



# Prompt Engineering - an Art, a Science, or your next Job Title?

**Maxim Salnikov**

Developer Productivity Lead

Microsoft

# I'm Maxim Salnikov

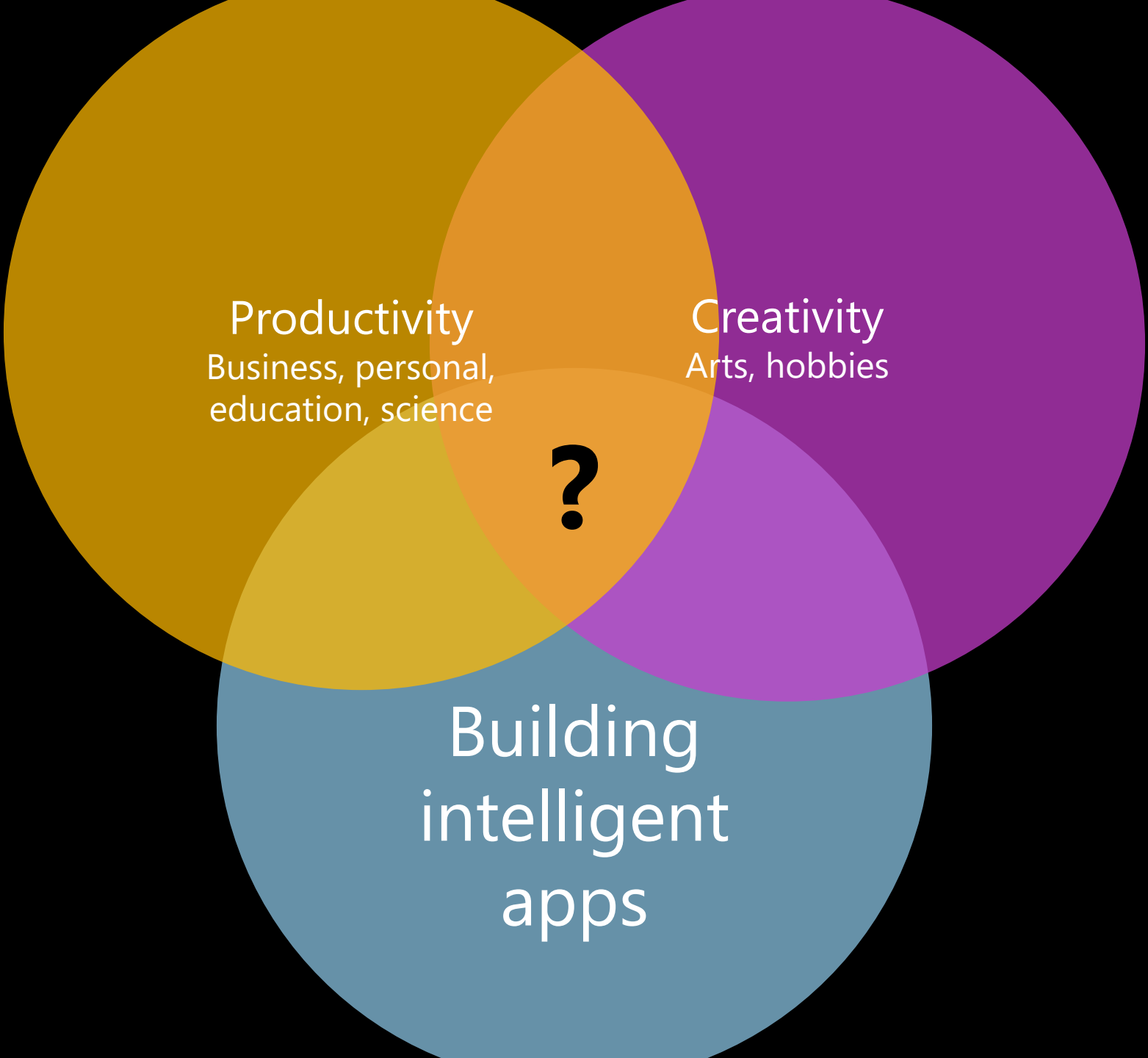
*Helping developers to succeed with the Dev Tools, Cloud & AI in Microsoft*



- Building on web platform since 90s
- Organizing developer communities and technical conferences
- Speaking, training, blogging: **Webdev**, **Cloud**, Generative AI, **Prompt Engineering**

**> 1 Billion**

generative AI users



Productivity  
Business, personal,  
education, science

Creativity  
Arts, hobbies

?

