among agents that share an answer, using a hybrid of NLI-based contradiction detection and embedding similarity. CARA introduces no additional LLM calls and operates post hoc on traces produced by ordinary system inference. Applying CARA to a standard three-agent debate system on MedQA-USMLE ($N=500$) and MedThink-Bench ($N=499$), we find that debate reduces explicit contradictions between agents ($CR \downarrow$) while *simultaneously* reducing the semantic similarity of their reasoning chains ($SIM \downarrow$). Debate produces surface harmony while reasoning diverges further: the empirical signature of the consistency illusion. The effect is pronounced in open-ended reasoning, where the unconstrained reasoning space gives divergence room to surface: on MedThink-Bench it is large and replicates across both backbones (Cohen’s $d=-0.30$ for Qwen, -1.32 for Llama-3), and it is muted on closed-form multiple-choice (MedQA), consistent with the narrower reasoning surface.

To test whether this failure is fundamental to multi-agent debate or merely a protocol-level deficiency, we propose the **Grounded Debate Protocol** (GDP), a prompt-level intervention requiring each reasoning step to pair a CLAIM with a named GROUND and forcing explicit STANCE statements toward other agents’ claims. GDP produces large, consistent alignment gains across all three conditions tested (Cohen’s $d=+1.43$ on D1 Qwen, $+1.62$ on D2 Qwen, and $+1.99$ on D2 Llama-3). Together, our results identify a previously unmeasured failure mode in multi-agent med-

ical QA, name and quantify it, and demonstrate that lightweight prompt-level grounding substantially improves cross-agent reasoning alignment. CARA and GDP are designed as diagnostic and auditing instruments for reasoning alignment in multi-agent medical QA under symmetric debate: they audit the reasoning behind consensus alongside accuracy, rather than serving as general correctness measures or domain-agnostic methods. We release the code and computed CARA scores publicly¹.

2 Related Work

Multi-agent systems for medical QA. A growing family of medical MAS—MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), MCC (Sun et al., 2026), TeamMedAgents (Mishra et al., 2025), the foundational debate framework of Du et al. (2024), and clinical-simulation systems like Agent Hospital (Li et al., 2024a) and Agent-Clinic (Schmidgall et al., 2026)—deploy multiple LLM instances that debate, consult, or vote on medical questions, reporting accuracy gains over single-agent baselines on benchmarks such as MedQA. A common thread unites these architectures: they treat multi-agent answer-level consensus as a proxy for reliable clinical reasoning. None, however, audits whether agents that converge on the same option actually share consistent underlying rationales.

Critiques of multi-agent deliberation. A growing body of work questions whether multi-agent debate improves reasoning beyond what voting or repeated sampling already provides. Choi et al. (2025) prove formally that debate dynamics form a martingale—gains are attributable to majority voting, not deliberation—a conclusion reinforced empirically by Zhu et al. (2025), Li et al. (2024b), Wang et al. (2024), Tran and Kiela (2026), and Zhang et al. (2025). Behavioral studies document specific failure modes: agents may reach consensus by copying explanations (Pitre et al., 2025), switch answers under sycophantic pressure (Yao et al., 2025), or degrade in multi-turn dialogue (Laban et al., 2026). Most closely related, the problem drift literature (Becker et al., 2026) measures how debate performance degrades over turns; CARA asks a complementary question: when agreement does emerge, is the underlying reasoning aligned? Prior critiques attack the debate *process*; our work

¹Code and data: <https://github.com/wangxiaoyang0412/consistency-illusion-code>

attacks an assumption about the debate *outcome*—that surface consensus implies reasoning alignment.

Evaluation of multi-agent reasoning. A distinct evaluation lineage measures whether reported rationales support or reflect a single agent’s answer. Lanham et al. (2023) introduced CoT truncation and mistake-injection probes; Matton et al. (2025) extended these to medical QA, and Afolabi et al. (2026) show that reasoning steps often do not causally drive predictions. Automated metrics include ROSCOE (Golovneva et al., 2023), NLI-based faithfulness checking (Falke et al., 2019), and CC-SHAP (Parcalabescu and Frank, 2024), but these are designed for single-trace verification of individual model outputs. Even the methods that do apply to multi-agent settings verify each trace in isolation; to our knowledge, none measures cross-agent reasoning alignment among agents that share an answer. How multi-agent aggregation affects collective reasoning coherence has therefore remained uncharacterized.

3 CARA: Measuring Reasoning Alignment

CARA measures whether agents that agree on an answer also agree on the reasoning behind it—a quantity that answer accuracy and single-agent faithfulness leave unmeasured. It is *orthogonal* to correctness: rather than checking whether a rationale is medically right, it asks whether agents who reach the same answer reach it for compatible reasons.

3.1 Agreement Set

We run N agents on a clinical question q . Each agent A_i produces an answer a_i and a reasoning chain $R_i = (r_i^1, \dots, r_i^{K_i})$ of K_i steps. The system answer $\hat{a}$ is the majority vote over $\{a_i\}$. We define the **agreement set** as the agents who voted for $\hat{a}$:

$$\mathcal{S}(q) = \{i : a_i = \hat{a}\}. \quad (1)$$

We compute CARA only when $|\mathcal{S}(q)| \geq 2$. This restriction is the point: it isolates the safety-critical case where surface agreement may hide divergent reasoning.

3.2 Step-Level Alignment

We split each rationale R_i into steps using a deterministic text-segmentation pipeline. For each ordered pair (A_i, A_j) in the agreement set with $i \neq j$,

we build an alignment matrix $\mathbf{M}_{ij} \in \mathbb{R}^{K_i \times K_j}$ with entries $\mathbf{M}_{ij}[k, l] = \text{align}(r_i^k, r_j^l)$.

Our metric is CARA-HYB; CARA-NLI and CARA-SIM are its two components, reported separately in Table 9 to expose what drives alignment. **CARA-NLI** returns an NLI label in $\{-1, 0, +1\}$ for contradiction, neutral, or entailment. **CARA-SIM** returns the cosine similarity of step embeddings. **CARA-HYB** combines both, letting contradictions override similarity:

$$\text{align}_{\text{hyb}}(r_i^k, r_j^l) = \begin{cases} -1 & \text{if } P_{\text{NLI}} > \tau, \\ \cos(\mathbf{e}_i^k, \mathbf{e}_j^l) & \text{otherwise,} \end{cases} \quad (2)$$

where P_{NLI} is the NLI contradiction probability for (r_i^k, r_j^l) , $\mathbf{e}_i^k$ is the L2-normalized embedding of r_i^k , and τ is the contradiction threshold. This hybrid addresses the known insensitivity of embeddings to negation (Marelli et al., 2014). For validation, **CARA-LLM** uses GPT-4o as an LLM judge to calibrate CARA-HYB against holistic human-style judgments, reported in Section 5.3.

3.3 Question-Level Aggregation and Diagnostics

Best-match alignment. Because agents produce different numbers of steps ($K_i \neq K_j$), alignment is asymmetric. We define **best-match** alignment by matching each step in A_i to its best match in A_j :

$$\text{match}_j(r_i^k) = \max_{l \in \{1, \dots, K_j\}} \text{align}(r_i^k, r_j^l). \quad (3)$$

Symmetrized pairwise score. The pairwise CARA score symmetrizes by averaging both directions:

$$\text{CARA}(i, j) = \frac{1}{2} \left[\frac{1}{K_i} \sum_{k=1}^{K_i} \text{match}_j(r_i^k) + \frac{1}{K_j} \sum_{l=1}^{K_j} \text{match}_i(r_j^l) \right]. \quad (4)$$

Question-level aggregation. The question-level score averages over all pairs in the agreement set:

$$\text{CARA}(q) = \frac{2}{|\mathcal{S}(q)| (|\mathcal{S}(q)| - 1)} \times \sum_{\substack{i < j \\ i, j \in \mathcal{S}(q)}} \text{CARA}(i, j). \quad (5)$$

Variants with negative range (CARA-NLI, CARA-HYB) are rescaled to $[0, 1]$ via $(\text{CARA}+1)/2$. Corpus-level scores are the mean of $\text{CARA}(q)$ over questions with $|\mathcal{S}(q)| \geq 2$.

Contradiction Rate (CR). We also report the **Contradiction Rate**: the fraction of best-matched step pairs in the agreement set flagged as contradictions by the NLI filter.

$$CR(q) = \frac{|\{(i, j, k) : i \neq j, \text{match}_j(r_i^k) = -1\}|}{\sum_{i \in \mathcal{S}(q)} (|\mathcal{S}(q)| - 1) K_i} \quad (6)$$

True alignment shows up as semantic similarity rising (CARA-SIM $\uparrow$) with contradictions falling (CR $\downarrow$). When the two diverge—contradictions fall but similarity also falls—we see the consistency illusion.

4 Grounded Debate Protocol: A Protocol-Level Intervention

4.1 Motivation and Protocol Design

CARA shows that free-form debate can produce answer agreement without aligned reasoning, a result we establish in Section 6. GDP tests whether this failure is an unresolvable feature of language-model interaction or merely a protocol-level deficiency. Standard debate prompts allow three failure modes: vague clinical claims; sycophantic answer adoption without engaging with peer evidence (Yao et al., 2025); and explanation copying with topic drift (Pitre et al., 2025; Becker et al., 2026). The common root is that nothing in standard prompts requires agents to commit to specific reasoning steps or substantively engage with other agents’ claims.

To address this, GDP enforces a structured, prompt-level communication protocol. It requires each agent to structure its output into up to three machine-parseable fields:

- **CLAIM**: A singular, atomic, and falsifiable clinical assertion.
- **GROUND**: a named medical fact, mechanism, or guideline that supports the claim (e.g., disease, drug target, receptor pathway).
- **STANCE** (*debate rounds only*): one of {AGREE, DISAGREE, EXTEND} toward a specific claim from another agent, with a one-sentence justification; a DISAGREE requires a counter-GROUND.

Agents emit CLAIM+GROUND pairs in the independent round (r_0) and add STANCE in the debate round (r_1). An anti-sycophancy rule instructs each agent to switch answers between r_0 and r_1 only after seeing a more compelling GROUND from another agent or finding a factual error in its own reasoning. A worked clinical example contrasting an unstructured debate log with a synchronized GDP trace is provided in Appendix K.

