

Epidemiology of Large Language Models: A Benchmark for Observational Distribution Knowledge

Drago Plečko¹, Patrik Okanović², Torsten Hoefler², Elias Bareinboim¹

¹Causal AI Lab, Columbia University
{dplecko, eb}@cs.columbia.edu

²Department of Computer Science, ETH Zurich
{patrik.okanovic, torsten.hoefler}@inf.ethz.ch

Abstract

Artificial intelligence (AI) systems hold great promise for advancing various scientific disciplines, and are increasingly used in real-world applications. Despite their remarkable progress, further capabilities are expected in order to achieve more general types of intelligence. A critical distinction in this context is between *factual* knowledge, which can be evaluated against true or false answers (e.g., “what is the capital of England?”), and *probabilistic* knowledge, which reflects probabilistic properties of the real world (e.g., “what is the sex of a computer science graduate in the US?”). In this paper, our goal is to build a benchmark for understanding the capabilities of large language models (LLMs) in terms of knowledge of *observational* distributions. Here, the term observational refers to Pearl’s Causal Hierarchy (PCH), to make explicit that we are interested in observational probabilistic knowledge (Layer 1 of the PCH), as opposed to interventional (Layer 2) or counterfactual (Layer 3) knowledge. Given that LLMs are trained on vast amounts of text, it may be plausible to think that they internalize aspects of real-world observational distributions. Indeed, this idea has gained traction, with LLMs being touted as powerful and universal approximators of real-world distributions. At the same time, classical results in statistics, known under the term curse of dimensionality, highlight fundamental challenges in learning distributions in high dimensions, challenging the notion of universal distributional learning. In this paper, we develop the first benchmark to directly test this hypothesis, evaluating whether LLMs have access to empirical observational distributions (corresponding to Layer 1 of Pearl’s Hierarchy) describing real-world populations across domains such as economics, health, education, and social behavior.

1 Introduction

Artificial intelligence (AI) systems hold great potential to accelerate many scientific disciplines including healthcare [27, 44], education [24], and economics [11]. Automated systems based on AI are also increasingly used in a wide variety of real-world settings. Yet, despite rapid progress and broad adoption of these systems, many core capabilities required for robust and reliable AI still need to be developed further, on the path to reaching artificial human intelligence [14, 29]. One of the hallmark features expected from AI systems is the ability to reason probabilistically about the world, and not only to mimic human intuition and language. Such capabilities are needed for systems capable of assisting scientific discovery, advancing our understanding of complex phenomena, and help societal decision-making.

In this context, it is helpful to distinguish between two types of knowledge embedded in AI systems: *factual* knowledge, which can be evaluated based on true or false answers based on questions for which there is an agreed upon answer (e.g., “what is the capital of England?”); and *probabilistic*

knowledge, which reflects the inherent uncertainties of the world we operate in (e.g., question such as “what is the sex of a computer science graduate in the US?”). Much of the current evaluations of large language models (LLMs), currently the prevailing paradigm in AI research [13], has focused on factual knowledge, i.e., question answering, fact recall, etc. At the same, the degree to which LLMs encode probabilistic knowledge about the real world remains relatively underexplored.

In this work, our focus is on *observational* probabilistic knowledge, where the term observational refers to the first layer of Pearl’s Causal Hierarchy (PCH, for short) [38, 8]. The PCH distinguishes different types of probabilistic knowledge, corresponding to different cognitive capabilities of (AI) systems: Layer 1 of *observation* (seeing), which includes reasoning about observed events and correlations between variables; Layer 2 of *intervention* (doing), which includes inferences about systems in which causal interventions are performed; and Layer 3 of *counterfactual reasoning* (imagining, see Fig. 2 left), which allows inference for hypothesized events for which the required pre-conditions may not have materialized in reality. Important results from previous work demonstrate that, in absence of appropriate assumptions, moving across different layers of PCH is provably impossible. The result known as the Causal Hierarchy Theorem (CHT, [8]) states that observational knowledge (Layer 1) about a system underdetermines its interventional (Layer 2) and counterfactual (Layer 3) behavior. With causal modeling and assumptions, inferences across layers may become feasible. A classical example is causal modeling that uses observational data, combined with structural assumptions, to infer interventional or counterfactual distributions [38, 7]. Therefore, if one can show that LLMs do not have access to observational knowledge (Layer 1), based on the CHT, one may also be skeptical of these models’ capabilities for performing reliable statements about interventions or counterfactuals.

Given the abundance of observational data, it is plausible that LLMs, trained on vast corpora of text data, have internalized aspects of real-world observational distributions. Indeed, strong claims in this debate have emerged, with OpenAI’s CEO Sam Altman suggesting that “humanity has found a universal way to approximate distributions,” [5] implying that large models serve as powerful approximators of the world, and may have access to any kind of Layer 1 knowledge.

While LLMs have been tremendously successful at various tasks, some caution is nonetheless warranted. Classical results in statistics highlight fundamental limitations of learning distributions in higher dimensions: the famous *curse of dimensionality* results demonstrate theoretically that learning distributions becomes exponentially harder (in terms of required samples) as dimensionality grows. Seminal work by Charles Stone [45] establishes that estimation rates degrade sharply with dimensionality, with the optimal rate of convergence for an unknown differentiable regression function f being $\mathcal{O}(n^{-\frac{1}{2+d}})$ in any L^p norm, where d is the dimension of the input of f . Such results challenge the notion of universal distributional learning in higher dimensions, in stark contrast with the notion of LLMs working as universal approximators.

This tension raises a natural question: should one believe the optimism of universal approximation put forward by Sam Altman, or the caution grounded in statistical theory, such as the results of Charles Stone? In this paper, we attempt to shed light on this debate and examine different aspects of LLMs’ capabilities in approximating distributions. We introduce a benchmark specifically designed to assess whether LLMs have access to observational distributions describing real-world populations, across various domains including economics, health, education, crime, and social behavior. This is why our approach can be termed as epidemiology, from Greek *epi* (fall upon) + *demos* (population) + *logos* (knowledge) – the knowledge of what falls upon populations.

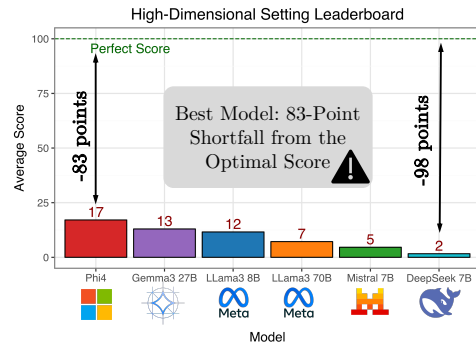


Figure 1: Benchmark leaderboard.

1.1 Contributions

Our contribution in this work is to construct a benchmark for evaluating the capabilities of large language models in terms of access to knowledge about observational distributions in the real-world.

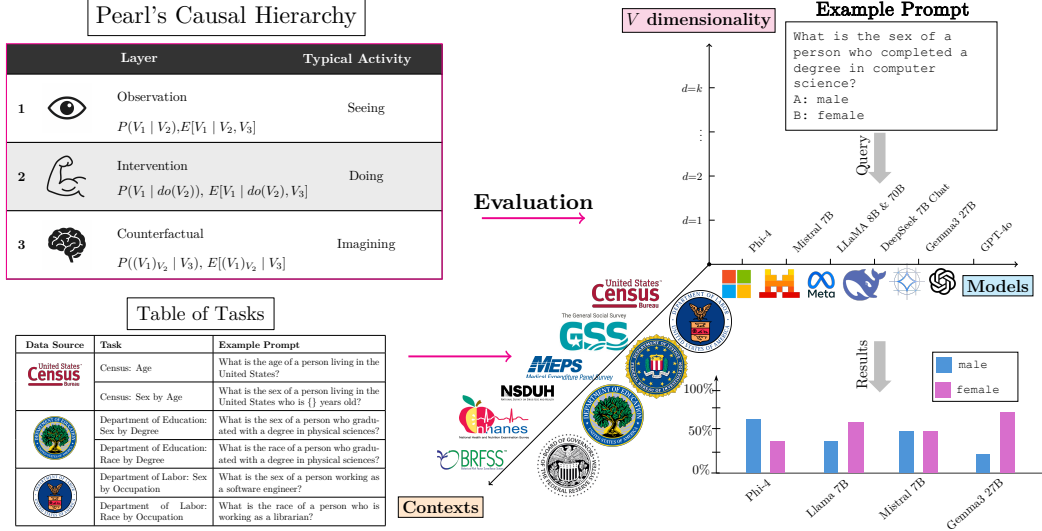


Figure 2: Overview figure for our benchmark.

