

Learning Personalized Prompts for Healthcare Guidance

Ruize Shi^{1*}, Hong Huang^{1*†}, Wei Zhou^{1*}, Kehan Yin^{1*}, Kai Zhao^{1,2}, Yun Zhao^{1,2,3}

¹Huazhong University of Science and Technology

²Tongji Medical College

³Maternal and Child Health Hospital of Hubei Province

{rzshi, honghuang, weizhou2021, kehanyin, kai_zhao, 2019fy0004}@hust.edu.cn

Abstract

The rapid development of *large language models* (LLMs) has transformed many industries, including healthcare. In practice, hospitals and patients increasingly seek LLM-based systems capable of interpreting personal health records and providing healthcare guidance. However, existing approaches mainly rely on general medical knowledge and often fail to account for individual variability, limiting their ability to provide personalized guidance. To address this, we propose *personalized prompt learning* (PPL), a framework that learns individualized prompts to guide LLMs in generating personalized healthcare recommendations. PPL constructs initial personalized prompts by leveraging both self-informed patient information and peer-informed signals derived from clinically similar cases. These prompts are then refined using *reinforcement learning* (RL) to better align the generated responses with physician recommendations written for each patient. PPL operates with hard prompts, enabling seamless integration with proprietary LLMs without modifying the underlying models. We evaluate PPL on real-world obstetrics and gynecology data. The results show that our approach produces more personalized healthcare guidance and wins 97 out of 100 comparisons in expert evaluation, demonstrating its potential for broader healthcare applications. Our code is publicly available at <https://github.com/CGCL-codes/PPL>.

1 Introduction

In recent years, the explosive growth of generative artificial intelligence has profoundly impacted various industries, drawing widespread attention to

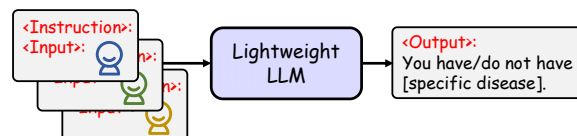


Figure 1: Prior works on personalized medical LLMs often target single disease settings, where personalization mainly comes from differences in patient inputs while overlooking personalization at the prompt level.

large language models (LLMs) (Fang et al., 2024; Qu et al., 2025; Zhao et al., 2026). LLMs have been applied in many fields, including healthcare, where they provide medical knowledge and valuable insights (Thirunavukarasu et al., 2023; Shah et al., 2023). Increasingly, patients and healthcare providers rely on LLM-based systems to interpret personal health records and obtain medical guidance (Lan et al., 2025; Zheng et al., 2025).

This growing adoption highlights an urgent demand for personalization in healthcare guidance with LLMs, as both patients and healthcare providers require guidance tailored to individual needs (Zhang et al., 2024; Liu et al., 2025). While some studies claim to have developed such LLMs (Yang et al., 2023; Zhu et al., 2023; Zhang et al., 2024), these models are typically restricted to a single disease and personalize only based on limited variations in patient data, as shown in Figure 1. Specifically, they focus on fine-tuning (Li and Liang, 2021; Lester et al., 2021) lightweight LLMs (Touvron et al., 2023; GLM et al., 2024) with knowledge about a specific disease, without the ability to adapt to individual patients. Thus, prior approaches provide only limited forms of personalization and cannot fully capture patient level variability in real-world healthcare scenarios.

As such, there remains a pressing need for effective mechanisms to enable personalization in LLM healthcare guidance, raising two key challenges: (1) *How to support multiple diseases while adapt-*

*Authors are affiliated with the National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology.

†Hong Huang is the corresponding author.

ing to individual patients? Most existing works rely on models designed for a single disease and provide only limited personalization based on variations in patient inputs. In reality, patients present with a wide range of conditions. Thus, enabling LLMs to provide guidance across diverse diseases while still tailoring responses to individual patients remains challenging. (2) *How can personalization be achieved automatically?* Current methods typically extract limited patient information and insert it into fixed prompt templates to guide LLM outputs. Such approaches cannot easily adapt to evolving patient records, contextual differences, or variations across underlying LLMs. Achieving fine-grained personalization therefore requires methods that can automatically refine and adjust the prompts used to guide the LLM for each patient.

To this end, we propose *personalized prompt learning* (PPL), a framework that learns personalized prompts to guide LLMs in generating tailored healthcare guidance. PPL analyzes patient history to extract personalized information and incorporates insights from similar patients via collaborative filtering. Using self-informed and peer-informed insights, it constructs coarse-grained personalized prompts, which are then refined through *reinforcement learning* (RL), where physician replies for each patient serve as the optimization signal. The final prompts are fed into a high-quality LLM to guide the generation of customized responses. Note that prompt refinement considers the characteristics of the downstream LLM, and the personalized prompts are hard prompts (Gu et al., 2023; Schulhoff et al., 2024). Thus, PPL can be reused across different LLMs and seamlessly integrated with proprietary LLMs with extensive medical knowledge. Our contributions are summarized as follows:

- We study how personalized prompts can improve healthcare guidance generated by LLMs. While previous methods achieve only basic personalization, our method enables more refined guidance customized to individual patients.
- We propose a framework named PPL, which uses RL to automatically adjust prompts based on patient information and downstream LLMs, guiding proprietary LLMs with extensive medical knowledge to offer tailored responses.
- Experiments on real-world obstetrics and gynecology data show that PPL significantly enhances personalization in proprietary LLMs and outperforms fine-tuned LLMs, as confirmed by

both automatic metrics and expert evaluation.

2 Related Work

2.1 Personalized LLMs

Recently, personalization in LLMs has attracted attention (Chen et al., 2024; Liu et al., 2025), particularly in recommendation systems (Wu et al., 2024). Some researchers focus on combining user history to improve recommendation quality (Woźniak et al., 2024; Shang et al., 2024; Shen et al., 2024), while others leverage reinforcement learning from human feedback to understand users’ needs and enhance personalization (Li et al., 2024b; Jang et al., 2023). However, these methods often rely on item embedding, making them difficult to apply in healthcare. Meanwhile, many LLMs have been applied in the healthcare domain and have shown promising results (Aththanagoda et al., 2025; Lan et al., 2025; Zheng et al., 2025). Some studies attempt to develop personalized medical LLMs, yet most focus on advice or treatment for specific diseases (Benary et al., 2023; Yang et al., 2023; Stade et al., 2024). Specifically, they fine-tune LLMs to learn general treatment strategies for a specific disease, achieving personalization at the disease level rather than tailoring prompts to individual patients. Moreover, they usually fine-tune lightweight LLMs, making limited use of proprietary LLMs with more extensive medical knowledge.