4.2 Integration Properties

GDP is a prompt-level change: the CLAIM/GROUND/STANCE format is added through the system message, with agent routing and the number of LLM calls unchanged. The only overhead is output length: tokens grow by $\sim 9\%$ from the explicit field markers. Applied to the symmetric debate of Du et al. (2024), this gives M3-GDP, which we evaluate in Section 5.1. Traces are parsed post-hoc by regex on the field markers; this parsing supports logging and analysis only, as CARA operates on the raw response text regardless of parse outcome.

5 Experimental Setup

5.1 Study Design

We evaluate whether debate-induced answer consensus is accompanied by reasoning alignment among agents that converge on the same answer. Table 10 summarizes our experimental matrix of three systems, each with three agents under majority voting: **M3** is symmetric one-round debate (Du et al., 2024), producing pre-debate (r_0) and post-debate (r_1) traces; **M4** is an independent-vote control with no cross-agent communication, specifically isolating the sampling-and-voting contribution from debate (Choi et al., 2025); **M3-GDP** keeps the M3 architecture and the same LLM-call budget but rewrites all prompts to require the GDP CLAIM+GROUND+STANCE format (Section 4). Full configurations and prompts are in Appendix J.

5.2 Datasets and Backbones

D1: MedQA-USMLE (Jin et al., 2021) provides closed-form four-option USMLE-style multiple-choice questions; we sample $N=500$ stratified by exam stage with a fixed seed. **D2: MedThink-Bench** (Zhou et al., 2025) spans ten medical domains with 3–10 answer options and expert-annotated reasoning trajectories (“Scoring Points”); we use $N=499$ of 500 available items (one excluded for incompatible option format). D2’s wider option space gives reasoning more room to diverge than D1. Primary trials use Qwen 2.5 72B Instruct (Qwen et al., 2025); we replicate on D2 ($N=100$) with Llama 3.3 70B Instruct (Grattafiori et al., 2024) to test the effect-direction consistency rather than cross-backbone statistical significance.

5.3 CARA Computation and Validation

CARA-HYB implementation. CARA-HYB decomposes each agent response into reasoning steps and computes pairwise alignment via two complementary signals: an NLI hard-filter on contradictions (DeBERTa-v3 (Laurer et al., 2023)) and a sentence-embedding cosine on non-contradictory step pairs (Stella (Zhang et al., 2024)). The pipeline runs post hoc with no additional LLM calls; model versions, threshold, decomposition rules, and runtime are in Appendix E.

Agent-expert validation (CARA-GT). For D2 only, CARA-GT applies the same pipeline with one agent replaced by the expert Scoring-Points reference, measuring agent-to-expert alignment and *expert coverage*, detailed in Appendix C.

LLM-judge calibration (CARA-LLM). CARA-LLM validates CARA-HYB against GPT-4o-as-judge (OpenAI, 2024) on a sample of agent pairs: the scores show strict tercile monotonicity (and a moderate linear correlation), establishing convergent validity across two operationalizations of alignment—step-level semantic overlap (CARA-HYB) vs. holistic logical coherence (GPT-4o judge). Sample design, correlation values, and tercile statistics are in Appendix F.

Human-expert validation. Because reliable annotation of medical reasoning requires medically-trained experts, we validate on a deliberately focused set of 50 agreement-set pairs rated by three independent annotators, with the full setup in Appendix H. Inter-annotator agreement is high (Spearman $\rho \in [0.93, 0.99]$; 100% within-1 agreement), and aggregate ratings correlate positively with CARA-HYB (Spearman $\rho=0.26$) and CARA-SIM ($\rho=0.31$), with tercile means rising monotonically (4.00/4.65/4.71 for low/mid/high CARA-HYB strata). The bounded magnitude reflects range restriction within the agreement set (CARA-HYB $\in [0.76, 0.98]$), paralleling the CARA-LLM pattern.

5.4 Statistical Analysis

We report Cohen’s d with 95% percentile bootstrap CIs ($B=10,000$, seed 42); $|d|>0.8$ is Tier A, $0.5<|d|\leq 0.8$ Tier B, and $0.2<|d|\leq 0.5$ Tier C. Accuracy comparisons use continuity-corrected McNemar tests. Questions with fewer than two agents choosing the same answer are undefined; we report this undefined rate per system and check survivorship bias by imputing the undefined cases. Per-cell counts are in Appendix M, and the full imputation

in Appendix B and Section 6.2.

6 Results

We report CARA-HYB, CARA-SIM, and CR for all three systems on D1 and D2 with Qwen 2.5 72B, plus a cross-backbone replication on D2 ($N=100$) with Llama 3.3 70B. All comparisons are within-dataset and within-backbone; we do not compare absolute CARA values across datasets.

6.1 Main Results and Dual-Layer Decomposition

The rank ordering is identical on D1 and D2: M3-GDP r1 is the highest cell (0.952/0.953), and M3 r1 is the *lowest* (0.892/0.881)—standard debate alignment falls below independent voting.

For M3 from r0 to r1, *both* CR and SIM decrease (CR 0.104 $\rightarrow$ 0.035 on D1, 0.144 $\rightarrow$ 0.054 on D2; SIM 0.801 $\rightarrow$ 0.787 on D1, 0.801 $\rightarrow$ 0.769 on D2); fewer contradictions with less semantic alignment is the *consistency illusion* (mechanism in Section 7).

In contrast, GDP debate produces the genuine-alignment signature. From r0 to r1, SIM *rises* (0.835 $\rightarrow$ 0.912 on D1; 0.836 $\rightarrow$ 0.914 on D2) while CR *falls* (0.117 $\rightarrow$ 0.098 on D1; 0.137 $\rightarrow$ 0.127 on D2). Both protocols reduce contradictions, but only GDP simultaneously increases reasoning similarity—CR $\downarrow$ with SIM $\uparrow$, the true-alignment pattern defined in Section 3 and the mirror image of the illusion. A D1 bootstrap test underscores that M3’s low CR is *anomalous*: GDP r1 CR exceeds M3 r1 ($\Delta=+0.063$, CI [+0.042, +0.087]), and M3 r1 CR falls even below independent voting ($\Delta=-0.072$ vs. M4, CI [−0.089, −0.056])—standard debate suppresses visible contradictions without aligning reasoning.

6.2 GDP Effect and the Consistency Illusion

Figure 2 shows the three key pairwise comparisons as a forest plot of Cohen’s d across all three dataset-backbone combinations.

GDP delivers a large alignment gain on both datasets: $d=+1.43$ on D1 and $d=+1.62$ on D2 against M3_{r1}, rising to $d=+2.26$ and $d=+2.22$ against independent voting (all four CIs above zero). Standard debate, in contrast, is alignment-neutral on D1 ($d=-0.08$, CI contains zero) and alignment-degrading on D2 ($d=-0.30$, CI [−0.43, −0.17]). The D1–D2 gap tracks both question complexity and undefined-question selection bias. The GDP effect splits in two:

System	Round	D1 (MedQA, $N=500$)			D2 (MedThink, $N=499$)		
		HYB	SIM	CR	HYB	SIM	CR
M3-GDP	r1 (debate)	0.952	0.912	0.098	0.953	0.914	0.127
M3-GDP	r0 (indep.)	0.912	0.835	0.117	0.912	0.836	0.137
M4	r0 (indep.)	0.895	0.801	0.108	0.895	0.801	0.161
M3	r0 (indep.)	0.895	0.801	0.104	0.896	0.801	0.144
M3	r1 (debate)	0.892	0.787	0.035 [†]	0.881	0.769	0.054 [†]

Table 1: CARA-HYB, CARA-SIM, and CR per cell on D1 and D2. Higher HYB and SIM, and lower CR, indicate better alignment. **Green row** = M3-GDP r1 (highest alignment); **Red row** = M3 r1 (lowest, consistency illusion). Per-cell HYB SDs are 0.021–0.059. N ranges 424–500 (D1) and 428–499 (D2) due to undefined questions. [†]M3 r1 CR is deflated by survivorship bias (Section 6.2).

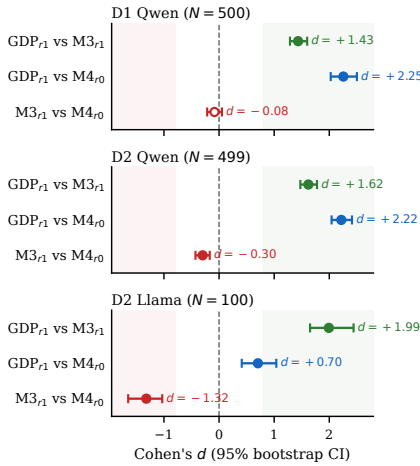


Figure 2: Cohen’s d for three key pairwise comparisons, one panel per dataset–backbone. Error bars: 95% bootstrap CIs ($B=10,000$); filled markers exclude zero; shading marks Tier A ($|d|>0.8$). GDP comparisons (top two rows) reach Tier A across all conditions; the consistency-illusion comparison (bottom) is significantly negative on D2 (both backbones).

CLAIM+GROUND format alone (GDP_{r0} vs. M4_{r0}) yields $d=+0.62$ on D1; adding STANCE-mediated debate (GDP_{r1} vs. GDP_{r0}) yields $d=+1.51$, $2.4\times$ the format component but only meaningful when there are structured claims to reference. On survivorship-bias sensitivity, M3 r1 excludes 15.2% (D1) and 14.2% (D2) of questions where no majority emerged. These exclusions are non-random—harder questions survive less often—so the sample skews toward easier cases. Under worst-case imputation (CARA-HYB = 0.50, CR=1.0 for every undefined question, both far beyond any observed value), M3 r1’s adjusted CARA-HYB stays below M4 r0 (0.832 vs. 0.895 on D1; 0.827 vs. 0.895 on D2): the consistency-illusion conclusion is robust. The CR finding, however, reverses (0.182>0.108 on D1; 0.188>0.161 on D2), so we

interpret CR alongside the undefined rate.

