Tables2Traces: Distilling Tabular Data to Improve LLM Reasoning in Healthcare

Mikkel Werling
University of Copenhagen
mikkel.werling@regionh.dk

Nabeel Seedat
University of Cambridge
ns741@cam.ac.uk

Jiashuo Liu
Tsinghua University
liujiashuo77@gmail.com

Lars Grønlykke
University of Copenhagen
lars.groenlykke@regionh.dk

Carsten Utoft Niemann
University of Copenhagen
carsten.utoft.niemann@regionh.dk

Mihaela van der Schaar
University of Cambridge
mv472@cam.ac.uk

Rudi Agius
University of Copenhagen
rudi.agius.01@regionh.dk

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. We introduce **Tables2Traces**, a framework that transforms tabular records into contrastive, case-based reasoning traces for LLM fine-tuning. This converts data traditionally used for prediction into structured reasoning supervision, introducing a new paradigm complementary to text-based QA fine-tuning. Crucially, this paradigm is orthogonal to text-based QA supervision: 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 MedMCQA, it achieves a 7.2% relative improvement and outperforms the base model in 17 of 21 specialties.

AI for Tabular Data Workshop at EurIPS 2025

Large language models (LLMs) have achieved remarkable performance across reasoning tasks, from mathematics [2] to medical question answering [16], largely due to large-scale pretraining and fine-tuning on text-based reasoning datasets [12, 20].

However, in many high-value domains, knowledge is stored in *tabular format* such as electronic health records or enterprise spreadsheets. Such datasets encode complex relationships and decision logic but lack the textual form required for LLM fine-tuning [21, 10]. Closing this modality gap would unlock the reasoning signals latent in tabular data.

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 elicit reasoning traces that capture the latent knowledge contained in those structures.

To address C1-C2, we introduce **Tables2Traces**, the first end-to-end framework that transforms tabular data into contrastive reasoning traces for LLM fine-tuning. More specifically, Tables2Traces addresses the following key research question:

> Can the latent knowledge embedded in 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?

1 Related Work

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

LLM Fine-Tuning. Fine-tuning LLMs on reasoning or instruction datasets (e.g., GSM8K, Self-Instruct) enhances generalization [20, 19]. DeepSeek-R1 [5] extends this by training on curated reasoning traces via Generalized Reinforcement Preference Optimization (GRPO). We follow this paradigm but derive reasoning traces directly from tabular data, requiring neither natural-language corpora nor human feedback [12, 15].

LLMs for tabular data. Models such as TaBERT, TAPEX, and TURL [21, 10, 3] learn joint text–table embeddings for QA, while TabNet and FT-Transformer [1, 4] focus on direct prediction. Recent TabLLM and UniPredict [7, 18] serialize tables for zero-shot tasks. In contrast, **Tables2Traces** uses tabular data as *reasoning supervision* to enhance LLM reasoning itself.

2 Method

2.1 Problem Formulation

We consider a tabular dataset $\mathcal{D} = (x_i, y_i)_{i=1}^N$, where each record $x_i \in \mathbb{R}^d$ has a binary label $y_i \in 0, 1$. Our goal is to map $\mathcal{D}$ into a corpus of contrastive prompts and reasoning traces,

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

where P_i describes a structured scenario in text and R_i is a generated reasoning trace. Fine-tuning an LLM on $\mathcal{C}$ aims to transfer the relational knowledge in tabular data into high-level reasoning ability. We demonstrate this on medical data but the approach is in principle domain-agnostic.

2.2 TABLES2TRACES

We propose Tables2Traces as a framework that realizes this mapping function, Π. In doing so, we address (C1) representation and (C2) trace elicitation. The underlying algorithm is outlined in Algorithm 1.

(C1) Representation. A deterministic encoder ϕ serializes each record into a short textual case. Column names are rendered as human-readable phrases, numerical values given with units, and missing values made explicit. These summaries form $\mathcal{C}_{\text{simple}} = \phi(x_i)$ which we use as training data for the Tables2Traces (simple) variant.

(C2) Trace Elicitation. Each anchor x_i is paired with its nearest neighbors using Gower distance; one neighbor with the same label as the anchor and one with the opposite label. Together, these form a contrastive triplet $\tau_i = (s_i^{(0)}, s_i, s_i^{(1)})$. A fixed prompt template π elicits differential, evaluative, and counterfactual reasoning from a frozen generator LLM $\mathcal{L}$, producing a trace $R_i = \mathcal{L}(\pi(\tau_i))$. Aggregating (P_i, R_i) pairs yields the corpus $\mathcal{C}$. We provide representative examples and full prompt templates in the Appendix.

Supervised Fine-tuning. We then fine-tune a downstream target LLM using standard supervised learning on the dataset $\mathcal{C}$ using standard language-modelling loss:

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

Extensibility. While demonstrated on binary clinical data, Tables2Traces can be applied to any tabular domain given a suitable contrastive pairing strategy (e.g., clustering or label binning).

Algorithm 1 TABLES2TRACES: From Tabular Data to Reasoning Corpus

Require: Tabular dataset D, frozen LLM $\mathcal{L}$
Output: Reasoning corpus $\mathcal{C}$
for each $(x_i, y_i) \in D$ **do**
  $s_i \leftarrow \phi(x_i)$ ▷ Representation
  for $y \in \{0, 1\}$ **do**
    $x_i^{(y)} \leftarrow \arg\min_{x: y_x = y} \text{Gower}(x, x_i)$
  end for
  $\tau_i \leftarrow \left(\phi(x_i^{(0)}), s_i, \phi(x_i^{(1)})\right)$ ▷ Contrastive triple
  $P_i \leftarrow \pi(\tau_i)$ ▷ Compose prompt
  $R_i \leftarrow \text{POSTPROCESS}\left(\mathcal{L}(P_i)\right)$ ▷ Trace elicitation
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(P_i, R_i)\}$
end for
return $\mathcal{C}$

3 Experiments

We evaluate Tables2Traces as a mechanism for converting structured tabular data into reasoning supervision for LLMs. Our objective is to determine whether this supervision improves medical QA performance and to analyze generalization across question types and domains.

Data. We use a subset of the UK Biobank [17] containing 105,299 cardiovascular patients, each represented by 32 demographic, medication, and laboratory variables. All data were collected under appropriate ethical approvals [14].

Setup. We compare two variants: (1) **Tables2Traces (simple)**, which converts each record into a patient narrative; and (2) **Tables2Traces**, which adds contrastive prompts using nearest-neighbor pairs. Both are trained on 90% synthetic traces and 10% QA-format examples from `MedQA-Mixtral-CoT` [6] for task alignment. We fine-tune 7B–8B parameter LLaMA and Qwen models using the Open-R1 framework [8]. Training details and 7B results appear in Appendix B.

Evaluation. We evaluate models on `MedQA` [9] and `MedMCQA` [13], reporting accuracy over 10 runs. We compare to the base model and to **Aloe** [6], a QA-optimized and resource intensive upper bound trained on extensive curated reasoning traces. Additional analyses—category-level results, domain transfer, embedding-space visualizations, and clinician review—are included in the Supplement.

We evaluate Tables2Traces across three aspects: overall performance, domain generalization, and comparison to QA-optimized baselines (see Supplement for full analyses).

3.1 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`, questions are grouped into 18 clinical categories using `DeepSeek-R1`. For `MedMCQA`, we use the dataset-provided labels and evaluate on the public validation set. To ensure comparability, we restrict evaluation to questions with a single correct answer.

