ASSIGNMENT 2: HARMONIES OF PROMPT: UNVEILING THE ARTISTRY IN ENGINEERING

Yangkuan, Li 223040028
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
223040028@link.cuhk.edu.cn

1 INTRODUCTION

The Importance of the Task In the current medical field, pharmacists play a crucial role, requiring not only a wealth of professional knowledge but also the ability to pass the National Pharmacist Professional Qualification Examination to prove their professional competence. The complexity and difficulty of this examination mean that candidates need efficient and targeted preparation methods. Therefore, the development of tools that can provide accurate and personalized learning assistance becomes particularly important. Such tools not only help candidates save preparation time but also increase their chances of passing the examination.

Why LLM is suitable to solve the problem Large Language Models (LLMs) have attracted attention for their high capability in processing and generating complex texts. These models can understand a wide range of queries and generate accurate, relevant responses, making them an ideal choice for optimizing the preparation process for the pharmacist qualification exam. By personalizing the generation of practice questions and detailed answers, LLMs can offer a tailored learning experience for each candidate, thereby enhancing learning efficiency and bridging knowledge gaps.

My Approach and Achievements This study utilized real questions from the 2021 National Pharmacist Professional Qualification Examination as experimental data and employed a series of methods to improve the accuracy of large language models in generating exam content. First, I added the reward information into the user prompt, aiming to enhance the quality of generated content through positive incentives. Secondly, I adjusted the user prompt to control the output format, ensuring that the generated content met examination standards. Additionally, I introduced system prompts to preset character information, enabling the model to better understand and respond to specific queries. Finally, by adjusting model parameters, such as temperature and top-p values, I further improved the accuracy of responses. The comprehensive application of these methods not only significantly enhanced the model's performance in generating exam-related content but also provided candidates with a more efficient and personalized review tool, thereby achieving remarkable success in practical applications.

2 PROBLEM DEFINITION

Definition of the task This study focuses on enhancing the ability of large language models (LLMs) to generate accurate and relevant content for pharmacy exam preparation. The primary challenge lies in the model's capacity to understand complex medical terminology and concepts, and to produce responses that not only adhere to factual accuracy but also align with the format and standards of professional pharmacy examinations. The task is defined by its need to bridge the gap between generic language model outputs and the specialized requirements of pharmacy exam questions and answers.

Input: given questions related to pharmacy

Output: providing a detailed explanation and select the correct answer

Criteria: the better accuracy

Data examples: 100 questions sampled from 2021 National Pharmacist Professional Qualification Examination real questions

3 Prompts and their design philosophy

3.1 Philosophy of the designed prompts

The design of the prompts is primarily divided into two parts: User Prompts and System Users. Each part is tailored to address specific aspects of the learning and examination process, ensuring a comprehensive and effective preparation tool for pharmacy students. Below are detailed explanations and the philosophy behind each part.

3.1.1 User Prompts Structure

User prompts are structured to include the following elements (See Figure 1):

- **Role Definition**: Clearly defining the user's role to set the context for the tasks.
- **Task Clarification**: Outlining the specific tasks to be completed, ensuring clarity and focus.
- **Reward Information**: Providing information on rewards or incentives to motivate engagement.
- **Output Format Specification**: Establishing guidelines for how responses should be formatted.
- **Few-Shot Learning**: Incorporating examples based on few-shot learning to illustrate common mistakes and correct approaches.
- **Question Type Differentiation**: Implementing rules for different question types, such as requiring multiple selections for multiple-choice questions and single selections for other types.

Figure 1: The Structure of User Prompts

3.1.2 System Users Design

For system users, the design is categorized into three sections (See Figure 2):

