Metric-Fair Prompting: Treating Similar Samples Similarly

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce *Metric-Fair Prompting*, a fairness-aware prompting framework that steers large language models (LLMs) to make decisions analogously to a margin-based binary classifier under a metric-fairness constraint. In our formulation of multiple-choice medical QA, each *(question, option)* pair is treated as a binary instance with label $+1$ (correct) or -1 (incorrect). To promote *individual fairness*—treating similar instances similarly—we compute question similarity using NLP embeddings and solve items in *joint pairs of similar questions* rather than in isolation. The prompt enforces a global decision protocol: extract decisive clinical features, map each $(\text{question}, \text{option})$ to a half-space score $f(x)$, use the margin $|f(x)|$ as confidence, and impose a Lipschitz-style constraint so that similar inputs receive similar scores and, hence, consistent outcomes. This joint, metric-fair perspective encourages cross-item consistency, reduces near-boundary errors, and preserves fairness by avoiding reliance on non-deterministic demographic attributes. Evaluated on the **MedQA (US)** benchmark, Metric-Fair Prompting improves performance over standard single-item prompting, demonstrating that fairness-guided, margin-oriented reasoning can enhance LLM accuracy on high-stakes clinical multiple-choice questions.

1 Introduction

Machine learning systems used for decision support can systematically disadvantage certain populations if fairness is not addressed [6, 2, 3]. These concerns extend to large language models (LLMs), which are increasingly applied in high-stakes domains such as clinical decision making and medical examinations. In such settings, it is crucial to promote *individual fairness*, treating similar instances similarly, and to base predictions on clinically determinative features rather than sensitive attributes (e.g., age, sex, race) unless those attributes are explicitly and directly relevant to the clinical task.

We study multiple-choice medical question answering (MedQA) [10], formulating each *(question, option)* pair as a binary instance ($+1$ correct, -1 incorrect). We introduce *Metric-Fair Prompting*, a prompting framework that guides an LLM to act analogously to a margin-based classifier under a metric fairness constraint. Concretely, we (i) compute similarity among questions using text embeddings to identify pairs of *deterministically* similar items; (ii) present similar questions jointly so the model can enforce cross-item consistency; and (iii) ask the model to extract decisive clinical features and map each pair $(\text{question}, \text{option})$ to a half-space score $f(x)$, using $|f(x)|$ as a margin (confidence). A Lipschitz-style constraint operationalizes fairness: if two inputs x, x' are close under a task-relevant metric $d(\cdot, \cdot)$, then their scores should remain close, encouraging consistent outcomes for clinically similar cases [6].

Our approach complements prior prompt-engineering methods that improve reasoning via intermediate structure, such as chain-of-thought [18], self-consistency [17], and search-based prompting

