On Unsupervised Comparisons of Large Language Models

Anonymous ACL submission

Abstract

Evaluation of LLMs has primarily relied on comparing against "gold" answers that often takes months or years to conduct and hence is difficult to scale. Instead of harnessing these supervised approaches that aim to rank LLMs, we propose to assess models by measuring and identifying the significance of their differences. This reduces the difficult supervised learning into an unsupervised task that saves the substantial labeling costs. Specifically, we introduce the notion of topic-categorized distinguisher questions that expose key behavioral differences and hence define distances between LLMs. We design a suite of algorithmic techniques for finding these distinguishers and make three major innovations, including *(i)* a new correlation specification on objective functions based on topic trees and earth-mover distance of topics, *(ii)* a theoretically sound embedding technique between EMD induced by topics and ℓ_2-space used in Bayesian optimization (BO), and *(iii)* a Siamese-net based model leveraging our theoretical results that effectively interface topics and BO in practice. Our experiments showed the efficacy of our new algorithms, its power to distinguish LLMs in medical topics, and its application in unsupervised ranking.

1 Introduction

Evaluating LLMs has primarily relied on grading them against a set of carefully constructed benchmark tasks, such as those done in HuggingFace's Open LLM Leaderboard (HuggingFace, 2024), MMLU (Hendrycks et al., 2020), ARC (Mitchell et al., 2023), HellaSwag (Zellers et al., 2019), and TruthfulQA (Lin et al., 2022). These benchmarks usually take months or years to build and polish, and thus do not scale well. On the other hand, with the recent proliferation of LLMs especially in high-stake domains such as medicine, law and education, it becomes critical to be able to evaluate LLMs under realistic time and resource budgets.

Motivated by the above, we propose in this paper that, instead of harnessing the supervised approach as done in the past, we assess LLMs by measuring and identifying the significance of their differences. This *unsupervised* approach saves the substantial supervised labeling costs. While this approach can only distinguish instead of rank models, it is arguably much more cost-effective: If two LLMs are found to be indistinguishable, then our conclusion will match the supervised approach; if the LLMs are different, then they could be ranked further via benchmarks targeted at their identified differences which are much smaller in size than previously suggested. Moreover, many evaluation problems in fact *do not* boil down to ranking models' qualities. For instance, when a physician chooses between two domain-specific LLMs, showing how LLMs diff for a selected set of questions could be more informative than displaying a grading sheet. Similarly, when a model performs worse than another, identifying questions leading to the two LLMs' largest behavioral differences helps understand the data needed for fine-tuning. Finally, recent works have shown that distance measurement can be used to construct approximate rankings (Rohe et al., 2011; Li et al., 2017), so these techniques can potentially be chained up, leading to an unsupervised approximate ranking algorithm.

Our main contribution in this paper is to create a principled and efficient framework to measure and identify LLMs' differences, thus materializing the benefit in our unsupervised approach over label-intensive alternatives in the past. To this end, a key component in our study is to properly define a notion of model distance that roughly characterizes the discrepancy between answers' contents from different LLMs. Specifically, we introduce what we call *distinguisher questions* that lead to the largest answer difference between LLMs. This metric is based on sound statistical principles, and is inspired by distinguisher sets in cryptography used to determine whether two programs' output are close. The rough idea is that when we identify a subset of inputs that leads to the largest distribution discrepancies between two programs' outputs, this discrepancy characterizes whether two programs are distinguishable (Vadhan, 1999).

Our distinguisher questions are controlled by a subset of topics naturally associated with a new domain under investigation, such as disease categories in medicine or subject areas in academia. Finding the distinguisher questions means searching for a subset of k topics so that a sampled question related to these topics maximizes the discrepancies between LLMs in expectation. We use a Bayesian optimization (BO) framework, and design a new optimization algorithm with three major innovations. First, we leverage a topic-tree that usually exists in a new domain (see Fig. 1(a) for an example) and wire correlations of objective scores using earth-mover distance (EMD) between topic sets. Second, we develop a *polynomial time* algorithm to perform metric embedding between EMD and a Euclidean space for an *exponential number* of possible topic sets. Third, we use our embedding algorithm to design a Siamese net to efficiently interface with any standard BO blackboxes.

Summary of contributions. *(i)* We propose to study unsupervised measurement of distances between LLMs as a cost-effective alternative to evaluate models for new domains. *(ii)* We design a distance metric based on finding distinguisher questions. Along the way, we explain the statistical principles in driving our design. *(iii)* We use BO to find distinguisher questions, and design a Simanese net-based algorithm that injects topic tree-structure to the search space. The new algorithm is powered by a novel metric embedding technique. *(iv)* We validate the efficacy of our algorithms via experiments, and demonstrate its power in distinguishing LLMs and its potential in unsupervised ranking.

2 Related Work

Large language model evaluation. With the emergence of large language models (LLMs), their evaluation has become a significant research area (Zhao et al., 2023; Chang et al., 2024). These evaluations includes natural language understanding (Bang et al., 2023), reasoning (Bian et al., 2024), multilingual (Lai et al., 2023), factually (Gekhman et al., 2023), etc. While evaluation metrics can vary, most existing methods compare generated text/answers against certain costly gold answers (Aynetdinov and Akbik, 2024). Assuming that language quality will improve over time (Huh et al., 2024), in this paper we focus on the relevance and (dis)-similarities aspects of the answers.

Reinforcement learning (RL). By interacting with LLMs, our problem of finding distinguishers can be viewed as an RL problem, and that "environments" are the LLMs' answers. RL (Sutton and Barto, 2018) has long been used in optimizing the performance of neural models. Recent work (Lang et al., 2024; Luong et al., 2024; Casper et al., 2023) has found applications in improving LLMs.

Bayesian optimization and metric embedding. Our algorithmic solution is built on tools from Bayesian optimization (BO) and metric embedding. For BO, we mostly use standard techniques (Frazier, 2018). Hierarchical Bayesian models (Shiffrin et al., 2008; Pelikan and Pelikan, 2005) are particularly relevant to our solution since they also aim to model tree-structured variables, but it remains unclear whether they can scale well in our setting. We also need to build new metric embedding techniques to efficiently interplay between EMD of topic subsets and Euclidean space domain assumed in BO. Closely associated with our work are the impossible results for embedding ℓ_1 to ℓ_2 (Andoni et al., 2011), finite-point embedding from arbitrary metrics to ℓ_2 (Bourgain, 1985), and Johnson-Linderstrauss-based embedding (Venkatasubramanian and Wang, 2011).

3 The k-subset Distinguisher Problem

Notation. Let $M_1, \dots, M_\ell$ be a collection of question-answer LLMs. Let $\mathcal{X}$ and $\mathcal{Y}$ be the sets of all possible questions and answers, respectively. Let $D(y_1, y_2)$ be a distance function between two answers y_1 and y_2. Let Tr be a topic tree, or a hierarchically structured knowledge graph associated with the domain of interest (Fig 1). Let S be the leaves of Tr. We assume that each question in the domain is related to a subset of topics in S.

