



# Prompt engineering for RAG



A large, bold, black font title 'Prompt engineering for RAG' is centered. The word 'RAG' is enclosed in a purple oval. Above the title, there's a yellow hand holding a pen emoji on the left, and a white speech bubble with three dots emoji on the right. A small lightbulb emoji is positioned between 'engineering' and 'RAG'.

5 latest research-backed methods



**Joanna Stoffregen**

co-founder & AI Product Lead  
[@Labsbit.ai](https://www.labsbit.ai)



Last week we looked into 4 prompt engineering techniques for RAG

Let's look into 5 more

# Technique

# How it works

1

## ExpertPrompting

Instructing LLMs to respond as distinguished experts

2

## Step-Back Prompting

Asking a step-back question first before responding

3

## Rephrase and Respond RaR

Rephrasing prompts before generating response

4

## Plan-and-Solve Prompting

Dividing tasks into sub-tasks and performing them according to a plan

5

## System 2 Attention

LLM rewrites the prompt to remove irrelevant information

1

# Expert Prompting

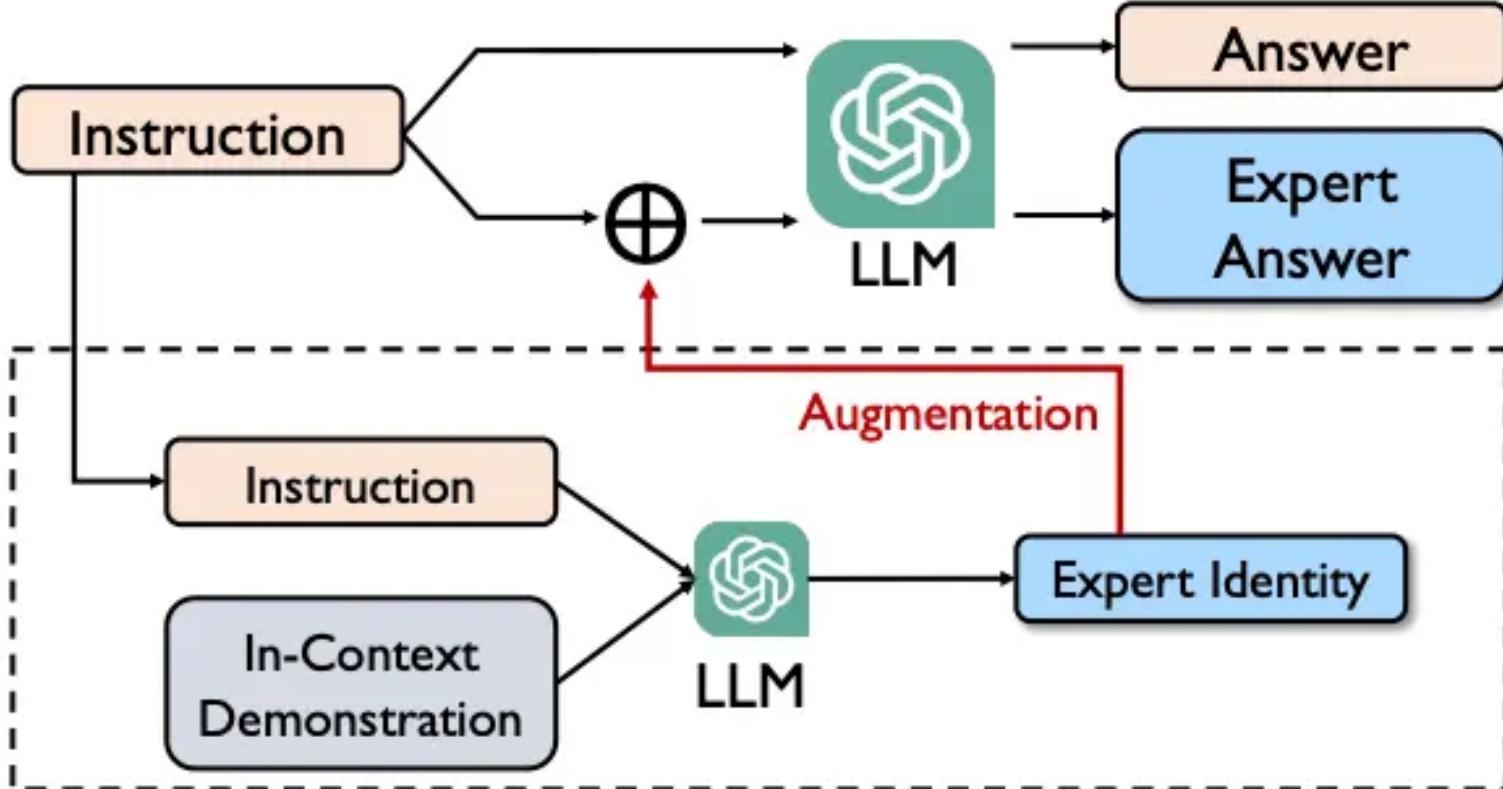
*Expert Prompting: Instructing Large Language  
Models to be Distinguished Experts*



