Uncertainty Quantification for Clinical Outcome Predictions with (Large) Language Models

Zizhang Chen
Brandeis University
zizhang2@brandeis.edu

Peizhao Li
GE HealthCare
peizhaoli05@gmail.com

Xiaomeng Dong
GE HealthCare
Xiaomeng.Dong@gehealthcare.com

Pengyu Hong
Brandeis University
hongpeng@brandeis.edu

Abstract

To facilitate healthcare delivery, language models (LMs) have significant potential for clinical prediction tasks using electronic health records (EHRs). However, in these high-stakes applications, unreliable decisions can result in high costs due to compromised patient safety and ethical concerns, thus increasing the need for good uncertainty modeling of automated clinical predictions. To address this, we consider uncertainty quantification of LMs for EHR tasks in both white-box and black-box settings. We first quantify uncertainty in white-box models, where we have access to model parameters and output logits. We show that an effective reduction of model uncertainty can be achieved by using the proposed multi-tasking and ensemble methods in EHRs. Continuing with this idea, we extend our approach to black-box settings, including popular proprietary LMs such as GPT-4. We validate our framework using longitudinal clinical data from over 6,000 patients across ten clinical prediction tasks. Results show that ensembling methods and multi-task prediction prompts reduce uncertainty across different scenarios. These findings increase model transparency in white-box and black-box settings, thereby advancing reliable AI healthcare. Our code is publically available at https://github.com/Cyrus9721/EHR_Uncertainty.

1 Introduction

Language models, such as (Steinberg et al., 2021; Theodorou et al., 2023; Steinberg et al., 2024) have emerged to be an efficient tool in the domain of EHR tasks. These models, extensively trained on diverse sources of clinical data, such as physician notes and longitudinal medical codes, have demonstrated remarkable effectiveness in predicting clinical outcomes. Despite their capabilities, measuring and reducing the uncertainties of these models in EHR tasks is crucial for ensuring patient safety, as clinicians can avoid interventions that the model indicates are uncertain and potentially hazardous. In addition, quantifying the uncertainties in clinical tasks can enhance the reliability of AI-driven medical decision-making systems (Begoli et al., 2019).

To address this challenge, leveraging the transparency of model parameters, we utilize established uncertainty metrics and propose to combine them with ensembling and multi-tasking approaches to effectively quantify and mitigate uncertainties in EHR tasks for these white-box language models. Recently, large language models have embarked on demonstrating their utility in clinical-related tasks, including EHR prediction tasks (Wornow et al., 2023b), analyzing radiology report examinations (Jeblick et al., 2024) and medical reasoning (Liévin et al., 2024). However, the encapsulation of modern Large Language Models, typically offered as API services with restricted access to internal model parameters and prediction probabilities, impedes the direct application of traditional uncertainty quantification methods. To overcome this limitation, We redefine uncertainty quantification as a post-hoc approach by analyzing the distribution of answers generated repeatedly from our designed prompts for clinical prediction tasks. Inspired by the effectiveness of our proposed methods in reducing model uncertainty for white-box LMs, we adapted and applied ensembling and multi-tasking methods to the black-box settings.

The main contributions of this paper are summarized as follows:

- We propose a multi-tasking method and a model ensembling approach to reduce model uncertainties for the white-box language model for clinical predictions using medical code sequences.
- We redefine the uncertainty quantification in EHR prediction tasks using black-box LLMs.
- We adapted our proposed two methods from white-box LM settings to black-box LLM set-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7527–7538
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

tings using natural languages and demonstrated their effectiveness in reducing uncertainties.

2 Background and Related Work

Uncertainty Quantification in Clinical Tasks Uncertainty quantification has emerged as a critical component in clinical tasks, particularly in safety-critical fields such as clinical decision-making (Begoli et al., 2019; Chen et al., 2021; Tomašev et al., 2021) and medical imaging (Edupuganti et al., 2020; Lambert et al., 2024). Current methods involve applying Bayesian approaches (Dusenberry et al., 2020; Jahmunah et al., 2023), ensembling methods (Mimori et al., 2021; Abe et al., 2024) and test-time augmentations (Ayhan et al., 2020) to reduce model uncertainties. In addition, (Uy, 2022; Rodman et al., 2023; Gao et al., 2024) examines the pre-test and post-test probabilities on clinical tasks for better decision-making. This work investigates model uncertainty on structured, longitudinal EHR datasets for clinical outcome predictions.

Langeuage Models for Clinical tasks Language models are transforming healthcare as an advanced tool for analyzing vast amounts of clinical data. Bert-based Models, like (Huang et al., 2019; Rasmy et al., 2021), have demonstrated their effectiveness in optimizing the disease treatment plan (Wang et al., 2023), clinical outcome prediction, etc. The recent success of large language models in reasoning and planning has prompted the community to adopt them for EHR tasks (Yang et al., 2022; Yoon et al., 2025). Uncertainty estimation is crucial for employing LLMs in electronic health record (EHR) tasks, as it helps mitigate the risk of false positives that could lead to inappropriate treatments or interventions (Savage et al., 2025). This paper mainly focuses on estimating and reducing the uncertainties for proprietary LLMs.

Uncertainty Quantification with LLMs The increasing reliance on black-box large language models (LLMs) such as GPT-4 (Achiam et al., 2023), Claude 3 (Anthropic, 2023), and Gemini (Team et al., 2023) in commercial applications has introduced complex challenges in Uncertainty Quantification. Due to the closed nature of LLMs, typically offered as API services, traditional uncertainty quantification methods that require access to model parameters are not applicable. To overcome these challenges, recent research (Kuhn et al., 2023; Lin et al., 2023; Xiong et al., 2024) has developed innovative techniques that estimate uncertainty based directly on the text outputs from LLMs, bypassing the need for internal data. Notably, Kuhn et al.(2023) have proposed semantic entropy as a new metric for quantifying uncertainty in LLMs, which capitalizes on the semantic equivalence across varying expressions. Subsequent studies (Lin et al., 2023; Xiong et al., 2024) have further advanced these approaches, crafting sophisticated methods that enhance black-box UQ through strategic prompting, sampling, and result aggregation. The most recent work (Geng et al., 2024) comprehensively surveys confidence estimation and uncertainty calibration methods for LLMs. It outlines several real-world applications to build more responsible AIs, including Hallucination detection and mitigation (Manakul et al., 2023; Varshney et al., 2023) and constructing an efficient Retrieval Augmented Generation (RAG) pipeline (Jiang et al., 2023).

