BEYOND PRESENTATION: TOWARD FORMAT-ROBUST MEDICAL EVIDENCE REASONING IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical evidence-source preference, where LLMs systematically favor certain evidence sources when reasoning about competing diagnoses, is an underexplored factor in reliable medical reasoning. However, whether such preferences reflect evidential content or are confounded by evidence presentation remains unclear. To investigate this question, we build PMC-ConForm, a benchmark of medical evidence conflicts from PMC-Patients case reports where evidence source and presentation format vary independently. Through detailed empirical analyses, we find that LLMs exhibit systematic medical source preferences, yet these preferences can shift substantially and even reverse under content-preserving format changes, indicating that evidence prioritization partly reflects presentation rather than evidence itself. To address this issue, we propose **FORMEDER**, a cross-format consensus self-distillation framework for format-robust medical evidence reasoning. We re-render the same evidence into multiple content-preserving formats and aggregate the frozen model's predictions into a soft consensus target shared across presentations. We then distill the model toward this target under each format while introducing a contraction objective to suppress residual format-dependent variation. The framework requires no external annotation and uses a single presentation at inference. Extensive experiments on eight LLMs demonstrate that FORMEDER substantially mitigates format-induced preference shifts and enhances cross-format choice consistency while faithfully preserving presentation-invariant judgments, and generalizes to unseen presentation formats.

1 INTRODUCTION

Large language models (LLMs) are increasingly applied to medical question answering (Singhal et al., 2023; 2025; Nori et al., 2023) and clinical reasoning (Thirunavukarasu et al., 2023; Hager et al., 2024), where a single case often involves heterogeneous evidence from pathology, imaging, laboratory tests, physical examinations, and patient-reported records. When different sources provide competing support for diagnostic hypotheses, models must appropriately weigh and integrate their evidence (Xu et al., 2024; Wan et al., 2024; Wu et al., 2024). In this process, LLMs may systematically assign greater diagnostic influence to some medical evidence sources than to others, a behavior we refer to as medical evidence-source preference. Such preferences are not inherently problematic, as different evidence sources may legitimately carry different diagnostic value.

The key question is whether an observed preference reflects these evidence-dependent differences. In medical reasoning, evidence source and presentation format are often naturally coupled, with laboratory results represented as structured fields or tables, pathology reports following sectioned templates, and patient-reported information expressed as free-form text. This coupling becomes particularly important given prior findings that LLM predictions can be sensitive to meaning-preserving changes in presentation (Sclar et al., 2024; Liu et al., 2026a), suggesting that an observed source preference may reflect not only how a model weighs the underlying evidence, but also how that evidence is presented. This ambiguity motivates our central research question:



To what extent are observed medical evidence-source preferences driven by the evidence itself rather than its presentation format?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

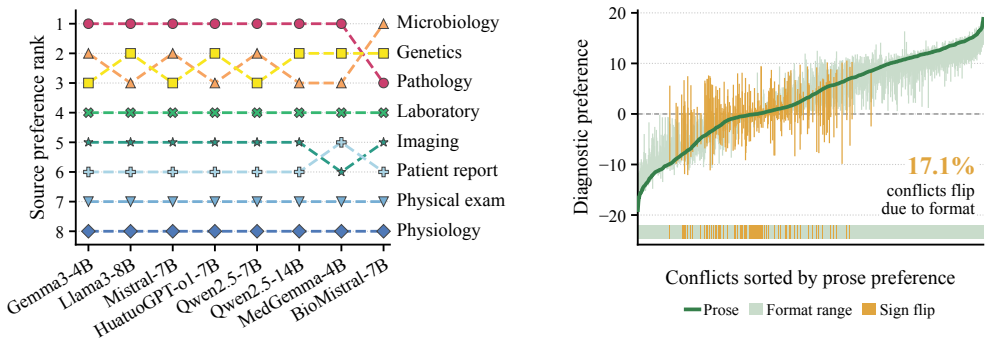

Figure 1: **Left:** Evidence-source preference rankings across eight LLMs under prose presentation. **Right:** MedGemma-4B's diagnostic preferences across content-preserving presentations; colored ranges show preference variation, with zero crossings indicating diagnosis reversals.

To investigate this question, we construct PMC-**ConForm** (**Con**flicts across **Form**ats), a medical evidence-conflict benchmark built from the case reports in PMC-Patients (Zhao et al., 2023), in which evidence source and presentation format can be varied independently. Our empirical analysis reveals two key findings. (1) ***Systematic Medical Evidence Preference***: when conflicting medical evidence is presented in a common format, LLMs exhibit clear, systematic, and highly consistent preferences across eight different medical evidence sources, as shown in Figure 1 (left). (2) ***Format-Induced Instability***: when the same evidence is re-rendered into content-preserving presentations, these preferences can shift substantially and even reverse, as shown in Figure 1 (right). This contrast shows that even a highly systematic source preference does not necessarily reflect the underlying evidence, as it can still depend strongly on how that evidence is presented.

These findings call for a model that keeps its diagnostic preference unchanged when only the presentation changes (**invariance**) while retaining the presentation-invariant component of its original judgment (**preservation**). The difficulty is estimating this component, since every presentation may carry its own format effect and no gold label specifies it. We therefore treat content-preserving presentations as counterfactual views of the same evidence and aggregate the model's predictions across them into an instance-specific reference that favors no single presentation.

Building on this insight, we propose **FORMEDER**, a three-stage self-distillation framework using Cross-**For**mat Consensus for **Med**ical **E**vidence **R**easoning. ① ***Counterfactual Evidence View Modeling*** re-renders each conflict into content-preserving views, defining presentation variation to remove. ② ***Cross-Format Consensus Inference*** averages the frozen model's predictions across views to estimate the judgment they share. ③ ***Consensus-Preserving Robust Alignment*** distills each view toward this consensus while contracting cross-view preference gaps, preserving this shared judgment and reducing presentation sensitivity. All supervision comes from the model's own predictions, requiring no external annotation, and FORMEDER uses a single presentation at inference.

To demonstrate the effectiveness of FORMEDER, we conduct experiments on eight LLMs, spanning general-purpose and domain-adapted models, across diverse presentation formats. Our results show that FORMEDER substantially reduces diagnostic preference shifts caused by presentation changes and improves decision consistency across formats, while preserving evidence-specific diagnostic behavior. Furthermore, FORMEDER generalizes to unseen presentation formats, and exposing the model to more training formats improves this generalization, demonstrating that the learned robustness transfers beyond the observed presentations. Finally, ablation studies validate the complementary roles of consensus distillation and residual contraction, with consensus distillation preserving shared judgments and residual contraction suppressing remaining presentation-induced variation.

2 PRELIMINARIES

Medical Evidence Conflict. We study medical reasoning settings in which evidence records from different sources support competing diagnoses. Given a medical question q, we consider a conflict between two evidence records E_A and E_B, which come from evidence sources s_A and s_B, respectively. The two records support different diagnoses D_A and D_B. Each record can be rendered in different presentation formats. We denote the format of E_A by f and the format of E_B by g.

Formally, we define the input of a conflict as

$$x_{AB}^{f,g} = (q, \mathcal{R}_f(E_A), \mathcal{R}_g(E_B), \{D_A, D_B, D_{\mathrm{N}}\}), \tag{1}$$

where $\mathcal{R}_f(\cdot)$ renders an evidence record in format f, and D_{N} is the option *Neither of the above*. We focus on content-preserving renderings, which keep the evidence statements and source identity of a record intact and change only how the record is organized. As a result, all renderings of a record E convey the same evidence. We collect them into the set of counterfactual renderings

$$\mathcal{V}(E) = \{\mathcal{R}_f(E) \mid f \in \mathcal{F}\}, \tag{2}$$

where $\mathcal{F}$ is the set of presentation formats we consider.

In our experiments, $\mathcal{F}$ includes prose, sectioned text, key-value fields, tables, and JSON. During evaluation, we balance the order of the two records and the order of the answer options to control for position bias. Throughout the paper, A and B refer to the underlying evidence records rather than to their displayed positions.

Diagnostic Preference. For the input $x = x_{AB}^{f,g}$ of a conflict, we denote the model's prediction by $\widehat{D}(x) \in \{D_A, D_B, D_{\mathrm{N}}\}$. The discrete prediction does not show how strongly the model prefers one diagnosis over the other. We therefore also measure the direction and strength of this preference using the probabilities that the model assigns to the two competing diagnoses.

Let p_A and p_B be the probabilities of the two competing diagnoses under a fixed answer-scoring protocol. We define the diagnostic preference score as

$$M_{AB}(f, g) = \log p_A - \log p_B = \log \frac{p_A}{p_B}. \tag{3}$$

A positive score indicates a preference for D_A and a negative score indicates a preference for D_B. This score captures the relative preference between the two diagnoses independently of their total probability mass. Moreover, changes on the log scale have a consistent interpretation across preference levels; for example, a change of $\ln 2 \approx 0.69$ corresponds to a twofold change in the odds. We use this score to quantify both evidence-source preference and its sensitivity to presentation changes.

Format-Induced Preference Shift. To test whether the diagnostic preference stays stable under presentation changes, we compare the same evidence conflict under two presentations that differ in the format of a single record. Suppose that we re-render only E_A from a reference format f_0 (prose in our experiments) into another format f and keep E_B in format g. We define the resulting format-induced preference shift as

$$\Delta M_A(f; g) = M_{AB}(f, g) - M_{AB}(f_0, g). \tag{4}$$

The two inputs contain exactly the same evidence, so this paired comparison isolates the effect of presentation format. We define ΔM_B symmetrically for the case where only E_B is re-rendered. For both shifts, a positive value means that the format change increases the preference for the diagnosis supported by the re-rendered record, and a negative value means that it decreases this preference.

In our analysis, we aggregate ΔM_A and ΔM_B over the position configurations. This aggregated measure lets us test whether an observed evidence-source preference reflects the evidence content alone or also depends systematically on how the evidence is presented across formats.

3 MOTIVATION

We begin by studying how LLMs weigh medical evidence sources when they resolve conflicting diagnoses, and whether the resulting source preferences reflect the evidence itself. For this analysis, we use the analysis set of PMC-ConForm, constructed from PMC-Patients case reports and covering all pairwise combinations of eight evidence sources across heterogeneous evidence types. Each conflict pairs clinically matched records supporting distinct diagnoses and is verified against the original case narratives. We independently vary evidence source and presentation format to distinguish source-dependent preference from presentation-induced variation (Appendix B).

Systematic Preference across Medical Evidence Sources. We first isolate source-dependent preference by rendering conflicting records in prose, thereby removing variation in presentation format.

