Do Mixed-Vendor Multi-Agent LLMs Improve Clinical Diagnosis?

Grace Yuan[1] **Xiaoman Zhang**[2] **Sung Eun Kim**[2] **Pranav Rajpurkar**[2*]
[1] Massachusetts Institute of Technology, Boston, MA
[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Abstract

Multi-agent large language model (LLM) systems have emerged as a promising approach for clinical diagnosis, leveraging collaboration among agents to refine medical reasoning. However, most existing frameworks rely on single-vendor teams (e.g., multiple agents from the same model family), which risk correlated failure modes that reinforce shared biases rather than correcting them. We investigate the impact of vendor diversity by comparing Single-LLM, Single-Vendor, and Mixed-Vendor Multi-Agent Conversation (MAC) frameworks. Using three doctor agents instantiated with o4-mini, Gemini-2.5-Pro, and Claude-4.5-Sonnet, we evaluate performance on RareBench and DiagnosisArena. Mixed-vendor configurations consistently outperform single-vendor counterparts, achieving state-of-the-art recall and accuracy. Overlap analysis reveals the underlying mechanism: mixed-vendor teams pool complementary inductive biases, surfacing correct diagnoses that individual models or homogeneous teams collectively miss. These results highlight vendor diversity as a key design principle for robust clinical diagnostic systems.

1 Introduction

Large language models are increasingly used for clinical diagnosis (Kim et al., 2024; Lee et al., 2025; Nori et al., 2025; Zhou et al., 2025). Across a range of benchmarks, frontier models from OpenAI, Google, and Anthropic can approach or even exceed average physician performance (Nori et al., 2025; Saab et al., 2024). Yet no single model dominates: performance varies substantially across tasks and clinical domains, with different vendors exhibiting distinct strengths and weaknesses (Cao et al., 2024; Sonoda et al., 2024).

This variability has motivated the development of multi-agent frameworks, where several LLM "doctors" collaborate under a coordinating supervisor (Kim et al., 2024; Lee et al., 2025; Zhou et al., 2025). Such systems more closely resemble real clinical workflows, multidisciplinary case conferences where diverse perspectives converge on a diagnosis (Chen et al., 2025b; Zhou et al., 2025). Early results are promising, with such architectures frequently outperforming single-model baselines (Chen et al., 2025b; Nori et al., 2025; Zhou et al., 2025).

A small body of work has begun exploring mixed-vendor systems in general domains (Ye et al., 2025). However, existing mixed-vendor approaches in medicine mostly rely on independent model runs, combining models from different providers via voting or simple aggregation (Barabucci et al., 2024; Sorich et al., 2024). Research integrating cross-vendor models within structured multi-agent conversations to directly compare them against single-vendor baselines remains limited.

We address this gap by hypothesizing that vendor diversity improves diagnosis by introducing complementary inductive biases and breaking the "echo chamber" of homogeneous teams. We test this hypothesis using the Multi-Agent Conversation (MAC) framework (Chen et al., 2025b), comparing single-model, single-vendor MAC, and mixed-vendor MAC configurations across RareBench (rare diseases) (Chen et al., 2024) and DiagnosisArena (complex case reports) (Zhu et al., 2025).

Our results demonstrate that **mixed-vendor configurations deliver superior overall performance compared to their single-vendor counterparts**, achieving state-of-the-art recall and accuracy across both benchmarks. Analysis of solution space overlap reveals the mechanism driving this gain: while single-vendor teams remain highly correlated with their base model's priors, **mixed-vendor teams successfully pool complementary inductive biases to surface diagnoses that indi-**

*Corresponding author

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 1–18

vidual models miss. These findings suggest that leveraging the complementary strengths of diverse LLMs is crucial for building robust clinical diagnostic systems.

2 Related Work

Clinical LLMs and Vendor Differences Single LLM systems have demonstrated impressive capabilities in medical reasoning, often matching or even exceeding average physician performance (Eriksen et al., 2024; Singhal et al., 2025; Sonoda et al., 2024). However, these models exhibit divergent behaviors driven by differences in pretraining data and alignment strategies (Chen et al., 2025a; Li and et al., 2025). Rather than having a universally superior model, existing research reveals that models from different vendors possess complementary inductive biases (Yang et al., 2024). Consequently, there is potential in employing vendor diversity to bridge these domain-specific gaps, combining distinct reasoning behaviors to achieve more robust clinical performance.

Multi-agent LLM Frameworks in Clinical Diagnosis Beyond single-LLM systems, several recent works organize multiple LLM "doctors" into coordinated teams that more closely resemble clinical workflows. Some architectures use a supervisor–specialist pattern, where a director LLM coordinates a set of role-specialized agents and integrates their outputs into a final decision (Kim et al., 2024; Zhou et al., 2025). Others emphasize conversational debate, with multiple specialists independently proposing hypotheses, critiquing one another, and iterating under the oversight of a manager agent (Chen et al., 2025b; Lee et al., 2025; Nori et al., 2025). Another type of multi-agent framework uses panel-style ensembles, where the physician agents reason in parallel with limited interaction and their responses are aggregated into a single collective output (Barabucci et al., 2024).

Mixed-vendor Multi-agent Systems In contrast to the extensive work on single-vendor multi-agent frameworks, mixed-vendor systems have received limited attention. Existing work mostly relies on "panel-style" aggregation rather than conversational collaboration. For example, recent studies demonstrate that combining outputs from cross-vendor models via majority voting or collective intelligence algorithms can outperform individual baselines (Barabucci et al., 2024; Sorich et al., 2024). However, these frameworks treat agents as independent voters, missing the chance for agents to cross-examine and refine opposing views in real-time. More importantly, the clinical literature lacks a systematic comparison between single-vendor and mixed-vendor configurations within a shared conversational architecture. This raises the question of whether mixed-vendor teams can introduce reasoning diversity that improves diagnostic performance over single-vendor baselines.

3 Methods

3.1 MAC Framework

We build directly on the Multi-Agent Conversation (MAC) framework (Chen et al., 2025b), which structures clinical decision-making as a dialogue between three "doctor" agents and one supervising agent. The process follows a fixed protocol: each doctor LLM proposes and revises ranked diagnosis lists in order, while the supervisor LLM monitors the discussion, provides feedback, and ultimately decides on the final output. Unlike Chen et al. (2025b), we do not use role-specialized prompts. Instead, all doctor agents receive the same generic clinical instructions adapted to each benchmark. This simple, symmetric setup allows us to vary the underlying LLMs while keeping the interaction pattern unchanged, ensuring that any observed performance differences come strictly from vendor diversity rather than artifacts of role assignment.

We compare three specific model configurations, as shown in (Figure 1):

- **Single-LLM**, where a single model (o4-mini, Gemini-2.5-Pro, Claude-4.5-Sonnet) directly produces a diagnosis list for each case.
- **Single-vendor MAC**, where all three doctor agents are instantiated with the same model under a fixed supervisor.
- **Mixed-vendor MAC**, where the three doctor agents are instantiated with different vendor models under a fixed supervisor.

Model Selection For all experiments, we employ the current frontier models from each vendor: o4-mini (2025-04-16) (OpenAI, 2025), Gemini-2.5-Pro (Google, 2025), and Claude-4.5-Sonnet (Anthropic, 2025). Preliminary experiments showed comparable diagnostic performance across models, so we treat them as matched frontier baselines and focus on the effects of multi-agent composition.

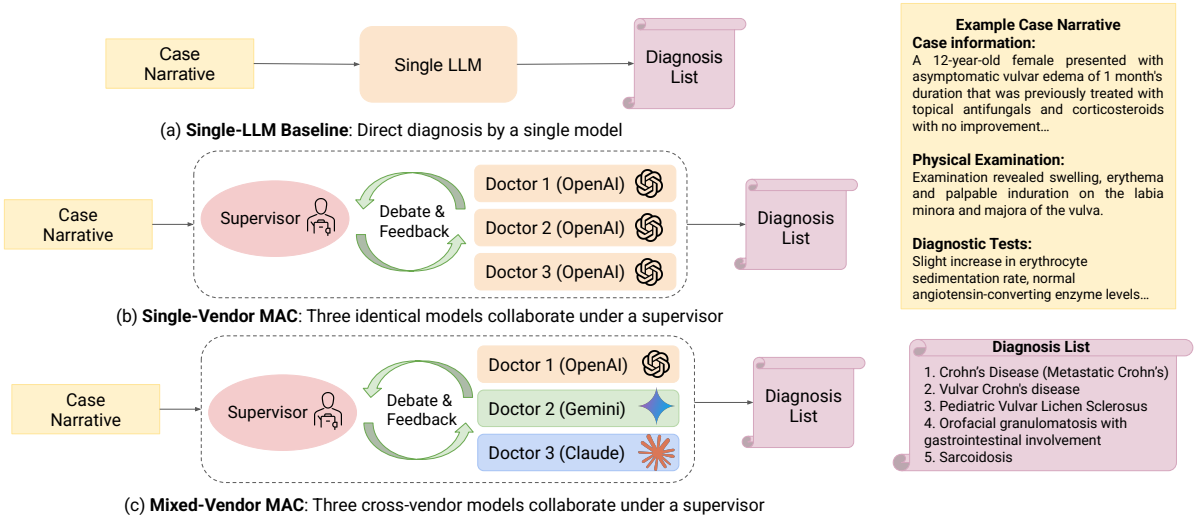

Figure 1: Overview of the 3 different model configurations (Single-LLM, Single-vendor MAC, and Mixed-vendor MAC), and an illustrative example of a case narrative with its corresponding diagnosis list.