3 Predictions in Electronic Health Records

Clinical Tasks We present our findings across a range of clinical tasks using the newly published EHRSHOT dataset (Wornow et al., 2023a). The EHRSHOT dataset encompasses **structured**, **longitudinal** clinical data extracted from the electronic health records of 6,739 patients at Stanford Medicine, featuring over 40 million clinical events. Within the framework of EHRSHOT, we explore 10 EHR prediction sub-tasks organized into three distinct categories. *(i) General Operational Outcomes*: including long length-of-stay: predicting whether a patient's length of stay will exceed seven days, and ICU Transfer: predicting whether the patient will be transferred to the ICU on the same day of admission. *(ii) Lab Test results*: This task involves predicting the normalcy of lab test results immediately before their official release. The lab tests covered Thrombocytopenia, Hyperkalemia, Hypoglycemia, Hyponatremia, and Anemia. *(iii) New Diagnose Diseases*: This involves forecasting whether the patient will be first diagnosed with specific diseases within the next year from the date of discharge. Diseases tracked include Hypertension, Hyperlipidemia, and Acute Myocardial Infarction.

EHR Tasks Formulation The EHR prediction tasks can be formulated as follows: consider patient p's data represented as $\{C_p, Y_p\}$. Here, medical sequence C_p=$\{(c_1, t_1), (c_2, t_2), ..., (c_n, t_n)\}$. denotes the complete clinical events for pa-

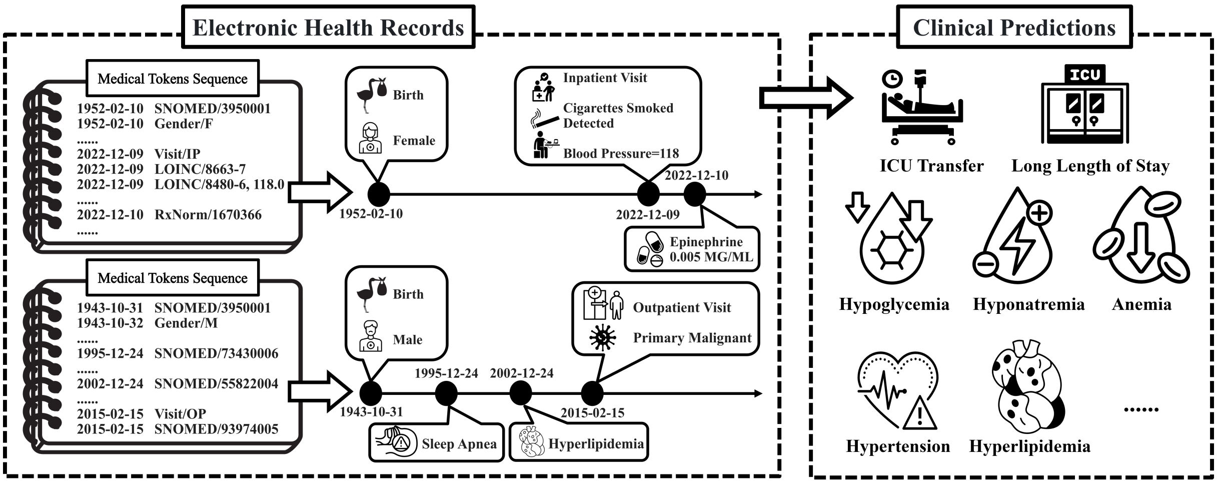

Figure 1: EHR predictions with medical codes sequences. **Left**: structured, longitudinal medical tokens, each code is in OMOP format (Reich et al., 2024) and associated with a specific time point. We translate these codes into natural languages that describe a patient's timeline. **Right**: The interpreted EHR data can be used for severing various clinical applications such as Long Length of Stay or Hypoglycemia predictions.

tient p arranged chronologically. Each code c_i represents a specific medical event at time point t_i in Observational Medical Outcomes Partnership (OMOP) format (Reich et al., 2024). $Y_p = \{y_1, y_2, ...y_m\}_{m \in \{t_1, t_2, ...t_n\}}$ is a set of labels indicating the clinical outcomes. At time t_k, our objective is to predict label y_{t_k}, with a truncated medical sequence Here, we denote $C_{t_{k-1}} = \{(c_1, t_1), (c_2, t_2), ...(c_{k-1}, t_{k-1})\}$, which consists of data up to, but not include time point t_k.

4 Modeling Uncertainties for Clinical Outcome Predictions in EHR

We consider two settings for uncertainty modeling: white-box and black-box. In white-box settings, the access to model's parameters and output probabilities/logits is available. We employ four widely recognized metrics: Brier Score (Rufibach, 2010), Expected Calibration Error (Naeini et al., 2015), Adaptive Expected Calibration Error (Nixon et al., 2019), and Negative Log Likelihood. We also employ two additional methods, Deep Ensemble and Monte Carlo Dropout, to reduce the model uncertainties for clinical predictions. We detail the formulation of the uncertainty metrics and implementing uncertainty methods in section 4.1. Reducing the model uncertainties can enhance the model's trustworthiness, especially in clinical decision-making. In the black-box setting, where the intrinsic model parameters are unavailable, one approach to quantify uncertainty is calculating entropy-based metrics on a repeatedly generated answer set. Then, these metrics are used to predict whether to rely on the model's response or not (Filos et al., 2019; Kuhn et al., 2023). We utilize uncertainty metrics to quantify the trustworthiness of the proprietary model in EHR tasks.