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

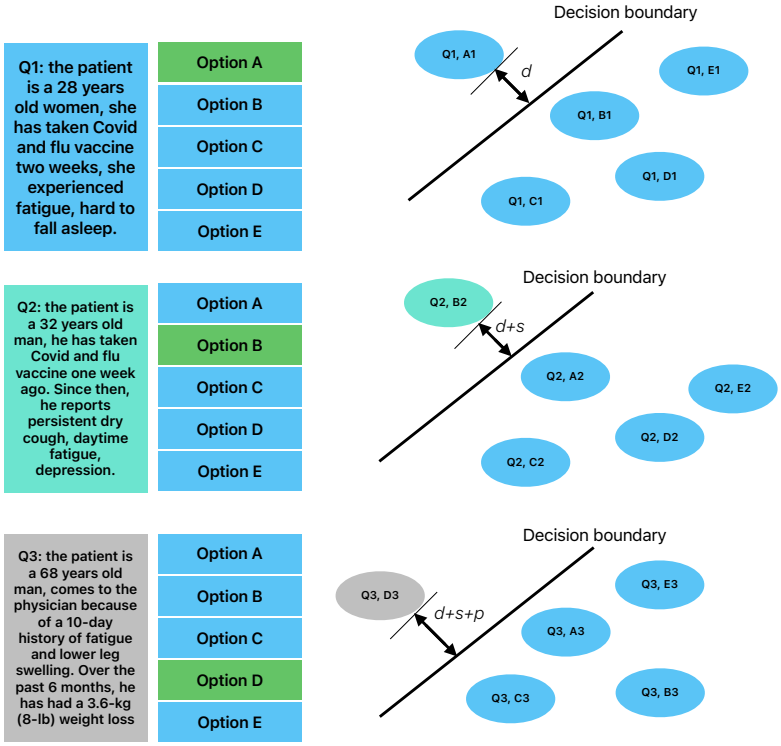

Figure 1: Geometric view of Metric-Fair Prompting. Questions 1 and 2 are highly similar (small metric distance d); their correct options lie on the same side of the decision boundary with nearby margins d and $d + s$ ($s > 0$ small). Question 3 is less similar to Question 1 (larger distance $d + s$, $p > 0$): its correct option remains in the same half-space but with a more separated margin $d + s + p$. The metric-fair (Lipschitz-like) constraint encourages similar items to receive similar scores and thus consistent decisions.

(e.g., Tree-of-Thoughts [20] and ReAct [21]). Unlike these techniques, which typically treat items independently and optimize intra-item reasoning, Metric-Fair Prompting explicitly introduces an *inter-item* coupling via a similarity metric, thereby promoting fairness (through stability to small, clinically irrelevant changes) and improving robustness on near-boundary decisions.

Contributions. (1) We propose a fairness-aware prompting framework that treats MedQA as binary classification over (question, option) pairs and enforces a metric-based Lipschitz constraint to encourage individual fairness. (2) We introduce a joint-inference protocol that feeds *pairs of similar questions* to the LLM, enabling cross-item consistency and reducing near-boundary errors. (3) On MedQA (US), the proposed protocol improves accuracy over single-item prompting (details in §**??**), illustrating that fairness-guided, margin-oriented reasoning can enhance LLM performance in clinical multiple-choice settings.

Figure 1 illustrates the motivation. Questions 1 and 2 exhibit high similarity under the embedding metric (small d); their stems and correct options map to nearby points in the margin half-space. By contrast, Question 3 is less similar to Question 1 (larger d); although its correct option falls on the same side of the decision boundary, it lies farther from Question 1's point in feature space.

2 Problem Setting

We study metric-fair learning on a domain $\mathcal{X}$ endowed with a similarity metric $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$. A learning algorithm receives d and an i.i.d. sample from a distribution $\mathcal{D}$ over labeled examples

Table 1: MedQA–US examples: two patients with similar deterministic features yield similar correct options.

Item	Content
Question 1 (Q1)	A 62-year-old man presents with 5 days of fatigue, fever, and chills. He has a 9-month history of hand pain and stiffness and started a new medication 3 months ago; prior meds included ibuprofen, prednisone, and hydroxychloroquine. He does not smoke or drink. Exam: subcutaneous nodule at left elbow, old joint destruction with boutonnière deformity, no active synovitis. Labs: Hb 10.5 g/dL, WBC 3500/mm^3, platelets 100,000/mm^3. Which of the following is most likely to have prevented these laboratory abnormalities?
Options for Q1	A) Cobalamin B) Amifostine C) Pyridoxine D) Leucovorin E) Mesna
Question 2 (Q2)	A 58-year-old woman presents with 1 week of worsening fatigue and a 1-year history of hand pain and stiffness. She started a new medication 4 months ago; prior meds included ibuprofen, prednisone, and hydroxychloroquine. Exam: subcutaneous nodule at left elbow, old joint destruction with Boutonnière deformity. Labs: Hb 10.1 g/dL, WBC 3400/mm^3, platelets 101,000/mm^3; methylmalonic acid normal. Which of the following could have prevented these laboratory abnormalities?
Options for Q2	A) Vitamin B6 B) Vitamin B12 C) Amifostine D) 2-Mercaptoethanesulfonate E) Leucovorin
Deterministic features	RA phenotype + new DMARD started months earlier + pancytopenia (anemia, leukopenia, thrombocytopenia) + normal MMA $\Rightarrow$ methotrexate-related folate pathway toxicity. Prevention: folate supplementation or folinic acid (leucovorin) rescue.
Correct answers	Q1: D Q2: E