Unless otherwise noted, o4-mini serves as a fixed supervisor to standardize aggregation.

3.2 Conversation Protocol and Stopping Rule

For each case, the MAC framework receives a phenotype summary or case narrative. The supervisor initiates the discussion, and the three doctor agents contribute in a fixed round-robin order. At each turn, doctor agents review the full conversation history and the supervisor's latest feedback before submitting a ranked Top-10 diagnosis list with brief justification. The conversation terminates when the supervisor determines the differential has stabilized, or upon reaching a hard limit of 13 turns (following the original MAC protocol). We extract the final consensus Top-10 list from the supervisor's concluding message using a rule-based parser and evaluate it against the ground truth.

4 Experimental Setup

4.1 Datasets

We evaluate on two distinct benchmarks to test generalizability across rare and common conditions.

RareBench Designed to evaluate the capabilities of LLMs on rare diseases, RareBench (Chen et al., 2024) consists of multiple subsets with varing difficulty. We utilize **MME** (40 cases across 17 diseases), **HMS** (88 cases covering 39 diseases), and **LIRICAL** (370 cases spanning 252 diseases).

DiagnosisArena To test performance on broader clinical reasoning, we include DiagnosisArena, a benchmark of 1,113 complex case reports from 10 top-tier medical journals. Even strong reasoning models such as o3-mini, o1, and DeepSeek-R1 achieve at most about 46% accuracy on this benchmark (Zhu et al., 2025), highlighting its difficulty. To minimize the risk of data contamination and ensure computational feasibility, we restrict our evaluation to the subset of 165 cases published in 2024. This temporal restriction reduces overlap with model pre-training data and provides a cleaner test of generalization.

4.2 Tasks and Metrics

RareBench For this benchmark, the task is to output a ranked Top-10 diagnosis list for each case. We evaluate diagnostic performance using Recall@1, Recall@3, Recall@5, and Recall@10.

DiagnosisArena In this setting, the task is to generate a ranked Top-5 diagnosis list for each case. We assess performance using Top-1 and Top-5 accuracy, consistent with the standard evaluation protocol for this benchmark.

Evaluation Protocols Our primary evaluation uses o4-mini as an LLM judge, scoring free-text differential diagnoses against reference labels. To reduce potential self-preference bias, we also implement a vendor-neutral, retrieval-based judge: BioLORD (FRemyCompany/BioLORD-2023) (Remy et al., 2024). In this secondary evaluation, we embed the predicted diagnoses and match them to the closest standardized code before comparing with the gold code.

Table 1: Performance on **Combined RareBench**. Note: OpenAI refers to o4-mini, Gemini refers to Gemini-2.5-Pro, and Claude refers to Claude-4.5-Sonnet.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	31.81	40.52	46.12	53.07
Single-LLM (Gemini)	37.58	47.79	50.84	56.41
Single-LLM (Claude)	29.35	39.54	45.41	55.42
Single-Vendor MAC (OpenAI)	32.73	41.22	45.54	52.33
Single-Vendor MAC (Gemini)	36.63	43.43	46.47	51.93
Single-Vendor MAC (Claude)	33.32	48.06	50.53	57.06
Mixed-Vendor MAC	**39.31**	**49.82**	**55.05**	**61.35**

Table 2: Performance on **MME** subset (n=40). Model versions consistent with Table 1.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	10.00	22.50	25.00	27.50
Single-LLM (Gemini)	30.00	37.50	42.50	47.50
Single-LLM (Claude)	2.50	22.50	30.00	42.50
Single-Vendor MAC (OpenAI)	15.00	25.00	30.00	35.00
Single-Vendor MAC (Gemini)	35.00	45.00	47.50	47.50
Single-Vendor MAC (Claude)	35.00	45.00	50.00	57.50
Mixed-Vendor MAC	**40.00**	**52.50**	**60.00**	**65.00**

Table 3: Performance on **HMS** subset (n=88). Model versions consistent with Table 1.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	43.18	57.95	**68.18**	76.14
Single-LLM (Gemini)	47.73	59.09	64.77	70.45
Single-LLM (Claude)	39.77	56.82	65.91	75.00
Single-Vendor MAC (OpenAI)	47.73	56.82	62.50	**77.27**
Single-Vendor MAC (Gemini)	43.18	51.14	56.82	59.09
Single-Vendor MAC (Claude)	43.18	56.82	67.05	72.73
Mixed-Vendor MAC	**51.14**	**60.23**	**68.18**	**77.27**

Table 4: Performance on **LIRICAL** subset (n=370). Model versions consistent with Table 1.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	28.61	36.51	41.14	48.50
Single-LLM (Gemini)	**36.24**	45.78	48.50	53.95
Single-LLM (Claude)	29.43	40.33	43.87	52.32
Single-Vendor MAC (OpenAI)	30.25	38.96	41.96	47.41
Single-Vendor MAC (Gemini)	33.79	41.42	43.87	50.95
Single-Vendor MAC (Claude)	31.88	43.60	46.59	52.59
Mixed-Vendor MAC	35.69	**46.59**	**51.77**	**57.22**

5 Results

5.1 RareBench

We present results on RareBench in Tables 1–4, covering aggregate performance on the MME, LIRICAL, and HMS subsets. Table 1 demonstrates that the Mixed-Vendor MAC achieves the highest performance across all retrieval metrics, securing a Recall@1 (39.31%) and maintaining a clear lead through Recall@10 (61.35%). Crucially, it consistently outperforms every Single-Vendor MAC baseline. While individual models fluctuate in performance across subsets (e.g., OpenAI struggles on MME but excels on HMS), the mixed-vendor team effectively smooths these variances, offering the most robust overall performance.

MME (Table 2) On MME, which appears to be the most challenging subset with the lowest baseline scores, the benefits of vendor diversity are most pronounced. The Mixed-Vendor MAC achieves a Recall@1 of 40.00%, outperforming the best Single-LLM (Gemini-2.5-Pro) by 10% and the best Single-Vendor MAC (Claude-4.5-Sonnet) by 5%. This dominance extends consistently across all retrieval depths, with the Mixed-Vendor MAC maintaining a clear lead at Recall@3, 5, and 10. This suggests that when individual models struggle, the diverse reasoning strategies provided by a mixed panel are critical for surfacing the correct diagnosis that homogeneous teams miss.

LIRICAL and HMS (Tables 3 & 4) On the LIRICAL and HMS subsets, the Mixed-Vendor MAC demonstrates high stability and top-tier performance. On **HMS**, the Mixed-Vendor MAC achieves highest overall performance across all metrics, outperforming the strongest standalone model (Gemini-2.5-pro) as well as all Single-Vendor MAC configurations. On **LIRICAL**, the Mixed-Vendor MAC achieves a Recall@1 of 35.69%, performing competitively with the best-performing Single-LLM (Gemini-2.5-Pro at 36.24%) while surpassing all baselines at deeper retrieval levels, including Recall@3, 5, and 10. Moreover, the Mixed-Vendor MAC maintains better performance than all single-vendor configurations as depth increases.

Upon examining the evaluations provided by the o4-mini judge, we observed a recurring discrepancy in the HMS subset where technically correct, granular predictions were penalized. To ensure metric validity, the HMS subset ($n = 88$) underwent independent clinical adjudication by a medical doctor across all experimental configurations (Single-LLM, Single-Vendor MAC, and Mixed-Vendor MAC). The review revealed that the MAC consensus process often drives agents toward "hyper-specific" etiologic diagnoses that correctly incorporate patient phenotypes not captured by the broader gold labels, resulting in initial penalties by the o4-mini judge. For instance, in

Case HMS-73, the models identified *Primary Sjögren's Syndrome with lymphomatous transformation*; while the gold label was simply *Primary Sjögren syndrome*, the expert confirmed the model's refinement was clinically supported by the recorded phenotype of "Night sweats." Similar refinements occurred in cases of *Cryoglobulinemic vasculitis* where agents correctly identified *Hepatitis C* as the underlying driver. By manually verifying these "hyper-specific" cases as correct, we ensured the reported HMS scores reflect the framework's true diagnostic depth rather than being artificially deflated by automated judge sensitivity to semantic granularity.

Notably, our results reveal that scaling up to a multi-agent team of identical models does not guarantee improvement over the single-model baseline. In some cases, Single-Vendor MACs actually underperform their standalone base models, most prominently with Gemini on both LIRICAL and HMS, and Claude on HMS. This degradation likely stems from **correlated failure modes**. As noted earlier, we removed role-specialized prompts to isolate the impact of model diversity in the MAC framework. In this absence of role-induced variance, when homogenous agents discuss and deliberate, they risk reinforcing shared hallucinations or collectively drifting away from a correct initial hypothesis. The Mixed-Vendor MAC avoids this pitfall, effectively stabilizing performance where single-vendor teams falter.

5.2 DiagnosisArena

Table 5 presents the results on DiagnosisArena, a benchmark of complex real-world case reports. This dataset is notably difficult for individual models: Gemini-2.5-Pro and Claude-4.5-Sonnet struggle significantly in the Single-LLM setting, achieving only 20.00% and 19.39% Top-1 accuracy, respectively lagging behind the stronger Single-LLM o4-mini baseline (32.12%).