4.1 Modeling and Reducing Uncertainties in White-box Settings

Clinical Prediction with BERT-Based Language Models In our *white box* setting for EHR predictions, we first follow the settings of (Steinberg et al., 2021; Wornow et al., 2023a) to generate sequence embeddings $\{e_i\}_{i=1}^{k-1}$ from medical code sequence $C_{t_{k-1}}$ using the CLMBR-T-base, a foundation model pre-trained on 2.57 million deidentified structured patient records from the private Stanford HealthCare Data Warehouse. The CLMBR-T-base model is pre-trained autoregressively to predict the subsequent medical code based on their prior medical sequence. The published CLMBR-T-base model includes encoders that generate representations of a patient's medical tokens. It contains embedding layers that first map up to 65536 unique medical tokens into a hidden dimension space of 768, then followed by a 12 stacked transformer layer with a fixed context window of 496 tokens. Multiple token representations could exist at the time point t_{k-1}; however, the last possible representation e_{k-1} is used for the downstream EHR predictions. According to (Wornow et al., 2023a), this selection strategy not only maximizes the utilization of available information but also helps prevent data leakage by ensuring that future data do not influence prediction. We subsequently trained decoders using the embedding e_{k-1} for each downstream EHR prediction task as introduced in section 3. We then assess model uncertainties and implement the uncertainty methods across decoders.

Uncertainty Metrics We adopt Brier Score, Expected Calibration Error, Adaptive Expected Calibration Error, and Negative Log Likelihood to quantify and assess the model uncertainties in EHR prediction tasks within our *white-box* settings. We explain the formulations for these metrics: Brier Score **(BS)** is formulated to evaluate the accuracy of model's probabilistic predictions:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2. \tag{1}$$

where y_i represents the actual label and $\hat{p}_i$ is the predicted probability of a clinical outcome for each case i. A lower Brier score reflects the higher confidence of the model in its classification predictions.

Expected Calibration Error/Adaptive Expected Calibration Error **(ECE/aECE)** is used to measure the calibration error of the classification models:

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{2}$$

which calculate the difference between predicted probabilities and actual outcomes in M bins, which are formed by dividing predicted probabilities into a series of intervals. Here, B_m are prediction bins and acc(B_m) and conf(B_m) represent the precision and the average predicted probability within each bin, respectively. ECE uses fixed-width bins, while aECE adjusts the bin widths based on the data distribution. A lower ECE/aECE suggests that a probabilistic model is well-calibrated, indicating a close correspondence between the predicted probabilities and the actual clinical outcomes.

Negative log-likelihood **(NLL)** measures the probability of the actual data given the model parameters, and representation set x_i:

$$\text{NLL} = -\sum_{i=1}^{N}\log(p(y_i|x_i)). \tag{3}$$

A lower Negative log-likelihood indicates that the model assigns high probabilities to the correct outcomes, implying high confidence in its predictions. aECE/ECE directly measures the model calibration error but does not penalize the model for being unconfident. The BS measures the confidence level for the prediction, and NLL significantly penalizes the model for being overconfident about false predictions. Integrating the above three metrics measures the overall model uncertainties.

Uncertainty Methods We implement *two uncertainty methods* and propose *one framework* to quantify and reduce model uncertainty in the decoders used for clinical predictions. We introduce two uncertainty methods: Monte Carlo Dropout (Gal and Ghahramani, 2016) and Deep Ensemble (Lakshminarayanan et al., 2017; Rahaman et al., 2021). Monte Carlo Dropout applies dropout not only during the training of neural networks but also during inference stages. This approach approximates Bayesian inference in deep Gaussian processes and allows the model to generate a distribution of predictions. Thus, it quantifies the uncertainty by allowing the network to express its confidence level through the variability of its predictions under different neuron configurations. Deep ensembles are created by training multiple versions of the same decoders, with variations only in random seeds and hyperparameters. The predictions of individual models are aggregated to produce a final prediction. This approach mitigates the inherent bias on the models, thereby reducing the overall uncertainty of the model. In addition, we propose a simple yet effective multitasking framework to predict multiple clinical tasks within the same category (*General Operation Outcome*, *Lab Test Results*, and *New Diagnose Diseases*) presented simultaneously in section 3. For simplicity, we formulate the multitasking framework as follows:

$$\{y^{h_i}_{t_{k-1}}\}_{h_i \in H} = f(e_{k-1} \oplus e_{h_i}). \tag{4}$$

Here, e_{k-1} is the representation embeddings for the clinical sequences $C_{t_{k-1}}$, e_{h_i} denotes the task-specific embeddings that are combined with e_{k-1} to distinguish among different clinical subtasks, h_i refers to a specific clinical task, and H represents the clinical tasks within the same category.

We present our results table 3. Observing Deep Ensemble's multitasking approach reduces uncertainties with white-box models by a large margin in EHR tasks. We are compelled to ask:

Can uncertainty reduction apply to proprietary LLMs?

This propels us to investigate specialized uncertainty quantification for the proprietary **black-box** models like GPTs in EHR prediction tasks.

4.2 Transferring to Black-box Setting

Clinical Prediction with GPTs In our *black-box* settings where the model parameters are opaque, we transform the patient's medical code sequence $C_{t_{k-1}}$ into free-form languages by generating text descriptions for each medical token. This process leverages information from the Athena Ontology Database (Hripcsak et al., 2015; Reich et al., 2024), where the concept of each medical code is clearly defined, categorized and described. We then adapted the descriptions by retaining only the most recent medical code details to fit within the input context length constraints of the GPT models. We use $S_{t_{k-1}}$ to denote the text descriptions of the medical sequences of a patient. We construct prompt $P_{t_{k-1}}$=$\{G, S_{t_{k-1}}, \Omega, O\}$ based on $S_{t_{k-1}}$, where G denotes general prompts which specify the role and scenario LLM acts, Ω specifies the clinical tasks to be performed, O defines the output formats of the GPT's response. We instruct the LLM to repeatedly generate a set of n responses R=$\{r_1, r_2, ..., r_n\}$ from prompt $P_{t_{k-1}}$. We then conduct post-hoc uncertainty quantification over the response set R. In the following sections, we detail the methods for quantifying and reducing the uncertainties in white-box LM settings and proprietary LLM settings for EHR prediction tasks.

Uncertainty Quantification for EHR tasks with GPTs We introduce the uncertainty quantification method and develop the corresponding metrics explicitly designed for proprietary large-language models, which are applied in the context of clinical prediction tasks. Similarly to the existing pipeline on proprietary LLM uncertainty quantification (Kuhn et al., 2023; Liu et al., 2024), we characterize black-box uncertainty quantification methods as post hoc approaches, required by the limited access to internal model parameters.