Results. Table 1 summarizes overall accuracy on both MedQA and MedMCQA. As shown in Appendix E.1 and Appendix E.2, Tables2Traces consistently outperforms the base model across most categories in both benchmarks. On `MedQA`, it improves accuracy in 16 / 18 categories. On `MedMCQA`, gains appear in 17 / 21 categories (81%). Small negative shifts occur only in low-sample categories ($N<20$).

Table 1: Average accuracy across MedQA and MedMCQA (8B models).

Benchmark	Base	Tables2Traces	% Change
MedQA	0.47 ± 0.01	$\mathbf{0.51 \pm 0.01}$	**+9.19% ↑**
MedMCQA	0.42 ± 0.01	$\mathbf{0.45 \pm 0.01}$	**+7.49% ↑**

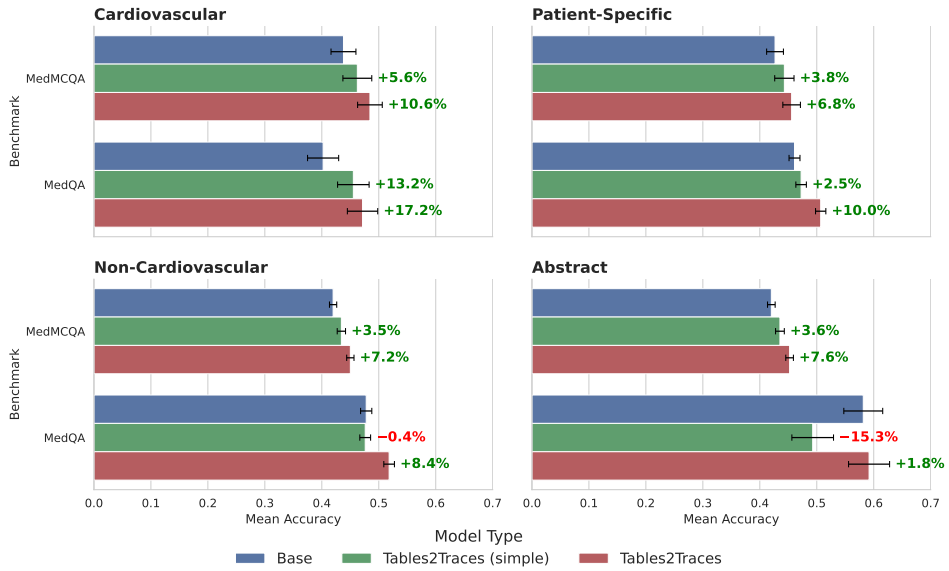

Figure 1: Accuracy comparison across model variants. Subplots show mean accuracy on `MedQA` and `MedMCQA`, with error bars for standard error. Percentages indicate improvement over the Base model. Top row: in-domain (cardiovascular, patient-specific); bottom row: out-of-domain (non-cardiovascular, abstract).

3.2 What types of questions benefit from tabular supervision?

Goal. To assess whether tabular supervision derived from a single clinical domain (cardiovascular) and patient-specific traces can transfer to unseen specialties and abstract medical knowledge.

Setup. We assess generalization along two axes: (1) **Domain**—partitioning questions into *cardiovascular* vs. *non-cardiovascular* 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 improves cardiovascular (+17.2 %) and patient-specific (+10.0%) subsets, with smaller gains on abstract questions (+1.8%). The simpler variant underperforms on abstract items (–15.3 %), indicating overfitting to format. In `MedMCQA`, gains are broader, improving both patient-specific (+6.8%) and abstract (+7.6%) subsets. UMAP visualizations in Appendix L corroborate these findings, showing that Tables2Traces improves coverage across semantically distant question regions.

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

Goal. We benchmark Tables2Traces against **Aloe** [6], a state-of-the-art medical QA model trained on over 750 k curated question–answer pairs with preference optimization. Aloe serves as a *task-specific upper bound*, whereas Tables2Traces relies primarily on reasoning traces generated from tabular data and only 10 k QA examples for format alignment.

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

Results. While Aloe attains the highest absolute accuracy, Tables2Traces closes a large share of the gap despite using roughly **75× less QA supervision**. On `MedQA`, it achieves +17% gain on cardiovascular questions (Aloe: +26%) and +10% on patient-specific questions (Aloe: +25 %) (Appendix G). Similar trends appear on `MedMCQA`. Both models struggle on abstract, non-patient-specific questions, indicating these remain structurally challenging.

4 Discussion

Tables2Traces introduces a new supervision paradigm: using tabular data as a source of structured reasoning signals. Using Tables2Traces improves medical QA performance even when trained on a single clinical domain. Because trace generation is a one-time preprocessing step, the resulting corpus can be reused across models with minimal compute. We show that tabular data can serve as a new *modality for reasoning supervision*, enabling domain adaptation and scalable alignment beyond text-based corpora.

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.

References

[1] S. Ö. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[2] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[3] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

[4] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

[5] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[6] A. K. Gururajan, E. Lopez-Cuena, J. Bayarri-Planas, A. Tormos, D. Hinjos, P. Bernabeu-Perez, A. Arias-Duart, P. A. Martin-Torres, L. Urcelay-Ganzabal, M. Gonzalez-Mallo, et al. Aloe: A family of fine-tuned open healthcare llms. *arXiv preprint arXiv:2405.01886*, 2024.

[7] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

[8] HuggingFace. Open r1: A fully open reproduction of deepseek-r1, January 2025.

[9] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[10] Q. Liu, B. Chen, J. Guo, M. Ziyadi, Z. Lin, W. Chen, and J.-G. Lou. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*, 2021.

[11] OpenAI. Openai text-embedding-3-large, 2023. `https://platform.openai.com/docs/guides/embeddings`.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[13] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[14] L. J. Palmer. Uk biobank: bank on it. *The Lancet*, 369(9578):1980–1982, 2007.

[15] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[16] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[17] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. 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.

[18] R. Wang, Z. Wang, and J. Sun. Unipredict: Large language models are universal tabular classifiers. *arXiv preprint arXiv:2310.03266*, 2023.

[19] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. 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)*, pages 13484–13508, 2023.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[21] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, 2020.

Part I

Tables2Traces Appendix

Table of Contents

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 P. 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.

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 [8], 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 2). All datasets are processed using 48 parallel workers.

Table 2: Training runtimes for each model variant.

Model Variant	Architecture	Runtime
Tables2Traces	8B (LLaMA)	20h 37m
Tables2Traces (simple)	8B (LLaMA)	9h 24m
Tables2Traces	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.

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.

Counterfactual Task Generation

```
### 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.

Patient-specific Categorization

```
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}}

Example 3:
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}
Answer:
```

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.

Medical Domain Categorization

```
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

---QUESTION---
{question}

---YOUR RESPONSE (MUST MATCH EXACTLY ONE FORMAT ABOVE)---
```

C.7 Evaluation Prompts: MedQA

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

MedQA Evaluation Prompt

```
**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

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!

_____

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 A1c (HbA1c) 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/m^2. 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 [...] µmol/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. </think>

Model Response (Redacted)

The patient is a [...]-year-old [...] [...] with a body mass index (BMI) of [...] kg/m² (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/[...]m^2 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.