- (i) We curate a total of 10 datasets, ranging from healthcare, health behavior, and education, to labor, consumer spending, and crime statistics. For each large scale dataset, which describes the population level statistics in the United States, we extract a set of questions to test whether a language model has knowledge about the population encoded in the respective dataset.
- (ii) We design a modular evaluation framework that facilitates easy benchmarking of additional models, making it straightforward to add datasets, questions, or models to our benchmark.

A preview of our results is shown in Fig. 1 indicating that over a range of models, the performance of the current generation of LLMs in terms of knowledge of real-world distributions is quite poor. As we argue later in the text, access to observational distributions is a fundamental step towards performing interventional (Layer 2) or counterfactual (Layer 3) inference, implying that are our results also implicitly challenge the abilities of LLMs in higher layers of Pearl’s Hierarchy.

Related Literature. A growing body of literature investigates aspects of knowledge and generalization abilities of LLMs. The two works most related to ours are that of Santurkar et al. [43], which examines the responses on language models on Pew Research Center’s American Trends Panel [39], in order to investigate if LLMs exhibit responses of a specific demographic group. Another related work is that of Dominguez-Olmedo et al. [18], that examines the responses of language models on the American Community Survey (ACS) [50], and compares them to the real responses of individual. Our work takes a broader view, and attempts to build a benchmark for systematically analyzing the knowledge of models on observational distributions (Layer 1 of the PCH), across a wide range of datasets, and distributions over variables with different dimensions.

Other works investigate the ability of models to perform correct probabilistic inferences, given information in forms of probability tables or probability conditionals [22, 28, 33]. This line of work is distinct from ours, since our benchmark aims to evaluate the availability of the true probabilities to the model, rather than the validity of inferences over these probabilities. We also mention the work of Zhao et al. [54], aimed at testing factual correctness of model responses on various domains, including knowledge about the world, math, or reasoning. In parallel, other lines of work use the term faithfulness while evaluating language models, usually referring to faithfulness of model responses to the the actual source material where the information is available. For instance, [55] investigates whether responses of models are affected by knowledge conflict, and whether they exhibit appropriate response abstentions. This work is focused on testing factual knowledge, as opposed to descriptive knowledge investigated by our benchmark. Another body of work looks at faithfulness of model explanations, that is, whether the explanation cited by the model is in fact compatible with the model’s actual reasoning steps [3, 32], differing from the approach and the goal of our work.

2 Benchmark Axes

We use the language of structural causal models (SCMs) to motivate some of the reasoning in this work [38]. A structural causal model (SCM) is defined as:

Definition 1 (Structural Causal Model (SCM) [38]). *A structural causal model $\mathcal{M}$ is a 4-tuple $\langle V, U, \mathcal{F}, P(u) \rangle$, where*

1. $U = \{U_1, \dots, U_n\}$ is a set of exogenous variables, also called background variables, that are determined by factors outside the model;
2. $V = \{V_1, \dots, V_n\}$ is a set of endogenous (observed) variables, that are determined by variables in the model (i.e., by the variables in $U \cup V$);
3. $\mathcal{F} = \{f_{V_1}, \dots, f_{V_n}\}$ is the set of variable assignment functions determining V , $V_i \leftarrow f_{V_i}(\text{pa}(V_i), U_i)$, where $\text{pa}(V_i) \subseteq V \setminus V_i$ and $U_i \subseteq U$ are the functional arguments of f_{V_i} ;
4. $P(u)$ is a distribution over the exogenous variables U .

The assignment mechanisms $\mathcal{F}$ determine how each of the observed variables V_i attains its value, based on other observed variables and the latent variables U . Together with the probability distribution $P(u)$ over the exogenous variables U , it specifies the entire behavior of the underlying phenomenon, meaning that it fully specifies all observational, interventional, and counterfactual distributions [8].

The remainder of the section is organized according to Fig. 2 and the different axes appearing therein. We first discuss observational knowledge and its different dimensionalities (Sec. 2.1), followed by the description of datasets that are used (Sec. 2.2), and then the tasks constructed for eliciting models' capabilities (Sec. 2.2.1). In Sec. 2.3 we discuss different models currently included in the benchmark.

2.1 Axis 1: Dimensionality of Layer 1 (Observational) Knowledge

Our focus in this paper is on evaluating observational (Layer 1) knowledge available in AI systems. The *observational distribution* of the underlying phenomenon can be defined through the SCM itself:

Definition 2 (Observational Distribution [8]). *An SCM $\mathcal{M}$ that is a 4-tuple $\langle V, U, \mathcal{F}, P(u) \rangle$ induces a joint probability distribution $P(V)$ such that for each $Y \subseteq V$,*

$$P^{\mathcal{M}}(Y = y) = \sum_u \mathbb{1}(Y(u) = y) P(U = u), \quad (1)$$

where $Y(u)$ is the solution for Y after evaluating $\mathcal{F}$ with $U = u$.

In the above definition, $Y \subseteq V$ should be thought of as multidimensional. A fixed value of exogenous $U = u$ is called a *unit*, and can be thought of as a individual in the population. To obtain the observational distribution $P(Y = y)$, we go over all units u , and add up the probability mass $P(U = u)$ of all units for which Y attains the value y . The observational distribution $P(V)$ captures all probabilistic relationships among observed variables under passive observation. On top of the distribution, we can also consider other types of quantities, such as (conditional) moments. In particular, we may consider different types of knowledge, such as:

- (i) **Marginal expectations:** e.g., $\mathbb{E}[V_1]$, the average value of a single variable.
- (ii) **Marginal distributions:** e.g., $P(V_1)$, the full distribution of a variable.
- (iii) **Conditional expectations:** e.g., $\mathbb{E}[V_1|V_2]$, the average value conditioned on another variable.
- (iv) **Conditional distributions:** e.g., $P(V_1|V_2)$, the distribution of one variable given another.
- (v) **Joint distributions:** e.g., $P(V_1, \dots, V_k)$, the distribution of over the variables $V_1, \dots, V_k$.

The above types of knowledge can be viewed as a hierarchy, with the level of required knowledge being progressively more refined. Our goal can be described as follows: we design tasks probing the above types of knowledge, based on large-scale datasets describing populations. For instance, we refer to a dataset $\mathcal{D}$, and consider the distribution over covariates $P(V)$ implied by $\mathcal{D}$. Then, we elicit responses from an LLM about the same distribution. The model's distribution is labeled $\tilde{P}(V)$. Our goal is to then compare various aspects of $\tilde{P}(V)$ and $P(V)$, according to types of knowledge (i) (v)

In this way, we can systematically evaluate whether LLMs internalize such quantities over real-world populations. In fact, we can formally argue that testing knowledge on observational distributions is a critical first step for understanding knowledge on causal capabilities more broadly:

Proposition 1 (Causal Hierarchy Theorem [8]). *Let $\mathcal{M}$ be an SCM, $P(V)$ be its observational distribution, and $\mathcal{A}$ a set of causal assumptions encoded in the form of a causal diagram [8] or ignorability statements [41][38]. Then, we can say that*

- (a) *in absence of $\mathcal{A}$, $P(V)$ underdetermines interventional and counterfactual distributions,*
- (b) *in absence of $P(V)$, $\mathcal{A}$ underdetermines interventional and counterfactual distributions.*

The first part of the above proposition is the one commonly considered in the causal inference literature – in short, it states that, in absence of causal assumptions, it is generally impossible to provide any guarantees for inference over interventional or counterfactual distributions. The second part states that in absence of the observational distribution, no inferences can be made for higher layers of the PCH, leading to the following important corollary:

Corollary 2 (No Observational Distribution $\implies$ No Layer 2/3 Inference). *If a model’s distribution $\tilde{P}(V)$ differs from the true $P(V)$, no guarantees can be provided for the validity of the model’s interventional or counterfactual inferences.*

The above corollary captures an important motivation for the benchmark constructed in this paper – even if we assume that an LLM has access to correct causal knowledge (a point we do not investigate in this manuscript), it still may not be able to perform causal inferences in case its observational distribution does not match the one in the real world. Therefore, testing whether language models have access to observational knowledge (required for causal reasoning) is an important first step since a negative answer warrants skepticism about the model’s capabilities for performing claims in higher layers of the PCH.