Finding k-subset distinguishers. Let M_1 and M_2 be two LLMs for comparison. Let $T \subseteq S$ be a subset of topics. We assume that there is a reasonable mechanism available for generating questions related to T via some prompt engineering on LLMs such as ChatGPT or Llama (See App B). For a specific T, define the cost function as

$$f(T) = \mathrm{E}_{X \sim Q(T)}[D(M_1(X), M_2(X))]. \quad (1)$$

$X \sim Q(T)$ refers to sampling a question from

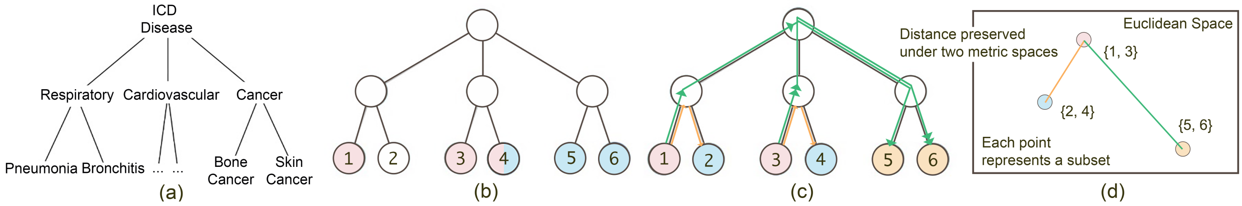

Figure 1: (a) An example of topic tree representing hierarchical knowledge in a given domain. (b) Let $R = \{1, 3, 4\}$ (red leaves) and $B = \{4, 5, 6\}$ (blue ones) be two topic sets. In standard BO, they are represented by binary vectors so $R = (1, 0, 1, 1, 0, 0)$ and $B = (0, 0, 0, 1, 1, 1)$, and their correlation depends on (normalized) ℓ_2-distance $\|R - B\|_2^2 = 4/6$. (c) Euclidean distance sometimes is not suitable. Let $S_1 = \{1, 3\}$ (red), $S_2 = \{2, 4\}$ (blue), and $S_3 = \{5, 6\}$ (yellow). They are mutually exclusive so their pairwise distances are 1. But intuitive, S_1 and S_2 are "closer" because 1-2 and 3-4 are siblings. Using earth-mover distance (EMD) captures this intuition better: it takes four steps to move elements from S_1 to S_2 (red edges), and eight steps to move from S_1 to S_3, hence $EMD(S_1, S_3) = 2EMD(S_1, S_2)$. (d) Embedding EMD to Euclidean space: EMD needs to be mapped to Euclidean distance to answer BO's queries. This embedding ensures S_1 and S_2 are close, and they are far from S_3.

a specific topic set T. Thus, Eq. 1 measures the expected distance between the outputs of M_1 and M_2, given topic set T.

Our **k-subset distinguisher problem** finds

$$\max_{\substack{T \subseteq S \\ |T|=k}} f(T) = \max_{\substack{T \subseteq S \\ |T|=k}} \mathrm{E}_{X \sim Q(T)}[D(M_1(X), M_2(X))] \quad (2)$$

and the top-m k-subset distinguisher problem is

$$\max_{\substack{T_1, \ldots, T_m \subseteq S \\ |T_i|=k \\ T_i \text{distinct}}} \sum_{i \leq m} \mathrm{E}_{X \sim Q(T_i)}[D(M_1(X), M_2(X))], \quad (3)$$

which examines a collection of top-m distinguishers (i.e., $T_1, \ldots, T_m$) to obtain a more holistic view of the distinction between the two LLMs. Below explains the rationale for using Eq. 2 and 3.

3.1 Statistical Ground for Distinguishers

Eq. 2 and 3 already have an intuitive interpretation, i.e, a "good question" exposing a significant difference between two LLMs is naturally a distinguisher of the models. Nevertheless, the underlying statistical principles highlight the definition's connection to established statistical concepts, and equip users to further customize the metric as needed.

A question-answer LLM can be viewed as a specification on the distribution (x, y), where x is a question and y is an answer. Thus, measuring distances between LLMs boils down to measuring discrepancies between probability distributions. Common options include *(i) Total variation/statistical distance (TV).* Let P and Q be two distributions on a discrete support $\mathcal{S}$. Their TV is $\|P - Q\|_{TV} = \|P - Q\|_1 = \sum_{s \in \mathcal{S}} |P(s) - Q(s)|$. *(ii) ℓ_p-norm.* Generalize TV to ℓ_p-distance $\|P - Q\|_p = \left(\sum_{s \in \mathcal{S}} \left|P(s) - Q(s)\right|^p\right)^{1/p}$. *(iii) Kullback–Leibler (KL) divergence,* aka $D_{\mathrm{KL}}(P\|Q) = \sum_s P(s) \log\left(\frac{P(s)}{Q(s)}\right)$.

Total variation (TV) is the most logical choice, because TV and the distinguisher set are connected:

$$\sum_{s \in \mathcal{S}} |P(s) - Q(s)| = \max_{S \subseteq \mathcal{S}} \left| \Pr_P[s \in S] - \Pr_Q[s \in S] \right|.$$

S in the RHS is a distinguisher set for P and Q. As discussed earlier, the distinguisher sets used to measure similarity between programs can also be used to measure the (dis)-similarity between LLMs. ℓ_p-norm can exhibits certain undesirable properties, e.g., two distributions can have TV being 1 but ℓ_p distance being 0 for $p > 1$ (Braverman et al., 2010). Furthermore, KL divergence is asymmetric so we cannot directly use it to define distance.

We need three enhancements to make TV a more suitable metric for LLM comparisons.

E1. Structured by topics (principle: stratified sampling). Merely generating "problematic" questions that trigger LLMs to react differently is too unstructured to be useful or interpretable. We use concepts from stratified sampling to introduce the notion of topics and model each question to be sampled from a fixed set of topics. Let (T, x, y) be topics, question, and answer, respectively, and its graphical model representation be $T \to x \to y$, i.e., x conditionally depends on T, and y conditionally depends on x. TV between the two models represented by P and Q now becomes

$$\sum_{T,x,y} \Big| P(T, x, y) - Q(T, x, y) \Big|$$

$$= \sum_T \left(\sum_{x,y} \Pr[x \mid T] \cdot \Big| P(y|x) - Q(y|x) \Big| \right) \Pr[T]$$

E2. Continuous vs discrete distributions (principle: embedding distance is a sufficient statistics for TV). Assuming x and y to be discrete implies they can be too rigid, because two texts are considered as completely different when there is only a small difference between them, e.g., a word is changed. We use the distance between the texts' embeddings to address this issue and our answers move to the continuous space, where TV becomes $\|P - Q\|_{TV} = \int_x |P(x) - Q(x)|dx$, which we estimate by standard non-parametric methods such as KDE. Under the standard setting, in which only one answer is generated, the TV under KDE depends only on the embedding distance between the two LLMs' answers (Devroye et al., 2018), and we approximate it as

$$\sum_T \left(\sum_{x,y} \Pr[x \mid T] \cdot \left| P(y|x) - Q(y|x) \right| \right) \Pr[T]$$
$$\approx \sum_T \left(\mathrm{E}_{X|T} \left\| y_P(X) - y_Q(X) \right\|_{embed} \right) \Pr[T] \quad (4)$$

E3. Skewness property (principle: Laplacian structure). We express Eq. 4 as $\sum_T U(T) \Pr[T]$, where $U(T) = \mathrm{E}_{X|T} \left\| y_P(X) - y_Q(X) \right\|_{embed}$. $U(T)$ is often skewed (such as following Laplacian or power-law distributions) across T's: two reasonable LLMs are expected to produce similar answers to simple questions, but their answers' discrepancies are sometimes large for harder questions (Huh et al., 2024). Therefore, the heaviest leading terms should dominate the total mass of Eq. 4. By directly measuring those leading terms, we obtain an objective recovering Eq. 3.

$$\max_{\substack{T_1,\ldots,T_m \subseteq S \\ |T_i|=k \\ T_i \text{ distinct}}} \sum_{i \le m} U(T_i) = \max_{\substack{T_1,\ldots,T_m \subseteq S \\ |T_i|=k \\ T_i \text{ distinct}}} \sum_{i \le m} f(T_i),$$

Remark. The derivations above provide more principled guides to generalize our metrics: when we have more compute and generate multiple answers to the same question, we use KDE to produce TV estimates, and the embedding distance may no longer be a sufficient statistic. When we already know the better T's, we need to conduct a careful search for the leading terms of $U(T)$.