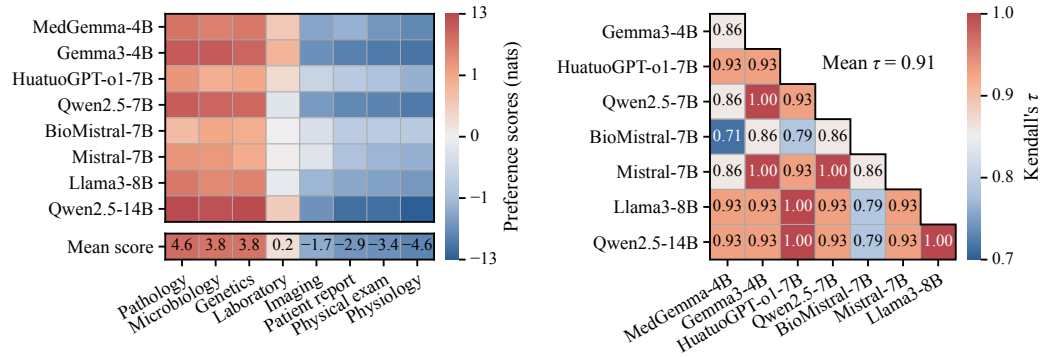

Figure 2: **Left:** Preference scores of eight evidence sources across eight LLMs in prose, with larger scores indicating stronger preference and mean scores in the bottom row. **Right:** Pairwise Kendall's τ between models source rankings, where higher values indicate stronger cross-model agreement.

We evaluate all 28 pairs formed by the eight evidence sources and summarize the results with a normalized preference score for each source, where a larger score indicates greater diagnostic weight.

All eight LLMs exhibit clear and highly consistent preferences across medical evidence sources. As Figure 2 (left) shows, pathology, microbiology, and genetics consistently receive greater diagnostic weight, whereas physiology and physical examination rank near the bottom. Figure 2 (right) further quantifies this agreement using Kendall's τ between model pairs. Across the 28 pairs, the mean τ is 0.91 and the minimum is 0.71, while six pairs produce identical rankings with $\tau = 1$. Domain primarily adaptation changes the strength of these preferences more than their ordering.

Observation 1 (Systematic Medical Evidence Preference)

LLMs exhibit systematic preferences among medical evidence sources, with highly consistent rankings across general-purpose and domain-adapted models.

Preference Instability under Presentation Changes. Observation 1 establishes systematic source preferences, but whether they reflect evidence alone remains unclear. If so, changing only the presentation of the same evidence should leave diagnostic preference unchanged. To test this, we construct paired counterfactual inputs that keep evidence statements and source identities fixed while re-rendering one or both records into sectioned text, key-value fields, tables, or JSON, yielding 13 presentation states per conflict. We then measure the resulting format-induced preference shift.

Presentation changes alone can substantially alter diagnostic preference despite unchanged evidence content. Figure 1 (right) illustrates this behavior for MedGemma-4B, where presentations can shift preference over a wide range. A format range crossing zero indicates a reversal in the favored diagnosis. The same pattern holds across all eight LLMs, as shown in Appendix Figure 8. Depending on the model, 9.8% to 17.1% of conflicts have at least one presentation that reverses the preference.

Format Effects Vary across Models and Evidence Sources. We next ask whether presentation changes affect medical evidence consistently across settings, to determine whether a universal format effect exists. For each format, we compare source-specific preference shifts across models and evidence sources, measuring how often these effects differ significantly or have opposite directions. Source-specific shifts use the same content-preserving comparison as in Appendix C.2.

Format effects vary substantially across both models and evidence sources. Figure 3 (a,b) shows that different effects are common across models for the same source and across sources within the same model. Figure 3 (c,d) further shows that these effects can reverse direction, with the same format increasing evidence influence in one setting while decreasing it in another. Thus, presentation does not introduce a single fixed bias, making a universal format-level correction unreliable.

Observation 2 (Format-Induced Instability)

Evidence-source preference is not determined by evidence content alone, as content-preserving format changes can substantially shift or even reverse diagnostic preference, with effects varying across models and evidence sources.

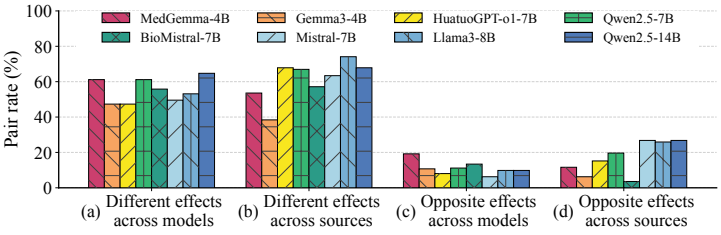

Figure 3: Observed rates of pairwise differences in format effects across models and evidence sources. Panels (a,b) show significantly different effects, while panels (c,d) show effects with opposite signs.

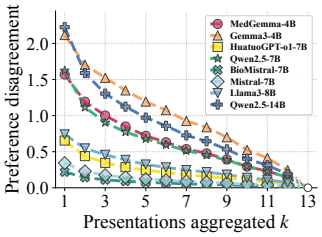

Figure 4: Preference disagreement consistently decreases with k across all eight LLMs.

These findings raise a key question: what judgment should a format-robust model preserve? Using one presentation as reference may retain format-specific variation, while fixed correction cannot account for heterogeneous effects. We therefore seek a stable, instance-specific reference without privileging any presentation. For each conflict, we construct multiple content-preserving presentations that keep evidence, source identities, and candidate diagnoses fixed. Although individual predictions may vary, they describe the same underlying evidence, so cross-format aggregation provides an instance-specific reference that reduces dependence on any single presentation.

Aggregating more presentations makes this reference increasingly stable across different settings. For each conflict, we compare pooled predictions from two randomly sampled subsets of k presentations and measure the absolute difference between their diagnostic preference scores. As Figure 4 shows, disagreement decreases steadily with k across all eight LLMs. These results show that cross-format aggregation becomes progressively less sensitive to which presentations are selected.

Insight (Cross-Format Consensus)

Aggregating predictions across content-preserving presentations yields an instance-specific reference that becomes more stable as more presentations are included.

4 FORMEDER

FORMEDER aims to remove presentation-specific variation from diagnostic judgments, without deciding which evidence source is clinically correct. This requires two properties. *Invariance* requires stable predictions across presentations, and *preservation* retains the presentation-invariant component of the original judgment. Each alone is trivial, since ignoring the evidence gives perfect invariance and the original model preserves its own judgments. Satisfying both therefore requires a reference judgment shared across presentations of the same evidence. It cannot come from a gold label, which a conflict lacks, or from one presentation, which may carry its own format effect.

FORMEDER derives this reference directly from the model itself through three stages (Figure 5). ❶ **Counterfactual Evidence View Modeling** (Section 4.1) defines the variation to ignore by re-rendering a conflict into counterfactual views of the evidence. ❷ **Cross-Format Consensus Inference** (Section 4.2) estimates the judgment to keep by averaging the predictions of the frozen model θ_0 over these views into a consensus that favors no single presentation. ❸ **Consensus-Preserving Robust Alignment** (Section 4.3) trains the model toward this consensus under every view for preservation and contracts the residual preference gaps between views for invariance. The framework needs no external annotation, and the trained model reads a single presentation at inference.

4.1 COUNTERFACTUAL EVIDENCE VIEW MODELING

The first stage isolates presentation variation. For each evidence conflict, we independently re-render each record while keeping the evidence statements, source identities, candidate diagnoses, and record and option positions fixed. This isolates presentation as the only changing factor.

Concretely, we sample a non-prose format $f \in \mathcal{F}$ and render each record either in prose P or in f, which gives the neighborhood

$$\mathcal{N}_f = \{(P, P), (f, P), (P, f), (f, f)\}, \tag{5}$$

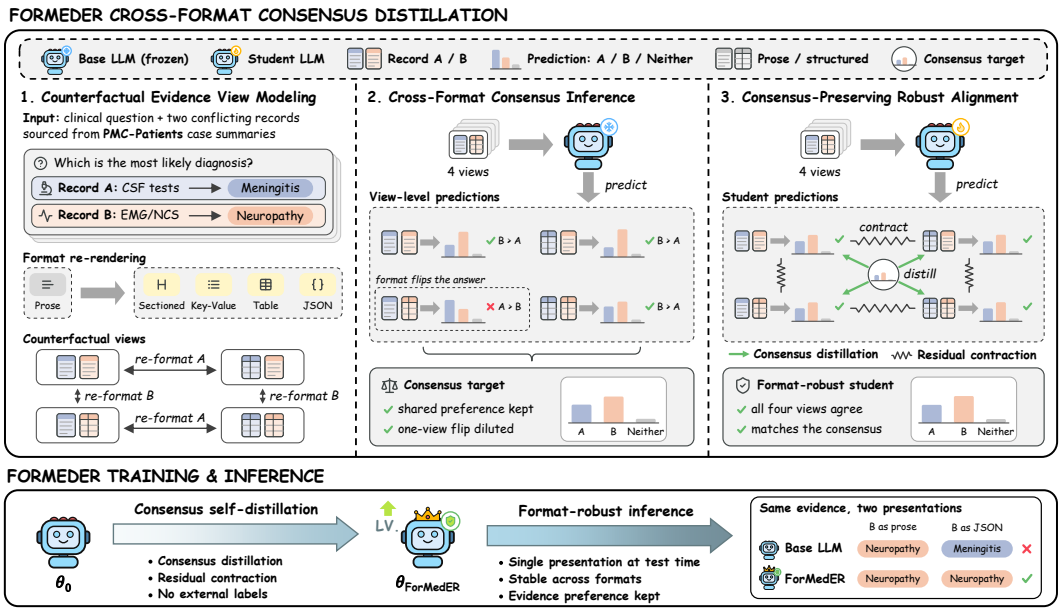

Figure 5: **Overview of FORMEDER.** Frozen-model predictions on four counterfactual views form a consensus target, and the student is trained with consensus distillation and residual contraction.

where the two coordinates give the formats of E_A and E_B. We write x^v for the input of view $v \in \mathcal{N}_f$. For notational simplicity, we omit the fixed instance i and position configuration o until Equation 10. The four views form a 2×2 grid whose edges each change the format of only one record, so the preference change along an edge is exactly a format-induced preference shift.

4.2 CROSS-FORMAT CONSENSUS INFERENCE

The second stage estimates the judgment to preserve. Since the four views contain the same evidence but may yield different predictions, we use their shared judgment rather than any single, format-specific prediction as the reference. This avoids privileging any particular presentation.

Each frozen prediction $p_{\theta_0}(\cdot \mid x^v)$ is a distribution over D_A, D_B, D_{N}, and the remaining probability mass, indexed by diagnosis identity rather than by option letter. The consensus target is the average of the four frozen predictions,

$$q_f = \frac{1}{|\mathcal{N}_f|} \sum_{v \in \mathcal{N}_f} p_{\theta_0}(\cdot \mid x^v). \tag{6}$$

Because all views receive equal weight, a preference that the views share survives in q_f, clinically justified or not, while a deviation that appears in only one view is diluted. Moreover, q_f is the unique distribution q that minimizes $\sum_{v \in \mathcal{N}_f} D_{\mathrm{KL}}(p_{\theta_0}(\cdot \mid x^v) \,\|\, q)$, and its preference score always lies between the smallest and the largest score of the four views (Proposition D.1, proved in Appendix D).