2.2 Hard Prompt Optimization

Early work (Shin et al., 2020) on hard prompt optimization is conducted on pre-trained models. Later, GrIPS (Prasad et al., 2023) and FluentPrompt (Shi et al., 2023) are applied to GPT-2 (Radford et al., 2019), where GrIPS is a gradient-free approach operating at phrase level, while FluentPrompt uses Langevin dynamics. PEZ (Wen et al., 2023) optimizes hard prompt through soft prompt as intermediaries. Several works explore modifying hard prompts through RL (Kaelbling et al., 1996; Wang et al., 2022), and RLPrompt (Deng et al., 2022) is among the earliest efforts. Based on RLPrompt, PIN (Choi et al., 2024) enhances prompt interpretability by sparse Tsallis entropy regularization (Lee et al., 2018). PRewrite (Kong et al., 2024) further uses a retrieval strategy to determine the most effective prompt. Rewriter (Li et al., 2024a) combines supervised learning with RL to rewrite prompts. Recent work (Wang et al., 2025) explores optimizing prompts in continuous representation

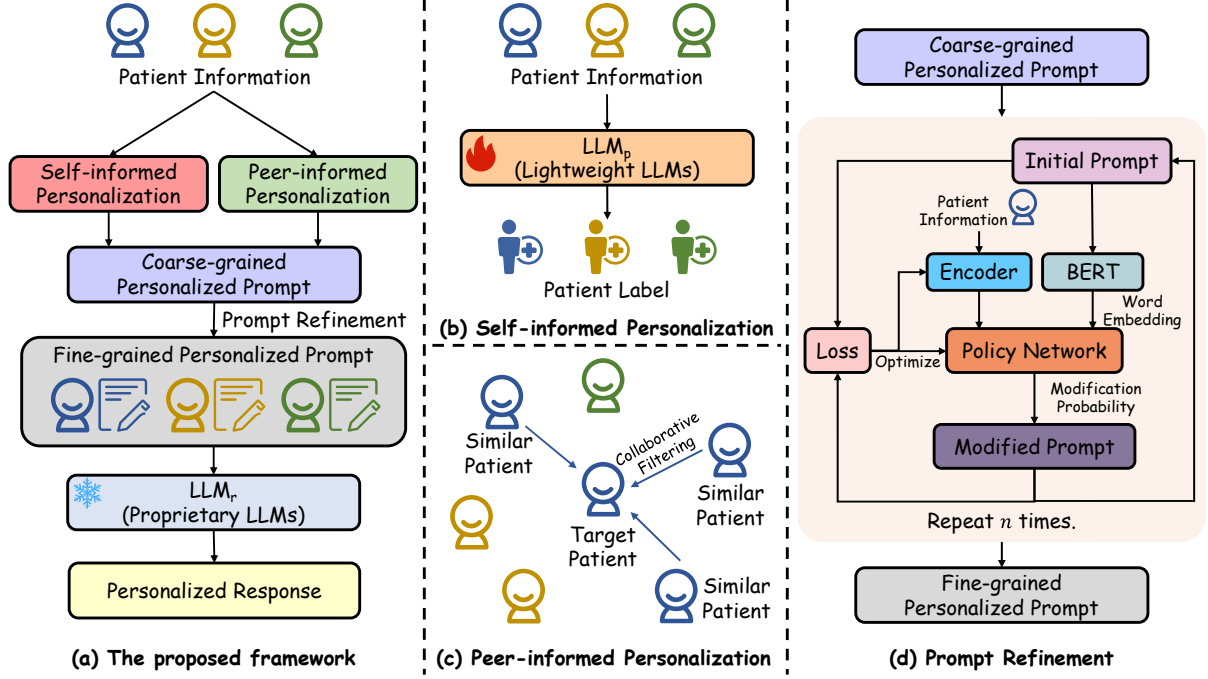


Figure 2: Overview of PPL. The framework constructs personalized prompts using patient records and similar cases, and refines them through RL to guide LLMs in generating personalized responses.

spaces. However, these methods focus on optimal prompts for specific tasks, rather than designing personalized prompts tailored to individual users.

3 Method Preparation

We denote a healthcare dataset as $\mathcal{H} = \{\mathcal{X}, \mathcal{Y}, \mathcal{R}\}$, where $\mathcal{X} = (x_1, x_2, \dots, x_N)$ contains N patients, $\mathcal{Y} = (y_1, y_2, \dots, y_N)$ denotes patient labels, such as diseases, or a normal state, and $\mathcal{R} = (r_1, r_2, \dots, r_N)$ is physician recommendations written for each patient, which serve as reference responses and reflect personalized healthcare guidance. Each patient $x_i \in \mathbb{R}^{N_{x_i} \times N_m}$, where N_{x_i} is the health check counts and N_m denotes the health examination metrics. Since patients may have different numbers of health checks, we use subscripts for differentiation.

Based on this, we extract patient specific information from health records and leverage signals from similar patients. RL is then used to refine the prompt so that the generated responses better align with physician recommendations, thereby achieving patient level personalization.

Personalized prompt is $P = (p_1, p_2, \dots, p_{N_p})$, where p_i represents a word. The prompt is represented in textual form, allowing it to be compatible with closed-source LLMs without modifications to their architectures. Furthermore, $P_i \neq P_j, i \neq j$

for different patients, meaning that each patient is associated with a distinct prompt. In addition, the prompt length may vary across patients.

4 The Proposed Method

We provide a detailed introduction to our proposed PPL. As illustrated in Figure 2, PPL guides LLMs to generate personalized responses via tailored prompts. PPL first employs a predictor to extract personalized information from the patient medical records (self-informed personalization). It then retrieves information from similar patients to further enrich the personalization (peer-informed personalization). Using both sources of information, PPL generates an initial coarse-grained personalized prompt. Through RL, these prompts are iteratively refined into fine-grained prompts tailored to each individual, where the reward encourages the generated responses to align with tailored responses written by physicians for each patient. The proposed PPL does not modify the parameters or architecture of the underlying LLM, enabling seamless integration with proprietary LLMs that possess extensive medical knowledge.

4.1 Extraction of Personalized Information

PPL leverages personalized information to build tailored prompts that guide LLMs in generating individualized responses. To achieve this, we extract

<Instruction>: Please evaluate the condition of patient based on its health examination indicators from different dates.

<Input>: The health examination indicators for patient include ... (depending on the dataset.)
The health data for the patient is $[\mathcal{X}]$.

<Output>: Your current condition is $[\mathcal{Y}]$.

Figure 3: Example prompt for the predictor.

personalized information from two complementary sources: self-informed personalization derived from the target patient’s historical health records, and peer-informed personalization obtained from clinically similar patients.

Self-informed Personalization. In our work, a patient $x_i \in \mathbb{R}^{N_{x_i} \times N_m}$ consists of multiple health check results from different dates, and the number of checks varies for each individual. To efficiently handle this variable-length data, we use LoRA (Hu et al., 2022) to fine-tune LLMs as predictors, denoted as LLM_p . Specifically, given a patient’s health check data and label, we format the data as shown in Figure 3 for instruction tuning.

After fine-tuning, LLM_p is able to make predictions based on the historical health check data of patient $\mathcal{X}$, which can be formally defined as:

$$\hat{\mathcal{Y}} = LLM_p(\mathcal{X}), \quad (1)$$

where $\hat{\mathcal{Y}}$ denotes the predicted label. Despite varying numbers of health checkups, the historical health data can be treated as time series. We evaluate several predictors, including traditional time series methods and LLMs (see Appendix A.3), and ultimately adopt LLaVA1.5-7B (Liu et al., 2024) as our predictor, since it delivers the best results even though it is primarily a multimodal LLM.

