TABLES2TRACES: DISTILLING TABULAR DATA TO IMPROVE LLM REASONING IN HEALTHCARE

Anonymous authorsPaper under double-blind review

000

001

002003004

010 011

012

013

014

015

016

017

018

019

021

024

025

026

027

028

029

031

032

033 034

035

037

040

041

042

043

044

046

047

048

051

052

ABSTRACT

Large language models (LLMs) excel at reasoning when fine-tuned on curated text corpora, but many domains, such as medicine, primarily store knowledge in structured tabular data. Despite its richness, tabular data has been largely overlooked as a source of reasoning supervision. Interpreting such data requires structured, relational reasoning across features and outcomes, not just surface-level pattern matching. In practice, this mirrors clinical decision making, where doctors often compare patients with similar characteristics and reason about why their outcomes diverge. We introduce **Tables2Traces**, the first framework to enable improved reasoning from raw tabular data by generating contrastive, case-based reasoning traces for model fine-tuning. This establishes a new supervision paradigm: converting tabular records, traditionally used only for prediction, into structured reasoning signals that can serve as an effective new source of supervision for LLMs. Crucially, this paradigm is orthogonal to text-based QA supervision: rather than competing with curated corpora, it unlocks an abundant and low-cost modality that complements existing approaches. Using only cardiovascular patient records, Tables2Traces yields relative gains of 17.2% on in-domain MedQA questions and 8.4% out-of-domain, improving accuracy in 15 of 17 clinical categories. On MedM-CQA, it achieves a 7.2% relative improvement and outperforms the base model in 17 of 21 specialties. These gains are driven by a lightweight, domain-agnostic pipeline that elicits structured reasoning via contrastive and counterfactual prompts. Compared to training on narrative patient descriptions, Tables2Traces generalizes more effectively across question types and medical specialties, showing that even limited tabular data can serve as a scalable and complementary source of reasoning supervision for LLMs.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across reasoning tasks, from multi-step mathematics (Cobbe et al., 2021) to medical question answering (Singhal et al., 2023). These advances are typically attributed to large-scale pretraining followed by supervised fine-tuning on datasets already structured as text-based reasoning tasks (Ouyang et al., 2022; Wei et al., 2022a).

In many domains, however, knowledge is stored in *structured*, *non-linguistic formats* such as electronic health records, lab results, insurance claims, finance spreadsheets, or scientific measurements. Such datasets encode complex relationships and decision logic but lack the textual form required for LLM fine-tuning (Yin et al., 2020; Liu et al., 2021). Closing this modality gap would unlock the reasoning signals latent in these data sources.

Consider a clinician assessing cardiovascular risk from patient records. A row of clinical features (e.g., age, LDL, diabetes, blood pressure) supports reasoning such as: "patients over 60 with high LDL and diabetes are at elevated risk, even if blood pressure is normal." Clinicians perform this reasoning intuitively, yet there is currently no systematic way to pass this knowledge to an LLM. Existing strategies rely on curated QA corpora (Puri et al., 2020), table-embedding models (e.g., TaBERT (Yin et al., 2020), TAPEX (Liu et al., 2021)), or lightweight adapters (Li & Liang, 2021; Hu et al., 2022), but none directly translate raw tabular data into reasoning supervision.

This raises two challenges. **(C1) Representation**: how to represent each row of features into a coherent format suitable for reasoning while preserving feature relationships.

(C2) Trace elicitation: how to automatically generate reasoning traces that capture the latent knowledge contained in the structured data.

To address C1-C2, we introduce **Tables2Traces**, the first end-to-end framework to transform the latent knowledge embedded in tabular data to reasoning traces that can then be used to fine-tune and improve an LLM. More specifically, Tables2Traces addresses the following key research question:

Can the latent knowledge embedded in structured tabular data be reformulated into reasoning tasks that LLMs can learn from—and does fine-tuning on such synthetic examples improve reasoning in both in-domain and out-of-domain settings?

Addressing this question offers dual benefits. **For domain experts**, it offers a path to adapt LLMs using structured datasets to which they already have access. This is especially valuable in fields like medicine, where data privacy and regulatory constraints often prevent data sharing. Practitioners can extract reasoning supervision directly from their own tabular data, effectively imbuing LLMs with local contextual knowledge.

For the LLM research community, this approach provides a new and complementary modality that contains rich domain knowledge for supervision. Although most fine-tuning datasets are human-annotated or LLM-synthesized from existing text corpora (Gururajan et al., 2024), we demonstrate that abundant tabular data can also provide useful reasoning supervision. This reframes tabular data as a rich and domain-specific supervision source and contributes to the growing data-centric shift in LLM development. Crucially, to the best of our knowledge, we are the first to demonstrate that latent knowledge encoded in tabular datasets can be reformulated into reasoning tasks that improve LLMs.

In doing so, we make the following contributions:

Contributions. ① Conceptually: We introduce a novel LLM supervision paradigm of transforming tabular datasets into structured reasoning traces. This introduces a new pathway for domain adaptation using data that has traditionally been excluded from LLM training pipelines. ② Methodologically: We present Tables2Traces, a lightweight and modular pipeline that generates contrastive questions and multi-step reasoning traces from labeled tabular data without requiring QA corpora, human annotation, or task-specific templates. ③ Empirically: We evaluate on MedQA and MedMCQA, using 105k synthetic traces and 10k QA examples. Tables2Traces yields strong in-domain gains (+17.2%) and generalizes out-of-domain (+8.4% on MedQA, +7.2% on MedMCQA), showcasing the effectiveness of tabular supervision. ④ Analytically: We find that Tables2Traces closes part of the gap to a state-of-the-art model, *Aloe*, despite using only 2% of its QA data. We show how contrastive supervision improves generalization across question types and embedding space regions.

2 RELATED WORK

This work engages with works on LLM fine-tuning and LLMs for tabular data.

LLM Fine-Tuning. Prior work demonstrates that fine-tuning LLMs on structured reasoning datasets (e.g. GSM8K) can significantly improve problem-solving capabilities, which can then be enhanced via chain-of-thought prompting (Wei et al., 2022b). Instruction-tuning (e.g. Self-Instruct (Wang et al., 2023b)) further show that training with human-style prompts and reasoning traces enhances generalization across unseen tasks. More recently, DeepSeek-R1 (Guo et al., 2025) introduced a large-scale framework for fine-tuning LLMs on curated reasoning traces using the Generalized Reinforcement Preference Optimization (GRPO). By combining diverse, high-quality reasoning traces with fine-tuning, DeepSeek-R1 demonstrated strong improvements. Our work builds on this paradigm by generating structured reasoning traces not from existing text corpora, but from raw tabular datasets, enabling fine-tuning in domain-specific settings.

In parallel, alignment methods such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), direct preference optimization (DPO) (Rafailov et al., 2023), and reward modeling (Christiano et al., 2017) have shown that models benefit from being tuned on outputs aligned with human preferences. However, all of these methods assume access to a large corpus of natural language examples. In contrast, we synthesize supervision from structured data, which is abundant but underutilized in current LLM pipelines. Our work is orthogonal to these and serves to highlight the potential of structured tabular data as a new source of supervision.

LLMs for tabular data. Research on modeling structured tables with language models has largely focused on two paradigms: semantic table understanding and supervised prediction. For the former, models like TaBERT (Yin et al., 2020), TAPEX (Liu et al., 2021), and TURL (Deng et al., 2022) learn joint text-table representations for question answering and schema reasoning. For the latter task of prediction, architectures such as TabNet (Arik & Pfister, 2021) and FT-Transformer (Gorishniy et al., 2021) are optimized for predictive modeling over tabular features.

More recent methods, such as TabLLM (Hegselmann et al., 2023) and UniPredict (Wang et al., 2023a), explore using general-purpose LLMs with serialized tabular input, enabling LLMs to perform tabular prediction (in particular, highlighting their zero-shot classification capabilities).

However, these methods tackle the fundamentally different problem of LLMs understanding tables or LLMs being used as tabular predictors. In contrast, we focus on answering the question of how to use tabular structured data (and the knowledge contained therein) to enhance the reasoning capabilities of LLMs within the relevant problem domain. i.e. how can we use tabular data on cardiovascular patients to improve an LLMs capabilities to reason about cardiovascular problems or even more general medical questions.

3 METHOD

108

110

111

112

113 114

115

116

117

118

119

120

121

122

123 124

125 126

127 128

129

130

131

132

133 134 135

136

137

138

139

141

142

143

144 145

146 147

148

149

150 151

152

153

154

155

156

157 158

159

160

161

3.1 Problem Formulation

We assume a tabular dataset $\mathcal{D} = \{x_1, \dots, x_N\}$ where each row $x_i \in \mathbb{R}^d$ is a structured record (e.g., a clinical case) and is associated with a binary label $y_i \in \{0,1\}$. While such tabular datasets encode rich domain knowledge, they do not naturally align with the data formats LLMs are typically trained on. We hence seek a mapping that converts this structure into contrastive prompts and structured reasoning traces suitable for supervised fine-tuning of LLMs:

$$\Pi: \mathcal{D} = \left\{ (x_i, y_i) \right\}_{i=1}^N \longrightarrow \mathcal{C} = \left\{ (P_i, R_i) \right\}_{i=1}^M \tag{1}$$

 $\Pi: \mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \longrightarrow \mathcal{C} = \{(P_i, R_i)\}_{i=1}^M$ (1) where each prompt P_i describes a clinical scenario in natural language and each trace R_i is a structured reasoning trace generated by an LLM.

Ultimately, the goal is that fine-tuning a target LLM with parameters Θ on \mathcal{C} (derived from tabular data), can teach the model to learn high-level reasoning behaviors from the structure of the data, without requiring domain-specific logic or annotation. In doing so, this can improve LLM performance on related text-based tasks such as Q&A. To anchor ideas, we hypothesize that eliciting reasoning traces from tabular medical data and then fine-tuning an LLM on these traces should improve an LLM's capabilities on medical Q&A tasks.

In this work, we primarily focus on the role of medical tabular data, hence our focus is on eliciting clinical reasoning capabilities from a structured medical dataset.

3.2 TABLES2TRACES

We propose Tables2Traces as a framework that realizes this mapping function Π . The underlying algorithm is outlined in Algorithm 1. In particular, it allows us to provide solutions to overcome the challenges of (C1) Representation and (C2) Trace elicitation.

(C1) Representation. Our first challenge is how to represent the tabular data in a suitable format, prior to eliciting reasoning traces. A deterministic encoder ϕ translates each tabular row x_i into a compact textual patient description. Column headers are normalized into human-interpretable phrases (e.g. ldl_chol→low-density lipoprotein cholesterol), numerical values are rendered with units, and missing entries are declared explicitly. Ultimately, ϕ (operationalized with an LLM), transforms tabular rows into fluent text-based summaries. The result is a corpus of textual case descriptions $C_{\text{simple}} = \{\phi(x_i)\}_{i=1}^N$, which we use as training data for the Tables2Traces (simple) variant.

- (C2) Trace Elicitation. Once we have the data in a suitable, represented textual format, we wish to elicit the appropriate knowledge and reasoning from the data. We do so as follows:
- ► Contrastive Neighbor Selection. Clinical reasoning frequently involves comparative analysis, i.e., why did this patient die, whereas a similar patient survived? To elicit a similar contrastive reasoning

163

164

165 166

167

168 169

170

171

172

173 174

175

176

177

178 179

181

182 183

185

186

187

188

189

190

191

192

193

194

195

196

197

199

200

201

202203

204

205206207

208209

210

211

212

213

214

215

for each anchor example x_i , we retrieve its nearest survivor $x_i^{(0)}$ and nearest deceased $x_i^{(1)}$ based on the Gower distance 1 , as it respects heterogeneous feature types. We use this to form a narrative triplet $\tau_i = (s_i^{(0)}, s_i, s_i^{(1)})$, which corresponds to a contrastive decision used as supervision for the LLM

- ▶ Reasoning extraction via prompt design. We wrap τ_i in a fixed prompt template P_i (see Appendix C.3) that explicitly outlines the reasoning competencies we want the model to learn:
 - 1. **Differential reasoning.** Compare the TARGET to each neighbour and list decisive feature differences. This elicits differential reasoning.
 - 2. **Label Plausibility.** State whether the recorded outcome is clinically plausible. By evaluating the ground-truth label, this elicits reasoning to recognise spurious or noisy supervision.
 - 3. **Counterfactual Planning.** We prompt the model to suggest one *minimal* feature edit that would reverse the outcome and justify why. This injects a causal inductive bias: the model must isolate an actionable feature (e.g. reduce LDL to < 100 mg/dL) and explain its physiological effect. This mirrors clinical reasoning in which a clinician asks, "What intervention would have saved this patient?"
- ▶ Trace Extraction and Corpus Assembly. We pass the constructed prompt P_i to the frozen LLM \mathcal{L} (e.g. Deepseek-R1) and extract the generated reasoning trace R_i . Collecting every pair yields the corpus \mathcal{C} , wherein we have converted the tabular data into structured reasoning traces.

Supervised Fine-tuning. We then fine-tune a downstream target LLM using standard supervised learning on the dataset $\{(P_i, R_i)\}$. Each training example consists of: **Input:** the contrastive prompt P_i (including the three case descriptions) and **Output:** the generated reasoning trace R_i .

We fine-tune the model on a dataset $\mathcal{R}' = \{(P_i, R_i)\}_{i=1}^M$ consisting of M prompt-response pairs, where each P_i is a contrastive input prompt and R_i is the corresponding reasoning trace. 90% of these samples are synthetic traces generated from tabular data and 10% are multiple-choice QA-format examples (e.g., MedQA). The QA examples were not contained in the evaluation datasets. Training on the QA subset alone performs on par with the base model (Appendix L).

We hold out 5% of the overall dataset for evaluation. Prompt templates and representative traces are shown in Appendices C and D.

Algorithm 1 TABLES2TRACES: From Tabular Data to Reasoning Corpus

```
Require: Tabular dataset D, frozen LLM \mathcal{L}
 1: Output: Reasoning corpus C
 2: for each (x_i, y_i) \in D do
             s_i \leftarrow \phi(x_i)
                                                                    ▶ Representation
 3:
            for y \in \{0,1\} do
 4:
           From y \in \{0, 1\} we x_i^{(y)} \leftarrow \arg\min_{x:y_x=y} \operatorname{Gower}(x, x_i) end for \tau_i \leftarrow \left(\phi(x_i^{(0)}), s_i, \phi(x_i^{(1)})\right) \Rightarrow \operatorname{Compose prompt}
 5:
 6:
 8:
                                                                R_i \leftarrow \text{POSTPROCESS}(\mathcal{L}(P_i)) \triangleright \text{Trace elicitation}
 9:
            \mathcal{C} \leftarrow \mathcal{C} \cup \{(P_i, R_i)\}
10:
11: end for
12: return C
```

Let \mathcal{L}_{LM} denote the language modeling loss. The fine-tuning objective is then:

$$\min_{\theta} \sum_{(P,R) \in \mathcal{R}'} \mathcal{L}_{LM}(R \mid P; \theta),$$

where θ are the parameters of the language model.

Extensibility. Although we apply the method in a clinical setting using binary outcomes, the framework is not domain-specific. The key requirement is the ability to construct meaningful contrastive pairs from structured data. In principle, the approach can be extended to non-clinical domains (e.g., finance, education) and to non-binary targets, by adapting the neighborhood selection strategy (e.g., through clustering or label binning). Exploring these extensions remains an important direction for future work.