Counterfactual Input (Redacted) - Part 1

```
### Role ###
Clinical AI analyzing patient outcomes using contrastive case pairs.

### Input Data ###
1) Target patient (labeled Dead)
2) Nearest neighbor who survived
3) Nearest neighbor who died

=== CLOSEST SURVIVOR ===
**Patient Description**
The patient is a **[...]-year-old [...] male**
with a body mass index (BMI) of **[...] kg/m^2** (weight: [...] kg,
height: [...] cm)), categorizing him as **overweight**,
a significant modifiable risk factor for cardiovascular disease (CVD).
He reports a **[...]-cigarette/day smoking history**, a major independent risk factor
for atherosclerotic CVD and thromboembolic events.

**Cardiovascular and Comorbidity Profile**:
- **Atrial fibrillation (AF)** is confirmed (diagnosis code present),
elevating his risk of thromboembolic complications, including stroke.
- **No diabetes mellitus** (HbA1c: [...]%, fasting glucose: [...] mmol/L))
or chronic kidney disease (CKD) (serum creatinine: [...] mumol/L,
cystatin C: [...] mg/L, urine microalbumin: [...] mg/L).
- **Uncontrolled hypertension** (systolic BP: [...] mmHg) is evident, with no current use
of anti-hypertensive medications, suggesting suboptimal risk factor management.
- **Hyperlipidemia** is present (total cholesterol: [...] mmol/L, HDL: [...] mmol/L,
triglycerides: [...] mmol/L), with an estimated LDL-C of **~[...]
mmol/L** (Friedewald equation), indicative of significant dyslipidemia.
Despite this, no lipid-lowering therapy is documented.

**Inflammatory and Metabolic Markers**:
- **C-reactive protein (CRP)** is within normal limits ([...] mg/L),
suggesting no acute systemic inflammation.
- **Uric acid** levels are borderline elevated ([...] mumol/L),
though below the threshold for clinical hyperuricemia.

**Additional Risk Stratification**:
- **No family history of CVD**, autoimmune disease (e.g., rheumatoid arthritis, SLE),
or severe mental illness.
- **Absence of diabetic, antihypertensive, or lipid-lowering pharmacotherapy**
highlights potential undertreatment of modifiable CVD risk factors.

**Clinical Synthesis**:
This patient presents with **high-risk cardiovascular profile** driven by **age,
smoking, untreated hypertension, and significant hypercholesterolemia**,
compounded by **AF-related thromboembolic risk**.
The absence of diabetes or CKD does not mitigate his overall risk,
as his ASCVD (atherosclerotic cardiovascular disease)
risk score would likely place him in a high-risk category.
Urgent interventions should prioritize **smoking cessation, BP control
(target <130/80 mmHg per guidelines), and statin therapy**
(high-intensity statin indicated for LDL-C reduction >50%).
**Anticoagulation for AF** (CHA_2DS_2-VASc score >=2 given age >=[...] and hypertension)
should be evaluated to mitigate stroke risk.
Close monitoring of renal function (cystatin C-based eGFR) and
lipid profiles is warranted to guide therapeutic efficacy and adherence.
```

Counterfactual Input (Redacted) - Part 2

```
=== CLOSEST DEATH ===
The patient is a [...]-year-old [...] male with a body mass index (BMI) of
[...] kg/m^2 (weight: [...] kg, height: [...] cm), categorizing him as overweight.
He is an active smoker with a significant tobacco exposure of [...] cigarettes/day,
a major independent risk factor for atherosclerotic cardiovascular disease (ASCVD).
His medical history is notable for atrial fibrillation (AFib),
a critical arrhythmia conferring a 5-fold increased risk of thromboembolic events,
but no evidence of chronic kidney disease (CKD),
diabetes mellitus (DM), or autoimmune disorders.

**Cardiometabolic Profile:**
- **Hypertension:** Uncontrolled stage 2 hypertension (systolic BP: [...] mmHg)
without current antihypertensive therapy. This elevates his 10-year ASCVD risk
substantially, particularly when combined with smoking.
- **Lipid Abnormalities:** Borderline-high total cholesterol ([...] mmol/L) with
elevated calculated LDL-C (~[...] mmol/L via Friedewald equation) and suboptimal
HDL-C ([...] mmol/L), consistent with atherogenic dyslipidemia.
Triglycerides are within normal limits ([...] mmol/L).
- **Glycemic Status:** Normoglycemic (fasting glucose: [...] mmol/L;
HbA1c: [...]%), excluding DM.

**Renal & Inflammatory Markers:**
- Preserved renal function: Serum creatinine [...] mumol/L (eGFR ~[...] mL/min/[...]m^2
by CKD-EPI), cystatin C [...] mg/L, and normoalbuminuria (urine microalbumin: [...] mg/L).
- Mild systemic inflammation: CRP [...] mg/L, potentially reflecting endothelial
dysfunction from smoking or subclinical atherosclerosis.

**Additional Risk Stratification:**
- Absence of protective factors: No lipid-lowering, antihypertensive,
or anticoagulant use despite AFib and ASCVD risk factors.
- No familial predisposition to CVD or contributory medication use
(e.g., corticosteroids, antipsychotics).

**Clinical Implications:**
This patient exhibits a high-risk phenotype for ASCVD
and thromboembolism due to the confluence of modifiable factors: smoking,
untreated hypertension, dyslipidemia, and AFib. The absence of diabetes
or CKD does not mitigate this risk, as his SCORE2-Diabetes equation-adjusted
10-year CVD risk likely exceeds 10%. Urgent interventions should include
smoking cessation, BP control (target <130/80 mmHg per ACC/AHA guidelines),
high-intensity statin therapy (e.g., atorvastatin 40-80 mg),
and CHA_2DS_2-VASc-guided anticoagulation for AFib. Weight optimization and
serial monitoring of inflammatory markers (e.g., CRP)
and cardiac troponins could further refine prognostication.

=== TARGET PATIENT ===
The patient is a [...]-year-old [...] male 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 2 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: [...] mumol/L; eGFR >[...] mL/min/[...] 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 ([...] mumol/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 >=2) 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.
```

Counterfactual Input (Redacted) - Part 3

```
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.

### 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 (e.g., "Target aligns with NN1 in [X] but NN2 in [Y]")

2. Label Evaluation:
   a) Assess if Dead is correct
   b) Confidence score (1-5):
      5 = All evidence strongly agrees
      4 = Most evidence agrees
      3 = Mixed evidence
      2 = Minimal supporting evidence
      1 = No discernible pattern

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: <Matches Survivor/Matches Death/Uncertain>
   2) Decisive factors:
      1) <Feature>: Target vs NN1 vs NN2
      2) <Feature>: Target vs NN1 vs NN2

2. Label assessment:
   1) Correctness: <Correct/Incorrect/Edge Case>
   2) Confidence: <1-5 with brief rationale>

3. Counterfactual:
   1) Modification: <Feature + change>
   2) Outcome: <Survived/Died>
   3) Evidence: <"Matches [NN] where [feature]=[value]">

### Quality Rules ###
- If confidence =<2, state: "Uncertain because [reason]"
- Counterfactuals must reference features actually present in NNs
- Use original feature names from input data
```

Model Reasoning (Redacted)

