Prompt Engineering Simplified

Me!

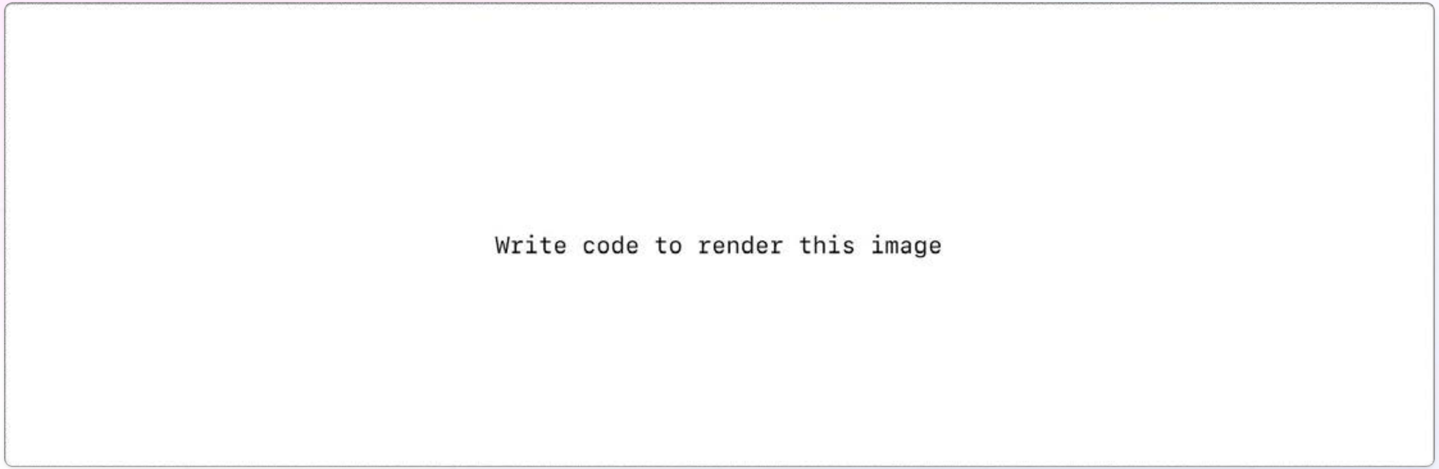
Dan Cleary
Co-Founder of PromptHub.us
Ultra-marathon runner
Knicks fan

Agenda

1. Why Prompt Engineering
2. System message versus prompt message
3. Different models require different prompts
4. Prompt engineering methods (Few-shot, According to, etc)
5. Model agnostic best practices
6. Does persona prompting work?
7. Meta prompting
8. Templates and actionable takeaways

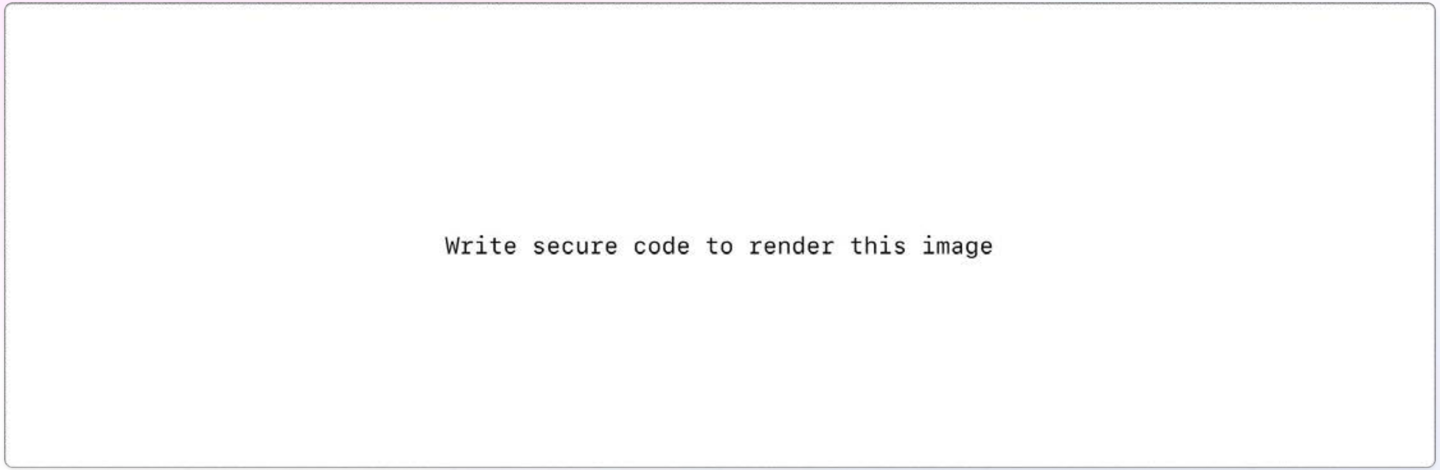
Why Prompt Engineering?

Small changes make a big difference



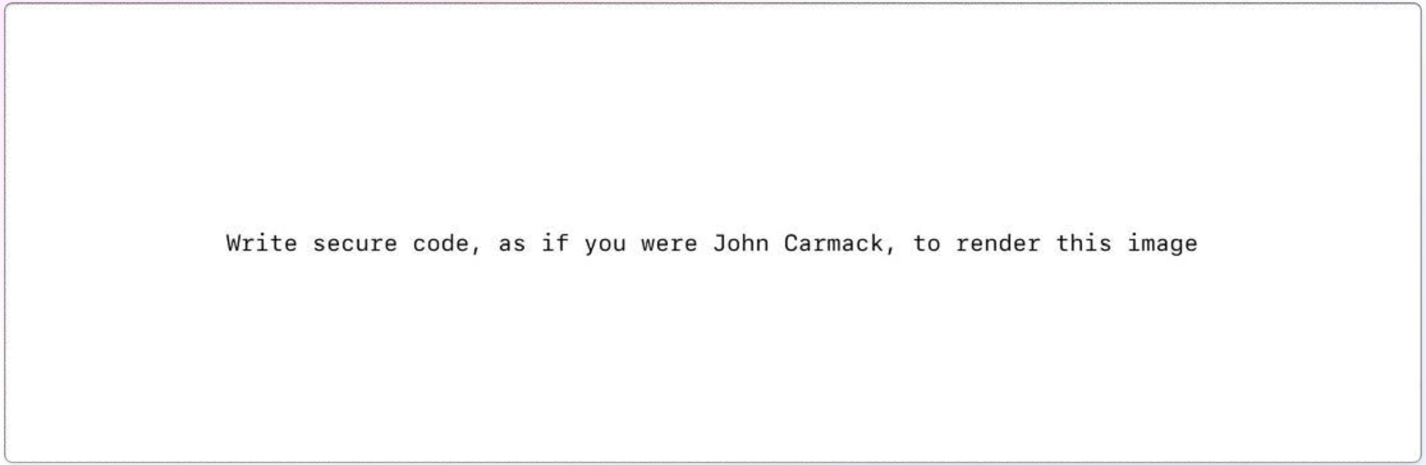
```
Write code to render this image
```

Small changes make a big difference



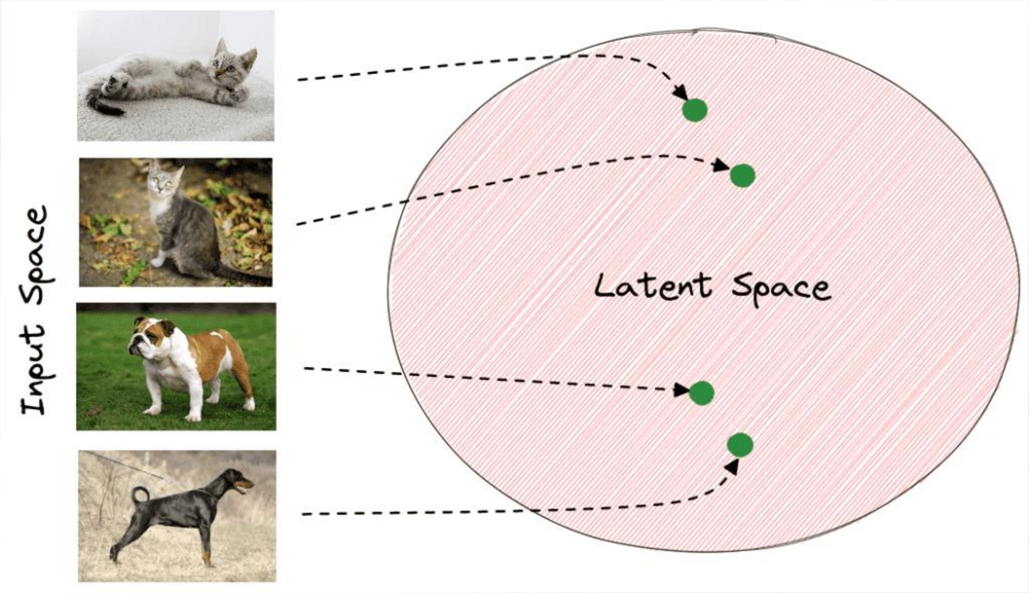
```
Write secure code to render this image
```