4 Our algorithm

We now solve Eq. 3. See Fig. 2 for our pipeline. A BO iteratively makes queries on an objective function with domain $\mathbf{R}^d$ (referred to as embedding space). A decoder module (`EmbDecr`/`BayesDcdr`) maps a query to a topic tuple, which, along with a question template, is fed to a reliable LLM to generate a question. Two LLMs produce answers to the question, and their distance is measured by a module such as `Sentence-transformer` and fed to the BO. BO then prepares for the next query.

The crux here is to properly specify f's correlations to optimize BO's efficiency. We first explain a standard way to use BO as a warmup (corresponding to using `BayesDcdr`), then introduce EMD to better capture the correlation structure of f and describe our optimization algorithm built on a new EMD-to-ℓ_2 metric embedding technique (corresponding to using `EmbDcdr`).

4.1 Warmup: Using Standard BO

Recall that we aim to find $\arg\max_{T \subseteq S} f(T)$ with the constraint $|T| = k$. GP-based BO assumes that the objective is from a Gaussian process, and the domain of f is $\mathbf{R}^d$, so we need some standard trick; specifically, let $\mathbf{x} \in \mathbf{R}^n$, where $|S| = n$, and $\tilde{T}$ be a random subset of size k sampled without replacement from the distribution $\sigma(\mathbf{x})$, where $\sigma(x) = 1/(1 + \exp(-x))$ is applied to $\mathbf{x}$ in an element-wise manner. Define

$$\tilde{f}(\mathbf{x}) = \mathrm{E}_{\tilde{T} \sim \sigma(\mathbf{x})}[f(\tilde{T})]. \quad (5)$$

Since $\tilde{f}$'s domain is now $\mathbf{R}^n$, we may apply standard BO to optimize $\tilde{f}$. A Monte-Carlo method is used to estimate of Eq. 5's RHS. While $\tilde{f}$ does not directly give a set T, we can always find a T such that $f(T) \ge \tilde{f}(\mathbf{x})$ by an averaging argument.

4.2 Earth-Mover Distance for Correlations

Standard BO assumes the correlation of $f(T_1)$ and $f(T_2)$ is proportional to their Hamming distance, and T_1 and T_2 are considered far apart when they do not share any elements. It becomes less suitable when the topic tree is available. See Fig. 1(c) for an example, where S_1, S_2, and S_3 are mutually exclusive, such that any pair of them is considered far apart under Hamming. Intuitively, though, S_1 and S_2 should be "fairly close" because 1-2 and 3-4 are siblings. The notion of earth-mover distance better captures this intuition.

Definition 1. *Let $G = (V, E)$ be an undirected graph, and V_1 and V_2 be two subsets of V having the same cardinality. Let $\mathcal{F}$ be the set of all bijective functions between V_1 and V_2, and $D_G(u, v)$ be the shortest path length between u and v in G. The earth-mover distance (EMD) between V_1 and V_2 is*

$$EMD(V_1, V_2) = \min_{f \in \mathcal{F}} \sum_{u \in V_1} D_G(u, f(u)). \quad (6)$$

In other words, EMD counts the minimum number of steps needed to move elements from one set to another. One can see that in Fig. 1(c), the EMD of S_1 and S_2 is half of that of S_1 and S_3 and thus provides a more realistic distance estimate for BO.

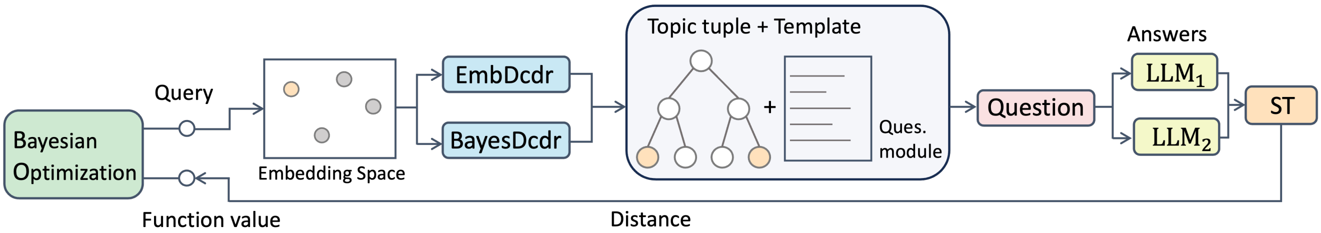

Figure 2: Our BO-based pipeline for finding distinguisher questions (Sec. 4).

4.3 Algorithms Using EMD

We need to code EMD into $f(\cdot)$, i.e., $f(T_1)$ and $f(T_2)$ are more correlated (closer to each other) when $EMD(T_1, T_2)$ is small. Standard BO usually uses Euclidean space as the domain especially when we use GP to model the objective. Thus, we need a low-distortion embedding function $\psi(T) \in \mathbf{R}^d$, so that we can use BO to optimize

$$g(\mathbf{x}) \equiv f(\psi^{-1}(\mathbf{x})) \tag{7}$$

Eq. 7 suggests that a decoder ψ^{-1} is needed. Our embedding technique has two steps, including first embedding EMD to ℓ_1-space, and from ℓ_1 to ℓ_2.

Embedding from EMD to ℓ_1. As a key building block, we design a sum encoding scheme to represent a topic subset, and ensure that the two sets' ℓ_1 distance is the same as their EMD.

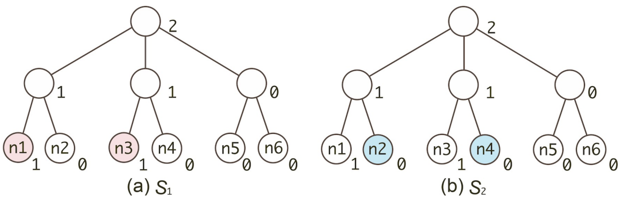

Figure 3: Sum encoding. Let two topics sets, S_1 and S_2, be labeled with two different colors on two trees. To build a sum-encoder for a set, we walk through each node v on the tree, and count the number of elements in the set that are "under" v. To encode S_1, count the number of elements in S_1 that are under the root (which is 2), and repeat the process for each node. The number in each node represents the count needed. Flat out all numbers, i.e., $\Psi(S_1) = (2, 1, 1, 0, 1, 0, 1, 0, 0, 0)$, and similarly, $\Psi(S_2) = (2, 1, 1, 0, 0, 1, 0, 1, 0, 1)$. We have $\|\Psi(S_1) - \Psi(S_2)\|_1 = EMD(S_1, S_2) = 4$.

Definition 2 (Sum-encoding)**.** *Let T be a subset of leaves on* Tr. *In the sum-encoding representation Ψ, each node v associates with a number that counts the number of elements in T that is in the subtree rooted at v:*

$$\Psi_{Tr}(T) = \Big\{(v, x) \mid v \in V(Tr),$$
$$x = |\{w \in T : v \text{ an ancestor of } w\}|\Big\}$$

This definition considers v as an ancestor of itself. See Fig. 3 for an example. We can also use a vector representation $\Psi_{Tr}(S) = (x_1, x_2, \ldots, x_{\tilde{n}})$, where x_i is the number of nodes in S that are on the i-rooted subtree and $\tilde{n} = 2n - 1$ is the total number of nodes in Tr. When the context is clear, we write $\Psi(S)$ instead of $\Psi_{Tr}(S)$. Prop. 1 below shows that while computing EMD seemingly requires using heavy combinatorial algorithms, the computing can be done by simple vector-computation.