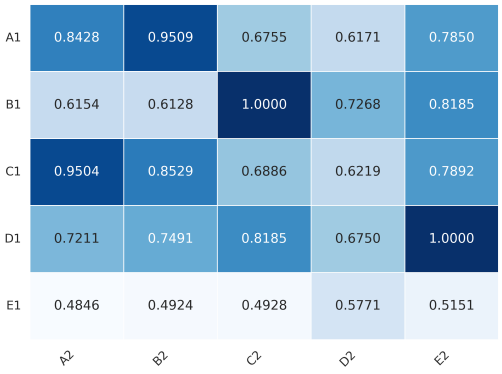

Figure 2: Correlations between the options of Question 1 and Question 2 in Table 1 by Qwen3-8B embedding.

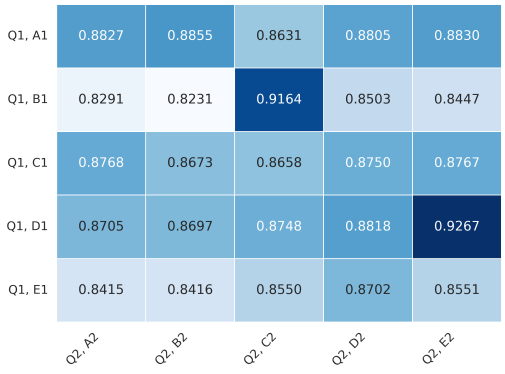

Figure 3: Correlations between questions and options from Table 1 by Qwen3-8B embedding.

$(x, y) \in \mathcal{X} \times \{\pm 1\}$ and outputs a classifier. To accommodate fairness, we focus on *probabilistic* classifiers $h : \mathcal{X} \to [0, 1]$ and interpret $h(x)$ as the probability of label $+1$ (so the probability of -1 is $1 - h(x)$). We refer to such probabilistic classifiers as *predictors*.

Our fairness principle is *treat similar individuals similarly*: two individuals that are close under d should receive similar predictive distributions. Formally, we view a classifier as a randomized mapping $M : \mathcal{X} \to \Delta(\{\pm 1\})$ assigning to each x a distribution $M(x)$ over outcomes. We require a Lipschitz-type constraint that upper-bounds the statistical distance between $M(x)$ and $M(y)$ by their feature-space distance $d(x, y)$.

Definition 1 (Lipschitz mapping)**.** *Let D be a statistical distance on distributions (e.g., total variation). A mapping $M : \mathcal{X} \to \Delta(\{\pm 1\})$ satisfies the (D, d)-Lipschitz property if for all $x, y \in \mathcal{X}$,*

$$D\big(M(x), M(y)\big) \;\le\; d(x, y). \tag{1}$$

When D and d are clear from context we simply say that M is Lipschitz*.*

Given a loss function $L : \mathcal{X} \times \{\pm 1\} \to \mathbb{R}$, our goal is to find a mapping M that minimizes expected loss subject to the Lipschitz fairness constraint. This setup naturally yields the optimization problem:

$$\min_{M:\mathcal{X}\to\Delta(\{\pm1\})} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{E}_{\hat{y}\sim M(x)}\big[L(x,\hat{y})\big] \quad \text{s.t.} \quad D\big(M(x), M(y)\big) \le d(x,y) \;\; \forall x, y.$$

When you make wrong prediction, there is an error. The goal is to after providing answers for all the questions, the loss is minimized. And you have developed a good halfspace that separates the correct options from incorrect options.

For intuition we connect to large-margin classification. Consider a binary classifier with a score $f : \mathcal{X} \to \mathbb{R}$ and margin-based decision rule $\mathrm{sign}(f(x))$. The margin $|f(x)|$ quantifies confidence and induces a *half-space* separating the two classes. In our application to multiple-choice medical QA, each *option* paired with its *question* is treated as an input x, and the classifier maps $(\text{question}, \text{option})$ to $\{0, 1\}$ (incorrect vs. correct). Metric fairness requires that if two question–option pairs are similar under d, then their predictive distributions (and thus their distances to the decision boundary) should also be similar; conversely, dissimilar pairs may legitimately receive different predictions.