Peer-informed Personalization. To further enrich personalized information, in addition to directly using the original patient data for predictions, PPL extracts information from similar patients. First, PPL utilizes an encoder to map the patient data into the same space. Formally, for each patient x_i , it can be represented as follows:

$$\hat{x}_i = \text{encoder}(x'_i), \quad (2)$$

$$x'_i = \text{padding}\left(\sum_{j=1}^{N_{x_i}} \|x_{ij}\|\right), \quad (3)$$

where $\text{padding}(\cdot)$ refers to filling in missing values, as the number of health checkups N_{x_i} varies between patients, and we apply zero padding here for simplicity. $\hat{x}_i \in \mathbb{R}^{1 \times d}$ denotes the mapped data, and d indicates the mapped space dimension. Furthermore, x_{ij} represents the j -th health check result of patient x_i , and $\|$ signifies the concatenation. The primary purpose of the encoder is to standardize the dimensions of the input data. To ensure efficiency, we employ the *multi-layer perceptron* (MLP) (Rosenblatt, 1958) here.

PPL then leverages the encoded data to retrieve similar patients to the target patient based on collaborative filtering (Linden et al., 2003). In detail, since the labels in the training set are known, we calculate the cosine similarity between the target patient x_i and each patient x_j in the training set, which is computed as follows:

$$w_{ij} = \frac{\hat{x}_i \cdot \hat{x}_j}{\|\hat{x}_i\|_2 \cdot \|\hat{x}_j\|_2}, \quad (4)$$

where w_{ij} indicates the similarity between patients x_i and x_j , and $\|\cdot\|_2$ is the Euclidean norm. PPL leverages the top- k similar patients to enhance the target patient’s personalized information for constructing a tailored prompt. k is a hyper-parameter, which we explore in detail in Section 5.6.

4.2 Generation of Personalized Prompt

Based on the self-informed and peer-informed personalized information, PPL constructs tailored prompts to guide LLMs in generating customized responses. We first format and create an initial prompt, then refine this prompt through RL, ultimately producing unique prompts for each patient.

Coarse-grained Personalized Prompt. Similar to previous works, our initial personalized prompt includes background and input data, as illustrated in Figure 4. Specifically, the background part provides an overview of the input data and the downstream task, helping the LLM to understand the context and objectives. The input information is then populated with the personalized data extracted earlier, covering both the target patient’s prediction outcomes and the recommendations from similar patients. Note that the obtained initial prompts already incorporate a coarse level of personalization since the input data is tailored to each individual, resulting in unique prompts for each patient.

Fine-grained Personalized Prompt. To achieve a finer level of personalized prompts, we employ RL to refine the initial prompt. We design a word-level

<Instruction>: You are an assistant with expertise in medical knowledge.

<Input>: I am a patient. Please provide advice based on my prediction results and the data from patients similar to me.

Based on my previous checkup results, the medical deep learning model predicts that my current condition is [Y]. In addition, the hospital analyzed the similarity of physical examination indicators and identified information about past patients with similar results to mine. Their discharge diagnoses are as follows (similarity range: 0-1):

The diagnosis result is [Y], with a similarity score of [W].
... (with a total of [k] entries.)

Figure 4: Example of the coarse-grained personalized prompt. The labels for similar patients are ground truth values, as these records come from past cases.

policy network that adjusts selected words to better guide LLMs in generating personalized responses. Since each initial prompt varies in length, PPL operates at the word level, unlike previous works that treat the prompt as a single fixed input.

We formulate optimization as a Markov Decision Process (Bellman, 1957). PPL performs n editing steps on the initial personalized prompt, where n is a hyper-parameter controlling refinement depth. We define the state set as $\mathcal{S} = \{s_0, s_1, \dots, s_{n-1}\}$, where $s_i, i \in [0, n-1]$ denotes the personalized prompt after the i -th optimization. Moreover, s_0 is the initial state, i.e., the initial prompt (Figure 4). After that, PPL introduces a policy network $f(\cdot)$ to determine the modification probability $\pi(e|s)$ for each word under a given state s . Formally, this can be described as follows:

$$\pi(e|s) = \text{softmax}(f(E)), \quad (5)$$

$$E = \text{BERT}(s) = (e_1, e_2, \dots, e_{N_s}), \quad (6)$$

where $\text{softmax}(\cdot)$ denotes normalization and E is the embedding of the personalized prompt at state s . The personalized prompt is textual, thus, PPL first encodes it using BERT (Devlin et al., 2019). Moreover, e indicates word embeddings, where each embedding has the same dimension, in contrast, the number of words N_s across states may vary. PPL samples a word from the probability distribution $\pi(e|s)$ and removes it to obtain the next state. This editing process allows the policy network to explore different prompt variants and select those that better align the generated responses with physician guidance. This process is repeated n times, each time deleting only one word, until

reaching the final state s_{n-1} , which serves as the personalized prompt P .

Nonetheless, the transition probability relies heavily on word embeddings, resulting in limited personalization. In detail, while these embeddings include contextual information, they still lack sufficient granularity to differentiate between individual patients. To address this, PPL incorporates patient information along with sentence representations. Formally, Eq.(5) can be reformulated as follows:

$$\pi(e_i|s) = \frac{\exp(f(e_i \parallel \bar{E} \parallel \hat{x}))}{\sum_{j=1}^{N_s} \exp(f(e_j \parallel \bar{E} \parallel \hat{x}))}, \quad (7)$$

$$\bar{E} = \text{mean}(E) = \frac{e_1 + e_2 + \dots + e_{N_s}}{N_s}, \quad (8)$$

where $i \in [1, \dots, N_s]$ and $\parallel$ denotes concatenation. $\bar{E}$ captures the global semantics of the prompt, while $\hat{x}$ denotes the patient representation obtained from the encoder in Eq.(2). With this refinement, the prompt modification considers patient information and overall semantics, enabling the generation of a more personalized prompt. For simplicity, we still employ the MLP as the policy network.

After n iterations, we obtain the personalized prompt P . We define the reward using BERTScore (Zhang et al., 2020), which is as follows:

$$\text{Reward} = \text{BS}(\hat{r}, r) - \text{BS}(\hat{r}_0, r), \quad (9)$$

$$\hat{r} = \text{LLM}_r(P), \quad (10)$$

where LLM_r indicates the LLM used to generate replies and $\hat{r}$ is its response to the personalized prompt, $\hat{r}_0 = \text{LLM}_r(s_0)$ is the response to the initial personalized prompt. Moreover, $\text{BS}(\cdot)$ stands for BERTScore, and r denotes the reference responses provided by physicians for each patient, which reflect personalized medical guidance in clinical practice. The reward is computed once based on the final refined prompt by comparing the generated response with the physician reference response using BERTScore. Since physician responses are personalized for each patient, this reward encourages the learned prompts to capture guidance for individual patients. The reward function also considers the downstream LLM, meaning that the personalized prompt is adaptively revised in alignment with LLM_r . As the personalized prompt is in text format, we avoid any modifications to LLM_r , enabling compatibility with proprietary LLMs. In this case, the loss function is defined as:

$$\mathcal{L} = - \sum_{i=0}^{n-1} \log(\pi(e_j|s_i)) \cdot \text{Reward}, \quad (11)$$

where $\pi(e_j|s_i)$ is the probability of deleting word e_j in the state s_i . Moreover, $j \in [1, N_{s_i}]$ denotes a chosen index, following the probability distribution $\pi(e|s_i)$. Finally, we optimize our proposed PPL by Adam (Kingma and Ba, 2015).