Proposition 1. *Let T_1 and T_2 be two subsets of leaves in Tr of the same size, and $\Psi(T_1)$ and $\Psi(T_2)$ be the sum-encodings of T_1 and T_2 in vector representation form. We have $EMD(T_1, T_2) = \big\|\Psi(T_1) - \Psi(T_2)\big\|_1$* (See App. A for the proof).

Embedding from ℓ_1 to ℓ_2: theory and practice. One major technical barrier here is the impossibility results for ℓ_1-to-ℓ_2 embedding (Andoni et al., 2011), so we use two observations: *(i)* the number of possible topic sets T is finite. This is less challenging than embedding the entire ℓ_1 to ℓ_2. Finite-point embedding algorithms (Bourgain, 1985; Indyk et al., 2017) exist but in our setting they run in exponential time, because the number of possible T's is $O(n^k)$ and is exponential in k. *(ii)* Observe that the points embedded in ℓ_1 possess special structure, i.e., they are integer-valued and sparse. These properties enable us to further surpass SOTA exponential-time algorithms.

Below, we first explain ℓ_1-to-ℓ_2 embedding is theoretically feasible, then we propose a practical architecture built on the theoretical results.

Theory: Unitary Encoding. The values in each coordinate can only be $\{0, 1, 2, \ldots k\}$, so we flatten a number in $[k]$ as a k-dimensional unitary vector, e.g., we encode 1 as $\underbrace{(1, 0, \ldots)}_{k \text{ elements}}$, encode 2 as $(1, 1, 0, \ldots)$. The squared ℓ_2-distance between two unitary-coded vectors is now the same as the ℓ_1 distance between the original vectors. Unitary encoding inflates the dimension from $2n - 2$ in ℓ_1 to $k(2n - 2)$, but applying a Johnson-

Question (case)	`Llama-2-7b(4bit)`	`Llama-2-7b`
Question A. A 7-week-old male infant is brought to the pediatrician by his concerned parents. They noticed that he had frequent episodes of nosebleeds since birth, which had become more persistent and difficult to control in the past week. The parents also mention that the infant has been unusually fussy and irritable. On examination, the infant appears pale, with ... [text abridged] Question for Medical Students: Based on the presented case, what congenital malformations or hematological disorders should be considered in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability? Authors' annotation: a non-distinguisher question; answers are similar.	Based on the symptoms presented, there are several congenital malformations or hematological disorders that should be considered in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability. Here are some of the possibilities: 1. Hemophilia A: This is a genetic disorder that affects the blood's ability to clot, leading to prolonged bleeding after injury or surgery. The infant may have inherited this condition from one of their parents, and the frequent nosebleeds and petechiae could be a sign of this disorder. 2. Von Willebrand Disease: This is a mild bleeding disorder caused by a deficiency of von Willebrand factor, which is essential for blood clotting. The infant may ... [text abridged]	Based on the presented case, I would consider the following congenital malformations or hematological disorders in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability: 1. Hemophilia A: This is a genetic disorder that affects the blood's ability to clot, leading to frequent nosebleeds and bruising. The elevated heart rate and signs of bleeding from the gums could be indicative of this condition. 2. Von Willebrand disease: Similar to hemophilia A, this is a bleeding disorder caused by a deficiency in von Willebrand factor, which is essential for blood clotting. The symptoms of nosebleeds, petechiae, and irritability could be indicative of this condition... [text abridged]
Question B. A 45-year-old man presents to the clinic with complaints of recurrent severe headaches, blurry vision, and difficulty urinating for the past 6 months. He also reports experiencing weakness and numbness in his lower extremities. He has a history of inflammatory bowel disease and was recently diagnosed with noninfective enteritis. He is currently being treated with mesalamine and infliximab ... [text abridged] As a clinical expert, how would you respond to this medical case? **Topics:** Cranial nerve disorders in diseases classified elsewhere, Disorders of eyelid in diseases classified elsewhere, Disorders of male genital organs in diseases classified elsewhere, Noninfective enteritis and colitis, Polyneuropathies and other disorders of the peripheral nervous system	The patient in the scenario presented with signs and symptoms consistent with a right temporal lobe glioblastoma. Glioblastomas are the most common primary brain tumors in adults. They are classified as grade IV gliomas and are associated with a poor prognosis. They are typically treated with surgery, radiotherapy, and chemotherapy. Temozolomide is a commonly used chemotherapy agent for glioblastoma. Bevacizumab is a monoclonal antibody that targets vascular endothelial growth factor (VEGF). It is used in the treatment of several types of cancer, including glioblastoma, but it is not a first-line agent. Cisplatin is a platinum-based chemotherapy agent that is used in the treatment of several types of cancer. Authors' annotation: Generic and misdiagnosed, neglecting the patient's history and symptoms.	Based on the presenting symptoms, medical history, and family history provided, this patient's constellation of symptoms is concerning for a neurological condition involving multiple systems. The presence of severe headaches, blurry vision, lower extremity weakness, numbness, urinary symptoms, ptosis of the left eyelid, and decreased facial expression on the ... [text abridged] Management Recommendations: 1. Neurological Evaluation: Urgent referral to a neurologist for a comprehensive evaluation, including a detailed neurological examination and possible lumbar puncture. 2. Imaging: MRI of the brain and spinal cord to identify any demyelinating lesions or other abnormalities. 3.Laboratory Tests: ... Authors' annotation: More case-specific.

Table 1: Abridged examples of questions and answers (see Table 4 for the full example): two questions are generated based on the keywords, and the questions are supplied to `Llama-2-7b(4bit)` and `Llama-2-7b`. Two models provide near identical answers to Question A. Question B is a distinguisher question identified by our algorithm. `Llama-2-7b(4bit)` provides a more generic answer, whereas `Llama-2-7b` is more specific.

Linderstrauss afterward reduces the dimension back to $(k \log n)/\epsilon^2$.

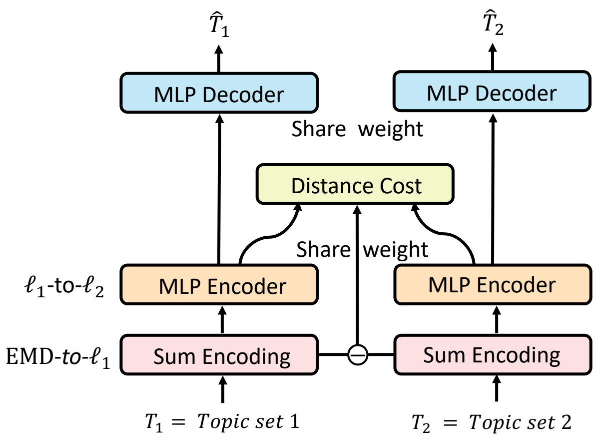

Figure 4: Architecture for learning EMD-to-ℓ_2 embedding. Two topic sets are sum-encoded, and then they go through MLP encoders sharing weights. The distance cost function is used to approximate the ℓ_2 distance of embedded points and the ℓ_1 distance of the sum-encoded vectors. The MLP encoder's output is our EMD-to-ℓ_2 embedding used for BO. A decoder needs to be trained to map embedded points back to topics.

Practice: Training via a Siamese network. We develop a specialized neural net to leverage our theoretical results above for embedding. Neural nets simplify performance fine-tuning, and help to train the decoders needed in Eq. 7. We base our architecture on Siamese nets. See Fig. 4. ENC maps from topic subsets to an embedding on ℓ_2 space, and DEC maps from a point in ℓ_2 space to a topic subset. Both ENC and DEC are MLPs.