1. For a clinical prediction task on an EHR for the patient p, we generate a patient text description S_p from its medical sequence C_p.

2. Construct the prompt P=$\{G, S_p, \Omega, O\}$ by integrating the text description with the role and scenario prompt G, the clinical task prompt Ω, and the output format prompt O.

3. For clinical tasks with a specific prediction time t_k, we customized the prompt P accordingly: first, all clinical events after the time point t_k are excluded to prevent data leakage; second, we limit the number of clinical events included to fit within the context length constraints of the Large Language Model. We denote the tailored clinical events descriptions as $S_{t_{k-1}}$ and the customized prompt as $P_{t_{k-1}}$

User: What is code "CVX/20?"

ChatGPT4:

Answer 1: The term "CVX/20" doesn't refer to a widely recognized standard or concept ...

Answer 2: The code "CVX/20" typically refers to a vaccine code in the context of medical records ... However, I'd need more context to provide a specific answer about this code ...

User: What is: "diphtheria, tetanus toxoids and acellular pertussis vaccine," can you briefly explain?

ChatGPT4:

Answer: The "Diphtheria, Tetanus Toxoids and Acellular Pertussis Vaccine," commonly abbreviated as DTaP, is a combination vaccine that protects against three serious diseases caused by bacteria ... This vaccine is crucial for preventing these diseases in children, who are particularly susceptible to serious complications from them ...

Table 1: A card of a sample dialogue between a user and ChatGPT4 regarding medical codes and their descriptions: GPTs have a limited understanding of OMOP medical tokens. The first prompt is a direct ask for the OMOP code. The second prompt converts the OMOP codes into clinical descriptions using the Athena Ontology Database. We observe that GPT does not fully understand the medical codes used in various formats across hospitals and organizations (Soroush et al., 2024), while it can relatively grasp the clinical meanings denoted by these tokens.

4. Instruct Large Langugage Models to generate n responses for the prompt $P_{t_{k-1}}$. Consequently, we obtain a set of responses $R_{t_{k-1}} = \{r^1_{t_{k-1}}, r^2_{t_{k-1}}, ..., r^n_{t_{k-1}}\}$. Here, $r^i_{t_{k-1}}$ represents the outputs of i^{th} response.

5. Calculate the uncertainty score U from the response repeatedly generated $R_{t_{k-1}}$.

In Step 1, we expand each medical token with the corresponding time point $(c_i, t_i){\in}C_p$ into free-formed languages that describe medical events at time t_i. We describe and demonstrate the significance of this conversion in section 4.2.

For steps 2 and 3, we constructed our prompts into four parts: *(i)* A role-playing prompt G with task-specific instructions; here, we instruct GPT to act as experienced doctors capable of providing clinical insights. *(ii)* Clinical descriptions S_p that indicate the patient p's medical events. We begin by extracting the patient's demographic information and calculating the patient's age at the prediction time from C_p using the OMOP Common Data Model (Hripcsak et al., 2016; Reich et al., 2024). *(iii)* Questions and Clinical Tasks Ω to be answered. We develop task-specific prompts for each of the ten clinical tasks within the EHRSHOT dataset. *(iv)* Output format prompt O that requests

GPT to provide restricted responses for clinical tasks (E.g., "Yes, the patient will be transferred to ICU." or "No." otherwise. For predicting whether the patient will be transferred to ICU on the day on admission). We present section 4.2 to demonstrate our prompt design for guiding GPTs in EHR tasks. Due to the lengthy context of clinical descriptions, we repeat our task questions at both the beginning and ending of a prompt, utilizing the Needle-In-A-Haystack (NIAH) method with GPT models. *For steps 4 and 5*, given a set of repeatedly generated responses: $R_{t_{k-1}} = \{r^1_{t_{k-1}}, r^2_{t_{k-1}}, ..., r^n_{t_{k-1}}\}$. We adopt the methodology outlined in (Kuhn et al., 2023) to compute Class Entropy $U(R_{t_{k-1}})$ as *Uncertainty Score*. $U(R_{t_{k-1}})$'s formula presented as:

$$U(R_{t_{k-1}}) = -\sum_{a_i \in A} P(a_i) \log P(a_i), \qquad (5)$$

where $a_i \in A$ is the clinical outcome label predicted by GPTs, the probability $P(a_i)$ of each clinical class i is determined by the frequency of occurrences of class a_i within the answer class set A. Following (Kuhn et al., 2023; Lin et al., 2023), we construct our Uncertainty Metric using the Uncertainty Score $U(R_{t_{k-1}})$ to predict whether LLM can correctly generate an answer. We employ the area under the receiver operating characteristic curve (AUROC) as the metric for assessing uncertainty.

4.3 Reducing Uncertainties for proprietary black-box models

In this section, we employ two approaches to reduce the uncertainty in EHR predictions generated by GPTs. Our first approach involves ensembling clinical predictions from multiple GPT models. For a specific prompt generated from a patient's sequence of medical codes, we repeatedly generate response sets from GPT-3.5-Turbo and GPT-4. We then combine two sets of GPT responses and compute the uncertainty score U. This idea is drawn from the proven efficacy of ensemble methods in reducing uncertainty for white-box deep learning models, as substantiated by key studies (Lakshminarayanan et al., 2017; Rahaman et al., 2021; Abe et al., 2023) and our empirical findings in table 3 with clinical EHR tasks. Similarly, based on our empirical findings from section 5.3, where predicting multiple clinical tasks within the same category can marginally reduce the uncertainty of the white box model. We extend this methodology to large language models of the GPT style. Our second approach involves instructing GPT models to generate predictions for several EHR tasks within the same clinical task category in a single-generation process. Similarly to eq. (4), we formulate the multi-task framework for GPTs as follows:

$$\{y^{h_i}\}_{h_i \in H} = LLM(\{G, S, \Omega, O\}). \qquad (6)$$

Here, h_i refers to a specific clinical task, and H represents the set of all clinical tasks within the same category in section 3. $\{G, S, \Omega, O\}$ is our constructed prompt for the clinical tasks S. $\{y^{h_i}\}$ represents the ensemble of EHR predictions we prompt the LLM to generate simultaneously.