4.3 CONSENSUS-PRESERVING ROBUST ALIGNMENT

The third stage aligns every view with the consensus. The student p_θ is initialized from the frozen teacher θ_0, making FORMEDER a form of self-distillation. We use a view-level distillation loss for preservation and an edge-level contraction loss for invariance.

Consensus Distillation on Views. We distill the consensus into every view,

$$\mathcal{L}_{\mathrm{cons}}(f) = \frac{1}{|\mathcal{N}_f|} \sum_{v \in \mathcal{N}_f} D_{\mathrm{KL}}\left(q_f \,\|\, p_\theta(\cdot \mid x^v)\right). \tag{7}$$

We supervise each view separately to prevent opposite deviations from canceling in the average. Since q_f is conflict-specific, the loss preserves instance-dependent judgments rather than pulling all conflicts toward a common prediction, thereby retaining variation across conflicts.

Residual Contraction on Edges. Consensus distillation does not directly constrain pairwise preference shifts across views. We therefore penalize these shifts along counterfactual edges. The neighborhood has four edges, each connecting two views that differ only in the format of one record,

$$\mathcal{E}_f = \{((P,P),(f,P)), ((P,P),(P,f)), ((f,P),(f,f)), ((P,f),(f,f))\}, \tag{8}$$

and the contraction loss is the mean preference gap over these edges,

$$\mathcal{L}_{\mathrm{ctr}}(f) = \frac{1}{|\mathcal{E}_f|} \sum_{(u,v)\in\mathcal{E}_f} |M_\theta(x^u) - M_\theta(x^v)|\,. \tag{9}$$

The gap along each edge is exactly a format-induced preference shift, so this term acts on the same quantity that Section 2 measures.

Joint Objective. Restoring the instance and position indices, we combine the two terms as

$$\mathcal{L}_{\textsc{ForMedER}} = \mathbb{E}_{i,o,f}\left[\mathcal{L}_{\mathrm{cons}}(i,o,f) + \lambda \mathcal{L}_{\mathrm{ctr}}(i,o,f)\right], \tag{10}$$

where λ balances preservation and invariance. Their complementarity is formalized in the following proposition, with proof in Appendix D.

Proposition 1 (Complementary Guarantees). *Fix a neighborhood $\mathcal{N}_f$ with consensus target q_f, and write $M_v = M_\theta(x^v)$ for the preference score of the student on view v. Let $s_f = q_f(D_A) + q_f(D_B)$ and $\pi_f = q_f(D_A)/s_f$. The first bound below holds for every student, and the second holds for every flattened student, that is, one with $M_v = 0$ on all four views,*

$$\begin{aligned} \underbrace{\max_{u,v\in\mathcal{N}_f} |M_u - M_v|}_{\textit{largest preference gap}} &\le 2 \quad \underbrace{\mathcal{L}_{\mathrm{ctr}}(f)}_{\textit{contraction loss}}\,, \\ \underbrace{\mathcal{L}_{\mathrm{cons}}(f)}_{\textit{consensus loss}} &\ge \quad \underbrace{s_f\big(\ln 2 - H_b(\pi_f)\big)}_{\textit{strength of the consensus preference}}\,, \end{aligned} \tag{11}$$

where $H_b(\pi) = -\pi\ln\pi - (1-\pi)\ln(1-\pi)$ is the binary entropy. The second bound is positive whenever q_f prefers one diagnosis. Neither guarantee transfers to the other term, since the consensus loss can be arbitrarily small while two views differ by an arbitrarily large preference gap, and the contraction loss is zero for every flattened student.

Thus, contraction bounds cross-view preference variation, while distillation prevents invariant but uninformative solutions. Section 5.3 verifies both effects empirically.

5 Experiments

Benchmark. We train on 3,928 instances from the train set of PMC-ConForm, select all settings on the 1,322 validation instances for model selection, and report the main results on the 1,322 test instances. The three splits share no patients with each other or with the analysis set, and details are provided in Appendix B.2. Each instance is rendered in prose and four structured formats f (sectioned, key-value, table, and JSON) under four position configurations o. Four additional formats (YAML, bullet lists, INI, and outline) are used only to evaluate generalization to unseen formats.

Models and Baselines. We evaluate eight LLMs, three domain-adapted models with their base models, MedGemma-4B (Sellergren et al., 2025) with Gemma3-4B (Gemma Team, 2025), HuatuoGPT-o1-7B (Chen et al., 2024) with Qwen2.5-7B (Yang et al., 2024), and BioMistral-7B (Labrak et al., 2024) with Mistral-7B (Jiang et al., 2023), and Llama3-8B (Grattafiori et al., 2024) and Qwen2.5-14B (Yang et al., 2024). We compare ForMedER with the frozen model and three *target-only* baselines, which share its teacher, adapter, and update budget but have no contraction term. *Per-rendering KD* (Hinton et al., 2015) and *Prose-reference KD* distill each view toward the frozen prediction on that view and on the all-prose view, respectively, and *Format-augmented SFT* trains every view on the frozen all-prose answer. Detailed baseline configurations are provided in Appendix E.1.

Metrics. We evaluate models from three perspectives. For each format edge $(u,v) \in \mathcal{E}_f$ in Equation 8, where the two inputs differ only in the presentation format of one record, and each single-record question x_A^f, we compute three metrics:

$$\mathrm{FPS} = \mathbb{E}\,|M(x^u) - M(x^v)|, \;\; \mathrm{FCC} = \Pr[\widehat{D}(x^u) = \widehat{D}(x^v)], \;\; \mathrm{RA} = \Pr[\widehat{D}(x_A^f) = D_A\,\forall f], \tag{12}$$

Table 1: Main results on the PMC-ConForm test set, with the best and second-best in **red** and **blue**.

Metrics	Methods	MedGemma 4B	Gemma3 4B	HuatuoGPT o1-7B	Qwen2.5 7B	BioMistral 7B	Mistral 7B	Llama3 8B	Qwen2.5 14B
FPS (nats) ↓	Frozen	1.7131	2.9976	0.8178	2.2400	0.2656	0.3867	0.7883	3.6525
	Per-rendering KD	1.7131	2.9976	0.8178	2.2400	0.2656	0.3867	0.7883	3.6525
	Prose-reference KD	0.4399	**0.3291**	**0.4050**	**0.5858**	**0.0729**	**0.1239**	**0.2946**	**0.6055**
	Format-augmented SFT	**0.4320**	0.3479	0.5024	0.6498	0.1825	0.1997	0.4053	0.6392
	FORMEDER (Ours)	**0.0767**	**0.0769**	**0.0336**	**0.0736**	**0.0248**	**0.0312**	**0.0602**	**0.1206**
FCC (%) ↑	Frozen	87.66	89.12	81.83	89.16	89.08	92.84	88.50	89.46
	Per-rendering KD	87.66	89.12	81.83	89.16	89.08	92.84	88.50	89.46
	Prose-reference KD	**92.03**	**93.03**	84.99	**92.82**	95.14	96.11	94.05	**92.99**
	Format-augmented SFT	91.84	92.94	**89.31**	92.29	**96.53**	**96.77**	**94.96**	92.66
	FORMEDER (Ours)	**96.11**	**96.43**	**93.86**	**97.41**	**95.59**	**96.95**	**97.43**	**96.33**
RA (%) ↑	Frozen	53.49	51.36	**40.62**	39.79	42.93	66.06	27.46	41.26
	Per-rendering KD	53.49	51.36	**40.62**	39.79	42.93	66.06	27.46	41.26
	Prose-reference KD	56.06	**52.80**	39.79	38.82	48.84	**70.53**	28.45	**41.35**
	Format-augmented SFT	**56.19**	51.42	39.08	**43.86**	**60.40**	**69.68**	**28.91**	40.37
	FORMEDER (Ours)	**56.68**	**53.77**	**43.33**	**43.70**	**61.66**	67.18	**29.60**	**42.47**

FPS (**Format Preference Shift**) measures the preference shift caused by format changes, with lower values indicating stronger format robustness. FCC (**Format Choice Consistency**) measures prediction consistency across formats, with higher values indicating more stable choices. RA (**Robust Accuracy**) measures whether the model preserves evidence-specific diagnostic behavior without competing records. FPS and FCC are averaged over instances, position configurations, formats, and edges to capture overall format sensitivity across different rendering conditions.

Implementation Details. Every method trains one rank-64 LoRA adapter (Hu et al., 2022) per backbone for 375 updates, with a learning rate of 5×10^{-5} and an effective batch size of 16. We select $\lambda = 4$ on MedGemma-4B and apply it unchanged to all eight backbones. Further training configurations and evaluation protocols are provided in Appendix E.2 and Appendix F.1.

5.1 MAIN RESULTS

Table 1 shows that FORMEDER consistently improves format robustness across all eight backbones. It achieves the lowest FPS on every model, reducing format-induced preference shifts by 10.7× to 39.0× compared with the frozen model and by 2.9× to 12.1× compared with the best target-only baseline. Unlike target-only baselines that only align predictions toward a fixed target, FORMEDER more effectively reduces residual disagreement across counterfactual presentations.

The reduced preference shifts lead to more stable diagnostic choices, with FORMEDER achieving the highest FCC on seven of eight backbones and improving over the frozen model by 4.11 to 12.03 points. Meanwhile, FORMEDER preserves evidence-specific behavior, achieving the best RA on six backbones and improving over the frozen model on all eight backbones, including the only improvement on HuatuoGPT-o1-7B among all methods. These results show that FORMEDER can simultaneously reduce presentation-induced variation and preserve the model's diagnostic judgment.

5.2 GENERALIZATION TO UNSEEN FORMATS

To test whether the robustness learned by FORMEDER transfers to unseen formats, we train FORMEDER with two to five formats, progressively adding key-value, table, and JSON to prose and sectioned text, and evaluate it on unseen formats (YAML, bullet lists, INI, and outline) on the test set. Figure 6 reports FPS and FCC averaged over unseen formats, together with RA measuring the fraction of single-record questions answered correctly under prose and four unseen formats.

FORMEDER generalizes beyond the observed formats, reducing unseen-format FPS relative to the frozen model on all eight backbones when trained on all five formats, demonstrating that learned invariance transfers across presentations. Detailed comparisons with the frozen model on unseen formats are provided in Appendix G.1. Moreover, increasing the number of training formats improves this generalization, with lower FPS and higher FCC on unseen formats across backbones. RA also remains stable or improves as more formats are introduced, indicating that broader format exposure enhances robustness without compromising evidence-specific diagnostic behavior.