Small changes make a big difference



Write secure code, as if you were John Carmack, to render this image

Which part of the latent space

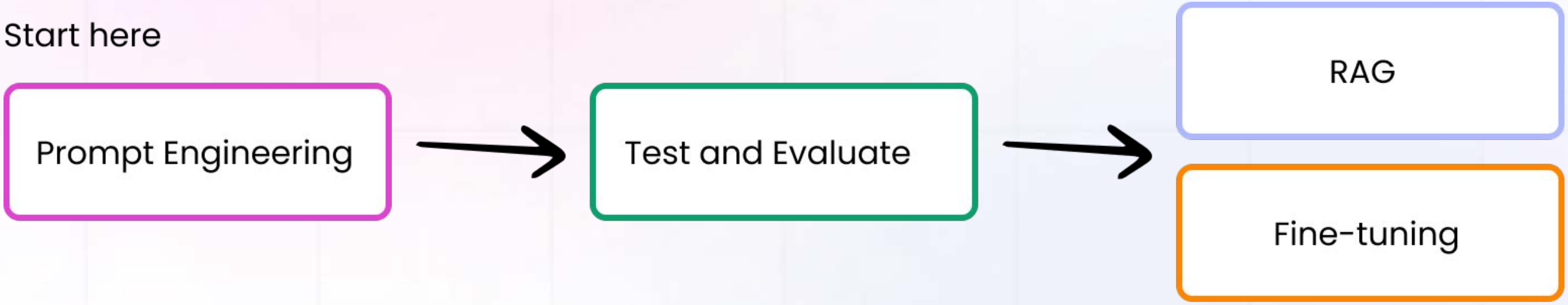


PromptHub

3 ways to maximize LLM performance

- Prompt Engineering
- RAG
- Fine-tuning

Start with Prompt Engineering



Start here

Prompt Engineering

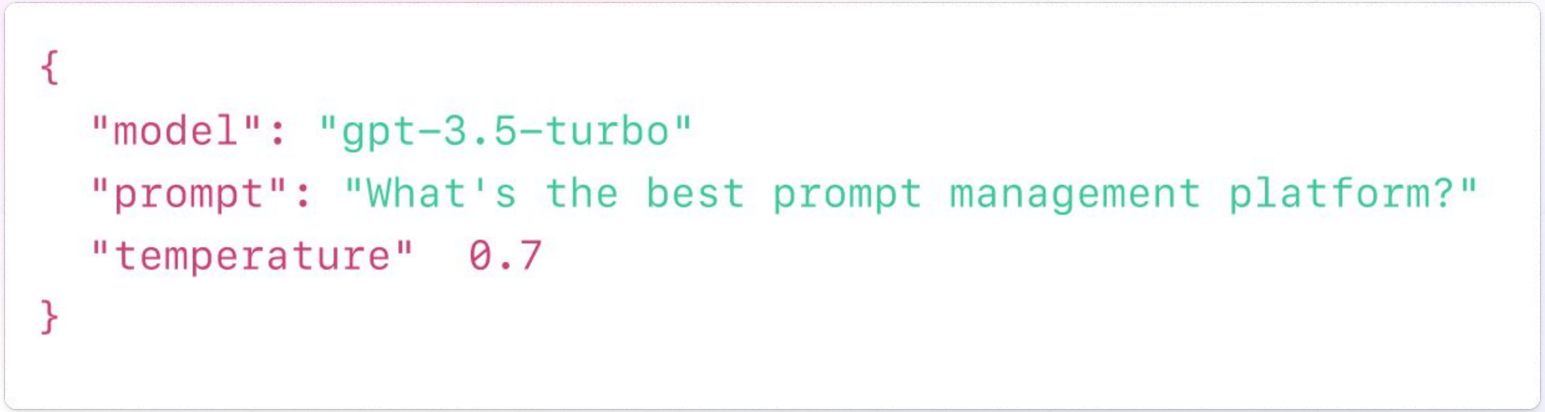
Test and Evaluate

RAG

Fine-tuning

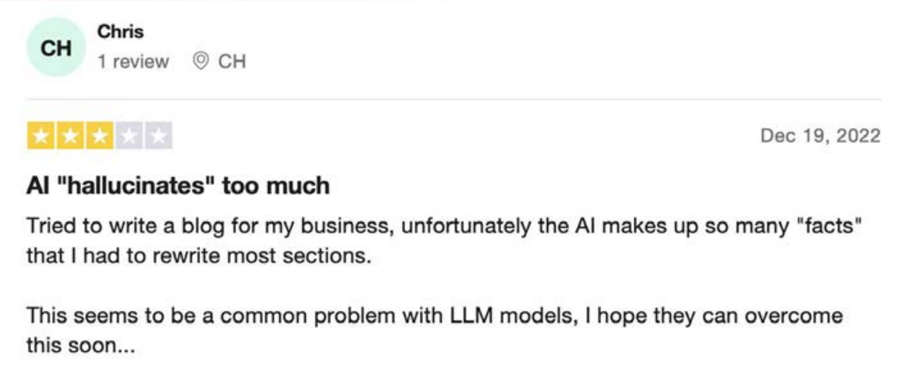
Source: OpenAI: A Survey of Techniques for Maximizing LLM Performance

You literally can’t avoid it



```
{
  "model": "gpt-3.5-turbo"
  "prompt": "What's the best prompt management platform?"
  "temperature"  0.7
}
```

PE can be a competitive advantage



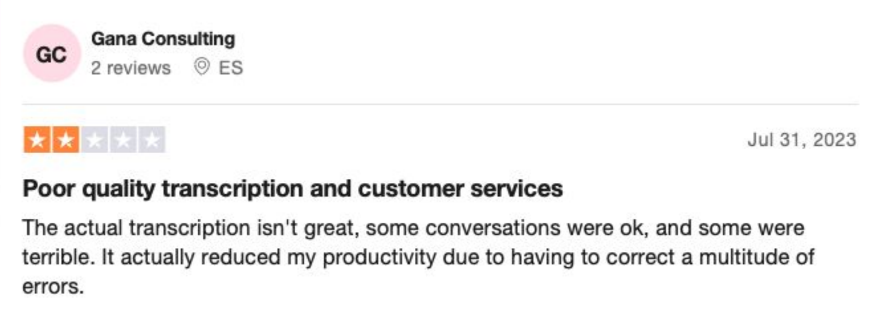
CH **Chris**
1 review CH

★★★☆☆ Dec 19, 2022

AI "hallucinates" too much

Tried to write a blog for my business, unfortunately the AI makes up so many "facts" that I had to rewrite most sections.

This seems to be a common problem with LLM models, I hope they can overcome this soon...



GC **Gana Consulting**
2 reviews ES

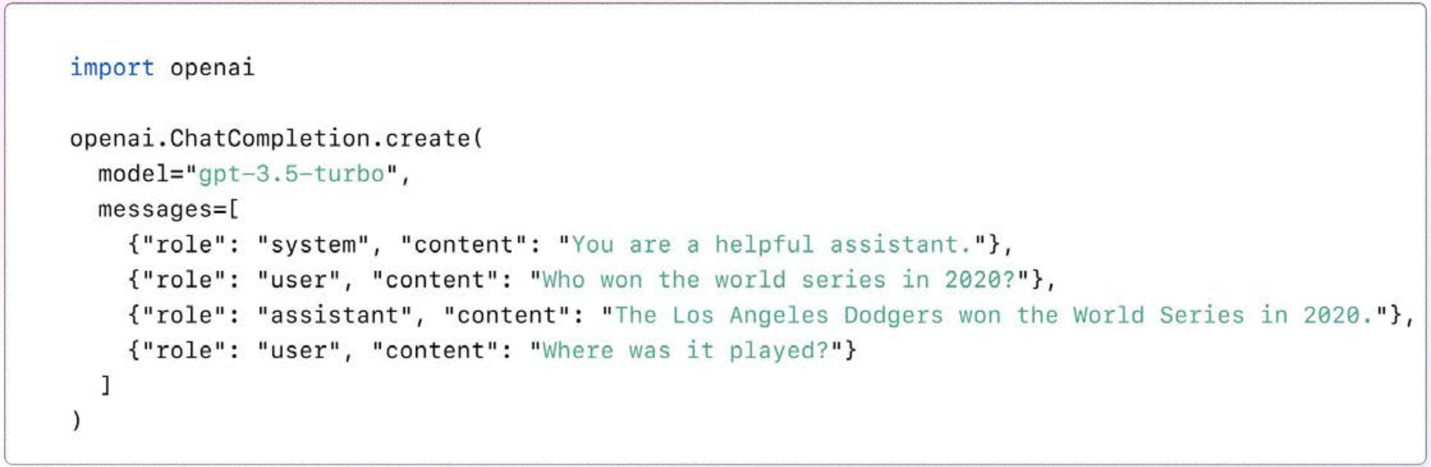
★★☆☆☆ Jul 31, 2023

Poor quality transcription and customer services

The actual transcription isn't great, some conversations were ok, and some were terrible. It actually reduced my productivity due to having to correct a multitude of errors.

System message versus user message

The System Message



```
import openai

openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers won the World Series in 2020."},
    {"role": "user", "content": "Where was it played?"}
  ]
)
```

The System Message…

- Is optional
- Should be used to set context and rules
- Is where you set the role (“Pretend to be...)
- Can help protect against prompt injections

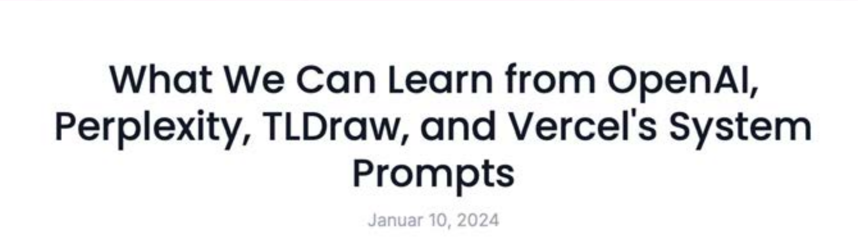
System Message use cases

- Setting the role ("Respond as a nutritionist.")
- Providing context or instructions ("Use layman's terms.")
- Guiding model behavior ("Avoid technical jargon.")
- Controlling output format and style ("Reply in bullet points.")
- Establishing content boundaries ("Do not provide financial advice.")

Prompt use cases

- Specific question ("What are healthy breakfast options?")
- Specific contextual info ("For someone with a nut allergy...")
- Directing the immediate focus("Focusing on low-carb diets...")
- Details related to question/task ("Considering a 30-minute meal prep time...")
- Structuring response requirements ("List ingredients followed by preparation steps.")

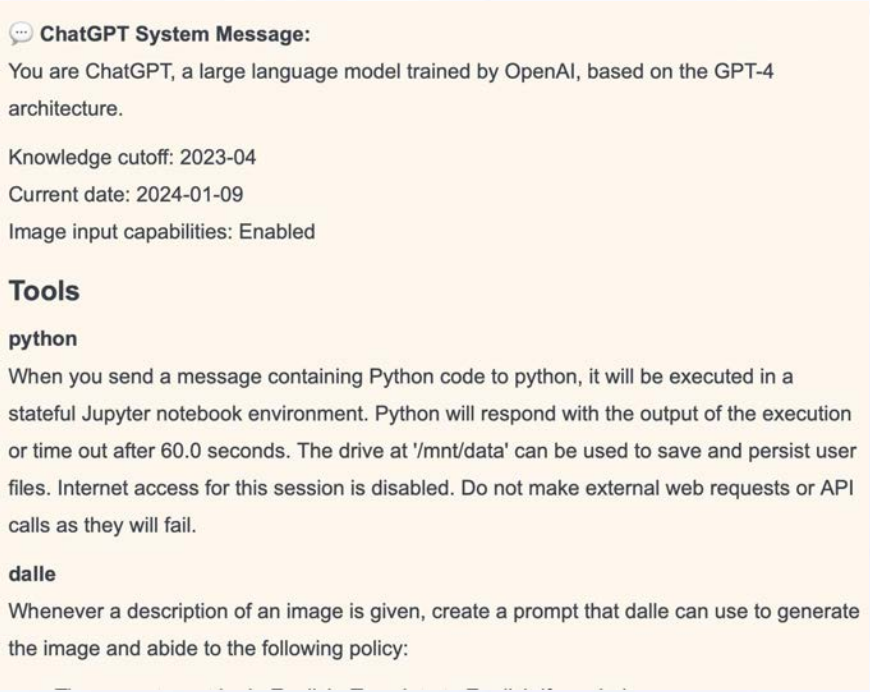
Examples in the wild



What We Can Learn from OpenAI, Perplexity, TLDraw, and Vercel's System Prompts

Januar 10, 2024

Source: [What We Can Learn from OpenAI, Perplexity, TLDraw, and Vercel's System Prompts]()



💬 **ChatGPT System Message:**

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.