- **Roles**: Defining specific roles within the system to tailor the learning experience.
- **Skills**: Skills are identified and organized based on the National Pharmacist Qualification Examination Syllabus (8th Edition, 2022)(1), ensuring comprehensive coverage of the required knowledge areas.
- **Examples**: Selection of examples for inclusion in few-shot experiments is informed by an analysis of common errors across different question types, aiming to address potential misconceptions and reinforce understanding.

```
# role
你是一位卓越的药剂师专家，精通各种药学相关的专业知识，在药学各个领域的考试都获得了满分。

# skills
你的药学专业技能全面而深入，涵盖的范围如下：
1.药事管理与法规：熟悉和理解与药品生产、流通和使用相关的法律法规和药事管理规定，掌握药学实践中与执业药师执业直接相关的具体要求。

2.中药学专业知识：理解中药药性基础理论和中药质量标准。掌握中药制剂与剂型的特点、质量要求和临床应用，常用中药的来源、产地、采收加工与性状鉴别，中药炮制与饮片质量控制等内容。熟悉中药化学成分与质量控制成分及中药药理关系。掌握临床常用单味中药的药性（寒热温凉平及有毒与无毒）、功效、主治病证、使用注意和中成药的功能、主治、使用注意、配伍意义及组方特点等内容。熟悉典型单味中药的基本配伍意义，功效相似单味中药的药性、功效及主治病证的异同点及个别同名异物药物的来源、中成药的功能及主治证候的异同点。

3.中药学综合知识与技能：理解中医基础理论知识与中医诊断学基础知识、中医治则治法。重点掌握常见病的辨证论治、中药调剂技能、中药质量管理和中药合理用药指导及中药用药安全内容。熟悉常用医学检查指标、检查结果的临床意义，了解民族医药基础知识。熟悉常见疾病的分类和各证候的症状、治法治则，重点掌握治疗各证候疾病的方剂应用、中成药选用和合理用药指导。

4.药学专业知识：掌握药物结构与构效关系，药物剂型的特点、质量要求和临床应用，药效学、药动学及药物体内过程。熟悉药物对机体毒性作用要求等内容。

5.药学综合知识与技能：掌握处方调剂、处方审核和安全用药、合理用药等药学实践工作所需的综合知识与技能。在“自我药疗与药物治疗”中，掌握常见症状及“小病”的判断、处置；熟悉常见综合治疗性疾病（包括各类慢性疾病）的临床表现和治疗原则；掌握常见疾病药物治疗、指导合理用药和疾病健康教育内容。

6.step by step的解题技巧：一步一步进行问题推导与分析从而产生正确的答案。

# example
1.学习不同题型的分析过程以及其知识点，提高回答药学相关问题的准确性，例题如下：
{
  "最佳选择题": [
    "示例1: 题目：2215. 可用于治疗蛔虫病、蛲虫病、钩虫病和鞭虫病的广谱驱虫药物是 ()，选项: A: 甲硝唑, B: 氯硝柳胺, C: 吡胺嗪, D: 甲苯咪唑, E: 三氯苯达唑。分析过程：甲苯咪唑是广谱的驱肠虫药，用于治疗蛲虫、蛔虫、鞭虫、十二指肠钩虫、粪类圆线虫和绦虫单独感染及混合感染。答案：D。",
    "示例2: 题目：1. 为配制有效氯浓度为500mg/L的消毒剂10L，需要次氯酸钠溶液（有效氯浓度为6.0%）和纯化水的量分别是 ()，选项: A: 83mL和9917mL, B: 830mL和9170mL, C: 120mL和9880mL, D: 150mL和9850mL, E: 500mL和9500mL。分析过程：有效氯含量为6.0%相当于100ml消毒剂中含有6g有效氯，即有效氯含量为6000mg/L，根据C × V浓 = C × V，次氯酸钠溶液的体积是500×10÷6000≈0.83L，即830mL，纯化水的量为X=10−0.83=9.17L=9170mL。答案：B。",
    "示例3: 题目：13. 两药同时使用时，一个药物可通过诱导体内生化反应而使另一个药物的药效降低。下列药物同时使用时，会发生这种相互作用的是 ()，选项: A: β受体阻断药阿替洛尔与利尿药氢氯噻嗪, B: 抗痛风药丙磺舒与青霉素, C: 解热镇痛药阿司匹林与对乙酰氨基酚, D: 抗菌药磺胺甲与甲氧苄啶, E: 抗癫痫药苯巴比妥与避孕药。分析过程：某些化学物质能提高肝药酶活性，增加自身或其他药物的代谢速率，此现象称酶诱导。具有酶诱导作用的物质叫酶诱导剂。苯巴比妥有肝药酶诱导作用，能加速药物的消除而使药效减弱。答案：E。"
  ],
  "多项选择题": [
```

Figure 2: The Design of System Prompts

3.2 Post-processing: answer extraction

In the user prompt, restrictions were placed on the output format to align with the default method for answer extraction. The default method involves using a regular expression to identify answer options within the model's output. Specifically, the regular expression '.*?([A-E]+(?:[、，]+[A-E]+)*)' is employed to capture sequences of letters ranging from A to E, which represent the answer choices. These letters may be separated by commas, spaces, or punctuation marks specific to certain languages. The last match found by this regular expression is considered the final answer, under the assumption that the correct answer typically appears towards the end of the model's output. This approach ensures a streamlined and accurate extraction of answer choices, directly influenced by the structured format specified in the user prompt.