6.3 Cross-Backbone Validation

We replicate all three systems on D2 ($N=100$) with Llama 3.3 70B Instruct. All five key comparisons agree in sign across both backbones (Table 14): GDP is Tier A positive (Llama $d=+1.99$, Qwen $d=+1.62$), the illusion is negative (Llama $d=-1.32$, Tier A; Qwen $d=-0.30$, Tier C), standard debate degrades alignment $r0\rightarrow r1$ while GDP debate improves it, and the format effect is positive. The illusion is stronger on Llama partly because Llama’s M3 r1 undefined rate (2%) is far lower than Qwen’s (14%), so more hard cases survive into the measured sample.

6.4 Agent–Expert Alignment (CARA-GT)

CARA-GT replaces one agent with the expert Scoring-Points reference (D2 only; Section 5.3). It mirrors the inter-agent finding: GDP moves agents *toward* expert reasoning ($d=+0.32$, CI $[+0.20, +0.45]$), and standard debate moves them *away* ($d=-0.21$, CI $[-0.34, -0.08]$). Expert-coverage tells a sharper story (Figure 3): standard debate loses 8.6pp of expert scoring-point coverage; GDP recovers 5.4pp.

6.5 Accuracy

CARA measures reasoning alignment, not answer correctness; majority-vote accuracy is reported separately. Standard debate does not change accuracy on either dataset (D1 $\Delta=+0.4\text{pp}$, D2 $\Delta=0.0\text{pp}$; both $p>0.85$), consistent with the martingale characterization (Choi et al., 2025). GDP shifts accuracy by -2.4pp on Qwen (both D1 and D2) and -6.0pp on Llama D2; none reaches significance at $\alpha=0.05$. D2’s lower absolute accuracy ($\sim 41\%$ baseline) reflects its ~ 7.5 average answer options versus D1’s 4, giving an $\approx 3\times$ random-performance ratio. Case-level tracing of every answer change (RQ5,

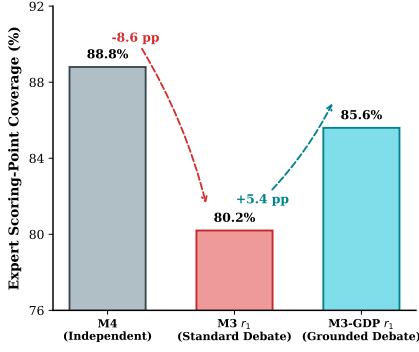


Figure 3: Expert scoring-point coverage on D2: standard debate (M3 r_1) drops 8.6pp from the independent-voting baseline; GDP (M3-GDP r_1) recovers 5.4pp.

Section 7; Appendix L) localizes where these shifts originate.

7 Analysis

We answer four research questions about the consistency illusion, the mechanism by which GDP improves it, and the generality of the findings.

RQ1: What mechanism produces the consistency illusion, and how does GDP improve it? **Answer:** *Standard debate creates the illusion via two failures—contradiction smoothing without reasoning convergence, and sycophantic convergence—and GDP reverses both via a format effect ($d=+0.62$) amplified by a STANCE-mediated debate interaction effect ($d=+1.51$, $2.4\times$ the format component; extended derivation in Appendix O).*

Standard debate drives $CR \downarrow + SIM \downarrow$ (Section 6.1) via two mechanisms. First, revision pressure removes contradictory steps without replacing them with shared reasoning—*subtraction without alignment*—so CR drops while SIM also drops. Second, agents adopt each other’s answers without adopting the underlying reasoning, sometimes severely: M3 r_1 fails consensus on 14–15% of questions on both datasets. GDP reverses both via two complementary effects. The *format effect* alone, before any debate (GDP $_{r0}$ vs. M4 $_{r0}$ on D1: $SIM +0.034$, $d=+0.62$), shows that much of the apparent misalignment originates from underspecified output format rather than genuine reasoning disagreement. The dominant *debate interaction effect* ($d=+1.51$ on D1) operates because STANCE statements force agents to share reasoning vocabulary and logical structure with their interlocutors, driving SIM up by $+0.077$ on D1 and $+0.078$ on

D2 from r_0 to r_1 . Format is a precondition: without structured claims, STANCE has no targets.

RQ2: Beyond CARA, what behavioral indicators support the illusion finding? **Answer:** *Standard debate destabilizes consensus $\sim 38\times$ more than independent voting on D1 and produces 5% “reasoning collapse” zero-step agents on D2; GDP eliminates both.*

Two independent behavioral indicators support the illusion finding without depending on agreement-set selection. First, the *undefined rate* (no majority emerges) jumps from 0.4% before debate (M3 r_0) to 15.2% after (M3 r_1) on D1 ($38\times$) and $5.6\% \rightarrow 14.2\%$ on D2: debate itself manufactures disagreement; grounding removes it (M3-GDP r_1 : 0.4%). Second, *reasoning collapse*—agents emitting zero extractable steps after debate, the behavioral signature of sycophantic convergence—affects 5.0% of M3 r_1 agreement-set memberships on D2 ($N=100$) vs. 0.3% for M3-GDP r_1 . GDP’s CLAIM+GROUND requirement prevents both failure modes.

RQ3: Which failure modes does GDP structurally eliminate? **Answer:** *GDP eliminates FM1 (complementary reasoning) and FM4 (sycophantic convergence) on both datasets, shifting residual misalignment toward the low-priority baseline FM6 (partial overlap, 62–64% of remaining low-CARA-HYB cases).*

We examine the 50 lowest-CARA-HYB cases per system where each still reaches the correct answer, classifying by deterministic rules over CARA sub-metrics; the full classification rules are in Appendix L. Table 2 groups the six modes into HIGH (FM1, FM4), MEDIUM (FM2, FM3, FM5), and LOW (FM6) audit-priority tiers.

Four findings replicate across the two datasets. First, GDP **eliminates sycophantic convergence (FM4)**: $11 \rightarrow 0$ on D1, $15 \rightarrow 0$ on D2; the CLAIM+GROUND requirement structurally prevents empty answer adoption. Second, GDP **eliminates complementary reasoning (FM1)**: $23 \rightarrow 0$ on D1, $18 \rightarrow 0$ on D2; structured grounding forces convergence on shared medical facts. Expert-coverage analysis on D2 further resolves FM1 into benign complementarity (FM1a, 10/18, union coverage above baseline) and divergence-with-gaps (FM1b, 8/18), with GDP consolidating complementary evidence rather than suppressing it (coverage $0.856 \rightarrow 0.921$; Appendix L). Third, residual misalignment under GDP **shifts to FM6 (partial**

Audit priority	Mode	Definition	M3 r1		GDP r1	
			D1	D2	D1	D2
HIGH	FM1	Complementary reasoning: different, non-contradictory reasoning paths.	23	18	0	0
	FM4	Sycophantic convergence: agent adopts the majority answer with zero reasoning steps.	11	15	0	0
MEDIUM	FM2	Granularity mismatch: step-count ratio >2 (depth difference, not factual clash).	12	15	5	2
	FM3	Contradictory premises: agents cite conflicting facts but converge on the correct answer.	2	1	12	15
	FM5	Terminology divergence: different medical terms used for the same concept.	2	1	0	0
LOW	FM6	Partial overlap: residual baseline misalignment without extreme signals.	0	0	31	32
High-priority modes eliminated by GDP (FM1 + FM4):			34	33	0	0

Table 2: Failure mode taxonomy on the 50 lowest-CARA-HYB correct-answer cases per system, with definitions and counts on both datasets. GDP *eliminates* both high-priority modes (FM1 + FM4: 34/50 and 33/50 under M3 → 0/50 under GDP on D1 and D2 respectively), shifting residual misalignment toward the low-priority baseline FM6. FM3’s rise under GDP is consistent with the CR analysis in §6.1: the structured format surfaces contradictions that vague free text would hide.

overlap): 31/50 on D1, 32/50 on D2 (62–64%)—a low-priority baseline remaining after high-priority modes are removed. Fourth, **FM3 contradictory premises rises under GDP** (2→12 on D1, 1→15 on D2), consistent with the CR analysis in Section 6.1: the structured format surfaces contradictions that vague free text would hide.

A qualitative case study (MQA_0447 hypovolemic shock, where agents misstate SVR direction before self-correcting) is in Appendix L.

RQ4: Do the findings generalize across datasets and backbones? **Answer:** *Yes—the GDP effect replicates with Tier-A magnitudes across all three dataset–backbone conditions (Table 14), with the system rank ordering identical throughout.*

The rank ordering $M3\text{-}GDP_{r_1} \succ M3\text{-}GDP_{r_0} \succ M4_{r_0} \approx M3_{r_0} \succ M3_{r_1}$ is identical across all three dataset–backbone combinations. The illusion magnitude varies across conditions ($d = -0.08, -0.30, -1.32$) for two reasons. First, D1’s fixed 4-option format leaves less room for reasoning to diverge. Second, M3 r1 undefined rates differ (15.2%, 14.2%, 2.0%): a lower undefined rate leaves more hard cases in the sample, strengthening the measured illusion.

RQ5: Where do answer changes originate? **Answer:** *Case-level tracing shows answer changes concentrate before any agent interaction (23/31 on D1), carry unfounded-consensus signals at 2–4× the background contradiction rate, and involve the*

debate round—the source of the alignment gain—in at most 5/31 cases in the strictest reading.

CARA’s per-question diagnostics make answer-level shifts individually traceable. For every question that GDP r1 answers incorrectly while M3 r1 answers correctly (31 on D1, 36 on D2; the paired comparison also contains 17 and 24 changes in the opposite direction, giving the non-significant net differences in Table 6), we check two properties: whether the M3 consensus carried an *unfounded-consensus signal* (inter-agent contradiction, expert coverage below 0.75, or a zero-step agent), and at which round the GDP answer first diverged. The signals checkable on D1 flag 45% of overturned answers; on D2, where the expert-coverage signal is also available, the rate is 69%, and overturned answers show 2–4× the contradiction rate of answers GDP retains. Round-of-origin tracing sharpens the picture: 23 of 31 D1 differences are already present at the independent round under the CLAIM+GROUND prompt, before any agent interaction, and crossing the two dimensions yields a complete decomposition (Appendix L). The strictest residual scenario—an answer unflagged by any signal that survives GDP r0 and is overturned during grounded debate—occurs in 5 of 31 cases: the STANCE-mediated round that produces the alignment gain introduces almost none of the answer changes.

