APRIL: Annotations for Policy evaluation with Reliable Inference from LLMs

Aishwarya Mandyam AM2@STANFORD.EDU
Stanford University

Kalyani Limaye LIMAYK@STANFORD.EDU
Stanford University

Barbara E. Engelhardt* BARBARAE@STANFORD.EDU
Stanford University, The Gladstone Institutes

Emily Alsentzer* EALSENTZER@STANFORD.EDU
Stanford University

Abstract

Off-policy evaluation (OPE) estimates the value of a contextual bandit policy prior to deployment. As such, OPE plays a critical role in ensuring safety in high-stakes domains such as healthcare. However, standard OPE approaches are limited by the size and coverage of the behavior dataset. While previous work has explored using expert-labeled counterfactual annotations to enhance dataset coverage, obtaining such annotations is expensive, limiting the scalability of prior approaches. We propose leveraging large language models (LLMs) to generate counterfactual annotations for OPE in medical domains. Our method uses domain knowledge to guide LLMs in predicting how key clinical features evolve under alternate treatments. These predicted features can then be transformed using known reward functions to create counterfactual annotations. We first evaluate the ability of several LLMs to predict clinical features across two patient subsets in MIMIC-IV, finding that state-of-the-art LLMs achieve comparable performance. Building on this capacity to predict clinical features, we generate LLM-based counterfactual annotations and incorporate them into an OPE estimator. Our empirical results analyze the benefits of counterfactual annotations under varying degrees of shift between the behavior and target policies. We find that in most cases, the LLM-based counterfactual annotations significantly improve OPE estimates up to a point. We provide an entropy-based metric to identify when additional annotations cease to be useful. Our results demonstrate that LLM-based counterfactual annotations offer a scalable approach for addressing coverage limitations in healthcare datasets, enabling safer deployment of decision-making policies in clinical settings.

Keywords: off-policy evaluation, synthetic datasets, contextual bandits

Data and Code Availability This paper uses data from the broadly available MIMIC-IV dataset. Code is available on Github.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Off-policy evaluation (OPE) methods estimate the value of a new (target) contextual bandit policy using a behavior dataset of samples collected under a distinct behavior policy (Sutton and Barto, 2018). OPE can be particularly useful in high-stakes domains such as healthcare, where evaluating policies by directly deploying them is either impossible or unethical. Standard approaches to OPE include importance sampling (Precup et al., 2000), the direct method (Beygelzimer and Langford, 2009), and doubly robust approaches (Dudik et al., 2014). However, the performance of OPE estimators is inherently limited by the coverage of the behavior dataset. When the target policy takes actions that are underobserved in the behavior dataset, standard OPE methods cannot reliably estimate the value of these actions, leading to inaccurate policy value estimates.

To address this, recent work proposes augmenting the behavior dataset with counterfactual annotations (Tang and Wiens, 2023). A counterfactual annotation is a prediction of the scalar reward resulting from an action unobserved in the behavior dataset. For example, if a patient received 20mEq of potassium, a counterfactual annotation would predict the reward had the patient instead received 40mEq. Two strategies have been developed to incorporate such annotations into OPE: one augments an importance sampling–based estimator (Tang and Wiens, 2023), and the other augments a doubly robust estimator (Mandyam et al., 2024). Both demonstrate that incorporating counterfactual annotations can improve OPE estimates, but these approaches rely on human experts (e.g., clinicians) to provide the annotations, which is costly and difficult to scale.

To address this, we propose a pipeline to source counterfactual annotations for OPE in clinical settings using large language models (LLMs). LLMs have the ability to reason effectively about medical domains, with the capacity to answer medical questions (Singhal et al., 2023b), perform differential patient diagnoses (Nori et al., 2025), and reason about medical images (Zhou et al., 2025). Our approach leverages LLMs to predict clinical features of interest such as downstream laboratory measurements; we then incorporate these predictions into known reward functions to produce synthetically generated counterfactual annotations.

We evaluate our proposed framework on two clinical tasks: intravenous (IV) potassium and sodium repletion. Both are critical procedures in clinical practice, where large errors in administration can lead to adverse outcomes (Voldby and Brandstrup, 2016). Furthermore, these are routine procedures with well-established guidelines for treatment and reasonably predictable treatment response curves, making them especially tractable settings for applying contextual bandit algorithms. We construct corresponding patient datasets from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which contains electronic health records (EHR) for patients admitted to the Beth Israel Deaconess Medical Center (Johnson et al., 2024, 2023; Goldberger et al., 2000). We first assess the ability of several LLMs to predict relevant clinical features, including serum potassium and sodium values. Using clinically motivated reward functions, we then transform these predictions into counterfactual annotations. Our results show that LLM-generated counterfactual annotations improve OPE estimates, particularly under large distribution shifts between the behavior and target policies.

Our contributions follow:

- **We perform OPE with LLM-generated counterfactual annotations in a multi-cohort setting** using MIMIC-IV. We systematically evaluate multiple general-purpose LLMs for their accuracy in predicting downstream clinical features.

- **We show that incorporating LLM-generated annotations can significantly improve OPE estimates**, reducing RMSE relative to baselines and confirming prior findings in real-world data.

- **We demonstrate that additional counterfactual annotations offer diminishing returns**, a phenomenon captured quantitatively via the marginal entropy over the action distribution.

2. Preliminaries

2.1. Problem setting

We adopt a contextual bandit setting, as potassium and sodium repletion are short-horizon decisions whose outcomes can be observed within a single timestep. A contextual bandit setting is represented as $(\mathcal{S}, \mathcal{A}, \mathcal{R}, d_0)$, where $\mathcal{S}$ is the discrete context space, $\mathcal{A}$ is the discrete action space, $\mathcal{R}$ is the reward distribution, and d_0 is the initial context distribution. The reward function $R : S \times A \to [0, 1]$ assigns a scalar reward between 0 and 1. Our goal is to evaluate a target contextual bandit policy π_e by estimating its value $v(\pi_e) = \mathbb{E}_{s\sim d_0, a\sim \pi_e,)}\left[R(s, a)\right]$ using a behavior dataset. The behavior dataset consists of samples $D = \{s_i, a_i, r_i\}_{i=1}^{N}$, where the actions are sampled from a behavior policy π_b.

2.2. Off-policy evaluation estimators

Many OPE estimators fall into three broad categories: importance sampling (IS), the direct method (DM), and doubly robust (DR) estimators. IS methods (Precup et al., 2000) re-weigh each sample in the behavior dataset using an inverse propensity score (IPS) $\frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}$. The second class includes direct-method (DM) approaches (Beygelzimer and Langford, 2009), which learn a reward model $\hat{R}$ from the behavior dataset, and use the model to simulate the returns of samples from the target policy. The final category includes doubly-robust (DR) approaches (Dudik et al., 2014; Jiang and Li, 2016), which combine strategies from IS and DM approaches, providing favorable theoretical guarantees when either the IPS ratio is known or the reward model is of high quality.

Recent work has proposed supplementing the behavior dataset with counterfactual annotations solicited from an expert. Tang and Wiens (2023) introduce an IS-based estimator and demonstrate that counterfactual annotations can improve OPE estimates when the annotations are of high quality. Mandyam et al. (2024) extends this to a doubly robust setting, mitigating the negative impacts of noisy or imperfect annotations. Both approaches assume that counterfactual annotations are expert-labeled, which limits the scalability of the proposed approaches. Other work has proposed using a variational auto-encoder to generate synthetic trajectories, thus enriching state–action coverage of the behavior dataset and tightening variance bounds (Gao et al., 2024). Our work builds on these approaches, identifying a scalable alternative to creating counterfactual annotations.