<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 other risks. </think>

Model Response (Redacted)

```
### 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 the MedQA benchmark for Base and Tables2Traces.

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

E Category-Level Results

E.1 MedQA

To analyze category-specific effects, we report per-specialty performance on the MedQA benchmark in Table 3. Tables2Traces improves accuracy across **16 of 18 categories** (89%), despite being fine-tuned only on cardiovascular tabular data. Largest relative gains are observed in *Renal/Genitourinary* (+29.65%), *Hematologic* (+18.98%), and *Cardiovascular* (+17.21%), with minor declines in small-sample categories such as *Other* (–4.88%) and *Pediatric* (–1.96%). Overall accuracy increases by **+9.19%** compared to the base model. These results indicate that reasoning supervision extracted from structured tabular data transfers effectively to diverse medical subfields, even those not represented in the training domain.

Table 4: 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	Base	$\mathbf{0.36 \pm 0.07}$	$\mathbf{0.88 \pm 0.07}$	$\mathbf{0.29 \pm 0.09}$	$\mathbf{0.08 \pm 0.06}$	
(N = 24)	Tables2Traces	$\mathbf{0.36 \pm 0.07}$	0.83 ± 0.08	$\mathbf{0.29 \pm 0.09}$	$\mathbf{0.08 \pm 0.06}$	-1.15% ↓
Anatomy	Base	0.36 ± 0.02	$\mathbf{0.86 \pm 0.03}$	0.26 ± 0.04	0.02 ± 0.01	
(N = 147)	Tables2Traces	$\mathbf{0.40 \pm 0.02}$	$\mathbf{0.88 \pm 0.03}$	$\mathbf{0.31 \pm 0.04}$	$\mathbf{0.07 \pm 0.02}$	+11.91% ↑
Biochemistry	Base	0.57 ± 0.03	0.90 ± 0.03	$\mathbf{0.58 \pm 0.04}$	0.11 ± 0.03	
(N = 122)	Tables2Traces	$\mathbf{0.59 \pm 0.03}$	$\mathbf{0.95 \pm 0.02}$	0.56 ± 0.05	$\mathbf{0.17 \pm 0.03}$	+3.62% ↑
Dental	Base	0.35 ± 0.01	0.82 ± 0.01	0.26 ± 0.02	0.05 ± 0.01	
(N = 845)	Tables2Traces	$\mathbf{0.39 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.06 \pm 0.01}$	+9.24% ↑
ENT	Base	0.39 ± 0.05	$\mathbf{0.92 \pm 0.04}$	0.26 ± 0.07	$\mathbf{0.08 \pm 0.04}$	
(N = 39)	Tables2Traces	$\mathbf{0.45 \pm 0.05}$	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.36 \pm 0.08}$	$\mathbf{0.08 \pm 0.04}$	+16.56% ↑
Forensic Medicine	Base	$\mathbf{0.41 \pm 0.05}$	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.32 \pm 0.07}$	0.09 ± 0.04	
(N = 44)	Tables2Traces	$\mathbf{0.41 \pm 0.05}$	$\mathbf{0.89 \pm 0.05}$	0.30 ± 0.07	$\mathbf{0.14 \pm 0.05}$	-1.10% ↓
Gynaecology & Obstetrics	Base	0.40 ± 0.03	0.81 ± 0.03	0.32 ± 0.04	$\mathbf{0.09 \pm 0.02}$	
(N = 154)	Tables2Traces	$\mathbf{0.42 \pm 0.03}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.38 \pm 0.04}$	0.08 ± 0.02	+4.03% ↑
Medicine	Base	0.44 ± 0.03	0.84 ± 0.03	0.39 ± 0.04	0.12 ± 0.02	
(N = 185)	Tables2Traces	$\mathbf{0.50 \pm 0.03}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.45 \pm 0.04}$	$\mathbf{0.15 \pm 0.03}$	+12.17% ↑
Microbiology	Base	0.45 ± 0.04	0.84 ± 0.04	$\mathbf{0.35 \pm 0.06}$	0.11 ± 0.04	
(N = 74)	Tables2Traces	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.35 \pm 0.06}$	$\mathbf{0.15 \pm 0.04}$	+7.55% ↑
Ophthalmology	Base	0.40 ± 0.05	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.30 \pm 0.07}$	0.14 ± 0.05	
(N = 43)	Tables2Traces	$\mathbf{0.41 \pm 0.05}$	0.88 ± 0.05	$\mathbf{0.30 \pm 0.07}$	$\mathbf{0.16 \pm 0.06}$	+1.72% ↑
Orthopaedics	Base	$\mathbf{0.40 \pm 0.08}$	$\mathbf{0.87 \pm 0.09}$	$\mathbf{0.53 \pm 0.13}$	$\mathbf{0.00 \pm 0.00}$	
(N = 15)	Tables2Traces	0.34 ± 0.08	0.80 ± 0.11	0.27 ± 0.12	$\mathbf{0.00 \pm 0.00}$	-15.00% ↓
Pathology	Base	0.51 ± 0.02	0.89 ± 0.02	0.44 ± 0.03	0.11 ± 0.02	
(N = 259)	Tables2Traces	$\mathbf{0.54 \pm 0.02}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.53 \pm 0.03}$	$\mathbf{0.16 \pm 0.04}$	+5.82% ↑
Pediatrics	Base	0.44 ± 0.03	0.82 ± 0.03	$\mathbf{0.39 \pm 0.04}$	0.09 ± 0.02	