3 Method

Motivation. In complex reasoning tasks such as medical examinations, distinct items can exhibit substantial semantic and clinical overlap. As illustrated in Table 1, two stems may share *deterministic* features (e.g., key signs, pathognomonic labs), and consequently their correct options tend to align. We operationalize this by computing similarity between questions and between *(question, option)* pairs using sentence embeddings (e.g., Qwen3-4B). Empirically, the correct option for a given stem has higher similarity to the stem than distractors, and stems that are similar under the embedding metric $d(\cdot, \cdot)$ often share clinically consistent answer patterns.

Overview. We propose *Metric-Fair Prompting*, a joint-inference protocol that treats each $(\text{question}, \text{option})$ as a binary instance and guides the LLM to behave like a margin-based classifier under a metric-fairness constraint. Let $x = \phi(\text{question}, \text{option}) \in \mathcal{X}$ be a feature representation and $f : \mathcal{X} \to \mathbb{R}$ a scoring function (half-space). The predicted label is $y = \mathbf{1}\{f(x) > 0\}$ with confidence $|f(x)|$. To promote individual fairness (*treat similar instances similarly*), we impose a Lipschitz-like stability: for all $x, x' \in \mathcal{X}$,

$$\big|f(x) - f(x')\big| \;\le\; L\, d(x, x'),$$

where d is a task-relevant similarity metric and $L > 0$ is a constant. Operationally, similar stems (and their options) should receive similar scores and hence consistent outcomes, unless a clear clinical contradiction exists.

Pipeline. Given a pool of inference items, we follow five steps:

(i) **Pair selection.** Embed all stems; for each stem select its nearest neighbor under $d(\cdot, \cdot)$ to form a two-item batch (high-similarity pair).

(ii) **Metric fairness.** Instruct the LLM that similar items should yield similar decisions (Lipschitz-like constraint), and that decisions must rely on clinically determinative features rather than sensitive attributes.

(iii) **Margin/half-space reasoning.** For each $(\text{question}, \text{option})$, score $f(x)$, first eliminating clear negatives (large negative margin), then resolving near-boundary candidates using decisive clinical discriminators (guidelines, pathognomonic findings, contraindications).

(iv) **Cross-item consistency.** Reconcile near ties within the pair by preferring choices that maintain consistency across similar items under $d(\cdot, \cdot)$.

(v) **Strict output.** Emit only machine-parsable results (e.g., JSON with `{"index": i, "answer": "A|B|C|D|E"}` for $i \in \{1, 2\}$).

Jointly presenting similar items introduces an inter-item coupling that stabilizes decisions near the boundary, encourages fairness via Lipschitz stability, and reduces reliance on spurious cues. This margin-oriented, metric-fair prompting improves robustness on clinically proximate stems while preserving a strict, parseable interface for evaluation.

Table 2: MedQA (US) test accuracy (%). Qwen3-14B with Metric-Fair Prompting substantially outperforms single-item prompting.

Model / Prompt	Single-item	Metric-Fair (two-item)
Qwen3-14B	68.0	**84.0**

4 Experiments

Setup. All experiments were conducted on a single NVIDIA RTX 6000 Ada.[1] We evaluate Qwen models [1] from HuggingFace (via `transformers`) with Unsloth optimizations for efficient inference. Unless stated otherwise, we use a low temperature (T=0.2) and greedy decoding (`do_sample=False`) to stabilize multiple–choice predictions.

Dataset. We use the MedQA (US) test split (N=1,273 items), each with one correct option (A–E). We frame each *(question, option)* as a binary instance (correct vs. incorrect).

Similarity and pairing. We embed every question with Qwen3-4B embeddings and compute cosine similarity. For each question q, we select its nearest neighbor q' (excluding itself) to form a *two-item* batch (q, q'). This produces N pairs (some questions may appear in multiple pairs as a neighbor). The top-3 example pairs by similarity have scores 0.9612, 0.9020, and 0.8314.