2.3. Generative models can encode medical knowledge

LLMs have shown impressive general medical reasoning capabilities. Models fine-tuned on web and biomedical corpora now match or surpass physicians on multiple-choice benchmarks such as MedQA (Jin et al., 2020). DeepMind's Med-PaLM 2 (Singhal et al., 2023a) and Gemini models (Saab et al., 2024) illustrate that scaling and instruction tuning can boost performance across a range of clinical knowledge tasks. However, these works center on general medical knowledge questions rather than reasoning about individual patient clinical trajectories, which is the focus of our work.

More granular, patient-specific LLM applications are beginning to emerge, including reasoning about how laboratory values evolve over a patient trajectory. Bhasuran et al. (2025) explore differential-diagnosis generation from brief clinical vignettes, highlighting the importance of structured patient summaries to improve LLM outputs. He et al. (2024) evaluate the ability to generate accurate and safe responses to patient lab-result inquiries using prompt engineering and detailed quality evaluation metrics. These studies suggest that LLMs can reason about patients when provided with curated input and task framing (Wei et al., 2022; Chung et al., 2022). Our work leverages prompting strategies that build on those used in these works to guide LLMs in generating patient-specific counterfactual annotations.

2.4. Synthetic data for machine learning

In our setting, supervision comes from both a real-world dataset and a noisier set of synthetic data. Using a noisy secondary dataset is a common paradigm in supervised learning, and methods to mitigate the covariance shift between the datasets have been extensively studied in the *robust machine learning* literature. Earlier results introduced transfer learning techniques to learn features from secondary datasets while mitigating issues with higher-variance samples (Krizhevsky et al., 2012; Pan and Yang, 2010; Ben-David et al., 2006; Sugiyama et al., 2007; Quiñonero-Candela et al., 2008). Other methods such as prediction-powered inference (Angelopoulos et al., 2023) explicitly correct for possible biases that result from the introduction of synthetic samples.

3. Methods

3.1. Dataset

We conduct our analysis using the MIMIC-IV dataset, partitioned into two subsets of non-ICU patients. The first subset includes all patients who received IV potassium, and the second includes all patients who received IV hypertonic (3%) saline. For each patient in the potassium and sodium subsets, we represent the clinical context as a feature vector comprising 15 variables that characterize the patient's state four hours prior to treatment (i.e., administration of potassium or saline). We focus on a four-hour window because this corresponds to the highest frequency of electrolyte administration observed in our dataset, with patients receiving electrolytes at most once every four hours. The features used to represent the clinical context include laboratory results, vital signs, administered medications, and static covariates such as age and gender. A complete list of features is provided in Appendix A. The action space corresponds to the administered dosage, represented in milliequivalents (mEq). For potassium administration, the dosage action space is $A = \{0, 10, 20, 40\}$. For sodium (i.e., hypertonic saline) administration,

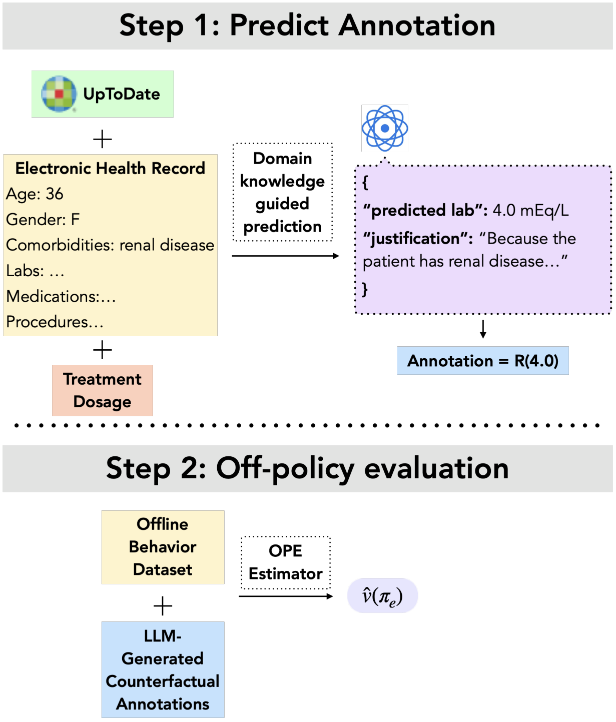

Step 1: Predict Annotation
UpToDate
+
Electronic Health Record
Age: 36
Gender: F
Comorbidities: renal disease
Labs: ...
Medications:...
Procedures...
+
Treatment Dosage
Domain knowledge guided prediction
{
"predicted lab": 4.0 mEq/L
"justification": "Because the patient has renal disease..."
}
Annotation = R(4.0)
Step 2: Off-policy evaluation
Offline Behavior Dataset
+
LLM-Generated Counterfactual Annotations
OPE Estimator
$\hat{v}(\pi_e)$

Figure 1: **Our work improves OPE estimates using LLM-generated counterfactual annotations.** We first query counterfactual annotations using domain knowledge guided prediction. We calculate the annotations using a known reward function R. Finally, we incorporate the counterfactual annotations and offline behavior dataset to learn an OPE estimate $\hat{v}(\pi_e)$.

dosages are discretized to accommodate our assumption of a discrete action space, yielding an action space $A = \{0, 100, 200, 300, 400, 500\}$.

Similar to prior work (Prasad, 2020), we adopt a reward function defined as a function of the clinical context observed following the administration of a treatment dosage. Specifically, the scalar reward depends on a single laboratory measurement in the next observed context. For patients who receive IV potassium, this is a serum potassium lab, and for patients who receive IV hypertonic saline, this is a serum sodium lab. The reward function

$$R(x) = \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{x-a}{2.5}\right)^2\right), & x < a \\ 1, & a \leq x \leq b \\ \exp\left(-\frac{1}{2}\left(\frac{x-b}{2.5}\right)^2\right), & x > b \end{cases}$$

takes as input a laboratory value x, and uses the lower bound (a) and upper bound (b) of the reference range to calculate a scalar reward. This reward function design reflects the clinical goal of repletion which is to bring a patient's electrolyte level into the normal range and keep it there, while smoothly penalizing deviations outside the range. Visual representations of the reward function can be seen in Appendix Figure 4.

3.2. Generating Counterfactual Annotations using LLMs

As described in Section 3.1, the reward functions for each decision-making task are functions of a single lab value. Therefore, generating a counterfactual annotation requires predicting the specified lab value under a counterfactual treatment dosage.

We construct prompts that include information about the patient's clinical state, a paragraph that cites the most relevant features for lab value prediction sourced from UpToDate (Kluwer, n.d.), and a query about the lab value had an alternative treatment dosage been administered (example in Appendix B). Building on prior work (Hegselmann et al., 2025), we organize the patient's information into categories such as comorbidities, laboratory results, and medications to create a structured text representation of the clinical state. Including features from UpToDate guides the LLM toward clinically relevant information, as EHRs often contain extraneous data that may not be predictive. To ensure structured outputs, we restrict the LLM's response to a JSON object containing two keys: the predicted lab value, and a justification for the prediction. This format allows for straightforward extraction of the numerical lab value and facilitates verification of the LLM's reasoning.