Role: Assuming you are an experienced doctor. Based on the descriptions of the patient's age, demographics, and medical events provided, Use your knowledge and reasoning to predict whether $\{Tasks\}$.
Chain of thoughts:
1. Review Patient Profile: Analyzing age, sex, and medical history ...
2. Evaluate Current Symptoms: Identify vital signs outside the normal range ...
......

Patient age and demographic information:
The patient was 36 years old at the discharge time. The patient has the following demographic information: ...
Medical Events:
On June 12, 2014:
One clinical drug event, "Oxycodone hydrochloride 5 MG Oral Tablet" was recorded.
One clinical Drug event, "Acetaminophen 10 MG/ML Injectable Solution" recorded.
On June 13, 2014:
One clinical drug event, "Oxycodone hydrochloride 5 MG Oral Tablet" was recorded.
Four measurement events, "Systolic blood pressure" was recorded with values: 131.0, 127.0, 133.0, 143.0.
......

Tasks: $\{Tasks\}$.

Output format:
Please answer with "Yes" or "No".

Table 2: An example card of prompts for EHR tasks, structured into four parts. The first section comprises a general prompt incorporating role-playing and chain-of-thought reasoning using GPTs. The second section comprises clinical descriptions in natural languages converted from medical codes. The third and fourth sections describe the clinical task and output restrictions.

5 Experiment

5.1 Data Setup

Following (Wornow et al., 2023a), we use the EHRSHOT benchmark to set up our experiments on structured longitudinal EHR datasets. We selected ten clinical tasks to evaluate the model uncertainty of the white-box model in clinical pre-

	Single-tasking Baseline				Single-tasking Deep Ensemble				Single-tasking MC dropout			
EHR Task / U.Q Metric	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓
Long Length of Stay	0.5645	1.5253	0.2575	0.2575	0.4942	0.8746	0.1661	0.1696	0.5416	1.3656	0.2232	0.2265
ICU Transfer	0.1388	0.4266	0.0530	0.0527	0.1013	0.3291	0.0425	0.0411	0.1108	0.4432	0.0500	0.0424
Thrombocytopenia	0.5002	0.6935	0.4861	0.4857	0.0395	0.1364	0.0183	0.0219	0.5004	0.6938	0.4858	0.4849
Hyperkalemia	0.5006	0.6939	0.4809	0.4796	0.0496	0.1772	0.0252	0.0272	0.5002	0.6933	0.4806	0.4801
Hypoglycemia	0.6345	1.4092	0.2727	0.2706	0.5160	0.7986	0.1357	0.1357	0.6178	1.1984	0.2399	0.2400
Hyponatremia	0.7079	1.6924	0.3162	0.3162	0.5613	0.9057	0.1748	0.1748	0.6701	1.3312	0.2745	0.2745
Anemia	0.7212	1.7921	0.3213	0.3211	0.5888	0.9668	0.1986	0.1986	0.6673	1.4020	0.2750	0.2737
Hypertension	0.2763	0.8392	0.1149	0.1134	0.2587	0.5458	0.0905	0.0893	0.2756	0.7987	0.1181	0.1101
Hyperlipidemia	0.2495	0.8229	0.1157	0.1111	0.2187	0.5106	0.0825	0.0810	0.2453	0.7004	0.1051	0.0967
Acute MI	0.0769	0.2480	0.0315	0.0282	0.0690	0.1676	0.0302	0.0226	0.0900	0.2308	0.0339	0.0268

	Multi-tasking Baseline				Multi-tasking Ensemble				Multi-tasking MC dropout			
EHR Task / U.Q Metric	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓	*Brier* ↓	*NLL* ↓	*ECE* ↓	*aECE* ↓
Long Length of Stay	0.6215	1.5723	0.2798	0.2789	0.5257	0.9421	0.1890	0.1896	0.5703	1.3140	0.2381	0.2381
ICU Transfer	0.0993	0.7288	0.0477	0.0466	0.0922	0.3905	0.0399	0.0382	0.0969	0.6399	0.0455	0.0418
Thrombocytopenia	0.0344	0.1532	0.0199	0.0227	0.0301	0.0947	0.0108	0.0203	0.0369	0.1293	0.0231	0.0300
Hyperkalemia	0.0459	0.1692	0.0246	0.0284	0.0408	0.1197	0.0155	0.0234	0.0474	0.1490	0.0272	0.0325
Hypoglycemia	0.5001	0.6945	0.2230	0.2208	0.5020	0.6972	0.0787	0.0882	0.4959	0.6929	0.2119	0.2067
Hyponatremia	0.5649	1.1548	0.2236	0.2225	0.5047	0.8144	0.1702	0.1701	0.5242	0.9476	0.1767	0.1757
Anemia	0.6797	1.3099	0.2856	0.2824	0.5605	0.8118	0.1586	0.1586	0.6386	1.0853	0.2379	0.2366
Hypertension	0.2778	0.9824	0.1213	0.1205	0.2549	0.5810	0.0903	0.0793	0.2601	0.8215	0.1075	0.1019
Hyperlipidemia	0.2503	1.0269	0.1151	0.1147	0.2410	0.6209	0.0936	0.0927	0.2283	0.9220	0.1023	0.0944
Acute MI	0.0636	0.3306	0.0299	0.0258	0.0615	0.1892	0.0276	0.0252	0.0675	0.3405	0.0292	0.0289

Table 3: Uncertainty Metrics for clinical predictions with BERT-Based Language Models on white-box settings. **Top**: Uncertainty metrics for single task settings. Ten decoders are trained for each clinical baseline task. **Bottom**: Uncertainty metrics for multi-task settings. Three decoders are trained for each baseline clinical task category. **left**: Baseline, **Middle**: Deep Ensembles, **Right**: MC Dropout. We adopted single-tasking without ensembles or MC dropouts **Top left** as the baseline. We compare the baseline with the other five methods, including a combination of the multi-tasking setting and the Ensembling and MC dropout methods. We highlight the chunks with the lowest uncertainty metrics.