¹Alternative distance metrics are possible - see Appendix M.

4 EXPERIMENTS

We evaluate Tables2Traces as a mechanism for transforming structured tabular data into effective reasoning supervision for LLMs. Our goal is to assess whether this supervision improves medical QA performance, and to analyze where performance differences across question types, medical domains, and benchmarks.

Data. We use a subset of the UK Biobank (Sudlow et al., 2015) comprising 105,299 individuals aged 40 and above, all diagnosed with cardiovascular disease. Each patient is represented by 32 variables spanning demographics, medication usage, lab results, and comorbidities. All data were collected under appropriate ethical approvals and informed consent (Palmer, 2007).

Setup. We evaluate two tabular supervision variants: (1) **Tables2Traces** (**simple**), which converts each row into a standalone patient narrative; and (2) **Tables2Traces**, which adds contrastive prompts using nearest-neighbor pairs. Each variant includes 90% synthetic traces and 10% QA-format examples from HPAI-BSC/MedQA-Mixtral-CoT (Gururajan et al., 2024).

We fine-tune both 8 billion parameter (8B) (DeepSeek-R1-Distill-LLaMA-8B) and 7 billion (7B) parameter (DeepSeek-R1-Distill-Qwen-7B) models using the Open-R1 framework (Face, 2025). Training configurations are detailed in Appendix B and results for 7B models are provided separately in Appendix H.

Evaluation. We evaluate on two medical QA benchmarks: MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). Accuracy is reported under four aggregation schemes (average, best-of-n, majority vote, worst-of-n). Results are averaged over 10 stochastic runs; error bars show the standard error of the mean. We compare both fine-tuned variants against a **Base** model without fine-tuning.

We also report **Aloe** (Gururajan et al., 2024), a strong medical QA system trained with resource-intensive, manually curated supervision across more than twenty datasets (e.g., synthetic chain-of-thought, guideline-based answers, adversarial/preference tuning). Importantly, Aloe is *not a reasoning model*, so it is not aligned with our supervision signal. Our approach is orthogonal: Tables2Traces derives reasoning supervision automatically from structured tabular data without manual labels. For completeness, we fine-tuned Aloe with Tables2Traces; performance did not improve (Appendix G), consistent with this misalignment.

All 8B models use the same LLaMA-8B backbone and identical inference settings, and we apply identical chain-of-thought prompts at test time (Appendix C.7-C.8). Differences therefore reflect supervision rather than prompting. As a qualitative sanity check of the supervision signal, two cardiologists independently reviewed 10 randomly sampled traces and confirmed that none received a *Concerning* safety rating (Appendix O; protocol in Appendix N).

We assess performance across the following dimensions:

- 1. **Performance:** Are gains consistent across clinical subdomains and benchmarks? Section 4.1 evaluates performance across 18 clinical categories in MedQA and 21 in MedM-CQA to identify where training is most effective and whether performance generalizes.
- 2. **Domain generalization: What types of questions benefit from tabular supervision?** Section 4.2 examines both *domain transfer* (e.g., cardiovascular → neurology) and *format transfer* (e.g., patient-specific → abstract) to determine what kinds of questions benefit most.
- 3. Locating successes and failures: Where do models succeed or fail in embedding space? Section 4.3 uses UMAP visualizations to localize model performance across semantic regions of the question embedding space.
- 4. **Upper bound comparison: How does performance compare to a QA-optimized model?** Sec. 4.4 compares our method to *Aloe*, a model trained on large-scale QA data, including the training set from both MedQA and MedMCQA. This contextualizes how far Tables2Traces can go with primarily tabular supervision compared to a task-optimized upper bound.

²We consider these smaller model sizes as these are realistic LLM sizes for fine-tuning in clinical settings given compute limitations.

Table 1: Per-category evaluation metrics on the MedQA benchmark for Base and Tables2Traces.

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Cardiovascular (N = 130)	Base Tables2Traces	0.40 ± 0.03 0.47 ± 0.03	0.86 ± 0.03 0.91 ± 0.03	0.31 ± 0.04 0.42 ± 0.04	$\begin{array}{c} 0.06 \pm 0.02 \\ 0.06 \pm 0.02 \end{array}$	+17.21% ↑
Dermatologic (N = 17)	Base Tables2Traces	0.59 ± 0.08 0.60 ± 0.08	0.94 ± 0.06 0.88 ± 0.08	0.53 ± 0.12 0.59 ± 0.12	0.06 ± 0.06 0.12 ± 0.08	+0.99%↑
Endocrine/Metabolic (N = 179)	Base Tables2Traces	0.49 ± 0.03 0.51 ± 0.02	0.89 ± 0.02 0.91 ± 0.02	0.45 ± 0.04 0.46 ± 0.04	0.13 ± 0.03 0.10 ± 0.02	+4.71% ↑
Gastrointestinal (N = 86)	Base Tables2Traces	0.47 ± 0.04 0.50 ± 0.04	0.87 ± 0.04 0.91 ± 0.03	0.40 ± 0.05 0.47 ± 0.05	0.12 ± 0.04 0.08 ± 0.03	+6.72% ↑
Hematologic (N = 68)	Base Tables2Traces	0.40 ± 0.04 0.48 ± 0.04	0.84 ± 0.04 0.91 ± 0.04	0.34 ± 0.06 0.43 ± 0.06	0.04 ± 0.03 0.07 ± 0.03	+18.98% ↑
Immunologic (N = 81)	Base Tables2Traces	0.51 ± 0.04 0.54 ± 0.04	0.85 ± 0.04 0.94 ± 0.03	0.47 ± 0.06 0.46 ± 0.06	0.22 ± 0.05 0.17 ± 0.04	+6.80% ↑
Infectious (N = 176)	Base Tables2Traces	0.48 ± 0.03 0.53 ± 0.02	0.92 ± 0.02 0.94 ± 0.02	0.41 ± 0.04 0.45 ± 0.04	$0.11 \pm 0.02 \\ 0.11 \pm 0.02$	+9.73% ↑
Musculoskeletal (N = 45)	Base Tables2Traces	0.49 ± 0.05 0.51 ± 0.04	0.89 ± 0.05 0.96 ± 0.03	0.49 ± 0.07 0.40 ± 0.07	0.04 ± 0.03 0.07 ± 0.04	+4.07% ↑
Neurological (N = 77)	Base Tables2Traces	0.47 ± 0.04 0.54 ± 0.04	0.86 ± 0.04 0.95 ± 0.03	0.42 ± 0.06 0.48 ± 0.06	0.09 ± 0.03 0.14 ± 0.04	+15.15% ↑
Obstetrics/Gynecology (N = 70)	Base Tables2Traces	0.46 ± 0.04 0.47 ± 0.03	0.90 ± 0.04 0.94 ± 0.03	0.39 ± 0.06 0.40 ± 0.06	0.09 ± 0.03 0.03 ± 0.02	+2.80% ↑
Oncology (<i>N</i> = 72)	Base Tables2Traces	0.53 ± 0.04 0.56 ± 0.04	0.92 ± 0.03 0.93 ± 0.03	0.47 ± 0.06 0.53 ± 0.06	0.11 ± 0.04 0.14 ± 0.04	+5.82% ↑
Other (N = 31)	Base Tables2Traces	0.53 ± 0.07 0.50 ± 0.07	0.77 ± 0.08 0.87 ± 0.06	0.45 ± 0.09 0.42 ± 0.09	0.23 ± 0.08 0.19 ± 0.07	-4.88%↓
Pediatric (<i>N</i> = 13)	Base Tables2Traces	$0.39 \pm 0.09 \\ 0.39 \pm 0.05$	0.77 ± 0.12 1.00 ± 0.00	0.39 ± 0.14 0.31 ± 0.13	$\begin{array}{c} 0.00 \pm 0.00 \\ 0.00 \pm 0.00 \end{array}$	-1.96%↓
Psychiatric (N = 52)	Base Tables2Traces	0.59 ± 0.05 0.62 ± 0.05	0.94 ± 0.03 0.90 ± 0.04	0.54 ± 0.07 0.61 ± 0.07	0.23 ± 0.06 0.21 ± 0.06	+5.57% ↑
Renal/Genitourinary (N = 54)	Base Tables2Traces	0.37 ± 0.04 0.48 ± 0.04	0.85 ± 0.05 0.96 ± 0.03	0.26 ± 0.06 0.41 ± 0.07	0.04 ± 0.03 0.09 ± 0.04	+29.65% ↑
Respiratory (N = 54)	Base Tables2Traces	0.49 ± 0.04 0.50 ± 0.04	0.91 ± 0.04 0.94 ± 0.03	0.43 ± 0.07 0.46 ± 0.07	0.09 ± 0.04 0.11 ± 0.04	+2.28% ↑
Toxicology (N = 68)	Base Tables2Traces	0.43 ± 0.04 0.52 ± 0.04	0.79 ± 0.05 0.91 ± 0.04	0.41 ± 0.06 0.47 ± 0.06	0.06 ± 0.03 0.09 ± 0.04	+20.68%↑
Overall (<i>N</i> = 1273)	Base Tables2Traces	0.47 ± 0.01 0.51 ± 0.01	0.88 ± 0.01 0.93 ± 0.01	0.41 ± 0.01 0.46 ± 0.01	0.11 ± 0.01 0.10 ± 0.01	+9.19% ↑

4.1 ARE GAINS CONSISTENT ACROSS CLINICAL SUBDOMAINS AND BENCHMARKS?

Goal. Assess whether improvements from tabular supervision generalize across diverse clinical categories, and whether these gains hold across both MedQA and MedMCQA benchmarks.

Setup. For MedQA, we assign each question to one of 18 clinical categories using DeepSeek-R1. For MedMCQA, we use the dataset assigned category labels and evaluate on the public validation set (as the test set is not accessible). We restrict evaluation to questions with a single correct answer to ensure consistency with the MedQA setup and to allow for accurate, per-question performance analysis.

Results. As shown in Table 1 and Appendix E, Tables2Traces consistently outperforms the base model across a majority of clinical categories in both datasets. On MedQA, we observe improvements in 16 out of 18 categories (89%), with the largest relative gains in *Renal/Genitourinary* (+29.65%), *Hematologic* (+18.98%), and *Cardiovascular* (+17.21%). On MedMCQA, Tables2Traces improves performance in 17 out of 21 categories (81%), including strong gains in *Psychiatry* (+31.71%), *ENT* (+16.56%), and *Anatomy* (+11.91%). Categories showing drops in performance e.g., *Skin* (-24.39%) and *Orthopaedics* (-15.00%) have very few test samples (N=11 and N=15), making these estimates statistically unreliable.

Takeaway. Tables2Traces yields robust, cross-domain improvements on both benchmarks, improving in over 80% of clinical categories. The few observed declines are isolated to small and noisy subsets.

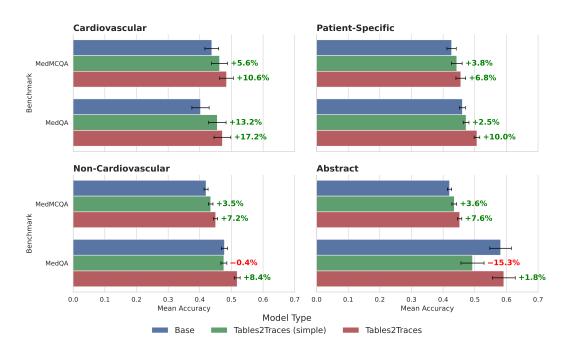


Figure 1: Accuracy comparison across model variants. Subplots show mean accuracy on MedQA and MedMCQA, with error bars for standard error. Percentages indicate improvement over the Base model. Top row: in-domain (cardiovascular, patient-specific); bottom row: out-of-domain (non-cardiovascular, abstract).

4.2 What types of questions benefit from tabular supervision?

Goal. To assess whether supervision derived from a single clinical domain (cardiovascular) and a single input style (patient-specific reasoning traces) can improve performance on both out-of-domain specialties and abstract medical knowledge.

Setup. We assess generalization along two axes: (1) **Domain**—partitioning questions into *cardiovas-cular* vs. *non-cardiovas-cular* using model-inferred labels; and (2) **Format**—classifying questions as *patient-specific* or *abstract*, based on whether they describe concrete cases (e.g., 45-year-old man) or general concepts. MedQA is predominantly patient-specific (92.3%), while MedMCQA is mostly abstract (83.7%) (see Appendix F).

Results. In MedQA, Tables2Traces shows strong gains on in-domain (cardiovascular, +17.2%) and patient-specific questions (+10.0%), with smaller gains on abstract questions (+1.8%) (Figure 1). However, Tables2Traces (simple) performs worse on abstract questions (-15.3%), possibly suggesting overfitting to training format. In MedMCQA, we observe broader generalization: Tables2Traces improves both patient-specific (+6.8%) and abstract (+7.6%) subsets.

Takeaway. Tabular supervision supports generalization beyond its source domain and format. Without counterfactual reasoning, tabular supervision can overfit to its patient-specific training data, resulting in performance degradation in out-of-domain tasks. With counterfactual reasoning, Tables2Traces improves performance even on abstract, unfamiliar questions.

4.3 Where do models succeed or fail in embedding space?

Goal. Visualize how supervision strategies impact the semantic generalization of medical questions.

Setup. We embed all MedQA and MedMCQA questions using text-embedding-3-large (OpenAI, 2023) and reduce dimensionality via UMAP. Each point corresponds to a question, colored by clinical category. Background shading shows smoothed relative accuracy of the fine-tuned model compared to the **Base** model. Figure 2 shows two panels: (a) **Tables2Traces**, and (b) **Tables2Traces** (simple). A complementary plot for MedMCQA is provided in Appendix E.

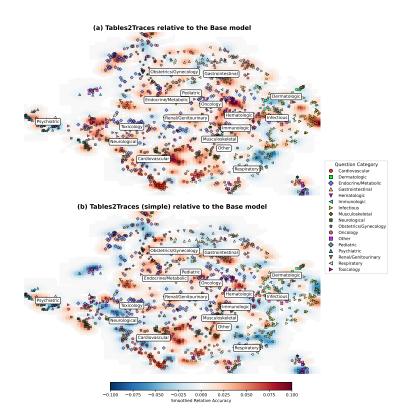


Figure 2: UMAP visualization of MedQA test questions comparing model performance to the **Base** model. Each point is a question, embedded using text-embedding-3-large (OpenAI, 2023), and annotated by medical category using distinct marker shapes and colors. Accuracy is smoothed using a Gaussian filter ($\sigma=1.5$). The background heatmap shows relative performance: red indicates improvement, blue indicates degradation. Cluster labels mark category centroids. (a) Tables2Traces achieves broad gains across much of the question space. (b) Tables2Traces (simple) yields localized improvements, but also shows notable drops in performance in several regions.

Results. Both models show localized gains within the cardiovascular region. However, only Tables2Traces generalizes effectively across distant clusters whereas the patient-style model (b) overfits to regions that closely resemble its training format. Peripheral zones, often containing abstract or non-patient-centered questions (e.g., *Biochemistry*, *Social Medicine*), show degradation under the simple model but improved performance under Tables2Traces. These patterns mirror our quantitative results and extend to the MedMCQA visualization, where we observe similar trends in spatial generalization behavior.

Takeaway. Contrastive reasoning traces lead to broader semantic generalization, increasing performance across diverse question types and topics. In contrast, models trained only on patient description data tend to overfit and struggle in abstract or semantically distant regions of the question space.