Is the gain just structural? Because CARA-HYB scores step-pair similarity, GDP’s structured

CLAIM/GROUND/STANCE format could in principle inflate scores mechanically—through shared template vocabulary or cleaner step parsing—rather than through better reasoning alignment. First, the format-only effect ($d=+0.62$) is less than half the full GDP gain; the dominant component ($d=+1.51$) is the STANCE-mediated debate interaction over content, not template. Second, CARA-GT (Section 6.4) shows GDP also aligns agents with the expert Scoring-Points reference ($d=+0.32$), which is not in GDP format. Third, GPT-4o judges M3-GDP r1 reasoning as more aligned than M3 r1 on the same 50 agreement-set pairs (4.86 vs. 4.50 on a 5-point scale, Cohen’s $d=+0.75$, 95% bootstrap CI on Δ $[+0.08, +0.65]$); since the LLM judge evaluates holistic logical coherence rather than template structure, this is direct evidence against pure structural confound. Fourth, stripping the literal CLAIM/GROUND/STANCE tokens from GDP traces and recomputing CARA-HYB attenuates paired Cohen’s d_z by only 6–10% across all three conditions (Appendix G). Fifth, a generic-structure control on the same 100 D2 questions (Appendix G) shows that structure alone does not reproduce the gain: numbered steps without named grounding or stance do not exceed the unstructured baseline at the independent round (0.863 vs. 0.894) and account for roughly 11% of GDP’s debate-round gain, while removing the anti-sycophancy clause shifts CARA-HYB by only 0.004.

8 Conclusion

We identified the *consistency illusion* in multi-agent medical QA: standard debate reduces detectable contradictions between agents while simultaneously decreasing the semantic similarity of their reasoning—agents appear to agree more but actually reason less consistently. We introduced CARA, a family of metrics for cross-agent reasoning alignment, and applied it on MedQA-USMLE and MedThink-Bench: the illusion is negligible on D1 ($d=-0.08$) but significantly negative on D2 (Qwen $d=-0.30$; Llama-3 $d=-1.32$). We proposed the Grounded Debate Protocol (GDP), a prompt-level intervention requiring CLAIM+GROUND+STANCE, which delivers Tier-A alignment gains across all conditions ($d=+1.43$ to $+1.99$), eliminates two failure modes (complementary reasoning and sycophantic convergence) on both datasets and adds no LLM calls. Together, our results challenge the assumption that multi-agent

consensus indicates reasoning reliability, and show that protocol-level grounding can close this gap, positioning CARA and GDP as diagnostic and auditing instruments for reasoning alignment in medical multi-agent QA under symmetric debate. In deployment, high CARA with high expert coverage marks corroborated consensus, while low CARA surfaces the cases that warrant audit. The structural sensitivity we observe—larger illusion where reasoning has more room to diverge—suggests grounding is most valuable on open-ended clinical tasks and proprietary backbones.

Limitations

While we evaluate CARA and GDP across two medical QA benchmarks (D1 MedQA-USMLE $N=500$ and D2 MedThink-Bench $N=499$) and two open-weight backbones (Qwen 2.5 72B and Llama 3.3 70B), with Tier-A effect sizes ($d=+1.43$ to $+1.99$) replicated across all conditions, our analysis has several limitations.

CARA’s agreement-set definition relies on discrete answer matching, which is well-defined for multiple-choice benchmarks but would require adaptation for free-form clinical generation tasks such as differential diagnosis or treatment-plan drafting; extending the metric to such settings is a natural next step. Our cross-backbone validation covers two open-weight 70B-class models at comparable scales; proprietary models (e.g., GPT-4o, Claude, Gemini) would require an identical-instance experimental setup that is outside the scope of this work.

CARA-HYB is a step-level semantic-alignment measure designed to operate within the agreement set; the GPT-4o oracle (Appendix F) and human-expert validation (Appendix H) anchor its interpretation, with both showing the predicted monotonicity across CARA-HYB terciles. Extending the metric to non-agreement-set settings (e.g., disagreement-pair audit) and to domains beyond clinical reasoning is left for future work.

CARA and GDP target cross-agent reasoning alignment, a property orthogonal to answer accuracy; GDP is accordingly intended as an alignment-auditing and -improving intervention, and we make no claim that it improves accuracy. Across both backbones and datasets, majority-vote accuracy under GDP does not differ significantly from the standard debate baseline (McNemar tests; Table 6)—consistent with prior evidence that multi-agent de-

bate itself does not reliably improve accuracy (Choi et al., 2025). The Llama 3.3 comparison is based on $N=100$ and is underpowered to detect small accuracy effects; an adequately powered analysis of the alignment–accuracy relationship is left for future work.

We study a single, widely-used debate protocol—the symmetric, single-round design of Du et al. (2024), the canonical multi-agent debate baseline. Whether the consistency illusion and GDP’s correction generalize to other topologies (asymmetric roles, additional rounds, larger agent sets) is an important direction for future work.

Ethics Statement

This study uses publicly available medical QA benchmarks (MedQA-USMLE and MedThink-Bench) that contain no patient-identifiable information. All LLM inference uses open-weight models (Qwen 2.5 72B, Llama 3.3 70B) running on institutional compute infrastructure. CARA-LLM validation uses the OpenAI API (GPT-4o) for 80 evaluation pairs. No clinical participants are involved. Three medically-trained annotators rated 50 agent-pair outputs from the agreement set for metric validation; this is an annotation task on model outputs and does not constitute human-subjects research.

We emphasize that CARA measures reasoning *alignment*, not reasoning *correctness*. High CARA scores indicate that agents reason consistently, not that they reason correctly. CARA should be used as a diagnostic tool alongside accuracy metrics, not as a replacement for clinical validation. Multi-agent medical QA systems evaluated in this work are research prototypes and are not intended for clinical deployment.

Acknowledgments

This work was supported in part by the National Science Foundation under the IIS-2235548, and by the Department of Defense under the Grant DoD W91XWH-05-1-023. This material is based upon work supported by (while serving at) the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- Halimat Afolabi, Zainab Afolabi, Elizabeth Friel, Jude Roberts, Antonio Ji-Xu, Lloyd Chen, Egheosa Ogbomo, Emiliomo Imevbore, Phil Eneje, Wissal El Ouahidi, Aaron Sohal, Alisa Kennan, Shreya Srivastava, Anirudh Vairavan, Laura Napitu, and Katie McClure. 2026. Faithful or just plausible? evaluating the faithfulness of closed-source LLMs in medical reasoning. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pages 1562–1591. PMLR.
- Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2026. Stay focused: Problem drift in multi-agent debate. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5068–5102, Rabat, Morocco. Association for Computational Linguistics.
- Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38. Spotlight.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.
- Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421. Originally arXiv:2009.13081 (2020).

- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. [LLMs get lost in multi-turn conversation](#). In *The Fourteenth International Conference on Learning Representations (ICLR)*. Oral.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.
- Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024b. [More agents is all you need](#). *Transactions on Machine Learning Research*.
- Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. [SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment](#). In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.
- Katie Matton, Robert Ness, John Gutttag, and Emre Kiciman. 2025. [Walk the talk? measuring the faithfulness of large language model explanations](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Pranav Pushkar Mishra, Mohammad Arvan, and Mohan Zalake. 2025. TeamMedAgents: Pareto-efficient multi-agent medical reasoning through teamwork theory. *arXiv preprint arXiv:2508.08115*.
- OpenAI. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Letitia Parcalabescu and Anette Frank. 2024. [On measuring faithfulness or self-consistency of natural language explanations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [AgentClinic: a multimodal benchmark for tool-using clinical AI agents](#). *npj Digital Medicine*, 9(1).
- Xinti Sun, Qiyang Hong, Mengyan Zhang, Yuyan Li, Tingwei Chen, Zigeng Huang, Guihan Liang, Wenjun Tang, Sulin Xu, Xiaolin Ni, Junling Pang, Peixing Wan, and Erping Long. 2026. [Model confrontation and collaboration: A debate intelligence framework for enhancing medical reasoning in large language models](#). *Cell Reports Medicine*, 7(1).
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.
- Dat Tran and Douwe Kiela. 2026. [Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets](#). *arXiv preprint arXiv:2604.02460*.
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. [Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.
- Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. [Peacemaker or troublemaker: How sycophancy shapes multi-agent debate](#). *Preprint*, arXiv:2509.23055.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025. [Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity](#). *Preprint*, arXiv:2502.08788.

Shuang Zhou, Wenya Xie, Jiayi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. 2025. [Automating expert-level medical reasoning evaluation of large language models](#). *npj Digital Medicine*.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. [MedAgentBoard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks](#). *Advances in Neural Information Processing Systems*, 38.

A Sample-Size Stability

A practical concern for any automated metric is whether scores are stable as the sample grows. Table 3 compares earlier pilot CARA-HYB estimates (D1 $N=200$, D2 pooled $N=100$) against the full sample (D1 $N=500$, D2 $N=499$). All ten cell-round-dataset combinations show deltas within ± 0.005 (maximum: 0.005 for M3 r0 on D2), with five of ten cells unchanged at the third decimal place. This stability has two implications. First, the metric is well-conditioned at the operational sample sizes used in the paper—further scaling would not change the qualitative conclusions. Second, the pilot estimates were already reliable point estimates rather than artifacts of small-sample noise; the decision to scale up at $N=200$ accurately anticipated the final $N=500$ effect sizes.

System	Round	D1		D2	
		Pilot	Δ	Pilot	Δ
M4	r0	0.895	+0.000	0.894	+0.001
M3	r0	0.896	-0.001	0.891	+0.005
M3	r1	0.891	+0.000	0.878	+0.003
M3-GDP	r0	0.914	-0.002	0.913	-0.001
M3-GDP	r1	0.951	+0.001	0.951	+0.001

Table 3: Pilot $\rightarrow$ full-sample CARA-HYB stability. All deltas are within ± 0.005 on both datasets, confirming that CARA-HYB is not sensitive to sample fluctuations at this scale.