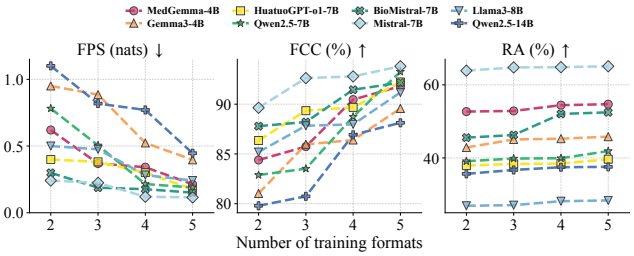

Figure 6: Generalization of FORMEDER to four unseen formats (YAML, bullet lists, INI, and outline).

Table 2: Ablation of consensus targets and residual contraction.

Variants	FPS (nats) ↓	FCC (%) ↑	RA (%) ↑
Frozen	1.7131	87.66	53.49
Prose target	0.1532	95.39	56.46
Random-format target	0.2189	93.97	54.76
Majority-vote target	0.1405	95.88	55.19
w/o residual contraction	0.4088	91.01	54.15
w/o consensus distillation	✗	✗	✗
FORMEDER (Ours)	0.0767	96.11	56.68

5.3 Ablation Study

Effect of the Consensus Target. We study whether the consensus target provides a better reference than simpler alternatives. Keeping contraction unchanged, we replace the consensus target with three alternatives: the frozen prediction on the all-prose view (*prose target*), the frozen prediction on the randomly sampled structured-format view (*random-format target*), and the majority answer among the four frozen views as a one-hot target (*majority-vote target*). As shown in Table 2, all three alternatives improve over the frozen model, showing that each, combined with contraction, reduces format sensitivity. Compared with these alternatives, FORMEDER achieves the best results across all metrics, with the lowest FPS (0.0767), highest FCC (96.11%), and highest RA (56.68%). This indicates that aggregating predictions across presentations provides a more reliable target, as it avoids relying on a single presentation while preserving shared evidence-specific behavior.

Effect of Residual Contraction. We further remove residual contraction to evaluate its contribution beyond consensus distillation. Without residual contraction, FPS increases from 0.0767 to 0.4088 and FCC decreases from 96.11% to 91.01%, showing that consensus distillation alone cannot eliminate format-induced preference shifts. Meanwhile, removing consensus distillation leads to output degeneration, where the model reduces cross-format differences at the cost of evidence-specific diagnostic behavior. These results show that consensus distillation preserves desired behavior, while residual contraction directly reduces residual format sensitivity across counterfactual views.

6 Related Work

Presentation effects on evidence preference. Predictions depend on prompt format (Sclar et al., 2024; He et al., 2024), medical answer format (Taveekitworachai et al., 2026), and table versus chart evidence (Ho et al., 2026). Under conflicting evidence, preferences vary with source (Schuster et al., 2026), evidence format (Liu et al., 2026a), and medical-document order (Han et al., 2026). We fix the content of two conflicting clinical records, the question, and the answer format, then vary their textual layout and measure paired shifts in diagnostic preference without gold conflict labels.

Learning invariant predictions. Multi-format supervision improves question-format robustness (Liu et al., 2026b), while Format-Adapter selects reasoning formats for answer aggregation (Wang et al., 2025). Consistency training aligns perturbed predictions (Xie et al., 2020), and distillation preserves teacher outputs (Hinton et al., 2015; Liusie et al., 2024). FORMEDER distills the frozen consensus and contracts residual shifts for inference from one rendering; Appendix A compares their training objectives, treatment of formats, and inference requirements.

7 Conclusion

In this paper, we investigate how the presentation format of medical evidence affects LLMs' evidence-source preferences. Through PMC-ConForm, we reveal that LLMs exhibit systematic source preferences that can shift or even reverse under content-preserving format changes. To mitigate this instability, we propose FORMEDER, a self-distillation framework that aligns each presentation with the model's cross-format consensus while contracting residual preference gaps. Extensive experiments on eight LLMs demonstrate that FORMEDER substantially improves format robustness while preserving presentation-invariant judgments, and generalizes to unseen formats.

AI USE STATEMENT

Generative AI assisted manuscript revision, exploration of related work, and parts of data construction and analysis. Generative models were also used as the base models in our experiments.

ETHICS STATEMENT

We study constructed conflicts from clinical case reports. Since paired records may come from different patients, they do not define a joint clinical gold diagnosis, and model preferences should not be interpreted as clinical advice.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the code, data, and analysis resources necessary to reproduce our experiments.

REFERENCES

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024.

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. Bias-augmented consistency training reduces biased reasoning in chain-of-thought, 2024. URL https://arxiv.org/abs/2403.05518.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1607–1616, 2018.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019. doi: 10.1145/3306618.3317950.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Abinitha Gourabathina, Walter Gerych, Eileen Pan, and Marzyeh Ghassemi. The medium is the message: How non-clinical information shapes clinical decisions in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1805–1828, 2025. doi: 10.1145/3715275.3732121.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. doi: 10.1038/s41591-024-03097-1. URL https://doi.org/10.1038/s41591-024-03097-1.

Yikun Han, Mengfei Lan, and Halil Kilicoglu. When evidence conflicts: Uncertainty and order effects in retrieval-augmented biomedical question answering. In *BioNLP 2026*, pp. 630–643, 2026. doi: 10.18653/v1/2026.bionlp-1.50.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on LLM performance? *arXiv preprint arXiv:2411.10541*, 2024.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 5549–5581, 2023.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *npj Digital Medicine*, 9(1), 2026. doi: 10.1038/s41746-026-02915-9.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Format matters: The robustness of multimodal LLMs in reviewing evidence from tables and charts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31014–31022, 2026. doi: 10.1609/aaai.v40i37.40361. URL https://ojs.aaai.org/index.php/AAAI/article/view/40361.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34279–34303, 2025. doi: 10.18653/v1/2025.emnlp-main.1738.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Jiwoo Lee, Yousun Chung, Dae-Hyun Ko, Eun Young Song, and Hyungsuk Kim. Interpreting antibody identification tests with large language models: Critical role of tabular data formatting. *Transfusion*, 2026. doi: 10.1111/trf.70364.

Jiacheng Liu, Mayi Xu, Qiankun Pi, Wenli Li, Ming Zhong, Yuanyuan Zhu, Mengchi Liu, and Tieyun Qian. Format as a prior: Quantifying and analyzing bias in LLMs for heterogeneous data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32060–32068, 2026a. doi: 10.1609/aaai.v40i38.40477.

June M. Liu, Shaomian Zheng, He Cao, Dingnan Jin, Qing Cui, and Jun Zhou. Improving cross-format robustness in language models with multi-format training. *arXiv preprint arXiv:2606.11643*, 2026b. URL https://arxiv.org/abs/2606.11643.

Adian Liusie, Yassir Fathullah, and Mark Gales. Teacher-student training for debiasing: General permutation debiasing for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1376–1387. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.81. URL https://aclanthology.org/2024.findings-acl.81/.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/tpami.2018.2858821.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in Hilbert space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3351–3361, 2020.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 529–537, 2025. doi: 10.18653/v1/2025.naacl-srw.51.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems, 2023. URL https://arxiv.org/abs/2303.13375.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19844–19863, 2024. doi: 10.18653/v1/2024.emnlp-main.1109.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, 2026. doi: 10.18653/v1/2026.acl-long.1357.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets LLM: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024. doi: 10.1145/3616855.3635752. URL https://doi.org/10.1145/3616855.3635752.

Pittawat Taveekitworachai, Natpatchara Pongjirapat, Krittaphas Chaisutyakorn, Piyalitt Ittichaiwong, Tossaporn Saengja, and Kunat Pipatanakul. On the robustness of answer formats in medical reasoning models. *arXiv preprint arXiv:2509.20866*, 2026. URL https://arxiv.org/abs/2509.20866.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8. URL https://doi.org/10.1038/s41591-023-02448-8.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6287–6310, 2024. doi: 10.18653/v1/2024.findings-acl.375.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7468–7484, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

Dingzirui Wang, Xuanliang Zhang, Rongyu Cao, Longxu Dou, Xianzhen Luo, Yingwei Ma, Qingfu Zhu, Wanxiang Che, Binhua Li, Fei Huang, and Yongbin Li. Format-Adapter: Improving reasoning capability of LLMs by adapting suitable format. *arXiv preprint arXiv:2506.23133*, 2025. URL https://arxiv.org/abs/2506.23133.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. URL https://arxiv.org/abs/2404.10198.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/1904.12848.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. EHRStruct: A comprehensive benchmark framework for evaluating large language models on structured electronic health record tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34340–34348, 2026. doi: 10.1609/aaai.v40i40.40731.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1): 909, 2023. doi: 10.1038/s41597-023-02814-8.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12697–12706, 2021.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2613–2626, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.192. URL https://aclanthology.org/2022.findings-emnlp.192/.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, 2024.

APPENDIX

Contents

A Additional Related Work **14**

B Benchmark and Data **15**
- B.1 Analysis Set 15
- B.2 Train, Validation, and Test Sets 15
- B.3 Content-Preserving Renderers 16
- B.4 Models and Answer Scoring 17

C Source Preference and Format Sensitivity **18**
- C.1 Source Preference and Latent Scores 18
- C.2 Format Sensitivity 19

D Proofs **20**

E Implementation Details of FORMEDER **22**
- E.1 Baselines 22
- E.2 Training Configuration 22
- E.3 Ablation Targets 22

F Evaluation Protocol **23**
- F.1 Metrics and Pairing 23
- F.2 FPS Against Rescaled Frozen Scores 23

G Additional Results **23**
- G.1 Number of Training Formats 23
- G.2 Preservation Analyses 24
- G.3 Case Study 24

H Limitations **24**

Terminology. A *conflict*, or *instance*, pairs two records supporting different diagnoses. A *format* is a layout type; a *rendering*, $\mathcal{R}_f(E)$, is one record in that format. A *presentation* (or *presentation state*) specifies both records' formats (f, g); a *view* belongs to a training neighborhood $\mathcal{N}_f$. A *position configuration* fixes record and option order. FORMEDER combines *consensus distillation* ($\mathcal{L}_{\text{cons}}$) and *residual contraction* ($\mathcal{L}_{\text{ctr}}$) to preserve the *presentation-invariant judgment*, the component independent of presentation. *Evidence-specific diagnostic behavior* means choosing a record's supported diagnosis without competing evidence; RA measures correctness across formats.

A ADDITIONAL RELATED WORK

Source credibility and evidence preference. LLMs favor query relevance over human-valued stylistic features (Wan et al., 2024); credibility-aware generation and source-reliability estimation incorporate context credibility (Pan et al., 2024; Hwang et al., 2025). Models prefer institutional to individual or social-media sources, though repetition can reverse this preference (Schuster et al., 2026). When conflicting claims are rewritten into different data formats, models favor text and knowledge graphs over infoboxes and tables (Liu et al., 2026a), and reversing the order of two contradictory medical documents flips 11.4–25.2% of predictions (Han et al., 2026). PMC-ConForm isolates format effects by varying layout with evidence and source identities fixed.