4.4 How does performance compare to a QA-optimized model?

Goal. To benchmark Tables2Traces against a strong QA-tuned model, we compare it to **Aloe**—a state-of-the-art medical LLM trained on over 750,000 QA pairs from datasets including MedQA and MedMCQA, and further optimized through preference and adversarial feedback tuning (Gururajan et al., 2024). Rather than a direct competitor, Aloe represents a *task-optimized upper bound* built from large-scale QA supervision. Our comparison asks how far Tables2Traces can go using primarily tabular reasoning traces and only a small amount of QA data.

Setup. We evaluate **Base**, **Tables2Traces**, and **Aloe** across four subsets: *cardiovascular*, *non-cardiovascular*, *patient-specific*, and *abstract*, using both MedQA and MedMCQA (Figure 3). Aloe is

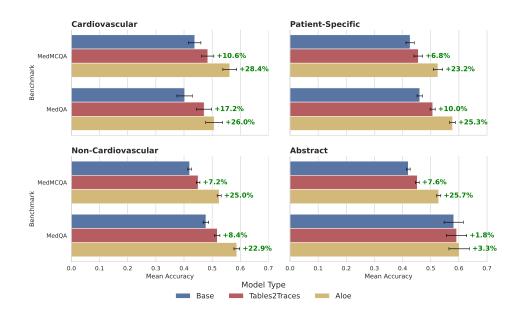


Figure 3: Accuracy comparison between **Base**, **Tables2Traces**, and **Aloe** across question categories, grouped by benchmark (MedMCQA, MedQA). Each subplot reports mean accuracy with standard error bars. Percentage improvements are relative to the Base model. Top row: in-domain categories. Bottom row: out-of-domain generalization.

trained with 750k QA-format samples, while Tables2Traces uses only 10k QA-format examples (for alignment), relying primarily on 105k reasoning traces generated from tabular data.

Results. As expected, Aloe achieves the highest accuracy in all categories. However, Tables2Traces closes a substantial portion of the gap despite using only 1.3% of the QA supervision data used by Aloe. On MedQA, Tables2Traces improves +17.2% on cardiovascular questions (vs. Aloe's +26.0%) and +9.6% on patient-specific questions (vs. +25.3%). In MedMCQA, Tables2Traces achieves +10.6% and +6.8% improvements on cardiovascular and patient-specific questions respectively. Notably, even Aloe shows limited gains on abstract, non-patient-specific questions across both benchmarks, suggesting these are structurally more challenging and underrepresented during training.

Takeaway. Tables2Traces achieves strong generalization with minimal QA supervision. Despite being trained using 75× fewer QA samples compared to Aloe, our approach closes a substantial portion of the performance gap. This highlights the value of clinical reasoning traces from tabular data as a scalable, interpretable, and cost-effective alternative to large-scale QA corpora.

5 Discussion

Tables2Traces offers a scalable approach for converting tabular data into contrastive reasoning traces, improving LLM performance on medical QA even when trained on a single clinical domain. Trace generation is a one-time preprocessing step; once the corpus is produced, it can be reused across models, and in practice fine-tuning dominates the total compute. By leveraging structured tabular medical data, the method introduces a new and broadly applicable source of supervision that promotes causal and actionable reasoning.

While our focus is on binary outcomes and cardiovascular data, the approach is modular and adaptable to other settings. Generalization to multi-class labels or other data types may require modest adjustments to the contrastive sampling process. Our evaluation is based on two established medical QA benchmarks and a large-scale clinical dataset; assessing performance in other domains or institutions is a natural next step. Finally, while we observe consistent improvements, evaluating trace fidelity in real-world settings remains an important direction for future work. More broadly, while tabular supervision can improve access to data used for domain adaptation, it also raises risks if synthetic traces reflect dataset bias or are applied without validation.

ETHICS STATEMENT

This work uses de-identified UK Biobank data accessed under approved use; all participants provided informed consent and data collection was overseen by the UK Biobank ethics framework. Our method, Tables2Traces, generates synthetic reasoning traces from structured records to fine-tune language models for research purposes only. The models and traces are *not* clinical devices and must not be used for diagnosis or treatment. To gauge plausibility and safety, two cardiologists qualitatively reviewed 10 randomly sampled traces independently using a structured rubric. Clinician review confirmed no safety concerns but did note overconfidence by the model, reflecting the inherent limitations of synthetic data. The cardiology experts noted that outcomes may depend on factors not present in the tabular snapshot; our traces are therefore positioned as research-only supervision signals, not calibrated risk assessments or clinical guidance. We provide an overview of their comments in Appendix O. We release prompts and code to support auditability. Finally, our evaluation is restricted to public medical QA benchmarks and does not involve individual-level deployment or decision support.

REPRODUCIBILITY STATEMENT

All implementation details, prompts, hyperparameters, and evaluation procedures are documented in the Appendix. Upon acceptance we will release the full codebase and configs to reproduce preprocessing, trace generation, fine-tuning, and evaluation, together with exact seeds and scripts that render all tables and figures. Results on public benchmarks (MedQA, MedMCQA) are reproducible with our released scripts and seeds. UK Biobank data cannot be shared; researchers with approved access can regenerate the training traces using our scripts and instructions.

REFERENCES

- Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.
- Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.
- Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.
- Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, et al. Aloe: A family of fine-tuned open healthcare llms. *arXiv* preprint arXiv:2405.01886, 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 *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
 - Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
 - Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv* preprint arXiv:2101.00190, 2021.
 - Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*, 2021.
 - OpenAI. Openai text-embedding-3-large, 2023. https://platform.openai.com/docs/guides/embeddings.
 - Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
 - Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.
 - Lyle J Palmer. Uk biobank: bank on it. The Lancet, 369(9578):1980–1982, 2007.
 - Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*, 2020.
 - Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
 - Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
 - Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015.
 - Ruiyu Wang, Zifeng Wang, and Jimeng Sun. Unipredict: Large language models are universal tabular classifiers. *arXiv preprint arXiv:2310.03266*, 2023a.
 - Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, 2023b.
 - Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022a.
 - Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.
 - Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, 2020.

APPENDIX

A EVALUATION SETUP

We use a standardized evaluation pipeline across all models and benchmarks. Each multiple-choice question is formatted using the appropriate chat template (e.g., using standard templates like AutoTokenizer.apply_chat_template) and fed into the model for completion. Evaluation is performed using the vLLM framework with sampling-based generation (temperature=0.6, top_p=0.85, n=3 completions per prompt, frequency_penalty=1.5, presence_penalty=0.9, max_tokens=32768). We extract the final answer (A-D) from the generated output using robust regex-based parsing, and fall back to the reasoning text if a clean answer is not present after multiple attempts.

We stop generation using model-specific stop tokens (e.g., </s> for LLaMA, <|EOT|> for Qwen), as well as answer-format strings (e.g., "Answer: A"). All completions are post-processed using a training-aware cleaner to remove template artifacts (e.g., "Assistant: "headers). For models fine-tuned on structured reasoning traces, we additionally parse the <think>...</think> block and extract the final prediction from the trailing answer segment.

The pipeline includes automatic retries for failed generations, and safely extracts answers even under high sampling variability. This setup ensures consistent evaluation across all models and supports multi-sample decoding strategies such as best-of-n, majority vote, and worst-of-n.

A.1 EVALUATION METRICS

All performance metrics are aggregated from 10 independent inference runs per model. For each test question, we collect a binary correctness label (extracted using a robust regex-based parsing) for each of the 10 completions and compute the following evaluation metrics:

- Average Accuracy: The average correctness across the 10 runs for each question.
- **Best-of-***n*: The question is marked correct if at least one of the 10 completions is correct.
- **Majority Vote**: The question is marked correct if a majority of the 10 completions are correct. In the case of a tie, the outcome defaults to incorrect.
- Worst-of-n: The question is marked correct only if all 10 completions are correct.

Category-level and overall scores are computed by averaging across all test questions per category. Error bars represent the standard error of the mean (SEM) across test examples. Additionally, we report the relative percent change in average accuracy compared with the Base model. In all results tables, the best-performing model is shown in bold for each metric within each category. If multiple models have the same value after rounding, all are shown in bold.

B TRAINING CONFIGURATION DETAILS

All models are fine-tuned using Open-R1's supervised fine-tuning pipeline (Face, 2025), with a single epoch of training on 4×A100 80GB GPUs. We use FlashAttention-2 and bfloat16 precision for all experiments. Below, we describe shared configurations and model-specific differences.

B.1 Shared Configuration

- **Precision:** bfloat16 with FlashAttention-2
- Epochs: 1 full pass over the training set
- Batch Size: 2 per device, 8 gradient accumulation steps
- **Optimizer:** AdamW with learning rate = 5e-6, cosine decay (min LR ratio = 0.1), weight decay = 0.0001
- Max Sequence Length: 32,768 tokens
- Evaluation: Every 500 steps on the test split
- Checkpointing: Saved every 500 steps, keep only latest
- Logging: Via wandb, every 5 steps
- Seed: 42
- **Gradient Checkpointing:** Enabled (non-reentrant)
- System Prompt:

```
You are a helpful AI Assistant that provides well-reasoned and detailed responses.

You first think about the reasoning process as an internal monologue and then provide the user with the answer.

Respond in the following format:

<think> ... </think>
<answer> ... </answer>
```

• Chat Template: Modified to include reasoning tags (<think>...</think>) in the completion and exclude them from the prefill.

B.2 MODEL VARIANTS

We fine-tune two architectures on two dataset variants, resulting in four total models:

Model Architecture	Training Data
DeepSeek-R1-Distill-Qwen-7B	Patient Descriptions (Tables2Traces (simple))
DeepSeek-R1-Distill-Qwen-7B	Counterfactual Traces (Tables2Traces)
DeepSeek-R1-Distill-Llama-8B	Patient Descriptions (Tables2Traces (simple))
DeepSeek-R1-Distill-Llama-8B	Counterfactual Traces (Tables2Traces)

The **patient descriptions** dataset consists of direct narrative renderings of individual tabular rows, while the **counterfactual traces** dataset includes contrastive triplets with structured reasoning (as described in Section 3). All datasets are processed using 48 parallel workers.

Table 2: Training runtimes for each model variant.

Model Variant	Architecture	Runtime
Tables2Traces Tables2Traces (simple)	8B (LLaMA) 8B (LLaMA)	20h 37m 9h 24m
Tables2Traces (simple)	7B (Qwen)	19h 52m
Tables2Traces (simple)	7B (Qwen)	9h 18m

C PROMPT TEMPLATES

This section documents all prompt templates used during dataset construction, training, and evaluation. Strings enclosed in curly brackets (e.g., {column_names}) represent placeholders that are dynamically replaced with instance-specific values at runtime, similar to Python f-strings.

C.1 COLUMN NAME MAPPING (TABLE REPRESENTATION)

Purpose: Transform raw or abbreviated column headers into clinically accurate feature names.

Placeholders: column_names is replaced with a list of all columns of the dataset.

Column Name Mapping

You are a powerful AI with expertise in medicine.

You are given a dataset with columns that relate to patients where each patient is a row and each column contains different information pertaining to the patient.

As your first task, you are tasked with converting a list of column names that are possibly abbreviated or not easy to understand into a fully understandable name for medical professionals.

Please provide the output as a Python dictionary.

The list of column names is: {column_names}

C.2 PATIENT DESCRIPTION GENERATION

Purpose: Convert structured patient rows into fluent narrative case descriptions.

Placeholders: json_file is replaced with a json-file containing the column names as keys and the values of columns as values.

758

759

760

761

762

763

764

765 766

767 768 769

770 771

772

773

774

775

776

777

778 779

781

782 783

784

785

786

787

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803 804

805

807

809

```
Patient Description Generation

You are a powerful AI with expertise in medicine.
Your task is to generate a detailed and exhaustive text description for a patient.
You are given all the patient information in a json-format, which contains the clinical attributes and the results from laboratory tests from real world patients.
The patients in question are patients with cardiovascular disease.
The reader of the description is an expert within this particular medical domain.
The language used in the description should reflect your domain expertise and your medical reasoning capabilities.
Please provide as many details as possible.
You should ONLY include the patient description!

The json-file containing the information from the patient: {json_file}
```

C.3 CONTRASTIVE REASONING AND COUNTERFACTUAL TRACES

Purpose: Generate reasoning traces comparing a target patient to contrasting neighbors.

Placeholders: target_outcome is the outcome (Alive / Dead) for the target patient. survivor_description is the text description of the nearest neighbor to the target patient who had the outcome "Alive". survivor_description is the text description of the nearest neighbor to the target patient who had the outcome "Dead". target_description is the text description target patient. All text descriptions are derived using the Patient Description Generation prompt in Appendix C.2.

```
### Role ###
Clinical AI analyzing patient outcomes using contrastive case pairs.
### Input Data ###
1) Target patient (labeled {target_outcome})
2) Nearest neighbor who survived
3) Nearest neighbor who died
=== CLOSEST SURVIVOR ===
{survivor description}
=== CLOSEST DEATH ===
{death_description}
=== TARGET PATIENT ===
{target_description}
### Required Analysis ###
1. Comparison:
   a) Identify 1-3 decisive differences between target and NNs
   b) Focus on features present in ALL THREE cases
  c) Flag any conflicting evidence
2. Label Evaluation:
   a) Assess if {target outcome} is correct
   b) Confidence score (1-5)
3. Counterfactual:
   a) Modify one feature present in NNs
   b) Predict outcome change
   c) Justify using specific NN evidence
### Response Format ###
1. Comparison:
   1) Outcome alignment: <...>
   2) Decisive factors: ...
2. Label assessment:
   1) Correctness: <...>
   2) Confidence: <...>
3. Counterfactual:
   1) Modification: <...>
   2) Outcome: <...>
   3) Evidence: <...>
```

C.4 CATEGORIZATION: PATIENT-SPECIFIC VS. ABSTRACT

Purpose: Categorize questions as either patient-specific or abstract.

Placeholders: question is the specific question to be categorized.

```
You are a clinical reasoning expert.
Your task is to determine whether a multiple-choice medical question is *patient-specific*
Definitions:
- A question is **patient-specific** if it describes
a particular patient case -- including their symptoms, medical history, age, lab results, etc.
These questions simulate real-life clinical decision-making.
- A question is **not patient-specific** if it
asks about general medical knowledge or includes references to people (e.g., doctors, nurses)
but *not to a patient's condition*.
Return: {{"patient_specific": true}} or {{"patient_specific": false}}
Examples:
Example 1:
Question: A 67-year-old man presents with sudden chest pain and shortness of breath.
Which of the following is the most likely diagnosis?
Answer: {{"patient_specific": true}}
Example 2:
Question: What is the most common cause of mitral stenosis worldwide?
Answer: {{"patient_specific": false}}
Question: A physician enters the operating room without washing his hands.
What is the correct protocol in this situation?
Answer: {{"patient_specific": false}}
Now classify the following question:
Question: {question}
```

C.5 CATEGORIZATION: CARDIOVASCULAR VS. NON-CARDIOVASCULAR (MEDMCQA)

Purpose: Categorize MedMCQA questions as either cardiovascular related or not.