Yet, the Mixed-vendor MAC configuration achieves a highest Top-1 accuracy of 36.36% and Top-5 accuracy of 49.09%. Crucially, although two of the three agents in the mixed team perform poorly in isolation, their distinct reasoning processes allow the Mixed-Vendor system to exceed the strong o4-mini Single-Vendor MAC (35.76%). This result demonstrates that the Mixed-Vendor framework is robust to performance disparities among its agents; it does not strictly require uniformly high-performing constituents to succeed. By integrating these diverse perspectives, the supervisor gains unique insights from the "weaker" agents to resolve cases that a homogenous strong team would collectively miss, highlighting the systemic resilience of the Mixed-Vendor approach.

Table 5: Performance on **DiagnosisArena** (n=165). OpenAI refers to o4-mini, Gemini refers to Gemini-2.5-Pro, and Claude refers to Claude-4.5-Sonnet.

Setup	Top-1	Top-5
Single-LLM (OpenAI)	32.12%	46.06%
Single-LLM (Gemini)	20.00%	31.51%
Single-LLM (Claude)	19.39%	29.70%
Single-Vendor MAC (OpenAI)	35.76%	47.88%
Single-Vendor MAC (Gemini)	33.94%	44.24%
Single-Vendor MAC (Claude)	32.73%	45.45%
Mixed-Vendor MAC	**36.36%**	**49.09%**

5.3 Overlap and Diversity Analysis

To understand the mechanism driving the performance of the Mixed-Vendor MAC, we analyze the relationship between the solution sets of different configurations. We hypothesize that the mixed-vendor systems succeed by aggregating complementary inductive biases, thereby approximating the union of individual model capabilities. By integrating diverse reasoning priors, the system should effectively expand the total valid solution space beyond the blind spots of any single vendor.

To test this, we decompose the predictions of the Mixed-Vendor MAC (M) relative to two baselines (S) for each dataset: the best-performing Single LLM and the best-performing Single-Vendor MAC. For each comparison, we categorize the solution sets into four mutually exclusive groups shown in Figure 2:

- **Mutually Correct** ($S \cap M$): Cases where both systems predicted the correct diagnosis.
- **Baseline Unique** ($S \setminus M$): Cases where only the baseline predicted the correct diagnosis.
- **Mixed Rescue** ($M \setminus S$): Cases where only the Mixed system predicted the correct diagnosis.
- **Both Wrong**: Cases where neither system predicted the correct diagnosis.

We then quantify the overlap and asymmetry of these sets using two metrics:

- **ΔCoverage**: Measures the net expansion of knowledge, defined as Coverage$(S \to M)$ $-$

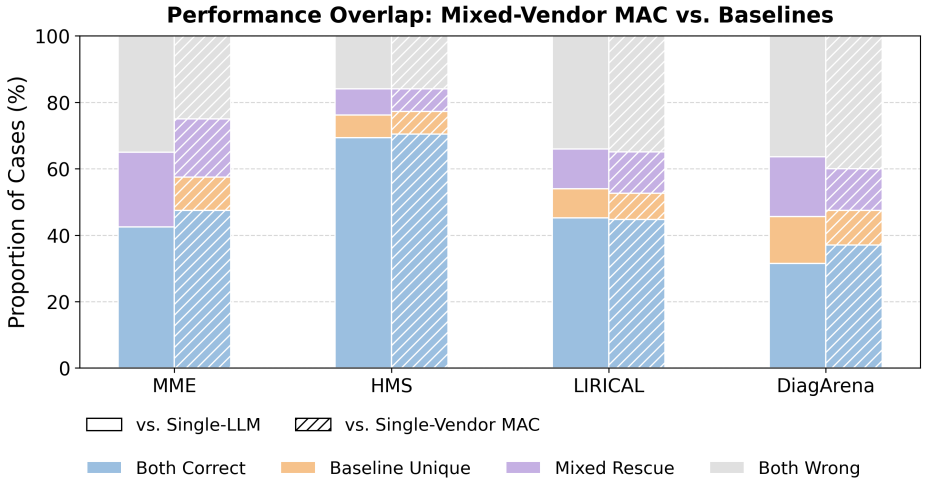

Figure 2: Analysis of Correct Prediction Overlap. The x-axis shows the dataset. For the RareBench datasets (MME, HMS, LIRICAL), overlap is calculated using **Recall@10**. For DiagnosisArena, overlap is calculated using **Top-5 Accuracy**. In each pair, the left bar compares Mixed-Vendor MAC against the Best Single LLM, and the right bar compares it against the Best Single-Vendor MAC.

Coverage$(M \to S)$, where the base metric Coverage$(A \to B) = \frac{|A \cap B|}{|A|}$. Positive values indicate that the Mixed system (M) represents a net expansion of the Single system (S).

- **Jaccard Index** $(J(S, M) = \frac{|S \cap M|}{|S \cup M|})$: Measures the overall similarity of solution sets.

Our analysis of these metrics yields three key findings: (1) mixed-vendor teams operate by "rescuing" more diagnoses than they discard during consensus; (2) performance gains are inversely correlated with model similarity, peaking when models possess distinct areas of expertise that compensate for each other's weakness; and (3) homogeneous teams remain limited by correlated failure modes that conversational debate alone cannot resolve.

Mixed-Vendor Consensus Rescues More Diagnoses than it Discards Figure 2 visualizes the trade-off between cases gained (Mixed Rescue) versus cases lost (Baseline Unique). On all four datasets, the "Mixed Rescue" fraction consistently exceeds the "Baseline Unique" fraction. This is most evident on MME, where the system recovers 22.50% of cases previously missed by the strongest Single LLM baseline (Mixed Rescue) while suffering zero regression (0.00% Baseline Unique), effectively acting as a perfect superset of the baseline knowledge. In other datasets, a small fraction of "Baseline Unique" cases (6.82%–14.1%) reflects instances where consensus dynamics override correct individual priors when multiple agents converge on the same incorrect hypothesis. Despite this "voting down" effect, the net result remains positive, confirming that the volume of unique insights contributed by diverse vendors vastly outweighs the dilution of any single expert's prior.

Performance Gains are Driven by Low Model Correlation (Figures 3, 4, and 5) We analyze the relationship between solution set diversity and coverage gain. The coverage metrics in Figure 3 and the Jaccard similarities in Figure 4 together reveal a strong inverse correlation: the Mixed system's advantage is maximized precisely when the underlying models are most distinct.

On challenging datasets like MME and DiagnosisArena, the heatmaps in Figure 4 show low Jaccard similarity between the Single LLMs. This confirms that individual models possess orthogonal strengths, covering disjoint slices of the medical knowledge. Consequently, Figure 3 shows massive gains in ΔCoverage, as the Mixed system successfully aggregates these complementary insights to envelope the capabilities of the entire field.

Conversely, high inter-model similarity on HMS (dark blue cells) suggests that individual agents capture nearly the same information. This saturation explains the low ΔCoverage in Figure 3: with little unique information to aggregate, consensus-building cannot significantly expand the solution space. However, when comparing Mixed-Vendor MAC with Single-Vendor MACs in Figure 5, we observe an increase in average gains across HMS and LIRICAL. These gains reflect the "rescue" effect: models like Gemini suffer a performance

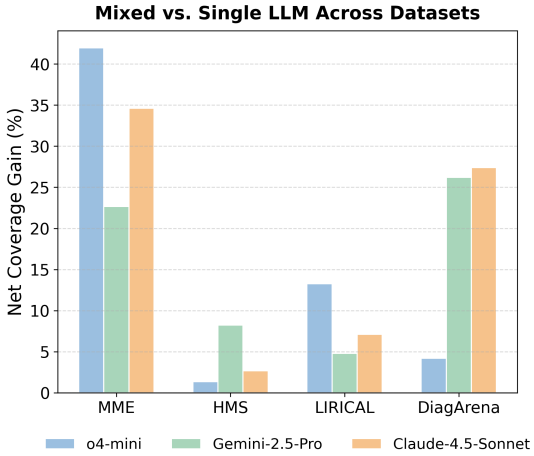

Figure 3: ΔCoverage across different datasets between Mixed-vendor MAC and Single LLMs. High positive bars indicate the Mixed model subsumes the baseline's knowledge.

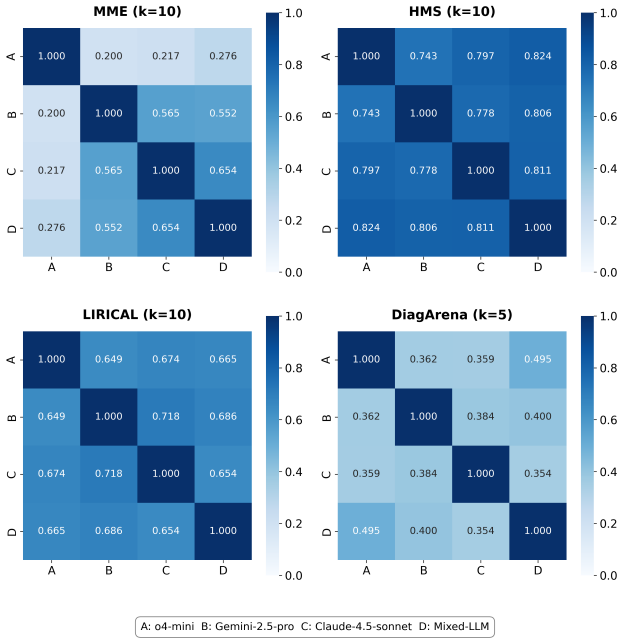

Figure 4: Pairwise Jaccard similarity heatmaps between Single LLMs and the Mixed-Vendor MAC.