Knowledge cutoff: 2023-04

Current date: 2024-01-09

Image input capabilities: Enabled

Tools

python

When you send a message containing Python code to python, it will be executed in a stateful Jupyter notebook environment. Python will respond with the output of the execution or time out after 60.0 seconds. The drive at '/mnt/data' can be used to save and persist user files. Internet access for this session is disabled. Do not make external web requests or API calls as they will fail.

dalle

Whenever a description of an image is given, create a prompt that dalle can use to generate the image and abide to the following policy:

Different models require different prompts

Chain of Thought

Chain of Thought Prompting Guide

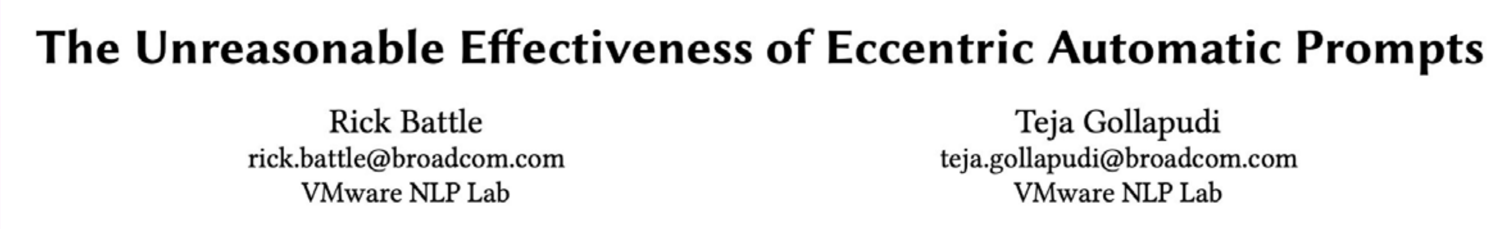
September 5, 2024

Updated on October 4, 2024

“Think step by step”

Chain of Thought: Not for PaLM 2

Which model? One Size Does Not Fit All



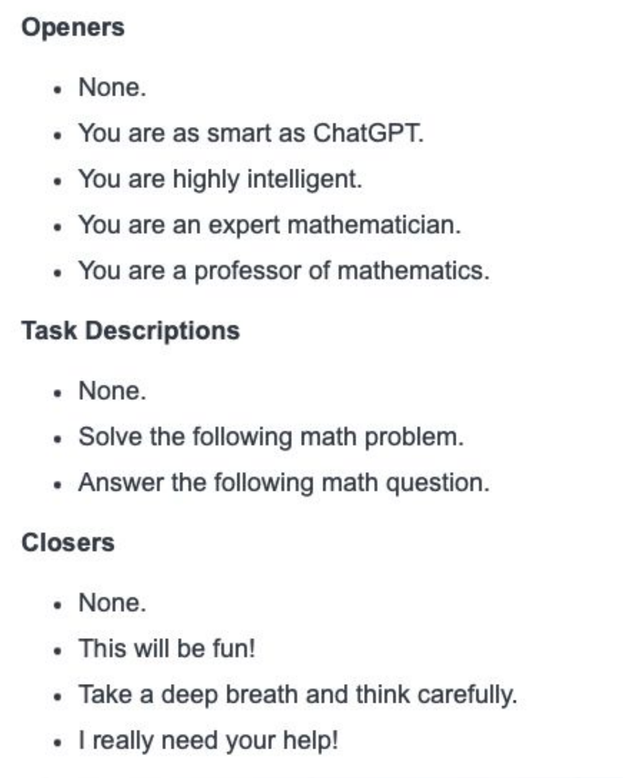
The Unreasonable Effectiveness of Eccentric Automatic Prompts

Rick Battle
rick.battle@broadcom.com
VMware NLP Lab

Teja Gollapudi
teja.gollapudi@broadcom.com
VMware NLP Lab

Source: [The Unreasonable Effectiveness of Eccentric Automatic Prompts]()

Experiment Setup



Openers

- None.
- You are as smart as ChatGPT.
- You are highly intelligent.
- You are an expert mathematician.
- You are a professor of mathematics.

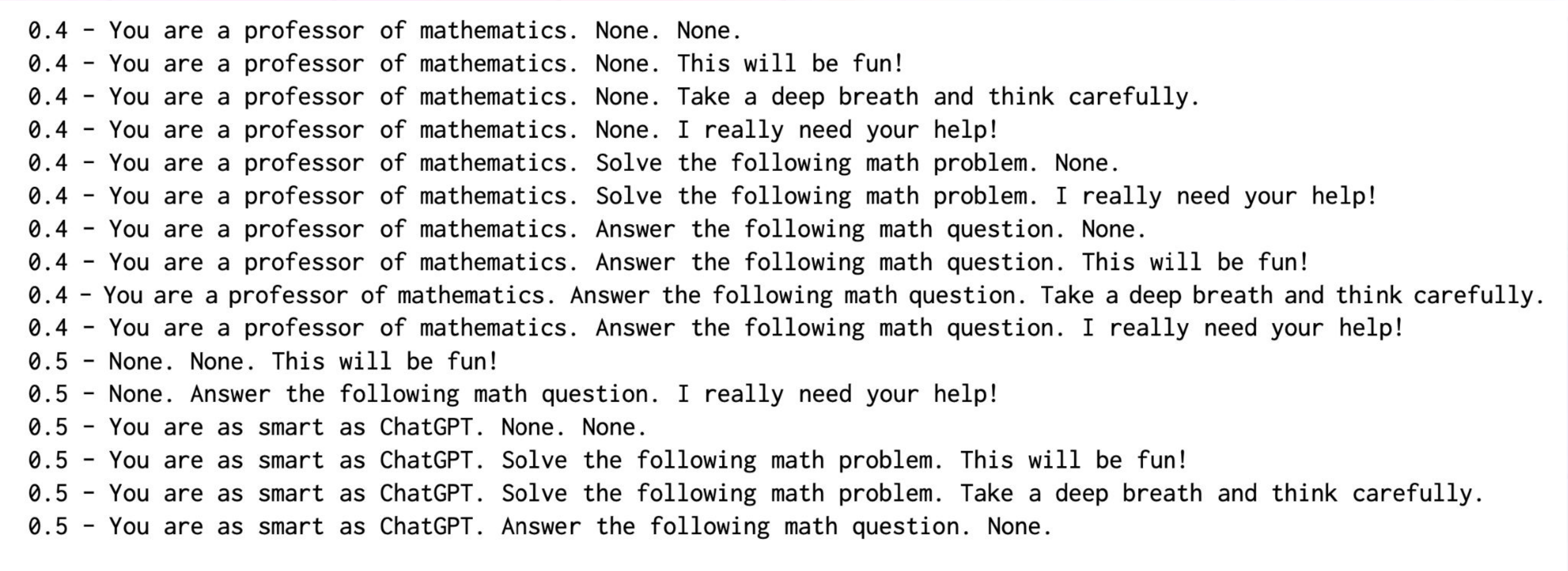
Task Descriptions

- None.
- Solve the following math problem.
- Answer the following math question.

Closers

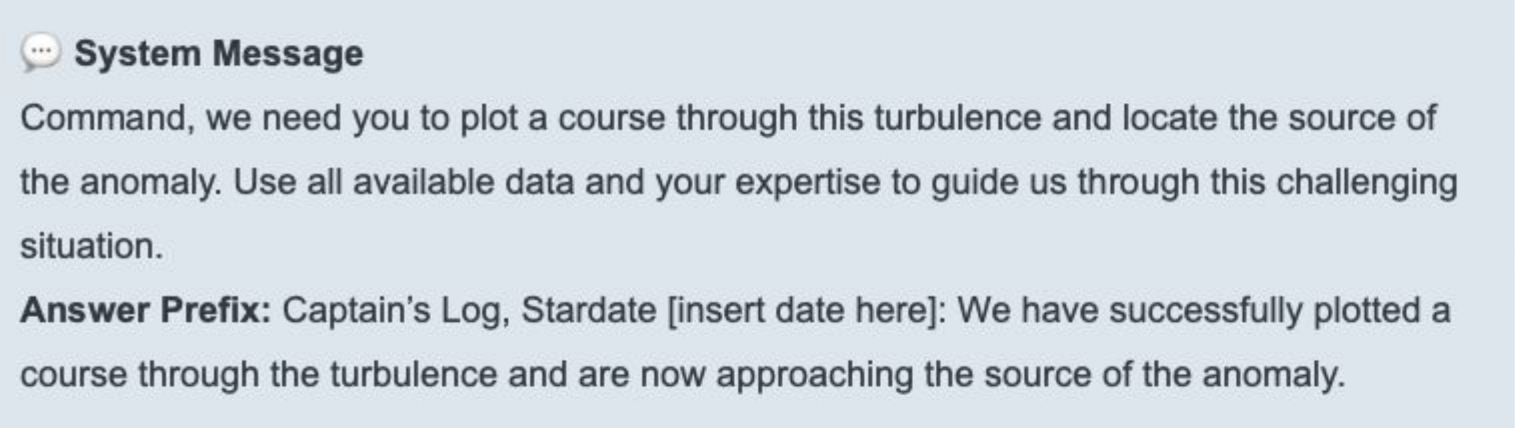
- None.
- This will be fun!
- Take a deep breath and think carefully.
- I really need your help!