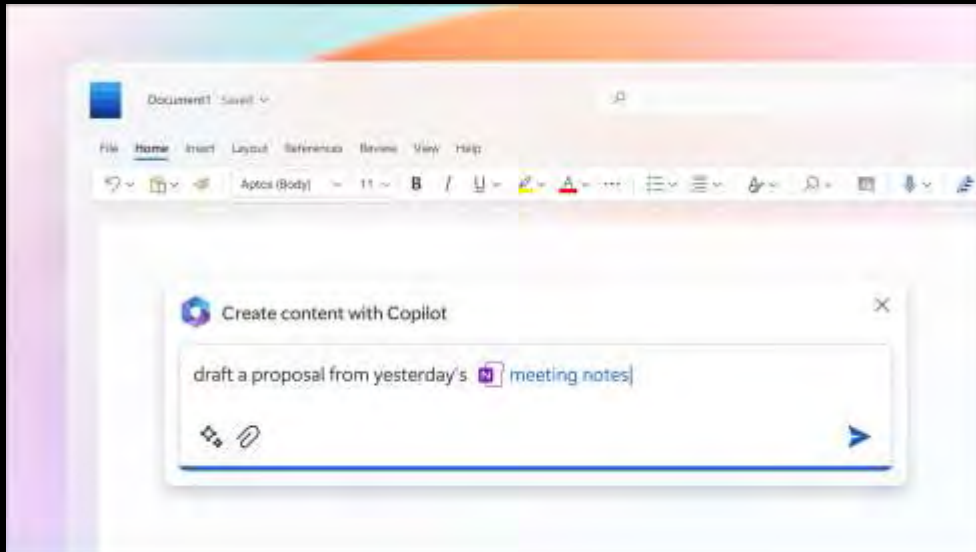
Building  
intelligent  
apps

# One common thing...



Midjourney Bot ✓ BOT Today at 1:24 PM

A white cute alien giraffe in astronaut suit. View of full body from different angles, in the style of childrens book illustration, white background --ar 16:8 - Variations (Strong) by



webmaxru

Create a webpage with 3 columns, header, and footer



GitHub Copilot

To create a webpage with 3 columns, header, and footer, you can use HTML and CSS. Here's an example code snippet that you can use:

```
<!DOCTYPE html>
```

You are an expert in creating domain-specific languages with syntax as valid javascript. You are also an expert in advanced DJ techniques on mixing music tracks together. Create a DSL for DJs to describe and automate mixing multiple tracks in real-time with basic and advanced DJ techniques included.



# Emerging model landscape

- Anthropic Claude 3
- Databricks DBRX
- Meta Llama 3
- Mistral Mixtral 8x22B
- Google Gemini 1.5
- RWKV Eagle 7B
- Microsoft Phi-2
- X Grok-1
- OpenAI GPT-4
- Cerebras Cerebras-GPT
- Cohere Command R
- Bloomberg BloombergGPT
- LAION OpenAssistant
- ...

GPT-o1

# Prompt Engineering



Process of designing, refining, and optimizing input prompts to guide a model toward producing more accurate outputs while keeping cost efficiency

# Prompt anatomy

Write a product launch email for new headphones.

This product is AI-powered and priced at \$79.99.