## 1

# ExpertPrompting

This method helps improves LLMs performance by treating them like experts in specific areas.



## Step 1

For each task, using "in context learning" the system synthesizes a detailed, customized description of an "expert identity." This description tells the language model what kind of expert it should act like for that particular question or topic.

## Step 2

Then, the model answers questions based on this expert identity, rather than just giving standard responses. This makes the model use its knowledge in a way that's more focused and similar to how an actual expert would respond.

Let's see an example →

## Instruction: Describe the structure of an atom.

### Expert Identity:

You are a physicist specialized in atomic structure and the properties of matter at the atomic level. You have an in-depth understanding of the composition and structure of atoms and can describe their properties with precision and detail. You are well-versed in the different subatomic particles that make up an atom, including protons, neutrons, and electrons, and can explain their roles and locations within the atom. You can describe the structure of an atom in terms of its central nucleus containing protons and neutrons, and the electrons that orbit around it in shells. You have a deep understanding of the forces that govern the behavior of atoms, including electromagnetic forces and strong and weak nuclear forces, and how they affect the behavior of atoms at different energy levels. Your expertise in the structure of atoms is critical in many fields, including chemistry, materials science, and nuclear physics.

### gpt-3.5 answer:

An atom consists of a nucleus, which is made up of positively charged protons and neutral neutrons, surrounded by negatively charged electrons that orbit the nucleus. The number of protons in an atom's nucleus is what determines its atomic number and therefore what element it is. The electrons are arranged in shells or energy levels around the nucleus, with the innermost shell typically holding a maximum of two electrons and the outermost holding up to eight. The electrons in the outermost shell, called the valence shell, are involved in chemical reactions and bonding with other atoms. The size of an atom is determined by the distance between the nucleus and the outermost electrons.

### gpt-3.5 answer w/ expert identity:

The structure of an atom can be thought of as comprising of three main components: a central nucleus containing protons and neutrons, and electrons that orbit around it in distinct energy levels or shells. The nucleus is the dense, positively charged center of the atom, while the electrons are negatively charged particles that orbit around the nucleus at varying distances. The number of protons present in the nucleus determines the atomic number of the atom, which in turn defines its unique chemical properties and place in the periodic table of elements. The number of neutrons present in the nucleus can vary, giving rise to isotopes of the atom. The electrons occupy distinct energy levels or shells around the nucleus, with those closest to the nucleus having lower energy levels than those further away. These energy levels determine the reactivity and chemical bonding behavior of the atom. The overall structure of an atom is governed by the electromagnetic force that holds the negatively charged electrons in orbit around the positively charged nucleus, and the strong nuclear force that binds the protons and neutrons together in the nucleus.



Check: The opensource ChatBot built with ExpertPrompting which achieves 96% of ChatGPT's capability.  
[github.com/OFA-Sys/ExpertLLaMA](https://github.com/OFA-Sys/ExpertLLaMA)

2

# Step-Back Prompting

*Take a Step Back: Evoking Reasoning via  
Abstraction in Large Language Models*



## 2

# Step-Back Prompting

A technique from Google Deepmind, first asks the LLM a "step-back" question that is more abstract and high-level, before asking the original specific question.