Prompting protocol. We apply the *Metric-Fair Prompting* template in Table 3: (i) jointly read both questions; (ii) enforce a metric-fair (Lipschitz-like) constraint over similar items; (iii) use a margin/half-space decision rule over $(\text{question}, \text{option})$ features; (iv) reconcile near-boundary choices to maintain cross-item consistency; and (v) output strict JSON containing only the two answers.

Conflict resolution. Because a question can appear in multiple pairs, it may receive two predictions. When predictions disagree, we trigger a light-weight *review prompt* that asks the model to re-evaluate both items jointly and to output (A–E) with a scalar confidence. We keep the answer with higher confidence; if confidences tie, we prefer the answer with larger decision margin (when available) or fall back to the original single-item prediction.

Baselines and metric. The main baseline is *single-item prompting* (standard instruction, one question at a time). We report **accuracy** (%) on the test split.

Results. Metric-Fair Prompting with Qwen3-14B improves accuracy from $\mathbf{68}\%$ (single-item) to $\mathbf{84}\%$ (two-item, metric-fair, joint inference), demonstrating that coupling similar items via a fairness-aware, margin-oriented protocol yields substantial gains on near-boundary decisions and promotes cross-item consistency.

4.1 Qualitative Examples

We illustrate how Metric-Fair Prompting enforces cross-item consistency and fairness on three highly similar question pairs selected by cosine similarity of stem embeddings (Qwen3-4B). In each case, the model reads both items jointly, extracts decisive clinical/statistical features, and applies a margin-based decision with a Lipschitz-like stability constraint.

Near-duplicate clinical stems (cosine = 0.9612). Figure 4 shows two stems that are clinically indistinguishable with respect to decisive features (symptoms, exam, labs, biopsy). The only difference is age, which is *not* determinative for the underlying mechanism queried.

Shared evidence, different foci (cosine = 0.9020). In Figure 5, both stems reference the same study abstract but ask distinct—yet related—questions: one on interpretation of the standard error (sample size/variability), the other on choosing an inferential method for group differences.

[1] No multi-GPU or model parallelism was used.

Table 3: Metric-Fair Prompt with a Binary Margin Classifier.

Objective	Read *both* questions and their options jointly; decide one option (A–E) per question. Output `{"index": i, "answer": "A|B|C|D|E"}` for each $i \in \{1, 2\}$ (JSON only).
Formulation	Consider potential correlation between the two questions. Identify deterministic clinical features that may be shared by the patients/situations. Items that are similar under a task-relevant metric should receive similar decisions (fairness-by-similarity).
Feature selection	For each question, extract the most important features from the stem (signs, key labs/imaging, contraindications, guideline thresholds). For each option, form a feature representation $x = \phi(\text{question}, \text{option})$.
Margin-based classifier	Use a binary large-margin classifier $f : \mathcal{X} \to \mathbb{R}$ on x. The predicted label is $y = \mathbf{1}\{f(x) > 0\} \in \{0, 1\}$ (1 = correct), with confidence magnitude $\lvert f(x)\rvert$. Select, for each question, the option with the largest positive margin.
Fairness	Let $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a similarity metric on question–option pairs. Enforce a Lipschitz-like constraint so that similar inputs yield similar scores. If two questions are similar in decisive clinical features, prefer consistent answer patterns unless a clear clinical conflict exists.
Cross-item reconciliation	If two options (within or across the two questions) are near the decision boundary, re-check decisive discriminators and prefer the choice that maintains cross-item consistency under $d(\cdot, \cdot)$ and standard clinical guidance.
Output format	JSON only, no prose. Example: `[{"index": 1, "answer": "C"}, {"index": 2, "answer": "A"}]`

Methodological notes. For each question q, we retrieve its nearest neighbor q' by cosine similarity on normalized sentence embeddings. The pair (q, q') is fed to the LLM with the joint protocol (Table 3). The model internally clusters decisive features, eliminates clear negatives (large negative margins), and resolves near-boundary options while enforcing a Lipschitz-like stability: if $d(q, q')$ is small, then the score difference $\lvert f(x) - f(x')\rvert$ remains small, promoting consistent outputs unless contradicted by a decisive discriminator (e.g., a contraindication).