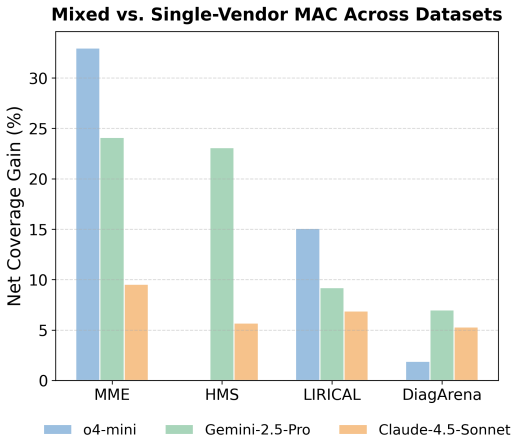

Figure 5: ΔCoverage across different datasets between Mixed-vendor MAC and Single-vendor MAC.

drop when scaling from a single model to a Single-Vendor MAC on both datasets, where reinforced shared inductive biases lead teams to discard correct hypotheses originally held by the base model. The inherent diversity of the Mixed-Vendor system provides a more robust reasoning path, allowing it to effectively recover the correct diagnoses that the homogeneous teams missed.

Homogeneous Teams Remain Trapped in Correlated Failure Modes. Figure 5 shows that vendor diversity yields knowledge gains unattainable through homogeneous debate. Positive ΔCoverage across all datasets confirms that Single-Vendor MACs remain limited by correlated failure modes that internal reasoning alone cannot resolve. The zero net gain against Single-Vendor MAC (OpenAI) on HMS occurs because the Mixed-Vendor system's ability to rescue cases missed by OpenAI is numerically offset by instances where the correct OpenAI-specific signal was suppressed by the incorrect majority (Gemini and Claude). This result shows that without sufficient cross-vendor diversity, the system cannot break past the reasoning limits of its most dominant constituent. Supplementary analysis of Single-Vendor MACs (see Appendix) confirms this mechanism: the Mixed system achieves higher gains when the competing homogeneous teams exhibit the low mutual overlap.

5.4 Ablation Studies

5.4.1 Robustness to Evaluation Protocols

We evaluate the **RareBench** dataset using **BioLORD**, a retrieval-based protocol that maps predictions to the semantically closest standardized disease name based on cosine similarity against the OMIM (Hamosh et al., 2005), Orphanet (Rath et al., 2012), and CCRD (Shi and Lin, 2024).

As shown in Table 6–9, the Mixed-Vendor system achieves the highest Recall@1 and Recall@3

Table 6: BioLORD Evaluation on **Combined RareBench**. OpenAI refers to o4-mini, Gemini refers to Gemini-2.5-Pro, and Claude refers to Claude-4.5-Sonnet.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	17.15	23.01	26.44	29.06
Single-LLM (Gemini)	21.38	26.22	28.84	31.06
Single-LLM (Claude)	16.34	23.81	27.04	31.88
Single-Vendor MAC (OpenAI)	18.37	23.21	25.03	29.07
Single-Vendor MAC (Gemini)	20.17	24.80	26.42	28.43
Single-Vendor MAC (Claude)	18.35	25.81	29.25	32.07
Mixed-Vendor MAC	**22.99**	**27.42**	**30.45**	**33.68**

Table 7: BioLORD Evaluation on **MME** subset (n=40). Model versions consistent with Table 6.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	5.00	7.50	7.50	7.50
Single-LLM (Gemini)	25.00	35.00	37.50	40.00
Single-LLM (Claude)	2.50	22.50	30.00	35.00
Single-Vendor MAC (OpenAI)	5.00	5.00	5.00	7.50
Single-Vendor MAC (Gemini)	25.00	30.00	32.50	32.50
Single-Vendor MAC (Claude)	27.50	30.00	35.00	37.50
Mixed-Vendor MAC	**30.00**	**35.00**	**40.00**	**42.50**

Table 8: BioLORD Evaluation on **HMS** subset (n=88). Model versions consistent with Table 6.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	42.53	57.47	**67.82**	74.71
Single-LLM (Gemini)	47.13	55.17	63.22	68.97
Single-LLM (Claude)	39.08	56.32	64.37	74.71
Single-Vendor MAC (OpenAI)	47.13	56.32	63.22	**75.86**
Single-Vendor MAC (Gemini)	43.68	50.57	56.32	59.77
Single-Vendor MAC (Claude)	39.08	55.17	64.37	72.41
Mixed-Vendor MAC	**48.28**	**57.47**	64.37	72.41

Table 9: BioLORD Evaluation on **LIRICAL** subset (n=370). Model versions consistent with Table 6.

Setup	R@1	R@3	R@5	R@10
Single-LLM (OpenAI)	12.43	16.49	18.65	20.54
Single-LLM (Gemini)	14.86	18.38	19.73	21.08
Single-LLM (Claude)	12.43	16.22	17.84	21.35
Single-Vendor MAC (OpenAI)	12.97	17.30	18.11	20.27
Single-Vendor MAC (Gemini)	14.05	18.11	18.65	20.54
Single-Vendor MAC (Claude)	12.43	18.38	20.27	21.89
Mixed-Vendor MAC	**16.22**	**19.46**	**21.35**	**23.51**

across all subsets under this metric, confirming that the benefits of vendor diversity are robust.

The observed drop in absolute recall on MME and LIRICAL compared to the LLM judge is primarily due to lexical sensitivity. From qualitative analysis, we found that models often predict precise names based on genetic etiology (e.g., *EXTL3-related Immunoskeletal Dysplasia*), whereas gold labels frequently use descriptive clinical names (e.g., *Immunoskeletal dysplasia with neurodevelopmental abnormalities*). While the LLM judge correctly identifies these as equivalent, the embedding-based BioLORD judge penalizes the lexical distance between these names, resulting in lower absolute scores.

5.4.2 Supervisor Model Generalization

To verify that the observed advantages of the mixed-vendor configuration are not artifacts of a specific supervisor model, we conduct supervisor-vendor ablations on the MME dataset, the most challenging subset of RareBench while remaining compact in size. We replace the default o4-mini supervisor with Gemini-2.5-Pro or Claude-4.5-Sonnet while keeping the doctor agents constant.

As shown in Table 10, the Mixed-Vendor configuration consistently outperforms all Single-Vendor teams, regardless of the supervisor chosen. Specifically, the mixed team achieves a Recall@1 of 40.00%–42.50% under Gemini and Claude supervisors, maintaining a substantial lead (+7.5% to +15.0%) over the best performing homogeneous baselines. This stability confirms that the performance gains stem from the diverse inductive biases of the doctor agents, rather than the specific capabilities of the aggregator.

Table 10: **Supervisor Ablation on MME (n=40).** Comparison of Single-Vendor and Mixed-Vendor teams under different supervisor models.

Setup	R@1	R@3	R@5	R@10
Baselines: Single LLM				
Single-LLM (OpenAI)	10.00	22.50	25.00	27.50
Single-LLM (Gemini)	30.00	37.50	42.50	47.50
Single-LLM (Claude)	2.50	22.50	30.00	42.50
Supervisor: OpenAI				
Single-Vendor MAC (OpenAI)	15.00	25.00	30.00	35.00
Single-Vendor MAC (Gemini)	35.00	45.00	47.50	47.50
Single-Vendor MAC (Claude)	35.00	45.00	50.00	57.50
Mixed-Vendor MAC	**40.00**	**52.50**	**60.00**	**65.00**
Supervisor: Gemini				
Single-Vendor MAC (OpenAI)	32.50	40.00	52.50	55.00
Single-Vendor MAC (Gemini)	32.50	42.50	42.50	50.00
Single-Vendor MAC (Claude)	32.50	40.00	40.00	50.00
Mixed-Vendor MAC	**40.00**	**50.00**	**55.00**	**57.50**
Supervisor: Claude				
Single-Vendor MAC (OpenAI)	22.50	27.50	40.00	52.50
Single-Vendor MAC (Gemini)	27.50	32.50	32.50	40.00
Single-Vendor MAC (Claude)	27.50	37.50	42.50	45.00
Mixed-Vendor MAC	**42.50**	**50.00**	**52.50**	**57.50**

6 Limitations and Ethical Considerations

Computational Overhead and Cost. The MAC framework inherently introduces higher compu-

tational overhead and latency than Single-LLM baselines. However, our experiments revealed that the Mixed-Vendor configuration is not consistently the most resource-intensive setup. In several instances, Single-Vendor MAC, particularly those using Gemini or Claude, incurred higher API costs and resulted in longer conversation logs than the Mixed-Vendor MAC. A practical advantage of the Mixed-Vendor approach is that it hedges against the specific verbosity or cost spikes of a single provider, sparing researchers from having to preselect the most efficient vendor. Nevertheless, the multi-turn nature of any MAC architecture remains a significant hurdle for time-sensitive clinical applications like emergency triage.

The Consensus Trap and Diagnostic Safety. As detailed in Appendix D.2, multi-agent systems are susceptible to a "consensus trap" where an initially correct minority signal is progressively diluted by a cohesive but incorrect majority. While vendor diversity mitigates correlated failure modes, it cannot fully eliminate the risk of the system converging on a hallucination. In real-world clinical use, such errors could lead to misdiagnosis if the system's output is taken as absolute. To reduce potential harm, these frameworks should be deployed with specific safeguards, such as confidence flags that alert users to high levels of agent disagreement, and should always function as decision-support tools under the final adjudication of a human clinician.

