

PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization





Xinyuan Wang^{1*}, Chenxi Li^{1*}, Zhen Wang^{12*}, Fan Bai⁵, Haotian Luo², Jiayou Zhang², Nebojsa Jojic², Eric P. Xing²⁴, Zhiting Hu¹

¹UC San Diego, ²MBZUAI, ³Microsfot Research, ⁴CMU ⁵Georgia Tech



Expert prompts: Highly effective, task-specific prompts that are often 1. Great empirical gains over 12 tasks spanning three domains: heavily engineered by domain experts

- Unleash the full potential of very large LLMs, e.g., GPT-3.5/4
- Spearhead the next era of prompt engineering, with more powerful LLMs that can understand intricate instructions

The following example shows how the expert prompt improves perf. with richer domain knowledge and structured guidance

Biomedical Task Input

Linkage studies in this family suggested a close linkage between the <u>c2 deficiency</u> gene and genes coding for B18, Dw2, and BfS antigens. "c2 deficiency" is a disease mention to be extracted



BIG-Bench Hard, domain-specific and general NLP tasks

	Domain-specific Tasks				General NLU Tasks			
	NCBI (F1)	Biosses	Med QA	Avg.	Subj	TREC	CB	Avg.
Human (ZS)	0.521	0.550	0.508	0.526	0.517	0.742	0.714	0.658
Human (FS)	0.447	0.625	0.492	0.521	0.740	0.742	0.429	0.637
CoT (ZS)	0.384	0.425	0.508	0.439	0.656	0.63	0.750	0.679
СоТ	0.376	0.675	0.542	0.531	0.670	0.784	0.643	0.699
GPT Agent	0.125	0.625	0.468	0.406	0.554	0.736	0.339	0.543
APE	0.576	0.700	0.470	0.582	0.696	0.834	0.804	0.778
PromptAgent	0.645	0.750	0.570	0.655	0.806	0.886	0.911	0.868

2. Transferrable expert prompts across various base models

3. Great balance of **exploration** efficiency and performance

Ordinary User Prompt Extract the disease or condition from the sentence, if any is mentioned.

Prompt From Sampling-Based Method If any disease or condition is mentioned in the sentence, extract it.

Expert-level Pr	ompt
Task Description	You're tasked with extracting diseases or conditions from the given sentence
Domain Knowledge	Avoid associated elements: inheritance patterns, genes or gene loci (like PAH)
Solution Guidance	Consider both specific diseases and broader categories, common abbreviations
Exception Handling	The term 'locus' should be recognized as a genomic location, not a disease name
Output Formatting	Provide the identified diseases in this format: {entity_1,entity_2,}

Ordinary User/Sampled Prompt Output

c2 deficiency gene

Expert Prompt Output c2 deficiency

No More Prompt Engineering? ジ

Manual expert prompt engineering?

- A unique blend of domain knowledge and intuition for LLMs
- Ad-hoc human-chatbot interactions with tedious trial-and-errors Existing automatic prompt engineering?
- Tend to overlook the depth of domain knowledge
- Struggle to efficiently explore the vast prompt space

Goal: Automatically craft expert-level prompts equivalent in quality to those handcrafted by domain experts





4. Ablation study on search variants, showing the **power of** strategic planning (MCTS)

MC	Beam	Greedy	MCTS (Ours
0.772	0.823	0.810	0.873
0.575	0.675	0.700	0.750
0.490	0.610	0.545	0.670
0.650	0.610	0.660	0.670
0.692	0.765	0.778	0.806
0.635	0.697	0.698	0.754
	MC 0.772 0.575 0.490 0.650 0.692 0.635	MCBeam0.7720.8230.5750.6750.4900.6100.6500.6100.6920.7650.6350.697	MCBeamGreedy0.7720.8230.8100.5750.6750.7000.4900.6100.5450.6500.6100.6600.6920.7650.7780.6350.6970.698

Method 0.85 Greedy-S Greedy-L 0.80 APE 0.75 Ours Accuracy 0.20 0.62 Task penguins 0.60 biosses geometry 0.55 causal 0.50 -🛑 subj 50 75 100 125 150 Exploration Efficiency (# of Explored Prompts)

5. Convergence analysis reveals a stale learning dynamic



6. A prompt evolution trajectory, showing how error feedbacks (highlighted colors) are accumulated into better expert prompts

 $>S_1$

 $\cdot S \gamma$

 $|a_1|$

Error Feedback: The language model should

Prompt: Extract the disease or condition from the sentence, if any \not s_0 is mentioned.

Key innovations (check more details in the paper):

- 1. Reframing prompt optimization as a strategic planning problem
- Efficiently explore prompt space with **lookahead and backtrack**
- 2. Self-reflection on model errors to mirror human's trial-and-error
- Effectively induce valuable domain insights with error feedbacks



(b) State Transition

Current Prompt: Please determine whether one sentence entails the next. Step1: Retrieve Errors from Base Model

Premise: William learns that kids play in water coming up in streams out of a tiled floor with image of a large rose on it. **Hypothesis**: William learns that kids are playing in water.

Label: Non-entailment **Prediction**: Entailment

Step 2: Generate Error Feedback (Action)

Meta-prompt 1: Summarize errors and suggest improvements

Error Feedback: Ignoring Context and Detail—The model might be overlooking the details of the premise 'kids play in water coming up in streams out of a tiled floor with an image of a large rose on it,', which directly implies the hypothesis.

Step 3: Update Prompt (State)

Meta-prompt 2: Given the error feedback, give me a better prompt

New Prompt: Compare the provided sentences ... Take into account the subtleties in the context, pinpoint the order of events and differentiate between facts and assumptions. If the hypothesis is a direct result of the premise, select 'entailment'

differentiate between diseases and the genes or other factors associated with those diseases. Consider a_0 different forms or variations of disease names, including abbreviations or short forms.

Error Feedback: The model misunderstood the difference between diseases and associated factors like proteins or pathways. Clarify the definition of disease and condition entities, stressing the exclusion of associated factors like genes or proteins. Specify that common biological terms or genetic locations (like locus) should not be mistaken as part of a disease specific name.

Error Feedback: The model wrongly identified "PAH" as a disease, while it's actually a gene. The model $|a_2|$ misunderstood the part "excluding any associated factors such as genes, proteins, or pathways". Explicitly emphasize the need to exclude associated factors like inheritance pattern (for example, autosomal dominant), genes (like PAH), proteins, or pathways when identifying diseases or conditions.

F1 score (test): 0.521

Prompt: Identify and extract all diseases or conditions mentioned in the sentence, taking care to distinguish between diseases and any associated factors like genes. ... Any variations or abbreviations of disease names should also be included.... **F1 score (test)**: 0.609

Prompt: You're tasked with identifying and extracting diseases or conditions as mentioned in the sentence, while carefully excluding any associated factors such as genes, proteins, or pathways. ... For clarity, the term 'locus' is not part of any disease name but represents a specific location in the genome. **F1 score (test)**: 0.622



Road Ahead with PromptAgent

- PromptAgent serves as **a principled framework** to study prompt optimization by unifying prompt sampling and rewarding
- Unlike discovering magic/local prompt variants, expert-level prompting is still an untapped area to solve challenging problems