For potassium administration, dosages are assumed to be delivered at a rate of 10, mEq/hr, and for sodium administration, at a rate of 30, mEq/hr, corresponding to the most common rates observed in MIMIC-IV. The prompt also specifies that the lab value should be predicted three hours after the IV infusion concludes; this corresponds to the average number of hours that the lab value post treatment administration was measured. Once the LLM predicts the lab value, it is converted into a scalar counterfactual annotation using the corresponding known reward function.

3.3. Incorporating Counterfactual Annotations into an OPE estimator

Once we generate counterfactual annotations, we must incorporate them into an OPE estimator. Prior methods for OPE with counterfactual annotations often assume that the IPS ratios are fully known. However, in this work, we must infer both π_b and π_e from finite sample sizes. To mitigate possible biases as a result of unknown IPS ratios, we choose to use a direct method estimator. The standard direct method estimator is

$$\hat{V}^{DM} = \sum_{s \in S} d_0(s) \sum_{a \in A} \pi(a|s)\hat{R}(s,a),$$

where $\hat{R}$ is a reward function estimate learned from the behavior dataset. When we have access to both a behavior dataset and counterfactual annotations, we choose to use modified version of the standard DM estimator suggested by prior work work (Mandyam et al., 2024),

$$\hat{V}^{DM^+} = \sum_{s \in S} d_0(s) \sum_{a \in A} \pi(a|s)\hat{R}^+(s,a),$$

where $\hat{R}^+$ is learned using both the behavior dataset and counterfactual annotations. In this work, we approximate both $\hat{R}$ and $\hat{R}^+$ using linear regression.

3.4. Evaluation Setup

A standard metric for assessing the accuracy of an OPE estimator is the root mean squared error (RMSE), defined as

$$\text{RMSE} = \sqrt{\mathbb{E}[(\hat{v}(\pi_e) - v(\pi_e))^2]},$$

where $\hat{v}(\pi_e)$ denotes the value estimated by the OPE method, and $v(\pi_e)$ is the true value of the target policy π_e. In practice, $v(\pi_e)$ is rarely available, which complicates the evaluation of OPE estimators in real-world settings.

To address this limitation in the MIMIC-IV dataset, we adopt a controlled evaluation strategy. We partition each dataset subset into disjoint behavior and target sub-cohorts, and infer corresponding policies via behavior cloning. Because the target sub-cohort contains observed rewards, we approximate the value of the cloned target policy by averaging these rewards. The fidelity of this approximation depends on how well the policies are cloned; to assess this, we evaluate the cloned policies' accuracy on a held-out validation set and find that they perform well in reproducing the observed treatment decisions (e.g., validation accuracy ¿ 90%). This gives us confidence that the averaged rewards in the target subset provide a reliable reference value against which to compute RMSE for different OPE estimators.

It is well known that the performance of OPE estimators depends considerably on the distribution shift between the behavior dataset and the samples induced by the target policy. To systematically study the effect of LLM-generated counterfactual annotations on an OPE method under varying degrees of distribution shift, we construct three behavior–target dataset pairs for each subset of patients from MIMIC-IV. The first pair splits by gender, with female patients forming the behavior dataset and male patients the target dataset. The second pair splits by comorbidity status: for potassium repletion, patients without renal disease form the behavior dataset and patients with renal disease the target dataset; for sodium repletion, the split is based on cirrhosis. We choose these comorbidities because their presence is likely to influence the patient's response to drug administration. The third pair separates patients by drug dosage, using low-dosage patients as the behavior dataset and high-dosage patients as the target dataset. These partitions are designed to reflect clinically meaningful subgroups while also inducing progressively larger divergences between the behavior and target policies. This allows us to evaluate when counterfactual annotations generated by LLMs yield improvements in OPE accuracy.

4. Experiments

Our empirical analyses seek to answer the following questions: **(1)** Can LLMs accurately predict downstream patient laboratory values after a treatment is administered? **(2)** Under what conditions do LLM-generated counterfactual annotations improve OPE estimates? **(3)** How do OPE estimates vary as the number of synthetic counterfactual annotations increases?

To address these questions, we use five LLMs spanning a range of parameter counts: OpenAI's `o1` (OpenAI et al., 2024), `o3-mini` (Zhang et al.), and `gpt-4o-mini` (Hurst et al., 2024), Google's `Gemini` 1.5 (Reid et al., 2024), and Anthropic's `Claude 3.7 Sonnet` (Cla). All models are hosted on an internal, sandboxed cluster to ensure HIPAA compliance with the MIMIC-IV dataset. All experiments were conducted with a temperature setting of zero whenever supported. For `o1` and `o3-mini`, which do not expose a temperature parameter, we use the default configuration.

4.1. LLMs can predict downstream lab values on real patient populations

We first evaluate whether LLMs can accurately predict serum potassium and serum sodium laboratory values. In realistic deployment, the target patient population may not be directly accessible, so we assess predictive performance using behavior datasets from each sub-cohort split. To generate predictions, we prompt the LLM following the procedure in Section 3.2, but instead of asking for counterfactual lab values, we request the lab value following the dosage administered in MIMIC-IV. Because the corresponding ground-truth lab values are observed in MIMIC-IV, we can directly quantify predictive accuracy. We evaluate accuracy using a weighted F1 score across clinically relevant categories of lab values (e.g., below reference range, within reference range, above reference range). The categories used to calculate the F1 score, and further details are reported in Section 3.1.

We find that LLMs can predict serum potassium and serum sodium lab values with clinically meaningful degrees of accuracy (Table 1, visualized in Appendix Figure 6). First, we note that serum potassium lab values are predicted more accurately than serum sodium lab values, likely due to the wider distribution and higher prevalence of outliers in sodium lab measurements. We also find that the performance of a given LLM remains consistent across cohorts within each prediction task, which suggests that predictive accuracy does not strongly depend on the underlying patient population. Finally, the differences in predictive accuracy across LLMs are modest, suggesting that multiple models are capable of producing reliable predictions of downstream lab values. Furthermore, these results demonstrate our proposed framework's ability to produce counterfactual annotations of reasonable quality. In particular, because LLMs can predict downstream lab values within a degree of accuracy that is clinically relevant, the resulting annotations are likely to be useful for OPE.

4.2. LLM-produced counterfactual annotations can improve OPE estimates

We next evaluate the utility of LLM-generated counterfactual annotations for OPE, following the setup in Section 3.4. We report results for the potassium repletion task in Figure 2, and for the sodium repletion task in Appendix Figure 7. Our results show that counterfactual annotations substantially improve OPE estimates in settings with large distribution shifts between the actions observed in the behavior and target policies. Across both the potassium and sodium repletion tasks, the reported RMSE reflects the relative difficulty of estimating $v(\pi_e)$ under each cohort split. For example, in the gender cohort split, where behavior and target policies are nearly identical, the RMSE is already near zero without counterfactual annotations, leaving little room for improvement. In contrast, in the dosage cohort split, where behavior and target policies have little overlap, the baseline RMSE of DM is highest, reflecting the difficulty of the task. Here, the incorporation of counterfactual annotations produces the largest reductions in RMSE, indicating that annotations are most valuable when behavior and target policies diverge strongly. Specifically, in the potassium dosage cohort, counterfactual annotations can reduce RMSE by 83%, and in the sodium dosage cohort by 49%.