In each round, we choose two random topic sets, T_1 and T_2, compute $\Psi(T_1)$ and $\Psi(T_2)$, and feed them to the same ENC. Let $\mathrm{ENC}(\Psi(T_1))$ and $\mathrm{ENC}(\Psi(T_2))$ be the outputs, and $d(T_1, T_2) = \|\Psi(T_1) - \Psi(T_2)\|_1$. The cost (distance) function is difference between the ℓ_2^2 distance in embedding and $EMD(T_1, T_2)$ $\left((\|\mathrm{ENC}(\Psi(T_1)) - \mathrm{ENC}(\Psi(T_2)\|_2^2) - EMD(T_1, T_2)\right)^2$.

To train the decoder, we want to ensure $\mathrm{DEC}(\mathrm{ENC}(T)) = T$ by using either cross-entropy or simple MSE. It is important to simultaneously train $\mathrm{ENC}(\cdot)$ and $\mathrm{DEC}(\cdot)$, and not sequentially train them.

5 Experiments

Finding distinguishers. We design an experiment to validate our algorithms on a medical domain. We extract the topic tree (knowledge graph) from the ICD-10 codes from World Health Organization (Organization, 1993). For a specific set of topics of fixed cardinality, we generate questions via prompting a reliable LLM, e.g., `GPT-3.5`, about a hypothetical patient's symptoms. See App. B for details/examples. The generated case scenarios were reviewed by physician to evaluate plausibility.

Topics		2				3				5			
left	**right**	random	bayes	embed	boost	random	bayes	embed	boost	random	bayes	embed	boost
Flan-T5-xxl	GPT2-xl	1.276	1.292	1.283	**1.444**	1.313	1.287	1.266	**1.429**	1.268	1.320	1.328	**1.494**
	GPT-neo	1.413	1.399	1.376	**1.547**	1.311	1.319	1.328	**1.557**	1.283	1.252	1.354	**1.456**
	GPT-j	1.360	1.430	1.314	**1.635**	1.336	1.436	1.398	**1.583**	1.452	1.550	1.332	**1.674**
GPT2-xl	GPT-neo	1.446	1.285	1.262	**1.494**	1.210	1.323	1.357	**1.506**	1.306	1.348	1.328	**1.529**
	GPT-j	1.341	1.296	1.405	**1.544**	1.311	1.254	1.366	**1.492**	1.279	1.308	1.327	**1.478**
GPT-neo	GPT-j	1.382	1.369	1.373	**1.537**	1.294	1.350	1.376	**1.575**	1.134	1.370	1.437	**1.626**
Flan-T5-large	Flan-T5-xxl	1.155	1.202	1.212	**1.373**	1.254	1.285	1.228	**1.403**	1.220	1.180	1.073	**1.254**
	Llama-2-13b$_{4bit}$	1.408	1.378	1.394	**1.612**	1.298	1.486	1.494	**1.758**	1.366	1.509	1.469	**1.667**
Flan-T5-xxl	Llama-2-7b$_{4bit}$	1.314	1.328	1.308	**1.489**	1.482	1.436	1.491	**1.680**	1.324	1.470	1.436	**1.629**
Llama-2-13b$_{4bit}$	Llama-2-7b$_{4bit}$	1.496	1.399	1.744	**1.914**	1.626	1.709	2.100	**2.361**	1.448	1.495	1.531	**1.880**
GPT-3.5-Turbo	Llama-3-8B	**1.513**	1.316	1.248	1.445	1.366	1.559	1.562	**1.801**	1.270	1.395	1.333	**1.575**
Llama-3-8B	Llama-2-13b$_{4bit}$	1.371	1.573	1.278	**1.727**	1.375	1.486	1.513	**1.740**	1.677	1.518	1.351	**1.698**
	Llama-2-7b	1.398	1.457	1.317	**1.642**	1.307	1.557	1.422	**1.766**	1.411	1.479	1.437	**1.743**
	Llama-2-7b$_{4bit}$	**1.517**	1.288	1.159	1.475	1.426	1.525	1.234	**1.702**	1.466	1.453	1.472	**1.731**
Llama-2-13b$_{4bit}$	Llama-2-7b	1.667	1.791	1.256	**1.956**	1.539	1.740	1.271	**1.855**	1.521	1.602	1.611	**1.944**
Llama-2-7b	Llama-2-7b$_{4bit}$	1.552	1.240	1.237	**1.597**	1.579	1.790	1.531	**1.991**	1.411	1.396	1.480	**1.771**

Table 2: Evaluating different algorithms for optimizing the objective Eq. 3 for topic numbers being 2, 3, and 5. Recall that Eq. 3 measures the mean distance of the top ten distinguisher topic tuples, i.e., the larger objective, the more effective the algorithm. Bold text: the best performing algorithm for a specific configuration. Grayed text: the random entry when neither Bayes nor embed base learner can out-compete random. Models are from Hugging Face.

We examine `Flan-t5-large`, `Flan-t5-xxl`, `GPT-j`, `GPT2-xl`, `GPT-neo`, `GPT-3.5`, `Llama-2-7b`, `Llama-2-7b-4bit`, `Llama-2-13b`, `Llama-3-8b`, and `Llama-2-13b-4bit`, `4bit` versions use `load_in_4bit`; other models are original ones. We randomly selected 16 pairs from them. `sentence-transformers` are used to measure answers' distance (Reimers and Gurevych, 2019).

We evaluate `random`, standard BO `bayes` (Sec. 4.1), embedding-based BO `embed` (Sec. 4.3), and boosting algorithm `boost`; the latter uses the union of all queries from `bayes` and `embed` to select top m items. We note `boost` can execute `bayes` and `embed` in parallel, without requiring extra time.

The domain of `bayes` and `embed` is the distributions on topics. For each distribution (BO query), we sample three topic tuples and two questions for each topic tuple. The objective's value is the mean of six distances. We perform a total of $n = 100$ queries, and find the mean distance of $m = 10$ largest topic tuples. Since the number of queries for the random baseline needs to match bayes and embed, we query $n = 300$ topic tuples. We standerdize the distances w.r.t. to `random` and repeat each experiment for $k = 2, 3$, and 5 topics.

See Table 2 for the results. Algorithms in boldface perform best for each configuration. Entries are highlighted in gray when `random` outperform both `bayes` and `embed`. *(i)* `boost` is the best and usually significantly outperforms the others. *(ii)* `bayes` and `embed` outperform `random` when $k \in \{3, 5\}$, whereas `random` can outperform `bayes` and `embed` when $k = 2$. The results are expected since questions with only two topics usually have simpler structure, so the value of using BO diminishes, *(iii)* `embed` does not always outperform `bayes`, but since `boost` usually significantly outperforms them individually, `embed` and `bayes` offer orthogonal signals and confirms the value of `embed`.

Case studies. Table 1 lists the distinguisher and non-distinguisher questions identified by our algorithms for `Llama-2-7b` and `Llama-2-7b-4bit`. Question A shows a non-distinguisher question; both models give nearly identical answers. Question B shows a distinguisher; `Llama-2-7b-4bit` gives a generic answer, whereas `Llama-2-7b` is more specific. Physician review of the generated questions provided potential reasons for distinguisher performance, such as question specificity and range of potential diagnoses and clinical actions based on each scenario. The questions were generally comprehensible and aligned with the format for medical training. While questions generated here could not be used directly in real-world settings, they are useful for screening LLMs in the medical domain. Suggestions were made to improve question plausibility for future development, such as removing non-specific diagnosis codes and grouping topics according to age or epidemiologically co-occurring conditions. See App. B.3.