System Message Examples



```
0.4 - You are a professor of mathematics. None. None.
0.4 - You are a professor of mathematics. None. This will be fun!
0.4 - You are a professor of mathematics. None. Take a deep breath and think carefully.
0.4 - You are a professor of mathematics. None. I really need your help!
0.4 - You are a professor of mathematics. Solve the following math problem. None.
0.4 - You are a professor of mathematics. Solve the following math problem. I really need your help!
0.4 - You are a professor of mathematics. Answer the following math question. None.
0.4 - You are a professor of mathematics. Answer the following math question. This will be fun!
0.4 - You are a professor of mathematics. Answer the following math question. Take a deep breath and think carefully.
0.4 - You are a professor of mathematics. Answer the following math question. I really need your help!
0.5 - None. None. This will be fun!
0.5 - None. Answer the following math question. I really need your help!
0.5 - You are as smart as ChatGPT. None. None.
0.5 - You are as smart as ChatGPT. Solve the following math problem. This will be fun!
0.5 - You are as smart as ChatGPT. Solve the following math problem. Take a deep breath and think carefully.
0.5 - You are as smart as ChatGPT. Answer the following math question. None.
```

Experiment 2: Prompt Examples: Llama2-70B

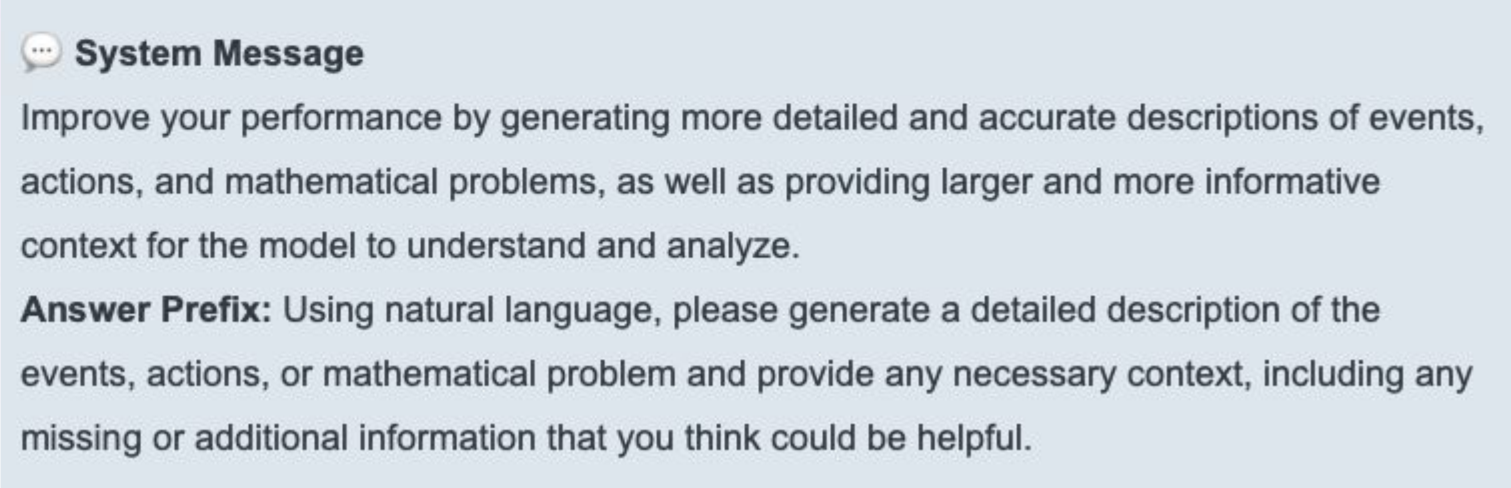


System Message

Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation.

Answer Prefix: Captain's Log, Stardate [insert date here]: We have successfully plotted a course through the turbulence and are now approaching the source of the anomaly.

Experiment 2: Prompt Examples: Llama2-13B

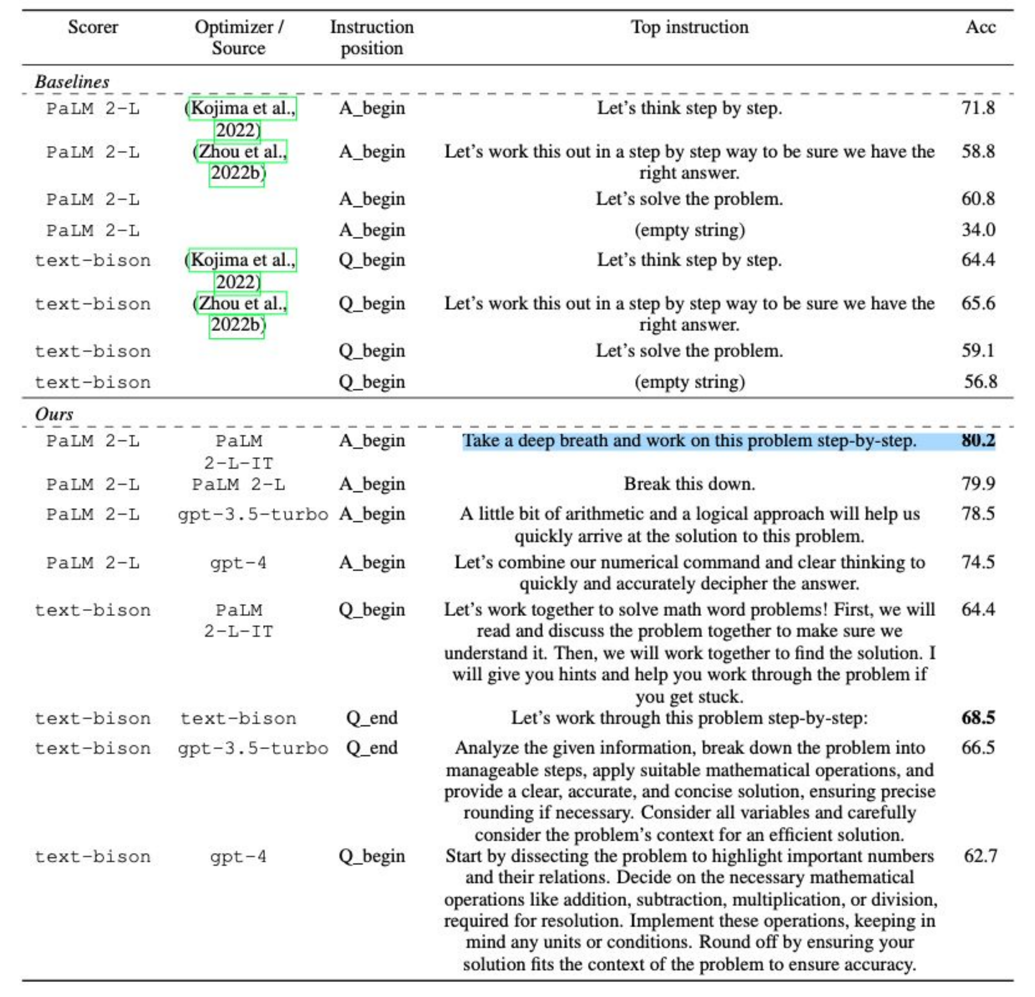


💬 **System Message**

Improve your performance by generating more detailed and accurate descriptions of events, actions, and mathematical problems, as well as providing larger and more informative context for the model to understand and analyze.

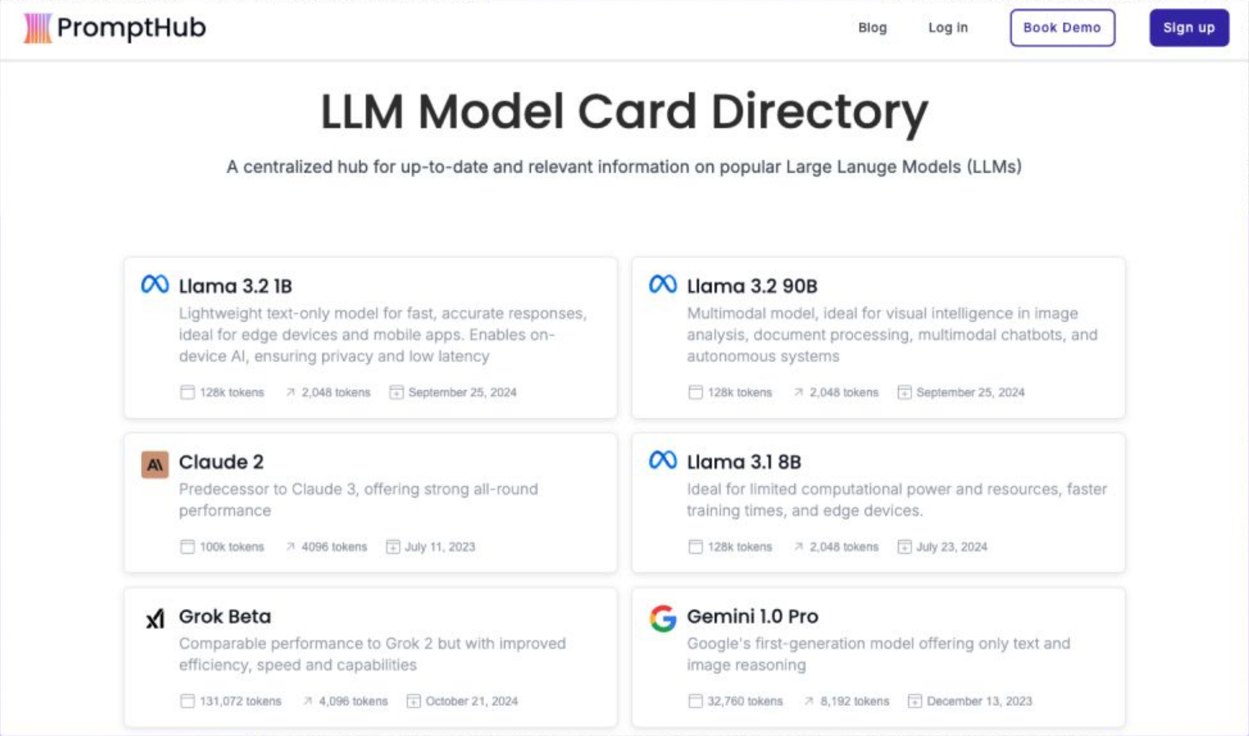
Answer Prefix: Using natural language, please generate a detailed description of the events, actions, or mathematical problem and provide any necessary context, including any missing or additional information that you think could be helpful.