White-box Data	#Patient	#Events	# Train/Val/Test
Operational Outcome	3,617	6,491	2,402 / 2,052 / 2,037
Lab Tests	5,691	152,331	59,983 / 44,928 / 47,420
New Diagnose	1,916	2,794	959 / 956 / 879

Black-box Data	#Patient	# Avg. Tokens	# Train/Val/Test
Operational Outcome	89	4,609	- / - / 100
Lab Tests	100	3,907	- / - / 100
New Diagnose	99	4,417	- / - / 100

Table 4: Data Statistics for both while-box and black-box settings. *Operational Outcome*, include Long Length of Stay and ICU Transfer. *Lab Tests* include Thrombocytopenia, Hyperkalemia, Hypoglycemia, Hyponatremia and Anemia. *New Diagnose* include Hypertension, Hyperlipidemia, and Acute MI. **Top**: Clinical medical codes in structured, longitudinal format for assessing white-box model uncertainties. **Bottom**: Natural languages converted from medical sequence, for assessing uncertainties of proprietary GPTs.

dictions. In the data preparation phase, we tailor the EHRSHOT dataset for each sub-task within a given task category. This approach ensures that, for each prediction category, every patient's medical sequence is associated with multiple corresponding labels at the same prediction time point. To assess the uncertainty of proprietary GPTs, we stochastically choose 100 medical sequences from each category of clinical tasks in the EHRSHOT database. The selection method ensures that each dataset includes a sufficient number of positive labels. We present our dataset information in Table 4.

5.2 White-box Model Results

This section presents the Uncertainty Quantification findings of using BERT-based language models for EHR tasks. We first generate sequential embeddings from structured, sequential medical codes using CLMBR-T-base. We then follow the setting of (Wornow et al., 2023a) to extract the medical codes' representation at prediction time point for downstream tasks. The prediction time point for General Operation Outcome tasks is at 11:59 pm on the day of admission and visits that last less than one day. For lab testing tasks, the prediction time point corresponds to one minute before the latest lab results are available. For New Diagnosis tasks, the prediction time is set to one minute before midnight on the day of discharge. We then trained a 2-layer Neural Network as decoders on the extracted representation embeddings for downstream tasks and reported the uncertainty metrics in the **top left** section of table 3. We then implement Deep Ensemble and MC Dropout to reduce model uncertainties, results presented in the **top middle** and **top right** sections of table 3, respectively. For Deep Ensemble, we set the number of ensembling models to 5. For MC Dropout, we set the dropout ratio to 0.5. Second, we implement a multi-tasking approach such that a decoder can give predictions for clinical tasks within the same category. We begin by generating embedding for each task and combining it with the representations. Again, we implement the Deep Ensembles and MC Dropout for the multi-tasking pipeline with the same hyperparameter. We present the multi-tasking uncertainty metrics for the at the **bottom** of Table 3.

Observations For both single-task settings and multi-task settings, Deep Ensemble consistently shows lower uncertainty metrics (measured by Brier score, NLL, ECE, and aECE) across most EHR tasks compared to the baseline. For new diagnosis tasks like "Thrombocytopenia" and "Hyperkalemia," Deep Ensemble improves considerably upon the baseline. The MC Dropout method also improves over the baseline but is generally less effective than the Deep Ensemble in reducing uncertainty. This indicates the effectiveness of the ensembling method in reducing the uncertainty of the model for clinical predictions, which prompted us to propose our initial methods to reduce the uncertainties of proprietary GPTs in section 4.3. In addition, the multi-task Deep Ensemble configuration consistently shows lower uncertainty metrics than the single-task configuration. Similar to Deep Ensemble, MC Dropout benefits from a multi-task setting, albeit with smaller margins of improvement. This indicates the benefits of our proposed multi-tasking method in EHR predictions, where tasks are within the same clinical category.

5.3 Black-box GPT Results

We conduct evaluations of our black-box uncertainty quantification methods using the specifically tailored EHRSHOT dataset, as detailed in table 4. For each task category, we began by filtering the EHRSHOT dataset based on prediction time points; we ensured that each instance of converted medical language had corresponding ground truth labels for all tasks within the same category. We then constructed the test set for each task category by stochastically sampling 100 entries from the tailored EHRSHOT dataset. Our sampling algorithm was designed to terminate once the data for all subtasks contained at least 12 positive labels. For evaluations, we employ GPT-4 and GPT-3.5 Turbo to generate responses. We repeatedly generate five responses for each constructed prompt. We then calculate the AUC score by cleaning the outputs and matching the generated answers with the ground-truth clinical outcomes. We then calculate the uncertainty metric $U.Q.$ in eq. (5) and use it to predict whether the response from GPTs is correct. Similar to the multitasking method described in section 5.2, we reformulated our prompts to request predictions for multiple clinical tasks within the same category. Furthermore, akin to the Deep Ensemble approach in the white-box setting, we aggregate the responses from multiple GPTs and calculate the AUCs and the uncertainty metrics. Our results are presented in Table 5.

Observations In evaluating model performances, GPTs show limited performance in predicting clinical outcomes from free-formed languages. In evaluation uncertainties, for GPT-3.5-Turbo and GPT-4, the ensembling methods score higher in UQ metrics in almost every case listed than in the single-model setting. In addition, GPT-4 ensembles outperform single-model significantly in UQ metrics. This indicates that ensembling models perform significantly better in quantifying thus reducing uncertainty across all clinical prediction tasks than their single counterparts. We observe minimal or no improvements in analyzing the UQ metric between single-tasking and multi-tasking settings within individual GPT models. Tasks such as Hypokalemia and Hyponatremia exhibit similar U.Q. scores regardless of whether they are approached through single or multi-task configurations within the same models. However, when multi-tasking is integrated with ensemble methods, we observe a marginal improvement in U.Q metrics. This indicates that combining multi-tasking approaches within ensemble frameworks substantially contributes to more reliable assessments of whether to trust LLM's output, thus reduce the uncertainties for EHR predictions.

6 Conclusion

In this work, we explored the quantification and reduction of uncertainty in clinical outcome predictions with EHR by harnessing white-box language models and black-box large language models. We focused on two main methodologies to mitigate uncertainty: ensemble methods, which combine predictions from multiple models, and multi-tasking, where models simultaneously predict multiple clinical outcomes. By transferring and adapting these methodologies originally developed for LMs to the realm of LLMs, we demonstrated reductions in uncertainties across white-box and black-box models.