Sensitivity to presentation. Prompt formatting (Sclar et al., 2024; Voronov et al., 2024) and serialization choices such as Markdown, JSON, or YAML (He et al., 2024) affect accuracy; mixed-style examples can improve robustness (Ngweta et al., 2025). Table serialization changes structural understanding (Sui et al., 2024). Ho et al. (2026) compare scientific claim verification from tables and charts representing the same evidence, finding stronger performance with tables. Medical predictions also depend on serialization of tabular features (Hegselmann et al., 2023), EHR events (Hegselmann et al., 2026; Yang et al., 2026), laboratory tables (Lee et al., 2026), and patient messages (Gourabathina et al., 2025). Taveekitworachai et al. (2026) study multiple-choice, open-ended, and ranked-list answers to medical questions; supervised fine-tuning produces more stable

behavior across these formats than reinforcement fine-tuning in their controlled comparisons. Our setting holds the answer format fixed and varies the textual layout of conflicting clinical records.

Training and selecting formats. Liu et al. (2026b) train on equivalent multiple-choice, true/false, fill-in-the-blank, and open-ended questions. Their FormatMix method expands a subset of items, selected randomly or by format sensitivity, into multiple question formats. Our Format-augmented SFT baseline applies hard-label supervision across evidence renderings, using the frozen prose answer as its target (Appendix E.1). Format-Adapter generates and selects reasoning formats, then aggregates answers from the selected formats (Wang et al., 2025). FORMEDER learns stable evidence preferences across layouts and processes one supplied rendering with a single adapted model at inference, avoiding format search and multiple answer generations for each input.

Consistency, invariance, and distillation. Consistency regularization aligns predictions under perturbations (Miyato et al., 2019; Xie et al., 2020), including alternative or biased prompts (Zhou et al., 2022; Chua et al., 2024); counterfactual augmentation and logit pairing target prediction invariance (Kaushik et al., 2020; Garg et al., 2019; Veitch et al., 2021). An invariance term alone admits collapse, guarded against here by a conflict-specific consensus target (Proposition 1). Distillation transfers teacher outputs (Hinton et al., 2015), including between identical architectures (Furlanello et al., 2018; Mobahi et al., 2020). The closest method distills a permutation-debiased teacher into a single-pass student (Liusie et al., 2024); FORMEDER instead pools counterfactual renderings and contracts residual preference gaps. Inference-time calibration uses additional predictions (Zhao et al., 2021; Zhou et al., 2024), whereas FORMEDER requires one forward pass per input.

B BENCHMARK AND DATA

B.1 ANALYSIS SET

PMC-ConForm comprises analysis data for Section 3 and train, validation, and test sets for Section 5 (Table 3). Records come from PMC-Patients summaries of PubMed Central case reports (Zhao et al., 2023). Its 1,310 analysis conflicts pair records from two patients with different diagnoses and no shared clinical gold diagnosis. Source labels and diagnoses are checked against the original case narrative, distinguishing patient genetic testing from pathogen typing. The eight categories of evidence source are pathology (Path), microbiology (Micro), genetics (Gen), imaging, laboratory tests (Lab), physiology (Phys), physical examination (Exam), and patient report (Patient).

Table 3: Dataset sizes of the four parts of PMC-ConForm.

Parts	Used in	Conflicts	Pair groups	Patients
Analysis set	Sec. 3	1,310	655	948
Train	Sec. 5	3,948	1,974	2,035
Validation	Sec. 5	1,322	661	971
Test	Sec. 5	1,322	661	966

Patients are matched by affected organ system and age group (pediatric, adult, or older adult), then screened for lexical similarity. Each pair group fixes two patients and two sources; exchanging which patient supplies each source yields two conflicts, each with one record per patient. A fixed reference model (Qwen3-32B) must select each record's diagnosis in isolation under both option orders; multi-option replies are rejected. We retain 655 groups spanning 948 unique patients, allowing patients to enter multiple groups. These yield 1,310 conflicts and jointly cover all 28 source pairs, with 14–94 conflicts each. No analysis patient or article overlaps the other splits (Appendix B.2). Age and sex are harmonized; diagnostic findings and sentence order stay fixed across presentations.

B.2 TRAIN, VALIDATION, AND TEST SETS

Train, validation, and test instances follow the analysis set's organ-system and age matching, lexical screening, and record-level verification, always pairing two different sources. Splitting connected components of patients who share records prevents patient overlap. Validation and test contain 1,322 instances in 465 and 464 components, respectively. Seven articles cross splits: two between train

and validation, three between train and test, and two between validation and test. Excluding 20 training instances from these articles leaves 3,928 with no article overlap with validation or test.

B.3 Content-Preserving Renderers

The input of every renderer is the ordered list of evidence sentences $(t_1, \ldots, t_K)$ of a record together with its source annotation. The renderers use this list without regenerating, summarizing, reordering, or rounding its content (Table 4). For the structured formats, a deterministic rule assigns each sentence a section label s_i from its source modality and lexical matches, for example *Technique*, *Comparison*, and *Findings* for imaging. Section indices never decrease along a record, so adding headings never moves a sentence relative to its neighbors, and repeated section names receive numbered keys k_i, shared by the key-value and table formats for the same record.

Table 4: Templates of the five content-preserving formats.

Formats	Rendering rule
Prose	Join $t_1, \ldots, t_K$ with single spaces.
Sectioned	Emit an uppercase `SECTION:` heading when s_i changes, followed by each t_i on its own line.
Key-value	Emit one `key: sentence` line per (k_i, t_i), in the original order.
Table	Use the Markdown header `| Field | Finding |`, separator `|---|---|`, and one `| key | sentence |` row per (k_i, t_i).
JSON	Place ordered `section`/`text` objects in a neutral `record` array, as shown below.

JSON uses the wrapper below with two sentence placeholders, one-space indentation, and unchanged non-ASCII characters. Each object pairs a section label with its original sentence.

JSON rendering template

```
{
 "record": [
  {"section": "<section_1>", "text": "<sentence_1>"},
  {"section": "<section_2>", "text": "<sentence_2>"}
 ]
}
```

Automatic content checks. We require exact recovery of the ordered evidence sentences, including numbers, units, and negations. These known templates can in principle be canonicalized to prose; an unknown format requires a suitable inverse based on its renderer and grammar.

Further formats. YAML, bullet-list, INI, and outline are reserved before training for evaluation. These deterministic templates preserve the evidence sentences and their order.

Unseen-format templates. The examples show two consecutive sentences in one section; each new section starts a block without reordering sentences. Placeholders use section labels and verbatim evidence sentences used by the five training renderers, preserving sentence order.

Unseen rendering templates

YAML

```
report:
  - section: <section>
    findings:
      - "<sentence_1>"
      - "<sentence_2>"
```

INI

```
[<section>]
item_1 = <sentence_1>
item_2 = <sentence_2>
```

Bullet list

```
**<Section>**
- <sentence_1>
- <sentence_2>
```

Outline

```
<section>
    <sentence_1>
    <sentence_2>
```

B.4 Models and Answer Scoring

Conflict prompts show two records, their diagnoses, and *Neither of the above*, without source-label lines. Preference scores use option-letter probabilities at the first answer position. Validation fixes the scored spellings: each letter occupies one token, and the three token IDs are distinct. BioMistral-7B and Mistral-7B use leading-space letters; the other backbones use bare letters. For FCC and RA, we parse generated answers and map option letters to diagnosis identities (Appendix F.1).

Checkpoint identifiers. Table 5 lists the exact repositories used for all eight backbones, including the specific Unsloth and NousResearch checkpoint distributions used for three of the eight models.

Table 5: Model checkpoint identifiers.

Backbone	Hugging Face checkpoint
MedGemma-4B	`unsloth/medgemma-4b-it`
Gemma3-4B	`unsloth/gemma-3-4b-it`
HuatuoGPT-o1-7B	`FreedomIntelligence/HuatuoGPT-o1-7B`
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`
BioMistral-7B	`BioMistral/BioMistral-7B`
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.1`
Llama3-8B	`NousResearch/Meta-Llama-3-8B-Instruct`
Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`
Reference model	`Qwen/Qwen3-32B`

Complete user prompts. Each checkpoint's chat template wraps the user messages below. Slots contain the rendered records and competing diagnosis names, without source-label lines. Record order and A/B order vary independently; C always denotes neither diagnosis. Single-record questions retain both competing diagnoses and *Neither* for each individual record shown alone.

Conflict question

```
Two records concerning one patient are shown below.  They do not
agree.
<rendered record 1>
<rendered record 2>
Which of the following is the most likely diagnosis for this patient?
A. <diagnosis A>
B. <diagnosis B>
C. Neither of the above
Read the records and answer with a single letter.  Do not explain your
reasoning.
```

Single-record question

```
One record concerning a patient is shown below.
<rendered record>
Which of the following is the most likely diagnosis for this patient?
A. <diagnosis A>
B. <diagnosis B>
C. Neither of the above
Read the record and answer with a single letter.  Do not explain your
reasoning.
```

Changing one or both records from prose to a structured format gives 13 presentation states:

$$\mathcal{V}_{13} = \{(P,P)\} \cup \bigcup_{f \in \mathcal{F} \setminus \{P\}} \{(f,P),(P,f),(f,f)\}, \tag{13}$$

where P denotes prose and $\mathcal{F} \setminus \{P\}$ contains the four structured formats. Each state is evaluated under four combinations of record order and option order, giving 52 configurations per conflict. All states of a model share the same all-prose reference when computing paired differences.

C Source Preference and Format Sensitivity

C.1 Source Preference and Latent Scores

Pairwise preference. For sources (X, Y), M_{XY} measures preference toward the diagnosis supported by X when both records are in prose. We average over the four position configurations, then over conflicts of the pair. Figure 7 reports all 28 pairs for eight models; positive values favor the first source. Colors span each model's range, with both extremes in bold. Pointwise 95% bootstrap intervals (2,000 draws, seed 20260915) exclude zero in 156 of 216 cells with at least 20 conflicts under conflict resampling, and in 86 of 112 cells with at least 20 groups under pair-group resampling.