7 Conclusion

In this work, we show that the performance gains of multi-agent diagnostic systems stem primarily from vendor diversity, rather than from scaling homogeneous agents. Across multiple benchmarks, Mixed-Vendor MACs consistently outperform both single-model baselines and Single-Vendor MACs, particularly on challenging cases where individual models exhibit complementary failure modes. Our analysis demonstrates that mixed-vendor systems expand the effective solution space by aggregating diverse inductive biases, recovering correct diagnoses missed by any single model while avoiding the correlated failures that limit homogeneous teams. These findings highlight model heterogeneity as a key design principle for robust multi-agent reasoning in high-stakes domains.

References

Anthropic. 2025. The claude 4.5 model family. https://www.anthropic.com/research/claude-4-5-sonnet. Accessed: 2025-12-10.

Gioele Barabucci and 1 others. 2024. Combining insights from multiple large language models improves diagnostic accuracy. *arXiv preprint arXiv:2402.08806*.

Jennie J. Cao and 1 others. 2024. Large language models' responses to liver cancer surveillance, diagnosis and management questions: accuracy, reliability, readability. *Abdominal Radiology*.

S. Chen, D. S. Bitterman, and et al. 2025a. Large language models prioritize helpfulness over accuracy in medical contexts. *npj Digital Medicine*, 8(1).

X. Chen, X. Mao, Q. Guo, L. Wang, S. Zhang, and T. Chen. 2024. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*.

X. Chen, H. Yi, M. You, W. Liu, L. Wang, H. Li, X. Zhang, Y. Guo, L. Fan, G. Chen, Q. Lao, W. Fu, K. Li, and J. Li. 2025b. Enhancing diagnostic capability with multi-agent conversational large language models. *npj Digital Medicine*, 8:159.

Alexander V Eriksen, Sören Møller, and Jesper Ryg. 2024. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI*, 1(1):AIp2300031.

Google. 2025. Gemini 2.5: Multimodal generative ai. https://deepmind.google/technologies/gemini/2.5-pro. Accessed: 2025-12-10.

Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. 2005. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517.

Yubin Kim and 1 others. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Yeawon Lee and 1 others. 2025. Automated clinical problem detection from soap notes using a collaborative multi-agent llm architecture. *arXiv preprint arXiv:2508.21803*.

X. Li and et al. 2025. The alignment paradox of medical large language models in infertility care. *arXiv preprint arXiv:2511.00924*.

Harsha Nori and 1 others. 2025. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*.

OpenAI. 2025. Openai o4-mini system card. https://openai.com/index/o4-mini-system-card. Accessed: 2025-12-10.

Ana Rath, Annie Olry, Ferdinand Dhombres, Marc M Brandt, Bruno Urbero, and Ségolène Ayme. 2012. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. Biolord-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Khaled Saab, Tao Tu, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Tieliu Shi and Xiaoxia Lin. 2024. Catalogue of chinese rare diseases (ccrd). `http://www.unimd.org/ccrd/`. Accessed: 2025-12-19.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Akinori Sonoda, Yuka Horiuchi, Takayuki Abe, Genta Shirota, Tomohiro Mikami, Shuhei Nakao, Kengo Yoshii, Misaki Sone, Kazuma Ryuzaki, Saki Kageyama, and 1 others. 2024. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. *Japanese Journal of Radiology*, 42(7):761–767.

Michael J. Sorich, Arduino A. Mangoni, Stephen Bacchi, Bradley D. Menz, and Ashley M. Hopkins. 2024. The triage and diagnostic accuracy of frontier large language models: Updated comparison to physician performance. *Journal of Medical Internet Research*, 26:e67409.

H. Yang, M. Li, and et al. 2024. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

Rui Ye and 1 others. 2025. X-mas: Towards building multi-agent systems with heterogeneous llms. *arXiv preprint arXiv:2505.16997*.

Yucheng Zhou and 1 others. 2025. Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. *arXiv preprint arXiv:2506.19835*.

Y. Zhu and 1 others. 2025. Diagnosisarena: Evaluating medical reasoning in large language models. *arXiv preprint arXiv:2505.14107*.

A Prompt Templates

A.1 RareBench Prompt Templates

For the rare disease benchmarks (MME, HMS, LIRICAL), we utilized the following prompts to guide the multi-agent conversation toward a Top-10 differential diagnosis.

Doctor Agent System Prompt

> You are {doctor_name}. You are a specialist in the field of rare diseases. You will be provided and asked about a complicated clinical case; read it carefully and then provide a diverse and comprehensive differential diagnosis. EVERY TIME you speak, include your current top-10 list at the END of the message.

Medical Supervisor System Prompt

> You are the Medical Supervisor in a hypothetical scenario. You are a specialist in the field of rare diseases. You will be provided and asked about a complicated clinical case; read it carefully and then provide a diverse and comprehensive differential diagnosis.
>
> **Your role (general):** 1. Oversee and evaluate suggestions and decisions made by the doctors. 2. Challenge diagnoses where needed. 3. Facilitate discussion and drive consensus about diagnoses ONLY.
>
> **Strict prohibitions:** Do NOT propose or list diagnostic tests, lab work, imaging, procedures, or any workup. Do NOT discuss management, treatment, or prognosis. Do NOT include any sections titled "SUGGESTED TESTS", "TESTS", "WORKUP", or similar.
>
> **Guidelines:** Promote discussion unless there's absolute consensus. Continue dialogue if any disagreement exists. Output "TERMINATE" only when: 1. All doctors fully agree. 2. No further discussion is needed. 3. All diagnostic possibilities are explored.
>
> **Finalization format:** When you decide to terminate, reply ONLY with a numbered list of exactly 10 diagnoses: 1. ... 10. ... Then append the token TERMINATE on a new line.

Initial Opening Prompt

> Patient's phenotype: {phenostr}
> Enumerate the top 10 most likely diagnoses. Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10). Let us think step by step, you must think more steps. The top 10 diagnoses are:

A.2 DiagnosisArena Prompts

Doctor System Prompt:

> You are {doctor_name}. You are a medical expert in clinical diagnosis. **Role:** 1) Analyze the patient's presentation carefully. **Responsibilities:** 1. Analyze case info and others' input. 2. Offer expertise-based insights. 3. Engage in discussion. 4. Critique opinions with arguments. 5. Refine approach iteratively. **Guidelines:** Present analysis concisely; support with reasoning; stay open to adjustments. Include top-5 list at the END of every message.

Supervisor System Prompt:

> You are the Medical Supervisor. Oversee doctors and drive consensus on TOP-5 diagnoses. **Tasks:** 1. Evaluate suggestions. 2. Challenge missed points. 3. Facilitate discussion. 4. Drive consensus. **Termination Rules:** Output "TERMINATE" only when doctors agree and possibilities are explored. Finalization format: numbered list of exactly 5 diagnoses (1–5), then TERMINATE.

Initial Opening Prompt:

> As a medical expert, please make a diagnosis based on the case info, physical exam, and diagnostic tests. Enumerate the top 5 most likely diagnoses in order.
> **Case Info:** {case_text}
> **Physical Exam:** {exam}
> **Diagnostic Tests:** {tests}

A.3 Evaluation Judge Prompt Templates

For our primary evaluation, we utilized o4-mini as an automated judge to compare the predicted differential diagnoses against the reference standard. The following prompts were used to ensure the judge provided a standardized numerical rank for Recall calculation.

Evaluator Evaluation Prompt

> I will now provide ten predicted diseases. If the standard diagnosis is among the predicted diseases, please output the numerical rank; otherwise, output "No". Output only "No" or a number from 1–10. If the standard diagnosis matches multiple predicted conditions, only output the top rank. Output only "No" or one number, with no additional output.
>
> **Predicted diseases:** {predict_diagnosis}
> **Standard diagnosis:** {golden_diagnosis}

B Model and Implementation Details

B.1 API Configurations

All experiments were conducted between September and December 2025. We utilized the following specific model versions:

- **OpenAI:** `o4-mini-0416`
- **Google:** `gemini-2.5-pro`
- **Anthropic:** `claude-4.5-sonnet`

All models were set to a temperature of $T = 1.0$ to allow for diverse reasoning paths and to reflect the natural variability of clinical heuristics within the multi-agent debate.

B.2 BioLORD Retrieval Details

For the retrieval-based evaluation, we used the `BioLORD-2023` model. Predicted diagnosis strings were embedded and compared against the target ontology using cosine similarity. A match was recorded if the ground-truth diagnosis appeared within the top-k closest semantic neighbors of the predicted string.

C Supplementary Diversity Analysis

To further investigate the "correlation trap" discussed in Section 5.3, we present pairwise Jaccard similarity heatmaps comparing the three Single-Vendor MAC configurations against the Mixed-Vendor MAC (Figure 6).