Limitations Our uncertainty quantification methods were validated using clinical prediction tasks containing longitudinal EHRs. While the results

	GPT-3.5 Single		GPT-3.5 Multi		GPT-4 Single		GPT-4 Multi		Ensemble Single		Ensemble Multi	
	Auc.↑	*U.Q.*↑	*Auc.*↑	*U.Q.*↑	*Auc.*↑	*U.Q.*↑	*Auc.*↑	*U.Q.*↑	*Auc.*↑	*U.Q.*↑	*Auc.*↑	*U.Q.*↑
Long Length of Stay	0.5430	0.4570	0.6153	0.4265	0.3614	0.4992	0.5125	0.4875	0.5461	0.8385	0.6166	0.7237
ICU Transfer	0.5047	0.5331	0.6853	0.3596	0.5938	0.5140	0.7083	0.4888	0.5538	0.6455	0.7552	0.4831
Thrombocytopenia	0.5327	0.4460	0.3745	0.5464	0.2917	0.5548	0.2062	0.5246	0.3235	0.6382	0.2973	0.5594
Hyperkalemia	0.4795	0.4821	0.3673	0.5988	0.2508	0.5094	0.4020	0.5125	0.2553	0.8470	0.3446	0.5948
Hypoglycemia	0.5404	0.4410	0.5416	0.4352	0.7131	0.4388	0.6688	0.4481	0.7052	0.5274	0.6386	0.5723
Hyponatremia	0.4593	0.5220	0.3189	0.6303	0.2652	0.5347	0.2844	0.4939	0.2745	0.6186	0.2437	0.6786
Anemia	0.4433	0.6100	0.2739	0.7254	0.1962	0.6877	0.2173	0.6044	0.2037	0.7722	0.2069	0.6825
Hypertension	0.4762	0.5238	0.5758	0.4707	0.7136	0.4488	0.7136	0.4222	0.6920	0.5606	0.7042	0.6804
Hyperlipidemia	0.5559	0.4203	0.5478	0.5032	0.6289	0.4077	0.6845	0.4127	0.7109	0.4014	0.6736	0.7338
Acute MI	0.5455	0.4545	0.4929	0.6379	0.7317	0.3394	0.6149	0.3436	0.6403	0.6053	0.6074	0.7335

Table 5: Uncertainty Metrics for Clinical Predictions with Proprietary GPTs. **Single** refers to the setting where answers are generated for one task at a time. **Multi** refers to the setting where multiple answers are generated simultaneously for tasks within the same clinical category. **Ensemble** refers to the approach that combines responses from multiple proprietary GPTs.

were promising, the generalizability of these methods to other distinct domains with limited data has not yet been tested. Future work may explore data from different cultures and the adaptation of these methods for broader applications beyond the current scope.

Acknowledgment

We would like to thank the three anonymous reviewers for their constructive questions and invaluable suggestions. This study is supported by GE Healthcare and is partially supported by NIH/NLM 1R01LM014239. Any opinions or conclusions in this paper are those of the authors and do not reflect the views of the funding agencies.

References

Taiga Abe, E Kelly Buchanan, Geoff Pleiss, and John P Cunningham. 2023. Pathologies of predictive diversity in deep ensembles. *arXiv preprint arXiv:2302.00704*.

Taiga Abe, E. Kelly Buchanan, Geoff Pleiss, and John Patrick Cunningham. 2024. Pathologies of predictive diversity in deep ensembles. *Transactions on Machine Learning Research*. Featured Certification.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2023. Introducing the claude-3 family.

Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen, and Philipp Berens. 2020. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical image analysis*, 64:101724.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

Chacha Chen, Junjie Liang, Fenglong Ma, Lucas Glass, Jimeng Sun, and Cao Xiao. 2021. Unite: Uncertainty-based health risk prediction leveraging multi-sourced data. In *Proceedings of the web conference 2021*, pages 217–226.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. 2020. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*.

Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. 2020. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. 2019. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. *Preprint at https://arxiv.org/abs/1912.10481*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy Miller, Danielle S Bitterman, Guanhua Chen, Anoop Mayampurath, Matthew M Churpek, and Majid Afshar. 2024. Position paper on diagnostic uncertainty estimation from large language models: Next-word probability is not pre-test probability. *medRxiv*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. 2015. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health*.

George Hripcsak, Patrick B Ryan, Jon D Duke, Nigam H Shah, Rae Woong Park, Vojtech Huser, Marc A Suchard, Martijn J Schuemie, Frank J DeFalco, Adler Perotte, et al. 2016. Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

V Jahmunah, Eddie Yin Kwee Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. 2023. Uncertainty quantification in densenet model using myocardial infarction ecg signals. *Computer Methods and Programs in Biomedicine*, 229:107308.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2024. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 34(5):2817–2825.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*.

Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. 2024. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, page 102830.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Takahiro Mimori, Keiko Sasada, Hirotaka Matsui, and Issei Sato. 2021. Diagnostic uncertainty calibration: Towards reliable machine predictions in medical domain. In *International Conference on Artificial Intelligence and Statistics*, pages 3664–3672. PMLR.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*.

Rahul Rahaman et al. 2021. Uncertainty quantification and deep ensembles. In *Advances in neural information processing systems*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*.

Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. 2024. Ohdsi standardized vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*.

Adam Rodman, Thomas A Buckley, Arjun K Manrai, and Daniel J Morgan. 2023. Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. *JAMA Network Open*.

Kaspar Rufibach. 2010. Use of brier score to assess binary predictions. *Journal of clinical epidemiology*.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. 2024. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. 2021. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. 2023. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, 14(1):5305.

Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo, et al. 2021. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6):2765–2787.

Elenore Judy B Uy. 2022. Key concepts in clinical epidemiology: Estimating pre-test probability. *Journal of Clinical Epidemiology*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. 2023a. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023b. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*.

WonJin Yoon, Shan Chen, Yanjun Gao, Zhanzhan Zhao, Dmitriy Dligach, Danielle S Bitterman, Majid Afshar, and Timothy Miller. 2025. Lcd benchmark: long clinical document benchmark on mortality prediction for language models. *Journal of the American Medical Informatics Association*.