Placeholders: question is the specific question to be categorized.

```
Cardiovascular Categorization

You are a medical assistant helping categorize medical questions.
Given a question, determine whether it pertains to cardiovascular diseases or not.

Only answer 'true' or 'false' depending on whether the core topic of the question involves cardiovascular systems, diseases, symptoms, diagnostics, or treatment.

Cardiovascular topics include (but are not limited to) conditions such as: hypertension, myocardial infarction, arrhythmias, heart failure, atherosclerosis, angina, or cardiac arrest.

Avoid false positives: questions mentioning blood pressure, heart rate, or medications like beta-blockers must still be relevant to cardiovascular context to count.

Output your answer in the following JSON format:

{{"cardiovascular_related": true}}

Question:
{question}
```

C.6 CATEGORIZATION: MEDICAL DOMAIN (MEDQA)

Purpose: Categorize the medical domain of MedQA questions.

Placeholders: question is the specific question to be categorized.

```
Analyze the medical question and respond EXACTLY as follows:
---STRICT RULES---
1. SINGLE HIGH-CONFIDENCE CATEGORY (>=0.7):
   • If ONE category scores >=0.7:
   "Category = Score"
   • If MULTIPLE categories score >=0.7:
   Choose ONLY THE HIGHEST SCORE (if tie, pick first alphabetically)
2. MULTIPLE LOW-CONFIDENCE CATEGORIES (all <0.7):
   • "Primary: Category1 = Score1, Secondary: Category2 = Score2, Tertiary: Category3 = Score3"
3. IRRELEVANT:
   • "None of the above = 1.0"
---VALID EXAMPLES---
• "Cardiovascular = 0.85"

    "Primary: Infectious = 0.6, Secondary: Hematologic = 0.3, Tertiary: Renal = 0.1"
    "None of the above = 1.0"

---CATEGORIES (ALPHABETICAL ORDER) ---
Cardiovascular, Dermatologic, Endocrine/Metabolic, Gastrointestinal, Hematologic,
Immunologic, Infectious, Musculoskeletal, Neurological, Obstetrics/Gynecology,
Oncology, Pediatric, Psychiatric, Renal/Genitourinary, Respiratory, Toxicology
---OUESTION---
{question}
---YOUR RESPONSE (MUST MATCH EXACTLY ONE FORMAT ABOVE) ---
```

C.7 EVALUATION PROMPTS: MEDQA

Purpose: Evaluate model on MedQA using reasoning-aware prompting.

```
**Role**: You are a medical knowledge expert.

**Task**: Analyze the following multiple-choice medical question by following these steps:

1. First, use critical clinical reasoning to think about the question step-by-step before giving a final answer.

2. After completing your reasoning, directly provide your final answer.

3. **IMPORTANT**: Do not provide any explanation beyond your answer in the final output.

**Response Format**:
[Your step-by-step reasoning goes here]
Answer: [Your final choice: A, B, C, or D]
```

C.8 EVALUATION PROMPTS: MEDMCQA

Purpose: Evaluate model on MedMCQA using structured reasoning steps.

```
MedMCQA Evaluation Prompt

**Role**: You are a medical knowledge expert.
**Task**: Analyze the following multiple-choice medical question by reasoning step-by-step before selecting the best answer.

Follow these steps:
1. Identify the topic and relevant concepts.
2. Recall or infer medical knowledge needed to solve the question.
3. Eliminate incorrect options and justify your final choice.
4. Provide your answer clearly and concisely.

**Response Format**:
[Step-by-step reasoning]
Answer: [A, B, C, or D]

**Constraints**:
- Do not explain anything after giving your final answer.
- Only choose a single option from A-D.
```

D QUALITATIVE EXAMPLES

972

973974975976

977

978

979

980

981

991

992

998 999

1000

1001

1002

1003

1004

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1023

1024

1025

This section provides representative examples of reasoning traces generated during training and evaluation. These examples illustrate the structure, fluency, and clinical plausibility of model outputs for both Tables2Traces and Tables2Traces (simple). All examples are drawn directly from the training data without curation or filtering. Specifically, the example shown here (both the patient description task and the counterfactual task) corresponds to the first available row in the dataset and is included with minimal editing beyond redaction for privacy.

D.1 QUALITATIVE EXAMPLE: PATIENT DESCRIPTION REASONING (REDACTED)

We include a representative example of a reasoning trace generated from a single patient description prompt. The input data has been redacted for privacy, but the structure of the trace is preserved. The resulting model response corresponds to the training data used for the Tables2Traces (simple) variant and as the generated patient descriptions used for generating the counterfactual tasks.

Patient JSON Input (Redacted)

You are a powerful AI with expertise in medicine. Your task is to generate a detailed and exhaustive text description for a patient. You are given all the patient information in a json-format, which contains the clinical attributes and the results from laboratory tests from real world patients. The patients in question are patients with cardiovascular disease. The reader of the description is an expert within this particular medical domain. The language used in the description should reflect your domain expertise and your medical reasoning capabilities. Please provide as many details as possible. You should ONLY include the patient description!

```
within this particular medical domain. The language used in the description should reflect your domain expertise and your medical
reasoning capabilities. Please provide as many details as possible. You should ONLY include the patient description!
The json-file containing the information from the patient:
 Sex
 Age (Years)
 Weight (Kilograms)
 Height (Centimeters)
 Smoking Amount (Cigarettes per Day)
 Atrial Fibrillation Diagnosis
 Chronic Kidney Disease Diagnosis
 Rheumatoid Arthritis Diagnosis
 Anti-diabetic Medication Use
                                                                     . . .
 Anti-hypertensive Medication Use
 Diabetes History
                                                                     . . .
 Lipid-lowering Medication Use
 Oral Contraceptive Use
 Blood Glucose Level (mmol/L)
 Hemoglobin Alc (HbAlc) Percentage
 White Blood Cell Count (x10^9/L)
  Serum Creatinine (µmol/L)
                                                                     . . .
 Triglyceride Level (mmol/L)
 Uric Acid Level (µmol/L)
                                                                     . . .
 Cystatin C Level (mg/L)
  Systolic Blood Pressure (mmHg)
                                                                     . . .
 Urine Microalbumin Concentration (mg/L)
                                                                     . . .
 C-Reactive Protein (CRP) Level (mg/L)
                                                                     . . .
 Family History of Cardiovascular Disease (CVD)
                                                                     . . .
 Atypical Antipsychotic Medication Use
                                                                     . . .
  Corticosteroid Medication Use
 Migraine History
                                                                     . . .
  Severe Mental Illness Diagnosis
  Systemic Lupus Erythematosus (SLE) Diagnosis
                                                                     . . .
 Total Cholesterol Level (mmol/L)
  High-Density Lipoprotein (HDL) Cholesterol Level (mmol/L)
 Ethnicity
```

Model Reasoning (Redacted)

<think> Okay, let's start by going through all the patient data provided. The patient is a [...]-year-old [...] [...] with a height of [...] cm and weight of [...] kg. Calculating his BMI: weight divided by height squared in meters. So [...] kg / ([...] m)² = [...] kg/m². That puts him in the overweight category, which is a risk factor for cardiovascular issues.

He's a smoker, [...] cigarettes per day. That's significant. Smoking is a major risk factor for cardiovascular diseases, including atherosclerosis and hypertension. His systolic blood pressure is [...] mmHg, which is way above the normal range. Stage 2 hypertension. But he's not on anti-hypertensive meds, which is concerning. Uncontrolled hypertension can lead to heart failure, stroke, etc.

Looking at his lab results. Blood glucose is [...] mmol/L, which is within normal fasting levels. HbA1c is [...]%, also normal, so no indication of diabetes. No history of diabetes either. So glucose metabolism seems okay. But his lipid profile: total cholesterol [...] mmol/L, which is a bit elevated. Triglycerides at [...] mmol/L, which is borderline high. HDL is [...] mmol/L, which is on the lower side for males. Low HDL and elevated triglycerides could contribute to atherosclerosis. But he's not on lipid-lowering meds, so that's another risk factor not being managed.

Serum creatinine is [...] µmol/L. Using the CKD-EPI equation for eGFR. For a [...]-year-old [...], serum creatinine of [...] converts to an eGFR around [...] mL/min/[...]m², which is normal. Cystatin C is [...] mg/L, which also correlates with normal kidney function. Urine microalbumin is [...] mg/L, which is within normal range. So no signs of chronic kidney disease, which aligns with the CKD diagnosis being [...].

CRP is [...] mg/L, which is normal. Low CRP suggests no significant inflammation. No rheumatoid arthritis, SLE, or other inflammatory conditions. Uric acid is [...] μ mol/L. Normal range for males is around 416, so his is within normal. Not a hyperuricemia case, so less concern for gout.

He has atrial fibrillation (diagnosis 1), which is a major risk for stroke and heart failure. Given his age and hypertension, that's a critical factor. No family history of CVD, so maybe environmental factors like smoking and hypertension are the main contributors here. No other comorbidities like severe mental illness or migraines.

Other meds: Not on anti-diabetics, lipid-lowering, or antihypertensives. So his AFib and hypertension are untreated, which is alarming. His high systolic BP could be exacerbating the atrial fibrillation, increasing stroke risk. The CHA_2DS_2 -VASc score would be [...], indicating anticoagulation should be considered.

In summary, this patient has multiple modifiable risk factors: smoking, obesity, uncontrolled hypertension, dyslipidemia. The presence of atrial fibrillation without anticoagulation or rate control meds significantly elevates his thromboembolic risk. The absence of diabetes or kidney disease is a positive note, but the untreated hypertension and smoking are critical areas needing intervention. His elevated systolic BP is a key driver for cardiovascular events. Management should focus on smoking cessation, initiating antihypertensive therapy, lipid management, and evaluating for anticoagulation given his AFib.

Model Response (Redacted)

The patient is a [...]-year-old [...] [...] with a body mass index (BMI) of [...] kg/m^2 (height: [...] cm; weight: [...] kg), categorizing him as overweight. He reports a significant smoking history of [...] cigarettes/day, a critical modifiable cardiovascular risk factor. Key clinical findings include **stage [...] hypertension** (systolic blood pressure: [...] mmHg) without current anti-hypertensive therapy, **atrial fibrillation** (diagnosis confirmed), and a lipid profile suggestive of mixed dyslipidemia (total cholesterol: [...] mmol/L, triglycerides: [...] mmol/L, HDL-C: [...] mmol/L). Notably, his LDL-C (calculated via Friedewald equation: [...] mmol/L) exceeds optimal targets, yet he is not on lipid-lowering therapy.

Metabolically, he exhibits normoglycemia (fasting glucose: [...] mmol/L; HbA1c: [...]%) without evidence of diabetes or anti-diabetic medication use. Renal function is preserved (serum creatinine: [...] µmol/L; eGFR >[...] mL/min/[...]m² by CKD-EPI; cystatin C: [...] mg/L; urine microalbumin: [...] mg/L), ruling out chronic kidney disease. Inflammatory markers are unremarkable (CRP: [...] mg/L), with no history of autoimmune or systemic inflammatory conditions. Uric acid levels ([...] µmol/L) are within normal limits, mitigating concern for gout-related contributions.

The absence of anti-hypertensive or anticoagulant therapy in the setting of atrial fibrillation and severe hypertension raises immediate concern for thromboembolic risk (CHA_2DS_2 -VASc score \leq [...]) and hypertensive end-organ damage. His lipid abnormalities—elevated triglycerides, low HDL-C, and borderline-high LDL-C—suggest residual atherogenic risk despite normal BMI-adjusted adiposity metrics. The combination of smoking, untreated hypertension, and dyslipidemia synergistically amplifies his cardiovascular risk profile, warranting aggressive risk factor modification.

No family history of CVD or secondary causes of hypertension (e.g., renal, endocrine) is documented. The absence of diabetes, chronic kidney disease, or inflammatory comorbidities narrows the primary drivers of his cardiovascular pathology to lifestyle factors (smoking, diet) and age-related vascular stiffening. Urgent interventions should prioritize smoking cessation, BP control (target <130/80 mmHg per ACC/AHA guidelines), statin initiation, and atrial fibrillation management with anticoagulation and rhythm control strategies.

D.2 QUALITATIVE EXAMPLE: COUNTERFACTUAL REASONING (REDACTED)