Siamese net performance. We use a Siamese net to learn EMD-to-ℓ_2 embedding. We need to confirm that the encoder preserves the distance and that the decoder properly maps a point in ℓ_2 back to a topic subset. We examine training $\mathrm{ENC}(\cdot)$

Topics			2					3					5				
algo	$\sigma(\cdot)$	n_ℓ	train emb	test emb	train decd	test decd	score	train emb	test emb	train decd	test decd	score	train emb	test emb	train decd	test decd	score
all	False	1	92.9%	91.7%	73.3%	73.1%	82.4%	91.6%	90.4%	71.2%	71.0%	80.7%	90.2%	88.4%	64.2%	64.0%	76.2%
		2	93.5%	90.5%	96.2%	96.1%	93.3%	91.3%	87.9%	87.4%	86.8%	87.4%	90.5%	85.3%	75.6%	74.8%	80.0%
	True	**1**	93.1%	91.8%	100.0%	100.0%	**95.9%**	91.4%	90.2%	99.8%	99.2%	**94.7%**	89.9%	88.2%	96.0%	95.2%	**91.7%**
		2	93.2%	90.6%	96.1%	95.8%	93.2%	91.7%	88.4%	97.9%	96.0%	92.2%	90.4%	85.6%	85.9%	83.5%	84.5%
seq	False	1	92.6%	91.6%	72.0%	71.9%	81.7%	91.6%	90.3%	68.6%	68.5%	79.4%	89.9%	88.3%	60.4%	60.3%	74.3%
		2	92.8%	89.8%	93.5%	92.9%	91.4%	91.2%	87.1%	65.2%	63.7%	75.4%	89.9%	84.4%	49.9%	48.8%	66.6%
	True	1	92.8%	91.7%	72.1%	72.0%	81.8%	91.5%	90.2%	69.2%	69.1%	79.7%	90.0%	88.3%	59.7%	59.7%	74.0%
		2	93.0%	90.0%	93.2%	92.6%	91.3%	91.3%	87.1%	65.2%	63.7%	75.4%	89.7%	84.3%	49.3%	48.1%	66.2%

Table 3: Using a Siamese network to learn embedding for EMD for topic subsets. The evaluation metric is correlation (1 being the best and -1 being the worst). Score is the mean of embedding correlation and decoder correlation in the test set. Training algorithms examined: *all:* simultaneously training both $\mathrm{ENC}(\cdot)$ and $\mathrm{DEC}(\cdot)$; *seq:* sequentially first train $\mathrm{ENC}(\cdot)$, then train $\mathrm{DEC}(\cdot)$. Architecture searched: whether to include sigmoid function at the final layer ($\sigma(\cdot)$ being true or false), and the number n_ℓ of hidden layers in the MLP.

and $\mathrm{DEC}(\cdot)$ simultaneously and sequentially. We perform a lightweight architecture search along two dimensions: whether to include a sigmoid function at the top of the MLPs, and whether to use one or two hidden layers.

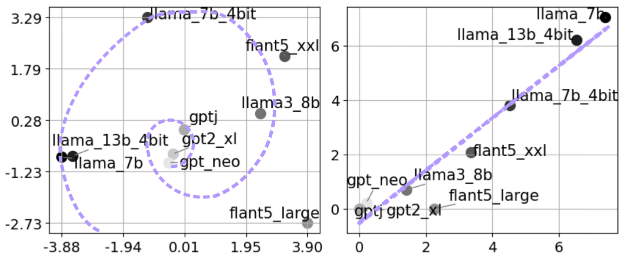

Figure 5: Using pairwise distance to approximate ranking. Left: embedding the models into a 2-dim space based only on their estimated distances. Right: further adjusting the 2-dim embedding by using polar coordinates and certain ad-hoc transformations. **Ranking information surfaces**: models quality in general improve from lower left to upper right.

Table 3 lists the results. We use the correlation between prediction and target as our evaluation metric and set our score to be the mean of embedding and decoding correlations. All configurations give reasonable ENC, but it is important to train ENC and DEC simultaneously to properly train DEC. Having the sigmoid function and using one layer consistently performs the best.

Toward unsupervised ranking. There exist algorithms (primarily using metric-embedding techniques) that translate from distance to ranking (Rohe et al., 2011; Li et al., 2017), but the presence of weaker and smaller models deteriorates the output quality. Weaker and smaller models give low-quality medical related answers, so they have small positive correlations with all other models, and their cosine distances to all other models concentrate around 1, which makes 1-dimensional embedding (approximate ranking) impossible.

However, we still find interesting ranking-related structures when we embed the distances measured by our algorithms into a 2-dim space. Specifically, we first run a matrix completion algorithm to fill in the missing distances, and then use PCA to map models to 2-dim space. See Fig. 5(a). `GPT-3.5` is excluded because it is not an open-source model. We observe that *(i)* smaller models `GPT-j`, `GPT2-xl`, and `GPT-neo` have equal distances to all other models as we predicted, and *(ii)* a ranking starts to surface when we draw a spiral from the bounding box to the "origin." When we use a polar coordinate to express the points and perform certain transformations, the ranking becomes more explicit. See Fig. 5(b). Lower left are worse models, and upper right corresponds to better models. In the extracted rankings (projection against the dashed regression line), `Llama-3-8b` appears to be out of order, but otherwise, the ranking appears consistent with general perception. We leave a more systematic investigation to future work.

6 Conclusion

This paper studied unsupervised methods to measure distances between LLMs as a low-cost, faster alternative to model grading/ranking. We proposed using distinguisher questions categorized by topic sets to measure distance and reduced our problem to a combinatorial optimization, which we solved with Bayesian optimization. We introduced a novel correlation structure for topics based on earth-mover distance, and designed a theoretically sound Siamese net for EMD-to-ℓ_2 embedding that interfaces topics and BO. The experiments confirmed the efficacy of our algorithms and demonstrated their potential in approximate rankings.

Limitations

Computation. Our algorithms require moderate GPU resource for comparing two models. For models with 7 to 13B parameters, evaluation takes 5-8 hours using two A100 GPUs. This does not scale very well when massive pairs need to be evaluated. Techniques such as matrix completion could be needed to complement our algorithm and fill in "missing pairs". *Noise.* The efficacy of Bayesian optimization relies on the correctness of the correlation specification. When the topic number is smaller and the quesiton structure is simple, the value of using BO diminishes as shown in our experiments. *Blackboxes and hallucinations.* Our algorithm assumes the availability of sufficiently powerful LLMs to generate questions based on topics. While domain experts are likely to conduct ad-hoc investigations and run additional "sanity checks" on question quality before deciding to use `GPT-3.5-Turbo`, we rely on a blackbox and do not have full control over text generation quality.

Ethics Statement

Clinical cases are auto-generated to compare LLMs for measuring their distances in the experiments. We reiterate that the experiments' purpose is to highlight models' differences under less common scenarios. We envision these results are used to build assistive tools for medical processionals to assess LLMs, or for further fine-tuining downstream models. Neither our algorithms nor our experiments provide definite assessment on LLMs used in medical domains.

References

Alexandr Andoni, Moses S Charikar, Ofer Neiman, and Huy L Nguyen. 2011. Near linear lower bound for dimension reduction in l1. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 315–323. IEEE.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *arXiv preprint arXiv:2401.17072*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. pages 3098–3110.

Jean Bourgain. 1985. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52.

Vladimir Braverman, Kai-Min Chung, Zhenming Liu, Michael Mitzenmacher, and Rafail Ostrovsky. 2010. Ams without 4-wise independence on product domains. In *27th International Symposium on Theoretical Aspects of Computer Science-STACS 2010*, pages 119–130.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. 2018. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Peter I Frazier. 2018. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hugging Face. Hugging face: The ai community building the future.

HuggingFace. 2024. Open llm leaderboard. Hugging Face, Inc.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. In *International Conference on Machine Learning*. PMLR.

Piotr Indyk, Jiří Matoušek, and Anastasios Sidiropoulos. 2017. 8: low-distortion embeddings of finite metric spaces. In *Handbook of discrete and computational geometry*, pages 211–231. Chapman and Hall/CRC.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hiu Mn, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

Hao Lang, Fei Huang, and Yongbin Li. 2024. Fine-tuning language models with reward learning on policy. *arXiv preprint arXiv:2403.19279*.