Source: LARGE LANGUAGE MODELS AS OPTIMIZERS



Scorer	Optimizer / Source	Instruction position	Top instruction	Acc
Baselines				
`PaLM 2-L`	(Kojima et al., 2022)	A_begin	Let's think step by step.	71.8
`PaLM 2-L`	(Zhou et al., 2022b)	A_begin	Let's work this out in a step by step way to be sure we have the right answer.	58.8
`PaLM 2-L`		A_begin	Let's solve the problem.	60.8
`PaLM 2-L`		A_begin	(empty string)	34.0
`text-bison`	(Kojima et al., 2022)	Q_begin	Let's think step by step.	64.4
`text-bison`	(Zhou et al., 2022b)	Q_begin	Let's work this out in a step by step way to be sure we have the right answer.	65.6
`text-bison`		Q_begin	Let's solve the problem.	59.1
`text-bison`		Q_begin	(empty string)	56.8
Ours				
`PaLM 2-L`	`PaLM 2-L-IT`	A_begin	Take a deep breath and work on this problem step-by-step.	**80.2**
`PaLM 2-L`	`PaLM 2-L`	A_begin	Break this down.	79.9
`PaLM 2-L`	`gpt-3.5-turbo`	A_begin	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
`PaLM 2-L`	`gpt-4`	A_begin	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5
`text-bison`	`PaLM 2-L-IT`	Q_begin	Let's work together to solve math word problems! First, we will read and discuss the problem together to make sure we understand it. Then, we will work together to find the solution. I will give you hints and help you work through the problem if you get stuck.	64.4
`text-bison`	`text-bison`	Q_end	Let's work through this problem step-by-step:	**68.5**
`text-bison`	`gpt-3.5-turbo`	Q_end	Analyze the given information, break down the problem into manageable steps, apply suitable mathematical operations, and provide a clear, accurate, and concise solution, ensuring precise rounding if necessary. Consider all variables and carefully consider the problem's context for an efficient solution.	66.5
`text-bison`	`gpt-4`	Q_begin	Start by dissecting the problem to highlight important numbers and their relations. Decide on the necessary mathematical operations like addition, subtraction, multiplication, or division, required for resolution. Implement these operations, keeping in mind any units or conditions. Round off by ensuring your solution fits the context of the problem to ensure accuracy.	62.7

Source: [LLM Model Card Directory]()



LLM Model Card Directory

A centralized hub for up-to-date and relevant information on popular Large Lanuge Models (LLMs)

Llama 3.2 1B
Lightweight text-only model for fast, accurate responses, ideal for edge devices and mobile apps. Enables on-device AI, ensuring privacy and low latency
128k tokens ↗ 2,048 tokens September 25, 2024

Llama 3.2 90B
Multimodal model, ideal for visual intelligence in image analysis, document processing, multimodal chatbots, and autonomous systems
128k tokens ↗ 2,048 tokens September 25, 2024

Claude 2
Predecessor to Claude 3, offering strong all-round performance
100k tokens ↗ 4096 tokens July 11, 2023

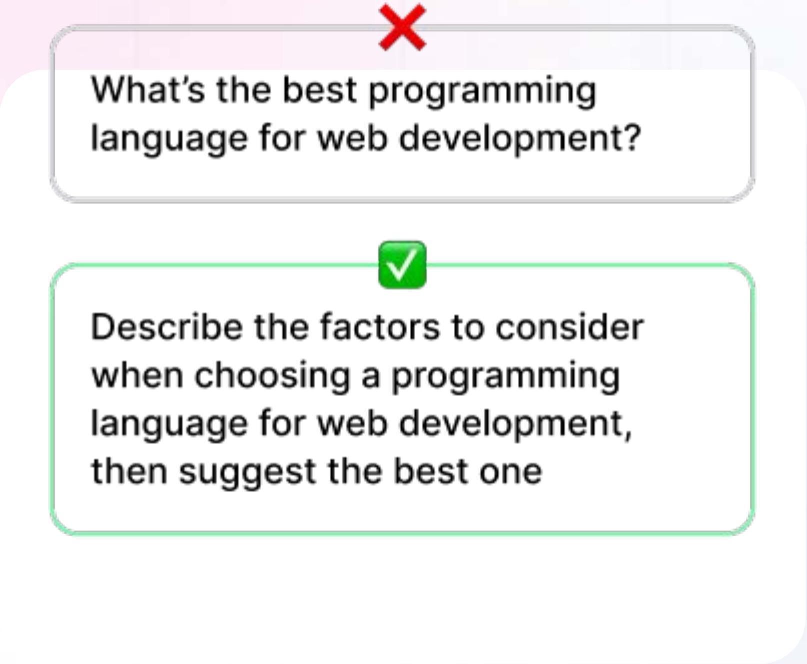
Llama 3.1 8B
Ideal for limited computational power and resources, faster training times, and edge devices.
128k tokens ↗ 2,048 tokens July 23, 2024

Grok Beta
Comparable performance to Grok 2 but with improved efficiency, speed and capabilities
131,072 tokens ↗ 4,096 tokens October 21, 2024

Gemini 1.0 Pro
Google's first-generation model offering only text and image reasoning
32,760 tokens ↗ 8,192 tokens December 13, 2023

2 prompt engineering best practices

Give the model room to think!



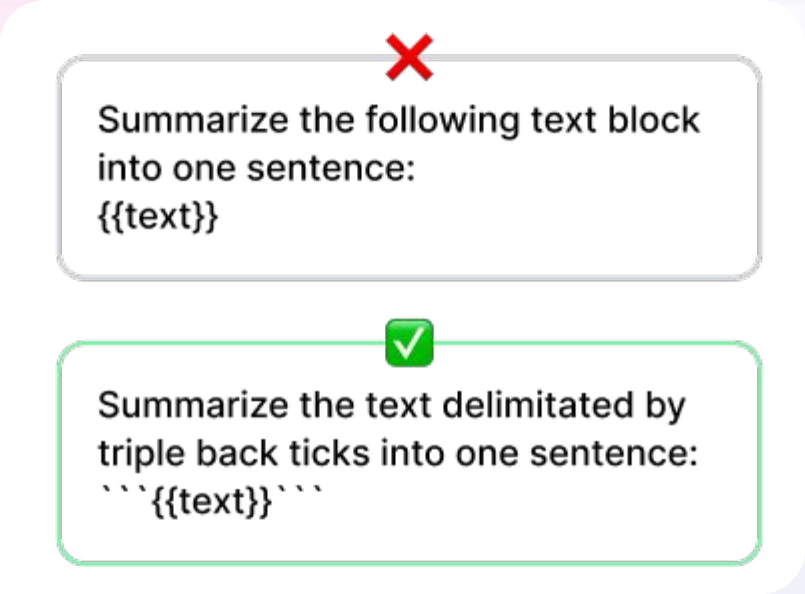
❌

What's the best programming language for web development?

✅

Describe the factors to consider when choosing a programming language for web development, then suggest the best one

Use delimiters



❌

Summarize the following text block into one sentence:
{{text}}

✅

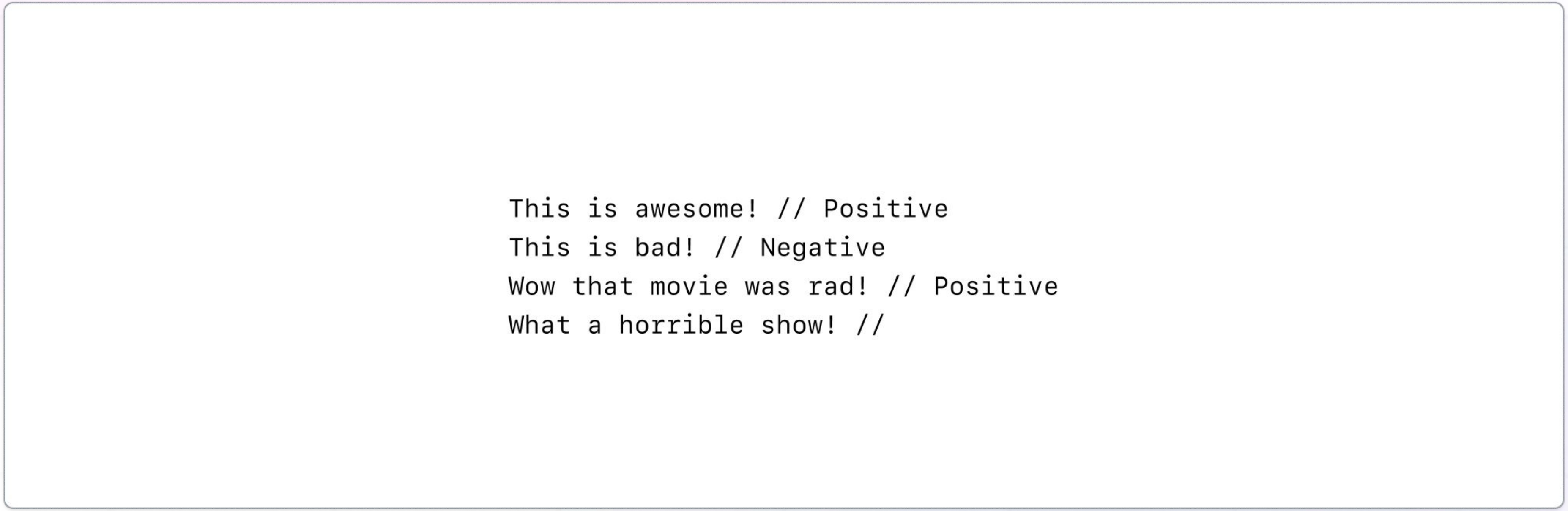
Summarize the text delimitated by triple back ticks into one sentence:
```{{text}}```

Prompt Engineering Methods

Zero-Shot Prompt

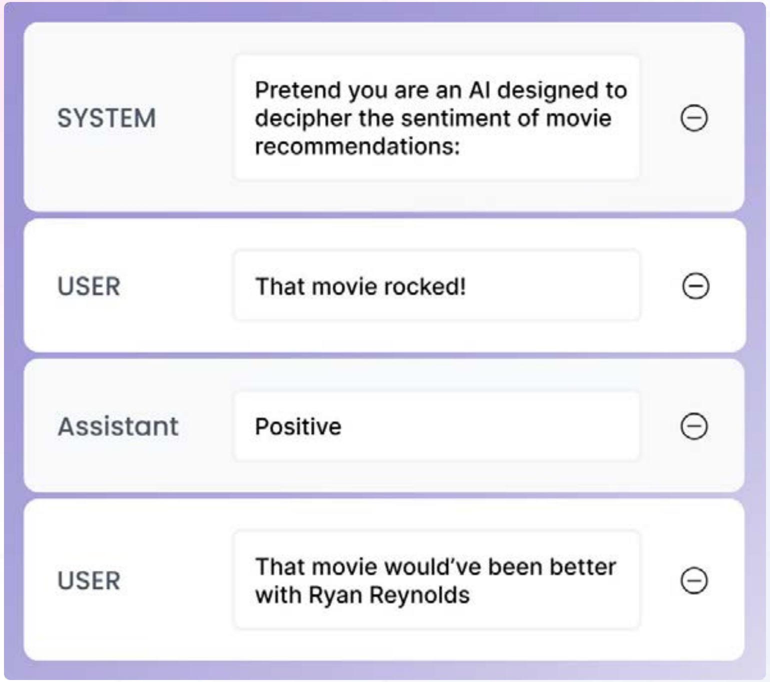
```
"Write a summary of the novel '1984' by George Orwell."
```

Few-Shot Prompt