Fairness and robustness. Across the examples, demographic attributes (e.g., age) are used only when explicitly clinically determinative; otherwise they are down-weighted by the prompt's fairness guard. The pairwise setting serves as a *regularizer* against spurious cues: when two similar items are solved jointly, option choices that are inconsistent across the pair are penalized by the metric constraint, improving reliability on near-boundary decisions.

Implementation notes. We normalize embeddings, use cosine similarity for pairing, and freeze model weights (no fine-tuning). All prompts produce strict JSON outputs to simplify parsing and evaluation. Hyperparameters (temperature, max tokens) are kept minimal to avoid confounding factors; we observed that lower temperatures further stabilize pairwise reconciliation.

5 Related Work

There is a growing body of work attempting to study the question of algorithmic discrimination. This literature is characterized by high-level distinction, group and individual notions of fairness. We also initroduce the prompt engineering related works.

Individual fairness posits that "similar individuals should be treated similarly" [6]. This powerful guarantees is formalized via a Lipschitz condition on the classifier mapping individuals to distributions over outcomes. Recent works study study different individual level fairness in the contexts of reinforcement and online learning. study different individual level fairness in the contexts of bandit [11]. [15] studies metric-fair active learning of homogeneous halfspaces, and show that under the distribution-dependent PAC learning model. Fairness and label efficiency can be achieved simultaneously.

Figure 4: Example of LLM output for two questions with cosine similarity 0. 9612 (embedding by Qwen3-4B embedding) given our prompt. The two patients are identical in clinical features, lab results and biopsy findings. The only difference is the age, which is not a distinguishing factor. Hence the correct option is same, "Adverse effect of anesthetic".

Group fairness notions assume the existence of a protected attribute (e.g. gender, race), which induces a partition of the instance space into some small number of groups. A fair classifier is one that achieves parit of some statistical measure across these groups. There are prominent measures include classification rates (statistical parity [7], calibration, and false positive or negative rates [8]. [19] incorporating the fairness notion of [8] into a statistical and computational theory of learning, and proposed a relaxation of the fairness definition to make it feasible to optimize the learning objective.

LMs are known to have already absorbed rich commonsense that makes it possible to propose reasonable plans conditioned on problem setting [4, 9, 16]. The idea of Chain-of-Thought wei 2022b is to enrich the few-shot exmaples with reasoning steps. There are many studies to improve the performance of DoT on complex stasks such as dynamic least-to-most prompting [5], active Prompt [4]. Tree of thought approach extends planning formlations by conidering multiple potentially feasible plans simultaneously at each problem-solving step [20]. [13] introduced the "self-reflection" mechanism, LMs provide feedback to their generation candidates. [12] introduces review steps cover the actions and states, deciding the next action. "self-guided decoding" followers a tree-search procedure with leaves sampled from stochastic beam search decoding. Our prompt teaches LMs to think like a machine learning algorithm with constraint and objective function, the constraint is the