We include a representative example of a reasoning trace generated from a counterfactual reasoning trace for the first row of the dataset. The input data has been redacted for privacy, but the structure of the trace is preserved. This format corresponds to the Tables2Traces variant. For readability, the input to the counterfactual task is shown in three parts. In practice, the full prompt is passed to the LLM as a single string. The split here is purely for presentation purposes.

```
1080
            Counterfactual Input (Redacted) - Part 1
1081
1082
             ### Role ###
            Clinical AI analyzing patient outcomes using contrastive case pairs.
1083
1084
            ### Input Data ###
            1) Target patient (labeled Dead)
1085
            2) Nearest neighbor who survived
1086
            3) Nearest neighbor who died
1087
            === CLOSEST SURVIVOR ===
1088
            **Patient Description**
            The patient is a **[...]-year-old [...] male**
1089
            with a body mass index (BMI) of **[...] kg/m^2** (weight: [...] kg,
1090
            height: [...] cm)), categorizing him as **overweight**,
            a significant modifiable risk factor for cardiovascular disease (CVD).
1091
            He reports a **[...]-cigarette/day smoking history**, a major independent risk factor
1092
            for atherosclerotic CVD and thromboembolic events.
1093
            **Cardiovascular and Comorbidity Profile**:
1094
              **Atrial fibrillation (AF) ** is confirmed (diagnosis code present),
            elevating his risk of thromboembolic complications, including stroke.
1095
             - **No diabetes mellitus** (HbA1c: [...]*, fasting glucose: [...] mmol/L))
1096
            or chronic kidney disease (CKD) (serum creatinine: [...] mumol/L,
            cystatin C: [...] mg/L, urine microalbumin: [...] mg/L).
1097
              **Uncontrolled hypertension** (systolic BP: [...] mmHg) is evident, with no current use
1098
            of anti-hypertensive medications, suggesting suboptimal risk factor management.
             - **Hyperlipidemia** is present (total cholesterol: [...] mmol/L, HDL: [...] mmol/L,
1099
            triglycerides: [...] mmol/L), with an estimated LDL-C of **~[...] mmol/L** (Friedewald equation), indicative of significant dyslipidemia.
1100
            Despite this, no lipid-lowering therapy is documented.
1101
1102
            **Inflammatory and Metabolic Markers**:
            - **C-reactive protein (CRP) ** is within normal limits ([...] mg/L),
1103
            suggesting no acute systemic inflammation.
1104
             - **Uric acid** levels are borderline elevated ([...] mumol/L),
            though below the threshold for clinical hyperuricemia.
1105
1106
            **Additional Risk Stratification**:
            - **No family history of CVD**, autoimmune disease (e.g., rheumatoid arthritis, SLE),
1107
            or severe mental illness.
1108
              **Absence of diabetic, antihypertensive, or lipid-lowering pharmacotherapy**
            highlights potential undertreatment of modifiable CVD risk factors.
1109
1110
             **Clinical Synthesis**:
            This patient presents with **high-risk cardiovascular profile** driven by **age,
1111
            smoking, untreated hypertension, and significant hypercholesterolemia \star\star ,
1112
            compounded by **AF-related thromboembolic risk**.
            The absence of diabetes or CKD does not mitigate his overall risk,
1113
            as his ASCVD (atherosclerotic cardiovascular disease)
1114
            risk score would likely place him in a high-risk category.
            Urgent interventions should prioritize **smoking cessation, BP control
1115
             (target <130/80 mmHg per guidelines), and statin therapy**
1116
             (high-intensity statin indicated for LDL-C reduction >50\%).
             **Anticoagulation for AF** (CHA_2DS_2-VASc score >=2 given age >=[...] and hypertension)
1117
            should be evaluated to mitigate stroke risk.
1118
            Close monitoring of renal function (cystatin C-based eGFR) and
            lipid profiles is warranted to guide therapeutic efficacy and adherence.
1119
1120
```

```
1134
            Counterfactual Input (Redacted) - Part 2
1135
1136
            === CLOSEST DEATH ===
            The patient is a [...]-year-old [...] male with a body mass index (BMI) of
1137
             [...] kg/m^2 (weight: [...] kg, height: [...] cm), categorizing him as overweight.
1138
            He is an active smoker with a significant tobacco exposure of [...] cigarettes/day,
            a major independent risk factor for atherosclerotic cardiovascular disease (ASCVD).
1139
            His medical history is notable for atrial fibrillation (AFib),
1140
            a critical arrhythmia conferring a 5-fold increased risk of thromboembolic events,
            but no evidence of chronic kidney disease (CKD),
1141
            diabetes mellitus (DM), or autoimmune disorders.
1142
             **Cardiometabolic Profile:**
1143
            - **Hypertension:** Uncontrolled stage 2 hypertension (systolic BP: [...] mmHg)
1144
            without current antihypertensive therapy. This elevates his 10-year ASCVD risk
            substantially, particularly when combined with smoking.
1145
              **Lipid Abnormalities:** Borderline-high total cholesterol ([...] mmol/L) with
1146
            elevated calculated LDL-C (\sim[...] mmol/L via Friedewald equation) and suboptimal
            \mbox{\sc HDL-C} ([...] \mbox{\sc mmol/L})\,\mbox{\sc ,} consistent with atherogenic dyslipidemia.
1147
            Triglycerides are within normal limits ([...] mmol/L).
1148
               **Glycemic Status:** Normoglycemic (fasting glucose: [...] mmol/L;
            HbA1c: [...]%), excluding DM.
1149
1150
            **Renal & Inflammatory Markers:**
              Preserved renal function: Serum creatinine [...] mumol/L (eGFR ~[...] mL/min/[...]m^2
1151
            by CKD-EPI), cystatin C [\ldots] mg/L, and normoalbuminuria (urine microalbumin: [\ldots] mg/L).
1152
              Mild systemic inflammation: CRP [...] mg/L, potentially reflecting endothelial
            dysfunction from smoking or subclinical atherosclerosis.
1153
1154
             **Additional Risk Stratification:**
             - Absence of protective factors: No lipid-lowering, antihypertensive,
1155
            or anticoagulant use despite AFib and ASCVD risk factors.
1156
             - No familial predisposition to CVD or contributory medication use
            (e.g., corticosteroids, antipsychotics).
1157
1158
             **Clinical Implications:**
            This patient exhibits a high-risk phenotype for ASCVD
1159
            and thromboembolism due to the confluence of modifiable factors: smoking,
1160
            untreated hypertension, dyslipidemia, and AFib. The absence of diabetes
            or CKD does not mitigate this risk, as his SCORE2-Diabetes equation-adjusted
1161
            10-year CVD risk likely exceeds 10%. Urgent interventions should include
1162
            smoking cessation, BP control (target <130/80 mmHg per ACC/AHA guidelines),
            high-intensity statin therapy (e.g., atorvastatin 40-80 \text{ mg}), and CHA_2DS_2-VASc-guided anticoagulation for AFib. Weight optimization and
1163
1164
            serial monitoring of inflammatory markers (e.g., CRP)
            and cardiac troponins could further refine prognostication.
1165
1166
            === TARGET PATIENT ===
            The patient is a [\ldots]-year-old [\ldots] male with a body mass index (BMI) of
1167
             [...] kg/m^2 (height: [...] cm; weight: [...] kg), categorizing him as overweight.
1168
            He reports a significant smoking history of [\ldots]
            cigarettes/day, a critical modifiable cardiovascular risk factor.
1169
            Key clinical findings include **stage 2 hypertension**
1170
             (systolic blood pressure: [...] mmHg) without current anti-hypertensive
            therapy, **atrial fibrillation** (diagnosis confirmed), and a lipid profile
1171
            suggestive of mixed dyslipidemia (total cholesterol:
1172
             [...] mmol/L, triglycerides: [...] mmol/L, HDL-C: [...] mmol/L).
            Notably, his LDL-C (calculated via Friedewald equation: [...] mmol/L) exceeds
1173
            optimal targets, yet he is not on lipid-lowering therapy.
1174
            \label{eq:metabolically, he exhibits normoglycemia (fasting glucose: [...] mmol/L; HbAlc: [...] %)
1175
            without evidence of diabetes or anti-diabetic medication use. Renal function is
1176
            preserved (serum creatinine: [...] mumol/L; eGFR >[...] mL/min/[...] by CKD-EPI;
            cystatin C: [...] mg/L; urine microalbumin: [...] mg/L), ruling out chronic kidney
1177
            disease. Inflammatory markers are unremarkable (CRP: [...] mg/L), with no history of
1178
            autoimmune or systemic inflammatory conditions. Uric acid levels ([...] mumol/L)
            are within normal limits, mitigating concern for gout-related contributions.
1179
1180
            The absence of anti-hypertensive or anticoagulant therapy in the setting of atrial
            fibrillation and severe hypertension raises immediate concern for
1181
            thromboembolic risk (CHA_2DS_2-VASc score >=2) and hypertensive end-organ damage.
1182
            His lipid abnormalities -- elevated triglycerides,
            low HDL-C, and borderline-high LDL-C--suggest
1183
            residual atherogenic risk despite normal BMI-adjusted adiposity metrics.
1184
            The combination of smoking, untreated hypertension, and dyslipidemia
            synergistically amplifies his cardiovascular risk profile,
1185
            warranting aggressive risk factor modification.
1186
```

```
1188
            Counterfactual Input (Redacted) - Part 3
1189
1190
            No family history of CVD or secondary
            causes of hypertension (e.g., renal, endocrine) is documented.
1191
            The absence of diabetes, chronic kidney disease, or inflammatory comorbidities
1192
            narrows the primary drivers of his cardiovascular pathology to lifestyle factors
            (smoking, diet) and age-related vascular stiffening. Urgent interventions
1193
            should prioritize smoking cessation, BP control (target <130/80 mmHg per
1194
            ACC/AHA guidelines), statin initiation, and atrial fibrillation management with
            anticoagulation and rhythm control strategies.
1195
1196
            ### Required Analysis ###
            1. Comparison:
1197
               a) Identify 1-3 decisive differences between target and NNs
1198
               b) Focus on features present in ALL THREE cases
               c) Flag any conflicting evidence (e.g., "Target aligns with NN1 in [X] but NN2 in [Y]")
1199
1200
            2. Label Evaluation:
               a) Assess if Dead is correct
1201
               b) Confidence score (1-5):
1202
                  5 = All evidence strongly agrees
                  4 = Most evidence agrees
1203
                  3 = Mixed evidence
1204
                  2 = Minimal supporting evidence
                  1 = No discernible pattern
1205
1206
            3. Counterfactual:
               a) Modify one feature present in NNs
1207
               b) Predict outcome change
1208
               c) Justify using specific NN evidence
1209
            ### Response Format ###
1210
            1. Comparison:
               1) Outcome alignment: <Matches Survivor/Matches Death/Uncertain>
1211
               2) Decisive factors:
1212
                  1) <Feature>: Target vs NN1 vs NN2
                  2) <Feature>: Target vs NN1 vs NN2
1213
1214
            2. Label assessment:
               1) Correctness: <Correct/Incorrect/Edge Case>
1215
               2) Confidence: <1-5 with brief rationale>
1216
            3. Counterfactual:
1217
               1) Modification: <Feature + change>
1218
               2) Outcome: <Survived/Died>
               3) Evidence: <"Matches [NN] where [feature]=[value]">
1219
1220
            ### Quality Rules ###
             If confidence =<2, state: "Uncertain because [reason]"</pre>
1221
            - Counterfactuals must reference features actually present in NNs
1222
            - Use original feature names from input data
1223
```

Model Reasoning (Redacted)

1242

1243 1244

1245

1246

1247

1248

1249

1250

1251 1252

1253

1254

1255

1257

1259

1261

1262

1263

1264

1265

1266

1267 1268

1270

1272

1274

1276

1291

1293

1295

<think> Okay, let's start by looking at the three patients: the target who died, the closest survivor (NN1), and the closest death (NN2). All three are [...] males in their [...], overweight, with a [...]-cigarette/day smoking history, atrial fibrillation, uncontrolled hypertension, and untreated hyperlipidemia. They also don't have diabetes or CKD.

First, comparing the target to the survivor and the other deceased. The target's systolic BP is [...] mmHg, which is higher than both NN1 ([...]) and NN2 ([...]). That's a big difference. High BP is a major risk factor, so maybe the target's much higher BP contributed to his death.

Next, looking at LDL-C levels. The target's LDL is [...] mmol/L, which is lower than NN1's [...] but higher than NN2's [...]. Wait, NN2's LDL was [...], so the target is actually lower than NN2. Hmm, but the survivor had the highest LDL. That's conflicting. Maybe LDL isn't the decisive factor here.

Triglycerides: Target has [...], which is higher than both NN1 ([...]) and NN2 ([...]). Elevated triglycerides can contribute to cardiovascular risk, so this might be a key difference. The target's higher triglycerides could have increased his risk compared to the other two.

So the main differences are systolic BP and triglycerides. The target's BP is the highest, and his triglycerides are higher than both NNs. Even though his LDL is in between, the other factors might be more critical.

For the label assessment, the target died. Both NNs had similar profiles but the target's BP and triglycerides are worse. The survivor had lower BP and lower triglycerides, while the deceased had lower BP than the target. So the higher BP in the target might explain the death. The confidence is high because BP and triglycerides are significant risk factors, and the target's values are worse. For the counterfactual, if we lower the target's systolic BP to [...] mmHg (like NN1), maybe he would have survived. NN1 had lower BP and survived, so reducing BP to that level might change the outcome. Alternatively, lowering triglycerides to match NN1 or NN2 could also help, but BP seems more directly impactful. The evidence from NN1 shows that with lower BP, survival occurred despite

Model Response (Redacted)

other risks. </think>

```
### 1. Comparison:
1) **Outcome alignment**: Matches Death
2) **Decisive factors**:
   1) **Systolic BP**: Target ([...] mmHg) > NN1 ([...] mmHg) > NN2 ([...] mmHg)
      - Target's severe hypertension (stage 2) exceeds both NNs,
      increasing acute cardiovascular event risk (e.g., stroke, myocardial infarction).
   2) **Triglycerides**: Target ([...] mmol/L) > NN2 ([...] mmol/L) > NN1 ([...] mmol/L)
      - Elevated triglycerides in the target suggest greater
      residual atherogenic risk despite comparable LDL-C levels.
### 2. Label assessment:
1) **Correctness**: Correct
2) **Confidence**: 4/5
    - Rationale: The target's **higher systolic BP** and **triglycerides** align
   with worse outcomes. Both NNs had lower BP and triglycerides, yet the target died.
   However, conflicting evidence exists in LDL-C
   (target: [...] mmol/L vs. NN1: [...] mmol/L),
   where the survivor paradoxically had higher LDL-C. This reduces confidence to 4.
### 3. Counterfactual:
1) **Modification**: Reduce systolic BP to [...] mmHg (NN1's level)
2) **Outcome**: Survived
3) **Evidence**: Matches **NN1** where systolic BP=[...] mmHg (untreated but lower
than target's [...] mmHg). NN1 survived despite similar AFib, smoking, and untreated
hyperlipidemia, suggesting BP control is a pivotal survivability factor.
```

Table 3: Per-category evaluation metrics on MedMCQA for Base and Tables2Trace (8B).

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Anaesthesia (N = 24)	Base Tables2Traces	$0.36 \pm 0.07 \\ 0.36 \pm 0.07$	0.88 ± 0.07 0.83 ± 0.08	$0.29 \pm 0.09 \\ 0.29 \pm 0.09$	$0.08 \pm 0.06 \\ 0.08 \pm 0.06$	-1.15% ↓
Anatomy (<i>N</i> = 147)	Base Tables2Traces	0.36 ± 0.02 0.40 ± 0.02	$0.86 \pm 0.03 \\ 0.88 \pm 0.03$	0.26 ± 0.04 0.31 ± 0.04	0.02 ± 0.01 0.07 ± 0.02	+11.91% ↑
Biochemistry (N = 122)	Base Tables2Traces	0.57 ± 0.03 0.59 ± 0.03	0.90 ± 0.03 0.95 ± 0.02	0.58 ± 0.04 0.56 ± 0.05	0.11 ± 0.03 0.17 ± 0.03	+3.62% ↑
Dental (<i>N</i> = 845)	Base Tables2Traces	0.35 ± 0.01 0.39 ± 0.01	0.82 ± 0.01 0.88 ± 0.01	0.26 ± 0.02 0.28 ± 0.02	$0.05 \pm 0.01 \\ 0.06 \pm 0.01$	+9.24% ↑
ENT (N = 39)	Base Tables2Traces	0.39 ± 0.05 0.45 ± 0.05	$0.92 \pm 0.04 \\ 0.92 \pm 0.04$	0.26 ± 0.07 0.36 ± 0.08	$0.08 \pm 0.04 \\ 0.08 \pm 0.04$	+16.56% ↑
Forensic Medicine (N = 44)	Base Tables2Traces	$0.41 \pm 0.05 \ 0.41 \pm 0.05$	$0.89 \pm 0.05 \\ 0.89 \pm 0.05$	0.32 ± 0.07 0.30 ± 0.07	0.09 ± 0.04 0.14 ± 0.05	-1.10% ↓
Gynaecology & Obstetrics (N = 154)	Base Tables2Traces	0.40 ± 0.03 0.42 ± 0.03	0.81 ± 0.03 0.82 ± 0.03	0.32 ± 0.04 0.38 ± 0.04	0.09 ± 0.02 0.08 ± 0.02	+4.03% ↑
Medicine (N = 185)	Base Tables2Traces	0.44 ± 0.03 0.50 ± 0.03	0.84 ± 0.03 0.88 ± 0.02	0.39 ± 0.04 0.45 ± 0.04	0.12 ± 0.02 0.15 ± 0.03	+12.17% ↑
Microbiology (N = 74)	Base Tables2Traces	0.45 ± 0.04 0.48 ± 0.04	0.84 ± 0.04 0.91 ± 0.03	$0.35 \pm 0.06 \\ 0.35 \pm 0.06$	0.11 ± 0.04 0.15 ± 0.04	+7.55% ↑
Ophthalmology (N = 43)	Base Tables2Traces	0.40 ± 0.05 0.41 ± 0.05	0.91 ± 0.04 0.88 ± 0.05	$0.30 \pm 0.07 \\ 0.30 \pm 0.07$	0.14 ± 0.05 0.16 ± 0.06	+1.72% ↑
Orthopaedics (N = 15)	Base Tables2Traces	0.40 ± 0.08 0.34 ± 0.08	0.87 ± 0.09 0.80 ± 0.11	0.53 ± 0.13 0.27 ± 0.12	$0.00 \pm 0.00 \\ 0.00 \pm 0.00$	-15.00% ↓
Pathology (N = 259)	Base Tables2Traces	0.51 ± 0.02 0.54 ± 0.02	0.89 ± 0.02 0.92 ± 0.02	0.44 ± 0.03 0.53 ± 0.03	0.11 ± 0.02 0.16 ± 0.04	+5.82% ↑
Pediatrics (N = 133)	Base Tables2Traces	0.44 ± 0.03 0.47 ± 0.03	0.82 ± 0.03 0.86 ± 0.03	0.39 ± 0.04 0.38 ± 0.04	0.09 ± 0.02 0.12 ± 0.03	+6.52% ↑
Pharmacology (N = 179)	Base Tables2Traces	0.52 ± 0.03 0.56 ± 0.02	0.90 ± 0.02 0.93 ± 0.02	0.46 ± 0.04 0.55 ± 0.04	0.17 ± 0.03 0.14 ± 0.03	+8.30% ↑
Physiology (N = 133)	Base Tables2Traces	0.46 ± 0.03 0.47 ± 0.03	$0.86 \pm 0.03 \\ 0.86 \pm 0.03$	$0.38 \pm 0.04 \\ 0.38 \pm 0.04$	0.16 ± 0.03 0.14 ± 0.03	+2.30% ↑
Psychiatry (N = 10)	Base Tables2Traces	0.41 ± 0.10 0.54 ± 0.09	0.80 ± 0.13 0.90 ± 0.10	0.30 ± 0.15 0.50 ± 0.17	$0.00 \pm 0.00 \\ 0.00 \pm 0.00$	+31.71% ↑
Radiology (N = 57)	Base Tables2Traces	0.49 ± 0.04 0.45 ± 0.04	0.93 ± 0.03 0.89 ± 0.04	0.40 ± 0.07 0.39 ± 0.07	0.05 ± 0.03 0.04 ± 0.02	-8.54% ↓
Skin (<i>N</i> = 11)	Base Tables2Traces	0.37 ± 0.08 0.28 ± 0.10	0.91 ± 0.09 0.73 ± 0.14	0.27 ± 0.14 0.18 ± 0.12	0.00 ± 0.00 0.09 ± 0.09	-24.39% ↓
Social & Preventive Medicine (N = 91)	Base Tables2Traces	0.44 ± 0.04 0.47 ± 0.04	0.81 ± 0.04 0.87 ± 0.04	0.34 ± 0.05 0.43 ± 0.05	0.10 ± 0.03 0.09 ± 0.03	+7.30% ↑
Surgery (N = 249)	Base Tables2Traces	0.41 ± 0.02 0.46 ± 0.02	$0.86 \pm 0.02 \\ 0.86 \pm 0.02$	0.35 ± 0.03 0.40 ± 0.03	$0.08 \pm 0.02 \\ 0.08 \pm 0.02$	+12.12% ↑
Unknown (N = 2)	Base Tables2Traces	0.30 ± 0.30 0.35 ± 0.35	$0.50 \pm 0.50 \\ 0.50 \pm 0.50$	$0.50 \pm 0.50 \\ 0.50 \pm 0.50$	$0.00 \pm 0.00 \\ 0.00 \pm 0.00$	+16.67% ↑
Overall (N = 2816)	Base Tables2Traces	0.42 ± 0.01 0.45 ± 0.01	0.85 ± 0.01 0.88 ± 0.01	0.35 ± 0.01 0.38 ± 0.01	0.09 ± 0.01 0.10 ± 0.01	+7.49% ↑