2.2 Axis 2: Datasets Describing Populations

As mentioned in the previous section, we will leverage large-scale datasets that describe population level statistics in order to establish the ground truth observational distribution $P(V)$. After this, we will probe the language model’s distribution $\tilde{P}(V)$ to draw comparisons. In this work, for the observational distributions $P(V)$, we make use of ten large, publicly available datasets that collectively describe diverse aspects of the US population, and are often considered representative of the national level statistics:

- (1) **American Community Survey (ACS) 2023** [49]: Conducted by the US Census Bureau, the ACS collects detailed demographic, social, economic, and housing data annually. In this dataset, our focus is on income, education, and employment information across demographics.
- (2) **National Health and Nutrition Examination Survey (NHANES) 2021-2023** [34]: Administered by the Centers for Disease Control and Prevention (CDC), NHANES combines interviews and physical exams to assess health and nutrition of individuals in the US. We investigate obesity, diabetes, and dietary habits across demographics.
- (3) **Behavioral Risk Factor Surveillance System (BRFSS) 2023** [15]: A telephone survey system run by the CDC, tracking health-related risk behaviors and conditions. We investigate exercise habits, diabetes, blood pressure, asthma, cholesterol, visual/auditive impairments by US states.
- (4) **Medical Expenditure Panel Survey (MEPS) 2023** [4]: A set of surveys conducted by the Agency for Healthcare Research and Quality (AHRQ), measuring health services use, expenditures, and insurance coverage. We investigate healthcare expenditure, utilization, and insurance across demographics.
- (5) **National Survey on Drug Use and Health (NSDUH) 2023** [46]: Collected by Substance Abuse and Mental Health Services Administration (SAMHSA), the NSDUH provides information on substance use and mental health in the US. We investigate alcohol, cigarette, marijuana, cocaine, and heroin usage across demographics.
- (6) **Survey of Consumer Finances (SCF) 2022** [12]: Sponsored by the Federal Reserve Board, the SCF provides detailed data on US household finances. We investigate food expenditure, home ownership, assets, and debt across demographics.

- (7) **General Social Survey (GSS) 2022** [35]: Conducted by the National Opinion Research Center (NORC) at the University of Chicago, the GSS collects data on social attitudes, behaviors, and demographics of U.S. adults. We investigate political views and party identification across age, sex, race, education, and income.
- (8) **Department of Education (IPEDS)** [51]: The Integrated Postsecondary Education Data System (IPEDS) collects data from colleges, universities, and technical schools. We investigate college degrees by sex.
- (9) **Department of Labor BLS Data 2023** [48]: The Bureau of Labor Statistics of the US Department of Labor provides detailed information about occupations, including worker demographics. We investigate occupations by sex and race.
- (10) **Federal Bureau of Investigation (FBI) Arrest Statistics (UCR Program)**: Compiled by the FBI’s Uniform Crime Reporting Program (UCR), this dataset contains arrest statistics across the US. We investigate crime rates by race and sex.

Each of the datasets either reports national-level statistics, or provides data on individuals. When individual data is available, sample weights are provided (determining how many persons in the overall US population the individual represents), allowing one to obtain representative results at the national level. Together, these datasets allow for a broad evaluation of models’ abilities to approximate population-level observational distributions across social, economic, educational, health, and behavioral domains. We next describe how different tasks in our benchmark are constructed.

2.2.1 Axis 2 Tasks: Responses and Scoring

Our benchmark consists of different *tasks*, and each task is associated with a dataset. The task is defined by the pair V_Y, V_X , representing the conditional distribution $P(V_Y | V_X)$ (here, V_X could be possibly empty). Each task is also accompanied by a natural language prompt template π (which takes a value of $V_X = v_X$ and returns a natural language question), and a set of valid answers in the domain of V_Y , labeled $\text{dom}(V_Y)$.

Example 1 (NSDUH: Marijuana Usage by Age). *The NSDUH dataset tracks various aspects of addiction and mental health for the US population. Suppose that V_Y represents whether a person ever used marijuana, and V_X represents age. In the benchmark, we are interested whether the LLM has access to the conditional distribution $P(V_Y | V_X)$. To test this, the accompanying prompt template π for the task is given by:*

$$\pi(v_X) = \text{“Has a person aged } \{v_X\} \text{ ever used marijuana?”} \quad (2)$$

Once a specified value $v_X = 16$ is chosen, this results in a prompt “Has a person aged 16 ever used marijuana?”, which is given to the model. In the prompt, the model is provided with a set of possible answers, which in this case amounts to $\text{dom}(V_Y) = \{\text{no}, \text{yes}\}$. For open-weights models, we can inspect the probabilities associated with each of the responses in the model’s next-token prediction. Alternatively, for closed models, the same question can be repeated multiple times, and we record all the provided answers, allowing us to reconstruct the distribution over answers. Finally, we compare the model’s distribution with the true $P(V_Y | V_X)$ distribution from the NSDUH dataset. $\square$

Intuitively, the difficulty of the task should increase with larger dimensions of V_Y and V_X . For instance, if $|V_Y| = 1, |V_X| = 0$, our task is to recover the 1-dimensional marginal distribution $P(V_Y)$. If $|V_Y| = 1, |V_X| > 0$, we are trying to recover a 1-dimensional conditional distribution $P(V_Y | V_X)$. The question templates π are carefully designed, to reflect the original questions asked to individuals during data acquisition. In total, we constructed 75 tasks across 10 datasets for $|V_Y| = 1, |V_X| = 1$ (we refer to this as the low-dimensional setting), and 94 tasks across 4 datasets for $|V_Y| = 1, |V_X| \in \{2, 3, 4, 5\}$ and V_Y binary (referred to as the high-dimensional setting). Fig. 2 (bottom left) offers some further examples, in addition to Ex. 1 while Appendix A contains the full list of tasks. We remark here that an alternative way of eliciting the model’s knowledge also exists, captured by the following example:

Example 1 (continued – NSDUH: Marijuana Usage by Age). *For the task of eliciting the distribution of marijuana usage by age group, a template prompt*

$$\pi^{\text{prob}}(v_X) = \text{“What is the probability that a person aged } \{v_X\} \text{ ever used marijuana?”} \quad (3)$$

may be used instead, representing an alternative way of eliciting $\tilde{P}(V_Y | V_X)$. $\square$

The comparison of the two prompting approaches (probabilistic vs. Q&A) is discussed in Appendix D while in the main text we focus on Q&A prompt templates as in Eq. 2.

Eliciting Responses. As mentioned, for each question the model is asked, there is a specific range of possible responses. Our approach to obtain the model answers is to label each response in $\text{dom}(V_Y)$ with labels $A, B, C, \dots$, and then inspect the conditional probability associated with tokens A, B, C , etc. Let $\mathcal{A}(v_Y)$ denote the letter of the answer v_Y . The model’s probability associated with each answer $v_Y \in \text{dom}(V_Y)$ corresponding to label $\mathcal{A}(v_Y)$ in the prompt is then simply computed as

$$\tilde{P}(V_Y = v_Y \mid V_X = v_X) \triangleq \frac{\tilde{P}^{\text{nt}}(\text{answer } \mathcal{A}(v_Y))}{\sum_{\tilde{v}_Y \in \text{dom}(V_Y)} \tilde{P}^{\text{nt}}(\text{answer } \mathcal{A}(\tilde{v}_Y))}, \quad (4)$$

where $\tilde{P}^{\text{nt}}$ is the model’s next-token probability function given the prompt.

Example 1 (continued – NSDUH: Marijuana Usage by Age). *Suppose that following the prompt*

Has a person aged 16 years ever used marijuana?
A. yes
B. no

we find the model’s next-token probabilities to be $\tilde{P}^{\text{nt}}(\text{answer } A) = 0.01$, $\tilde{P}^{\text{nt}}(\text{answer } B) = 0.03$. Then, the model’s probability of a positive outcome would be computed as:

$$P(V_Y = 1 \mid V_X = 16) = \frac{0.01}{0.01 + 0.03} = 25\%. \quad (5)$$

□

This is a known strategy for eliciting model responses used in numerous works on question answering [18, 23, 43]. As models may exhibit ordering bias [18], we consider all of the different $|\text{dom}(V_Y)|!$ permutations of labels $A, B, C, \dots$, over the answers $v_Y \in \text{dom}(V_Y)$, and average the probabilities accordingly (if the number of permutations $|\text{dom}(V_Y)|! > 120$, we consider 120 random permutations). The above way of eliciting responses may be used for all models with open-source weights, which allows us to access next-token prediction probabilities. An adaptation of this approach is used for closed models requiring API queries (see Sec. 2.3 for more details). In this case, the same prompt is given to the model, but we sample n_{mc} Monte Carlo samples, and then determine the model’s distribution $\tilde{P}$ based on the MC sample, using bin-counting. For the ground truth distribution P , in the low-dimensional setting we use a simple bin-counting estimator on the available dataset, whereas in the high-dimensional case we fit the distribution $P(V_Y \mid V_X)$ using `lightgbm` [30], ensuring that the actual values $P(V_Y = v_Y \mid V_X = v_X)$ are obtained out-of-sample (using cross-validation).