	MedGemma-4B	Gemma3-4B	HuatuoGPT-o1-7B	Qwen2.5-7B	BioMistral-7B	Mistral-7B	Llama3-8B	Qwen2.5-14B
Path / Micro	+0.17	-0.69	+0.36	+0.51	+0.01	+0.71	+1.21	+1.93
Path / Gen	+1.41	+3.53	+1.36	+3.62	+0.19	+1.25	+1.25	+2.29
Path / Imaging	+5.79	+12.05	+2.42	+10.06	+1.08	+2.15	+5.54	+17.67
Path / Lab	+3.42	+6.35	+0.61	+5.76	+0.67	+1.13	+3.45	+10.14
Path / Phys	**+7.30**	+12.85	+2.90	+12.70	+1.13	+2.27	+5.26	+20.62
Path / Exam	+7.19	+14.59	+2.34	+12.14	+1.32	+2.97	**+5.80**	+21.65
Path / Patient	+5.52	+12.36	+2.23	+10.85	+1.28	+2.09	+5.20	+17.54
Micro / Gen	**-3.74**	-3.12	**-1.88**	**-5.25**	**-0.46**	-0.24	**-2.09**	**-6.68**
Micro / Imaging	+6.25	+12.75	+1.65	+9.40	+1.48	+1.82	+4.20	+14.95
Micro / Lab	+3.42	+8.81	+0.99	+8.38	+1.39	+1.91	+3.75	+15.31
Micro / Phys	+5.10	+9.80	+2.27	+10.82	+1.21	+2.42	+3.70	+17.62
Micro / Exam	+6.93	**+15.50**	+2.55	**+12.72**	+1.95	+2.67	+5.11	+16.28
Micro / Patient	+3.22	+15.12	+1.83	+10.71	**+2.10**	**+3.24**	+3.87	+20.37
Gen / Imaging	+5.39	+11.71	+0.83	+7.90	+1.06	+1.41	+4.26	+19.54
Gen / Lab	+1.45	+3.97	+0.20	+4.18	+0.71	+0.79	+1.81	+5.03
Gen / Phys	+5.61	+9.58	+1.88	+9.54	+1.74	+2.22	+4.66	**+21.71**
Gen / Exam	+6.05	+10.60	+1.20	+7.90	+1.59	+2.03	+4.21	+17.64
Gen / Patient	+6.44	+10.86	**+3.11**	+10.48	+1.79	+1.86	+4.92	+19.19
Imaging / Lab	-1.34	-2.93	-0.41	-0.83	-0.18	-0.05	-0.66	+0.25
Imaging / Phys	+2.83	+5.90	+0.93	+4.25	+0.59	+1.89	+2.64	+14.73
Imaging / Exam	+0.65	+3.53	+0.04	+1.88	+0.28	+0.88	+0.94	+5.66
Imaging / Patient	-0.95	+2.80	-0.07	+0.75	+0.36	+0.78	+1.08	+5.38
Lab / Phys	+5.74	+10.20	+1.93	+7.60	+1.29	+2.09	+4.16	+16.62
Lab / Exam	+3.85	+6.45	+1.16	+4.46	+0.71	+1.29	+1.96	+9.08
Lab / Patient	+1.69	+5.86	+0.64	+3.77	+0.49	+0.77	+1.33	+8.99
Phys / Exam	-2.73	-2.46	-0.59	-2.36	-0.02	-0.31	-0.90	-4.21
Phys / Patient	-3.56	**-5.12**	-1.18	-3.04	-0.19	**-0.71**	-1.11	-5.15
Exam / Patient	-1.19	-1.37	-0.44	-0.91	-0.02	-0.68	-0.25	-0.75

Figure 7: Pairwise source preferences across eight LLMs under prose presentation.

Latent source scores. Figure 2 summarizes each model's 28 pairwise preferences by eight zero-sum source scores, fitted by equally weighted ordinary least squares (Jiang et al., 2011):

$$\hat{\boldsymbol{u}} = \arg \min_{\boldsymbol{u}: \sum_{s \in \mathcal{S}} u_s = 0} \sum_{(X,Y) \in \mathcal{E}} \left[M_{XY} - (u_X - u_Y) \right]^2, \tag{14}$$

where $\mathcal{S}$ contains eight sources and $\mathcal{E}$ their 28 pairs in a fixed orientation. With equal pair weights, $\hat{u}_X = \frac{1}{8} \sum_Y M_{XY}$, where $M_{YX} = -M_{XY}$ and $M_{XX} = 0$. Each score summarizes preference against all seven opponents, independent of pair sample sizes. A source's fitted rank reflects these combined margins, so a direct pairwise loss can coexist with a higher overall score.

Source rankings under structured presentation. We average each source-pair score over four position configurations with both records in the same structured format, then refit Equation 14 with equal pair weights. Each cell of Table 6 reports Kendall's τ between the structured and prose rankings of eight sources and the count of sign reversals among the 28 observed source-pair means.

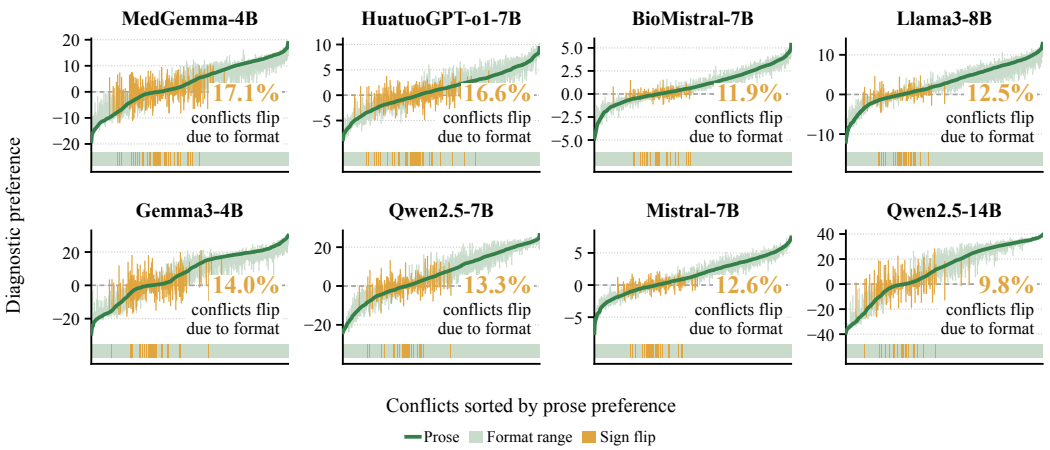

Figure 8: Preference ranges across 13 presentations; gold marks diagnosis reversals.

Table 6: Source-rank agreement and pairwise preference reversals.

Backbone	Sectioned	Key-value	Table	JSON
MedGemma-4B	1.000 / 0	1.000 / 0	1.000 / 0	0.929 / 1
Gemma3-4B	0.929 / 0	1.000 / 0	1.000 / 0	1.000 / 1
HuatuoGPT-o1-7B	0.929 / 0	0.929 / 1	0.929 / 1	0.929 / 1
Qwen2.5-7B	0.929 / 0	1.000 / 1	0.929 / 0	0.929 / 0
BioMistral-7B	0.857 / 1	0.857 / 1	0.929 / 1	0.857 / 0
Mistral-7B	1.000 / 0	1.000 / 1	1.000 / 1	0.929 / 1
Llama3-8B	1.000 / 0	1.000 / 0	1.000 / 0	0.929 / 0
Qwen2.5-14B	0.929 / 0	0.857 / 0	0.857 / 0	0.929 / 1

Across 32 model–format combinations, τ ranges from 0.857 to 1.000; 13 of 896 source-pair comparisons reverse sign (1.5%), all among their model's four smallest absolute prose gaps. Source rankings remain largely stable. Preferences reverse in 9.8%–17.1% of conflicts (Appendix C.2).

C.2 Format Sensitivity

Every conflict under re-rendering. Figure 8 extends Figure 1 (right) to eight models. Each line gives the all-prose preference; each bar spans the 13 states $\mathcal{V}_{13}$ (Equation 13), averaging four position configurations per state. A flip requires a strictly negative minimum and positive maximum; touching zero does not count. These conflict-level reversal rates range from 9.8% (128/1,310; Qwen2.5-14B) to 17.1% (224/1,310; MedGemma-4B). Panels sort conflicts by prose preference, use separate vertical scales, and pair domain-adapted and base models vertically in the first three columns.

Magnitude and source dependence. The analysis-set mean absolute format shift (FPS) ranges from 0.28 nats for BioMistral-7B to 2.48 nats for Qwen2.5-14B. Only 1.2%–4.4% of re-renderings leave the score unchanged, while 7.2%–69.8% change the odds between the diagnoses by at least a factor of two. Source-specific mean shifts range from 0.71 to 1.23 times each model's overall mean, with physiology the most stable source under re-rendering in seven of the eight frozen models.

Direction of the shifts. Table 7 averages signed shifts toward the re-rendered record over both records of 1,310 conflicts. JSON shifts preference away from that record by 0.87 nats on MedGemma-4B but toward it by 0.30 nats on Qwen2.5-7B. Grouping by format explains at most 3.4% of shift variance; adding the re-rendered source explains 2.9%–8.6%, and adding the competing source explains 5.5%–14.6%. The remaining conflict-level variation motivates the instance-specific reference in Section 3. Positive values favor the re-rendered record; negative values disfavor it. Gray entries have a 95% bootstrap interval over conflicts or pair groups that includes zero.

Table 7: Signed format effects in nats, pooled over evidence sources.

Backbone	Sectioned	Key-value	Table	JSON
MedGemma-4B	+0.01	−0.13	−0.69	−0.87
Gemma3-4B	+0.18	+0.12	−0.27	−0.60
HuatuoGPT-o1-7B	+0.14	+0.07	+0.01	+0.13
Qwen2.5-7B	+0.37	+0.24	+0.23	+0.30
BioMistral-7B	+0.05	+0.04	+0.00	+0.05
Mistral-7B	+0.06	+0.03	−0.01	+0.02
Llama3-8B	+0.11	+0.06	−0.08	−0.14
Qwen2.5-14B	+0.28	+0.12	+0.17	−0.65

D PROOFS

This section proves the two results used in Section 4: the consensus properties (Proposition D.1) and Proposition 1. Proposition D.2 further extends the second bound of Proposition 1 to partially flattened students. All logarithms are natural, so preference scores and divergences are measured in nats. Every distribution below is a vector over the four coordinates $(D_A, D_B, D_{\mathrm{N}}, \mathrm{rest})$ of Section 4.2, and every coordinate is positive because the probabilities come from a softmax. For a distribution p, we write $M(p) = \log p(D_A) - \log p(D_B)$ for its preference score.

Proposition D.1 (Properties of the Consensus Target). *Fix a neighborhood $\mathcal{N}_f$, let $p_v^0 = p_{\theta_0}(\cdot \mid x^v)$ for $v \in \mathcal{N}_f$, and let $q = q_f$ be the arithmetic mean of these frozen predictions (Equation 6).*
(a) The target q is the unique minimizer of $\frac{1}{|\mathcal{N}_f|}\sum_v D_{\mathrm{KL}}(p_v^0 \,\|\, r)$ over all distributions r.
(b) The target satisfies $\min_v M(p_v^0) \le M(q) \le \max_v M(p_v^0)$. Thus q preserves any preference shared by all views: if each favors D_A, so does q; the same holds for D_B.