E MEDMCQA CATEGORY-LEVEL RESULTS

To further evaluate generalization, we analyze performance across medical specialties on the MedM-CQA benchmark. As shown in Table 3, Tables2Traces improves performance across 17 of 21 categories, despite being fine-tuned exclusively on tabular data from a single clinical domain (cardiovascular). Notable gains appear in ENT (+16.56%), Social & Preventive Medicine (+16.67%), and Medicine (+12.71%), among others. While a few categories see drops (e.g., Skin, Orthopaedics), the overall gain is +7.49%. These results demonstrate that contrastive supervision derived from structured data can support generalization even to out-of-domain medical topics.

Table 4: Distribution of question types in MedQA and MedMCQA using LLM-based classification. Values are shown as raw counts and percentages of each dataset.

Benchmark	Patient-Specific	Abstract	Cardiovascular	Non-Cardiovascular
MedQA	1175 (92.3%)	98 (7.7%)	130 (10.2%)	1143 (89.8%)
MedMCQA	460 (16.3%)	2356 (83.7%)	226 (8.0%)	2590 (92.0%)

F QUESTION TYPE DISTRIBUTIONS

To better understand the nature of the questions in each benchmark, we classify them along two axes using an LLM-based approach: whether a question is *patient-specific* (referring to a concrete clinical case) or *abstract* (testing general medical knowledge), and whether it falls within the *cardiovascular* domain. As shown in Table 4, MedQA is overwhelmingly patient-specific (92.3%) and contains a small cardiovascular subset (10.2%). In contrast, MedMCQA is largely abstract (83.7%) and similarly skewed toward non-cardiovascular questions. This highlights the generalization challenge: our fine-tuned models, trained only on cardiovascular tabular data, are evaluated on questions that are mostly out-of-domain and structurally distinct.

Table 5: Aloe fine-tuning with Tables2Traces supervision. Means and standard error estimates over 10 inference runs.

Model	Avg Accuracy
Aloe	0.58 ± 0.01
Aloe + Tables2Traces	0.56 ± 0.01

G ALOE FINE-TUNING RESULTS

Aloe is a strong medical QA system trained on many curated datasets with synthetic chain-of-thought, guideline-based answers, and adversarial supervision. It is optimized for direct question answering rather than multi-step or counterfactual reasoning. We include Aloe as a point of contrast and test alignment: does reasoning supervision from Tables2Traces improve a QA-oriented model? We fine-tuned Aloe on the same Tables2Traces prompt—trace pairs and evaluated under identical test-time prompts and decoding settings as in the main experiments. Average accuracy decreases from 0.58 to 0.56 with the same standard error, indicating no benefit from reasoning-based supervision. This supports the claim that Tables2Traces is orthogonal to expensive QA curation and that QA-specific training is misaligned with reasoning traces.

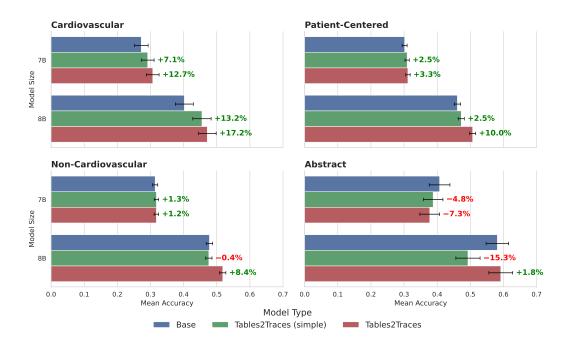


Figure 4: Accuracy on different question types in the MedQA benchmark across model sizes (7B and 8B) and fine-tuning methods. Tables2Traces yields large gains on cardiovascular and more modest gains patient-specific questions. On both cardiovascular and patient-specific questions both 7B and 8B models show consistent improvement. Minor gains are observed for non-cardiovascular questions except for the 8B Tables2Traces model. On abstract questions, all models underperform compared to the base model, except for the 8B Tables2Traces model. Values reflect relative improvement over the base model, with error bars denoting standard error across inference runs.

H RESULTS FROM QWEN-7B MODELS

To assess whether the benefits of Tables2Traces generalize across model scales, we replicate our main experiments using Qwen models with 7 billion parameters. These models are evaluated on the same MedQA and MedMCQA benchmarks, using identical training procedures as the 8B counterparts. Unlike the 8B results, however, we observe that Tables2Traces provides less consistent improvements at this smaller scale—particularly on out-of-domain or abstract questions. In some cases, performance even degrades relative to the base model.

It is important to note that this comparison involves both a change in model size ($8B \rightarrow 7B$) and architecture (LLaMA \rightarrow Qwen), so the effects cannot be attributed to scaling alone. These results suggest that both model capacity and architecture may influence the effectiveness of structured, trace-based supervision.

H.1 MEDQA

Figure 4 shows performance on the MedQA benchmark, stratified by question type and model size (7B vs. 8B). Tables2Traces yields substantial improvements on cardiovascular questions (up to +17.2%) and consistent gains on patient-specific questions, especially at the 8B scale. This suggests that structured reasoning supervision is particularly effective for case-based clinical reasoning tasks.

Performance on non-cardiovascular questions improves only modestly, and the Tables2Traces (simple) variant offers little benefit over the base model. For abstract questions, all 7B models underperform, and only Tables2Traces 8B retains accuracy. These results highlight the importance of contrastive, trace-based supervision for enabling models to generalize beyond narrowly defined training inputs.

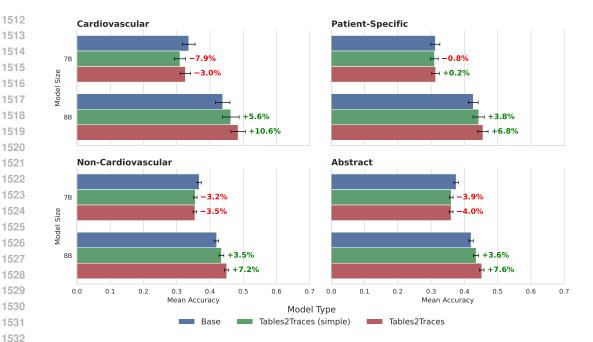


Figure 5: Accuracy on different question types in the MedMCQA benchmark across model sizes (7B and 8B) and fine-tuning methods. At 8B, Tables2Traces improves performance on all question types, including abstract and non-cardiovascular questions. In contrast, 7B models show inconsistent or negative gains, particularly for out-of-domain categories. These results suggest that contrastive supervision derived from tabular data is more effective at scale, and can generalize beyond the source domain when model capacity is sufficient. Values show relative accuracy improvements over the base model, with error bars denoting standard error across inference runs.

H.2 MEDMCQA

Figure 5 shows model performance on the MedMCQA benchmark, stratified by question type and model size. Tables2Traces yields gains at the 8B scale, improving accuracy on cardiovascular, abstract, and non-cardiovascular questions. Relative gains reach +10.6% on cardiovascular questions and +7.6% on abstract ones.

At the 7B scale, results are more mixed. Both Tables2Traces and Tables2Traces (simple) underperform the base model on most question types, suggesting that smaller models struggle to benefit from structured supervision alone. These findings reinforce the idea that contrastive, trace-based supervision is especially valuable when paired with sufficient model capacity.

Table 6: Per-category evaluation metrics on MedQA for Tables2Traces (simple) and Tables2Traces (8B). % Change refers to change in performance relative to the Base model.

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Cardiovascular (N = 130)	Tables2Traces (simple) Tables2Traces	0.46 ± 0.03 0.47 ± 0.03	0.86 ± 0.03 0.91 ± 0.03	$0.42 \pm 0.04 \\ 0.42 \pm 0.04$	$0.06 \pm 0.02 \\ 0.06 \pm 0.02$	+13.19% ↑ +17.21% ↑
Dermatologic (N = 17)	Tables2Traces (simple)	0.71 ± 0.09	0.94 ± 0.06	0.76 ± 0.11	0.29 ± 0.11	+18.81% ↑
	Tables2Traces	0.60 ± 0.08	0.88 ± 0.08	0.59 ± 0.12	0.12 ± 0.08	+0.99% ↑
Endocrine/Metabolic (N = 179)	Tables2Traces (simple) Tables2Traces	0.52 ± 0.02 0.51 ± 0.02	0.95 ± 0.02 0.91 ± 0.02	0.47 ± 0.04 0.46 ± 0.04	$0.10 \pm 0.02 \\ 0.10 \pm 0.02$	6.09% ↑ 4.71% ↑
Gastrointestinal (N = 86)	Tables2Traces (simple)	0.44 ± 0.04	0.88 ± 0.04	0.38 ± 0.05	0.09 ± 0.03	-5.72% ↓
	Tables2Traces	0.50 ± 0.04	0.91 ± 0.03	0.47 ± 0.05	0.08 ± 0.03	+6.72% ↑
Hematologic (N = 68)	Tables2Traces (simple)	0.38 ± 0.03	0.90 ± 0.04	0.25 ± 0.05	0.03 ± 0.02	-6.93% ↓
	Tables2Traces	0.48 ± 0.04	0.91 ± 0.04	0.43 ± 0.06	0.07 ± 0.03	+18.98% ↑
Immunologic (N = 81)	Tables2Traces (simple)	0.50 ± 0.04	0.93 ± 0.03	0.44 ± 0.06	0.09 ± 0.03	-2.67% ↓
	Tables2Traces	0.54 ± 0.04	0.94 ± 0.03	0.46 ± 0.06	0.17 ± 0.04	+6.80% ↑
Infectious (<i>N</i> = 176)	Tables2Traces (simple)	0.46 ± 0.02	0.92 ± 0.02	0.40 ± 0.04	0.06 ± 0.02	-3.44% ↓
	Tables2Traces	0.53 ± 0.02	0.94 ± 0.02	0.45 ± 0.04	0.11 ± 0.02	+9.73% ↑
Musculoskeletal (N = 45)	Tables2Traces (simple)	0.48 ± 0.05	0.89 ± 0.05	0.42 ± 0.07	0.04 ± 0.03	-2.71% ↓
	Tables2Traces	0.51 ± 0.04	0.96 ± 0.03	0.40 ± 0.07	0.07 ± 0.04	+4.07% ↑
Neurological (N = 77)	Tables2Traces (simple)	0.47 ± 0.04	0.86 ± 0.04	0.42 ± 0.06	0.09 ± 0.03	+6.89% ↑
	Tables2Traces	0.50 ± 0.04	0.90 ± 0.04	0.43 ± 0.06	0.05 ± 0.02	+15.15% ↑
Obstetrics/Gynecology (N = 70)	Tables2Traces (simple)	0.46 ± 0.04	0.93 ± 0.03	0.43 ± 0.06	0.07 ± 0.03	+0.93% ↑
	Tables2Traces	0.47 ± 0.03	0.94 ± 0.03	0.40 ± 0.06	0.03 ± 0.02	+2.80% ↑
Oncology (<i>N</i> = 72)	Tables2Traces (simple)	0.50 ± 0.04	0.86 ± 0.04	0.47 ± 0.06	0.15 ± 0.04	-4.76% ↓
	Tables2Traces	0.56 ± 0.04	0.93 ± 0.03	0.53 ± 0.06	0.14 ± 0.04	+5.82% ↑
Other (<i>N</i> = 31)	Tables2Traces (simple) Tables2Traces	0.45 ± 0.06 0.50 ± 0.07	0.84 ± 0.07 0.87 ± 0.06	0.45 ± 0.09 0.42 ± 0.09	0.10 ± 0.05 0.19 ± 0.07	-15.24% ↓ -4.88% ↓
Pediatric (<i>N</i> = 13)	Tables2Traces (simple) Tables2Traces	0.32 ± 0.07 0.39 ± 0.05	0.92 ± 0.08 1.00 ± 0.00	0.23 ± 0.12 0.31 ± 0.13	$\begin{array}{c} 0.00 \pm 0.00 \\ 0.00 \pm 0.00 \end{array}$	-19.61% ↓ -1.96% ↓
Psychiatric (N = 52)	Tables2Traces (simple) Tables2Traces	0.60 ± 0.05 0.62 ± 0.05	0.94 ± 0.03 0.90 ± 0.04	0.58 ± 0.07 0.61 ± 0.07	$0.21 \pm 0.06 \ 0.21 \pm 0.06$	+2.30% ↑ +5.57% ↑
Renal/Genitourinary (N = 54)	Tables2Traces (simple)	0.42 ± 0.04	0.93 ± 0.04	0.35 ± 0.07	0.06 ± 0.03	+15.08% ↑
	Tables2Traces	0.48 ± 0.04	0.96 ± 0.03	0.41 ± 0.07	0.09 ± 0.04	+29.65% ↑
Respiratory $(N = 54)$	Tables2Traces (simple) Tables2Traces	0.49 ± 0.04 0.50 ± 0.04	0.93 ± 0.04 0.94 ± 0.03	0.44 ± 0.07 0.46 ± 0.07	$0.11 \pm 0.04 \\ 0.11 \pm 0.04$	+0.76% ↑ +2.28% ↑
Toxicology $(N = 68)$	Tables2Traces (simple)	0.41 ± 0.04	0.93 ± 0.03	0.35 ± 0.06	0.03 ± 0.02	-6.10% ↓
	Tables2Traces	0.52 ± 0.04	0.91 ± 0.04	0.47 ± 0.06	0.09 ± 0.04	+20.68% ↑
Overall (<i>N</i> = 1273)	Tables2Traces (simple)	0.47 ± 0.01	0.91 ± 0.01	0.42 ± 0.01	0.08 ± 0.01	+0.82% ↑
	Tables2Traces	0.51 ± 0.01	0.93 ± 0.01	0.46 ± 0.01	0.10 ± 0.01	+9.19% ↑

I PER-CATEGORY RESULTS FROM TABLES2TRACES (SIMPLE)

Table 6 reports category-level results for both Tables2Traces and its ablated variant, Tables2Traces (simple), on the MedQA benchmark. Across most categories, the full method consistently outperforms the simple variant, highlighting the added value of contrastive and counterfactual reasoning supervision. However, the simple variant still delivers strong gains over the base model in several categories, including Cardiovascular (+13.19%), Renal/Genitourinary (+15.08%) and Neurological (+6.89%). This table complements the main figures by providing a more granular view of how each model variant performs across medical specialties.