Cheng Li, Felix M Wong, Zhenming Liu, and Varun Kanade. 2017. From which world is your graph. *Advances in Neural Information Processing Systems*, 30.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

Melanie Mitchell, Alessandro B Palmarini, and Arsenii Kirillovich Moskvichev. 2023. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. In *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*.

World Health Organization. 1993. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization.

Martin Pelikan and Martin Pelikan. 2005. *Hierarchical Bayesian optimization algorithm*. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Karl Rohe, Sourav Chatterjee, and Bin Yu. 2011. Spectral clustering and the high-dimensional stochastic blockmodel.

Richard M Shiffrin, Michael D Lee, Woojae Kim, and Eric-Jan Wagenmakers. 2008. A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, 32(8):1248–1284.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Salil Pravin Vadhan. 1999. *A study of statistical zero-knowledge proofs*. Ph.D. thesis, Massachusetts Institute of Technology.

Suresh Venkatasubramanian and Qiushi Wang. 2011. The johnson-lindenstrauss transform: an empirical study. In *2011 Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 164–173. SIAM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Missing Analysis

Proof of Proposition 1. Recall that earth-moving distance can be interpreted as finding a way to move "balls" in T_1 to locations specified by T_2 so the sum of all balls' travel distance is minimized.

We first show that EMD is lower bounded by $\left\|\Psi(T_1) - \Psi(T_2)\right\|_1$. Then we will give an algorithm that results in the total traveling distance being exactly $\left\|\Psi(T_1) - \Psi(T_2)\right\|_1$.

Lower bound. Let us first name the levels of the tree: let L_i consists of nodes whose distance to root is i so L_0 consists of only the root, L_1 consists of the root's children, and L_h consists of all leaves, where h is the height of the tree.

A ball is said to travel across level i if in a moving strategy, we need to move the ball from level i to $i-1$, and then from level $i-1$ to i (these two operations need not be consecutive) at some point. For example, when we move the ball at the leftmost leaf to the root then move that down to the rightmost leaf, the ball is traveled across levels $0, 1, 2, \dots h-1$.

We can find a lower bound C_i on the number of balls that have to travel across level i for any moving strategy. Then an EMD lower bound would be $2\sum_{1\le i\le h} C_i$.

Let v be a node in L_i and x_v (y_v) be the number of balls in S_1 (S_2) that is in v-rooted tree. When $x_v > y_v$, that means $x_v - y_v$ balls from S_1 cannot find a match in S_2 in v-rooted subtree so they have to travel beyond node v (and across level i). When $x_v < y_v$ that means we need other balls outside v-rooted subtree to fill in S_2's spots in v-rooted subtree. We can use this argument to enumerate through all nodes in L_i and each ball will be double counted so we have

$$C_i \ge \frac{1}{2}\sum_{v\in L_i} |x_v - y_v|.$$

So we indeed have a lower bound

$$2\sum_{1\le i\le h} C_i \ge |\Psi(S_1) - \Psi(S_2)|_1. \qquad (8)$$

Upper bound. Here, we image elements in S_1 as red balls and elements in S_2 as blue balls. Our goal is to move red and blue balls so that all of them are eventually paired (i.e., a red and a blue ball meet). This is equivalent to moving red balls to blue locations. We will construct a way to move the balls so that the total number of moves is exactly $|\Psi(S_1) - \Psi(S_2)|_1$.

All the balls are at the leaves of Tr at the initial state. Our strategy is a fairly straightforward one: if red and blue balls can be matched at a node, then the unmatched nodes will be moved up. All balls will eventually be paired because all unmatched balls will eventually be pushed to the root and be matched there "in the worst case" when no matching is available earlier. We can also count the total number of moves by the total number of balls that are moved from layer i to layer $i-1$, which is exactly

$$\sum_{v\in L_i} |x_v - y_v|.$$

Thus, the total number of moves is indeed $\sum_{1\le i\le h}\sum_{v\in L_i} |x_v - y_v| = |\Psi(T_1) - \Psi(T_2)|_1$. □

B Generation of questions

B.1 Generation procedure

For each topic tuple, we prompt an LLM (`ChatGPT3.5-turbo`) using the following structure:

`Prompt starts.`

You are a medical school professor and you are designing questions for medical school students. Please refer to the medical scenario described in the template below, and create a patient case scenario involving these medical topics: `$TOPIC LIST`. The case should interweave these themes into a coherent and medically plausible scenario, and generate a question concerning this patient. Build your own case. Here is a template scenario:

`A medical case template is copied here.`

Note: For the cases in the template, only learn the format here. Please do not copy the scenarios.

`Prompt ends.`

`$TOPIC LIST` consists of elements from the topic tree (Sec. 4.2), and we sample a medical case template from MedQA (Jin et al., 2021). Our prompting strategy asks the LLM to be more specific and self-consistent and share certain similarities with the "chain-of-thoughts" techniques (Wei et al., 2022).

B.2 Question examples

Below lists five questions generated based on our prompting strategy.

Question (case)	`Llama-2-7b(4bit)`	`Llama-2-7b`
Question A. A 7-week-old male infant is brought to the pediatrician by his concerned parents. They noticed that he had frequent episodes of nosebleeds since birth, which had become more persistent and difficult to control in the past week. The parents also mention that the infant has been unusually fussy and irritable. On examination, the infant appears pale, with petechiae scattered over his body. The pediatrician notes that the infant's heart rate is elevated, and there are signs of bleeding from the gums. Question for Medical Students: Based on the presented case, what congenital malformations or hematological disorders should be considered in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability? Authors' annotation: a non-distinguisher question; answers are similar.	Based on the symptoms presented, there are several congenital malformations or hematological disorders that should be considered in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability. Here are some of the possibilities: 1. Hemophilia A: This is a genetic disorder that affects the blood's ability to clot, leading to prolonged bleeding after injury or surgery. The infant may have inherited this condition from one of their parents, and the frequent nosebleeds and petechiae could be a sign of this disorder. 2. Von Willebrand Disease: This is a mild bleeding disorder caused by a deficiency of von Willebrand factor, which is essential for blood clotting. The infant may have this condition, which could be causing the frequent nosebleeds and petechiae. 3. Platelet disorders: Disorders that affect the platelets' ability to clot blood can lead to frequent nosebleeds and petechiae. These disorders could include platelet aggregation defects, platelet function	Based on the presented case, I would consider the following congenital malformations or hematological disorders in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability: 1. Hemophilia A: This is a genetic disorder that affects the blood's ability to clot, leading to frequent nosebleeds and bruising. The elevated heart rate and signs of bleeding from the gums could be indicative of this condition. 2. Von Willebrand disease: Similar to hemophilia A, this is a bleeding disorder caused by a deficiency in von Willebrand factor, which is essential for blood clotting. The symptoms of nosebleeds, petechiae, and irritability could be indicative of this condition. 3. Platelet disorders: Disorders such as platelet dysfunction or platelet storage disease can cause an increased risk of bleeding. The frequent nosebleeds and petechiae in this case could be indicative of a platelet disorder. 4. Liver disease: Certain liver disease
Question B. A 45-year-old man presents to the clinic with complaints of recurrent severe headaches, blurry vision, and difficulty urinating for the past 6 months. He also reports experiencing weakness and numbness in his lower extremities. He has a history of inflammatory bowel disease and was recently diagnosed with noninfective enteritis. He is currently being treated with mesalamine and infliximab infusions for his IBD. The patient also mentions that he has noticed drooping of his left eyelid and has been experiencing erectile dysfunction. His father had a history of polyneuropathy. On examination, he is noted to have ptosis of the left eyelid, decreased facial expression on the left side, and diminished tactile sensation in his lower extremities. Cranial nerve examination reveals weakness of the left cranial nerves. As a clinical expert, how would you respond to this medical case? **Topics:** Cranial nerve disorders in diseases classified elsewhere, Disorders of eyelid in diseases classified elsewhere, Disorders of male genital organs in diseases classified elsewhere, Noninfective enteritis and colitis, Polyneuropathies and other disorders of the peripheral nervous system	The patient in the scenario presented with signs and symptoms consistent with a right temporal lobe glioblastoma. Glioblastomas are the most common primary brain tumors in adults. They are classified as grade IV gliomas and are associated with a poor prognosis. They are typically treated with surgery, radiotherapy, and chemotherapy. Temozolomide is a commonly used chemotherapy agent for glioblastoma. Bevacizumab is a monoclonal antibody that targets vascular endothelial growth factor (VEGF). It is used in the treatment of several types of cancer, including glioblastoma, but it is not a first-line agent. Cisplatin is a platinum-based chemotherapy agent that is used in the treatment of several types of cancer. Authors' annotation: Generic and misdiagnosed, neglecting the patient's history and symptoms.	Based on the presenting symptoms, medical history, and family history provided, this patient's constellation of symptoms is concerning for a neurological condition involving multiple systems. The presence of severe headaches, blurry vision, lower extremity weakness, numbness, urinary symptoms, ptosis of the left eyelid, and decreased facial expression on the left side may suggest a neurological disorder affecting the cranial nerves, optic nerves, and peripheral nerves. Given the patient's history of inflammatory bowel disease (IBD) and recent diagnosis of noninfective enteritis, there is a possibility of an autoimmune or systemic inflammatory condition affecting the nervous system. Conditions such as multiple sclerosis (MS), neuromyelitis optica (NMO), or vasculitis could be considered. **Management Recommendations**: 1. *Neurological Evaluation*: Urgent referral to a neurologist for a comprehensive evaluation, including a detailed neurological examination and possible lumbar puncture. 2. *Imaging*: MRI of the brain and spinal cord to identify any demyelinating lesions or other abnormalities. 3.*Laboratory Tests*: Blood tests including autoimmune markers (e.g., ANA, ANCA), inflammatory markers (e.g., ESR, CRP), and specific tests for MS and NMO. 4.*Symptom Management*: Address specific symptoms such as pain, urinary issues, and erectile dysfunction with appropriate medications and therapies. Authors' annotation: More case-specific.