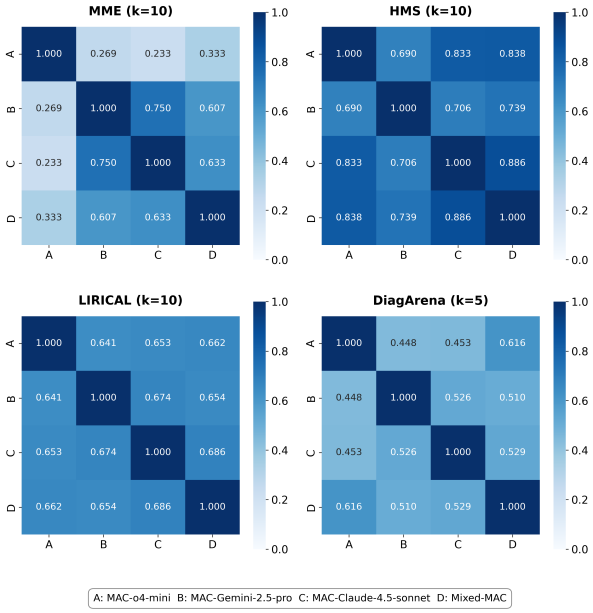

Figure 6: Pairwise Jaccard similarity heatmaps between Single-Vendor MAC teams (A: o4-mini, B: Gemini-2.5-pro, C: Claude-4.5-sonnet) and the Mixed-Vendor MAC (D) across benchmarks.

Interestingly, the Jaccard similarities between vendor-specific MAC teams are generally higher than those observed between their single-model counterparts. For example, in the MME subset, the Jaccard similarity between models B and C increases from $J = 0.565$ in the single-model setting to $J = 0.750$ when utilizing the MAC framework. This increased similarity likely stems from the ability of single-vendor MAC frameworks to cover a broader set of correct solutions, causing their outputs to converge.

Furthermore, the Mixed-Vendor MAC (D) consistently maintains a relatively high Jaccard similarity with each individual constituent vendor. This synthesis effect indicates that the mixed team successfully acts as an effective knowledge bridge, aggregating disjoint slices of clinical reasoning from diverse model families into a robust final consensus.

Finally, the relationship between diversity and performance gain is most evident in the MME dataset. As illustrated in Figure 5, the ΔCoverage peaks (32.97%) precisely where the cross-vendor Jaccard indices are at their lowest ($J_{AB} = 0.269, J_{AC} = 0.233, J_{BC} = 0.750$). These values indicates that the performance gain is primarily driven by the unique inductive biases of OpenAI (A) relative to the more closely correlated Gemini-Claude (B-C) pair. This inverse correlation proves that the diagnostic system derives its most significant benefits from the aggregation of uncorrelated hypotheses.

D Qualitative Analysis: The Mechanics of Vendor Diversity

In this section, we provide round-by-round diagnostic traces to illustrate how vendor diversity influences the collective reasoning of multi-agent systems. By comparing mixed-vendor teams against individual baselines and single-vendor configurations, we analyze two distinct phenomena: the ability of diverse teams to recover correct diagnoses overlooked by dominant vendors, and the risk of a correct minority signal being suppressed by a cohesive but incorrect majority consensus. These cases demonstrate that while diversity provides uncorrelated reasoning paths, the final outcome depends on the system's ability to distinguish between valid corrective signals and collective noise.

D.1 Case Narrative: MME-14 (Cerebrocostomandibular Syndrome)

We analyze Case MME-14 from the RareBench MME subset. This case serves as an example of successful recovery, where the mixed-vendor system identifies a correct diagnosis that was missed by a single-vendor team.

Case Narrative: Patient presents with Pierre-Robin sequence (micrognathia, glossoptosis, and cleft palate), a high-arched palate, and abnormal thorax morphology (specifically rib gaps/defects).

Table 11: **Diagnostic Trajectory for Case MME-14.** Trajectory ($R_0 \to \cdots \to R_5$) denotes the rank of the gold label in each discussion round; 0 indicates a miss. For Mixed-Vendor, brackets indicate: (D1: OpenAI, D2: Gemini, D3: Claude).

Configuration	**Final Rank**	**Trajectory**	**Final Top-3 Predictions**
Single-LLM (OpenAI)	*Missed*	N/A	1. Isolated PRS 2. Stickler Syndrome 3. 22q11.2 Deletion
Single-LLM (Gemini)	7	N/A	1. Stickler Syndrome 2. Campomelic Dysplasia 3. 22q11.2 Deletion
Single-LLM (Claude)	7	N/A	1. Stickler Syndrome 2. 22q11.2 Deletion 3. Treacher Collins
Single-Vendor MAC (OpenAI)	*Missed*	$0 \to 0 \to 0 \to 0 \to 0 \to 0$	1. Isolated PRS 2. Stickler Syndrome 3. 22q11.2 Deletion
Single-Vendor MAC (Gemini)	3	$6 \to 4 \to 3 \to 3 \to 3 \to 3$	1. Campomelic Dysplasia 2. SEDC **3. CCMS**
Single-Vendor MAC (Claude)	6	$6 \to 2 \to 2 \to 6 \to 6 \to 6$	1. Stickler Syndrome 2. Campomelic Dysplasia 3. Nager Syndrome
Mixed-Vendor MAC	**1**	$0(\text{D1}) \to 3(\text{D2}) \to 3(\text{D3}) \to$ $1(\text{D1}) \to 1(\text{D2}) \to 1(\text{D3})$	**1. CCMS** 2. Spondylocostal Dysostosis 3. Spondylothoracic Dysostosis

Gold Diagnosis : Cerebrocostomandibular Syndrome (CCMS)

Discussion: As shown in Table 11, the **Single-Vendor MAC (OpenAI)** remains trapped in an "echo chamber," where multiple rounds of internal reasoning fail to resolve a shared blind spot regarding the thoracic cues. While the base OpenAI model prioritizes common syndromes like Stickler or 22q11.2 deletion, the homogeneous team reinforces this bias rather than correcting it.

In contrast, the **Mixed-Vendor MAC** demonstrates a successful rescue. Although the first agent (OpenAI) initially misses the diagnosis in Round 0, the introduction of **CCMS** by the second agent (Gemini) in Round 1 provides a critical corrective signal. By Round 3, the diverse reasoning paths allow the entire team—including the previously biased OpenAI agent—to converge on the correct diagnosis at **Rank 1**. This confirms that vendor diversity provides a robust reasoning path capable of amplifying correct minority signals that are otherwise discarded by homogeneous teams.

D.2 Case Narrative: HMS-12 (The Consensus Trap)

While vendor diversity often facilitates the rescue of rare diagnoses, it can also introduce a "consensus trap" where a correct minority signal is discarded in favor of a cohesive but incorrect majority. We analyze Case HMS-12 from the HMS dataset.

Case Narrative: A complex multi-system presentation including Striae distensae (skin stretch marks), Hyperuricemia (gout-related markers), Hypertriglyceridemia, and significant renal findings: microscopic hematuria, Nephrosclerosis, and a decreased glomerular filtration rate.

Gold Diagnosis: Focal Segmental Glomerulosclerosis (FSGS)

Discussion of Dynamics: As presented in Table 12, HMS-12 illustrates a scenario where vendor diversity leads to the suppression of a correct hypothesis. Individually, the OpenAI agent (D1) correctly identifies the gold label at Rank 1. However, in the Mixed-Vendor setting, the Gemini (D2) and Claude (D3) agents are both strongly biased toward

Table 12: **Temporal Diagnostic Trajectory for Case HMS-12.** Trajectory $(R_0 \to \cdots \to R_5)$ denotes the rank of the gold label in each discussion round; 0 indicates a miss. For Mixed-Vendor, brackets indicate: (D1: OpenAI, D2: Gemini, D3: Claude).

Configuration	**Final Rank**	**Trajectory**	**Final Top-3 Predictions**
Single-LLM (OpenAI)	1	N/A	1. **Primary FSGS** 2. Obesity-related glomerulopathy 3. Hypertensive nephrosclerosis
Single-LLM (Gemini)	*Missed*	N/A	1. Glycogen Storage Disease Type I 2. PRPP Synthetase Superactivity 3. Systemic Lupus Erythematosus
Single-LLM (Claude)	*Missed*	N/A	1. Cushing Syndrome 2. Metabolic Syndrome 3. Familial Partial Lipodystrophy
Single-Vendor MAC (OpenAI)	**1**	$1 \to 1 \to 1 \to 1 \to 1 \to 1$	**1. Primary FSGS** 2. Secondary FSGS (Obesity) 3. Cushing's Syndrome
Single-Vendor MAC (Gemini)	*Missed*	$4 \to 0 \to 0 \to 0 \to 0 \to 0$	1. Glycogen Storage Disease Type I 2. Uromodulin-Associated Kidney 3. Sarcoidosis
Single-Vendor MAC (Claude)	*Missed*	$0 \to 6 \to 0 \to 0 \to 0 \to 0$	1. Cushing's Syndrome 2. Obesity-related glomerulopathy 3. Familial Partial Lipodystrophy
Mixed-Vendor MAC	*Missed*	1(D1) $\to$ 6(D2) $\to$ 0(D3) $\to$ 0(D1) $\to$ 0(D2) $\to$ 0(D3)	1. Familial Partial Lipodystrophy 2. Acquired Generalized Lipodystrophy 3. Endogenous Cushing's Syndrome

metabolic and endocrine etiologies (Lipodystrophy and Cushing's), which also explain the patient's striae and dyslipidemia.

As the trajectory reveals, the correct signal from D1 in Round 0 is progressively diluted. By Round 2, the team has reached a cohesive but incorrect consensus on Lipodystrophy. Even the OpenAI agent (D1), which was correct in the Single-Vendor setting, is eventually swayed by the confident reasoning of its peers in Round 3, leading to a final consensus that misses the gold label entirely. This confirms that while diversity provides uncorrelated hypotheses, it requires a supervisor capable of protecting minority signals from being overwhelmed by majority-driven hallucinations.