Table 7: Per-category evaluation metrics on MedMCQA for Base and Aloe (8B).

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Anaesthesia (N = 24)	Base Aloe	0.36 ± 0.07 0.46 ± 0.07	0.88 ± 0.07 0.79 ± 0.08	0.29 ± 0.09 0.46 ± 0.10	$0.08 \pm 0.06 \\ 0.08 \pm 0.06$	+26.44%↑
Anatomy (N = 147)	Base Aloe	0.36 ± 0.02 0.49 ± 0.03	$0.86 \pm 0.03 \\ 0.86 \pm 0.03$	0.26 ± 0.04 0.42 ± 0.04	0.02 ± 0.01 0.14 ± 0.03	+37.05%↑
Biochemistry (N = 122)	Base Aloe	0.57 ± 0.03 0.69 ± 0.03	0.90 ± 0.03 0.93 ± 0.02	0.58 ± 0.04 0.66 ± 0.04	0.11 ± 0.03 0.38 ± 0.04	+21.59%↑
Dental (N = 845)	Base Aloe	0.35 ± 0.01 0.41 ± 0.01	0.82 ± 0.01 0.84 ± 0.01	0.26 ± 0.02 0.34 ± 0.02	0.05 ± 0.01 0.11 ± 0.01	+15.61%↑
ENT (N = 39)	Base Aloe	0.39 ± 0.05 0.55 ± 0.06	0.92 ± 0.04 0.90 ± 0.05	0.26 ± 0.07 0.51 ± 0.08	0.08 ± 0.04 0.23 ± 0.07	+41.06%↑
Forensic Medicine (N = 44)	Base Aloe	0.41 ± 0.05 0.50 ± 0.05	$0.89 \pm 0.05 \\ 0.89 \pm 0.05$	0.32 ± 0.07 0.43 ± 0.08	0.09 ± 0.04 0.20 ± 0.06	+20.88%↑
Gynaecology & Obstetrics (N = 154)	Base Aloe	0.40 ± 0.03 0.53 ± 0.03	0.81 ± 0.03 0.89 ± 0.03	0.32 ± 0.04 0.46 ± 0.04	0.09 ± 0.02 0.21 ± 0.03	+30.43%↑
Medicine (<i>N</i> = 185)	Base Aloe	0.44 ± 0.03 0.58 ± 0.03	0.84 ± 0.03 0.90 ± 0.02	0.39 ± 0.04 0.54 ± 0.04	0.12 ± 0.02 0.26 ± 0.03	+29.93%↑
Microbiology (N = 74)	Base Aloe	0.45 ± 0.04 0.58 ± 0.04	0.84 ± 0.04 0.89 ± 0.04	0.35 ± 0.06 0.55 ± 0.06	0.11 ± 0.04 0.24 ± 0.05	+29.31%↑
Ophthalmology (N = 43)	Base Aloe	0.40 ± 0.05 0.54 ± 0.06	0.91 ± 0.04 0.93 ± 0.04	0.30 ± 0.07 0.49 ± 0.08	0.14 ± 0.05 0.23 ± 0.07	+34.48%↑
Orthopaedics (N = 15)	Base Aloe	0.40 ± 0.08 0.59 ± 0.08	0.87 ± 0.09 0.93 ± 0.07	0.53 ± 0.13 0.60 ± 0.13	0.00 ± 0.00 0.13 ± 0.09	+46.67%↑
Pathology (N = 259)	Base Aloe	0.51 ± 0.02 0.65 ± 0.02	0.89 ± 0.02 0.91 ± 0.02	0.44 ± 0.03 0.64 ± 0.03	0.11 ± 0.02 0.32 ± 0.03	+27.73%↑
Pediatrics (N = 133)	Base Aloe	0.44 ± 0.03 0.57 ± 0.03	0.82 ± 0.03 0.90 ± 0.03	0.39 ± 0.04 0.51 ± 0.04	0.09 ± 0.02 0.19 ± 0.03	+31.05%↑
Pharmacology (N = 179)	Base Aloe	0.52 ± 0.03 0.72 ± 0.03	0.90 ± 0.02 0.93 ± 0.02	0.46 ± 0.04 0.69 ± 0.03	0.17 ± 0.03 0.43 ± 0.04	+38.04%↑
Physiology (N = 133)	Base Aloe	0.46 ± 0.03 0.60 ± 0.03	0.86 ± 0.03 0.89 ± 0.03	0.38 ± 0.04 0.58 ± 0.04	0.16 ± 0.03 0.29 ± 0.04	+31.2%↑
Psychiatry (N = 10)	Base Aloe	0.41 ± 0.10 0.60 ± 0.13	0.80 ± 0.13 0.90 ± 0.10	0.30 ± 0.15 0.60 ± 0.16	0.00 ± 0.00 0.30 ± 0.15	+46.34%↑
Radiology (N = 57)	Base Aloe	0.49 ± 0.04 0.51 ± 0.05	0.93 ± 0.03 0.89 ± 0.04	0.40 ± 0.07 0.44 ± 0.07	0.05 ± 0.03 0.14 ± 0.05	+2.85%↑
Skin (N = 11)	Base Aloe	0.37 ± 0.08 0.47 ± 0.12	0.91 ± 0.09 0.73 ± 0.14	0.27 ± 0.14 0.36 ± 0.15	0.00 ± 0.00 0.18 ± 0.12	+26.83%↑
Social & Preventive Medicine (N = 91)	Base Aloe	0.44 ± 0.04 0.52 ± 0.04	0.81 ± 0.04 0.88 ± 0.03	0.34 ± 0.05 0.47 ± 0.05	0.10 ± 0.03 0.20 ± 0.04	+19.14%↑
Surgery (N = 249)	Base Aloe	0.41 ± 0.02 0.51 ± 0.02	$0.86 \pm 0.02 \\ 0.86 \pm 0.02$	0.35 ± 0.03 0.47 ± 0.03	0.08 ± 0.02 0.17 ± 0.02	+24.24%↑
Unknown (N = 2)	Base Aloe	0.30 ± 0.30 0.20 ± 0.20	$0.50 \pm 0.50 \\ 0.50 \pm 0.50$	0.50 ± 0.50 0.00 ± 0.00	$0.00 \pm 0.00 \\ 0.00 \pm 0.00$	-33.33%↓
Overall (<i>N</i> = 2816)	Base Aloe	0.42 ± 0.01 0.53 ± 0.01	0.85 ± 0.01 0.88 ± 0.01	0.35 ± 0.01 0.48 ± 0.01	0.09 ± 0.01 0.21 ± 0.01	+25.31%↑

J PER-CATEGORY RESULTS FROM ALOE

For completeness, we report a category-level breakdown of Aloe's performance on the MedMCQA benchmark in Table 7. Aloe achieves consistent improvements over the base model across nearly all medical specialties, with an overall relative gain of +25.31%. Gains are especially large in domains such as Psychiatry (+46.34%), Orthopaedics (+46.67%), and Pharmacology (+38.04%). Only one category (Unknown) shows a performance regression, but it notably only contains two questions. These results align with Aloe's strong overall performance and provide additional insight into which specialties benefit most from its QA-style supervision. We note that Aloe is an upper-bound baseline and that our work is best viewed as a complementary approach rather than a competing one.

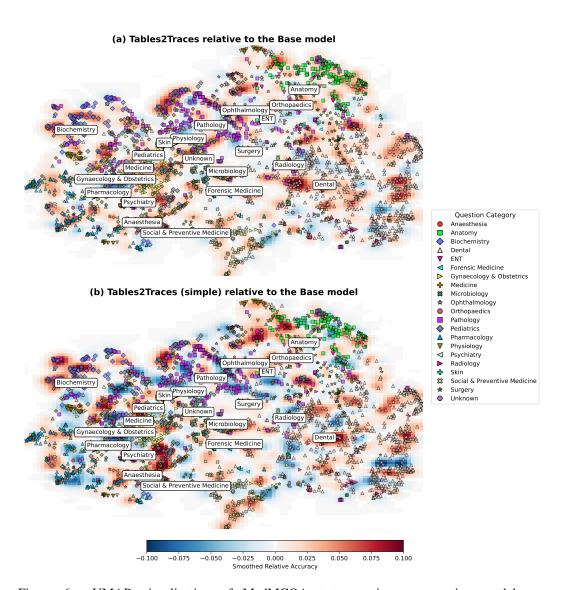


Figure 6: UMAP visualization of MedMCQA test questions, comparing model performance to the **Base** model. Each point represents a question, embedded using text-embedding-3-large (OpenAI, 2023), and annotated by medical category using distinct marker shapes and colors. The background heatmap reflects smoothed relative accuracy: red indicates improved performance, blue indicates degradation. (a) Tables2Traces shows consistent gains across diverse medical categories. (b) Tables2Traces (simple) displays more variable patterns, with several regions showing decreased performance. Cluster labels indicate category centroids.

K UMAP VISUALIZATION OF MEDMCQA

Figure 6 shows a UMAP projection of MedMCQA test questions, colored by medical category and overlaid with performance changes relative to the base model. As in MedQA, Tables2Traces (Figure 6a) shows widespread gains across the space. Notable improvements appear in regions corresponding to Anatomy, Dental, and Pathology.

In contrast, Tables2Traces (simple) (Figure 6b) demonstrates a more fragmented pattern. While some clusters benefit (e.g., Dental, Anatomy), others experience performance drops, particularly in Biochemistry and Pharmacology. These results further support the conclusion that structured contrastive supervision is critical for consistent generalization beyond the source domain.

Table 8: Per-category evaluation metrics on the MedQA benchmark for Base and QA-Finetuning (8B).

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Cardiovascular (N = 130)	Base QA-Finetuning	0.40 ± 0.03 0.39 ± 0.02	0.86 ± 0.03 0.89 ± 0.03	0.31 ± 0.04 0.30 ± 0.04	0.06 ± 0.02 0.03 ± 0.02	-3.63% ↓
Dermatologic (N = 17)	Base QA-Finetuning	0.59 ± 0.08 0.54 ± 0.07	$0.94 \pm 0.06 \\ 0.94 \pm 0.06$	$0.53 \pm 0.12 \ 0.53 \pm 0.12$	$\begin{array}{c} 0.06 \pm 0.06 \\ 0.06 \pm 0.06 \end{array}$	-9.9%↓
Endocrine/Metabolic (N = 179)	Base QA-Finetuning	0.49 ± 0.03 0.45 ± 0.02	0.89 ± 0.02 0.91 ± 0.02	0.45 ± 0.04 0.40 ± 0.04	0.13 ± 0.03 0.06 ± 0.02	-6.55% ↓
Gastrointestinal (N = 86)	Base QA-Finetuning	0.47 ± 0.04 0.45 ± 0.03	0.87 ± 0.04 0.88 ± 0.03	0.40 ± 0.05 0.37 ± 0.05	0.12 ± 0.04 0.07 ± 0.03	-3.73% ↓
Hematologic (N = 68)	Base QA-Finetuning	0.40 ± 0.04 0.42 ± 0.04	0.84 ± 0.04 0.87 ± 0.04	0.34 ± 0.06 0.37 ± 0.06	$0.04 \pm 0.03 \\ 0.04 \pm 0.03$	+5.11% ↑
Immunologic (N = 81)	Base QA-Finetuning	0.51 ± 0.04 0.50 ± 0.03	0.85 ± 0.04 0.94 ± 0.03	0.47 ± 0.06 0.43 ± 0.06	0.22 ± 0.05 0.10 ± 0.03	-2.43% ↓
Infectious (N = 176)	Base QA-Finetuning	0.48 ± 0.03 0.46 ± 0.02	0.92 ± 0.02 0.90 ± 0.02	0.41 ± 0.04 0.37 ± 0.04	0.11 ± 0.02 0.07 ± 0.02	-4.98%↓
Musculoskeletal (N = 45)	Base QA-Finetuning	0.49 ± 0.05 0.45 ± 0.04	0.89 ± 0.05 0.98 ± 0.02	0.49 ± 0.07 0.33 ± 0.07	0.04 ± 0.03 0.02 ± 0.02	-8.14% ↓
Neurological (N = 77)	Base QA-Finetuning	0.47 ± 0.04 0.51 ± 0.03	0.86 ± 0.04 0.96 ± 0.02	0.42 ± 0.06 0.44 ± 0.06	0.09 ± 0.03 0.06 ± 0.03	+8.26% ↑
Obstetrics/Gynecology (N = 70)	Base QA-Finetuning	0.46 ± 0.04 0.45 ± 0.03	0.90 ± 0.04 0.91 ± 0.03	$0.39 \pm 0.06 \\ 0.39 \pm 0.06$	0.09 ± 0.03 0.03 ± 0.02	-1.86% ↓
Oncology (N = 72)	Base QA-Finetuning	0.53 ± 0.04 0.46 ± 0.04	0.92 ± 0.03 0.90 ± 0.04	0.47 ± 0.06 0.44 ± 0.06	0.11 ± 0.04 0.06 ± 0.03	-13.23% ↓
Other (N = 31)	Base QA-Finetuning	0.53 ± 0.07 0.48 ± 0.06	0.77 ± 0.08 0.90 ± 0.05	$0.45 \pm 0.09 \\ 0.45 \pm 0.09$	0.23 ± 0.08 0.16 ± 0.07	-8.54% ↓
Pediatric (N = 13)	Base QA-Finetuning	0.39 ± 0.09 0.43 ± 0.09	0.77 ± 0.12 0.85 ± 0.10	0.39 ± 0.14 0.38 ± 0.14	0.00 ± 0.00 0.08 ± 0.08	+9.8%↑
Psychiatric (<i>N</i> = 52)	Base QA-Finetuning	0.59 ± 0.05 0.53 ± 0.04	0.94 ± 0.03 0.88 ± 0.04	0.54 ± 0.07 0.50 ± 0.07	0.23 ± 0.06 0.13 ± 0.05	-8.85% ↓
Renal/Genitourinary (N = 54)	Base QA-Finetuning	0.37 ± 0.04 0.42 ± 0.04	0.85 ± 0.05 0.93 ± 0.04	0.26 ± 0.06 0.30 ± 0.06	0.04 ± 0.03 0.06 ± 0.03	+13.57% ↑
Respiratory (N = 54)	Base QA-Finetuning	0.49 ± 0.04 0.50 ± 0.04	$0.91 \pm 0.04 \\ 0.91 \pm 0.04$	0.43 ± 0.07 0.48 ± 0.07	$0.09 \pm 0.04 \\ 0.09 \pm 0.04$	+2.66% ↑
Toxicology (N = 68)	Base QA-Finetuning	$0.43 \pm 0.04 \\ 0.43 \pm 0.03$	0.79 ± 0.05 0.93 ± 0.03	0.41 ± 0.06 0.29 ± 0.06	0.06 ± 0.03 0.03 ± 0.02	-1.69%↓
Overall (<i>N</i> = 1273)	Base QA-Finetuning	0.47 ± 0.01 0.46 ± 0.01	0.88 ± 0.01 0.91 ± 0.01	0.41 ± 0.01 0.39 ± 0.01	0.11 ± 0.01 0.06 ± 0.01	-3.17% ↓

L QA-ONLY ABLATION RESULTS

To assess whether standard QA-format supervision could account for the performance improvements observed in our full method, we conduct an ablation where the model is fine-tuned exclusively on the 10k QA-format examples used in the mixed setup. Importantly, these 10K QA-format examples do not overlap with the questions used for evaluation. Results are shown in Table 8 (MedQA) and Table 9 (MedMCQA).