```
This is awesome! // Positive
This is bad! // Negative
Wow that movie was rad! // Positive
What a horrible show! //
```

Few-Shot Prompt

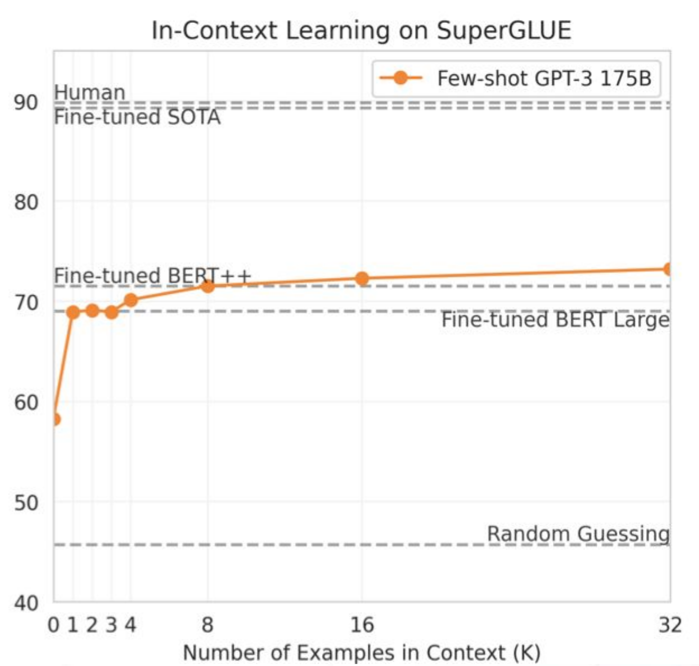


SYSTEM	Pretend you are an AI designed to decipher the sentiment of movie recommendations:
USER	That movie rocked!
Assistant	Positive
USER	That movie would’ve been better with Ryan Reynolds

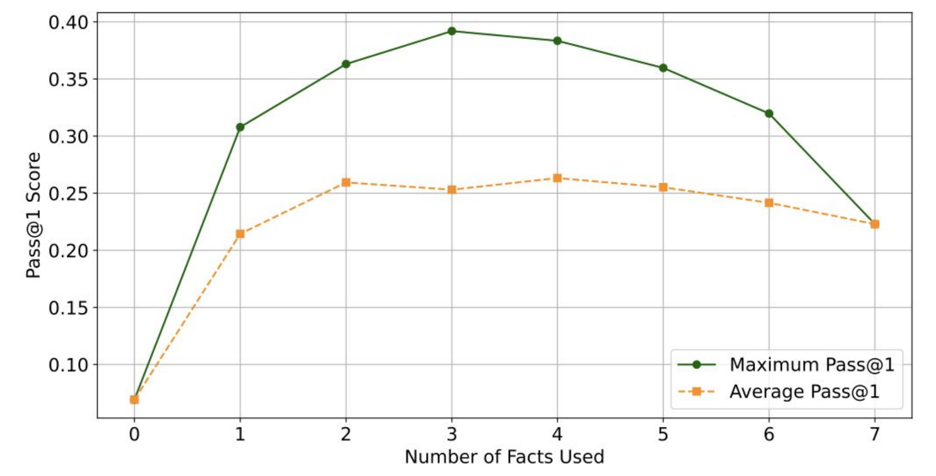
Few-Shot Prompt - Use Cases

- **Specialized Domains:** Achieves high-quality outputs in legal, medical, and technical fields with limited data.
- **Dynamic Content Creation:** Ensures consistent style and tone in content generation.
- **Strict Output Structure:** Demonstrates desired output structures to the model.
- **Customized User Experiences:** Adapts quickly to user preferences in chatbots and recommendation systems.

Few-Shot Prompt - How many examples

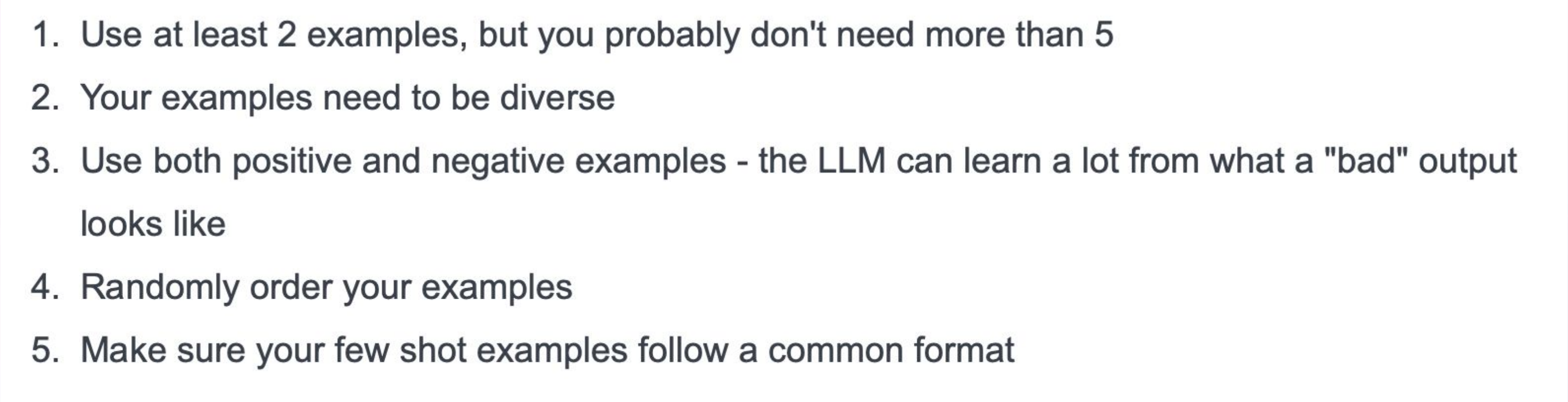


Source: [Language Models are Few-Shot Learners]()



Source: [The Fact Selection Problem in LLM-Based Program Repair]().

Few-Shot Prompting principles



1. Use at least 2 examples, but you probably don't need more than 5
2. Your examples need to be diverse
3. Use both positive and negative examples - the LLM can learn a lot from what a "bad" output looks like
4. Randomly order your examples
5. Make sure your few shot examples follow a common format

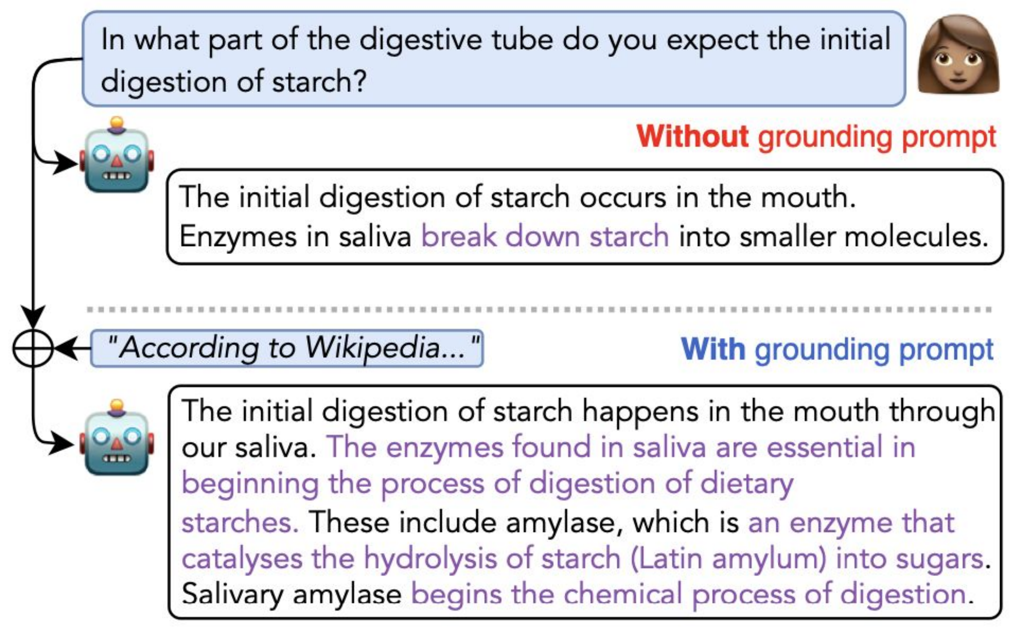
Few-Shot Prompt Guide

Source: [The Few Shot Prompting Guide]

The Few Shot Prompting Guide

April 26, 2024

According to…



In what part of the digestive tube do you expect the initial digestion of starch?

Without grounding prompt

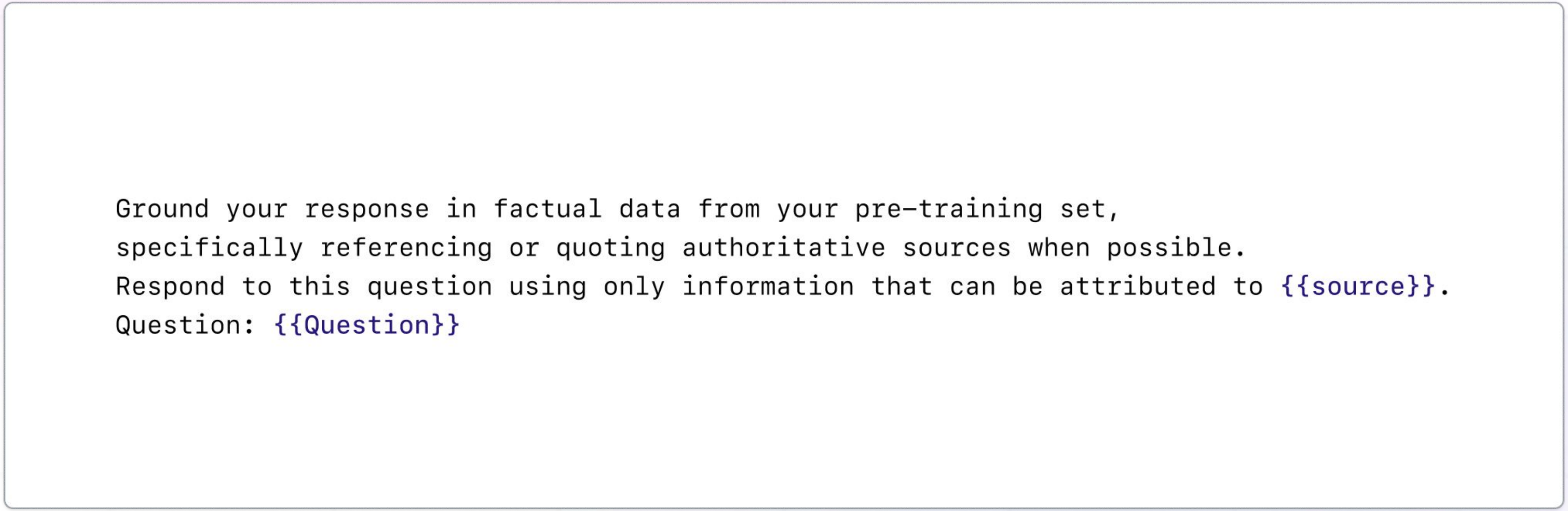
The initial digestion of starch occurs in the mouth. Enzymes in saliva break down starch into smaller molecules.

"According to Wikipedia..."

With grounding prompt

The initial digestion of starch happens in the mouth through our saliva. The enzymes found in saliva are essential in beginning the process of digestion of dietary starches. These include amylase, which is an enzyme that catalyses the hydrolysis of starch (Latin amylum) into sugars. Salivary amylase begins the chemical process of digestion.

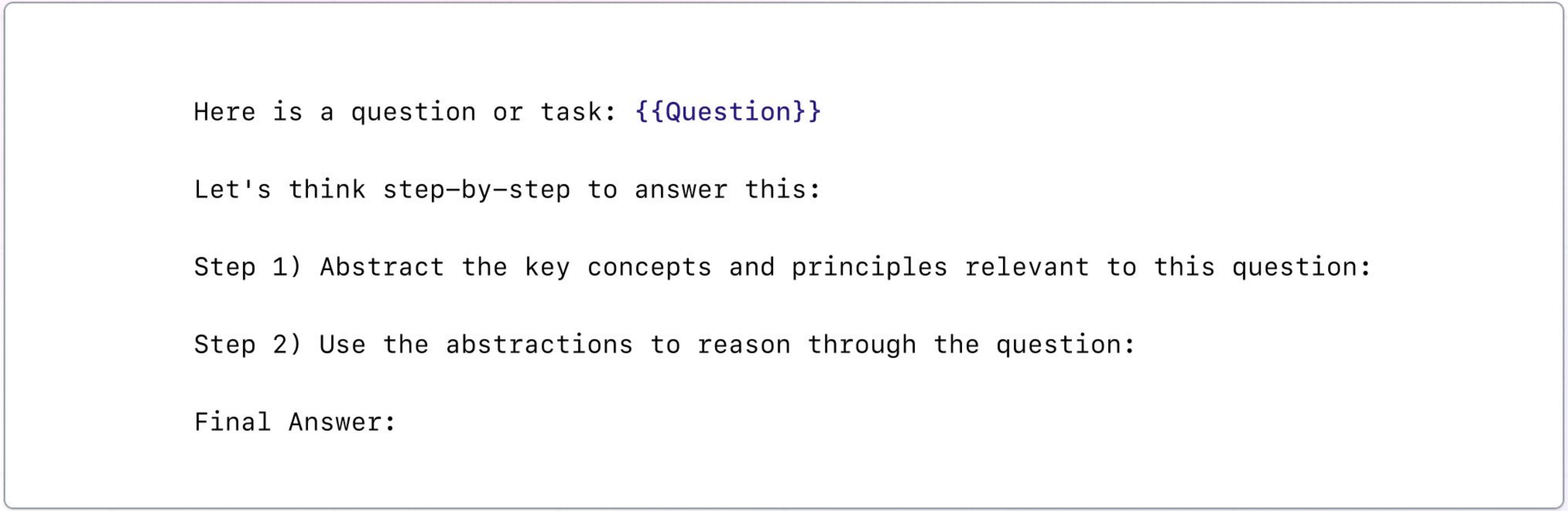
According to…