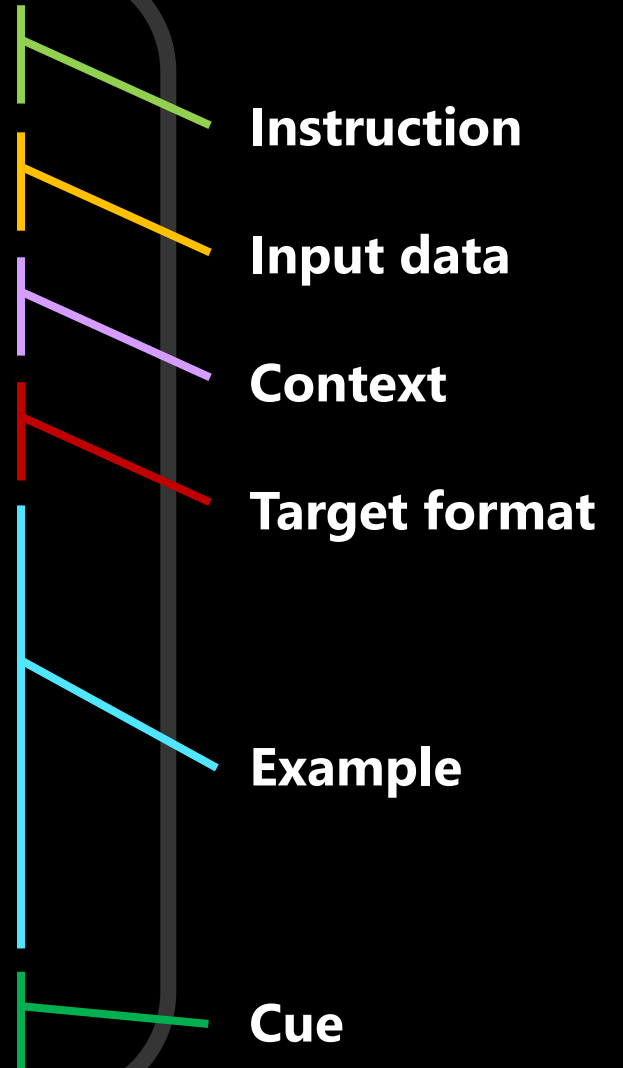
The tone is friendly and exciting.

Output as a JSON object with "subject" and "body" fields.

Example:

```
{  
  "subject": "Introducing Our New AI-Powered Headphones!",  
  "body": "We are thrilled to announce the launch of our new AI-  
powered headphones"  
}
```

JSON output:





# Tokens

Tokens	Characters
84	364

Write a product launch email for new headphones.

This product is AI-powered and priced at \$79.99.

The tone is friendly and exciting.

Output as a JSON object with "subject" and "body" fields.

Example:

```
{  
  "subject": "Introducing Our New AI-Powered Headphones!",  
  "body": "We are thrilled to announce the launch of our new AI-powered  
headphones"  
}
```

Output:

# Tokens and cost efficiency

(Prompt length + Completion length) x Number of API calls

Models	Context	Input (Per 1,000 tokens)	Output (Per 1,000 tokens)
GPT-3.5-Turbo-0125	16K	\$0.0005	\$0.0015
GPT-3.5-Turbo-Instruct	4K	\$0.0015	\$0.002
GPT-4-Turbo	128K	N/A	N/A
GPT-4-Turbo-Vision	128K	N/A	N/A
GPT-4	8K	\$0.03	\$0.06
GPT-4	32K	\$0.06	\$0.12

<https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service>

# Model selection strategy

1. Write the best prompt for the most expensive model
2. Downgrade to the cheaper model
3. Tune the prompt if needed



# Using multiple models

## 1. Mix the models

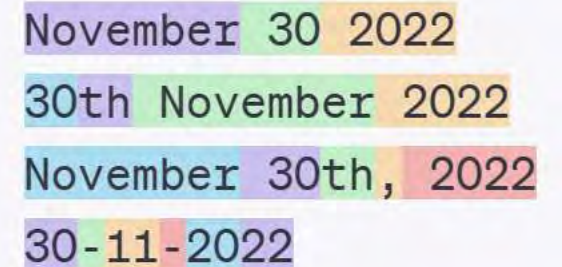
- Expensive: generation
- Cheap: summarization, classification