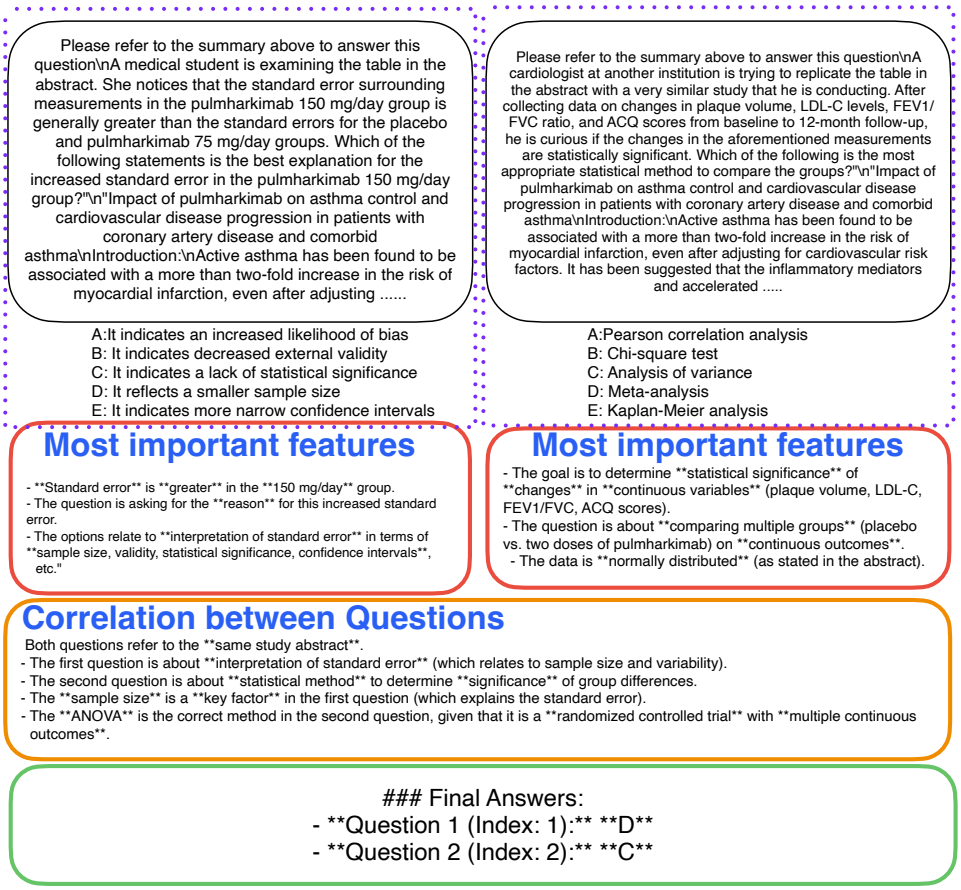

Figure 5: Example of LLM output for two questions with cosine similarity 0.9020 (embedding by Qwen3-4B embedding) given our prompt. Both questions refer to the same study abstract. The first question is about interpretation of standard error (which relates to sample size and variability). The second question is about statistical method to determine significance of group differences.

fairness metric. [14] embeds LMs in an algorithmic search procedure to help solve problems like question answering step-by-step, in which relevant paragraphs that might provide answers.

6 Societal Impact and Limitations

We presented *Metric-Fair Prompting*, a joint-inference framework that treats each $(\text{question}, \text{option})$ as a binary instance, applies a margin-based decision rule, and enforces a Lipschitz-style metric constraint to encourage individual fairness.

By encouraging *individual fairness*, Metric-Fair Prompting can reduce spurious variability in multiple–choice medical reasoning and promote decisions grounded in *clinically determinative* evidence rather than sensitive attributes. However, LLM outputs are not guaranteed to be correct or calibrated; the method must *not* be used for autonomous clinical care. Second, fairness is only as good as the similarity metric: embedding-based proximity can reflect corpus biases and omit decisive but rare clinical features, potentially yielding *fairness gerrymandering* across subpopulations.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[3] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020. Earlier version: arXiv:1811.08810 (2018).

[4] Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350, 2024.

[5] Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*, 2022.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

[7] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[8] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[9] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[10] Di Jin, Yifan Pan, Nazneen Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. In *Proceedings of EMNLP*, 2020.

[11] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Better fair algorithms for contextual bandits.

[12] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677, 2023.

[13] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[14] Imanol Schlag, Sainbayar Sukhbaatar, Asli Celikyilmaz, Wen-tau Yih, Jason Weston, Juergen Schmidhuber, and Xian Li. Large language model programs. 2023.

[15] Jie Shen, Nan Cui, and Jing Wang. Metric-fair active learning. In *International conference on machine learning*, pages 19809–19826. PMLR, 2022.

[16] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023.

[17] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*, 2023.

[18] Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[19] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on learning theory*, pages 1920–1953. PMLR, 2017.

[20] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[21] Shunyu Yao, Jeffrey Zhao, Dian Yu, Yuan Cao, Da Yu, Izhak Shafran, Tom Griffiths, and Karthik Narasimhan. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: [TODO]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theory

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [TODO]

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: public dataset

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: will release the code

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: public dataset

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: public dataset

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: only for polishing paper

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.