## The Algorithm concept

- Generate a Step back prompting based on the user's original question
- Collect data related to both the original query and stepback question
- Formulate a response by considering the gathered information from both queries

Check more here: [quickaitutorial.com/how-powerful-step-back-prompting-transforms-l1m-performance/](https://quickaitutorial.com/how-powerful-step-back-prompting-transforms-l1m-performance/)

Let's see an example





Example NOT using Step-Back Prompting technique

### Original Question

Estella Leopold went to which school between Aug 1954 and Nov 1954?



wrong answers



### Chain-of-Thought

... After that, she pursued graduate studies at the University of California, Berkeley, earning her master's degree in 1950 and her Ph.D. in 1955.

Considering this timeline, it is highly likely that Estella Leopold was attending the University of California, Berkeley, between August 1954 and November 1954...



### Original Answer

Estella Leopold attended the University of Wisconsin-Madison between August 1954 and November 1954





## Example using Step-Back Prompting technique

### Original Question

Estella Leopold went to which school between Aug 1954 and Nov 1954?



### Stepback Question

What was Estella Leopold's education history?



### Stepback Answer

B.S. in Botany, University of Wisconsin, Madison, 1948  
M.S. in Botany, University of California, Berkeley, 1950  
Ph.D. in Botany, Yale University, 1955



correct answer

### Final Answer

She was enrolled in the Ph.D. program in Botany at Yale from 1951 to 1955. Therefore, Estella Leopold was most likely attending Yale University between August 1954 and November 1954.



3

# Rephrase and Respond RaR

*Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves*



# 3 Rephrase and Respond (RaR)

A complementary technique to CoT, aims to enhance LLM's performance allowing them to rephrase and expand the prompts before generating a response.

## Original question

Was {person} born in an even day?

Was {person} born in an even month?

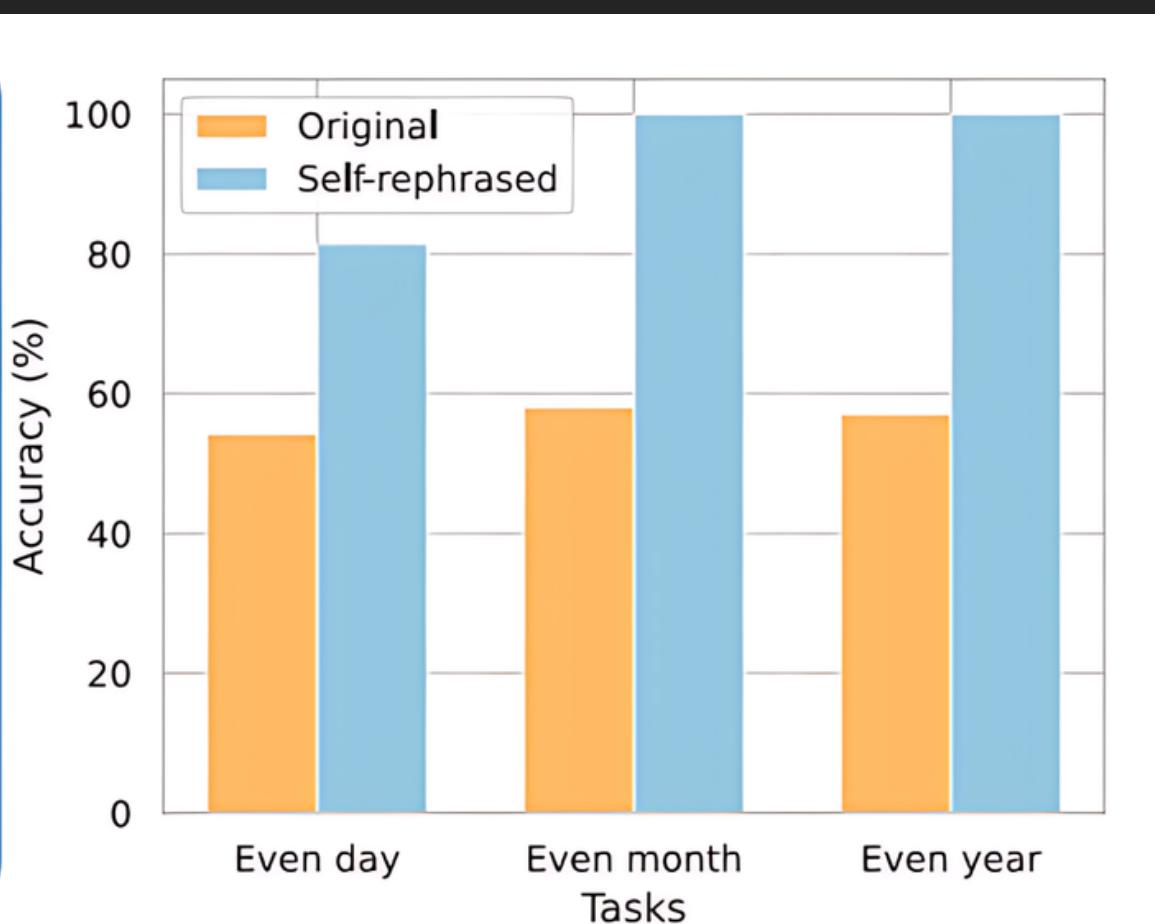
Was {person} born in an even day?