We further provide the pseudo-code of PPL, training strategies, and an interactive web interface in Appendix A.1.

5 Experiments

5.1 Experimental Setup

Datasets. We collect real-world obstetrics data from multiple hospitals in Wuhan city and surrounding areas from 2020 to 2022. Details on data processing are provided in the Appendix A.2. The final dataset consists of 38,817 records for 2,373 pregnant patients, with each record containing 35 examination metrics. The patient with the most records has 45 entries, and each patient is assigned a label and receives medical recommendations, with a total of 12 categories. We partition the data by year: data from 2021 and earlier form the training set, the first half of 2022 as the validation set, and the remaining data as the test set. This dataset is suitable for our task since it contains longitudinal health records together with medical recommendations written by physicians for each patient, which are used both as the RL optimization signal and as reference responses in evaluation. We use prenatal examination records available before delivery and retrieve similar patients from the training set, avoiding future information leakage.

Baselines. We evaluate PPL on multiple proprietary LLMs and compare it with several fine-tuned lightweight LLMs to examine whether prompt learning on strong proprietary LLMs can outperform directly fine-tuning lightweight LLMs. Our goal is not to compare model capacity across different scales, but to evaluate prompt learning as a practical alternative when proprietary LLMs cannot be fine-tuned. We do not compare with prompt optimization methods, since they mainly optimize shared prompts for tasks rather than personalized prompts. The baselines are as follows:

- **Proprietary LLMs.** We evaluate PPL on several LLMs, including Gemini1.5-pro (Team et al., 2024), GLM4 (GLM et al., 2024), GLM4-plus, GPT3.5-turbo (Brown et al., 2020), and GPT4 (Achiam et al., 2023).
- **Lightweight LLMs.** We compare several LLMs that are fine-tuned on the training data, includ-

```
<Instruction>: You are an assistant with expertise
in medical knowledge.
***
<Input>: I am a patient. Please assess my condition
and provide advice based on my medical test results.
The health examination indicators include ...
(depending on the dataset.)
Here are my health results from different dates: [x].
***
<Output>: [r].
```

Figure 5: Example prompt for evaluation, where <Output> is omitted for proprietary LLMs.

ing Llama3-8B (Dubey et al., 2024), GLM4-9B (GLM et al., 2024), Qwen2-7B (Bai et al., 2023), and LLaVA1.5-7B (Liu et al., 2024).

Experimental settings. We evaluate PPL on a server with an Intel Xeon Gold 5117 CPU, a Tesla V100 GPU (32 GB), and 256 GB RAM. The server uses Ubuntu 18.04 with CUDA 12.1, and our code is implemented in PyTorch 2.1.0. To reduce randomness, we report the average results of five runs.

For hyper-parameter settings, both the encoder and policy network in PPL are MLPs, each with hidden layer dimensions of 256 and a dropout rate of 0.4. Specifically, the encoder has 2 layers with an output dimension of 128, while the policy network has 3 layers with an output dimension of 1. The number of similar patients k and the number of prompt modification steps n are both set to 10, with a learning rate of 0.005. For PPL’s predictor and other fine-tuned LLMs, we follow a representative work (Zheng et al., 2024).

5.2 Performance Study

We evaluate whether PPL improves alignment between generated responses and physician recommendations written for each patient, and compare with fine-tuned lightweight LLMs. To this end, we use the prompt shown in Figure 5. For proprietary LLMs, we directly evaluate on the test data, as fine-tuning is not feasible. For lightweight LLMs, we fine-tune on the training data, followed by evaluation. We compare the outputs from each LLM with the physician reference responses, which reflect personalized guidance for each patient, using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020). These automatic metrics are mainly used to evaluate the lexical and semantic alignment between generated responses and physician reference responses. We

Table 1: Performance of PPL based on proprietary LLMs.

Model		BLEU-4	ROUGE-L	ROUGE-N		BERTScore		
				1	2	Precision	Recall	F1
Gemini 1.5-pro	before	6.49	15.46	27.57	6.19	62.79	65.15	63.94
	after	10.65	22.39	37.01	13.67	69.54	70.75	70.13
GLM4	before	8.39	19.46	30.93	9.95	63.06	64.95	63.99
	after	9.42	21.31	34.67	11.57	66.80	69.51	68.12
GLM4-plus	before	7.79	15.45	33.15	10.48	59.43	63.26	61.26
	after	9.04	18.47	36.22	11.96	65.20	69.58	67.31
GPT3.5-turbo	before	8.73	20.08	29.48	9.69	65.56	64.90	65.19
	after	9.39	22.84	32.41	10.49	69.25	68.59	68.89
GPT4	before	7.76	15.49	28.62	9.38	61.03	64.59	62.75
	after	10.88	22.69	39.03	14.30	69.80	71.21	70.50

Table 2: Comparison with lightweight LLMs.

Model	BLEU-4	ROUGE-L	ROUGE-N		BERTScore		
			1	2	Precision	Recall	F1
Llama3-8B	9.56	23.52	38.44	14.12	65.61	65.93	65.77
GLM4-9B	8.79	19.72	36.40	12.02	64.37	64.29	64.33
Qwen2-7B	8.03	19.78	36.64	12.18	64.38	64.22	64.30
LLaVA1.5-7B	7.45	16.61	33.26	13.59	61.30	63.29	62.28
PPL	10.88	22.69	39.03	14.30	69.80	71.21	70.50

further conduct human evaluation in Section 5.5 to assess the quality of personalized healthcare guidance from an expert perspective.

From Table 1, it is evident that PPL consistently improves the performance of general LLMs. By leveraging personalized prompts, all evaluated LLMs exhibit performance improvements exceeding 10% on several metrics. In addition, the GPT4-based PPL performs best, thus, we use GPT4 as the backbone in subsequent experiments.

To provide a more comprehensive evaluation, we also compare PPL with fine-tuned LLMs, which are presented in Table 2. We observe that PPL achieves the best overall performance in this comparison. However, it falls short on ROUGE-L, which is expected since this metric relies heavily on word-level overlap. For the lightweight LLMs, their vocabulary is closer to the reference responses since they are trained directly on the physician replies in the training data. In this task, BERTScore is particularly important since it measures semantic similarity between generated responses and physi-

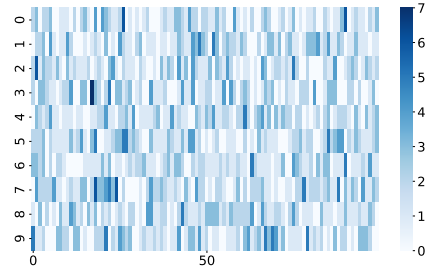


Figure 6: Heatmap of word modification counts, where the X-axis represents prompt word indices and the Y-axis represents refinement iterations.

cian recommendations. Notably, PPL achieves the highest BERTScore, indicating stronger semantic alignment with physician guidance.

5.3 Personalization Study

To illustrate how PPL generates personalized prompts for different patients, we analyze the word modification behavior of the policy network on the test set. Specifically, we record the indices of

Table 3: Examples for the personalization study of PPL.