## 2. Chain the models

- Summarize text before sending to the expensive one

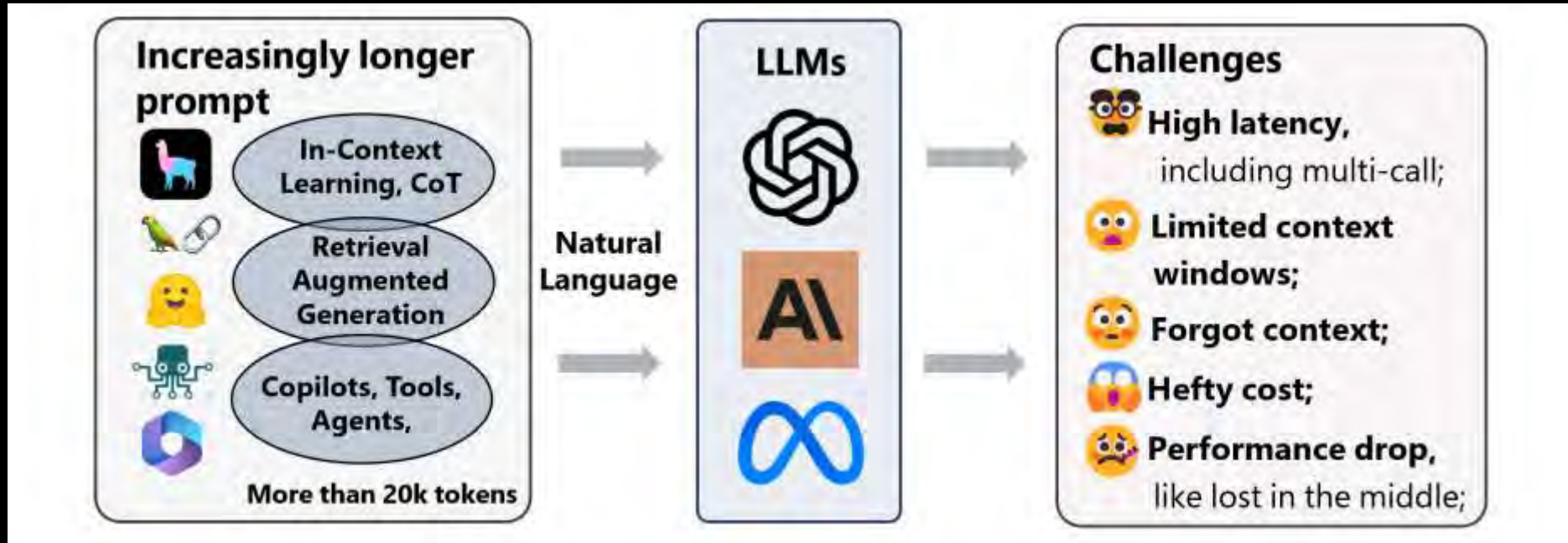
# Space efficiency

- Use whitespaces carefully
- Try various formats
- Tabular | data | is | space-efficient
- Language makes difference



November 30 2022  
30th November 2022  
November 30th, 2022  
30-11-2022

# LLMLingua: Prompt compressor



- Compact, well-trained language model (e.g., GPT2-small, LLaMA-7B) to identify and remove non-essential tokens in prompts
- Achieving up to 20x compression with minimal performance loss

# Compressed

Write product launch email new headphones.

product AI - powered priced \$ 79. 99.

tone friendly exciting.

Output JSON object subject " body " fields.

Example :

```
{  
  subject " Introducing New AI - Powered Headphones! ",  
  body " thrilled announce launch new AI - powered headphones  
}
```

Output :

70 vs 84 tokens (17%)

# General recommendations

- Be specific and clear
- Be descriptive, use examples
- Order matters
- Double down
- Give the model an "out"



# Technical recommendations

- Start with clear instructions and repeat the instructions at the end
- Add clear syntax: headers, separators
- Break the task down
- Adjust parameters: Temperature and Top probabilities

# Zero-shot VS few-shot prompts

## Zero-shot

The model predicts the answer given only a natural language description of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: What packages do you have in auto insurance?

Classified topic:

## Few-shot

In addition to the task description, the model sees a few examples of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?

Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.

Classified topic: 1

Customer question: {insert new question here}

Classified topic:

# Chain-of-Thought (Reasoning)

## Before

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

The annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day is **8,000,000** liters per year.

## After

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

**Let's think step by step and explain the calculation step by step.**

- Step 1: Calculate the total number of days in a year:  
There are 365 days in a year.
- Step 2: Calculate the total number of days the family is at home:  
The family is at home an average of 200 days per year.
- Step 3: Calculate the total water demand of the family:  
The family uses an average of 100 liters of water per day. Therefore, the total water demand of the family is  $100 \text{ liters} \times 200 \text{ days} = 20,000 \text{ liters per year}$ .
- Step 4: Calculate the annual water demand of the single-family household:  
The annual water demand of the single-family household is **20,000** liters per year.