## Self-rephrased question

Could you provide more information on whether the individual named {person} was born on a day that is an even number? This refers to dates such as the 2nd, 4th, 6th, 8th, and so on within a given month.

Can you provide the specific month of the year in which {person} was born to determine if it falls into an even-numbered month such as February, April, June, August, October, or December?

What is the birth year of {person} and is it an even number?



# How it works

1

## Rephrasing

The LLM analyzes the prompt and generates a paraphrased version, simplify complicated tasks or further explain things that weren't fully stated.



“Take the last letters of the words in ‘Edgar Bob’ and concatenate them.”

Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question.

Rephrasing  
LLM



Can you identify and extract the final letters in both the words that form ‘Edgar Bob’, and then join them together in the order they appear?

2

## Responding

Based on original and rephrased versions of the prompt, the model generates a response.



(original) Take the last letters of the words in ‘Edgar Bob’ and concatenate them.

(rephrased) Can you identify and extract the final letters in both the words that form ‘Edgar Bob’, and then join them together in the order they appear?

Responding  
LLM



The last letters in the words “Edgar Bob” are “r” and “b”. Concatenating them in the order they appear would be “rb”

4

# Plan-and-Solve Prompting

[\*Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models\*](#)



## 4

# Plan-and-Solve Prompting

## The problem

CoT prompting fails in three areas:

1. Calculations (7% failure rate in test examples)
2. Missing steps in a sequence of events (12% failure rate in test examples)
3. Semantic misunderstanding (27% failure rate in test examples)

## The solution

- These vulnerabilities are addressed by Plan-And-Solve (PS) prompting.
- The model analyzes the problem, breaks it down into smaller tasks and creates a plan.
- It then works through each task to solve the problem

Let's see compare the Zero-Shot-CoT to the Plan and Solve method



Zero-shot-CoT enables LLMs to generate multi-step reasoning with “Let’s think step by step”. But it can generate wrong reasoning steps if the problem is complex.

Q: In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?



wrong answer

First, 20% of the class enrolled in contemporary dance, so that means that 80% of the class is left.

Next, 25% of the remaining 80% enrolled in jazz dance, so that means that 25% of the original class enrolled in jazz dance. That leaves 55% of the original class.

Finally, the rest of the class (55%) enrolled in hip-hop dance. So, 55% of the entire class enrolled in hip-hop dance.





**Unlike Zero-shot-CoT, PS prompting first asks LLMs to devise a plan to solve the problem by generating a step-by-step plan and carrying out the plan to find the answer**

Q: In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?