(N = 133)	Tables2Traces	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.86 \pm 0.03}$	0.38 ± 0.04	$\mathbf{0.12 \pm 0.03}$	+6.52% ↑
Pharmacology	Base	0.52 ± 0.03	0.90 ± 0.02	0.46 ± 0.04	$\mathbf{0.17 \pm 0.03}$	
(N = 179)	Tables2Traces	$\mathbf{0.56 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.55 \pm 0.04}$	0.14 ± 0.03	+8.30% ↑
Physiology	Base	0.46 ± 0.03	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.38 \pm 0.04}$	$\mathbf{0.16 \pm 0.03}$	
(N = 133)	Tables2Traces	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.38 \pm 0.04}$	0.14 ± 0.03	+2.30% ↑
Psychiatry	Base	0.41 ± 0.10	0.80 ± 0.13	0.30 ± 0.15	$\mathbf{0.00 \pm 0.00}$	
(N = 10)	Tables2Traces	$\mathbf{0.54 \pm 0.09}$	$\mathbf{0.90 \pm 0.10}$	$\mathbf{0.50 \pm 0.17}$	$\mathbf{0.00 \pm 0.00}$	+31.71% ↑
Radiology	Base	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.40 \pm 0.07}$	$\mathbf{0.05 \pm 0.03}$	
(N = 57)	Tables2Traces	0.45 ± 0.04	0.89 ± 0.04	0.39 ± 0.07	0.04 ± 0.02	-8.54% ↓
Skin	Base	$\mathbf{0.37 \pm 0.08}$	$\mathbf{0.91 \pm 0.09}$	$\mathbf{0.27 \pm 0.14}$	0.00 ± 0.00	
(N = 11)	Tables2Traces	0.28 ± 0.10	0.73 ± 0.14	0.18 ± 0.12	$\mathbf{0.09 \pm 0.09}$	-24.39% ↓
Social & Preventive Medicine	Base	0.44 ± 0.04	0.81 ± 0.04	0.34 ± 0.05	$\mathbf{0.10 \pm 0.03}$	
(N = 91)	Tables2Traces	$\mathbf{0.47 \pm 0.04}$	$\mathbf{0.87 \pm 0.04}$	$\mathbf{0.43 \pm 0.05}$	0.09 ± 0.03	+7.30% ↑
Surgery	Base	0.41 ± 0.02	$\mathbf{0.86 \pm 0.02}$	0.35 ± 0.03	$\mathbf{0.08 \pm 0.02}$	
(N = 249)	Tables2Traces	$\mathbf{0.46 \pm 0.02}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.40 \pm 0.03}$	$\mathbf{0.08 \pm 0.02}$	+12.12% ↑
Unknown	Base	0.30 ± 0.30	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.00 \pm 0.00}$	
(N = 2)	Tables2Traces	$\mathbf{0.35 \pm 0.35}$	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.00 \pm 0.00}$	+16.67% ↑
Overall	Base	0.42 ± 0.01	0.85 ± 0.01	0.35 ± 0.01	0.09 ± 0.01	
(N = 2816)	Tables2Traces	$\mathbf{0.45 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	**+7.49% ↑**

E.2 MedMCQA

To further evaluate generalization, we analyze performance across medical specialties on the MedMCQA benchmark. As shown in Table 4, 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 5: 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 5, 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.

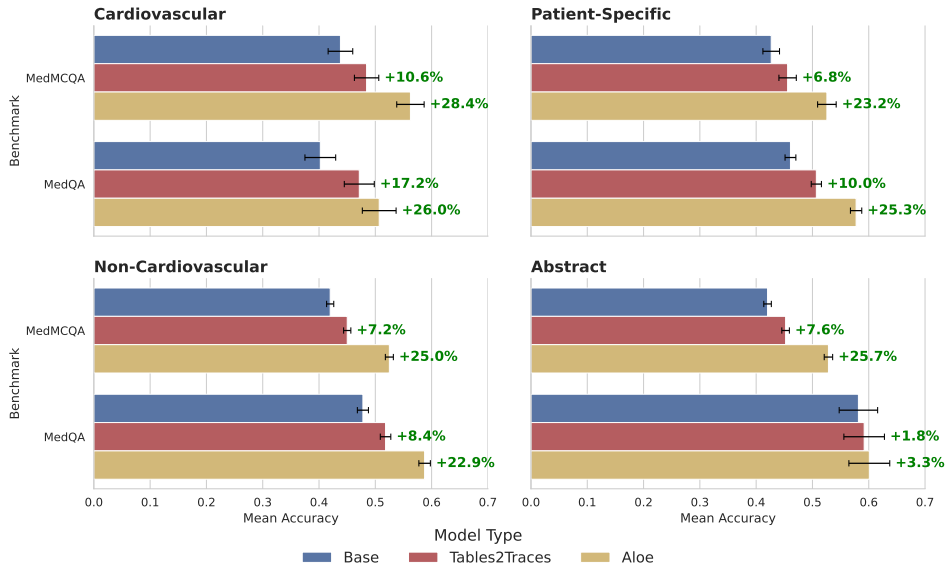

Figure 2: Accuracy comparison between **Base**, **Tables2Traces**, and **Aloe** across question categories, grouped by benchmark (MedMCQA, MedQA). Each subplot reports mean accuracy with standard error bars. Percentage improvements are relative to the Base model. Top row: in-domain categories. Bottom row: out-of-domain generalization.

G Accuracy comparison between Base, Tables2Traces and Aloe

Figure 2 compares the **Base**, **Tables2Traces**, and **Aloe** models across four subsets of medical QA questions: cardiovascular, non-cardiovascular, patient-specific, and abstract. Each subplot reports mean accuracy on `MedQA` and `MedMCQA`, with improvements relative to the Base model shown as percentages. As discussed in the main text, **Aloe** serves as a QA-optimized upper bound trained on over 750k curated QA samples, while **Tables2Traces** relies primarily on 105k tabular reasoning traces and only 10k QA-format examples. Despite this 75× smaller QA corpus, Tables2Traces closes much of the gap to Aloe—especially in cardiovascular and patient-specific subsets—highlighting that tabular reasoning traces provide a scalable, low-cost alternative to text-based supervision.

Table 6: 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

H 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.

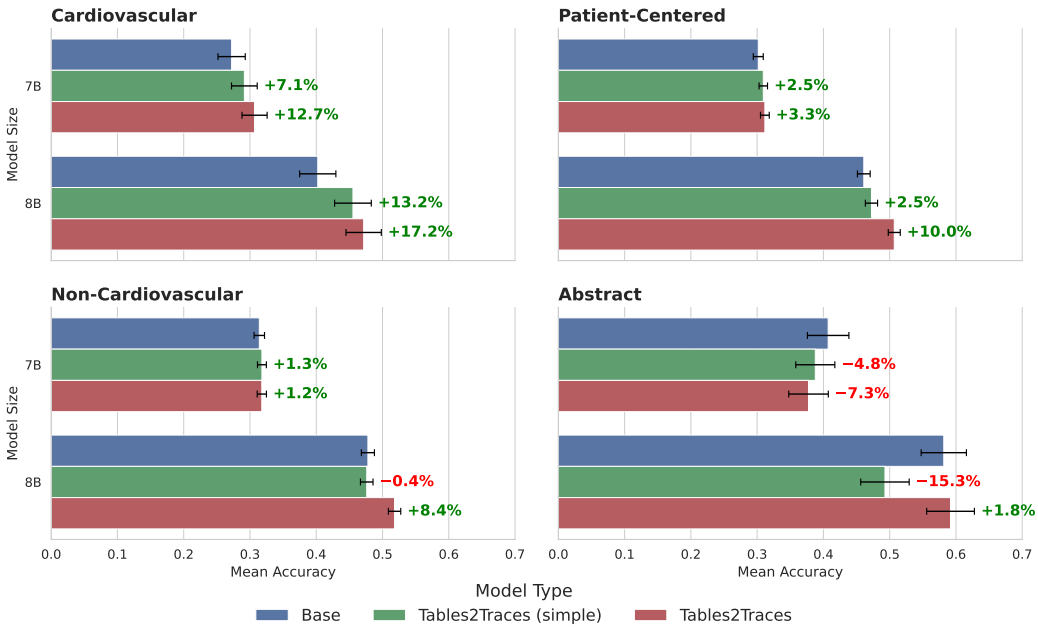