# Prompt Chaining

## Entity Extraction

Please **extract entities** from the following news article: 'The new iPhone model is set to be released next month. It has been highly anticipated by Apple fans and is expected to feature a larger screen and improved camera

```
{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}
```

## Summarization

Please **summarize** the information about the product

```
[{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}]
```

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

## Sentiment Analysis

Please provide a **sentiment** for the following text:

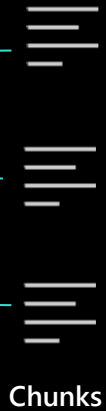
The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

**Positive**

# Prompt Chunking



Large Text  
Beyond  
Token  
Limit



Please summarize the following text:

Climate change is a pressing issue that needs to be addressed immediately. The increasing levels of carbon dioxide in the atmosphere are causing the Earth's temperature to rise, which is leading to more frequent and severe natural disasters. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources.

...  
+ many more sentences

**Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations**

Please summarize and extract topic:

**Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations**

Summary of chunk 2

Summary of Chunk 3

...

**The text is discussing global warming, which is the gradual increase in the overall temperature of the Earth's atmosphere caused by the burning of fossil fuels. This burning releases greenhouse gases, such as carbon dioxide, into the atmosphere which trap heat and cause the temperature to rise. The text also highlights the urgency of the issue of climate change and the need for governments to take action to reduce carbon emissions and invest in renewable energy sources. Failure to address this issue will have negative consequences for future generations.**

**Topic: Climate change and Global warming.**

# Directional Stimulus Prompting

## Input text

Article: (CNN) For the first time in eight years, a TV legend returned to doing what he does best. Contestants told to "come on down!" on the April 1 edition of "The Price Is Right" encountered not host Drew Carey but another familiar face in charge of the proceedings. Instead, there was Bob Barker, who hosted the TV game show for 35 years before stepping down in 2007. Looking spry at 91, Barker handled the first price-guessing game of the show, the classic "Lucky Seven," before turning hosting duties over to Carey, who finished up. Despite being away from the show for most of the past eight years, Barker didn't seem to miss a beat.

## Reference

Bob Barker returned to host "The Price Is Right" on Wednesday . Barker, 91, had retired as host in 2007 .

## Standard Prompting

Q: Summarize the above article briefly in 2-3 sentences.

## Model output

A: Bob Barker, the former host of \"The Price Is Right\", returned to the show for the first time in 8 years. He hosted the first game of the show, and then handed the hosting duties over to Drew Carey.

ROUGE-1: 34.48 👎

## Directional Stimulus Prompting

Q: Summarize the above article briefly in 2-3 sentences **based on the hint.**

**Hint:** Bob Barker; TV; April 1; "The Price Is Right"; 2007; 91.

## Model output

A: On April 1, Bob Barker returned to the TV show "The Price Is Right" after eight years . Looking spry at 91, Barker handled the first price-guessing game . Barker stepped down as host in 2007, but didn't seem to miss a beat .