We also find that the performance of DM^+ varies with the choice of LLM used to generate counterfactual annotations. In the potassium repletion task, annotations from `o1` yield the best performance as shown by lowest RMSE, whereas in the sodium repletion task, annotations from `gpt-4o-mini` and `o3-mini` yield the best performance. Although the best-performing LLM is not consistent across tasks or cohort splits, counterfactual annotations consistently

Task	Cohort	o1	gpt-4o-mini	o3-mini	Gemini	Claude 3.7
Potassium Repletion	Gender	0.856	0.809	0.854	0.858	**0.866**
	Comorbidity	0.871	0.787	0.869	0.872	**0.879**
	Dosage	0.878	0.791	0.876	0.879	**0.885**
Sodium Repletion	Gender	0.758	0.774	0.738	0.749	**0.776**
	Comorbidity	0.771	**0.801**	0.753	0.779	0.796
	Dosage	0.772	**0.809**	0.768	0.780	0.804

Table 1: **All LLMs perform comparably across potassium and sodium lab prediction.** Predictions are evaluated using weighted F1 scores across clinically relevant lab value categories. The best performing LLM within each cohort is in bold.

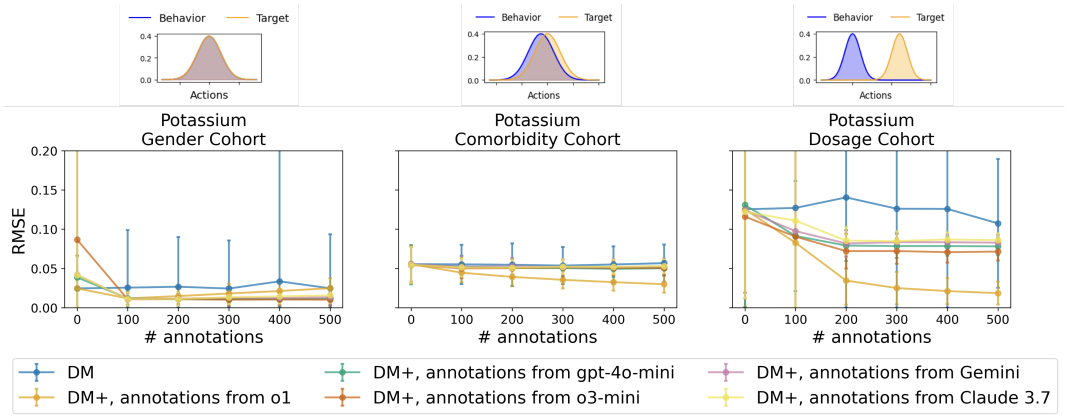


Figure 2: **LLM-generated counterfactual annotations improve OPE estimates in settings with high divergence between actions observed in behavior and target policies.** We report results for the potassium repletion task. Our baseline is a direct method estimator (blue) that does not use counterfactual annotations. The performance of DM^+ with annotations from each LLM is reported in the corresponding colors. Error bars represent standard error across 500 bootstrapped datasets sampled with replacement. Since RMSE is non-negative, the lower bound of the error bars is truncated at 0 where necessary. Figures above each plot demonstrate the difference in distribution of actions observed in the behavior and target policies.

reduce RMSE in the most challenging settings (e.g., dosage cohorts), regardless of the LLM used.

Finally, to assess statistical significance, we compare DM and DM^+ using a paired t-test, with DM^+ learned using 500 counterfactual annotations (Section 4.2). In nearly all settings, DM^+ achieves significantly lower RMSE than DM. The main exception is the gender split in both potassium and sodium tasks, where annotations from some LLMs do not yield a meaningful performance improvement. This outcome is expected, given the substantial overlap between behavior and target policies in the gender cohorts, which allows DM to perform well even without counterfactual annotations.

4.3. Additional counterfactual annotations yield marginal improvements in OPE estimates

A key consideration when using synthetic data in machine learning is determining the point at which adding further synthetic samples no longer provides benefits. In our setting, a single source of counterfactual annotations can generate at most $N \cdot (|\mathcal{A}| - 1)$ unique annotations, where $|\mathcal{A}|$ is the number of actions and N is the number of samples in the behavior dataset. When multiple sources are available, each source provides separate predictions for unobserved actions, which can either be combined or averaged. Direct combination increases the total number of an-

Task	Cohort	o1	gpt-4o-mini	o3-mini	Gemini	Claude 3.7
Potassium Repletion	Gender	5.4E-31	9.8E-03	3.0E-04	3.7E-01	2.4E-01
	Comorbidity	1.5E-94	1.3E-08	1.1E-07	7.7E-03	5.7E-04
	Dosage	1.3E-83	1.7E-14	1.9E-20	7.2E-11	1.4E-08
Sodium Repletion	Gender	1.9E-03	7.0E-04	1.5E-19	7.5E-14	6.6E-08
	Comorbidity	1.0E-07	2.0E-16	2.0E-03	1.0E-04	2.2E-05
	Dosage	2.8E-37	3.0E-15	3.2E-57	1.44E-44	7.79E-39

Table 2: **In most cohorts across both tasks, LLM-generated annotations significantly improve RMSE.** We compare RMSE distributions for DM and DM^+ with 500 counterfactual annotations using a paired t-test, and report p-values. P-values shown in red indicate results that are not statistically significant ($p \geq 0.05$) or cases where RMSE does not improve relative to DM ($t < 0$).

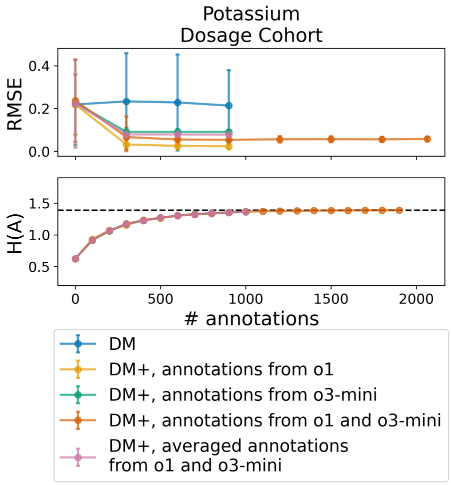


Figure 3: **Combining annotation sources yields limited returns.** (Top) We compare DM to DM^+ with annotations from the best-performing LLMs for potassium repletion in the dosage cohort, with two aggregation methods: pooling predictions and averaging annotations. Error bars show standard error over 500 bootstrapped datasets, truncated at 0. (Bottom) Marginal entropy over the action space $H(A)$ when adding counterfactual annotations to the behavior dataset for the potassium cohort. The dashed line marks the maximum possible entropy.

notations, whereas averaging maintains the same total count. We study both strategies for the potassium task (Figure 3) and sodium task (Figure 9).

We focus on the dosage cohort splits for both tasks, where counterfactual annotations have the greatest impact in reducing RMSE due to minimal overlap between the behavior and target policies. Specifically, we examine combinations of the two LLMs whose counterfactual annotations yield the best performance for DM^+: `o1` and `o3-mini` for the potassium task, and `Gemini` and `o3-mini` for the sodium task. We find that, while adding counterfactual annotations initially reduces OPE error, the improvement quickly plateaus as more annotations are included. Averaging multiple sources does not provide additional gains beyond the best-performing single source. For instance, in the potassium task, averaging annotations from `o1` and `o3-mini` yields OPE performance worse than using `o1` alone, though slightly better than `o3-mini` alone. Similarly, combining annotations without averaging, which nearly doubles the number of annotations, does not improve OPE estimates relative to a single source. These results indicate that substantially increasing the number of counterfactual annotations provides limited utility.