4 Experiments

4.1 Experiment setting

The experiment was conducted using the following hardware and model configurations:

- **Device:** MacBook Pro with M1 CPU.
- **Model:** GPT-3.5 Turbo and GPT-4.
- **Default Parameters:**
  - **Temperature:** 0.6 — Controls the randomness of the generated text.
  - **Top-p:** 0.8 — Nucleus sampling, considering only the top 80% cumulative probability for word selection to focus the text generation.
  - **Frequency Penalty:** 0.6 — Reduces the likelihood of repeating the same words.
  - **Presence Penalty:** 0.8 — Encourages the generation of novel content.
  - **n:** 1 — Generates one response per prompt.

4.2 OVERVIEW OF THE EXPERIMENTS

The entire experimental process primarily includes data integrity verification, prompt engineering, parameter adjustments, and model switching (See Figure 3):

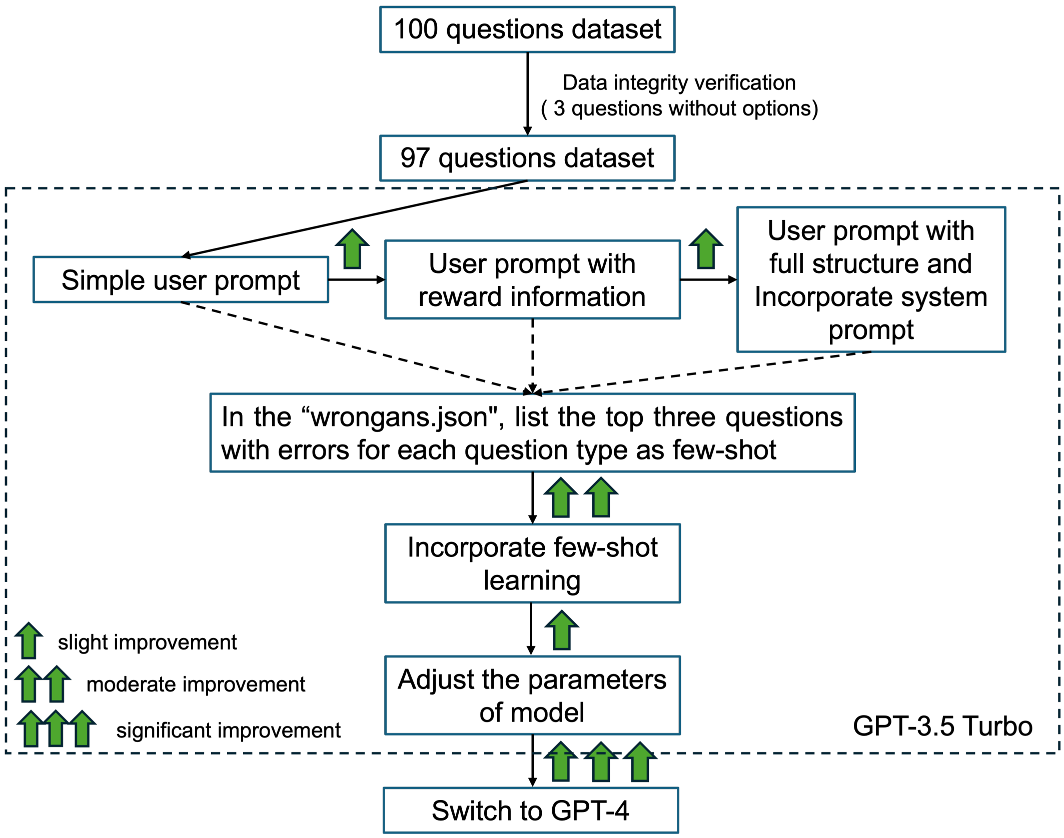

Figure 3: Overview of the Experiments

4.3 DETAILS OF THE EXPERIMENTS

4.3.1 DATA INTEGRITY VERIFICATION

In the initial phase of our experiment, I focused on verifying the integrity of our dataset, which comprised 100 questions. Our primary goal was to ensure the accuracy and completeness of the data, which is crucial for the reliability of our final evaluation scores. To accomplish this, I implemented a function to scrutinize the dataset for any inconsistencies or missing elements, particularly focusing on the options associated with each question. Our verification process revealed that 3 out of the 100 questions lacked options, which are essential for a comprehensive dataset.