ROUGE-1: 48.39 👍

Produced by small, optimized, fine-tuned LM



# The Cambridge Dictionary

## Word of the Year 2023 is...



# hallucinate

*verb*

---

**When an artificial intelligence hallucinates, it produces false information.**

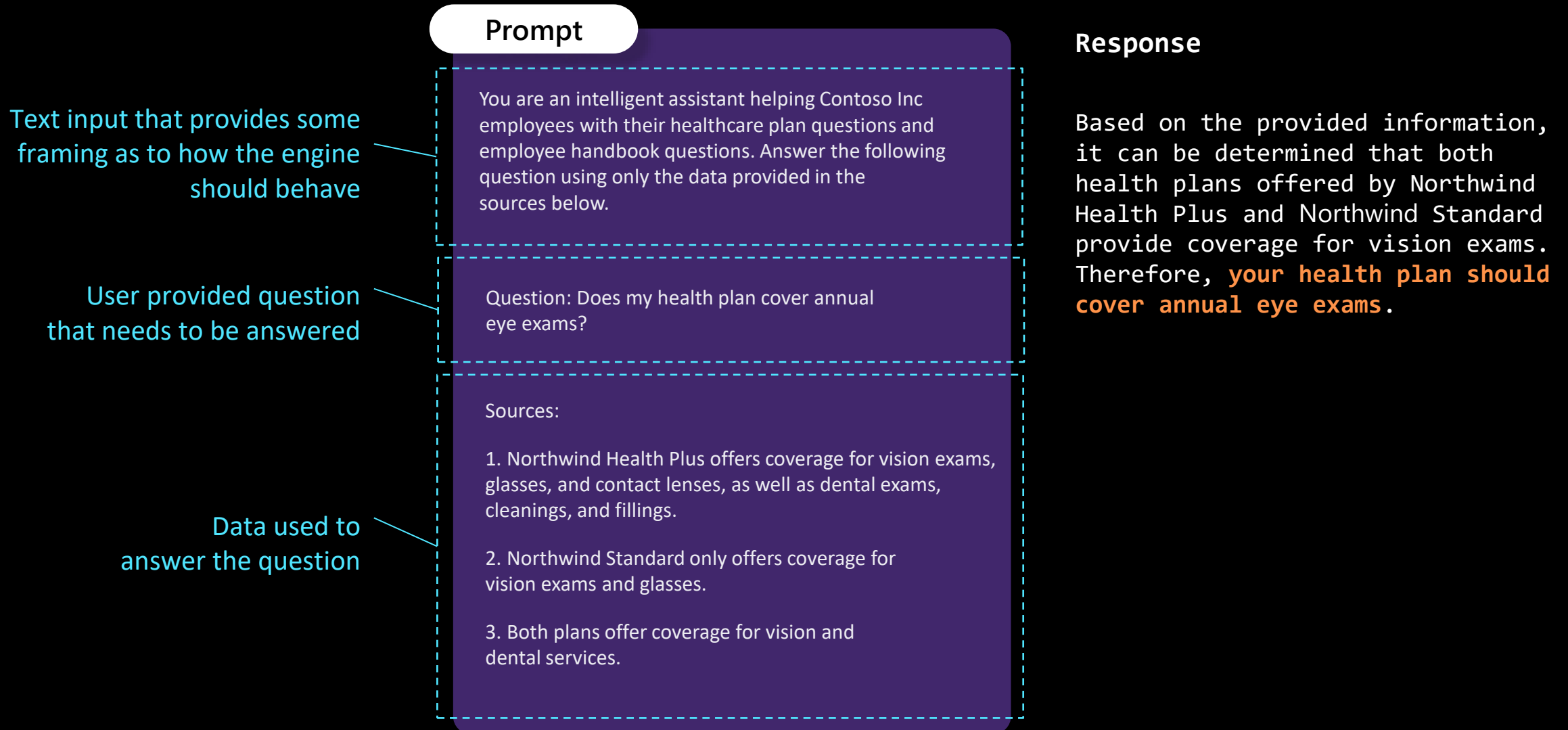


# Reducing hallucination

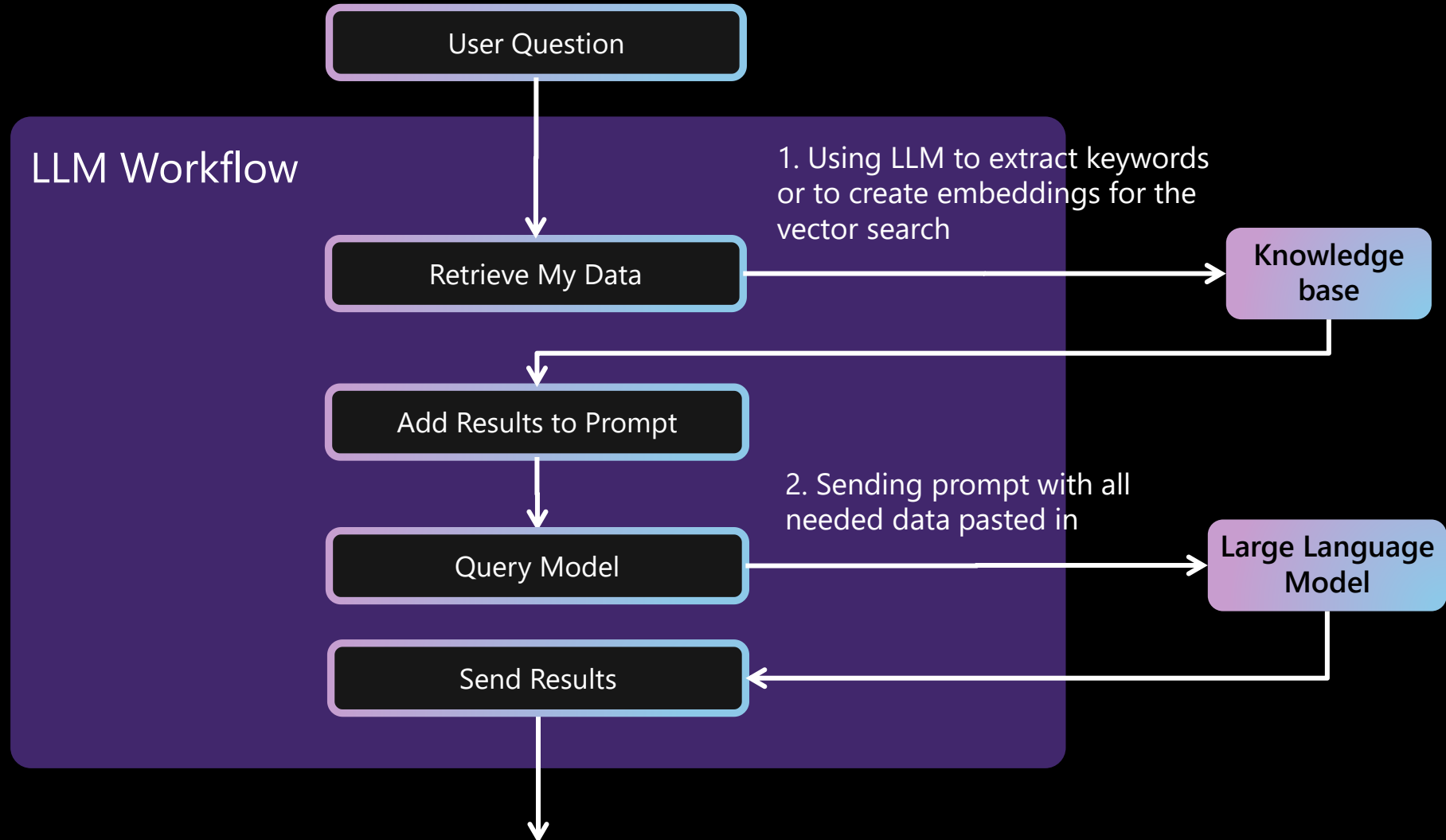
- Tell the model what you don't want
- Tell it what to say when it is not sure, say "I don't know"