Scoring Strategy. We next develop a scoring strategy for each task based on the true distribution $P(V_Y \mid V_X)$ and the model’s distribution $\tilde{P}(V_Y \mid V_X)$. For this purpose, we use the L_1 -norm (instead of KL divergence, to avoid sensitivity to low-probability events and support mismatch [19]), and define the following distributional distance:

$$D(\tilde{P} \parallel P; V_Y, V_X) = \sum_{v_X} \sum_{v_Y} \left| P(V_Y = v_Y \mid V_X = v_X) - \tilde{P}(V_Y = v_Y \mid V_X = v_X) \right| P(V_X = v_X) \quad (6)$$

The above notion of distance allows the evaluation of how far $P, \tilde{P}$ are for the specific conditional $V_Y \mid V_X$. To obtain a normalized score, we compare the distance $D(P \parallel \tilde{P})$ to the distance of P from a uniform distribution over the answers, labeled P^{unif} , defined as

$$P^{\text{unif}}(V_Y = v_Y \mid V_X = v_X) \triangleq \frac{1}{|\text{dom}(V_Y)|} \quad \forall v_X, v_Y. \quad (7)$$

In this context, we refer to the P^{unif} distribution as a random *baseline*. In the high-dimensional case, which focuses on binary V_Y , we add another baseline, namely a fixed 0/1 prediction depending on the dataset mean, given by $P^{0/1}(V_Y = 1 \mid V_X = v_X) \triangleq \mathbb{1}(\mathbb{E}_P[V_Y] > 0.5)$ for all v_Y, v_X , where $\mathbb{E}_P[V_Y]$ represent the true distribution mean. Therefore, the 0/1 baseline predicts a constant probability of 0 for all conditioning sets for outcomes with a marginal incidence $\mathbb{E}_P[V_Y] \leq 0.5$, and predicts a constant 1 whenever the marginal incidence is $\mathbb{E}_P[V_Y] > 0.5$. Such a baseline also

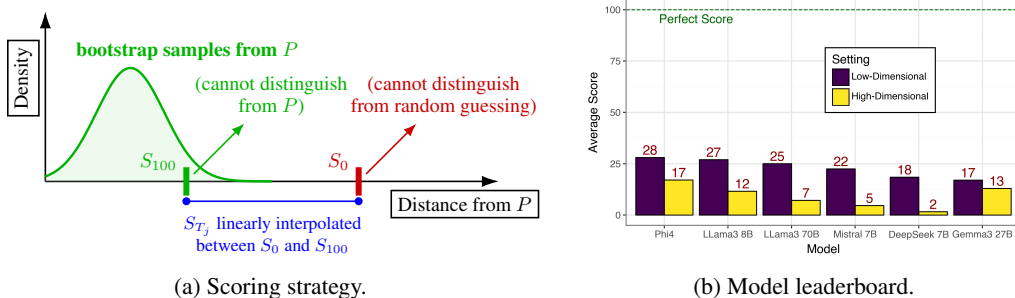


Figure 3: Scoring strategy and model performance across task groups.

corresponds to effectively no probabilistic knowledge. With these baselines in place, our scoring system is described next (illustrated in Fig. 3a). As many datasets consist of samples of the population (and do not survey the entire population), there is some uncertainty on the ground truth distribution P . In the scoring, we account for this fact, and proceed as follows. We draw bootstrap samples of size $|\mathcal{D}|$ (size of the dataset) labeled $\mathcal{D}^{(b)}$, and extract the bootstrap ground truth distribution $P^{(b)}$. We then look at the distance $D(P||P^{(b)})$ of P (based on the available data) and $P^{(b)}$ (obtained from bootstrapped data), across different samples (see green density in Fig. 3a). The upper 5% quantile of the spread of $D(P||P^{(b)})$ is labeled with the perfect score $S = 100$ – meaning that the score 100 is assigned if a model’s distribution cannot be statistically distinguished from the ground truth at the 5% significance level. For setting the score of 0, we use the minimum of the distances $D(P||P^{\text{unif}})$, $D(P||P^{0.1})$, with the latter only considered for binary outcomes. The model’s final score is then linearly interpolated between $S = 0$ and $S = 100$, as follows:

$$S_T = 100 \times \max \left(1 - \frac{D(\tilde{P}||P)}{\min \{D(P^{\text{unif}}||P), D(P^{0.1}||P)\}}, 0 \right). \quad (8)$$

2.3 Axis 3: Models

Models we consider can be divided in two groups: *open-weight models*, whose trained parameters are publicly available, and *closed-weight models*, whose parameters remain proprietary and can be accessed only via an API. We focus mainly on open models, but also consider some closed models to ensure our benchmark covers a broad range of state-of-the-art models.

Open-weight Models. Our open-weight models include both the original pretrained checkpoints, and *instruction-tuned* variants, which are further trained on human-generated instruction–response pairs to improve their ability to follow instructions in user prompts. Particularly, we evaluate the following models: Mistral 7B [26], LLaMA3 8B [20], LLaMA3 70B [20], Gemma3 27B [47], DeepSeek 7B [10], Phi-4 [1], and DeepSeek R1 32B [10]. Appendix B contains a detailed comparison of these models.

Closed-weight Models. Closed-weight models are available only through API-based access to proprietary weights, and include a state-of-the-art reasoning model. In particular, we evaluate Reasoning Language Models (RLMs) such as OpenAI’s o4-mini. These models augment standard LLM backbones with components like Monte Carlo Tree Search, policy and value networks, and reinforcement learning to perform explicit, multi-step reasoning [9]. We also include models with web access that use retrieval-augmented generation (RAG) [31], such as OpenAI’s GPT-4.1. These models pull in external documents at inference time—so as to assess how dynamic knowledge augmentation impacts performance. Appendix C contains the details on the performance of selected closed models.

2.4 Benchmark Construction

Our benchmark is designed with reproducibility and extensibility as key principles. All dataset constructions are made available, ensuring transparency and verifiability. The modular structure of the benchmark allows easy addition of both new datasets and new models, facilitating broader future

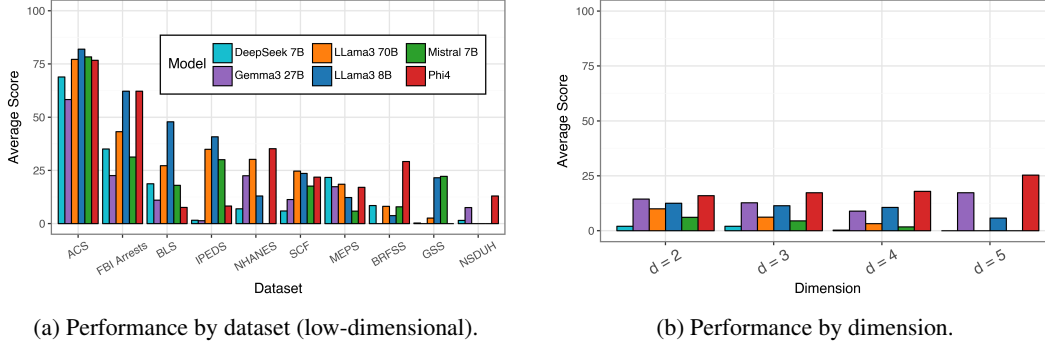


Figure 4: Model performance across (a) datasets; (b) dimension.

evaluations. Furthermore, model evaluation is streamlined: any model available through Hugging Face [53] can be benchmarked with a single line of code, allowing for rapid inclusion of diverse architectures. At the time of submission, our benchmark consists of 10 pre-processed datasets, 169 tasks (75 low-dimensional and 94 high-dimensional), with 12 open-weight models evaluated.

3 Key Insights and Observations

Our results summarizing the leaderboards for the low-dimensional and high-dimensional settings are shown in Figure 3b. The results indicate that the observational distributions $\hat{P}(V)$ encoded in the LLMs are closer to the ground truth, real-world observational distributions $P(V)$ than the uniform distribution $P^{\text{unif}}(V)$ since models achieve scores above zero. All models achieve scores better than uniform guessing, suggesting some access to L1 knowledge. However, overall, the models’ performance is rather poor, with the best model scoring an average of 28/100 points in the low-dimensional setting, and 17/100 in the high-dimensional setting. Therefore, capabilities of LLMs for encoding real-world observational distributions seem limited, and using such knowledge for downstream tasks should be done with care. We next analyze different aspects of model performance.