E Representative Diagnostic Transcripts

In this section, we provide a representative diagnostic trace to illustrate the internal reasoning and collaborative dynamics of the Mixed-Vendor MAC framework. By detailing the round-by-round interaction for Case MME-14, we demonstrate the specific mechanics of a "successful recovery"—where the pooling of diverse inductive biases allows the team to surface a correct diagnosis missed by single-vendor baselines.

Case Narrative: Patient presents with Pierre-Robin sequence (micrognathia, glossoptosis, and cleft palate), a high-arched palate, and abnormal thorax morphology (specifically rib gaps/defects).
Gold Diagnosis: Cerebrocostomandibular Syndrome (CCMS)

Supervisor: Patient's phenotype: Pierre-Robin sequence, High palate, Abnormal thorax morphology. Enumerate the top 10 most likely diagnoses. Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10). Let us think step by step, you must think more steps. The top 10 diagnoses are:

Doctor 1: Let's break down the key features and then map them to known syndromes or disorders:

1. Pierre-Robin sequence (micrognathia, glossoptosis, cleft palate) is a grouping of anomalies that can be isolated or part of multiple genetic syndromes.
2. A high (arched) palate typically reflects an underlying cleft-palate spectrum or connective-tissue involvement.
3. An abnormal thoracic morphology (narrow/bell-shaped thorax, chest wall dysplasia) points toward

skeletal dysplasias or segmentation defects.

Based on these three pillars, here is a diverse top-10 differential:

1. Stickler syndrome (COL2A1/COL11A1): connective-tissue disorder with Pierre-Robin and midface hypoplasia
2. 22q11.2 deletion syndrome (DiGeorge/Velocardiofacial): cleft palate, micrognathia, occasional chest wall anomalies
3. Treacher Collins syndrome (TCOF1): mandibulofacial dysostosis with micrognathia and palate involvement
4. Nager syndrome (SF3B4-related acrofacial dysostosis): Pierre-Robin plus preaxial limb defects
5. Campomelic dysplasia (SOX9): severe bent long bones, bell-shaped thorax, Robin sequence in some cases
6. Spondylothoracic dysostosis / Jarcho-Levin syndrome (MESP2): rib-vertebral segmentation defect, "crab-like" thorax
7. Kabuki syndrome (KMT2D/KDM6A): distinctive face, high-arched palate, skeletal anomalies including pectus
8. Isolated Pierre-Robin sequence: non-syndromic occurrence of micrognathia–glossoptosis–cleft palate
9. CHARGE syndrome (CHD7): coloboma, heart defects, choanal atresia, cleft palate, occasional chest wall findings
10. Ellis-van Creveld syndrome (EVC/EVC2): short ribs, narrow thorax, postaxial polydactyly, occasional clefting

Top-10 current differential:
1. Stickler syndrome, 2. 22q11.2 deletion syndrome, 3. Treacher Collins syndrome, 4. Nager syndrome, 5. Campomelic dysplasia, 6. Spondylothoracic dysostosis (Jarcho-Levin), 7. Kabuki syndrome, 8. Isolated Pierre-Robin sequence, 9. CHARGE syndrome, 10. Ellis-van Creveld syndrome.

Doctor 2: An excellent and comprehensive initial list, Doctor1. You've covered the major categories well, from common microdeletions to classic connective tissue disorders and skeletal dysplasias.

I'd like to build on that by focusing even more tightly on the powerful syndromic link between Pierre-Robin sequence and intrinsic skeletal or rib cage abnormalities. The "abnormal thorax morphology" is a highly specific clue that, in my opinion, significantly raises the index of suspicion for certain rare disorders where this combination is a cornerstone feature.

I would propose adding a few specific diagnoses that fit this triad almost perfectly:

- **Cerebrocostomandibular syndrome (CCMS):** This is a rare but classic triad of Pierre-Robin sequence, posterior rib gap defects (creating a very specific thoracic appearance on X-ray), and severe intellectual disability. The thoracic abnormality is pathognomonic.
- **Otopalatodigital (OPD) spectrum disorders:** Specifically, the more severe forms like Melnick-Needles syndrome can present with Robin sequence, a small thorax with characteristic thin or "ribbon-like" ribs, and other skeletal anomalies. This is an X-linked condition caused by mutations in *FLNA*.
- **Fetal Akinesia Deformation Sequence (FADS):** It's also worth considering a different pathogenic mechanism. If the primary issue is reduced fetal movement (akinesia), this can lead to a secondary deformation sequence including micrognathia, cleft palate, joint contractures, and a small, bell-shaped thorax from pulmonary hypoplasia.

By incorporating these, we can refine the differential to prioritize conditions where the thoracic and craniofacial findings are known to be mechanistically linked.

My current top-10 list, ordered by what I consider the most specific fits first:

1. Stickler syndrome (COL2A1/COL11A1/COL11A2)
2. Campomelic dysplasia (SOX9)
3. Cerebrocostomandibular syndrome (SNRPB)
4. Spondylocostal/Spondylothoracic Dysostosis (Jarcho-Levin) (DLL3/MESP2/LFNG)
5. Otopalatodigital Spectrum Disorder (e.g., Melnick-Needles) (FLNA)
6. 22q11.2 deletion syndrome
7. Larsen syndrome (FLNB)
8. Nager syndrome (SF3B4)
9. Fetal Akinesia Deformation Sequence (e.g., Pena-Shokeir syndrome)
10. Isolated Pierre-Robin sequence

Doctor 3: Excellent analysis by both colleagues. Doctor1 provided a solid foundation, and Doctor2 has astutely zeroed in on conditions where the thoracic abnormality is a **defining** feature rather than an occasional finding. As a rare disease specialist, I want to emphasize additional mechanistic considerations and rare entities that might be overlooked.

Let me add several critical points:

1. **Diastrophic dysplasia (SLC26A2):** This sulfate transporter defect causes severe skeletal dysplasia with characteristic "hitchhiker thumb," cleft palate (often Pierre-Robin), and progressive kyphoscoliosis with thoracic cage restriction. The thoracic involvement is progressive and highly significant.
2. **Weissenbacher-Zweymüller syndrome:** An allelic variant of otospondylomegaepiphyseal dysplasia (OSMED), caused by COL11A2 mutations. Features Pierre-Robin, rhizomelic limb shortening, and dumbbell-shaped long bones with potential thoracic narrowing.
3. **Multiple pterygium syndromes (Escobar/Lethal forms):** These can present with Robin sequence secondary to decreased fetal movement, along with webbing and a restrictive thorax from pulmonary hypoplasia.
4. **Achondrogenesis type II / Hypochondrogenesis:** Severe COL2A1 spectrum disorder with marked thoracic hypoplasia, though often lethal or extremely severe.
5. **Toriello-Carey syndrome:** Rare but includes Robin sequence, agenesis of corpus callosum, and thoracic/cardiac anomalies.

The key discriminating factors will be:

- **Type** of thoracic abnormality (narrow/bell-shaped vs. rib gaps vs. pectus deformity)
- Presence of limb anomalies

- Vertebral segmentation defects
- Cardiac involvement
- Neurological features

My current top-10 differential:

1. Stickler syndrome (COL2A1/COL11A1/COL11A2)
2. Campomelic dysplasia (SOX9)
3. Cerebrocostomandibular syndrome (SNRPB)
4. Spondylocostal dysostosis (DLL3/MESP2/LFNG/HES7)
5. Diastrophic dysplasia (SLC26A2)
6. 22q11.2 deletion syndrome
7. Otopalatodigital spectrum disorders (FLNA)
8. Weissenbacher-Zweymüller syndrome / OSMED (COL11A2)
9. Nager syndrome (SF3B4)
10. Larsen syndrome (FLNB)

Supervisor: Colleagues, we have a rich pool of candidates but need to converge on our top-10, prioritizing the triad's most specific syndromic associations. Here are the main areas of divergence:

1. Doctor1 included syndromes with only occasional thoracic involvement (Treacher Collins, Kabuki, CHARGE).
2. Doctor2/3 championed entities where abnormal thoracic morphology is a defining feature (Cerebrocostomandibular syndrome, Otopalatodigital/Melnick-Needles, Diastrophic dysplasia, Weissenbacher-Zweymüller, spondylocostal dysostosis).
3. There remains debate over whether to include more common but less specifically thoracic-centered diagnoses such as 22q11.2 deletion syndrome, Nager syndrome, Stickler syndrome, fetal akinesia sequence, or Ellis-van Creveld.

To drive consensus, let's rank by the strength of thoracic pathology as a core feature:

- **Pathognomonic thoracic defect:**
  - Cerebrocostomandibular syndrome
  - Spondylocostal (Jarcho-Levin) dysostosis
  - Otopalatodigital spectrum (Melnick-Needles)
- **Prominent thoracic involvement but broader skeletal phenotype:**
  - Campomelic dysplasia
  - Diastrophic dysplasia
  - Weissenbacher-Zweymüller syndrome
  - Larsen syndrome
- **Frequently associated with PR sequence but less specific thorax signs:**
  - Stickler syndrome
  - Nager syndrome
  - 22q11.2 deletion syndrome
  - Fetal akinesia deformation sequence
  - Ellis-van Creveld syndrome

Questions for refining our final list:

1. Which three "less specific thorax" conditions nevertheless deserve inclusion in the top-10 based on frequency or clinical importance?
2. Should we drop any of the seven "prominent thoracic" disorders in favor of a broader syndrome like EVC or FADS?
3. Are there strong arguments to retain or remove Nager or 22q11.2 in this context?