B Survivorship-Bias Sensitivity

M3 r1 excludes the 76/500 (15.2%) of D1 and 71/499 (14.2%) of D2 questions where no majority answer emerged after debate. These exclusions are non-random: harder questions are more likely to produce post-debate disagreement, so the surviving sample is biased toward inherently easier cases with higher baseline agreement. The exclusion mechanism differs by system—M3-GDP retains 99.6% of questions on both datasets—so direct comparison of M3 r1 against M3-GDP r1 (or against M4 r0, where M4 retains all D1 questions) involves comparing differently-selected populations. We assess sensitivity through imputation: for the undefined questions, we impute plausible CARA-HYB and CR values and recompute the cell mean. Three scenarios are reported in Table 4: *observed* (current measurement, excluding undefined), *neutral* (impute the M4 baseline value of that metric), and *worst-case* (impute CARA-HYB = 0.50 and CR = 1.0—both far beyond any value observed in any cell on either dataset).

Two conclusions are robust across both datasets. First, the CARA-HYB finding is robust: under all imputation scenarios—including the worst case, which assigns CARA-HYB = 0.50 (lower than any observed sample), M3 r1’s adjusted score remains at or below M4 r0. The consistency-illusion conclusion (M3 debate does not improve, and on D2 significantly degrades, alignment) therefore does not depend on the survivorship-biased sample. Second, the CR finding is fragile in the opposite direction: under worst-case imputation, M3 r1’s adjusted CR *exceeds* M4 r0’s on both datasets (0.182 > 0.108 on D1; 0.188 > 0.161 on D2), so M3 r1’s apparently low observed CR is at least partly an artifact of survivorship. Practically, this means CR should be interpreted alongside the undefined rate, not in isolation, and the consistency-illusion claim rests primarily on the joint movement of CR and SIM in the main CARA-HYB result (Section 6.1) rather than on CR alone.

C CARA-GT Per-Cell Results

D2 includes expert-annotated reasoning steps (*Scoring Points*) of 2–6 steps per question (mean 3.1), authored by clinical experts. CARA-GT measures alignment between each agent’s reasoning and the expert reference, computed via the same NLI+embedding pipeline used for CARA-HYB but with one agent replaced by the expert ref-

Scenario	D1 ($N=500$)		D2 ($N=499$)	
	HYB	CR	HYB	CR
M4 baseline	0.895	0.108	0.895	0.161
Observed (M3 r1)	0.892	0.035	0.881	0.054
Neutral impute	0.892	0.046	0.883	0.069
Worst-case	0.832	0.182	0.827	0.188

Table 4: Sensitivity of M3 r1 to imputation of undefined questions. *Neutral* imputes the M4 baseline value; *Worst-case* imputes HYB=0.50 and CR=1.0. The HYB conclusion (M3 r1 $\leq$ M4) is robust on both datasets; the CR direction reverses under worst-case on both datasets.

erence. Where CARA-HYB asks whether agents reason consistently *with each other*, CARA-GT asks whether they reason consistently *with experts*. CARA-GT is undefined when an agent produces zero extractable steps; this affects 35/2,495 pairwise records (1.4%), much cleaner than CARA-HYB’s 156/2,495 error rate (which additionally excludes agents who disagree on the final answer). Table 5 reports per-cell CARA-GT-HYB scores and expert-coverage rates on D2 ($N=499$).

GDP improves expert alignment; standard debate degrades it. The pairwise comparisons mirror the inter-agent CARA-HYB pattern. GDP_{r1} vs. M3_{r1}: Δ GT-HYB=+0.011, d =+0.32, 95% CI=[+0.20, +0.45] (excludes zero). M3_{r1} vs. M4_{r0}: Δ =−0.007, d =−0.21, CI=[−0.34, −0.08] (excludes zero). The dual evidence—GDP improves alignment with respect to experts, and standard debate degrades it—supports the interpretation that the inter-agent illusion has a substantive directional component: under standard debate, agents drift not merely apart from each other but *away* from the expert reasoning standard.

Expert coverage tells a sharper story. Coverage is the fraction of expert scoring points matched by at least one agent step at the contradiction-filtered alignment threshold. Standard debate loses 8.6 percentage points of expert coverage relative to independent voting (M4: 88.8% $\rightarrow$ M3 r1: 80.2%, CI=[−11.2, −6.1]): roughly one in eleven expert steps is dropped during debate. GDP debate recovers 5.4pp of that loss (M3 r1: 80.2% $\rightarrow$ M3-GDP r1: 85.6%, CI=[+2.7, +8.1]). Standard debate *drops* expert scoring points; GDP debate *keeps* them. This coverage gap is independent of the inter-agent CARA-HYB measurement and provides a second, expert-grounded line of evidence for the same conclusion: GDP produces reasoning that is

System	Round	GT-HYB	Coverage	N
M3-GDP	r1	0.813	85.6%	499
M3	r0	0.810	87.4%	499
M4	r0	0.809	88.8%	499
M3-GDP	r0	0.807	88.2%	499
M3	r1	0.802	80.2%	464

Table 5: CARA-GT on D2 ($N=499$): alignment between agent reasoning and expert *Scoring Points*. Coverage is the fraction of expert scoring points that match at least one agent step.

more consistent with both other agents and external standards.

D Accuracy and McNemar Tests

CARA measures reasoning alignment, not answer correctness, but we separately report majority-vote accuracy for context. Accuracy is computed by extracting the majority answer per agent at each round, then comparing to the gold answer with continuity-corrected McNemar tests on paired binary outcomes (correct/incorrect per question) and paired percentile bootstrap CIs ($B=10,000$, seed 42) on the accuracy delta. Table 6 reports majority-vote accuracy for all dataset–backbone combinations.

Findings. Standard debate (M3 r1) does not significantly change accuracy on either Qwen dataset (D1: Δ =+0.4pp, p =0.86; D2: Δ =0.0pp, p =0.90), consistent with the martingale characterization of debate dynamics: gains from sampling and voting are already realized by the independent baseline. D2 baseline accuracy ($\sim 41\%$) is much lower than D1 ($\sim 84\%$) because D2 questions have a mean of ~ 7.5 answer options (random baseline $\sim 13\%$), making $\sim 41\%$ approximately $3\times$ random on this benchmark—i.e., the model is performing non-trivially despite the lower headline number. Llama D2 accuracy ($N=100$, 49.0%) and Qwen D2 ($N=499$, 41.5%) are computed on different sample sizes of the same dataset, so we do not draw cross-backbone accuracy conclusions from this comparison.

E CARA-HYB Implementation

NLI component. DeBERTa-v3-large fine-tuned on MNLI, FEVER, ANLI, LingNLI, and WANLI (Laurer et al., 2023) (304M parameters). We threshold the contradiction probability at $\tau=0.7$ for the hybrid filter.

Dataset	Backbone	M4	M3 r1	GDP r1	Δ_{GDP}
D1 MedQA ($N=500$)	Qwen	83.8%	84.2%	81.4%	$-2.4 \text{ pp}^{\text{ns}}$
D2 MedThink ($N=499$)	Qwen	41.5%	41.5%	39.1%	$-2.4 \text{ pp}^{\text{ns}}$
D2 MedThink ($N=100$)	Llama	49.0%	47.0%	43.0%	$-6.0 \text{ pp}^{\text{ns}}$

Table 6: Majority-vote accuracy. $\Delta_{\text{GDP}} = \text{GDP}_{\text{r1}} - \text{M4}_{\text{r1}}$. “ns” = not significant at $\alpha=0.05$. On D1, GDP vs. M3: McNemar $p=0.061$; GDP vs. M4: $p=0.097$; M3 vs. M4: $p=0.86$. On D2 (Qwen), all pairwise McNemar tests yield $p>0.15$. On D2 (Llama, $N=100$), all pairwise tests yield $p>0.28$.

Embedding component. Stella-EN-1.5B-v5 (Zhang et al., 2024) (1.5B parameters), L2-normalized 1024-dimensional vectors; cosine similarity reduces to dot product.

Step decomposition. Deterministic regex pipeline: primary extraction of numbered lists (e.g., 1., 2)) and bullet points (–, *); sentence-level splitting as fallback for unstructured paragraph responses. The final answer line (e.g., ANSWER: X) is stripped before decomposition; resulting steps shorter than 20 characters are filtered. GDP’s CLAIM/GROUND/STANCE fields parse successfully in 99.2–100% of responses per condition-round cell (99.8% of 6,594 GDP responses overall). Two alternative step-level segmentation strategies—sentence-level splitting and removing the minimum-length filter—preserve the system ranking and both the illusion and repair signatures on both datasets.

Runtime. On a single NVIDIA RTX 4090 (24 GB), CARA processes $\sim 9,000$ traces per hour. The full D1 pipeline ($N=500$, 2,500 pairwise records) completes in ~ 8 minutes; D2 ($N=499$, 2,495 records) in ~ 8 minutes.

F CARA-LLM Oracle Validation

Design. To validate CARA-HYB against human-like judgment, we conduct a three-dimensional validation using GPT-4o (OpenAI, 2024) as an alignment judge. We evaluate 80 agent pairs: 50 agreement-set pairs (both agents selected the majority answer) and 30 disagreement pairs (agents selected different answers). The inclusion of disagreement pairs is critical: agreement-set pairs produce a ceiling effect in GPT-4o scores (skewed toward 4–5), and disagreement pairs break this ceiling by introducing genuinely misaligned reasoning (scores 2–3), expanding the effective scoring range. For each pair, GPT-4o assigns a 1–5 alignment score at temperature 0.0, repeated three times per pair. We report CARA-LLM as this rating normalized to $[0, 1]$ via $(\text{score} - 1)/4$; the per-system

head-to-head below is given on the raw 1–5 scale for interpretability.

Three-dimensional validation results.