Proof. (a) Using $q = |\mathcal{N}_f|^{-1}\sum_v p_v^0$, the mean divergence to any distribution r is

$$\frac{1}{|\mathcal{N}_f|}\sum_v D_{\mathrm{KL}}(p_v^0 \,\|\, r) = \frac{1}{|\mathcal{N}_f|}\sum_v \sum_y p_v^0(y)\log p_v^0(y) - \sum_y q(y)\log r(y).$$

Only the last term depends on r, and it equals $H(q) + D_{\mathrm{KL}}(q \,\|\, r)$, where $H(q)$ is the entropy of q. The expression is therefore minimized if and only if $D_{\mathrm{KL}}(q \,\|\, r) = 0$, that is, $r = q$.

(b) Let $a_v = p_v^0(D_A) > 0$ and $b_v = p_v^0(D_B) > 0$. The pooled odds can be written as

$$\frac{q(D_A)}{q(D_B)} = \frac{\sum_v a_v}{\sum_v b_v} = \sum_v w_v \frac{a_v}{b_v}, \qquad w_v = \frac{b_v}{\sum_u b_u}.$$

The nonnegative weights w_v sum to one, placing $q(D_A)/q(D_B)$ between the extreme view ratios a_v/b_v. Taking logarithms gives the bound on $M(q)$. If all views favor D_A, their ratios exceed one, as does the weighted mean; hence $M(q) > 0$. The case favoring D_B follows by symmetry. □

Proof of Proposition 1. We follow Section 4.3, writing $M_v = M_\theta(x^v)$ and $p_v = p_\theta(\cdot \mid x^v)$ for the student, $q = q_f$ for the consensus, $s = q(D_A) + q(D_B)$, $\pi = q(D_A)/s$, and H_b for binary entropy. We prove the two bounds and then show why neither objective alone provides both guarantees.

First bound. Contraction controls the gap between every pair of views. The four edges in $\mathcal{E}_f$ connect (P, P), (f, P), (f, f), and (P, f) in a cycle of length four. For two distinct views u and v, the cycle splits into two edge-disjoint paths from u to v. By the triangle inequality, $|M_u - M_v|$ is at most the sum of the edge gaps along each path. Adding these two path inequalities therefore gives

$$2\,|M_u - M_v| \le \sum_{(w,z)\in\mathcal{E}_f} |M_w - M_z| = 4\,\mathcal{L}_{\mathrm{ctr}}(f),$$

which proves the first bound of Equation 11. The bound also covers the diagonal pairs $((P, P), (f, f))$ and $((f, P), (P, f))$, in which both records change format and which no edge joins.

Second bound. Distillation penalizes a flattened preference. Let the student be flattened, so that $p_v(D_A) = p_v(D_B)$ for every view, and let $t_v = p_v(D_A) + p_v(D_B)$. Grouping D_A and D_B, the KL chain rule separates block-mass and within-block terms for each of the four student views:

$$D_{\mathrm{KL}}(q \,\|\, p_v) = D_{\mathrm{KL}}\big((s,\, q(D_{\mathrm{N}}),\, q(\mathrm{rest})) \,\|\, (t_v,\, p_v(D_{\mathrm{N}}),\, p_v(\mathrm{rest}))\big) + s\, D_{\mathrm{KL}}\big((\pi,\, 1-\pi) \,\|\, (\tfrac{1}{2},\, \tfrac{1}{2})\big).$$

The first term is nonnegative and vanishes if and only if the three block masses agree. The second term equals $s\,[\pi \ln(2\pi) + (1-\pi)\ln(2(1-\pi))] = s\,(\ln 2 - H_b(\pi))$. Averaging over the four views gives the second bound of Equation 11, with equality if and only if every p_v places the masses s, $q(D_{\mathrm{N}})$, and $q(\mathrm{rest})$ on the three blocks. Since $s > 0$ and $H_b(\pi) \le \ln 2$ with equality only at $\pi = 1/2$, the bound is positive if and only if q prefers one diagnosis, and since $s \le 1$ and $H_b \ge 0$, it never exceeds $\ln 2$. Finally, $\pi = \sigma(M(q))$ for the logistic function σ, and $H_b(\sigma(m))$ is even in m and decreasing in $|m|$, so the bound increases with $|M(q)|$ at fixed diagnosis mass s.

Last statement. Neither term alone gives both guarantees: distillation does not bound preference gaps, and contraction permits flattened preferences even when the consensus favors a diagnosis.

Distillation alone does not bound the gap. Fix $\Delta > 0$, set $a = e^{\Delta/4}$ and $c = (a + a^{-1})/2 \ge 1$, and take $0 < \delta < 1/(2c)$. Over $(D_A, D_B, D_{\mathrm{N}}, \mathrm{rest})$, define the target and student distributions

$$\begin{aligned} q &= \big(\delta,\ \delta,\ \tfrac{1-2\delta}{2},\ \tfrac{1-2\delta}{2}\big), \\ r^+ &= \big(a\delta,\ a^{-1}\delta,\ \tfrac{1-2c\delta}{2},\ \tfrac{1-2c\delta}{2}\big), \\ r^- &= \big(a^{-1}\delta,\ a\delta,\ \tfrac{1-2c\delta}{2},\ \tfrac{1-2c\delta}{2}\big). \end{aligned}$$

These positive vectors sum to one; q is a valid consensus when every frozen view returns q. The student preferences are $M(r^+) = \Delta/2$ and $M(r^-) = -\Delta/2$. In $D_{\mathrm{KL}}(q \| r^+)$, the diagnosis terms $-\delta \log a$ and $\delta \log a$ cancel, leaving only the contribution from the other two coordinates:

$$D_{\mathrm{KL}}(q \,\|\, r^+) = (1-2\delta)\log\frac{1-2\delta}{1-2c\delta} \le \frac{2(c-1)\delta}{1-2c\delta},$$

where we use $1 - 2\delta \le 1$, the nonnegativity of the logarithm for $c \ge 1$, and $\log(1+x) \le x$. The same bound holds for r^- by symmetry. Now let the student return r^+ on (P, P) and (f, f) and r^- on (f, P) and (P, f). Every edge of $\mathcal{E}_f$ then joins a view with r^+ to a view with r^-, so every edge gap equals Δ and $\mathcal{L}_{\mathrm{ctr}}(f) = \Delta$. Meanwhile, $\mathcal{L}_{\mathrm{cons}}(f) \le 2(c-1)\delta/(1-2c\delta)$, which tends to zero as $\delta \to 0$. For any $\epsilon > 0$, a small enough δ therefore gives $\mathcal{L}_{\mathrm{cons}}(f) \le \epsilon$ while adjacent views differ in preference score by Δ. The construction works because both diagnoses receive a small probability δ, so moving very little probability mass changes their log ratio by a large amount.

Contraction alone allows a flattened student. Constant M_v makes every summand of $\mathcal{L}_{\mathrm{ctr}}(f)$ zero. Every flattened student has zero contraction loss for any consensus, even with a large distillation bound. Contraction cannot distinguish preserving the consensus preference from erasing it. □

Proposition D.2 (Partial Flattening). *Let $s_f = q_f(D_A) + q_f(D_B)$, $\pi_f = q_f(D_A)/s_f$, and $d(a \,\|\, b) = a \ln \frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$ be the binary KL divergence. With logistic function σ, each student view has a binary-divergence lower bound on its relative diagnosis preference:*

$$D_{\mathrm{KL}}(q_f \,\|\, p_v) \ge s_f\, d\big(\pi_f \,\|\, \sigma(M_v)\big).$$

Consequently, if $|M_v| \le m < |M(q_f)|$ on all four views, with $m \ge 0$, then

$$\mathcal{L}_{\mathrm{cons}}(f) \ge s_f\, d\big(\pi_f \,\|\, \sigma(m\ \mathrm{sgn}\, M(q_f))\big) > 0.$$

This recovers the second bound of Proposition 1 at $m = 0$ and increases as m decreases.

Proof. Group coordinates into $\{D_A, D_B\}$, $\{D_{\mathrm{N}}\}$, and the rest. The KL chain rule gives a nonnegative divergence between block masses plus $s_f d(\pi_f \| \rho_v)$, with $\rho_v = p_v(D_A)/(p_v(D_A) + p_v(D_B)) = \sigma(M_v)$, proving the first inequality by dropping the nonnegative term. Next consider

$$d(\pi_f \| \sigma(t)) = -\pi_f t + \ln(1 + e^t) - H_b(\pi_f).$$

The convex function has derivative $\sigma(t) - \pi_f$ and its minimum at $t = M(q_f)$. For $M(q_f) > 0$, convexity implies that $M_v \le m < M(q_f)$ gives $d(\pi_f \| \sigma(M_v)) \ge d(\pi_f \| \sigma(m))$; negative $M(q_f)$ is symmetric. Averaging across views yields the bound and $d(\pi_f \| \frac{1}{2}) = \ln 2 - H_b(\pi_f)$ at $m = 0$. □

E IMPLEMENTATION DETAILS OF FORMEDER

E.1 BASELINES

The frozen backbone is the reference before adaptation. All trained baselines and the contraction-only ablation in Section 5.3 share FORMEDER's frozen teacher, four-view neighborhoods, sampler, adapter, learning rate, and update budget. Only their training objectives differ.

Per-rendering KD targets each view's frozen prediction. With teacher initialization and zero LoRA update ($B = 0$), matched deterministic forward passes give identical normalized four-coordinate distributions. KL divergence and its gradient are zero, so this control reproduces Frozen.

Prose-reference KD gives every view the frozen prediction on the all-prose view, so it supplies a single canonical judgment instead of a consensus. *Format-augmented SFT* supervises every view with the option letter that the frozen model answers under prose. The student therefore sees all four renderings of an instance with one hard label and no information about how close the decision was. *Contraction only* keeps the residual contraction of Equation 9 at the weight of FORMEDER and removes the distillation term, so it differs from FORMEDER in a single setting.

E.2 TRAINING CONFIGURATION

Table 8 gives shared settings. Each sampled instance, position, and structured format defines a four-view neighborhood (Equation 5). The teacher disables the backbone's adapter; its detached target averages the four predictions (Equation 6), without a moving-average teacher. Training minimizes KL divergence on all views plus λ times the mean absolute edge gap. The four coordinates represent the two record-supported diagnoses, *Neither*, and remaining-vocabulary mass. Gradient accumulation gives an effective batch of 16. Hyperparameters and answer-token conventions are fixed on validation data; all models use the listed seed and final checkpoint after 375 updates. Inference uses one adapted student.

Table 8: Training configuration.