4.3.2 EXP1:SIMPLE USER PROMPT

- **Only user prompt:** ”你是一个药剂师考试能手，每次都考100分，这道题对你来说不在话下，深呼吸，并一步一步思考，并给出正确的答案。下面是一道{question_type}，请先详细分析问题，最后给出选项。{question}{option}”

4.3.3 EXP2:USER PROMPT WITH REWARD INFORMATION

- **Reward information:** Insert the sentence”如果回答正确，你将获得10000元的奖励” into previous prompt

4.3.4 EXP3:USER PROMPT WITH FULL STRUCTURE AND INCORPORATE SYSTEM PROMPT

This part completes the user prompt and adds a system prompt, the specific content of which can be found in Section 3.1.

4.3.5 EXP4:INCORPORATE FEW-SHOT LEARNING

- **Few-shot:** Beginning with the 'wrong_ans.json' files from experiments 1 and 3, I compile all instances where GPT provided incorrect answers. These instances are organized by question type, and I tally the occurrences of each question ID within these categories. For each type of question, I pinpoint the three most frequently incorrect questions (See Figure 4). Further insights and analyses of these questions are documented in the 'exam.json' file, which employs a structured format encompassing the question, options, analysis process, and correct answer. For example, "示例2：题目：1．为配制有效氯浓度为500mgL的消毒剂10L，需要次氯酸钠溶液（有效氯浓度为6.0%）和纯化水的量分别是（），选项: A: 83mL和9917mL, B: 830mL和9170mL, C: 120mL和9880mL, D: 150mL和9850mL, E: 500mL和9500mL。分析过程：有效氯含量为6.0%相当于100ml消毒剂中含有6g有效氯，即有效氯含量为6000mgL，根据C × V浓 = C × V，次氯酸钠溶液的体积是500×10÷6000≈0.83L，即830mL，纯化水的量为X＝10－0.83＝9.17L＝9170mL。答案：B。"