- Moderate pair-level correlation** (Spearman $\rho=0.55$, $n=80$): CARA-HYB and GPT-4o capture overlapping but non-identical dimensions of alignment. The correlation is moderate rather than high because the two metrics operationalize alignment differently: CARA-HYB measures step-level semantic overlap via embeddings, while GPT-4o evaluates holistic logical coherence.
- Strict stratum monotonicity:** when pairs are binned by CARA-HYB terciles, mean CARA-LLM scores are 0.67 (low), 0.77 (mid), 0.82 (high). CARA-HYB reliably distinguishes alignment levels even where it does not predict individual GPT-4o scores precisely.
- Perfect inter-run reliability:** all 80 pairs receive identical GPT-4o scores across three independent runs (inter-run SD = 0.00), confirming deterministic evaluation at temperature 0.

Per-system head-to-head. Within the 50 agreement-set pairs, M3-GDP r1 receives a mean GPT-4o score of 4.86 ($n=28$, SD 0.35), vs. 4.50 for M3 r1 ($n=22$, SD 0.58); Cohen’s $d=+0.75$, 95% bootstrap CI on Δ $[+0.08, +0.65]$ excludes zero ($B=10,000$, seed 42). This direct head-to-head corroborates the inter-agent CARA-HYB ranking and complements the tercile monotonicity reported above.

Disagreement-pair divergence. On disagreement pairs (agents with different final answers), CARA-HYB assigns scores of 0.83–0.95 because medical reasoning steps have high embedding similarity even when conclusions differ, while GPT-4o assigns 2–3 out of 5 because it detects the logical incoherence of conflicting conclusions. This divergence is by design: CARA-HYB measures reasoning-step alignment *within the agreement set*, so its high scores on disagreement pairs reflect

genuine semantic overlap in medical reasoning vocabulary rather than a metric failure. This regime-based division of labor grounds the joint use we recommend in deployment: CARA-HYB for scalable in-agreement-set auditing, CARA-LLM as a holistic validation check on logical coherence.

G Label-Stripped CARA-HYB: Structural Confound Check

Motivation. CARA-HYB computes step-pair embedding cosine and NLI contradictions over reasoning steps extracted from agent responses. GDP requires each step to carry literal CLAIM/GROUND/STANCE label tokens, so a reasonable concern is that these template tokens themselves inflate the embedding component of CARA-HYB, making part of the GDP-vs-M3 gap a metric artifact of template self-similarity rather than substantive reasoning convergence. The format/debate decomposition (Section 7) and the CARA-GT result (Section 6.4) already weigh against this concern; this appendix reports the direct check.

Procedure. For each GDP trace (D1 Qwen, D2 Qwen, D2 Llama; both r_0 and r_1) we remove the literal label tokens CLAIM:, GROUND:, and STANCE: together with their Markdown bold variants (e.g., ****CLAIM****), preserving (i) the numbered-step markers (1., 2., ...) that M3 traces also use, (ii) the AGREE/DISAGREE/EXTEND content tokens that constitute the agent’s actual stance decision, and (iii) all free-text claim, ground, and stance content. We then re-run the unchanged CARA-HYB pipeline (DeBERTa-v3 NLI + Stella embeddings, $\tau=0.7$, seed 42) on the stripped traces; M3 and M4 cells are unchanged. Paired Cohen’s d_z is computed at the question level for GDP r1 vs. M3 r1, using the same script for both the original and the stripped condition.

Results. Table 7 reports the paired comparison across all three dataset-backbone conditions. Removing the literal label tokens attenuates Cohen’s d_z by only 6–10%. All three 95% bootstrap CI lower bounds remain above +0.76, and mean CARA-HYB itself drops by only ~ 0.005 across configurations. Approximately 90% of the alignment effect survives label removal, indicating that the literal template tokens are not the principal source of GDP’s gain on CARA-HYB.

Generic-structure control. Two additional arms on the same 100 D2 questions (pilot+extension subsample) isolate structure from grounding. An arm prompted with generic numbered steps—no named grounding, no stance—scores CARA-HYB 0.863 at the independent round (below the unstructured M4 baseline of 0.894, where CLAIM+GROUND reaches 0.913) and 0.886 in debate, an increment of +0.008 over standard debate (0.878) against GDP’s +0.073 (0.951): generic structure accounts for roughly 11% of the debate-round gain. An arm removing only the anti-sycophancy clause scores 0.947 vs. full GDP’s 0.951, consistent with that clause’s failure-mode role (Section 7, RQ3) rather than an alignment-magnitude driver.

Interpretation and limitations. The label-stripping experiment isolates one specific structural-confound mechanism—that the literal template tokens themselves drive embedding similarity. The data do not support this mechanism: $\sim 90\%$ of the effect persists after token removal.

Robustness to the NLI model. CARA-HYB is also robust to the choice of contradiction detector: substituting a cross-encoder NLI model (cross-encoder/nli-deberta-v3-large) for the default DeBERTa-MNLI on the D2 $N=100$ set leaves the illusion direction unchanged (M3 r1 < M4 r0) and Cohen’s d nearly identical (-0.36 vs. -0.38).

H Human-Expert Validation

Setup. Three independent annotators with medical training each rated 50 agent pairs drawn from the agreement set on the MedThink-Bench (D2) benchmark. The 50 pairs are stratified by CARA-HYB tercile (low/mid/high, ~ 17 pairs each). Each pair was rated on a 5-point alignment scale (5=near-identical reasoning, 4=substantially aligned, 3=partially aligned, 2=weakly aligned, 1=misaligned reasoning with the same final answer) and a 3-point confidence scale, following the annotation guidelines released with the code repository. Annotators saw only the question and the two agents’ reasoning steps; system identity (M3 vs. M3-GDP) and CARA-HYB scores were withheld to prevent bias.

Inter-annotator agreement. Pairwise Spearman correlation between annotators is consistently high (Table 8); 100% of pair-level disagreements are

Config	Original d_z	Stripped d_z	Stripped 95% CI
D1 MedQA (Qwen, $n=422$)	+0.97	+0.87	[+0.76, +0.99]
D2 MedThink (Qwen, $n=427$)	+1.16	+1.09	[+0.98, +1.21]
D2 MedThink (Llama, $n=98$)	+1.39	+1.29	[+0.96, +1.74]

Table 7: Label-stripped CARA-HYB confound check. Paired Cohen’s d_z between GDP r1 and M3 r1 on the agreement set, before and after removing literal CLAIM/GROUND/STANCE tokens from GDP traces. The effect attenuates by 6–10% but remains Tier B–A in all three conditions, with all CI lower bounds above +0.76.

bounded to ≤ 1 scale point, confirming task reliability.

Annotator pair	Spearman ρ	Exact agreement	Within-1
A1 vs. A2	0.934	76%	100%
A1 vs. A3	0.987	98%	100%
A2 vs. A3	0.927	74%	100%

Table 8: Inter-annotator agreement on the 50-pair human annotation.

Correlation with CARA metrics. Aggregate human ratings (mean of three annotators) correlate positively with CARA-HYB (Spearman $\rho=0.263$), CARA-SIM ($\rho=0.312$), and weakly with the contradiction rate ($\rho=0.137$). Each individual annotator shows a similar correlation with CARA-HYB ($\rho=0.21$ – 0.26), confirming the signal is not driven by any single annotator.

Tercile monotonicity. Aggregate ratings increase monotonically across CARA-HYB strata: 4.00 (low, $n=17$), 4.65 (mid, $n=17$), 4.71 (high, $n=16$). The same monotonic pattern is observed in the CARA-LLM GPT-4o oracle terciles (Figure 4), supporting a consistent ordering of CARA-HYB-graded alignment across two independent external judges. The low-to-mid gap is larger than the mid-to-high gap, reflecting a real shift in alignment quality at the lower end combined with a ceiling effect at the upper end (more than half of high-stratum pairs receive the maximum rating of 5).

Range restriction. All 50 pairs have $\text{CARA-HYB} \in [0.76, 0.98]$ because the metric is conditioned on the agreement set (Section 3.1); the “low” stratum mean is 0.864, far above the corpus-level low end. This narrow range bounds the achievable Spearman correlation against any external judge, paralleling the bounded correlation observed in the CARA-LLM GPT-4o oracle validation when restricted to the same agreement set.

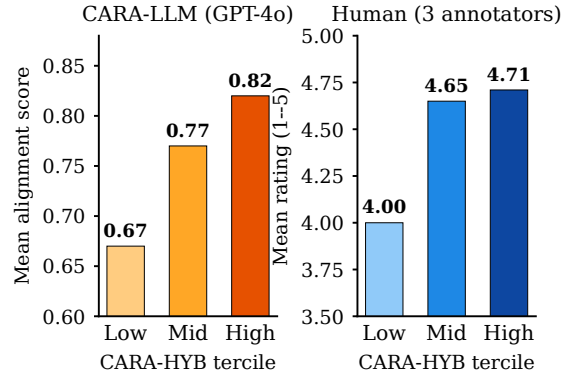


Figure 4: Tercile monotonicity across two external validators on the same 50-pair sample stratified by CARA-HYB. Both the GPT-4o oracle (CARA-LLM; Appendix F) and the mean of three medically-trained annotators show monotonic increase from low to mid to high CARA-HYB strata.

Annotation instructions. Annotators received the following written guidelines before the task (verbatim text is released with our code repository):

Task overview. You will see pairs of AI-generated medical reasoning traces. Both agents chose the same final answer; judge how aligned their reasoning is—not whether the answer is correct.

Rating 1 — Reasoning Alignment (1–5):

- **5 (Near-identical):** same claims, same facts, same logical sequence; minor wording differences only.
- **4 (Substantially aligned):** same core reasoning and key facts; one may add detail or reorganize.
- **3 (Partially aligned):** share key steps but diverge on intermediate justifications.
- **2 (Weakly aligned):** same answer through largely different reasoning; few overlapping steps.
- **1 (Misaligned):** same answer through fundamentally different or contradictory reasoning.

Rating 2 — Confidence (1–3): 3 confident, 2 somewhat confident, 1 low confidence (reasoning too vague to judge).

What NOT to judge. Do not evaluate answer correctness, writing quality, response format (structured vs. free text), or step count. Some responses use a CLAIM/GROUND/STANCE format; others use free text—judge reasoning content regardless of format. System identity and automated CARA scores are withheld to prevent bias.