Case	Details
Case 1	Similar cases: Uterine scarring and gestational diabetes. Response: “Your examination results are similar to multiple prior cases, especially those involving uterine scarring and gestational diabetes...”
Case 2	Similar cases: Gestational hypertension. Response: “Although the prediction result suggests premature rupture of membranes, gestational hypertension may also need attention...”
Case 3	Similar cases: Premature rupture of membranes. Response: “Among prior cases with examination results similar to yours, premature rupture of membranes appears frequently...”

All cases share the same predicted condition (PROM). “Similar cases” summarize the dominant diagnosis categories among retrieved similar patients. Response segments are directly extracted from PPL’s outputs, with irrelevant parts omitted.

Table 4: Ablation study of PPL.

Model				BLEU-4	ROUGE-L	ROUGE-N		BERTScore		
ID	SP	PP	PR			1	2	Precision	Recall	F1
1	✗	✓	✓	8.42	18.39	34.01	10.85	61.10	65.01	62.99
2	✓	✗	✓	9.29	21.38	34.54	11.51	65.92	68.61	67.24
3	✓	✓	✗	7.91	17.68	33.07	10.39	63.62	66.09	64.83
PPL	✓	✓	✓	10.88	22.69	39.03	14.30	69.80	71.21	70.50

ID indicates different ablation variants, and ‘✓’ and ‘✗’ indicate whether a module is included.

words removed during each refinement iteration. Due to space constraints, we report the modification counts for the first 100 words.

As shown in Figure 6, the frequency of word modifications varies significantly across different indices. This indicates that the policy network adapts the refinement process according to the prompt context and patient information, producing different prompts for different patients.

To further examine whether such prompt differences lead to personalized healthcare guidance, we conduct a targeted case study using three patients with the same predicted condition, i.e., *premature rupture of membranes* (PROM). Table 3 presents response segments generated by PPL that reflect the influence of peer-informed information. Although their predicted labels are identical, their examination histories lead to different retrieved similar patients, providing different peer-informed information for generating healthcare guidance.

These cases show that PPL does not simply generate guidance based on the predicted condition. Instead, it incorporates patient examination histories and peer-informed information, resulting in different healthcare guidance even for patients with the same predicted condition.

5.4 Ablation Study

We further assess the contributions of each module in PPL. In detail, we examine the roles of *self-*

Table 5: Human evaluation study of PPL.

Model	GLM4	GPT3.5-turbo	PPL
Vote Share	5.56	8.32	86.12
Wins	1	2	97

informed personalization (SP), *peer-informed personalization* (PP), and *prompt refinement* (PR) in guiding LLMs to generate personalized responses.

Table 4 shows that all modules of PPL contribute to its performance, confirming the effectiveness of the overall design. Compared with the other variants, the impact of PP (Variant 2) is smaller, since PP relies on the similarity of medical records, which may not always reflect true clinical similarity. As a result, even without this component, the model can still achieve reasonable performance.

5.5 Human Evaluation Study

To validate the practical utility of PPL, we conduct a human evaluation study. Given the high cost of manual assessment, we randomly sample 100 cases from our dataset and select GLM4 and GPT3.5-turbo as representative proprietary LLM baselines. The evaluation was conducted by 25 evaluators with medical backgrounds, including 7 licensed obstetricians, 3 PhD students, and 15 medical master students. For each case, participants are presented with anonymized outputs from different

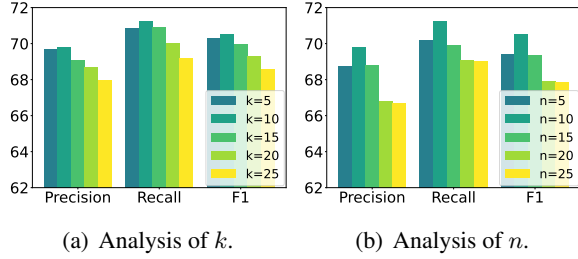


Figure 7: Hyper-parameter study. The three groups correspond to BERTScore precision, recall, and F1, and each group contains results for different k or n .

LLMs in randomized order and asked to identify the most appropriate healthcare guidance according to clinical relevance, usefulness, and clarity. The final preference is determined by majority voting among the participants.

As shown in Table 5, PPL is preferred in 97 out of 100 cases, outperforming the baselines and demonstrating the effectiveness of personalized healthcare guidance. Additional case studies are provided in Appendix A.4.

5.6 Hyper-parameter Study

We also investigate the impact of hyper-parameters on performance, focusing on the number of similar patients k and prompt modification iterations n .

The results in Figure 7 suggest that as k or n increases, the performance of PPL first improves and then declines. Increasing k beyond a certain point may introduce patients with lower similarity, whose records may introduce irrelevant or conflicting information, thereby weakening personalization. Likewise, for n , excessive modifications to the prompt may distort its semantics and reduce the reliability of the guidance. Moreover, n has a more significant impact on performance, which is consistent with the importance of prompt refinement observed in the ablation study. Based on these observations, we set k and n to 10 in our experiments.

6 Conclusion

We study personalization for LLMs in healthcare and propose the PPL framework. PPL constructs prompts tailored to each patient from their health data and refines these prompts through RL to better align with physician recommendations. The resulting personalized prompts guide LLMs to generate customized responses. Moreover, PPL can be applied to various LLMs, including proprietary models, without directly sending raw patient records.

Extensive experiments, evaluated using both automatic metrics and human experts, demonstrate the effectiveness of PPL and highlight its potential for advancing adaptive healthcare LLMs.

Limitations

While PPL explores personalized prompt learning for healthcare LLMs and demonstrates promising results, several limitations remain.

- **Dataset.** PPL learns personalized prompts by aligning generated responses with physician recommendations, while public healthcare datasets rarely contain both longitudinal patient records and physician guidance. We thus evaluate PPL on a single real-world obstetrics and gynecology dataset. Although this dataset reflects realistic clinical practice, it may not cover the full diversity of medical scenarios. Evaluating PPL on broader datasets from different hospitals could better assess its generalizability.
- **Prompt editing.** PPL refines prompts using a deletion-based editing strategy, which keeps the editing space stable and allows RL to explore prompt modifications efficiently. However, richer operations such as insertion or replacement could provide more flexible prompt optimization. In addition, deletion may occasionally produce prompts that appear semantically incoherent. Prior studies (Webson and Pavlick, 2022; Deng et al., 2022) report that LLMs can still generate effective responses under such prompts, possibly since LLMs interpret prompt signals differently from human readers.

Ethical Consideration

We conduct this study in collaboration with Tongji Medical College of HUST and Maternal and Child Health Hospital of Hubei Province, using fully anonymized retrospective medical records, with all identifiable patient and physician information removed prior to analysis. The dataset contains no direct personal identifiers, and all data are handled in accordance with standard privacy protection practices. Moreover, PPL avoids sending raw records to proprietary LLMs, and all data are used solely for academic research aimed at improving healthcare applications.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (No.62127808).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- AKNL Aththanagoda, KA Saneera Hemantha Kulathilake, and Nor Aniza Abdullah. 2025. Precision and personalization: how large language models redefining diagnostic accuracy in personalized medicine—a systematic literature review. *IEEE Journal of Biomedical and Health Informatics*, 30(4):3102–3117.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Richard Bellman. 1957. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684.
- Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.
- Yunseon Choi, Sangmin Bae, Seonghyun Ban, Minchan Jeong, Chuheng Zhang, Lei Song, Li Zhao, Jiang Bian, and Kee-Eung Kim. 2024. Hard prompts made interpretable: Sparse entropy regularization for prompt tuning with rl. In *Proceedings of the ACL 2024*, pages 8252–8271.
- Angus Dempster, François Petitjean, and Geoffrey I Webb. 2020. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.
- Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3369–3391.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Proceedings of the SIGIR 2024*, pages 2910–2914.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.
- Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.
- Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the ICLR 2015*.
- Weize Kong, Spurthi Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Prewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the ACL 2024*, pages 594–601.