Low-Dimensional Setting & Performance by Dataset. The performance of models on 75 low-dimensional tasks, comparing $\hat{P}(V_Y | V_X)$ with the ground truth $P(V_Y | V_X)$ for $|V_Y| = |V_X| = 1$ is shown in Fig. 3b (purple bars). Among a selection of models from different organizations (Meta, Microsoft, Google, DeepSeek AI, Mistral AI), the overall performance across models does not exhibit large differences, and all models perform poorly. Another interesting aspect is the performance of models across different datasets, visualized in Fig. 4a. The best performance is observed over the ACS Census data (questions on employment, education, and income across demographics), FBI Arrest Statistics (sex/race distributions over crime types), and Departments of Education (sex by graduation degree) and Labor (sex/race by occupation). The latter three datasets, in fact, offer these detailed statistics on their websites, meaning that models could have had access to the exact probability tables used for constructing our questions. Therefore, we hypothesize that improved performance may correspond to the actual data being available in the model’s training set. This observation also emphasizes the importance of the high-dimensional setting – in which we condition on multiple observed variables – as this knowledge is not readily available on the web.

High-Dimensional Setting & Performance by Dimension. The performance of models in the high-dimensional case, for $P(V_Y | V_X)$ with $|V_Y| = 1$, $|V_X| \in \{2, 3, 4, 5\}$, and V_Y binary, is shown in Fig. 3b (yellow bars). As the scores indicate, performance is again quite poor across models, and certainly worse than in the low-dimensional case. We also inspect the model performance with varying dimension (Fig. 4b). The pattern of the performances does not clearly indicate the curse of dimensionality, possibly due to the fact that each dimension d has a relatively small number of tasks, and additional tasks may need to be added to investigate this effect in more detail.

Implications on Causal Inferences. The recovery of high-dimensional conditional distributions investigated in this work may be seen as related to questions of causal inference. Commonly, to adjust for confounding by a set of covariates V_Z , data analysts may wish to adjust for the propensity of a

treatment V_X , that is, reweigh the samples with $P(V_X = v_X \mid V_Z = v_Z)^{-1}$ to obtain estimates of a desired potential outcome [42, 40]. The other strategy is to use a predictive model for the outcome variable V_Y , namely, a model for the distribution $P(V_Y \mid V_X, V_Z)$ in order to allow for or improve causal inference [6]. However, if these conditional distributions are not available to a model, eliciting them may be a futile exercise. Therefore, our work brings into question some recent approaches claiming to improve causal inferences using foundation models [17]. In our view, using powerful AI models for causal inferences may require a more sophisticated approach.

Limitations & Future Work. A limitation of our work is the limited number of models evaluated on our benchmark. However, by choosing a representative sample of models, we attempted to cover a wide range of models developed by different organizations. Furthermore, our emphasis on easy access and evaluation of models on the benchmark will facilitate the benchmarking of other models, currently left for future work. Another limitation is the fact that high-dimensional inference is investigated across 4 datasets, and the number of datasets and contexts for evaluation should grow over time. In terms of future work, we envision that our work will set the basis for evaluating capabilities of AI models for inference in Layers 2 & 3 of the PCH – to systematically test how observational distributions available to models affect downstream tasks such as causal or counterfactual inferences.

4 Conclusions

One of the initial motivations of this paper is the broad insight coming from the discipline of causal inference: to make claims about interventional or counterfactual aspects of reality (Layers 2 and 3 of the Pearl Causal Hierarchy), one needs two key ingredients – (i) observational (associational) data (Layer 1) and (ii) qualitative assumptions about the environment. Since AI systems aim to perform such inferences, we ask whether they even possess the first ingredient. (Of course, the question of whether LLMs are capable of answering probabilistic queries about the world is also interesting in its own right.) To that end, we introduce the first benchmark designed to systematically evaluate whether LLMs can provide observational information across diverse domains; or, put differently, whether their statistics are aligned with the world they are intended to model. Our benchmark assesses language models on a range of associational quantities for which the ground truth is relatively well established. These experiments test the hypothesis that LLMs can serve as a proxy for real-world associations and, in principle, support causal inference. However, results across a broad suite of models suggest that their statistics are consistently and substantively misaligned with the ground truth. We hope that these sobering findings motivate LLM developers to improve the statistical fidelity of their models – a necessary, if insufficient, step toward enabling credible causal reasoning.

Acknowledgements

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

- [1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] C. Agarwal, S. H. Tanneru, and H. Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.
- [4] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey. <https://www.meps.ahrq.gov/> 2023. U.S. Department of Health and Human Services.
- [5] S. Altman. The intelligence age. <https://ia.samaltman.com/> 2024. Accessed: 2025-04-28.

- [6] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- [7] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. In R. M. Shiffrin, editor, *Proceedings of the National Academy of Sciences*, volume 113, pages 7345–7352. National Academy of Sciences, 2016.
- [8] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl’s hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.
- [9] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, T. Sternal, M. Copik, G. Kwaśniewski, J. Müller, Łukasz Flis, H. Eberhard, H. Niewiadomski, and T. Hoeffer. Reasoning language models: A blueprint, 2025. URL <https://arxiv.org/abs/2501.11223>.
- [10] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [11] S. J. Bickley, H. F. Chan, and B. Torgler. Artificial intelligence in the field of economics. *Scientometrics*, 127(4):2055–2084, 2022.
- [12] Board of Governors of the Federal Reserve System. Survey of Consumer Finances. <https://www.federalreserve.gov/econres/scfindex.htm> 2022.
- [13] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [14] S. Bubeck, V. Chadrsekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [15] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. <https://www.cdc.gov/brfss/> 2023. U.S. Department of Health and Human Services.
- [16] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- [17] P. De Bartolomeis, J. Abad, G. Wang, K. Donhauser, R. M. Duch, F. Yang, and I. J. Dahabreh. Efficient randomized experiments using foundation models. *arXiv preprint arXiv:2502.04262*, 2025.
- [18] R. Dominguez-Olmedo, M. Hardt, and C. Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878, 2024.
- [19] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.
- [20] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [21] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [22] J. He-Yueya, G. Poesia, R. E. Wang, and N. D. Goodman. Solving math word problems by combining language models with symbolic solvers. *arXiv preprint arXiv:2304.09102*, 2023.
- [23] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

- [24] W. Holmes and I. Tuomi. State of the art and practice in ai in education. *European journal of education*, 57(4):542–570, 2022.
- [25] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.
- [26] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- [27] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.
- [28] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.
- [29] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.
- [31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [32] K. Matton, R. O. Ness, J. Guttag, and E. Kıcıman. Walk the talk? measuring the faithfulness of large language model explanations. *arXiv preprint arXiv:2504.14150*, 2025.
- [33] A. Nafar, K. B. Venable, and P. Kordjamshidi. Probabilistic reasoning in generative large language models. *arXiv e-prints*, pages arXiv–2402, 2024.
- [34] National Center for Health Statistics. National Health and Nutrition Examination Survey Data. <https://www.cdc.gov/nchs/nhanes/> 2023. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
- [35] NORC at the University of Chicago. General Social Survey (GSS). <https://gss.norc.org/>, 2022. Accessed April 30, 2025.
- [36] OpenAI. Openai o3 and o4-mini system card, 2025. URL <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: 2025-05-22.
- [37] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [38] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.
- [39] Pew Research Center. American trends panel, 2024. URL <https://www.pewresearch.org/american-trends-panel/>. Accessed: 2025-04-28.
- [40] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [41] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- [42] D. B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

- [43] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.
- [44] M. Y. Shaheen. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*, 2021.
- [45] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.
- [46] Substance Abuse and Mental Health Services Administration. National Survey on Drug Use and Health (NSDUH). <https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health>, 2023. U.S. Department of Health and Human Services.
- [47] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [48] U.S. Bureau of Labor Statistics. Labor Force Statistics. <https://www.bls.gov/>, 2023. U.S. Department of Labor.
- [49] U.S. Census Bureau. American Community Survey 1-Year Estimates, 2023. Retrieved from <https://data.census.gov/>.
- [50] U.S. Census Bureau. American community survey, 2024. URL <https://www.census.gov/programs-surveys/acs/>. Accessed: 2025-04-28.
- [51] U.S. Department of Education, National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS). <https://nces.ed.gov/ipeds>, 2023. Accessed April 30, 2025.
- [52] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [53] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [54] Y. Zhao, J. Zhang, I. Chern, S. Gao, P. Liu, J. He, et al. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36: 44502–44523, 2023.
- [55] W. Zhou, S. Zhang, H. Poon, and M. Chen. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*, 2023.