On MedQA, the QA-only model performs comparably or slightly worse than the base model (0.46 vs. 0.47 average accuracy), with inconsistent effects across clinical categories. On MedMCQA, the QA-only model performs worse than the base model overall (0.40 vs. 0.42 average accuracy) and shows negative or negligible gains across most categories. These results indicate that the QA examples alone do not explain the improvements observed in our main models.

Table 9: Per-category evaluation metrics on MedMCQA for Base and QA-finetuning (8B).

Category	Model Type	Avg Accuracy	Best-of-n	Majority Vote	Worst-of-n	% Change
Anaesthesia (N = 24)	Base QA-Finetuning	0.36 ± 0.07 0.25 ± 0.04	0.88 ± 0.07 0.83 ± 0.08	0.29 ± 0.09 0.04 ± 0.04	0.08 ± 0.06 0.00 ± 0.00	-29.89%↓
Anatomy (<i>N</i> = 147)	Base QA-Finetuning	$egin{array}{l} {f 0.36 \pm 0.02} \ 0.33 \pm 0.02 \end{array}$	0.86 ± 0.03 0.89 ± 0.03	$egin{array}{l} {f 0.26 \pm 0.04} \ 0.23 \pm 0.03 \end{array}$	0.02 ± 0.01 0.01 ± 0.01	-8.70%↓
Biochemistry $(N = 122)$	Base QA-Finetuning	0.57 ± 0.03 0.51 ± 0.03	$\begin{array}{c} 0.90 \pm 0.03 \\ 0.90 \pm 0.03 \end{array}$	0.58 ± 0.04 0.49 ± 0.05	0.11 ± 0.03 0.12 ± 0.03	-9.57%↓
Dental (<i>N</i> = 845)	Base QA-Finetuning	$egin{array}{l} {f 0.35 \pm 0.01} \ 0.34 \pm 0.01 \end{array}$	0.82 ± 0.01 0.86 ± 0.01	$egin{array}{c} {f 0.26 \pm 0.02} \ 0.23 \pm 0.01 \end{array}$	0.05 ± 0.01 0.03 ± 0.01	-4.90%↓
ENT (N = 39)	Base QA-Finetuning	0.39 ± 0.05 $\mathbf{0.42 \pm 0.05}$	0.92 ± 0.04 0.90 ± 0.05	0.26 ± 0.07 0.28 ± 0.07	$0.08 \pm 0.04 \\ 0.08 \pm 0.04$	+7.28%↑
Forensic Medicine (N = 44)	Base QA-Finetuning	0.41 ± 0.05 0.36 ± 0.05	0.89 ± 0.05 0.86 ± 0.05	0.32 ± 0.07 0.25 ± 0.07	0.09 ± 0.04 0.07 ± 0.04	-12.64%↓
Gynaecology & Obstetrics (N = 154)	Base QA-Finetuning	0.40 ± 0.03 0.37 ± 0.02	0.81 ± 0.03 0.86 ± 0.03	0.32 ± 0.04 0.31 ± 0.04	0.09 ± 0.02 0.03 ± 0.01	-7.41%↓
Medicine (N = 185)	Base QA-Finetuning	$0.44 \pm 0.03 \\ 0.44 \pm 0.02$	0.84 ± 0.03 0.87 ± 0.02	0.39 ± 0.04 0.36 ± 0.04	0.12 ± 0.02 0.09 ± 0.02	-0.73%↓
Microbiology (N = 74)	Base QA-Finetuning	0.45 ± 0.04 0.42 ± 0.04	0.84 ± 0.04 0.86 ± 0.04	0.35 ± 0.06 0.36 ± 0.06	0.11 ± 0.04 0.04 ± 0.02	-5.74%↓
Ophthalmology (N = 43)	Base QA-Finetuning	0.40 ± 0.05 0.41 ± 0.05	0.91 ± 0.04 0.93 ± 0.04	0.30 ± 0.07 0.33 ± 0.07	0.14 ± 0.05 0.05 ± 0.03	+0.57%↑
Orthopaedics (N = 15)	Base QA-Finetuning	0.40 ± 0.08 0.38 ± 0.07	$0.87 \pm 0.09 \\ 0.87 \pm 0.09$	0.53 ± 0.13 0.27 ± 0.12	0.00 ± 0.00 0.07 ± 0.07	-5.00%↓
Pathology (N = 259)	Base QA-Finetuning	0.51 ± 0.02 0.45 ± 0.02	0.89 ± 0.02 0.90 ± 0.02	0.44 ± 0.03 0.37 ± 0.03	0.11 ± 0.02 0.06 ± 0.01	-10.92%↓
Pediatrics (N = 133)	Base QA-Finetuning	$0.44 \pm 0.03 \\ 0.44 \pm 0.03$	0.82 ± 0.03 $\mathbf{0.87 \pm 0.03}$	0.39 ± 0.04 0.37 ± 0.04	0.09 ± 0.02 0.05 ± 0.02	+0.86%↑
Pharmacology (N = 179)	Base QA-Finetuning	0.52 ± 0.03 0.50 ± 0.02	$0.90 \pm 0.02 \\ 0.90 \pm 0.02$	0.46 ± 0.04 0.44 ± 0.04	0.17 ± 0.03 0.12 ± 0.02	-3.13%↓
Physiology (N = 133)	Base QA-Finetuning	0.46 ± 0.03 0.45 ± 0.03	0.86 ± 0.03 $\mathbf{0.87 \pm 0.03}$	$0.38 \pm 0.04 \\ 0.38 \pm 0.04$	0.16 ± 0.03 0.12 ± 0.03	-2.79%↓
Psychiatry (N = 10)	Base QA-Finetuning	0.41 ± 0.10 0 0.49 \pm 0.10	$0.80 \pm 0.13 \\ 0.80 \pm 0.13$	0.30 ± 0.15 0.50 ± 0.17	0.00 ± 0.00 0.00 ± 0.00	+19.51%↑
Radiology (N = 57)	Base QA-Finetuning	0.49 ± 0.04 0.45 ± 0.04	$0.93 \pm 0.03 \\ 0.93 \pm 0.03$	$0.40 \pm 0.07 \\ 0.40 \pm 0.07$	0.05 ± 0.03 0.04 ± 0.02	-8.54%↓
Skin (N = 11)	Base QA-Finetuning	0.37 ± 0.08 $\mathbf{0.39 \pm 0.10}$	0.91 ± 0.09 1.00 ± 0.00	$0.27 \pm 0.14 \\ 0.27 \pm 0.14$	0.00 ± 0.00 0.09 ± 0.09	+4.88%↑
Social & Preventive Medicine (N = 91)	Base QA-Finetuning	0.44 ± 0.04 0.43 ± 0.03	0.81 ± 0.04 $\mathbf{0.89 \pm 0.03}$	0.34 ± 0.05 0.32 ± 0.05	0.10 ± 0.03 0.08 ± 0.03	-2.52%↓
Surgery (N = 249)	Base QA-Finetuning	0.41 ± 0.02 0.40 ± 0.02	0.86 ± 0.02 $\mathbf{0.88 \pm 0.02}$	0.35 ± 0.03 0.31 ± 0.03	0.08 ± 0.02 0.05 ± 0.01	-2.93%↓
Unknown (N = 2)	Base QA-Finetuning	0.30 ± 0.30 $\mathbf{0.45 \pm 0.25}$	0.50 ± 0.50 1.00 ± 0.00	$0.50 \pm 0.50 \\ 0.50 \pm 0.50$	$0.00 \pm 0.00 \\ 0.00 \pm 0.00$	+50.00%↑
Overall (N = 2816)	Base QA-Finetuning	0.42 ± 0.01 0.40 ± 0.01	0.85 ± 0.01 0.88 ± 0.01	0.35 ± 0.01 0.31 ± 0.01	0.09 ± 0.01 0.06 ± 0.00	-5.10%↓

M DISTANCE METRIC CHOICE FOR CONTRASTIVE NEIGHBOR SELECTION

Rationale. We use the Gower distance because it is data-type agnostic and compares heterogeneous features (numeric, binary, categorical) without domain-specific encodings. It provides a simple, interpretable default for mixed clinical tables.

Alternatives. The pipeline is metric-agnostic. In principle, other choices can be substituted in the neighbor retrieval step, for example: (i) scaled Euclidean on normalized numeric features with one-hot categories, (ii) Hamming distance for categorical-only subsets, (iii) cosine distance on serialized or embedded representations, or (iv) learned metrics (e.g., Mahalanobis) if one wishes to tune feature weights.

Scope. A full comparison of distance functions is outside the scope of this paper. We adopt Gower as a simple and effective default for mixed-type data, and future work could explore learned or task-specific metrics.

N CLINICIAN EVALUATION PROTOCOL AND RUBRIC

Protocol. We randomly sampled 10 supervision traces from the training corpus. Two independent clinicians with cardiology expertise reviewed the same set, each completing a structured rubric for every trace without seeing the other's responses. Cases contained only de-identified, synthesized patient descriptions derived from tabular rows (anchor and neighbors). The clinicians were asked to rate each trace along five dimensions and optionally add a one-line comment. We report the normalized tallies in Tables 10–11.

Rubric (per trace). Each trace was rated on the following dimensions with the indicated discrete scale.

- 1. Overall clinical plausibility: Yes / Partially / No.
- 2. Unsafe or inappropriate recommendations: None | Minor | Concerning.
- 3. Appropriate weighting of key factors: Yes / Partially / No.
- 4. Comparative reasoning quality (why target vs. neighbor): Clear / Partial / Superficial.
- 5. **Uncertainty expression:** *Understated | Appropriate | Overstated.*
- 6. One-line comment (optional): free-text note (e.g., phrasing, missing considerations).

Guidance provided to raters.

- *Plausibility* asks whether the narrative could reasonably reflect clinical reasoning given only the provided variables.
- *Unsafe/inappropriate* flags any recommendation that would be clinically unsafe or clearly inappropriate in context; "Minor" covers low-risk or borderline phrasing.
- Weighting assesses whether major risk factors are emphasized appropriately relative to minor ones.
- Comparative reasoning evaluates whether differences between target and neighbors are identified and used to justify outcomes.
- Uncertainty evaluates acknowledgment of limits of the available variables (avoid overconfidence or implying hidden labels).

Limitations. This review is qualitative and small-scale (n=10 traces), with no rater training or adjudication; results should be interpreted as a plausibility/safety check for *research-only* supervision rather than clinical validation or calibrated risk assessment. Importantly, we also note the high disagreement between the two clinicians.

Table 10: Clinician (A): tally of ratings across 10 traces.

	Positive	Partial / Minor	Negative
Plausibility	5 (Yes)	4 (Partially)	1 (No)
Unsafe / inappropriate	7 (None)	3 (Minor)	0 (Concerning)
Weighting	5 (Yes)	4 (Partially)	1 (No)
Reasoning	5 (Clear)	4 (Partial)	1 (Superficial)
Uncertainty	5 (Appropriate)	_	5 (Overstated)

Table 11: Clinician (B): tally of ratings across 10 traces.

	Positive	Partial / Minor	Negative
Plausibility	1 (Yes)	9 (Partially)	0 (No)
Unsafe / inappropriate	4 (None)	6 (Minor)	0 (Concerning)
Weighting	0 (Yes)	8 (Partially)	2 (No)
Reasoning	0 (Clear)	4 (Partial)	6 (Superficial)
Uncertainty	0 (Appropriate)	_	10 (Overstated)

O CLINICIAN VALIDATION OF SUPERVISION TRACES

Out of the 10 randomly sampled traces, no trace received a "Concerning" safety rating, and only one traces was judged to be implausible by Clinician (A). In addition to these discrete measures, we also include a table containing the qualitative feedback from Clinician (B) on 10 randomly sampled traces. Clinician (A) only had one comment, stating that "CRP should be acted on. Recommend finding the cause of CRP 45, like cancer". We therefore only include a table for Clinician (B), who had additional comments for all traces. Tags indicate recurring themes; comments are lightly abridged for brevity.

Table 12: Clinician (B) qualitative review of 10 traces.

Patient	Issue tags	Clinician comment (abridged)
1	Vague phrasing; overconfidence	"BP 'way above' is not clinical phrasing—use concrete categories (e.g., stage 2 hypertension). Consider guideline scores (e.g., CHADS ₂ -VASc)."
2	Overstates intervention benefit; weighting	"Hyperlipidemia not that serious for a non-smoking woman without diabetes, even with grade 2 hypertension."
3	Speculative; conflicting factors; circular counterfactual	"Acknowledge conflict between anthropometric and biochemical factors; counterfactual goes in circles."
4	Partial weighting; overconfidence	"Reasoning partial; certainty overstated given available variables."
5	Baseline risk omitted	"Age not addressed correctly—baseline mortality risk."
6	Counterfactual focus mis- aligned	"BG change not the primary modifiable variable in this context; reasoning superficial."
7	Overconfidence; superficial	"Reasoning superficial; certainty overstated."
8	Misinterpretation of CRP	"Elevated CRP may reflect infection—don't treat as CVD risk alone."
9	Lab inconsistency note	"Glucose/HbA1c discrepancy is often seen (e.g., after a meal)."
10	Risk calibration; BP control	"CVD risk overstated; emphasize blood-pressure control (cf. risk charts/guidelines)."

P LLM USAGE

In accordance with the ICLR 2026 Author Guide, we disclose that large language models (LLMs) were used solely to assist with text polishing and minor formatting during paper preparation. LLMs were not involved in research ideation, experiment design, or result interpretation. All scientific contributions, methods, and conclusions are the work of the authors.