Annotator background and procedure. The three annotators are medically-trained professionals aged 30–35 (two female resident physicians and one male postdoctoral fellow), recruited through direct personal contact by the authors and participating as volunteers without compensation. All three annotators rated all 50 pairs, enabling the pairwise Spearman inter-annotator agreement reported above. Pairs were presented in randomized order, and system identity (M3 vs. M3-GDP), agent round, and CARA-HYB/CARA-SIM scores were withheld throughout. The task does not constitute human-subjects research—annotators rated machine-generated outputs only—and required no institutional review. The verbatim annotation guidelines, blank annotation sheet, and three completed per-annotator scored sheets are released with our code repository.

I CARA Variants and Aggregation Formulas

Variant table. The four CARA alignment variants used in the paper are summarized below.

Rescaling. For variants whose range includes negative values (CARA-NLI, CARA-HYB), we define a unit-interval rescaling:

$$\text{CARA}_{[0,1]} = \frac{\text{CARA} + 1}{2}. \quad (7)$$

Contradiction Rate. The Contradiction Rate diagnostic is defined in Eq. (6) of Section 3.3. The denominator $\sum_{i \in \mathcal{S}(q)} (|\mathcal{S}(q)| - 1) K_i$ equals $\sum_{i < j} (K_i + K_j)$, the total number of best-match operations across all directed agent pairs in the agreement set. CR captures the rate at which best-aligned step pairs across agreeing agents are nonetheless contradictory: a direct indicator of the consistency illusion.

J System Configurations

Configuration overview. We evaluate three system configurations under identical compute budgets. Each system uses three agents and aggregates their final answers by majority vote. “Round 0” refers to the independent step in which each agent answers without seeing the others’ responses; “Round 1” refers to the post-debate step in which each agent revises its answer after seeing all other agents’ Round 0 outputs. M4 has only Round 0 (the control condition with no debate); M3 and M3-GDP have both rounds. Majority voting is applied

to Round 1 answers for M3 and M3-GDP, and to Round 0 answers for M4.

Backbone serving environment. Primary trials use Qwen 2.5 72B Instruct AWQ (4-bit quantized) served via vLLM (Kwon et al., 2023) on 2×NVIDIA A40 GPUs (48 GB each). The cross-backbone replication uses Llama 3.3 70B Instruct AWQ on a separate pair of A40 GPUs with the same vLLM serving configuration. All experiments use the same decoding hyperparameters: temperature 0.7, top- p 0.9, and a maximum of 2,048 output tokens per agent response.

Standard M3 and M4 prompts. M3 and M4 share a 2-message Round 0 prompt. The system message is verbatim:

You are a medical specialist. Answer medical questions with step-by-step reasoning.

The user message contains the question text with inline answer options, followed by the suffix:

Think step by step and provide your answer in the format: ANSWER: X

M3 Round 1 extends this to a 4-message asymmetric prompt: the original system and user messages, the agent’s own Round 0 response as the assistant turn, and a final user turn presenting the other agents’ Round 0 responses with the instruction “*Using these as additional information, provide your answer in the format: ANSWER: X.*” Each agent sees only the other two agents’ responses in the debate user turn (never its own), preventing self-conditioning artifacts.

M3-GDP prompts. M3-GDP preserves the 2-message (Round 0) and 4-message (Round 1) prompt structure of M3 but replaces the system message with format requirements that enforce structured reasoning. The Round 0 system message requires CLAIM+GROUND for each reasoning step, where GROUND must name a specific medical fact, mechanism, or guideline supporting the claim. The Round 1 system message adds STANCE (AGREE/DISAGREE/EXTEND) with respect to each other agent’s claims, with the rule that any DISAGREE must be accompanied by a counter-GROUND. Round 1 also includes an explicit anti-sycophancy clause:

Do NOT change your answer simply because other agents disagree. Only change if you identify a specific factual error

Variant	Range	Definition
CARA-NLI	$\{-1, 0, +1\}$	NLI label: +1 entailment, 0 neutral, -1 contradiction.
CARA-SIM	$[-1, +1]$	Cosine similarity: $\cos(\text{emb}(r_i^k), \text{emb}(r_j^l))$.
CARA-HYB	$[-1, +1]$	Hybrid: if $P(\text{contradiction}) > \tau$ then -1 , else cosine similarity.
CARA-LLM	$[0, 1]$	GPT-4o 1–5 alignment rating, normalized to $[0, 1]$ via $(\text{score}-1)/4$.

Table 9: CARA alignment variants. CARA-HYB is the default metric used throughout; CARA-LLM serves as a validation oracle.

ID	System	Agents	Rounds	GDP
M3	Du et al. (2024) Debate	3	1	no
M4	Independent Vote	3	0	no
M3-GDP	Debate + GDP	3	1	yes

Table 10: System configurations. All systems use majority voting for answer aggregation. “Rounds” counts debate rounds; Round 0 (independent) is always present. M4 serves as a no-communication control.

in your own reasoning or receive a more compelling medical ground from another agent.

The full GDP Round 0 and Round 1 system messages, along with the debate user template, are released with the code repository.

K GDP Worked Example

We compare a standard debate (M3) trace and a GDP (M3-GDP) trace on the same clinical question. Both systems converge on the same correct answer; the contrast is in how the agents arrive there.

Clinical question. A patient presents with fasting plasma glucose of 140 mg/dL on two consecutive measurements. What is the most appropriate first-line management?
A. Insulin **B. Metformin** C. Lifestyle modification only D. Sulfonylurea
Gold answer: **B**.

M3 (Standard Debate) — Round 1

Agent A

The elevated fasting glucose is consistent with a diagnosis of diabetes. Considering the clinical picture, metformin would be appropriate. I agree with Agent B’s reasoning. *Answer: B.*

Agent B

The patient’s clinical picture supports metformin therapy. Given the lab values and the analysis Agent A provided, I support this approach. *Answer: B.*

Voted solution: B ✓ | Both agents converge with no medical citations; agreement is vague.

M3-GDP — Round 1

Agent A

CLAIM: The fasting glucose of 140 mg/dL meets ADA diagnostic criteria for T2DM.
GROUND: Fasting plasma glucose ≥ 126 mg/dL on two occasions confirms diabetes (ADA Standards of Care).
STANCE: DISAGREE with Agent B (insulin)—insulin is not first-line for uncomplicated T2DM without DKA. *Answer: B.*

Agent B

CLAIM: Metformin is the appropriate first-line agent for newly diagnosed T2DM.
GROUND: ADA Standards of Care recommend metformin first-line for adults with T2DM without contraindications, given its efficacy in reducing hepatic glucose production and favorable safety profile.
STANCE: AGREE with Agent A’s diagnostic claim; reaffirm metformin over insulin given absence of DKA or severe hyperglycemia. *Answer: B.*

Voted solution: B ✓ | Both agents cite ADA criteria and take explicit stances on each other’s claims; agreement is grounded.

The contrast is structural: M3 agents agree on the answer without committing to specific medical evidence or explicitly engaging with each other’s claims, whereas M3-GDP agents reach the same answer through named criteria (ADA guidelines) and explicit STANCE declarations. This pattern recurs at the dataset scale, producing the CARA-HYB and CARA-GT gaps reported in Section 6.1 and Appendix C.

L Failure Mode Taxonomy and Case Study

Failure-mode codebook. The 50 lowest-CARA-HYB correct-answer cases per system are classified using deterministic rules over CARA sub-metrics ($\text{CR} > 0.3$, step-count ratio > 2 , zero-step agents, $\text{SIM} < 0.80$).

Case study: MQA_0447 (hypovolemic shock). We illustrate FM3 (contradictory premises) with a D1 case in which the correct answer is reached through partially incorrect reasoning. The question

Code	Failure mode
FM1	COMPLEMENTARY REASONING: different but non-contradictory reasoning paths. Low SIM, low CR.
FM2	GRANULARITY MISMATCH: step-count ratio >2. Depth difference, not substantive disagreement.
FM3	CONTRADICTORY PREMISES: agents cite contradictory facts but converge on the correct answer.
FM4	SYCOPHANTIC CONVERGENCE: agent adopts the majority answer without substantive reasoning (zero steps).
FM5	TERMINOLOGY DIVERGENCE: different medical terms for the same concept.
FM6	PARTIAL OVERLAP: residual baseline misalignment without extreme signals.

Table 11: Six failure modes identified by classifying the 50 lowest-CARA-HYB correct-answer cases per system.

describes a 52-year-old man with variceal bleeding and hypovolemic shock. Two of three GDP agents in the agreement set initially state that systemic vascular resistance (SVR) *decreases* in hypovolemic shock—a clinically incorrect claim, since SVR *increases* as a compensatory response to maintain perfusion. Both agents then self-correct within the same response and select the correct treatment (volume resuscitation followed by endoscopic intervention). This case exemplifies the consistency illusion at the individual-fact level: the answer is correct, but the underlying reasoning contains explicit factual contradictions that the metric correctly surfaces. GDP’s structured format makes these contradictions *detectable*; under standard debate, the same misstatement would more likely be smoothed away in vague free text.

Answer-change decomposition. Table 12 crosses the two diagnostic dimensions from RQ5 for all 31 D1 correct-to-incorrect changes (M3 r1 → GDP r1). On D2, the additional expert-coverage signal flags 25/36 overturned answers (69%): 17 with inter-agent contradictions, 10 with expert coverage below 0.75, and 5 with zero-step agents (categories overlap). Overturned answers show contradiction rates of 0.094 vs. 0.023 for retained-correct answers on D1, and 0.087 vs. 0.039 on D2. The 17 unflagged D1 cases are indistinguishable from retained-correct questions on every M3-side diagnostic (15/17 unanimous consensus; mean CARA-HYB 0.900 vs. 0.893; mean steps per agent 3.63 vs. 2.38), and 12 of 17 diverge before any

interaction: they reflect prompt-stage answer shifts under the structured format (GDP r0 81.8% vs. M4 r0 83.8% on D1) rather than debate-stage overturning.