Supplementary Material for *Epidemiology of Large Language Models: A Benchmark for Observational Distribution Knowledge*

The source code for reproducing the benchmark can be found in the anonymized code repository <https://anonymous.4open.science/r/llm-epidemia-563D>. The repository includes a README file explaining all the steps for setting up the benchmark. For the Llama3 8B, Mistral 7B, and Phi4 models the experiments were run on a single NVIDIA H100 GPU. Evaluating both the low- and high-dimensional settings required <1 hour per model. Evaluating DeepSeek 7B (Chat version), Gemma3 27B, and LLama3 70B models was done on a node with four NVIDIA GH200 Grace Hopper chips, and running both settings required <1 hour of compute per model.

A Tasks

This appendix contains the full list of tasks considered in the benchmark, complementing the description of task construction from Sec. 2.2.1. In Tab. 1 the 75 tasks for the low-dimensional setting are listed, which are concerned with recovering $P(V_Y | V_X)$ for univariate V_X, V_Y . Tab. 2 lists all the 94 tasks for the high-dimensional setting, recovering $P(V_Y | V_X)$ for V_Y univariate and binary, and $|V_X| \in \{2, 3, 4, 5\}$.

Table 1: Low-dimensional tasks.

Task #	Task Name
1	Census: Employment Status by Sex
2	Census: Employment Status by Race
3	Census: Employment Status by Age
4	Census: Employer by Sex
5	Census: Employer by Race
6	Census: Employer by Age
7	Census: Salary by Sex
8	Census: Salary by Race
9	Census: Salary by Age
10	Census: Education by Sex
11	Census: Education by Race
12	Census: Education by Age
13	BRFSS: Exercise by State
14	BRFSS: Diabetes by State
15	BRFSS: High BP by State
16	BRFSS: Asthma by State
17	BRFSS: Cholesterol by State
18	BRFSS: Visual Impairments by State
19	BRFSS: Hearing Impairments by State
20	BRFSS: Heart Attack by State
21	BRFSS: Stroke by State
22	Department of Education: Sex by Type of Degree
23	FBI Crime Statistics: Sex by Crime Type
24	FBI Crime Statistics: Race by Crime Type
25	GSS: Political View by Age
26	GSS: Political View by Race
27	GSS: Political View by Education
28	GSS: Political View by Income
29	GSS: Political View by Sex
30	GSS: Party Affiliation by Age
31	GSS: Party Affiliation by Race
32	GSS: Party Affiliation by Education
33	GSS: Party Affiliation by Income
34	GSS: Party Affiliation by Sex
35	Department of Labor: Sex by Occupation
36	Department of Labor: Race by Occupation

Task #	Task Name
37	MEPS: Expenditure by Age Group
38	MEPS: Office-based Visits by Age Group
39	MEPS: Inpatient Visits by Age Group
40	MEPS: Dental Visits by Age Group
41	MEPS: Has Insurance by Age Group
42	MEPS: Expenditure by Race
43	MEPS: Office-based Visits by Race
44	MEPS: Inpatient Visits by Race
45	MEPS: Dental Visits by Race
46	MEPS: Has Insurance by Race
47	NHANES: Age by BMI Group
48	NHANES: Diabetes by BMI Group
49	NHANES: Diabetes by Age Group
50	NHANES: Weekly Alcohol Consumption by Age Group
51	NSDUH: Alcohol Use in Last Month by Age
52	NSDUH: Cigarette Use in Last Month by Age
53	NSDUH: Marijuana Ever Used by Age
54	NSDUH: Cocaine Ever Used by Age
55	NSDUH: Heroin Ever Used by Age
56	NSDUH: Alcohol Use in Last Month by Race
57	NSDUH: Cigarette Use in Last Month by Race
58	NSDUH: Marijuana Ever Used by Race
59	NSDUH: Cocaine Ever Used by Race
60	NSDUH: Heroin Ever Used by Race
61	SCF: Food Expenditure by Age Group
62	SCF: House Ownership by Age Group
63	SCF: Total Assets by Age Group
64	SCF: Debt by Age Group
65	SCF: Net Worth by Age Group
66	SCF: Food Expenditure by Race
67	SCF: House Ownership by Race
68	SCF: Total Assets by Race
69	SCF: Debt by Race
70	SCF: Net Worth by Race
71	SCF: Food Expenditure by Education
72	SCF: House Ownership by Education
73	SCF: Total Assets by Education
74	SCF: Debt by Education
75	SCF: Net Worth by Education

Table 2: High-dimensional tasks. The third column indicates the task dimension $d = |V_X|$, and the fourth column indicates whether the task is used for the evaluation of closed models in Appendix C

Task #	Task Name	d	Closed Eval
1	BRFSS: Diabetes by Sex, Race	2	✓
2	BRFSS: Diabetes by Age, Income	2	✓
3	BRFSS: Diabetes by Age, Race	2	✓
4	BRFSS: Diabetes by Sex, Income	2	✓
5	BRFSS: Diabetes by Age, Sex, Race	3	✓
6	BRFSS: Diabetes by Age, Race, Income	3	✗
7	BRFSS: Diabetes by Age, Education, Income	3	✓
8	BRFSS: Diabetes by Age, Sex, Income	3	✓
9	BRFSS: Diabetes by Age, Education, Sex, Race	4	✗
10	BRFSS: Diabetes by Age, Sex, Race, Income	4	✗
11	BRFSS: Diabetes by Age, Education, Race, Income	4	✗
12	BRFSS: Diabetes by Age, Education, Sex, Income	4	✗
13	BRFSS: Diabetes by Age, Education, Sex, Race, Income	5	✗

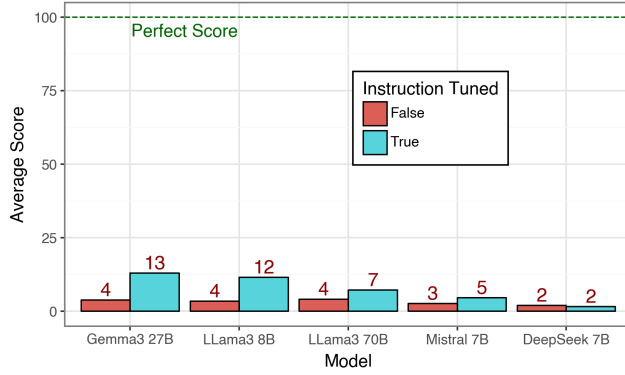
Task #	Task Name	d	Closed Eval
14	BRFSS: High Blood Pressure by Education, Race	2	✓
15	BRFSS: High Blood Pressure by Age, Race	2	✓
16	BRFSS: High Blood Pressure by Sex, Income	2	✓
17	BRFSS: High Blood Pressure by Race, Income	2	✓
18	BRFSS: High Blood Pressure by Age, Sex, Income	3	✓
19	BRFSS: High Blood Pressure by Sex, Race, Income	3	✓
20	BRFSS: High Blood Pressure by Age, Education, Race	3	✓
21	BRFSS: High Blood Pressure by Education, Sex, Income	3	✓
22	BRFSS: High Blood Pressure by Age, Sex, Race, Income	4	x
23	BRFSS: High Blood Pressure by Age, Education, Sex, Income	4	x
24	BRFSS: High Blood Pressure by Age, Education, Sex, Race	4	x
25	BRFSS: High Blood Pressure by Age, Education, Race, Income	4	x
26	BRFSS: High Blood Pressure by Age, Education, Sex, Race, Income	5	x
27	BRFSS: Depression by Race, Income	2	✓
28	BRFSS: Depression by Sex, Race	2	✓
29	BRFSS: Depression by Age, Income	2	✓
30	BRFSS: Depression by Education, Race	2	✓
31	BRFSS: Depression by Age, Race, Income	3	x
32	BRFSS: Depression by Education, Sex, Income	3	✓
33	BRFSS: Depression by Age, Sex, Income	3	✓
34	BRFSS: Depression by Sex, Race, Income	3	✓
35	BRFSS: Depression by Age, Education, Sex, Race	4	x
36	BRFSS: Depression by Age, Education, Sex, Income	4	x
37	BRFSS: Depression by Education, Sex, Race, Income	4	x
38	BRFSS: Depression by Age, Sex, Race, Income	4	x
39	BRFSS: Depression by Age, Education, Sex, Race, Income	5	x
40	MEPS: Health Insurance by Education, Sex	2	✓
41	MEPS: Health Insurance by Age, Sex	2	✓
42	MEPS: Health Insurance by Age, Race	2	x
43	MEPS: Health Insurance by Education, Race	2	✓
44	MEPS: Health Insurance by Age, Education	2	x
45	MEPS: Health Insurance by Sex, Race	2	✓
46	MEPS: Health Insurance by Education, Sex, Race	3	✓
47	MEPS: Health Insurance by Age, Sex, Race	3	x
48	MEPS: Health Insurance by Age, Education, Sex	3	x
49	MEPS: Health Insurance by Age, Education, Race	3	x
50	MEPS: Health Insurance by Age, Education, Sex, Race	4	x
51	NSDUH: Cigarette Use (Last 30d) by Education, Sex	2	✓
52	NSDUH: Cigarette Use (Last 30d) by Age, Sex	2	✓
53	NSDUH: Cigarette Use (Last 30d) by Age, Race	2	✓
54	NSDUH: Cigarette Use (Last 30d) by Education, Race	2	✓
55	NSDUH: Cigarette Use (Last 30d) by Age, Education	2	✓
56	NSDUH: Cigarette Use (Last 30d) by Sex, Race	2	✓
57	NSDUH: Cigarette Use (Last 30d) by Education, Sex, Race	3	✓
58	NSDUH: Cigarette Use (Last 30d) by Age, Sex, Race	3	✓
59	NSDUH: Cigarette Use (Last 30d) by Age, Education, Sex	3	✓
60	NSDUH: Cigarette Use (Last 30d) by Age, Education, Race	3	x
61	NSDUH: Cigarette Use (Last 30d) by Age, Education, Sex, Race	4	x
62	NSDUH: Marijuana Use by Education, Sex	2	✓
63	NSDUH: Marijuana Use by Age, Sex	2	✓
64	NSDUH: Marijuana Use by Age, Race	2	✓
65	NSDUH: Marijuana Use by Education, Race	2	✓
66	NSDUH: Marijuana Use by Age, Education	2	✓
67	NSDUH: Marijuana Use by Sex, Race	2	✓
68	NSDUH: Marijuana Use by Age, Education, Race	3	x
69	NSDUH: Marijuana Use by Age, Sex, Race	3	✓
70	NSDUH: Marijuana Use by Education, Sex, Race	3	✓
71	NSDUH: Marijuana Use by Age, Education, Sex	3	✓