- Wei Lan, Zhentao Tang, Mingyang Liu, Qingfeng Chen, Wei Peng, Yiping Phoebe Chen, and Yi Pan. 2025. The large language models on biomedical data analysis: a survey. *IEEE Journal of Biomedical and Health Informatics*.
- Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- Kyungjae Lee, Sungjoon Choi, and Songhwa Oh. 2018. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.
- Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024a. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, pages 3367–3378.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.
- Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024b. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. Grips: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864.
- Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, Hyojung Han, Sevien Schulhoff, et al. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.
- Nigam H Shah, David Entwistle, and Michael A Pfeffer. 2023. Creation and adoption of large language models in medicine. *Jama*, 330(9):866–869.
- Fu Shang, Fanyi Zhao, Mingxuan Zhang, Jun Sun, and Jiatu Shi. 2024. Personalized recommendation systems powered by large language models: Integrating semantic understanding and user preferences. *International Journal of Innovative Research in Engineering and Management*, 11(4):39–49.
- Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 3833–3843.
- WeiJia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2023. Toward human readable prompt tuning: Kubrick’s the shining is a good movie, and a good prompt too? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10994–11005.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

- Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral health-care: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincui Huang, Xin Xu, Bin Dai, and Qiguang Miao. 2022. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078.
- Yangkun Wang, Zihan Wang, and Jingbo Shang. 2025. Direct prompt optimization with continuous representations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652.
- Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344.
- Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025.
- Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520. IEEE.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.
- Hao Yang, Jiaxi Li, Siru Liu, Lei Du, Xiali Liu, Yong Huang, Qingke Shi, and Jialin Liu. 2023. Exploring the potential of large language models in personalized diabetes treatment strategies. *medRxiv*, pages 2023–06.
- Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. 2026. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627.
- Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2025. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 16(2):1015–1040.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.
- Lingxuan Zhu, Weiming Mou, and Rui Chen. 2023. Can the chatgpt and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *Journal of translational medicine*, 21(1):269.

Assay summary		Medical Test Results						
Assay name: Test 12345		[View Report]						
Patient ID: 123456								
Test Date: 2023-10-27								
Requester: Dr. Smith								
Test Details								
Test Name: Complete Blood Count (CBC)								
Reference Range: 12-18 months								
Results								
Parameter	Unit	Observed	Normal Range	Reference Range	Abnormal Results	Test Results	U/L	Reference
2023-10-27								
2023-10-26	mm3	250	100	60	True	5.5	3.5	5.0
2023-10-25	g/dL	12.5	12	10	True	4.5	3.5	4.0
2023-10-24	mm3	150	50	40	True	4.0	2.0	3.0
2023-10-23	mm3	180	60	50	True	3.5	2.5	3.0
2023-10-22	g/dL	10.0	10	8	True	4.0	3.0	3.5
2023-10-21	mm3	120	40	30	True	3.0	2.0	2.5
2023-10-20	g/dL	9.0	9	7	True	3.0	2.0	2.5
2023-10-19	mm3	110	30	20	True	2.5	1.5	2.0
2023-10-18	g/dL	8.0	8	6	True	2.5	1.5	2.0
2023-10-17	mm3	100	20	10	True	2.0	1.0	1.5
2023-10-16	g/dL	7.0	7	5	True	2.0	1.0	1.5
2023-10-15	mm3	90	10	5	True	1.5	0.5	1.0
2023-10-14	g/dL	6.0	6	4	True	1.5	0.5	1.0
2023-10-13	mm3	80	5	3	True	1.0	0.5	0.8
2023-10-12	g/dL	5.0	5	3	True	1.0	0.5	0.8
2023-10-11	mm3	70	4	2	True	0.8	0.4	0.6
2023-10-10	g/dL	4.0	4	2	True	0.8	0.4	0.6
2023-10-09	mm3	60	3	1	True	0.5	0.3	0.4
2023-10-08	g/dL	3.0	3	1	True	0.5	0.3	0.4

A.1 Details of PPL

We also present an interactive Web interface to demonstrate the practical use of PPL. Figure A1 shows a representative page displaying a patient’s historical medical test results across different dates. The top-left corner presents the basic information of the patient, including gender and height. The left panel allows users to switch between viewing basic information, medical test results, and recommendations generated by PPL. The interface has been briefly tested by internal users, suggesting that PPL can guide the LLM to produce tailored responses for individual patients.

We collect and process obstetrics and gynecology data from multiple hospitals in Wuhan city and surrounding areas, covering records from 2020 to 2022. The dataset includes longitudinal medical examination records of pregnant women from admission to discharge. It contains 194,345 entries, each with 302 examination indicators. Discharge diagnoses are used as labels, while physician recommendations serve as reference responses. *The use of this dataset complies with ethical standards, as described in the Ethical Consideration.*

Algorithm 1: The Proposed PPL

Output: Responses of test set $\tilde{R}$

316

Table A1: Predictor study.

Model	Micro-F1	Macro-F1
LSTM	21.60	5.51
CNN	20.89	6.42
ROCKET	36.05	20.58
Transformer	41.12	24.58
Llama3-8B	79.41	62.94
GLM4-9B	65.63	56.12
Qwen2-7B	80.05	69.86
LLaVA1.5-7B	82.58	74.13

level, albumin, total protein, indirect bilirubin, direct bilirubin, total bilirubin, alkaline phosphatase, alanine aminotransferase, prealbumin, total bile acid, large platelet count, plateletcrit, large platelet ratio, mean platelet volume, platelet distribution width, red blood cell distribution width (coefficient of variation and standard deviation), basophil count, eosinophil count, monocyte count, lymphocyte count, neutrophil count, basophil ratio, eosinophil ratio, monocyte ratio, lymphocyte ratio, neutrophil ratio, mean corpuscular hemoglobin concentration, mean corpuscular hemoglobin, hematocrit, white blood cell count, platelet count, hemoglobin, mean corpuscular volume, and red blood cell count.

Labels. The original dataset contains 455 conditions plus a "normal" category. We filter out conditions with insufficient samples and retain 12 labels: normal, uterine scar, acute fetal distress (heart type), pregnancy complicated with placental dysfunction, gestational hypertension, premature rupture of membranes, pregnancy complicated with hypothyroidism, uterine rupture, acute fetal distress (amniotic fluid type), umbilical cord entanglement, thalassemia, and gestational diabetes.

Final dataset. After removing incomplete records and retaining the selected indicators, the final dataset contains 38,817 entries from 2,373 pregnant women. The dataset includes 56 patients from 2020, 1,638 from 2021, and 679 from 2022. We split the dataset by year: data from 2021 and earlier form the training set, the first half of 2022 as the validation set, and the remaining data as test set.