Figure 3: Accuracy on different question types in the MedQA benchmark across model sizes (7B and 8B) and fine-tuning methods. Tables2Traces yields large gains on cardiovascular and more modest gains patient-specific questions. On both cardiovascular and patient-specific questions both 7B and 8B models 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.

I 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.

I.1 MedQA

Figure 3 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 4: 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.

I.2 MedMCQA

Figure 4 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 7: 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	Tables2Traces (simple)	0.46 ± 0.03	0.86 ± 0.03	$\mathbf{0.42 \pm 0.04}$	$\mathbf{0.06 \pm 0.02}$	+13.19% ↑
(N = 130)	Tables2Traces	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.42 \pm 0.04}$	$\mathbf{0.06 \pm 0.02}$	+17.21% ↑
Dermatologic	Tables2Traces (simple)	$\mathbf{0.71 \pm 0.09}$	$\mathbf{0.94 \pm 0.06}$	$\mathbf{0.76 \pm 0.11}$	$\mathbf{0.29 \pm 0.11}$	+18.81% ↑
(N = 17)	Tables2Traces	0.60 ± 0.08	0.88 ± 0.08	0.59 ± 0.12	0.12 ± 0.08	+0.99% ↑
Endocrine/Metabolic	Tables2Traces (simple)	$\mathbf{0.52 \pm 0.02}$	$\mathbf{0.95 \pm 0.02}$	$\mathbf{0.47 \pm 0.04}$	$\mathbf{0.10 \pm 0.02}$	6.09% ↑
(N = 179)	Tables2Traces	0.51 ± 0.02	0.91 ± 0.02	0.46 ± 0.04	$\mathbf{0.10 \pm 0.02}$	4.71% ↑
Gastrointestinal	Tables2Traces (simple)	0.44 ± 0.04	0.88 ± 0.04	0.38 ± 0.05	$\mathbf{0.09 \pm 0.03}$	-5.72% ↓
(N = 86)	Tables2Traces	$\mathbf{0.50 \pm 0.04}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.47 \pm 0.05}$	0.08 ± 0.03	+6.72% ↑
Hematologic	Tables2Traces (simple)	0.38 ± 0.03	0.90 ± 0.04	0.25 ± 0.05	0.03 ± 0.02	-6.93% ↓
(N = 68)	Tables2Traces	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.43 \pm 0.06}$	$\mathbf{0.07 \pm 0.03}$	+18.98% ↑
Immunologic	Tables2Traces (simple)	0.50 ± 0.04	0.93 ± 0.03	0.44 ± 0.06	0.09 ± 0.03	-2.67% ↓
(N = 81)	Tables2Traces	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.46 \pm 0.06}$	$\mathbf{0.17 \pm 0.04}$	+6.80% ↑
Infectious	Tables2Traces (simple)	0.46 ± 0.02	0.92 ± 0.02	0.40 ± 0.04	0.06 ± 0.02	-3.44% ↓
(N = 176)	Tables2Traces	$\mathbf{0.53 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.45 \pm 0.04}$	$\mathbf{0.11 \pm 0.02}$	+9.73% ↑
Musculoskeletal	Tables2Traces (simple)	0.48 ± 0.05	0.89 ± 0.05	$\mathbf{0.42 \pm 0.07}$	0.04 ± 0.03	-2.71% ↓
(N = 45)	Tables2Traces	$\mathbf{0.51 \pm 0.04}$	$\mathbf{0.96 \pm 0.03}$	0.40 ± 0.07	$\mathbf{0.07 \pm 0.04}$	+4.07% ↑
Neurological	Tables2Traces (simple)	0.47 ± 0.04	0.86 ± 0.04	0.42 ± 0.06	$\mathbf{0.09 \pm 0.03}$	+6.89% ↑
(N = 77)	Tables2Traces	$\mathbf{0.50 \pm 0.04}$	$\mathbf{0.90 \pm 0.04}$	$\mathbf{0.43 \pm 0.06}$	0.05 ± 0.02	+15.15% ↑
Obstetrics/Gynecology	Tables2Traces (simple)	0.46 ± 0.04	0.93 ± 0.03	$\mathbf{0.43 \pm 0.06}$	$\mathbf{0.07 \pm 0.03}$	+0.93% ↑
(N = 70)	Tables2Traces	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.94 \pm 0.03}$	0.40 ± 0.06	0.03 ± 0.02	+2.80% ↑
Oncology	Tables2Traces (simple)	0.50 ± 0.04	0.86 ± 0.04	0.47 ± 0.06	$\mathbf{0.15 \pm 0.04}$	-4.76% ↓
(N = 72)	Tables2Traces	$\mathbf{0.56 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.53 \pm 0.06}$	0.14 ± 0.04	+5.82% ↑
Other	Tables2Traces (simple)	0.45 ± 0.06	0.84 ± 0.07	$\mathbf{0.45 \pm 0.09}$	0.10 ± 0.05	-15.24% ↓
(N = 31)	Tables2Traces	$\mathbf{0.50 \pm 0.07}$	$\mathbf{0.87 \pm 0.06}$	0.42 ± 0.09	$\mathbf{0.19 \pm 0.07}$	-4.88% ↓
Pediatric	Tables2Traces (simple)	0.32 ± 0.07	0.92 ± 0.08	0.23 ± 0.12	$\mathbf{0.00 \pm 0.00}$	-19.61% ↓
(N = 13)	Tables2Traces	$\mathbf{0.39 \pm 0.05}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.31 \pm 0.13}$	$\mathbf{0.00 \pm 0.00}$	-1.96% ↓
Psychiatric	Tables2Traces (simple)	0.60 ± 0.05	$\mathbf{0.94 \pm 0.03}$	0.58 ± 0.07	$\mathbf{0.21 \pm 0.06}$	+2.30% ↑
(N = 52)	Tables2Traces	$\mathbf{0.62 \pm 0.05}$	0.90 ± 0.04	$\mathbf{0.61 \pm 0.07}$	$\mathbf{0.21 \pm 0.06}$	+5.57% ↑
Renal/Genitourinary	Tables2Traces (simple)	0.42 ± 0.04	0.93 ± 0.04	0.35 ± 0.07	0.06 ± 0.03	+15.08% ↑