Task #	Task Name	<i>d</i>	Closed Eval
72	NSDUH: Marijuana Use by Age, Education, Sex, Race	4	✗
73	NSDUH: Cocaine Use by Education, Race	2	✓
74	NSDUH: Cocaine Use by Education, Sex	2	✓
75	NSDUH: Cocaine Use by Age, Race	2	✓
76	NSDUH: Cocaine Use by Age, Education	2	✓
77	NSDUH: Cocaine Use by Sex, Race	2	✓
78	NSDUH: Cocaine Use by Age, Sex	2	✓
79	NSDUH: Cocaine Use by Age, Sex, Race	3	✓
80	NSDUH: Cocaine Use by Age, Education, Race	3	✗
81	NSDUH: Cocaine Use by Age, Education, Sex	3	✓
82	NSDUH: Cocaine Use by Education, Sex, Race	3	✓
83	NSDUH: Cocaine Use by Age, Education, Sex, Race	4	✗
84	SCF: Home Ownership by Education, Sex	2	✓
85	SCF: Home Ownership by Age, Sex	2	✓
86	SCF: Home Ownership by Age, Race	2	✓
87	SCF: Home Ownership by Education, Race	2	✓
88	SCF: Home Ownership by Age, Education	2	✓
89	SCF: Home Ownership by Sex, Race	2	✓
90	SCF: Home Ownership by Education, Sex, Race	3	✓
91	SCF: Home Ownership by Age, Sex, Race	3	✓
92	SCF: Home Ownership by Age, Education, Sex	3	✓
93	SCF: Home Ownership by Age, Education, Race	3	✗
94	SCF: Home Ownership by Age, Education, Sex, Race	4	✗

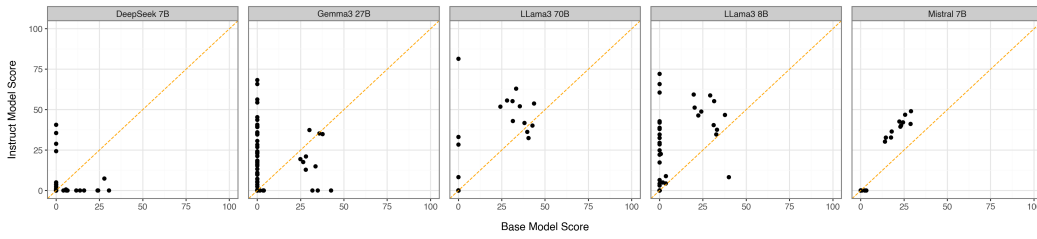
B Comparison of Base & Instruction-Tuned Models

In this appendix, we compare the performance of base models vs. instruction-tuned models optimized via either self-instruction [52] or supervised fine-tuning [37] followed by reinforcement learning with human feedback (RLHF) [16]. In the main text, we analyzed the performance of instruction-tuned models on our benchmark: Mistral 7B [26], LLaMA3 8B [20], LLaMA3 70B [20], Gemma3 27B [47], DeepSeek 7B [10], and Phi-4 [1]. For each of these models, we compare the model’s performance against the corresponding base (pre-trained) model. The only exception is the Phi-4 model, for which a base model is not released due to safety concerns [1].

For evaluating the models on our benchmark, we focus on the 94 tasks in the high-dimensional setting (see Sec. 2.2.1 and Tab. 2 for details). The results of comparing base vs. instruction-tuned models are shown in Fig. 5a, which again indicates poor performance across all models. We note that, for all families of models (here, a family refers to a pair of base and instruction-tuned models), instruction-tuned models perform equal or better than base models. Furthermore, a comparison of base model’s vs. instruction-tuned model’s scores for each task is shown in Fig. 5b. The figure indicates that for a number of tasks, instruction-tuning improves performance from a zero score to a non-zero score. Less commonly, instruction tuning makes performance worse (some instances are observed for DeepSeek and Gemma 3 model families).



(a) Comparing the performance of base and instruct models on high-dimensional tasks.



(b) Performance of base and instruct models by task and model family.

Figure 5: Performance of base vs. instruction-tuned models.

C Closed Models Performance

In this appendix, we look at how the performance of closed models compares to the performance of open-weight models. In particular, we evaluate two closed models from OpenAI – GPT-4.1 [2] and o4-mini [36], where the latter model is a reasoning language model (RLM) [9]. Since these models require access to an API, we evaluate them on a subset of the tasks in the high-dimensional setting (consisting of tasks with a univariate binary V_Y , and $|V_X| \in \{2, 3, 4, 5\}$). In particular, we select 63 out of 94 tasks (the selected tasks are indicated in the Closed Eval column of Tab. 2), focusing on tasks for which at most 250 queries need to be evaluated on the model.

To elicit the model’s distribution $\tilde{P}(V_Y | V_X)$, we use the likelihood-prompting technique, described in detail in Appendix D. We choose likelihood prompting, since closed models such as GPT-4.1 and o4-mini do not provide access to next-token prediction probabilities, meaning that question/answer-prompting used in the main text becomes possible only via Monte Carlo sampling. This, however, would increase the number of queries sent to the models and the associated cost by at least two orders of magnitude. This is due to the fact that even with $n_{mc} = 100$ Monte Carlo samples, the 95% confidence interval for a probability $\tilde{P}(V_Y = 1 | V_X = v_X) \in [0, 1]$ is only guaranteed to be reduced to a width of approximately 0.1, still reflecting a high level of uncertainty.