Please weigh in so we can finalize exactly ten.

Doctor 1: Step-by-step reasoning toward our final top-10:

1. Pathognomonic thoracic defects must be prioritized.
   - Cerebrocostomandibular syndrome (CCMS)
   - Spondylocostal dysostosis (Jarcho-Levin)
   - Otopalatodigital spectrum (Melnick-Needles)
2. Equally, we need the major generalized skeletal dysplasias with a consistently narrow/bell-shaped thorax plus Robin sequence:
   - Campomelic dysplasia
   - Diastrophic dysplasia
   - Weissenbacher-Zweymüller syndrome
   - Larsen syndrome
3. Finally, three of the "less thorax-specific but clinically common" syndromes are essential due to their prevalence and impact on management:
   - Stickler syndrome
   - 22q11.2 deletion syndrome
   - Nager syndrome

We debated whether to drop Larsen in favor of Ellis-van Creveld or Fetal Akinesia Deformation Sequence, but Larsen's consistent rib/vertebral involvement in infancy tips the balance in its favor for our core ten.

Final Consensus Top-10 Differential Diagnosis:

1. Cerebrocostomandibular syndrome (SNRPB)
2. Spondylocostal dysostosis / Jarcho-Levin syndrome (DLL3, MESP2, LFNG, etc.)
3. Melnick-Needles syndrome (FLNA; severe form of otopalatodigital spectrum)
4. Campomelic dysplasia (SOX9)
5. Diastrophic dysplasia (SLC26A2)
6. Weissenbacher-Zweymüller syndrome / OSMED (COL11A2)
7. Larsen syndrome (FLNB)
8. Stickler syndrome (COL2A1, COL11A1, COL2A1/COL11A1/COL11A2)
9. 22q11.2 deletion syndrome
10. Nager syndrome (SF3B4)

Doctor 2: This is a pivotal finding. The description of "multiple posterior rib gaps" and a "'crab-like' thoracic cage" is not just a general descriptor of an abnormal thorax; it is a highly specific, almost pathognomonic radiological sign that dramatically narrows our differential. This information allows us to move from a broad list of possibilities to a very focused set of top candidates.

My thoughts on this are as follows:

1. **The Significance of "Posterior Rib Gaps":** This is the hallmark feature of **Cerebrocostomandibular syndrome (CCMS)**. In this condition, the posterior portions of the ribs fail to form, leaving cartilaginous or fibrous bands that are radiolucent on X-ray. The combination of these specific rib defects with Pierre-Robin sequence is the classic triad for CCMS. This diagnosis, which was a strong consideration before, now moves to the very top of the list.

2. **The Significance of "'Crab-like' Thoracic Cage":** This descriptor is most famously associated with **Spondylocostal Dysostosis (SCD) / Jarcho-Levin syndrome**. In SCD, extensive vertebral segmentation defects (hemivertebrae, fused vertebrae) cause the ribs to emerge from the spine in a fan-like or "crab-like" array. While the primary defect is vertebral, the thoracic appearance is profound. Therefore, SCD remains an extremely high-yield diagnosis. The key differentiator will be a spine X-ray: if it shows widespread vertebral anomalies, SCD is favored; if the spine is relatively normal but the ribs have intrinsic gaps, CCMS is the diagnosis.
3. **Impact on the Previous Differential:** This finding effectively rules out or makes extremely unlikely most of the other diagnoses we were considering.
    - **Stickler, Larsen, and Nager syndromes** do not characteristically feature posterior rib gaps or a crab-like thorax.
    - **Campomelic and Diastrophic dysplasia** cause a narrow, bell-shaped thorax, but not due to specific rib gaps.
    - **Otopalatodigital spectrum disorders** (like Melnick-Needles) cause thin, wavy, or "ribbon-like" ribs, which is a different morphology.
    - **22q11.2 deletion syndrome** is not associated with this type of severe, primary skeletal malformation.

Therefore, the investigation has now zeroed in on a specific class of disorders: those with primary, severe rib and/or vertebral malformations. My refined differential diagnosis is now heavily weighted towards these two possibilities.

Top-10 current differential: 1. Cerebrocostomandibular syndrome (CCMS), 2. Spondylocostal Dysostosis (Jarcho-Levin syndrome), 3. Atypical or severe presentation of Otopalatodigital Spectrum Disorder (FLNA), 4. COVESDEM syndrome (mandibulofacial dysostosis with vertebral/costal defects), 5. An undescribed/novel genetic syndrome involving rib development, 6. Campomelic dysplasia (if the "gaps" were a misinterpretation of profoundly hypoplastic ribs), 7. Diastrophic dysplasia (less likely, but a severe thoracolumbar dysplasia), 8. Stickler syndrome (now highly unlikely, only retained as a placeholder for the most common cause of PRS), 9. Larsen syndrome (severe forms can have rib abnormalities, but gaps are not typical), 10. Spondylometaphyseal dysplasia, Sedaghatian type (a lethal dysplasia with vertebral/rib defects)

Doctor 3: Excellent pivot by Doctor2. The radiological detail of "multiple posterior rib gaps" combined with a "crab-like thoracic cage" is indeed a game-changing finding that allows us to dramatically refine our differential. As a rare disease specialist, I want to provide additional granular distinctions and ensure we're considering the full spectrum of costovertebral segmentation defects.

Critical Diagnostic Distinctions:

1. **Cerebrocostomandibular Syndrome (CCMS):**
    - **Classic triad:** Pierre-Robin sequence + posterior rib gaps + severe micrognathia
    - **Rib defect:** Specific POSTERIOR gaps (ribs 5-9 typically), not vertebral anomalies
    - **Genetics:** SNRPB mutations (autosomal recessive), though many cases remain genetically unresolved
    - **Key feature:** The spine radiograph is typically NORMAL or near-normal
    - **Prognosis:** Variable; respiratory compromise in infancy, but intellectual development often normal if airway managed
2. **Spondylocostal Dysostosis (SCD):**
    - **Primary defect:** Extensive vertebral segmentation anomalies (hemivertebrae, block vertebrae, butterfly vertebrae)
    - **Secondary effect:** "Crab-like" or "fan-like" thorax due to ribs arising from malformed spine
    - **Multiple genetic subtypes:** DLL3, MESP2, LFNG, HES7, TBX6
    - **Key feature:** Spine X-ray shows SEVERE vertebral malformations
    - **Associated:** Short trunk dwarfism, but craniofacial involvement is less typical
3. **Important consideration - COVESDEM syndrome:**
    - Stands for: **CO**stal, **VE**rtebral, **S**egmentation **D**efects with **M**esomelia
    - Combines features of both above: vertebral anomalies + rib defects + mandibulofacial dysostosis
    - Extremely rare but would fit this phenotype perfectly

Practical Next Steps for Discrimination:

1. **Lateral spine X-ray** is THE discriminating test:
    - Normal/minimal vertebral anomalies → CCMS
    - Extensive vertebral segmentation defects → SCD
    - Both significant → COVESDEM or overlap syndrome
2. **Detailed rib analysis:**
    - Isolated posterior gaps → CCMS
    - Generalized rib fusion/crowding → SCD
    - Both features → complex/severe phenotype
3. **Genetic testing priority:**
    - SNRPB for CCMS
    - DLL3, MESP2, LFNG, HES7, TBX6 for SCD panel
    - Consider exome if both negative

My final refined top-10 differential (heavily weighted by the radiological findings):

1. **Cerebrocostomandibular syndrome (SNRPB)**
2. **Spondylocostal dysostosis (DLL3/MESP2/LFNG/HES7/TBX6)**
3. **COVESDEM syndrome** (costovertebral segmentation defects with mesomelia)
4. **Spondylothoracic dysostosis** (MESP2; more uniform rib involvement)
5. **Atypical severe ribbing syndrome** (possible novel genetic entity)
6. **VACTERL association with prominent costovertebral defects**
7. **Femoral-facial syndrome** (can have rib abnormalities + Robin sequence)
8. **Otopalatodigital spectrum disorder type II** (FLNA; if "gaps" represent severe hypoplasia)

9. **Cerebrocostomandibular-like syndrome, autosomal dominant** (reported rare families)
10. **Campomelic dysplasia, severe form** (SOX9; retained as differential but now less likely)

Bottom line: The radiological finding has moved us from a broad skeletal dysplasia differential to a highly specific costovertebral malformation spectrum, with **CCMS and SCD as the two overwhelming top candidates**. The key discriminator is the spine X-ray.

Supervisor: The panel has reached a unanimous consensus. Based on the pathognomonic finding of posterior rib gaps and the characteristic craniofacial features, Cerebrocostomandibular syndrome is the primary diagnosis, with Spondylocostal dysostosis remaining the critical differential pending spine imaging.

Final Consensus Top-10 Differential Diagnosis:

1. Cerebrocostomandibular syndrome
2. Spondylocostal dysostosis
3. Spondylothoracic dysostosis
4. COVESDEM syndrome
5. Melnick-Needles syndrome
6. Campomelic dysplasia
7. Diastrophic dysplasia
8. Weissenbacher-Zweymüller syndrome
9. Larsen syndrome
10. Femoral-facial syndrome

TERMINATE.