(N = 54)	Tables2Traces	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.41 \pm 0.07}$	$\mathbf{0.09 \pm 0.04}$	+29.65% ↑
Respiratory	Tables2Traces (simple)	0.49 ± 0.04	0.93 ± 0.04	0.44 ± 0.07	$\mathbf{0.11 \pm 0.04}$	+0.76% ↑
(N = 54)	Tables2Traces	$\mathbf{0.50 \pm 0.04}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.46 \pm 0.07}$	$\mathbf{0.11 \pm 0.04}$	+2.28% ↑
Toxicology	Tables2Traces (simple)	0.41 ± 0.04	$\mathbf{0.93 \pm 0.03}$	0.35 ± 0.06	0.03 ± 0.02	-6.10% ↓
(N = 68)	Tables2Traces	$\mathbf{0.52 \pm 0.04}$	0.91 ± 0.04	$\mathbf{0.47 \pm 0.06}$	$\mathbf{0.09 \pm 0.04}$	+20.68% ↑
Overall	Tables2Traces (simple)	0.47 ± 0.01	0.91 ± 0.01	0.42 ± 0.01	0.08 ± 0.01	+0.82% ↑
(N = 1273)	Tables2Traces	$\mathbf{0.51 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.46 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	+9.19% ↑

J Per-category results from Tables2Traces (simple)

Table 7 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 8: 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	Base	0.36 ± 0.07	$\mathbf{0.88 \pm 0.07}$	0.29 ± 0.09	$\mathbf{0.08 \pm 0.06}$	
(N = 24)	Aloe	$\mathbf{0.46 \pm 0.07}$	0.79 ± 0.08	$\mathbf{0.46 \pm 0.10}$	$\mathbf{0.08 \pm 0.06}$	+26.44%↑
Anatomy	Base	0.36 ± 0.02	$\mathbf{0.86 \pm 0.03}$	0.26 ± 0.04	0.02 ± 0.01	
(N = 147)	Aloe	$\mathbf{0.49 \pm 0.03}$	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.42 \pm 0.04}$	$\mathbf{0.14 \pm 0.03}$	+37.05%↑
Biochemistry	Base	0.57 ± 0.03	0.90 ± 0.03	0.58 ± 0.04	0.11 ± 0.03	
(N = 122)	Aloe	$\mathbf{0.69 \pm 0.03}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.38 \pm 0.04}$	+21.59%↑
Dental	Base	0.35 ± 0.01	0.82 ± 0.01	0.26 ± 0.02	0.05 ± 0.01	
(N = 845)	Aloe	$\mathbf{0.41 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.34 \pm 0.02}$	$\mathbf{0.11 \pm 0.01}$	+15.61%↑
ENT	Base	0.39 ± 0.05	$\mathbf{0.92 \pm 0.04}$	0.26 ± 0.07	0.08 ± 0.04	
(N = 39)	Aloe	$\mathbf{0.55 \pm 0.06}$	0.90 ± 0.05	$\mathbf{0.51 \pm 0.08}$	$\mathbf{0.23 \pm 0.07}$	+41.06%↑
Forensic Medicine	Base	0.41 ± 0.05	$\mathbf{0.89 \pm 0.05}$	0.32 ± 0.07	0.09 ± 0.04	
(N = 44)	Aloe	$\mathbf{0.50 \pm 0.05}$	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.43 \pm 0.08}$	$\mathbf{0.20 \pm 0.06}$	+20.88%↑
Gynaecology & Obstetrics	Base	0.40 ± 0.03	0.81 ± 0.03	0.32 ± 0.04	0.09 ± 0.02	
(N = 154)	Aloe	$\mathbf{0.53 \pm 0.03}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.46 \pm 0.04}$	$\mathbf{0.21 \pm 0.03}$	+30.43%↑
Medicine	Base	0.44 ± 0.03	0.84 ± 0.03	0.39 ± 0.04	0.12 ± 0.02	
(N = 185)	Aloe	$\mathbf{0.58 \pm 0.03}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.26 \pm 0.03}$	+29.93%↑
Microbiology	Base	0.45 ± 0.04	0.84 ± 0.04	0.35 ± 0.06	0.11 ± 0.04	
(N = 74)	Aloe	$\mathbf{0.58 \pm 0.04}$	$\mathbf{0.89 \pm 0.04}$	$\mathbf{0.55 \pm 0.06}$	$\mathbf{0.24 \pm 0.05}$	+29.31%↑
Ophthalmology	Base	0.40 ± 0.05	0.91 ± 0.04	0.30 ± 0.07	0.14 ± 0.05	
(N = 43)	Aloe	$\mathbf{0.54 \pm 0.06}$	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.49 \pm 0.08}$	$\mathbf{0.23 \pm 0.07}$	+34.48%↑
Orthopaedics	Base	0.40 ± 0.08	0.87 ± 0.09	0.53 ± 0.13	0.00 ± 0.00	
(N = 15)	Aloe	$\mathbf{0.59 \pm 0.08}$	$\mathbf{0.93 \pm 0.07}$	$\mathbf{0.60 \pm 0.13}$	$\mathbf{0.13 \pm 0.09}$	+46.67%↑
Pathology	Base	0.51 ± 0.02	0.89 ± 0.02	0.44 ± 0.03	0.11 ± 0.02	
(N = 259)	Aloe	$\mathbf{0.65 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.64 \pm 0.03}$	$\mathbf{0.32 \pm 0.03}$	+27.73%↑
Pediatrics	Base	0.44 ± 0.03	0.82 ± 0.03	0.39 ± 0.04	0.09 ± 0.02	
(N = 133)	Aloe	$\mathbf{0.57 \pm 0.03}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.51 \pm 0.04}$	$\mathbf{0.19 \pm 0.03}$	+31.05%↑
Pharmacology	Base	0.52 ± 0.03	0.90 ± 0.02	0.46 ± 0.04	0.17 ± 0.03	
(N = 179)	Aloe	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.69 \pm 0.03}$	$\mathbf{0.43 \pm 0.04}$	+38.04%↑