A: *Let's first understand the problem and devise a plan to solve the problem.*

*Then, let's carry out the plan and solve the problem step by step.*



correct answer

**Plan:**

Step 1: Calculate the total number of students who enrolled in contemporary and jazz dance.

Step 2: Calculate the total number of students who enrolled in hip-hop dance.

Step 3: Calculate the percentage of students who enrolled in hip-hop dance.

**Solution:**

Step 1: 20% of 20 students is 4 students. 25% of the remaining 16 students is 4 students. So, a total of 8 students enrolled in contemporary and jazz dance.

Step 2: The remaining 12 students enrolled in hip-hop dance.

Step 3: The percentage of students who enrolled in hip-hop dance is  $12/20 = 60\%$ .



5

# System 2 Attention

*System 2 Attention (is something you might need too)*



# 5

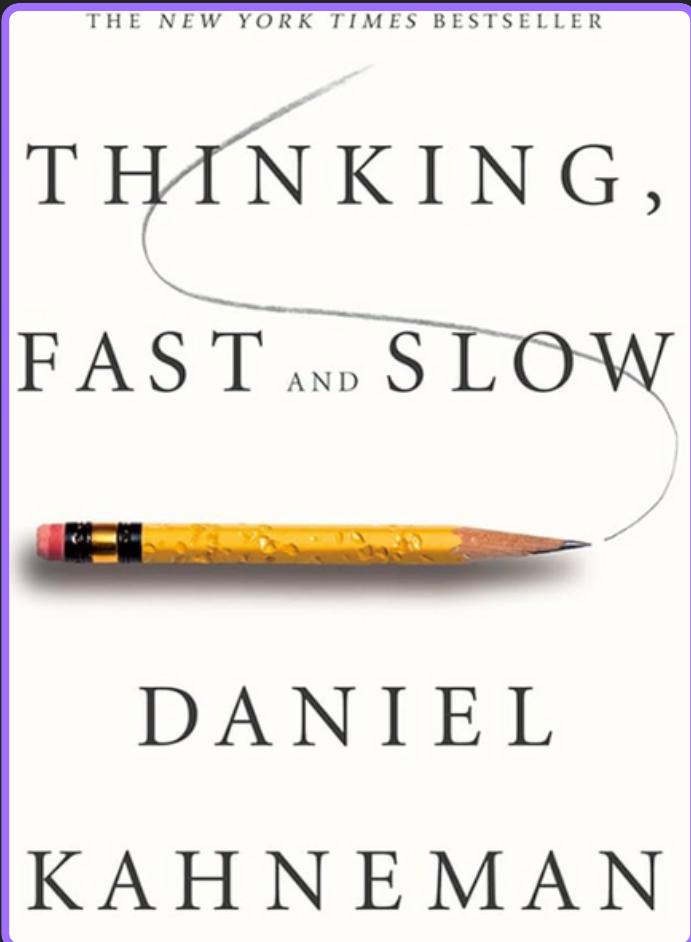
# System 2 Attention

## The problem

- Prompts for LLMs often contain irrelevant information that can distract the model and lead to errors, as the models tend to latch onto correlations in the input, even when they are not meaningful.

## The solution

- S2A rewrites the input prompt, removing the irrelevant parts and retaining only the relevant information.



- Inspired by the "System 1" (fast, heuristic-based) and "System 2" (slower, more deliberate) cognitive processes described by Daniel Kahneman.
- S2A acts as a "System 2" that takes control when the standard attention mechanism ("System 1") is prone to errors.



## Examples of how the LLM's responses are affected by irrelevant context

**Human:** Saratoga is a city in California.  
Saratoga has many parks.

In which city was San Jose's mayor Sam Liccardo born?