The performance of GPT-4.1 and o4-mini models on the 63 selected tasks is shown in Fig. 6. The

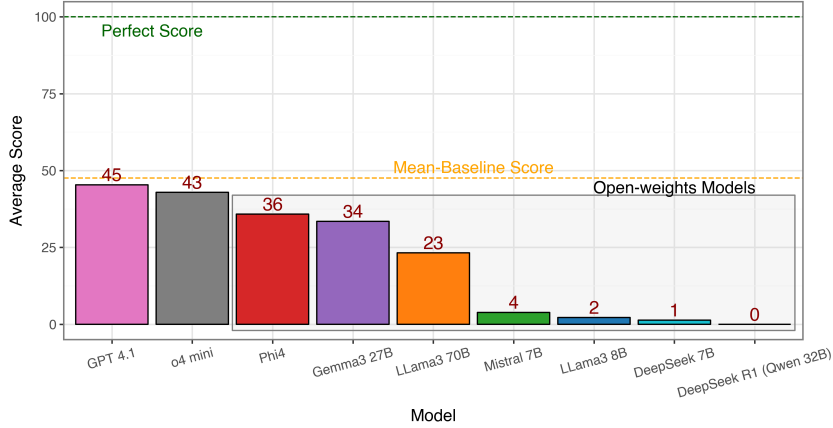


Figure 6: Performance of closed OpenAI models GPT-4.1 and o4-mini.

figure shows that GPT-4.1 and o4-mini outperform the evaluated open-weights models. This may be explained by the fact that GPT-4.1 is the largest model we evaluated, in terms of the parameter count (even though the number of parameters is not officially published). Even though OpenAI’s models outperform the evaluated open-weights models, they still do not perform well. In Fig. 6 the orange line indicates the performance of the so-called *mean-baseline*, which outputs the overall population mean $P^{\text{mean}}(V_Y = 1 | V_X = v_X) = \mathbb{E}_P[V_Y]$ for all values $v_X \in \text{dom}(V_X)$ (and extended discussion on this baseline can be found in Appendix D.1). Therefore, none of the models are able to outperform such a baseline.

Retrieval-Augmented Generation (RAG) [31]. Retrieval-augmented generation (RAG) is a technique in which an LLM is combined with an external retriever [31]. In this case, the model does not have to rely only on its internal knowledge, but relevant documents can be fetched from a collection of texts. To investigate the impact of RAG on our benchmark, we used the GPT 4.1 model with web access (the model can search the web for each query). To test how RAG affects performance, we focused on 11 tasks on which the GPT 4.1 model without web-based RAG had a score of 0 (corresponding to task indices 40, 43, 57, 62, 63, 64, 65, 66, 67, 70, and 71 in Tab. 2). Here, the expectation is that RAG would help the model performance, and would result in a higher score. However, the results for the GPT 4.1 model with web-based RAG still showed a zero score for each task. Therefore, perhaps surprisingly, for these tasks RAG did not result in improved performance. This finding warrants further investigation of how RAG may improve the performance of models on our benchmark. This investigation is left for future work.

D Likelihood vs. Question-Answering Prompts

In this appendix, we discuss an alternative way of eliciting model distributions $\tilde{P}(V_Y | V_X)$. We recall Ex. 1 from the main text, in which we were interested in marijuana usage (V_Y) over age groups (V_X). For this task, we used a prompt template

$$\pi(v_X) = \text{“Has a person aged } \{v_X\} \text{ ever used marijuana?”}, \quad (9)$$

which for different values of age v_X creates different questions posed to the model. With the question, the model is also given the choice of answers yes/no, which are labeled with letters A, B. For extracting the model’s probability $\tilde{P}(V_Y = 1 | V_X = v_X)$, we investigate the probabilities associated with letters A, B in the model’s next-token prediction distribution $\tilde{P}^{\text{nt}}$.

However, this approach, which we refer to as *question/answer-prompting* (QA-prompting, for short), is not the only way to elicit the model’s distribution $\tilde{P}(V_Y | V_X)$. An alternative way of eliciting this distribution is to use a different prompting strategy, which asks for the probability directly:

$$\pi^{\text{prob}}(v_X) = \text{“What is the probability that a person aged } \{v_X\} \text{ ever used marijuana?”} \quad (10)$$

This strategy, which we refer to as *likelihood-prompting*, is investigated in this appendix. To elicit a probability from the model, a range of responses is offered, namely responses in the following form:

- A. 0%
- B. 0%-5%
- C. 5%-10%
- ...
- U. 95%-100%
- V. 100%

We then again inspect the next-token prediction probabilities in $\tilde{P}^{\text{nt}}$, and choose the response with the highest probability:

$$\mathcal{A}^{\text{resp}} = \arg \max_{\mathcal{A} \in \{A, \dots, V\}} \tilde{P}^{\text{nt}}(\text{answer } \mathcal{A}). \quad (11)$$

The chosen response letter $\mathcal{A}^{\text{resp}}$ is then mapped to either an interval, or a fixed value of 0% or 100%. For assigning the final predicted probability, for intervals, we take the midpoint. For instance, if the answer B. 0%-5% is chosen, we set the predicted probability to 0.025, and so on. In this way, we elicit the probability $\tilde{P}(V_Y = 1 | V_X = v_X)$ of the model.

Model performances on the benchmark when using likelihood-prompting, on the 94 tasks in the high-dimensional setting with a binary one-dimensional V_Y , and $|V_X| \in \{2, 3, 4, 5\}$, are shown in Fig. 7. On top of the models used in the main text, we also include a version of DeepSeek R1 [21] distilled using Qwen2.5 32B [25]. The results in the figure can be compared with the performance

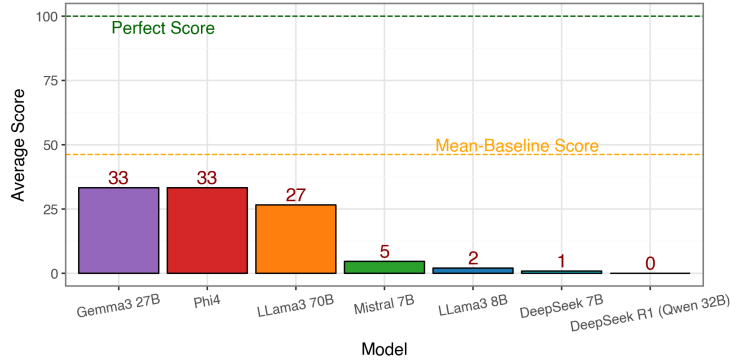


Figure 7: Models’ performance with likelihood-prompting on the 94 high-dimensional tasks.

of models on the same tasks via QA-prompting used in the main text. We note that the model performances, for some models, improve with likelihood-prompting (in particular, the Phi4 model

moves from an average score of 17/100 to 33/100, Gemma3 27B from 13/100 to 33/100, LLama3 70B from 7/100 to 27/100). At the same time, for some models, the performance with likelihood prompting becomes worse (LLama3 8B). However, overall, model performance is still quite poor, and we further contextualize this in the sequel.

D.1 Baselines – how good are the models?

In the evaluation of our benchmark, we used two *baselines*, which corresponded to the notion of zero probabilistic knowledge. The first baseline was the uniform baseline P^{unif} , which is in the binary case given by:

$$P^{\text{unif}}(V_Y = 1 \mid V_X = v_X) = \frac{1}{2} \forall v_X. \quad (12)$$

This baseline corresponds to random guessing. In the main text, we also discussed the 0/1 baseline, which outputs either a probability 0 for events with a marginal probability ≤ 0.5 , or a probability 1 for events with a marginal probability > 0.5 :

$$P^{0/1}(V_Y = 1 \mid V_X = v_X) = \mathbb{1}(\mathbb{E}_P[V_Y] > 0.5) \forall v_X, \quad (13)$$

where $\mathbb{E}_P[V_Y]$ is the true mean of V_Y (in the ground truth distribution P). In our scoring scheme, these baselines were used to determine the score $S = 0$ for a given task (see Sec. 2.2.1). However, another interesting baseline is the *mean-baseline*, defined by:

$$P^{\text{mean}}(V_Y = 1 \mid V_X = v_X) = \mathbb{E}_P[V_Y] \forall v_X. \quad (14)$$

This baseline outputs a fixed value (the marginal mean of the outcome V_Y) for each $v_X \in \text{dom}(V_X)$. In words, this baseline is aware of the marginal probability of the event V_Y , but has no knowledge about the *variation* of $P(V_Y = 1 \mid V_X = v_X)$ according to v_X around the marginal mean $\mathbb{E}_P[V_Y]$. This baseline corresponds to some probabilistic knowledge (marginal mean), but without any knowledge over how the probability varies across population subsets. Assigning a score of $S = 0$ to this baseline would make high scores on our benchmark even more difficult to achieve, and an argument could be made for including it in future iterations of the benchmark. In Fig. 7 the performance of the mean-baseline with respect to the current evaluation is shown (orange horizontal line), indicating an average score of 46/100. This means that the inclusion of this baseline would effectively render the scores of all models investigated in this manuscript close to 0. This observation once again confirms that the current generation of LLMs does not possess knowledge about observational distributions in the real world, regardless of the strategy (QA- or likelihood-prompting) we choose for eliciting the model distributions $\tilde{P}(V_Y \mid V_X)$.