A.3 Predictor Study

We further study the choice of predictor. Patient data can essentially be viewed as time series data, so we evaluate both LLMs and representative time series methods, including CNN (LeCun et al.,

1989), LSTM (Graves, 2012), ROCKET (Dempster et al., 2020), and Transformer (Vaswani et al., 2017). Since the number of medical visits varies across patients, zero padding is applied to align data.

Table A1 shows that fine-tuned LLMs consistently outperform traditional time series methods by a large margin. This may be since healthcare data are sensitive to sequence padding, which can affect the performance of time series methods. Among the LLMs, LLaVA1.5-7B achieves the best performance. Although it is a multimodal model, using only its text input already yields strong results. Thus, we adopt it as the predictor in our work.

A.4 Case Study

Here we provide several cases to illustrate how the proposed PPL guides LLMs to generate personalized responses. We compare the responses of GLM4 (GLM et al., 2024) and GPT4 (Achiam et al., 2023). Specifically, we present cases for two different patient types, a pregnant woman diagnosed with *Gestational Diabetes Mellitus* (GDM) (Figures A2, A3, and A4) and a pregnant woman with normal conditions (Figures A5, A6, and A7).

The results show that PPL can accurately identify the condition of patients (highlighted in green) and provide more tailored recommendations, demonstrating its ability to capture individual patient characteristics. In contrast, GLM4 and GPT4 tend to rely on individual numerical metrics and may fail to interpret the overall health status of patients. As a result, their responses are often more generalized. For example, in the case of the normal patient, both GLM4 and GPT4 incorrectly suggest potential conditions, which further highlights their limited ability to provide personalized healthcare guidance. These observations validate the effectiveness of PPL in generating patient-specific recommendations and suggest its potential for developing more personalized LLM-based healthcare systems.

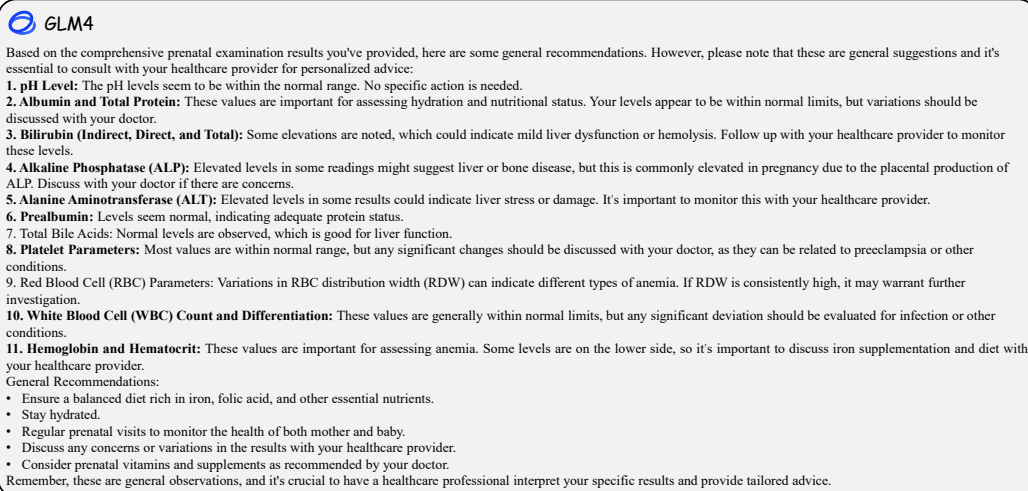


Figure A2: GLM4 response in the GDM case study.

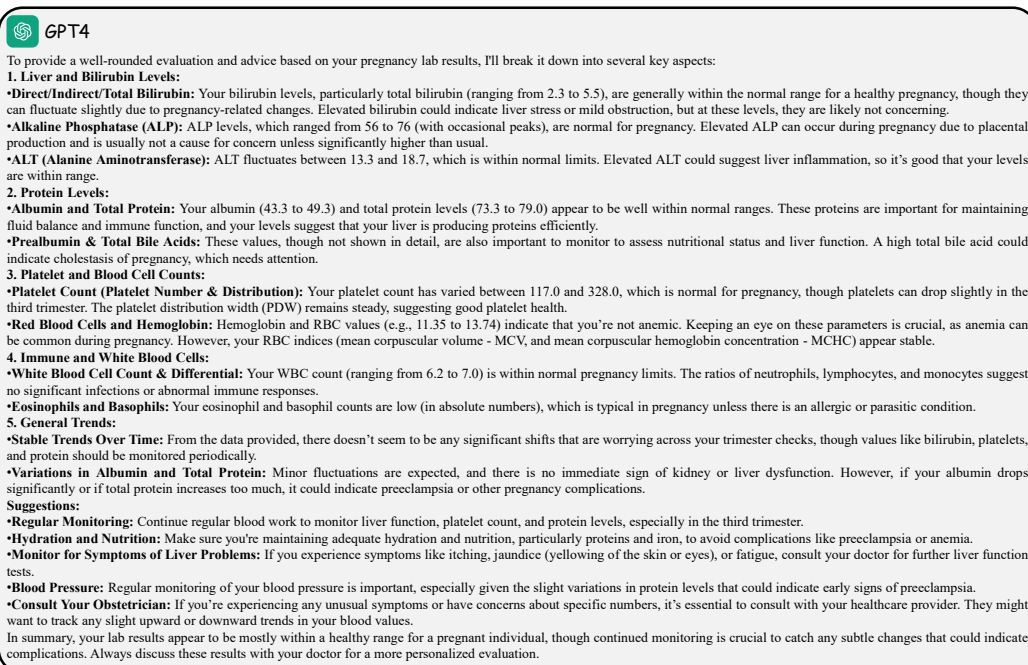


Figure A3: GPT4 response in the GDM case study.

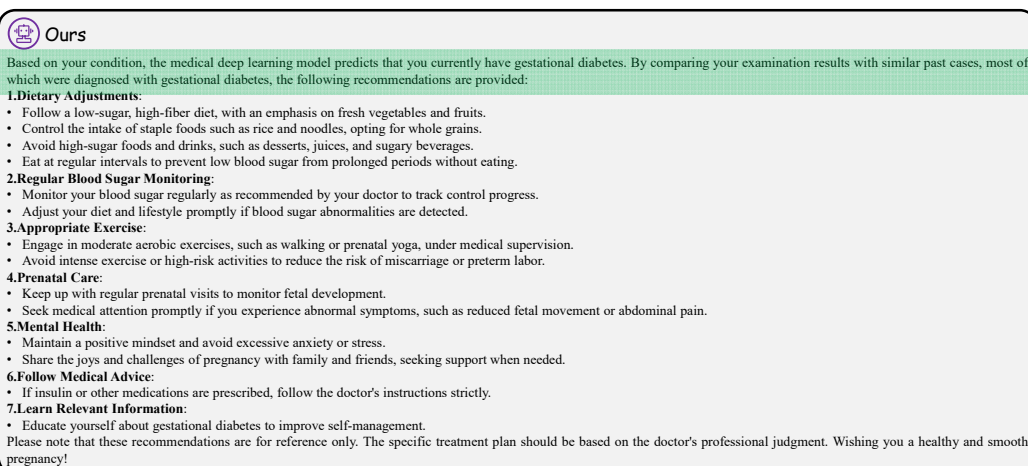


Figure A4: PPL response in the GDM case study.