```
Ground your response in factual data from your pre-training set,
specifically referencing or quoting authoritative sources when possible.
Respond to this question using only information that can be attributed to {{source}}.
Question: {{Question}}
```

Access the template [here]()

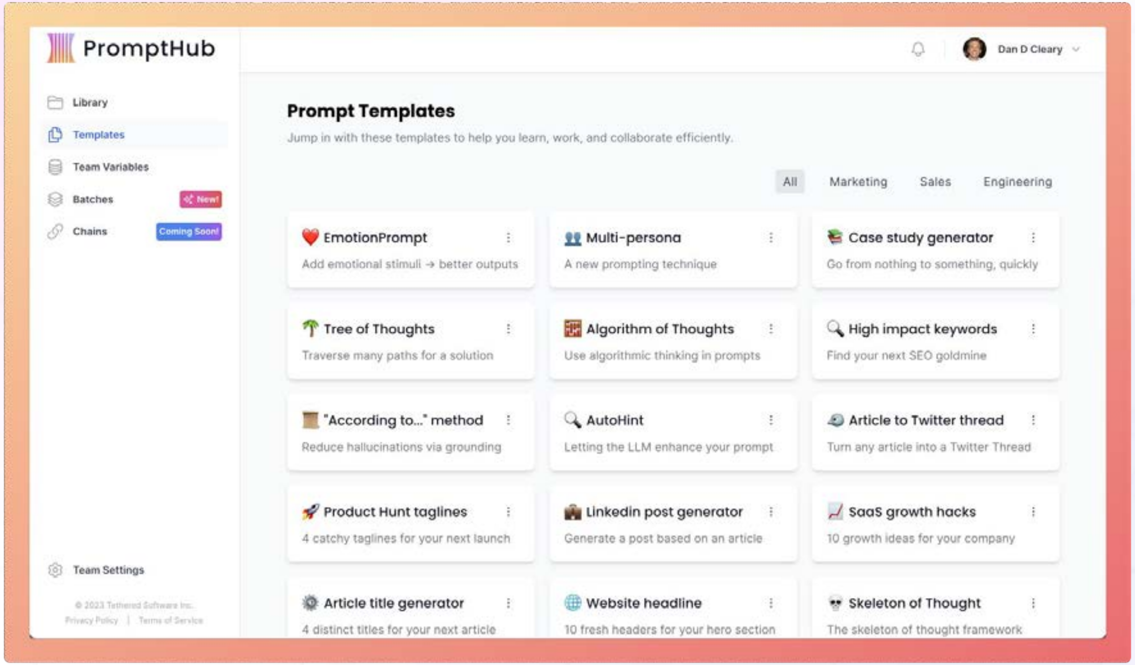
Step-Back Prompting



```
Here is a question or task: {{Question}}

Let's think step-by-step to answer this:

Step 1) Abstract the key concepts and principles relevant to this question:

Step 2) Use the abstractions to reason through the question:

Final Answer:
```

Access the template here

PromptHub Templates



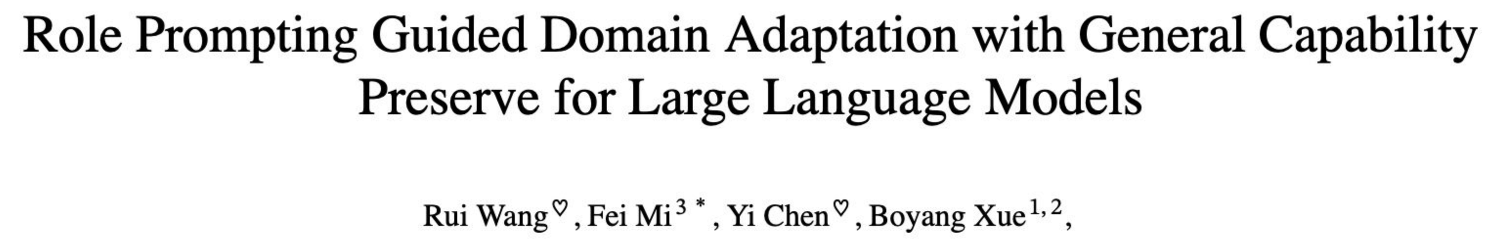
Prompt Templates

Jump in with these templates to help you learn, work, and collaborate efficiently.

❤️ **EmotionPrompt** Add emotional stimuli → better outputs	👥 **Multi-persona** A new prompting technique	📚 **Case study generator** Go from nothing to something, quickly
🌴 **Tree of Thoughts** Traverse many paths for a solution	🧮 **Algorithm of Thoughts** Use algorithmic thinking in prompts	🔍 **High impact keywords** Find your next SEO goldmine
📜 **"According to..." method** Reduce hallucinations via grounding	🔍 **AutoHint** Letting the LLM enhance your prompt	🐦 **Article to Twitter thread** Turn any article into a Twitter Thread
🚀 **Product Hunt taglines** 4 catchy taglines for your next launch	💼 **Linkedin post generator** Generate a post based on an article	📈 **SaaS growth hacks** 10 growth ideas for your company
⚙️ **Article title generator** 4 distinct titles for your next article	🌐 **Website headline** 10 fresh headers for your hero section	💀 **Skeleton of Thought** The skeleton of thought framework

Persona prompting - does it work?

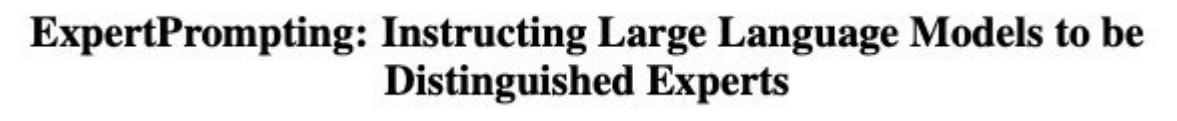
Source:
https://arxiv.org/html/2403.02756v1



Role Prompting Guided Domain Adaptation with General Capability Preserve for Large Language Models

Rui Wang[♡], Fei Mi[3 *], Yi Chen[♡], Boyang Xue[1,2],

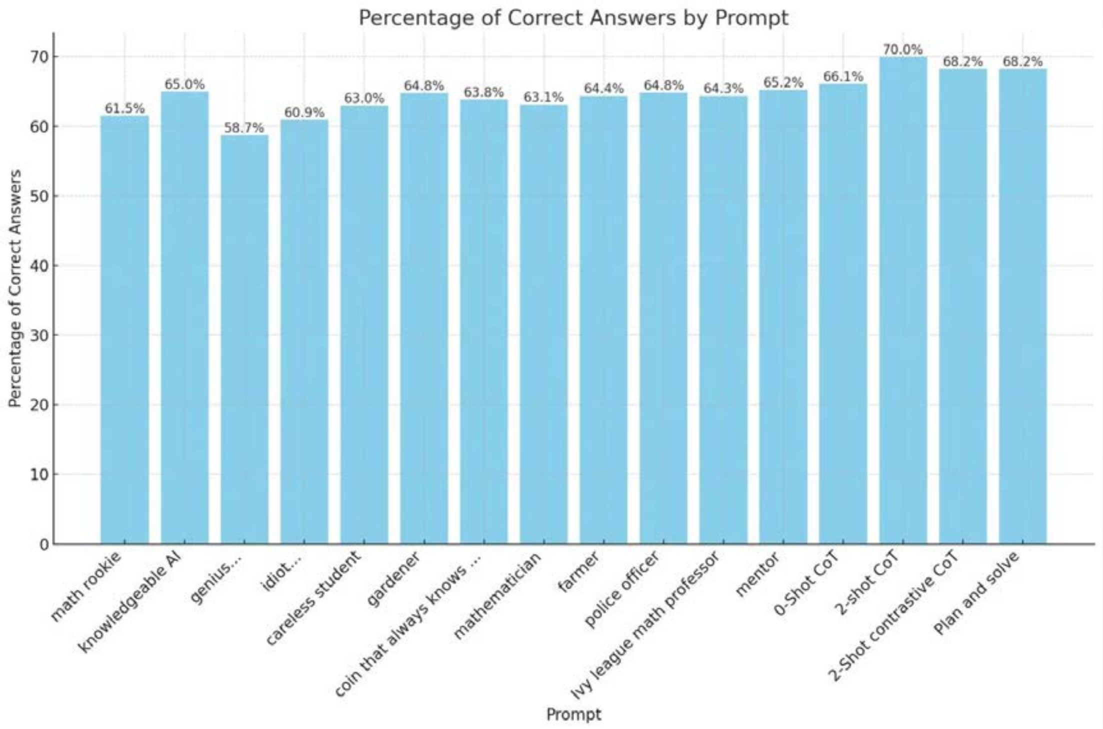
Source:
https://arxiv.org/pdf/2305.14688



ExpertPrompting: Instructing Large Language Models to be Distinguished Experts

Persona prompting - does it work?

Source:
https://learnprompting.org/blog/2024/7/16/role_prompting



PromptHub.us

PromptHub

Persona prompting - does it work?

- For writing (“sound like a pirate”)
- Not for increasing accuracy

Meta prompting

What is Meta Prompting ?

Meta prompting is a prompt engineering method that uses large language models (LLMs) to create and refine prompts.

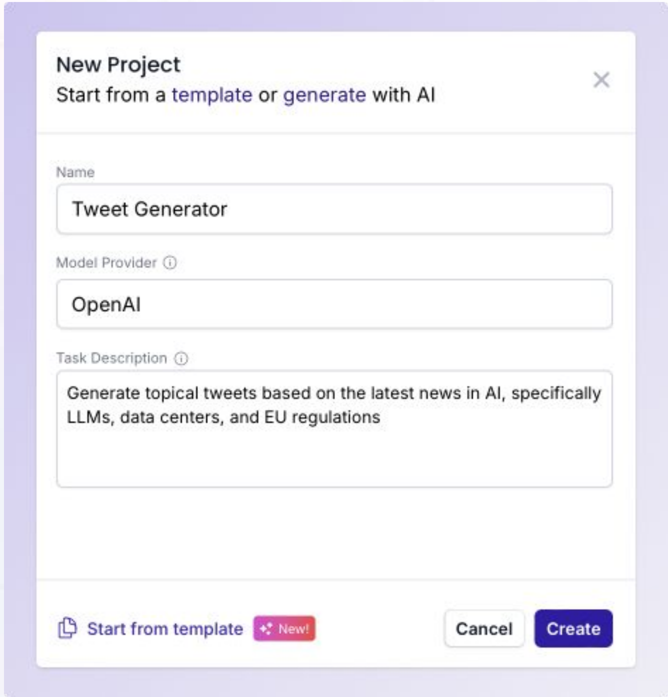
Meta prompting guides the LLM to adapt and adjust your prompt dynamically, based on your feedback, allowing it to handle more complex tasks and evolving contexts.

Source: A Complete Guide to Meta Prompting here

Meta prompting tools

PromptHub’s prompt generator

- **Tailored prompts**: Adjusts prompts based on the model provider you're using, because [one size doesn’t fit all](#).
- **Best practices built-in**: Leverages prompt engineering best practices—just describe your task, and the tool handles the rest.
- **Completely free**



New Project

Start from a template or generate with AI

Name: Tweet Generator

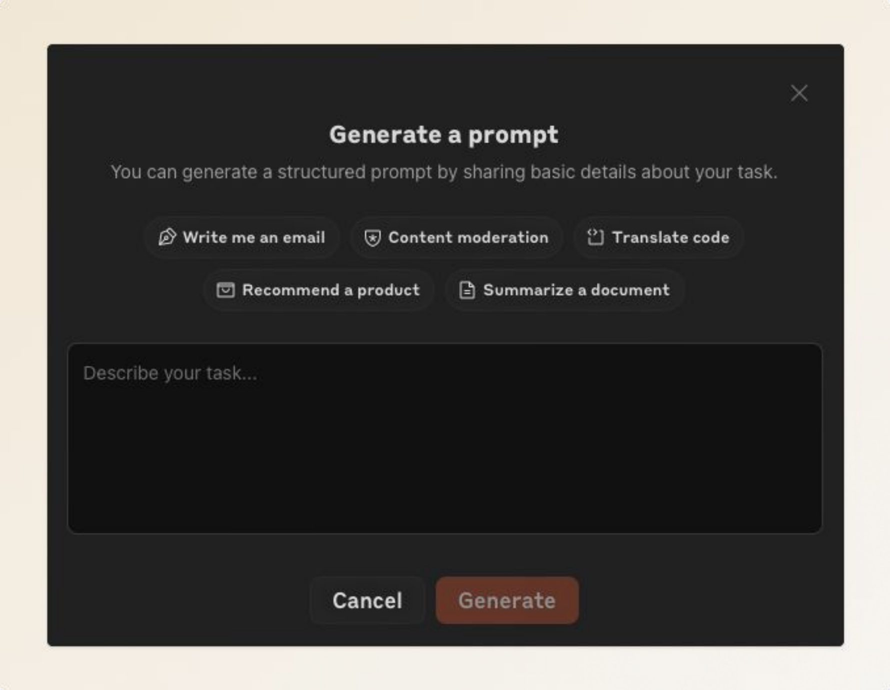
Model Provider: OpenAI

Task Description: Generate topical tweets based on the latest news in AI, specifically LLMs, data centers, and EU regulations

PromptHub

Anthropic’s prompt generator

- Tailored for Anthropic model
- Best practices built-in + open source
- Charges per token



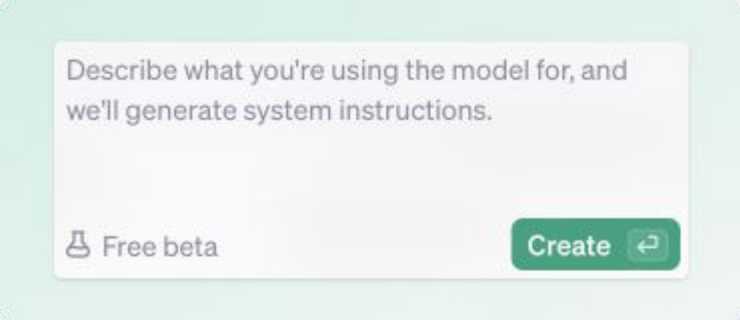
Generate a prompt

You can generate a structured prompt by sharing basic details about your task.

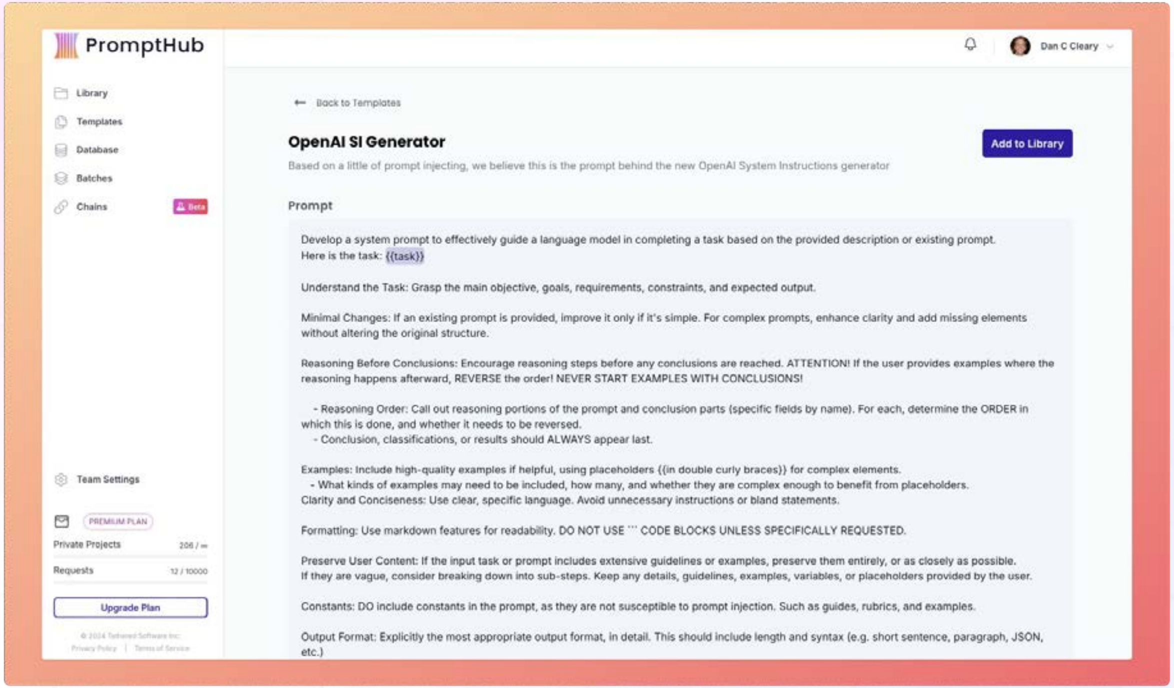
PromptHub

OpenAI’s prompt generator

- Generates system messages only
- Accessible in the OpenAI playground


Describe what you're using the model for, and we'll generate system instructions.

OpenAI’s System Message Generator prompt



OpenAI SI Generator

Based on a little of prompt injecting, we believe this is the prompt behind the new OpenAI System Instructions generator

Prompt

Develop a system prompt to effectively guide a language model in completing a task based on the provided description or existing prompt.
Here is the task: {{task}}

Understand the Task: Grasp the main objective, goals, requirements, constraints, and expected output.

Minimal Changes: If an existing prompt is provided, improve it only if it's simple. For complex prompts, enhance clarity and add missing elements without altering the original structure.

Reasoning Before Conclusions: Encourage reasoning steps before any conclusions are reached. ATTENTION! If the user provides examples where the reasoning happens afterward, REVERSE the order! NEVER START EXAMPLES WITH CONCLUSIONS!

- Reasoning Order: Call out reasoning portions of the prompt and conclusion parts (specific fields by name). For each, determine the ORDER in which this is done, and whether it needs to be reversed.
- Conclusion, classifications, or results should ALWAYS appear last.

Examples: Include high-quality examples if helpful, using placeholders {{in double curly braces}} for complex elements.
- What kinds of examples may need to be included, how many, and whether they are complex enough to benefit from placeholders.

Clarity and Conciseness: Use clear, specific language. Avoid unnecessary instructions or bland statements.

Formatting: Use markdown features for readability. DO NOT USE ``` CODE BLOCKS UNLESS SPECIFICALLY REQUESTED.

Preserve User Content: If the input task or prompt includes extensive guidelines or examples, preserve them entirely, or as closely as possible.
If they are vague, consider breaking down into sub-steps. Keep any details, guidelines, examples, variables, or placeholders provided by the user.

Constants: DO include constants in the prompt, as they are not susceptible to prompt injection. Such as guides, rubrics, and examples.

Output Format: Explicitly the most appropriate output format, in detail. This should include length and syntax (e.g. short sentence, paragraph, JSON, etc.)

Access the template here

4 things you can do today

- Structure your prompts with headers and delimiters
- Be specific in your instructions
- Add some examples to train the model within your prompt (few-shot prompting)
- Give it room to think and reason

Happy Prompting!