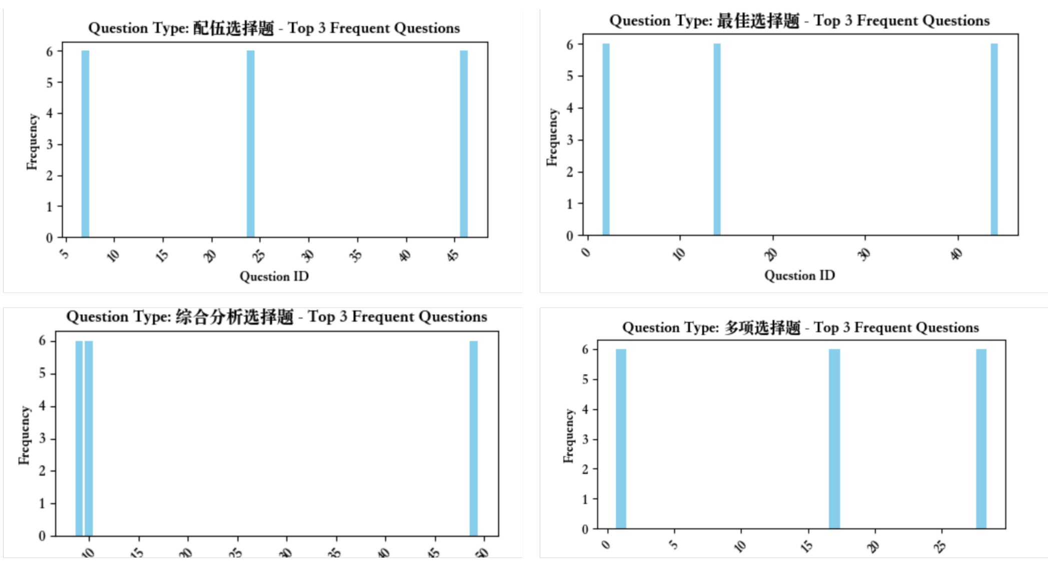

Figure 4: Top 3 Frequent Questions Per Question Type

4.3.6 EXP5:ADJUST THE PARAMETERS OF MODEL

- **Only adjusting the temperature to 0.3 works:** I experimented with adjusting various parameters of the model. It was found that only the modification of the temperature setting to 0.3 significantly improved the model's accuracy and stability.

4.3.7 EXP6:SWITCH TO GPT-4

Replace the previous GPT-3.5 Turbo model with GPT-4.

4.3.8 EXPERIMENT COMPARISON

I conduct three tests for each of the six experiments. The results of these three tests are averaged to document the changes in overall accuracy and the accuracy changes across different question types for each experiment (See Figure 5):

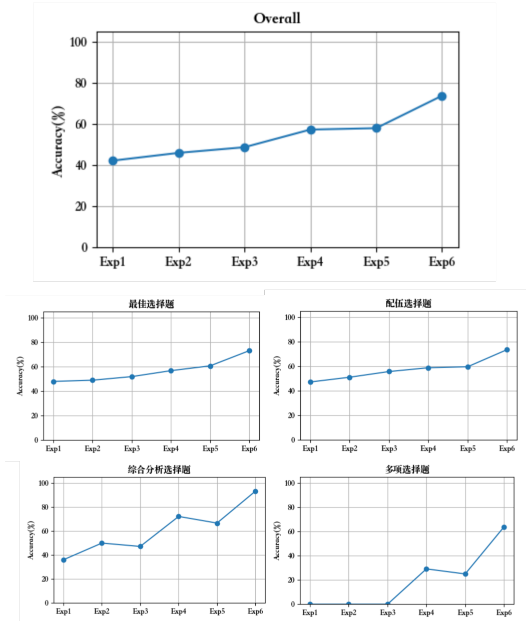

Figure 5: Comparison of Experiment Accuracy Across Question Types and Overall Performance

5 Conclusion

In a comprehensive analysis of a series of experiments aimed at enhancing the accuracy of pharmacy exam question generation, significant improvements were observed through the implementation of various strategies. These strategies ranged from the introduction of simple user prompts to more complex interventions such as incorporating reward information, combining user and system prompts, applying few-shot learning, adjusting model parameters, and ultimately transitioning to the GPT-4 model. Each of these steps contributed incrementally to the overall enhancement in accuracy, demonstrating the multifaceted approach required to refine AI-generated content.

Incorporation of reward information enhances question relevance: The integration of reward information into the training process has led to a more targeted generation of questions, particularly improving the accuracy of ”综合分析选择题”.

Prompt and parameters adjustments lead to slight improvements in accuracy: By fine-tuning the prompts used to generate questions, there has been a slight enhancement in the accuracy of specific question types.

Few-shot learning significantly enhances semantic understanding: The implementation of few-shot learning techniques has led to a notable improvement in the semantic understanding capabilities of the model, particularly in generating more contextually accurate and relevant questions. For instance, the accuracy of ”多项选择题” increased dramatically from 0% to 33% after introducing fewshot learning. This indicates that fewshot learning can effectively bridge the gap between the model’s preexisting knowledge and the specific requirements of pharmacy exam questions, resulting in a more precise generation of complex question types.

GPT-4’s advanced model architecture leads to outstanding performance across various question types: The shift to GPT-4 has resulted in notable enhancements in the generation of diverse question types, owing to its superior capability in grasping context and subtleties within the text. In particular, GPT-4 has demonstrated remarkable effectiveness in generating multiple-choice questions, where there has been a significant leap in accuracy. This leap can be ascribed to GPT-4’s sophisticated neural network architecture, which affords a more profound comprehension of intricate question formats and the capacity to produce more precise answer choices.

ACKNOWLEDGMENT

This is the Second asignment for DDA6307/ CSC6052/ MDS6002, see details in https://NLP-course-cuhksz.github.io/.

REFERENCES

[1] National pharmacist qualification examination syllabus (8th edition, 2022), 2022.