Physiology	Base	0.46 ± 0.03	0.86 ± 0.03	0.38 ± 0.04	0.16 ± 0.03	
(N = 133)	Aloe	$\mathbf{0.60 \pm 0.03}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.58 \pm 0.04}$	$\mathbf{0.29 \pm 0.04}$	+31.2%↑
Psychiatry	Base	0.41 ± 0.10	0.80 ± 0.13	0.30 ± 0.15	0.00 ± 0.00	
(N = 10)	Aloe	$\mathbf{0.60 \pm 0.13}$	$\mathbf{0.90 \pm 0.10}$	$\mathbf{0.60 \pm 0.16}$	$\mathbf{0.30 \pm 0.15}$	+46.34%↑
Radiology	Base	0.49 ± 0.04	$\mathbf{0.93 \pm 0.03}$	0.40 ± 0.07	0.05 ± 0.03	
(N = 57)	Aloe	$\mathbf{0.51 \pm 0.05}$	0.89 ± 0.04	$\mathbf{0.44 \pm 0.07}$	$\mathbf{0.14 \pm 0.05}$	+2.85%↑
Skin	Base	0.37 ± 0.08	$\mathbf{0.91 \pm 0.09}$	0.27 ± 0.14	0.00 ± 0.00	
(N = 11)	Aloe	$\mathbf{0.47 \pm 0.12}$	0.73 ± 0.14	$\mathbf{0.36 \pm 0.15}$	$\mathbf{0.18 \pm 0.12}$	+26.83%↑
Social & Preventive Medicine	Base	0.44 ± 0.04	0.81 ± 0.04	0.34 ± 0.05	0.10 ± 0.03	
(N = 91)	Aloe	$\mathbf{0.52 \pm 0.04}$	$\mathbf{0.88 \pm 0.03}$	$\mathbf{0.47 \pm 0.05}$	$\mathbf{0.20 \pm 0.04}$	+19.14%↑
Surgery	Base	0.41 ± 0.02	$\mathbf{0.86 \pm 0.02}$	0.35 ± 0.03	0.08 ± 0.02	
(N = 249)	Aloe	$\mathbf{0.51 \pm 0.02}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.17 \pm 0.02}$	+24.24%↑
Unknown	Base	$\mathbf{0.30 \pm 0.30}$	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.50 \pm 0.50}$	$\mathbf{0.00 \pm 0.00}$	
(N = 2)	Aloe	0.20 ± 0.20	$\mathbf{0.50 \pm 0.50}$	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$	-33.33%↓
Overall	Base	0.42 ± 0.01	0.85 ± 0.01	0.35 ± 0.01	0.09 ± 0.01	
(N = 2816)	Aloe	$\mathbf{0.53 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.21 \pm 0.01}$	+25.31%↑

K Per-category results from Aloe

For completeness, we report a category-level breakdown of Aloe's performance on the MedMCQA benchmark in Table 8. 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.

L UMAP visualizations of embedded questions

We embed all `MedQA` and `MedMCQA` questions using `text-embedding-3-large` [11] 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. Figures show two panels: (a) **Tables2Traces**, and (b) **Tables2Traces (simple)**

L.1 UMAP visualization of MedQA

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.

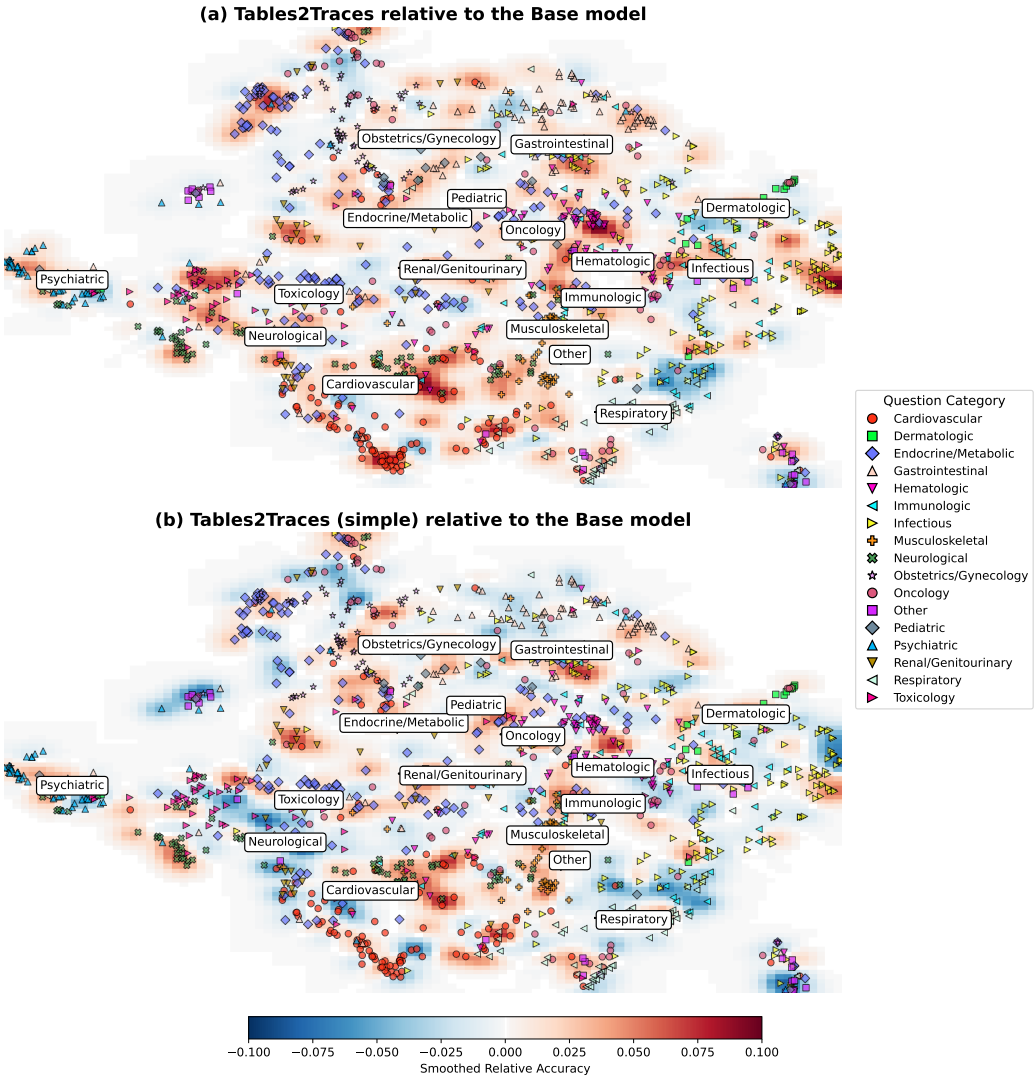

Figure 5: UMAP visualization of MedQA test questions comparing model performance to the **Base** model. Each point is a question, embedded using `text-embedding-3-large` [11], and annotated by medical category using distinct marker shapes and colors. Accuracy is smoothed using a Gaussian filter ($\sigma = 1.5$). 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.

L.2 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.

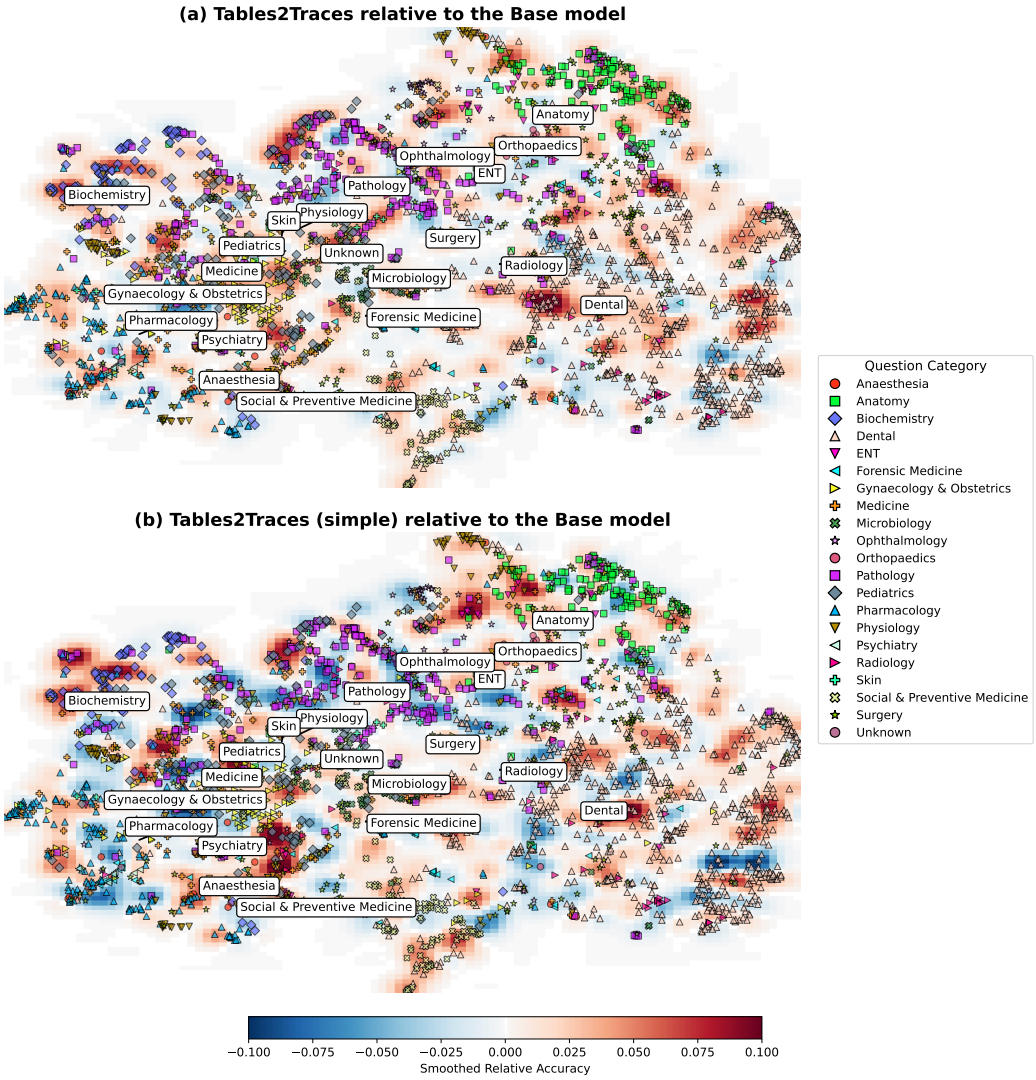

Figure 6: UMAP visualization of MedMCQA test questions, comparing model performance to the **Base** model. Each point represents a question, embedded using `text-embedding-3-large` [11], and annotated by medical category using distinct marker shapes and colors. The background heatmap reflects smoothed 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.

M 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 9 (MedQA) and Table 10 (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 the MedQA benchmark for Base and QA-Finetuning (8B).

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

N 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.

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

O 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 11–12.

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 over-confidence 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 11: 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 12: 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)**

P 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 13: 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 misaligned	"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)."