To quantify the effect of additional annotations, we compute the marginal entropy over the action distribution. Entropy measures the overall uncertainty or spread of actions in the dataset. Formally, the marginal entropy over the action distribution is $H(A) = -\sum_{a \in A} \hat{p}(a)\ln(\hat{p}(a))$ where $\hat{p}(a)$ is the probability of observing a given action a, estimated empirically. The maximum entropy occurs when all actions are equally frequent, in which case $H(A) = ln(|\mathcal{A}|)$. We observe that, as the number of annotations increases, the action coverage approaches the maximum entropy, and further annotations yield only marginal gains. In particular, at around 700 annotations for the potassium task and 500 annotations for the sodium task, OPE improvements have largely plateaued, and the marginal action entropy is already near its maximum, indicating that additional counterfactual annotations provide little further utility. This analysis suggests that marginal entropy over the action space is a proxy that may be used to determine when to stop generating counterfactual annotations.

5. Discussion

In this work, we present a scalable strategy for generating counterfactual annotations for OPE in clinical settings. We show that LLMs can reason over clinical contexts and predict downstream lab values, which in turn can be used to construct counterfactual annotations. Focusing on the potassium and sodium repletion tasks, we demonstrate that this approach leads to substantial improvements in OPE estimates, particularly when there is considerable divergence between the behavior and target policies. We recommend using LLM-generated annotations when there are known coverage gaps in the behavior dataset, and relying on an entropy-based metric to decide when additional counterfactual annotations are needed.

Limitations and Future Work. Our study is limited to reward functions that consider a single clinical feature. While our results provide evidence that LLMs can reliably predict these downstream lab values, future work should evaluate whether similar gains can be achieved for predicting more complex clinical outcomes.

Acknowledgments

AM was funded in part by a Stanford Data Science Fellowship. BEE was funded in part by grants from the Parker Institute for Cancer Immunology (PICI), the Chan-Zuckerberg Institute (CZI), the Biswas Family Foundation, NIH NHGRI R01 HG012967, and NIH NHGRI R01 HG013736. BEE is a CIFAR Fellow in the Multiscale Human Program. The authors would like to thank Dr. Chloe Stanwyck for support with designing and evaluating empirical procedures.

References

Claude 3.7 sonnet system card. URL https://api.semanticscholar.org/CorpusID:276612236.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference, 2023. URL https://arxiv.org/abs/2301.09633.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. volume 19, pages 137–144, 01 2006.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138, 2009.

Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine*, 8(1):166, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Xin Huang, Andrew Dai, Xuezhi Wang, Brian Lester, and et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1097–1105, 2014.

Ge Gao, Qitong Gao, Xi Yang, Song Ju, Miroslav Pajic, and Min Chi. On trajectory augmentations for off-policy evaluation. In *The Twelfth International Conference on Learning Representations*, 2024.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. ISSN 0009-7322. Online.

Zhe He, Balu Bhasuran, Qiao Jin, Shubo Tian, Karim Hanna, Cindy Shavor, Lisbeth Garcia Arguello, Patrick Murray, and Zhiyong Lu. Quality of answers of generative large language models vs peer patients for interpreting lab test results for lay patients: Evaluation study. *ArXiv*, pages arXiv–2402, 2024.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders, 2025. URL https://arxiv.org/abs/2502.17403.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alexandre Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, An drey Mishchenko, Angela Baek, Angela Jiang, An toine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, B. Ghorbani, Ben Leimberger, Ben Rossen, Benjamin Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Chris Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mély, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Phong Duc Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Hai-Biao Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Hee woo Jun, Hendrik Kirchner, Henrique Pondé de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, İbrahim Cihangir Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub W. Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Ryan Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quiñonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Joshua Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Ouyang Long, Louis Feuvrier, Lu Zhang, Lukasz Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Made laine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Ma teusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Ali Yatbaz, Mengxue Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Castro, Mikhail Pavlov, Miles Brundage, Miles

Wang, Mina Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Na talie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nikolas A. Tezak, Niko Felix, Nithanth Kudige, Nitish Shirish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Phil Tillet, Prafulla Dhariwal, Qim ing Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Raphael Gontijo Lopes, Raul Puri, Reah Miyara, Reimar H. Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Ramilevich Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *ArXiv*, abs/2410.21276, 2024. URL https://api.semanticscholar.org/CorpusID:273662196.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

Di Jin, Yansong Pan, Wenqiang Ouyang, and Caiming Xiong. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Alistair E. W. Johnson, Luca Bulgarelli, Li-wei Shen, and et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Alistair E. W. Johnson, Luca Bulgarelli, Tom J. Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo A. Celi, and Roger Mark. MIMIC-IV (version 3.1). https://physionet.org/content/mimiciv/3.1/, 2024. PhysioNet. https://doi.org/10.13026/kpb9-mt58.

Wolters Kluwer. Uptodate, n.d. URL https://www.uptodate.com.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. URL https://api.semanticscholar.org/CorpusID:195908774.

Aishwarya Mandyam, Shengpu Tang, Jiayu Yao, Jenna Wiens, and Barbara E. Engelhardt. Candor: Counterfactual annotated doubly robust off-policy evaluation, 2024. URL https://arxiv.org/abs/2412.08052.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela

Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Niranjani Prasad. *Methods for reinforcement learning in clinical decision support*. PhD thesis, Princeton University, 2020.

Doina Precup, Richard Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 06 2000.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 12 2008. ISBN 9780262255103. doi: 10.7551/mitpress/9780262170055.001.0001. URL https://doi.org/10.7551/mitpress/9780262170055.001.0001.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross Mcilroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Os car Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Au rko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo-Yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xi ance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Roșca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajku mar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, El-

speth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxi aoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mah moud Alnahlawi, Christo pher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Livio Baldini Soares, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Ha roon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clément Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Re beca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiří Šimša, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Lucas Dixon, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Nicholas FitzGerald, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Re-

pina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Ilia Shumailov, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Kather ine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Junwen Bai, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, Oriol Vinyals, and Alexandra Chronopoulou. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024. URL https://api.semanticscholar.org/CorpusID:268297180.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL https://arxiv.org/abs/2404.18416.

Karan Singhal, Shekoofeh Azizi, Tien-Ju Tu, and et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023b. URL https://arxiv.org/abs/2305.09617.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, dec 2007. ISSN 1532-4435.

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

Shengpu Tang and Jenna Wiens. Counterfactual-augmented importance sampling for semi-offline policy evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dsH244r9fA.