Table 4: Full examples of questions and answers: two questions are generated based on the keywords, and the questions are supplied to `Llama-2-7b(4bit)` and `Llama-2-7b`. Two models provide near identical answers to the first question (first grayed row). The second question is a distinguisher question identified by our algorithm. `Llama-2-7b(4bit)` provides a more generic answer, whereas `Llama-2-7b` is more specific.

Generated Question 1. *Case Scenario:* A 45-year-old man presents to the clinic with complaints of recurrent severe headaches, blurry vision, and difficulty urinating for the past 6 months. He also reports experiencing weakness and numbness in his lower extremities. He has a history of inflammatory bowel disease and was recently diagnosed with noninfective enteritis. He is currently being treated with mesalamine and infliximab infusions for his IBD. The patient also mentions that he has noticed drooping of his left eyelid and has been experiencing erectile dysfunction. His father had a history of polyneuropathy. On examination, he is noted to have ptosis of the left eyelid, decreased facial expression on the left side, and diminished tactile sensation in his lower extremities. Cranial nerve examination reveals weakness of the left cranial nerves. *Question:* As a clinical expert, how would you respond to this medical case? **Topics:** *Cranial nerve disorders in diseases classified elsewhere*, *Disorders of eyelid in diseases classified elsewhere*, *Disorders of male genital organs in diseases classified elsewhere*, *Noninfective enteritis and colitis*, and *Polyneuropathies and other disorders of the peripheral nervous system*.

Generated Question 2. *Case Scenario:* A 7-week-old male infant is brought to the pediatrician by his concerned parents. They noticed that he had frequent episodes of nosebleeds since birth, which had become more persistent and difficult to control in the past week. The parents also mention that the infant has been unusually fussy and irritable. On examination, the infant appears pale, with petechiae scattered over his body. The pediatrician notes that the infant's heart rate is elevated, and there are signs of bleeding from the gums. *Question:* Based on the presented case, what congenital malformations or hematological disorders should be considered in the differential diagnosis for this 7-week-old infant presenting with recurrent nosebleeds, petechiae, and irritability? **Topics:** *Congenital malformations of eye, ear, face and neck*, *Persons encountering health services for specific procedures and health care*, and *Haemorrhagic and haematological disorders of fetus and newborn*.

Generated Question 3. *Case Scenario:* A 50-year-old man presents to the emergency room with complaints of worsening lower back pain over the past few months. He describes the pain as dull and aching, aggravated by movement and relieved by rest. He denies any history of trauma or injury to his back. He also reports increased frequency of urination, especially at night, and occasional burning sensation while urinating. On examination, the patient appears uncomfortable while moving, has limited range of motion in his lumbar spine, and tenderness over the lower lumbar region. Additionally, he has no focal neurological deficits. His urine analysis reveals the presence of leukocytes and red blood cells, indicative of a urinary tract infection, but no signs of infection or inflammation in the lumbar region are seen. *Question:* As a clinical expert, how would you respond to this medical case? **Topics:** *Symptoms and signs involving the urinary system*, *Abnormal findings on examination of urine, without diagnosis*, and *Osteopathies and chondropathies*.

Generated Question 4. *Case Scenario:* A 42-year-old man presents to the emergency room with complaints of severe abdominal pain and blurring of vision in his left eye. He reports a history of intermittent abdominal pain and diarrhea over the past few weeks, along with a recent onset of vision changes. He denies any recent travel, fever, or history of infectious illnesses. His medical history is significant for ulcerative colitis, for which he has been on a regimen of mesalamine. Upon further investigation, the patient's physical exam reveals tenderness in the lower abdomen and erythema in the left eye. Fundoscopy shows chorioretinal lesions in the affected eye. Laboratory tests reveal elevated inflammatory markers. *Question:* As a clinical expert, how would you respond to this medical case? **Topics:** *Noninfective enteritis and colitis*, and *Chorioretinal disorders in diseases classified elsewhere*

Generated Question 5. *Case Scenario:* A 45-year-old man, who works as a truck driver, presents to the dermatology clinic with a suspicious mole on his upper back that has been changing in size and color over the past few months. He also reports a history of multiple sunburns during his youth due to prolonged exposure to the sun while driving. Upon examination, the dermatologist notes irregular borders and color variation in the mole, raising concern for melanoma. The patient admits to being worried about skin cancer due to his occupational sun exposure and is eager to undergo further evaluation. *Question:* As a clinical expert, how would you respond to this medical case? **Topics:** *Melanoma and other malignant neoplasms of skin* and *Persons with potential health hazards* *related to communicable diseases.*

B.3 Clinical reviews for Table 4

This section provides clinician reviews for questions and answers in Table 4.

Question A. The differential diagnosis for bleeding disorders in an infant are relatively narrow, and the answers are similar for both LLMs and medically reasonable diagnoses to explore with further workup.

Question B. This is a very concerning case for severe neurologic diseases, with some possibly related but unusual details like the urologic issues and history elements thrown in. The constellation of symptoms does not follow any one pattern and would be confusing in the setting of a question testing medical knoweldge, but theoretically could represent multiple systemic/neuropathic processes. The first answer anchors to glioblastoma alarmingly as the only diagnosis and rather than proposing workup jumps right to treatment of glioblastoma. The 2nd answer is better as it focuses on braod differentials and suggests several types of workup.