Round of first divergence	Flagged	Unflagged	Total
GDP r0 (independent)	11	12	23
GDP r1 (grounded debate)	3	5	8
Total	14	17	31

Table 12: D1 correct-to-incorrect answer changes (M3 r1 → GDP r1), crossed by unfounded-consensus signal status and round of first divergence. Most changes precede any agent interaction; the grounded-debate round accounts for 8/31, of which 5 are unflagged.

FM1 resolution by expert coverage. On D2, where the expert Scoring-Points reference permits a coverage check, the 18 FM1 cases resolve into two subtypes. **FM1a (complementary-grounded, 10/18):** the agents jointly cover *all* expert scoring points through different rationales (union coverage 0.856, above the 0.808 mean among correct-answer questions). **FM1b (divergent-with-gaps, 8/18):** the union leaves expert points uncovered. Both subtypes remain audit-relevant—individual agent–expert alignment averages ~ 0.60 and the consensus lacks cross-agent corroboration—and GDP consolidates rather than suppresses complementary evidence: on the same questions, union coverage rises from 0.856 to 0.921 and full-coverage cases from 10 to 13.

M Undefined Rate

The undefined rate—questions where no majority answer emerges—is an independent behavioral indicator that does not depend on agreement-set selection. Table 13 reports per-system rates on both datasets.

Standard debate destabilizes consensus. On D1, M4 (independent voting) achieves 100% consensus because agents sample independently from the same distribution: the probability of three-way disagreement among i.i.d. samples is low. Standard debate (M3 r1) drives the undefined rate to 15.2% on D1—a $38\times$ increase over the pre-debate round (M3 r0, 0.4%). On D2, where the harder question distribution gives even M4 a 5.8% undefined rate, M3 r1 still pushes this up to 14.2% ($2.5\times$ relative). GDP debate maintains consensus at near-M4 levels

System / Round	D1 ($N=500$)	D2 ($N=499$)
M3 r1	76 (15.2%)	71 (14.2%)
M3-GDP r0	3 (0.6%)	26 (5.2%)
M3 r0	2 (0.4%)	28 (5.6%)
M3-GDP r1	2 (0.4%)	2 (0.4%)
M4 r0	0 (0.0%)	29 (5.8%)

Table 13: Undefined rate per system–round–dataset. M3 r1 destabilizes consensus on $\sim 14\text{--}15\%$ of questions on both datasets; M3-GDP r1 retains essentially all of them.

(0.4% on both datasets) while substantially improving reasoning alignment.

Reasoning collapse. A related phenomenon is reasoning collapse: agents that produce zero extractable reasoning steps after debate (a behavioral signature of sycophantic convergence). On D2 $N=100$, M3 r1 yields zero-step agents in 5.0% of agreement-set memberships; M3-GDP r1 in only 0.3%. GDP’s CLAIM+GROUND requirement prevents this collapse because an agent cannot adopt an answer without producing a named ground supporting it.

N Consistency-Illusion Visualization

Figure 5 visualizes the (CR, SIM) trajectory of each system on D2 ($N=499$, Qwen 2.5 72B), making the dual movement of standard debate ($\text{CR} \downarrow + \text{SIM} \downarrow$) and the repair pattern under GDP ($\text{SIM} \uparrow$) directly observable. The two arrows correspond to the $r0 \rightarrow r1$ transitions under M3 (red) and M3-GDP (green); M4 is shown as a single point for reference. The figure complements the conceptual schematic of Figure 1 with the actual measured coordinates that anchor every claim made in Section 6.1.

O Mechanism Discussion

The condensed mechanism account in RQ1 (Section 7) merges two effects under one heading. Here we elaborate each with the original four-component analysis.

Contradiction smoothing without reasoning convergence (M3). During debate, agents see each other’s responses and revise their own. When agents disagree on factual claims, revision pressure causes them to soften or remove the conflicting statements, which reduces CR. However, this removal is *subtraction without alignment*: agents delete contradictory steps without replacing them

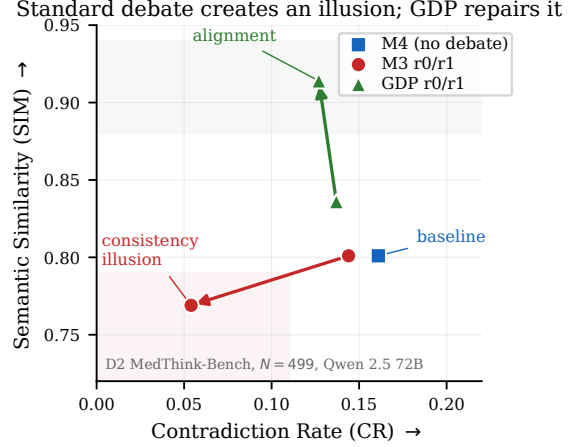


Figure 5: Empirical (CR, SIM) trajectory on D2 ($N=499$, Qwen 2.5 72B). M3 (red arrow) moves $r0 \rightarrow r1$ both $\text{CR} \downarrow$ and $\text{SIM} \downarrow$ (consistency illusion). GDP (green arrow) moves $r0 \rightarrow r1$ $\text{SIM} \uparrow$ while keeping CR informative.

with shared reasoning anchors that other agents have also committed to. The result is fewer contradictions but also less overlapping reasoning content, so SIM decreases. This pattern is stronger on D2 because the more open answer space (3–10 options, mean ~ 7.5) gives agents more room to diverge in reasoning while still landing on the same final answer.

Sycophantic convergence (M3). Agents may adopt each other’s *answers* without adopting the underlying reasoning—a form of answer-level sycophancy (Yao et al., 2025). This produces answer convergence (the system reports consensus) while reasoning chains remain divergent or become vaguer. The high M3 r1 undefined rate—15.2% on D1 and 14.2% on D2—shows that for a substantial minority of questions, agents’ positions are sufficiently incoherent that no majority emerges even after revision. The zero-step phenomenon (Appendix M) is the limiting case: an agent produces an answer with no extractable reasoning, indicating that revision has degenerated into pure imitation.

Format effect (GDP, pre-debate). The CLAIM+GROUND structure forces agents to decompose reasoning into named, verifiable steps. This produces more comparable reasoning units across agents even before any debate interaction: on D1, GDP_{r0} vs. M4_{r0} alone yields a format-effect $d=+0.62$ with SIM increase of $+0.034$. This isolates a structural cause for misalignment in standard debate: a substantial portion of apparent

misalignment originates from underspecified output format rather than from genuine reasoning disagreement.

Debate interaction effect (GDP, $r_0 \rightarrow r_1$). The STANCE mechanism in Round 1 requires each agent to explicitly address every claim made by the other agents. When an agent must state “I AGREE with Agent B’s claim that renal function is preserved, because...” or “I DISAGREE because...,” the resulting response necessarily shares reasoning vocabulary and logical structure with the other agents. This drives SIM upward (+0.077 on D1, +0.078 on D2 from r_0 to r_1 under GDP) at a magnitude $2.4\times$ larger than the format effect alone ($d=+1.51$ vs. $d=+0.62$), indicating that STANCE-mediated engagement is the primary mechanism. The format effect is a necessary precondition, however: without structured claims to reference, the STANCE mechanism has no well-defined targets.

P Cross-Dataset and Cross-Backbone Discussion

Comparison	Llama D2 ($N=100$)	Qwen D2 ($N=499$)	Dir.
GDP _{r1} vs. M3 _{r1}	+1.99	+1.62	✓
M3 _{r1} vs. M4 _{r0} (illusion)	-1.32	-0.30	✓
M3 debate Δ HYB ($r_0 \rightarrow r_1$)	-0.038	-0.014	✓
GDP debate Δ HYB ($r_0 \rightarrow r_1$)	+0.014	+0.040	✓
Format effect Δ HYB	+0.006	+0.017	✓

Table 14: Cross-backbone validation on D2 (referenced from Section 6.3). First two rows give Cohen’s d ; remaining rows give Δ CARA-HYB. All five comparisons are direction-consistent (“Dir.” ✓).

GDP effect generalization. The GDP effect replicates across all tested conditions with Tier A effect sizes: D1 Qwen ($N=500$): $d=+1.43$, $CI=[+1.29, +1.60]$; D2 Qwen ($N=499$): $d=+1.62$, $CI=[+1.48, +1.78]$; D2 Llama ($N=100$): $d=+1.99$, $CI=[+1.65, +2.44]$. The rank ordering of systems is identical across all three conditions: $M3\text{-}GDP_{r_1} \succ M3\text{-}GDP_{r_0} \succ M4_{r_0} \approx M3_{r_0} \succ M3_{r_1}$. This consistency across two structurally different datasets (fixed 4-option MCQ vs. variable 3–10-option clinical reasoning) and two independent backbone models supports the generalizability of both the consistency-illusion finding and the GDP repair effect.

Consistency-illusion magnitude variation. The consistency illusion magnitude varies more across

conditions: $d=-0.08$ (D1, Qwen), $d=-0.30$ (D2, Qwen), $d=-1.32$ (D2, Llama). Two factors plausibly drive this variation. First, D1’s fixed 4-option format constrains the space in which reasoning can diverge while answers remain matched, weakening the measurable illusion compared to D2’s mean ~ 7.5 -option open answer space. Second, M3 r_1 undefined rates differ markedly across conditions (15.2%, 14.2%, 2.0%), and a lower undefined rate preserves more of the worst-alignment cases in the measured sample, strengthening the apparent illusion magnitude. The third dataset (Llama D2) has the lowest undefined rate and the strongest illusion—an interpretive asymmetry we surface honestly via the sensitivity analysis in Appendix B rather than treating the Llama magnitude as a backbone-effect comparison.

Implications for deployment. Where the underlying multi-agent question allows substantial reasoning-space variation (open-ended diagnosis, differential reasoning, treatment selection with multiple acceptable plans), the consistency illusion is likely to be *larger* than on constrained MCQ benchmarks, and the GDP benefit correspondingly more valuable. The D1 vs. D2 magnitude difference is therefore not a defect of the metric or the intervention but a direct reflection of how much room the task gives reasoning to diverge.