Anders Voldby and Birgitte Brandstrup. Fluid therapy in the perioperative setting-a clinical review. *Journal of Intensive Care*, 4, 12 2016. doi: 10.1186/s40560-016-0154-3.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Brian Zhang, Eric Mitchell, Hongyu Ren, Kevin Lu, Max Schwarzer, Michelle Pokrass, Shengjia Zhao, Ted Sanders, Adam Tauman Kalai, Alexandre Passos, Benjamin Sokolowsky, Elaine Ya Le, Erik Ritter, Hao Sheng, Hanson Wang, Ilya Kostrikov, James Lee, Johannes Ferstad, Michael Lampe, Prashanth Radhakrishnan, Sean Fitzgerald, Sébastien Bubeck, Yann Dubois, Yu Bai, Andy Applebaum, Elizabeth Proehl, Evan Mays, Joel Parish, Kevin Liu, Leon Maksin, Leyton Ho, Miles Wang, Michele Wang, Olivia Watkins, Patrick Chao, Samuel Miserendino, Tejal Patwardhan, Antonia Woodford, Beth Hoover, Jake

Brill, Kelly Stirman, Neel Ajjarapu, Nick Turley, Nikunj Handa, Olivier Godement, Akshay Nathan, Alyssa Huang, Andy Wang, Ankit Gohel, Ben Eggers, Brian Yu, Bryan Ashley, Chengdu Huang, Davin Bogan, Emily Sokolova, Eric Horacek, Felipe Petroski Such, Jonah Cohen, Joshua Gross, Justin Becker, Kan Wu, Larry Lv, Lee Byron, Manoli Liodakis, Max Johnson, Mike Trpcic, Murat Yesildal, Rasmus Rygaard, R. J. Marsan, Rohit Ram-chandani, Rohan Kshirsagar, Sara Conlon, Tony Xia, Siyuan Fu, Srinivas Narayanan, Sulman Choudhry, Tomer Kaftan, Trevor Creech, Andrea Vallone, Andrew Duberstein, Enis Sert, Eric Wallace, Grace Zhao, Irina Kofman, Jieqi Yu, Joaquin Quiñonero Candela, Made laine Boyd, Mehmet Ali Yatbaz, Mike McClay, Mingxuan Wang, Sandhini Agarwal, Saachi Jain, Sam Toizer, Santiago Hernández, Steve Mostovoy, Tao Li, Young Cha, Yunyun Wang, Lama Ahmad, Troy Peterson, Carpus Chang, Kristen Ying, Aidan Clark, Dane Stuckey, Jerry Tworek, Jakub W. Pachocki, Johannes Heidecke, Kevin Weil, Liam Fedus, Mark Chen, Sam Altman, and Wojciech Zaremba. Openai o3-mini system card. URL https://api.semanticscholar.org/CorpusID:276119184.

Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation, 2025. URL https://arxiv.org/abs/2405.07988.

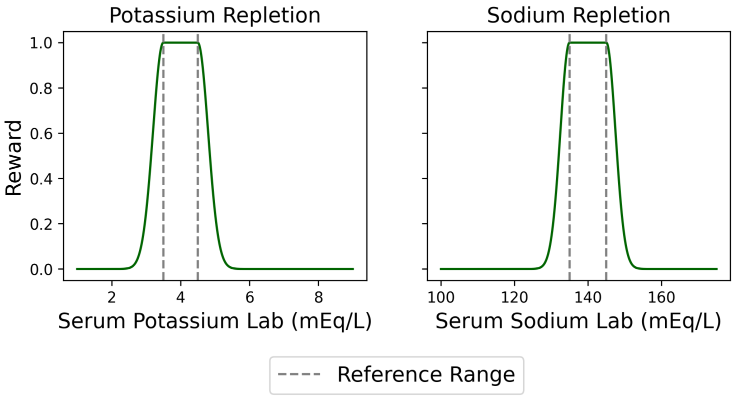


Figure 4: **Reward functions for both decision-making tasks are a function of the corresponding reference range.** Reward is bounded in the range $[0, 1]$, attaining its maximum when the lab value falls within the corresponding clinical reference range ($3.5 - 4.5$ mEq/L for serum potassium, and $135 - 145$ mEq/L for serum sodium). As the lab value deviates from this range, the reward decreases according to a Gaussian decay, with the lowest rewards assigned to critically low or high values.

Appendix A. MIMIC-IV dataset

The MIMIC-IV dataset consists of patient data for over 65,000 patients admitted to the ICU and over 200,000 patients admitted to the emergency department. This data is represented as electronic health records (EHRs), which capture a variety of information about each patient including static covariates such as age and gender, all hospital procedures and events such as lab measurements and administered medications, as well as indications of comorbidities. In this work, we consider non-ICU patients who have been administered either IV potassium, or IV hypertonic saline. We have 1622 patients who were administered IV potassium, and 1187 patients who were administered saline.

We represent each patient context using the following 15 features: age, gender, weight, height, heart rate, respiratory rate, oxygen saturation pulseoxymetry, systolic blood pressure, diastolic blood pressure, serum creatinine lab, administered NaCl 0.9%, administered dextrose 5%, administered propofol, administered norepinephrine, and administered insulin. We choose these features due to their relevance in being able to predict downstream serum potassium and serum sodium labs (Kluwer, n.d.). The reward function for both tasks is visualized in Figure 4.

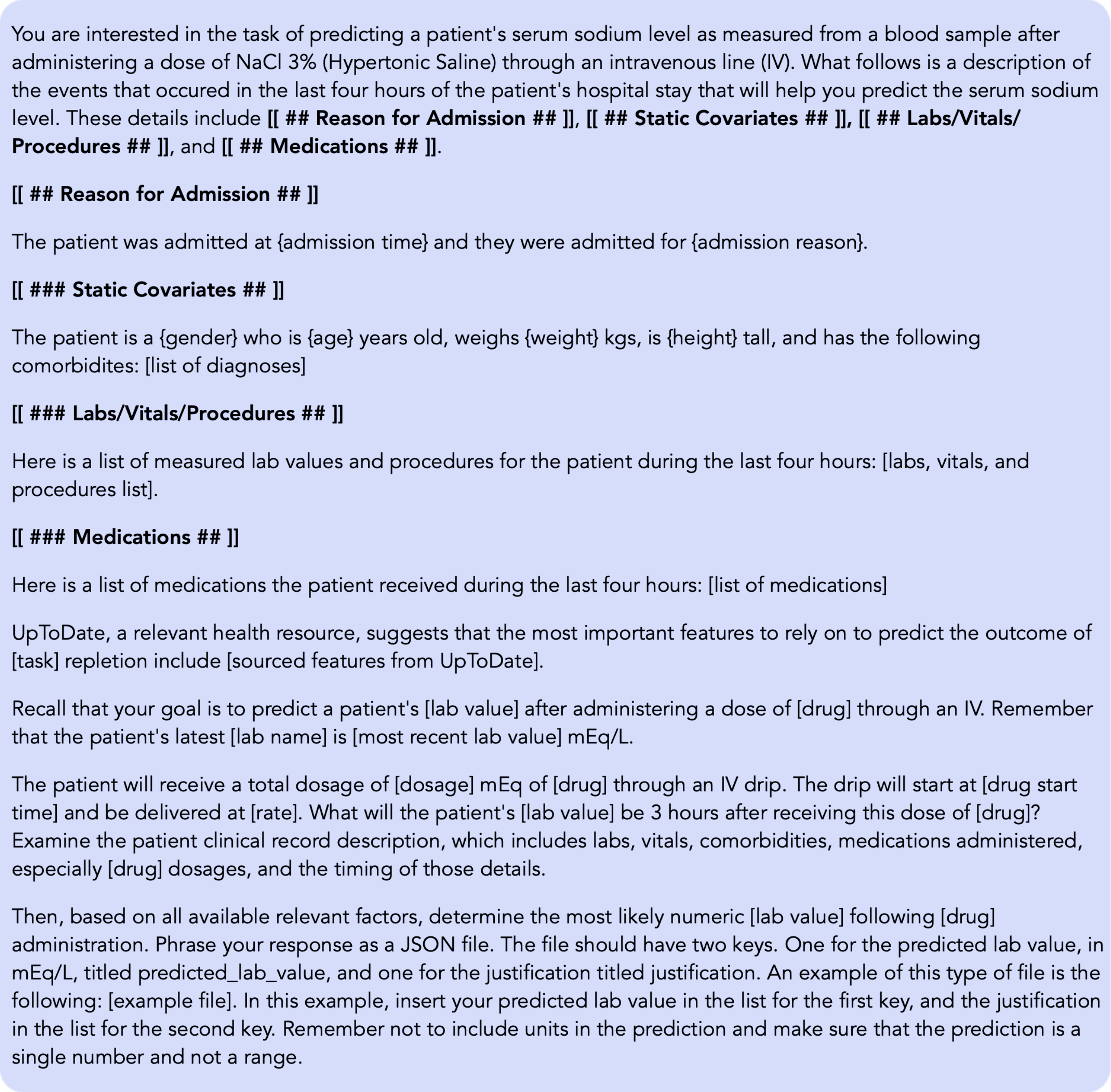
When we report the accuracy of the LLM in predicting downstream lab values, we use weighted F1 score. The classes of predictions for potassium (all in mEq/L) are $[< 3.2, >= 3.2$ and $< 5, >= 5$ and $< 6, >= 6]$. The classes of predictions for sodium (all in mEq/L) are $[< 118$, $>= 118$ and < 135, $>= 135$ and < 152, $>= 152$ and < 169, $>= 169]$.