A General performance of the CLMBR-T

We provide both the AUR score and the Accuracy score using the CLMBR-T embedding on the original EHRSHOT dataset (Wornow et al., 2023a) in table 6. We observe that though our multi-tasking framework can marginally reduce the uncertainty, it remains limited in improving the LM's performance. Still, we want to increase the trustworthiness of using (large) language models in clinical tasks. This finding inspires us to explore similar uncertainty quantification frameworks within the context of proprietary black-box models like GPTs.

	Single-tasking		**Multi-tasking**	
	Acc.	**Auc.**	**Acc.**	**Auc.**
Long Length of Stay	77.130	68.604	76.856	69.723
ICU Transfer	94.845	60.740	95.483	60.510
Thrombocytopenia	72.468	71.685	73.575	71.833
Hyperkalemia	95.010	55.972	95.042	55.204
Hypoglycemia	97.376	51.594	97.345	52.191
Hyponatremia	68.319	62.453	66.446	62.560
Anemia	86.949	84.706	86.965	82.662
Hypertension	83.148	55.927	82.114	56.411
Hyperlipidemia	82.916	56.579	83.219	56.259
Acute MI	89.939	55.319	90.503	54.655

Table 6: Performance metrics for CLMBR-T-base in single-task and multi-task settings. Acc. stands for accuracy score, and Auc. stands for Area Under the ROC Curve.

B Ablation study

We present the table for illustrations on the white-box model for each single clinical task. Here, we report the Brier score, NLL, ECE, and aECE metrics when the number of model m spans in a range of 1, 5, 10, and 50. Firstly, as the ensemble size increases from 1 to 10, we observe an overall decrease in the uncertainty scores. This suggests that using more models in the ensemble generally leads to better calibration and lower uncertainty in predictions. However, an interesting reversal of this trend is observed when the ensemble size becomes

Metric	**Brier (m = 1)**	**NLL (m = 1)**	**ECE (m = 1)**	**aECE (m = 1)**	**Brier (m = 5)**	**NLL (m = 5)**	**ECE (m = 5)**	**aECE (m = 5)**
Long Length of Stay	0.5645	1.5253	0.2575	0.2575	0.4942	0.8746	0.1661	0.1696
ICU Transfer	0.1388	0.4266	0.053	0.0527	0.1013	0.3291	0.0425	0.0411
Thrombocytopenia	0.7212	1.7921	0.3213	0.3211	0.5888	0.9668	0.1986	0.1986
Hyperkalemia	0.5006	0.6939	0.4809	0.4796	0.0496	0.1772	0.0252	0.0272
Hypoglycemia	0.5002	0.6935	0.4861	0.4857	0.0395	0.1364	0.0183	0.0219
Hyponatremia	0.6345	1.4092	0.2727	0.2706	0.516	0.7986	0.1357	0.1357
Anemia	0.7079	1.6924	0.3162	0.3162	0.5613	0.9057	0.1748	0.1748
Hypertension	0.2763	0.8392	0.1149	0.1134	0.2587	0.5458	0.0905	0.0893
Hyperlipidemia	0.2495	0.8229	0.1157	0.1111	0.2187	0.5106	0.0825	0.081
Acute MI	0.0769	0.248	0.0315	0.0282	0.069	0.1676	0.0302	0.0226
Metric	**Brier (m = 10)**	**NLL (m = 10)**	**ECE (m = 10)**	**aECE (m = 10)**	**Brier (m = 50)**	**NLL (m = 50)**	**ECE (m = 50)**	**aECE (m = 50)**
Long Length of Stay	0.4702	0.7053	0.1089	0.1145	0.4777	0.6706	0.0458	0.0577
ICU Transfer	0.1041	0.2827	0.0407	0.0395	0.1849	0.3296	0.1111	0.1117
Thrombocytopenia	0.5543	0.868	0.1728	0.1728	0.5037	0.7025	0.0737	0.0742
Hyperkalemia	0.0444	0.1716	0.0208	0.0227	0.0482	0.1322	0.0218	0.0283
Hypoglycemia	0.0332	0.1249	0.0149	0.0183	0.0363	0.1032	0.0156	0.0243
Hyponatremia	0.5057	0.7415	0.1145	0.1145	0.4923	0.6863	0.0388	0.0387
Anemia	0.5391	0.8062	0.1504	0.1504	0.5203	0.7187	0.0951	0.0951
Hypertension	0.2708	0.4673	0.072	0.079	0.3985	0.5891	0.1921	0.1921
Hyperlipidemia	0.2243	0.4262	0.0664	0.0643	0.3383	0.5231	0.1725	0.1725
Acute MI	0.0715	0.1572	0.0182	0.0254	0.1658	0.3215	0.2057	0.2057

Table 7: Metrics for ensembling methods across a various number of ensembles decoding models. Each column shows metrics for different model ensemble sizes (m = 1, 5, 10, 50).

very large (m = 50). At this point, the uncertainty scores start to increase again. This could indicate that while adding more models to the ensemble initially improves prediction reliability, there might be a point of diminishing returns where the addition of too many models leads to increased variability or overfitting to specific aspects of the training data, thereby increasing overall uncertainty.

C Implementation Challenges

We separately discuss the implementation challenges for white-box LMs and black-box LLMs. For the Bert-based model used for clinical predictions, we believe a general challenge (or uncertainty) is to generate the embeddings for a sequence of medical tokens. We adopted CLMBR-T, which pre-trained on 2.57 million deidentified patients' medical token sequences, for the clinical tasks. Unlike the general BERT-based model to perform classification tasks, the vector output corresponding to the [CLS] token is usually used to embed a sentence. The medical code sequence does not contain a [CLS] token. (Wornow et al., 2023a) proposes to use the representation of the last medical tokens to represent a sequence of medical codes. However, there is currently no consensus on the best method to represent sequences of medical codes. For proprietary LMs such as GPT-4, we believe the biggest real-world implementation challenge is the context length of the medical event descriptions converted from the medical code sequences. Since we cannot access the internal model parameter, we can only adopt post-hoc methods to analyze the repeatedly generated set of answers; therefore, with limited budgets, the number of experiments that can be performed is limited.