**text-davinci-003:** Sam Liccardo was born in Saratoga, California.

**GPT-3.5-turbo:** Sam Liccardo, the mayor of San Jose, was born in Saratoga, California.

**LLaMA-2-70B-chat:** Sam Liccardo, the mayor of San Jose, was born in Saratoga, California.

**Human:** Sunnyvale is a city in California. Sunnyvale has many parks. Sunnyvale city is close to the mountains. Many notable people are born in Sunnyvale.

In which city was San Jose's mayor Sam Liccardo born?

**text-davinci-003:** Sam Liccardo was born in Sunnyvale, California.

**GPT-3.5-turbo:** Sam Liccardo, the mayor of San Jose, was born in Sunnyvale.

**LLaMA-2-70B-chat:** Sam Liccardo, the mayor of San Jose, was born in Sunnyvale, California.



## How to use the System 2 Attention method

Given the following text by a user, extract the part that is unbiased and not their opinion, so that using that text alone would be good context for providing an unbiased answer to the question portion of the text.

Please include the actual question or query that the user is asking. Separate this into two categories labeled with “Unbiased text context (includes all content except user’s bias):” and “Question/Query (does not include user bias/preference):”.

Text by User: [ORIGINAL INPUT PROMPT]

**Problem:** Mary has 3 times as much candy as Megan. Mary then adds 10 more pieces of candy to her collection. Max has 1000 more books than Mary. If Megan has 5 pieces of candy, how many does Mary have in total?



An example from where a distracting sentence ("Max has 1000 more books than Mary") makes the LLM make a mistake

**LLaMA-2-70B-chat:** Solution: ...  
Max has 1000 more books than Mary, so  
Max has  $1000 + 3x + 10$  pieces of candy.  
2. Now, we can set up an equation using  
the information we have:  
 $1000 + 3x + 10 = 3(5) + 10$   
Simplifying the equation:  
 $1000 + 3x + 10 = 15 + 10$   
Solving for x:  
 $3x = 1000 - 10 - 15$   $3x = 985$   
Dividing both sides by 3:  
 $x = 985 / 3$   $x = 328$   
Therefore, Mary has 328 pieces of candy  
in total.  
Final answer (in numbers): 328

System 2 Attention  
regenerates the portion of  
the context it decides to  
pay attention to,  
successfully removing the  
distracting sentence, then  
hence answering correctly.