Appendix B. Prompts

Here we include the format of the prompt used to query downstream lab value predictions. The format is consistent across both the potassium and sodium repletion tasks, and varies only based on individual patient details. The prompt consists of five components: task information, static covariates, labs and medicines, domain information from UpToDate, and a prediction query. An example prompt is shown in Figure 5.

Appendix C. Additional Empirical Results

Here, we include additional empirical results to support our claims in the main text. First, we report figures that demonstrate the quality of downstream lab predictions across LLMs for both potassium and sodium lab



You are interested in the task of predicting a patient's serum sodium level as measured from a blood sample after administering a dose of NaCl 3% (Hypertonic Saline) through an intravenous line (IV). What follows is a description of the events that occured in the last four hours of the patient's hospital stay that will help you predict the serum sodium level. These details include **[[## Reason for Admission ##]]**, **[[## Static Covariates ##]]**, **[[## Labs/Vitals/Procedures ##]]**, and **[[## Medications ##]]**.

[[## Reason for Admission ##]]

The patient was admitted at {admission time} and they were admitted for {admission reason}.

[[### Static Covariates ##]]

The patient is a {gender} who is {age} years old, weighs {weight} kgs, is {height} tall, and has the following comorbidites: [list of diagnoses]

[[### Labs/Vitals/Procedures ##]]

Here is a list of measured lab values and procedures for the patient during the last four hours: [labs, vitals, and procedures list].

[[### Medications ##]]

Here is a list of medications the patient received during the last four hours: [list of medications]

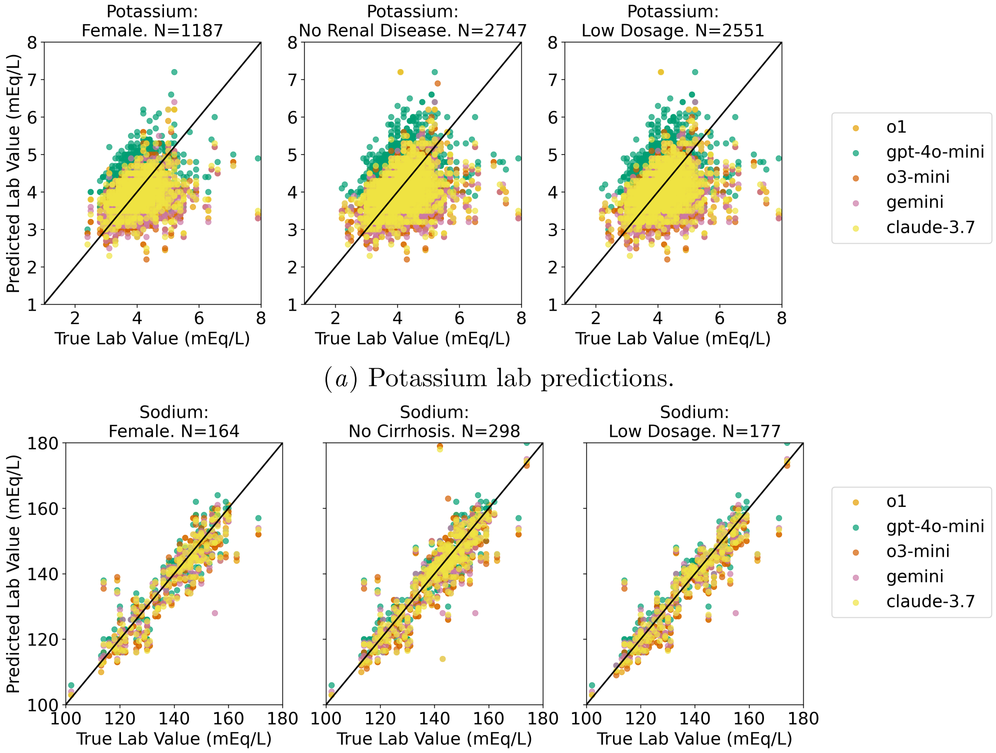
UpToDate, a relevant health resource, suggests that the most important features to rely on to predict the outcome of [task] repletion include [sourced features from UpToDate].

Recall that your goal is to predict a patient's [lab value] after administering a dose of [drug] through an IV. Remember that the patient's latest [lab name] is [most recent lab value] mEq/L.

The patient will receive a total dosage of [dosage] mEq of [drug] through an IV drip. The drip will start at [drug start time] and be delivered at [rate]. What will the patient's [lab value] be 3 hours after receiving this dose of [drug]? Examine the patient clinical record description, which includes labs, vitals, comorbidities, medications administered, especially [drug] dosages, and the timing of those details.

Then, based on all available relevant factors, determine the most likely numeric [lab value] following [drug] administration. Phrase your response as a JSON file. The file should have two keys. One for the predicted lab value, in mEq/L, titled predicted_lab_value, and one for the justification titled justification. An example of this type of file is the following: [example file]. In this example, insert your predicted lab value in the list for the first key, and the justification in the list for the second key. Remember not to include units in the prediction and make sure that the prediction is a single number and not a range.

Figure 5: **LLMs can be prompted to construct downstream lab value predictions.** The prompt contains separate components that first describe the patient's clinical state four hours prior to receiving treatment, and then contains instructions to perform the lab value prediction. The prompt includes relevant information from UpToDate, a clinical resource, to help an LLM identify which features in the medical record are most predictive.



(a) Potassium lab predictions.

(b) Sodium lab predictions.

Figure 6: **LLMs can accurately predict sodium and potassium lab values in MIMIC-IV.** Predictions are evaluated using weighted F1 scores across clinically relevant lab value categories. The black line denotes perfect agreement with ground truth, and predictions from a different LLM are reported in different colors.