Settings	Values
All trained models	
LoRA rank / scaling	64 / 128
LoRA dropout	0.05
Learning rate	5×10^{-5}
Updates	375
Effective batch size	16
Maximum input length	2,048 tokens
Training seed	20260908
Training instances	3,928
Objective of FORMEDER	
Consensus target	Arithmetic mean
Distillation weight	1
Contraction weight λ	4

Choosing the contraction weight. MedGemma-4B validation FPS, FCC, and prose-answer preservation determine λ. In Figure 9, FPS decreases from $\lambda = 1$ to 16, while FCC peaks at 95.60% for $\lambda = 4$. At $\lambda = 8$ and 16 (hatched), lower FPS coincides with lower FCC. We apply $\lambda = 4$ unchanged to all eight backbones (Section 5) and assess FCC, RA, and frozen prose-answer agreement alongside the scale control in Appendix F.2. Table 2 reports test ablations on MedGemma-4B.

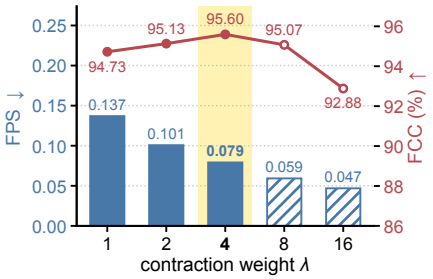

Figure 9: Validation λ sweep.

E.3 ABLATION TARGETS

Table 2 evaluates MedGemma-4B on test data. The prose-target variant replaces the consensus with the frozen all-prose prediction while retaining contraction. The random-format variant targets the frozen view with both records in the sampled structured format. Majority vote uses each view's argmax over four coordinates, including *Neither*, and distills a one-hot label on the most frequent answer; ties select the answer with highest mean probability. All three variants retain $\lambda = 4$ and share the other ablations' fixed seed, sampled instances, rendering formats, and neighborhood edges.

F Evaluation Protocol

F.1 Metrics and Pairing

Data splits and model selection. Validation data determine hyperparameters and answer-token conventions, including the sweep in Figure 9. Main results, ablations, and unseen-format generalization use the held-out test set; source-preference analyses use the separate analysis set.

Decoding and answer parsing. All models use greedy decoding with a maximum of 2,048 new tokens, stopping at the end-of-turn token. The parser uses an explicit final-answer line or a terminal standalone option letter. Reasoning-only letters and missing or ambiguous answers are unparseable.

Metrics. FPS and FCC average the 64 edge comparisons per instance (Equation 12). FPS evaluates the same mean absolute gap as $\mathcal{L}_{\mathrm{ctr}}$ on held-out instances. FCC compares parsed diagnosis identities: two *Neither* answers agree, whereas an unparseable reply makes the comparison inconsistent. RA requires correct single-record answers in prose and all four structured formats, replacing the latter with YAML, bullet lists, INI, and outline for unseen-format RA; *Neither* and unparseable replies are errors. FCC, RA, and unseen-format metrics are not directly optimized.

Hardware. All computation uses NVIDIA RTX PRO 6000. Trained and frozen models share inputs, renderings, orders, and scoring. Red and blue mark best and second-best performance (Table 1).

Computational budget. Each configuration uses 375 optimizer updates with an effective batch of 16 neighborhoods. At inference, one adapted model answers from a single rendering.

F.2 FPS Against Rescaled Frozen Scores

Uniform score shrinkage reduces FPS by the same factor. Let S be the mean absolute all-prose preference, averaging the four position configurations before taking the absolute value, and let $r = S_{\mathrm{ours}}/S_0$. The matched *shrink line* is $r\,\mathrm{FPS}_0$; dividing it by $\mathrm{FPS}_{\mathrm{ours}}$ gives the reduction beyond uniform shrinkage. In Table 9, FormedER retains 33%–58% of frozen prose strength, with FPS below the matched shrink line by a factor of 5.78–14.12. See Appendix G.2 for preservation results.

Table 9: Test-set preference strength and scale controls.

Backbone	S_0	S_{ours}	$100r$ (%)	Shrink line	$r\,\mathrm{FPS}_0/\mathrm{FPS}_{\mathrm{ours}}$
MedGemma-4B	6.25	2.63	42	0.7195	9.38
Gemma3-4B	11.64	4.19	36	1.0791	14.03
HuatuoGPT-o1-7B	2.89	1.68	58	0.4743	14.12
Qwen2.5-7B	9.36	3.84	41	0.9184	12.48
BioMistral-7B	0.94	0.51	54	0.1434	5.78
Mistral-7B	1.57	0.75	48	0.1856	5.95
Llama3-8B	2.84	1.48	52	0.4099	6.81
Qwen2.5-14B	15.79	5.21	33	1.2053	9.99

G Additional Results

G.1 Number of Training Formats

Table 10 reports the Figure 6 evaluation on YAML, bullet lists, INI, and outline in the test set. These formats are excluded from training and selection. Training starts with prose and sectioned text, then adds key-value, table, and JSON. Five-format training uses Table 1's checkpoint; smaller conditions vary only formats, sharing $\lambda = 4$, adapters, and the update budget (Appendix E.2).

FPS and FCC use the main evaluation's position configurations and edges, averaged over unseen formats. RA requires the same record's correct diagnosis in prose and all four unseen formats; frozen models provide the reference. Increasing training formats from two to five reduces mean FPS from 0.6116 to 0.2412 nats, raises FCC from 84.63% to 91.52%, and raises RA from 43.07% to 45.69%. RA improves as formats are added, with gains occurring at different stages across backbones. With all five training formats, every backbone has lower unseen-format FPS than its frozen reference.

Table 10: Test-set generalization to four unseen formats.

Metrics	Training formats	MedGemma 4B	Gemma3 4B	HuatuoGPT o1-7B	Qwen2.5 7B	BioMistral 7B	Mistral 7B	Llama3 8B	Qwen2.5 14B
FPS (nats) ↓	Frozen	1.9187	3.5372	0.8914	2.5760	0.3240	0.4486	0.9775	4.1273
	2: prose + sectioned	0.6201	0.9505	0.3988	0.7815	0.2993	0.2411	0.5001	1.1010
	3: + key-value	0.3692	0.8854	0.3816	0.5046	0.1880	0.2239	0.4756	0.8182
	4: + table	**0.3413**	**0.5233**	**0.3012**	**0.2151**	**0.1765**	**0.1202**	**0.2823**	**0.7705**
	5: + JSON	**0.2134**	**0.3967**	**0.1765**	**0.1895**	**0.1509**	**0.1150**	**0.2416**	**0.4462**
FCC (%) ↑	Frozen	82.37	79.18	80.94	81.06	86.83	88.29	83.46	77.64
	2: prose + sectioned	84.39	81.03	86.35	82.88	87.79	89.64	85.19	79.79
	3: + key-value	85.74	85.96	89.35	83.52	88.21	92.62	87.83	80.74
	4: + table	**90.46**	**86.40**	**89.68**	**88.74**	**91.45**	**92.79**	**87.97**	**86.92**
	5: + JSON	**91.81**	**89.57**	**92.23**	**93.26**	**92.17**	**93.81**	**91.18**	**88.13**
RA (%) ↑	Frozen	51.83	42.10	37.35	38.52	41.96	63.43	26.74	35.23
	2: prose + sectioned	52.69	42.85	37.90	39.11	45.56	63.88	26.91	35.65
	3: + key-value	52.86	45.05	38.37	39.83	46.27	64.77	27.08	36.71
	4: + table	**54.44**	**45.27**	**38.45**	**39.92**	**52.04**	**64.84**	**28.16**	**37.46**
	5: + JSON	**54.71**	**45.82**	**39.64**	**41.87**	**52.48**	**65.09**	**28.37**	**37.56**

G.2 Preservation Analyses

Preference preservation. Position-averaged prose scores remain strongly rank-correlated with frozen scores (Spearman's $\rho = 0.81$–0.94); prose strength retains 33%–58% of its frozen value (Table 9). Removing consensus distillation while retaining contraction at $\lambda = 4$ causes *output degeneration*: replies become unparseable as preference scores approach zero, marked by ✗ in Table 2.

Answer preservation. Across 5,288 prose comparisons per backbone (1,322 test instances in four position configurations), FORMEDER retains 87.3%–93.4% of frozen answers, slightly below prose-reference KD on every backbone. We compare diagnoses after mapping option letters.

G.3 Case Study

Figure 10 uses a single-patient vignette, separate from the benchmark. Imaging and reported symptoms support competing explanations for breathlessness. Only Record 1's layout changes; evidence, patient profile, and record and option orders stay fixed. Positive M favors A; negative M favors B.

H Limitations

Scope of source preferences. Each PMC-ConForm conflict pairs records from two different patients and therefore has no joint clinical gold diagnosis. Conflict prompts contain no explicit source labels, and source rankings are measured in prose. We therefore use source preference to describe systematic differences across records grouped by source category, rather than preference for source labels themselves. Our balanced construction controls patient- and diagnosis-level effects within each source pair, although record characteristics may still vary across source categories. Accordingly, the observed rankings should be interpreted as descriptive patterns associated with these categories.

What the design identifies. Our format analysis relies on within-instance comparisons in which the underlying records, sentences, demographics, options, and positions are fixed, while only their presentation changes. This design more directly isolates presentation-induced variation from content differences. FORMEDER likewise does not depend on source rankings or conflict-level gold labels: its target is the frozen model's cross-format consensus, FPS and FCC compare different renderings of the same instance, and RA evaluates single-record questions with well-defined targets.

Case Study on FORMEDER

(i) Before FORMEDER Training

`Input: both records in prose`

Patient
A 68-year-old patient presents with shortness of breath.

Record 1
The chest radiograph shows bilateral perihilar opacities and small bilateral pleural effusions. The cardiac silhouette is enlarged.

Record 2
The patient reports three days of fever and cough with yellow sputum. Breathlessness developed over the same period.

Question: Which diagnosis is most likely? Answer with a single letter.
A. Acute heart failure
B. Community-acquired pneumonia
C. Neither of the above

`Format change: Record 1 in key-value`

`Finding 1:` The chest radiograph shows bilateral perihilar opacities and small bilateral pleural effusions.
`Finding 2:` The cardiac silhouette is enlarged.

Record 2, all evidence sentences, and the record and option orders are unchanged.

`Frozen MedGemma-4B output`

Record 1 / Record 2	Answer	Preference M (nats)
Prose / Prose	**A**	$+1.34$
Key–value / Prose	**B**	-0.47

✗ *The diagnosis choice reverses when only the format changes.*

(ii) After FORMEDER Training

For each of the same two input presentations, the trained model selects diagnosis **A**.

`MedGemma-4B + ForMedER output`

Record 1 / Record 2	Answer	Preference M (nats)
Prose / Prose	**A**	$+0.62$
Key–value / Prose	**A**	$+0.54$

✓ *The named diagnosis stays consistent across the two presentations.*

Both trained preference margins remain positive and are closer together across the two presentations.

Figure 10: Choice stability under a format-only change to the same patient's records.