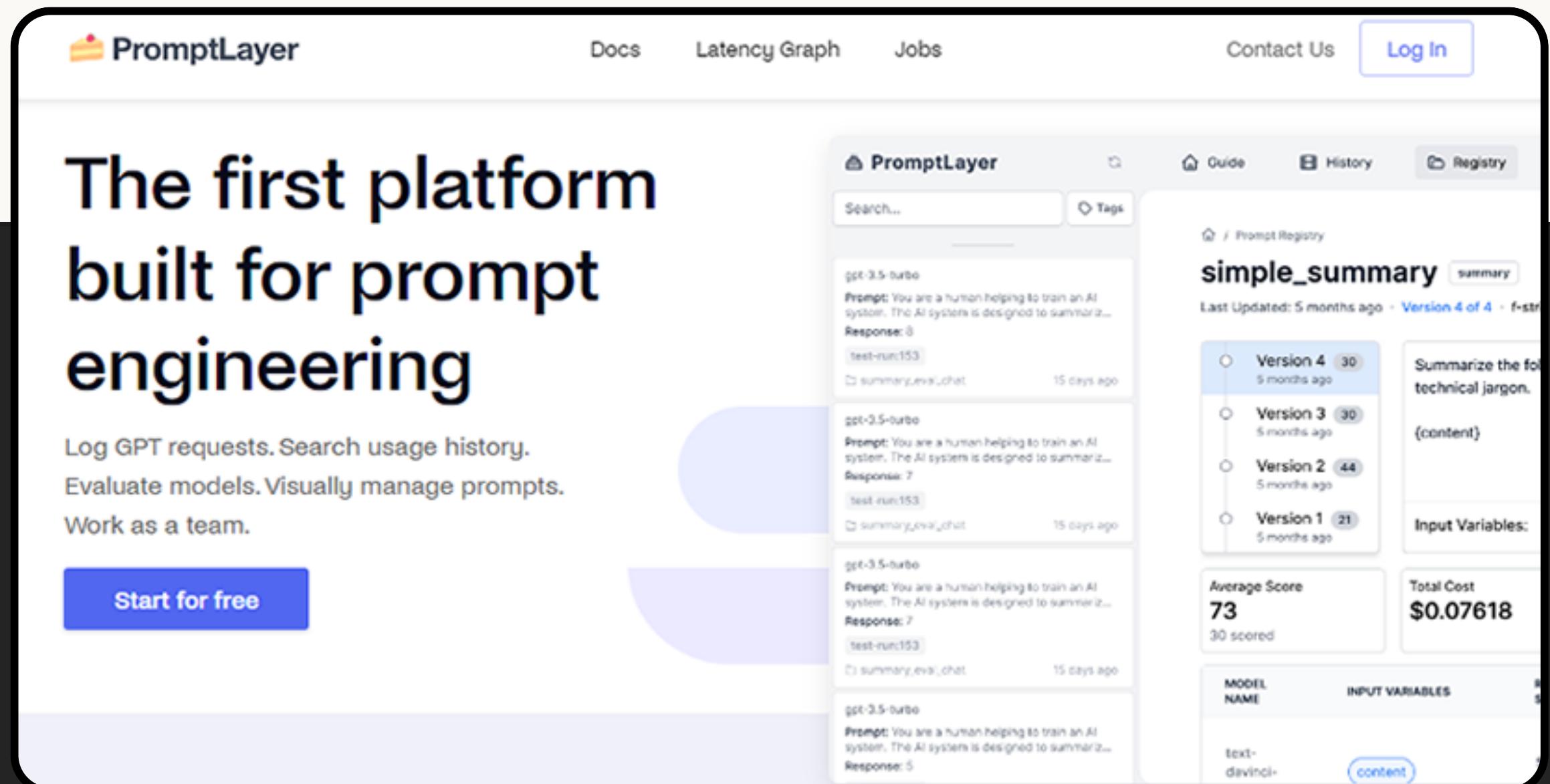
**Problem:** Mary has 3 times as much candy as Megan. Mary then adds 10 more pieces of candy to her collection. Max has 1000 more books than Mary. If Megan has 5 pieces of candy, how many does Mary have in total?

**S2A:** Mary has 3 times as much candy as Megan. Mary then adds 10 more pieces of candy to her collection. If Megan has 5 pieces of candy, how many does Mary have in total? Question: How many pieces of candy does Mary have in total?

**LLaMA-2-70B-chat:** Solution:  
Let's say Megan has 5 pieces of candy.  
Mary has 3 times as much candy, so Mary  
has  $3 \times 5 = 15$  pieces of candy. Mary then  
adds 10 more pieces of candy, so Mary  
has  $15 + 10 = 25$  pieces of candy.  
Final answer (in numbers): 25

# Bonus: 🍰 PromptLayer

## Batch test different versions of your prompts



The screenshot shows the PromptLayer platform interface. At the top, there's a navigation bar with links for "Docs", "Latency Graph", "Jobs", "Contact Us", and "Log In". Below this, a large heading reads "The first platform built for prompt engineering". Underneath the heading, there are three bullet points: "Log GPT requests. Search usage history.", "Evaluate models. Visually manage prompts.", and "Work as a team.". A prominent blue button labeled "Start for free" is located at the bottom left. On the right side, there's a detailed view of a "simple\_summary" prompt entry in the "Prompt Registry". The entry includes a search bar, a list of four versions (Version 4, Version 3, Version 2, Version 1) with their respective scores (30, 30, 44, 21), and a summary of the prompt: "Summarize the following text in one sentence. Remove all technical jargon. (content)". It also shows average scores (73) and total cost (\$0.07618). The interface is clean and modern, with a light gray background and blue accents.

I'm Joanna  
from **Labsbit.ai**

An AI Product Development  
& Automation Company



Reach out to learn more about how we  
can help you launch your RAG app