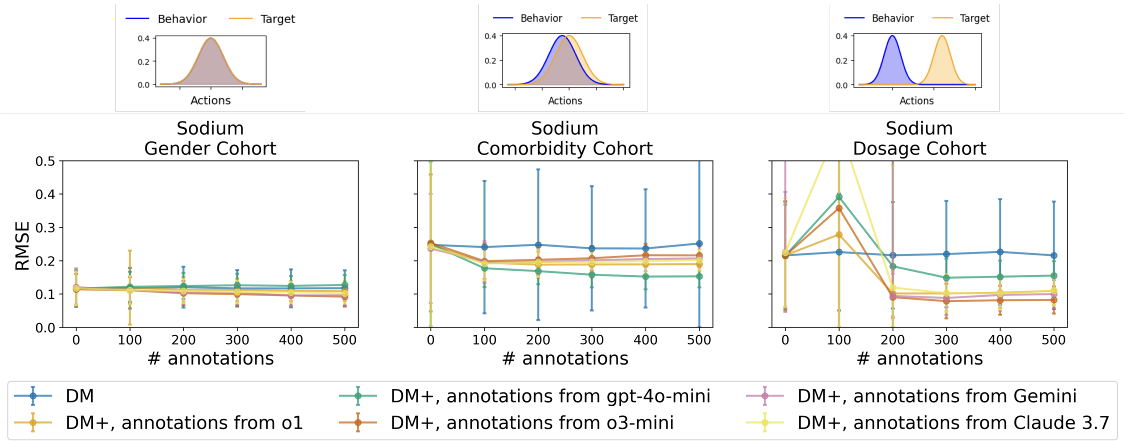


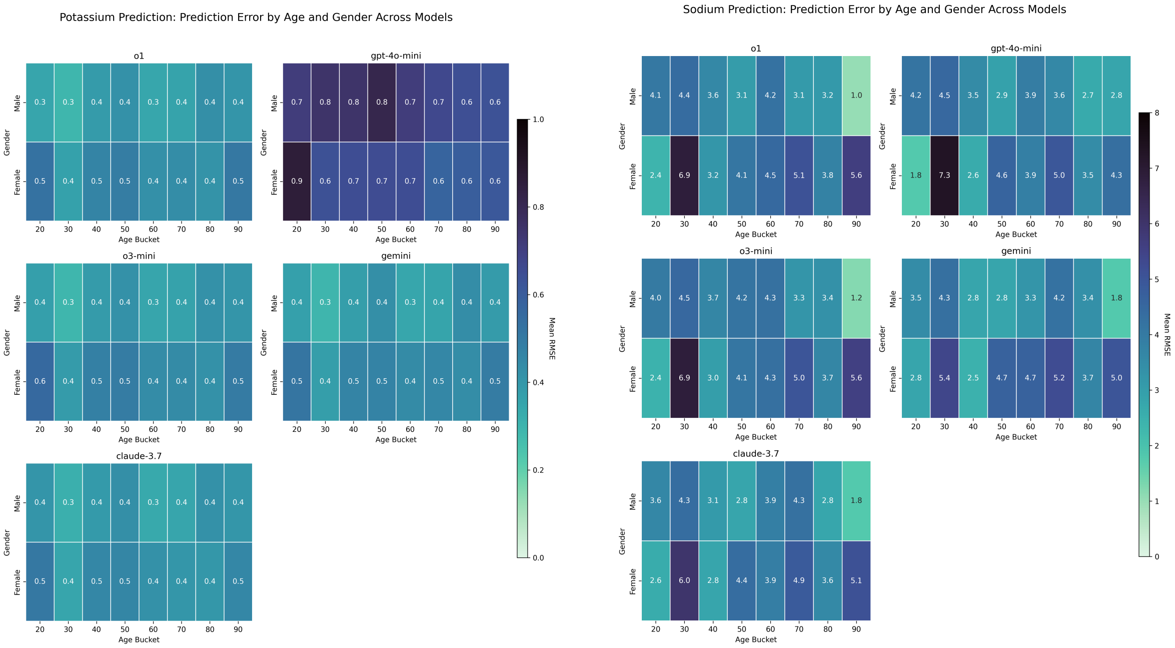
Figure 7: **LLM-generated counterfactual annotations improve OPE estimates for sodium repletion in settings with high divergence between behavior and target policies.** We report results using the DM baseline (blue) which uses no annotations, and DM^+ with annotations from different LLMs in other colors. Error bars represent standard error across 500 bootstrapped datasets, truncated at 0 when necessary.

predictions (Figure 6). Our results conclude similarly to those reported in Table 1, suggesting that LLMs can predict downstream lab values within clinically relevant degrees of error.

Furthermore, we investigate whether the age and gender of the patient affects the accuracy of the LLM in predicting potassium and sodium levels. We find that the prediction error varies substantially depending on the model and trends are not consistent given a patient's age or gender. (Figure 8).

Now we discuss the utility of LLM-generated counterfactual annotations within the sodium repletion task (Figure 7). Similar to the potassium repletion results, we find that LLM-generated counterfactual annotations help most when there is substantial divergence between the actions observed in the behavior and target policies. Just as in the potassium task results, the most improvement due to annotations occurs in the sodium dosage cohort.

Finally, we report entropy and further annotations results for the sodium repletion task, suggesting that, similar to the potassium repletion task, that more annotations may yield only marginal gains (Figure 9).



(a) Potassium Lab Prediction Errors

(b) Sodium Lab Prediction Errors

Figure 8: **Model prediction error varies depending on the patient's age and gender.** However, the trends are not consistent as observed for both sodium prediction error (Figure 8(b)) and potassium prediction error (Figure 8(a)).

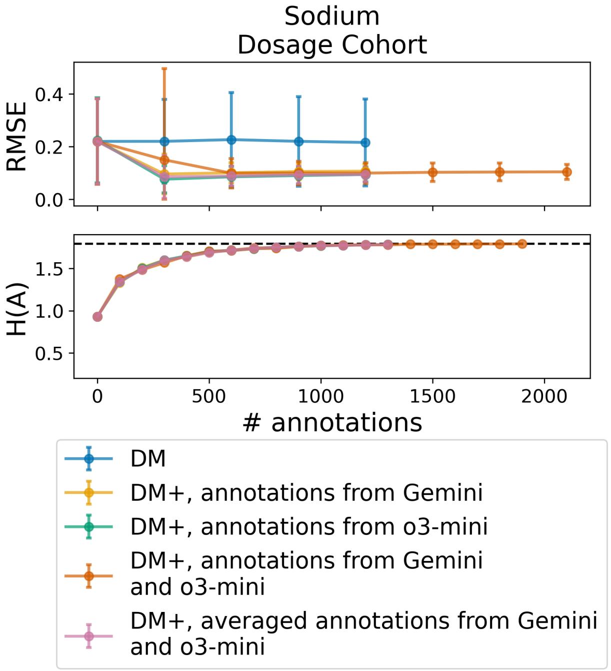


Figure 9: **Combining annotation sources yields limited returns in the sodium repletion task.** (Top) We compare DM^+ with the two best-performing LLMs for sodium repletion (yellow, green) and two aggregation methods: pooling predictions (orange) and averaging annotations (pink). Error bars show standard error over 500 bootstrapped datasets, truncated at 0. (Bottom) Marginal entropy over the action space $H(A)$ when adding counterfactual annotations to the behavior dataset for the sodium cohort. The horizontal dashed line marks the maximum possible entropy.