GLM4

Based on the prenatal check-up results you've provided, here are some general recommendations. However, please note that these are general suggestions, and it's crucial to consult with your healthcare provider for personalized advice.

- 1. pH (6.0, 8.0):** The pH level seems to be outside the normal range for blood (7.35-7.45). This could be a lab error or an indication of an underlying issue. Please follow up with your healthcare provider.
- 2. Albumin (38.0 g/L):** This is within the normal range, indicating good protein levels.
- 3. Total Protein (74.6 g/L):** This is also within the normal range.
- 4. Bilirubin (Indirect, Direct, and Total):** Your bilirubin levels are normal, which suggests that your liver is functioning well.
- 5. Alkaline Phosphatase (38.0, 120.0 U/L):** The last value is high and could indicate liver dysfunction or other conditions. Further investigation is needed.
- 6. Alanine Aminotransferase (ALT) (26.0 U/L):** This is within the normal range, suggesting no significant liver damage.
- 7. Total Bile Acids (0.8, 1.4 $\mu\text{mol/L}$):** The last value is slightly high and could be related to gallbladder or liver function. Follow up with your doctor.
- 8. Platelet Count and Related Parameters:** Most of these values are within normal ranges, but the last measurement of platelet count ($288.0 \times 10^9/\text{L}$) is on the lower side of normal. Monitor this closely.
- 9. Hemoglobin (119.0 g/L):** This is slightly low and could indicate anemia, which is common during pregnancy. Discuss with your doctor about iron supplementation.

Red Blood Cell (RBC) Count and Hematocrit: These values are within normal ranges.

- 10. White Blood Cell (WBC) Count ($326.0 \times 10^9/\text{L}$):** This is slightly elevated and could indicate an infection or inflammation. Monitor this with your healthcare provider.

General Recommendations:

- Follow up with your healthcare provider to discuss any abnormal results.
- Ensure a balanced diet rich in iron, folic acid, and other essential nutrients.
- Stay hydrated and engage in regular, gentle exercise as recommended by your doctor.
- Attend all scheduled prenatal appointments for monitoring and support.

Remember, these recommendations are not a substitute for professional medical advice. Always consult with your healthcare provider for specific guidance related to your pregnancy.

Figure A5: GLM4 response in the normal case study.

GPT4

Based on the medical test results provided, here are some general observations and recommendations for you as a pregnant woman:

- 1. pH (Acid-Base Balance) (6.0-8.0):** Your pH levels seem to vary within the normal range (6.0 to 8.0). Maintaining an optimal pH is crucial during pregnancy to ensure proper metabolic and respiratory function. Ensure you stay hydrated and follow a balanced diet with adequate electrolytes.
- 2. Albumin (38.0 g/L) and Total Protein (74.6 g/L):** Your albumin and total protein levels appear normal. These proteins are essential for maintaining proper blood volume and pressure. Ensure you are getting sufficient protein from your diet.
- 3. Bilirubin Levels (Direct, Indirect, and Total):** Your bilirubin levels have fluctuated but appear to remain within a generally normal range (direct bilirubin: $0.5\text{--}1.9 \mu\text{mol/L}$, indirect bilirubin: $6.4\text{--}6.9 \mu\text{mol/L}$). High bilirubin can indicate liver or bile duct problems, so it is important to monitor these levels regularly during pregnancy.
- 4. Alkaline Phosphatase (ALP):** Elevated levels of ALP (such as 120 U/L) may be expected in pregnancy, as it can come from the placenta. However, persistently high levels should be investigated further to rule out liver or bone conditions.
- 5. Alanine Aminotransferase (ALT):** This enzyme is an indicator of liver function. Your ALT levels are normal, suggesting there are no immediate concerns with liver function.
- 6. Prealbumin:** Levels seem to be in a healthy range. Prealbumin is a good marker for nutritional status, and it's important to maintain proper nutrition to support both your health and the development of your baby.
- 7. Platelets and Blood Cell Parameters:**
 - Your platelet count and volume parameters (e.g., Platelet Count, Mean Platelet Volume, Platelet Distribution Width) are generally within normal ranges. However, some variability in platelet count may be observed during pregnancy, and it's important to monitor for any abnormal changes that could suggest thrombocytopenia or other conditions.
 - Red blood cell parameters, such as hemoglobin, hematocrit, and red blood cell distribution width (RDW), are important for assessing anemia. Your levels appear to be within expected ranges, but keep track of any signs of anemia (e.g., fatigue or dizziness).
- 8. White Blood Cells and Differential Count:** Your white blood cell count (WBC) is in the normal range, and the differential count (e.g., neutrophils, lymphocytes) appears balanced, indicating no signs of infection or significant immune response at this time.
- 9. Other Recommendations:**
 - **Stay hydrated:** Maintaining proper fluid intake is important for kidney function and overall health during pregnancy.
 - **Monitor for symptoms:** Be mindful of any changes in your health, such as jaundice (yellowing of the skin or eyes), fatigue, or swelling, which may indicate liver, kidney, or blood-related issues.
 - **Nutritional balance:** Ensure you are eating a balanced diet rich in proteins, vitamins (especially iron and folic acid), and minerals to support your pregnancy. If you have concerns about your diet, consider consulting with a nutritionist.
 - **Regular check-ups:** Continue regular prenatal visits to monitor your blood work and other health indicators to ensure both you and your baby are doing well.

Please consult with your healthcare provider for a more personalized interpretation of your results and for any specific concerns based on your pregnancy stage.

Figure A6: GPT4 response in the normal case study.

Ours

Based on your pregnancy prediction results and the information from past patients with similar examination results, the following recommendations are provided:

- 1. Regular Prenatal Check-ups:** Although the prediction results are normal, regular prenatal visits are still very important to monitor the health of both the mother and the baby.
- 2. Monitor Fetal Movement:** Pay attention to fetal movement. Reduced movement may be a sign of fetal distress, and medical attention should be sought promptly.
- 3. Balanced Diet:** Ensure adequate nutrition, especially important nutrients such as protein, iron, and folic acid, which are vital for both the mother and the baby.
- 4. Moderate Exercise:** Moderate exercise can improve circulation and strengthen muscles, but avoid intense physical activity.
- 5. Maintain Healthy Lifestyle Habits:** Refrain from smoking and alcohol consumption, avoid exposure to harmful substances, and maintain good sleep quality.
- 6. Monitor Thyroid Function:** Although your current prediction is normal, thyroid dysfunction is closely related to both maternal and fetal health. Regular monitoring of thyroid function is recommended.
- 7. Monitor Fetal Heart Rate:** The fetal heart rate is an important indicator of fetal health. If any abnormalities are detected, seek medical attention immediately.
- 8. Prevent Premature Rupture of Membranes:** Avoid heavy lifting, intense physical activities, and sexual intercourse. Keep the genital area clean.
- 9. Maintain a Positive Mental Outlook:** Stay optimistic and avoid excessive stress and anxiety.

Please follow the doctor's advice and integrate these recommendations into your prenatal care and daily life. If you experience any discomfort or have concerns, seek medical attention promptly. Wishing you a smooth pregnancy and a healthy baby!

Figure A7: PPL response in the normal case study.