- "Do not make up facts"
- Discriminator that checks if all information needed to answer is available
- Step by step reasoning
- Ask the model to explain along with the answer
- **Dynamically finding and injecting relevant context into prompt**



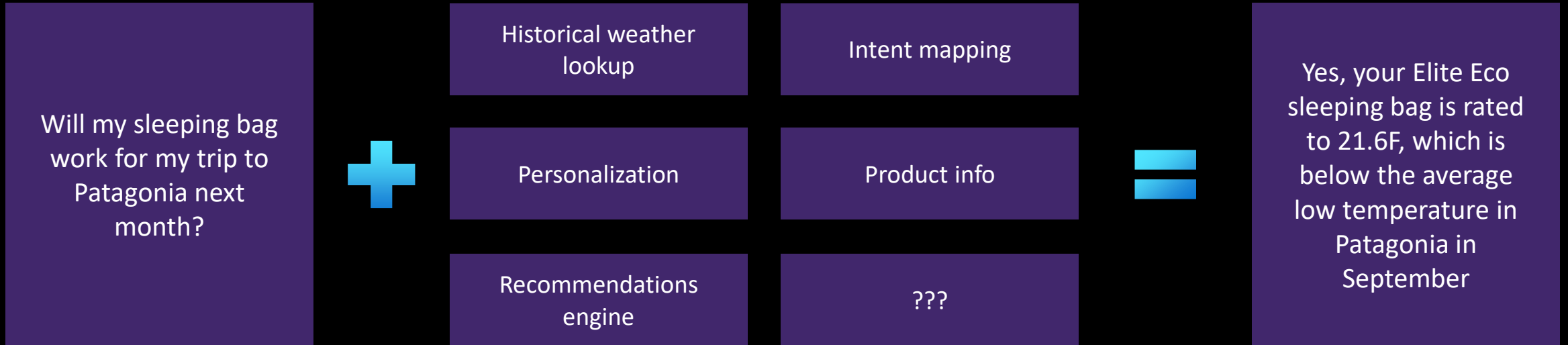
# Added Knowledge Prompting



# Retrieval Augmented Generation (RAG)



# More context



User input



Prompt engineering



LLM

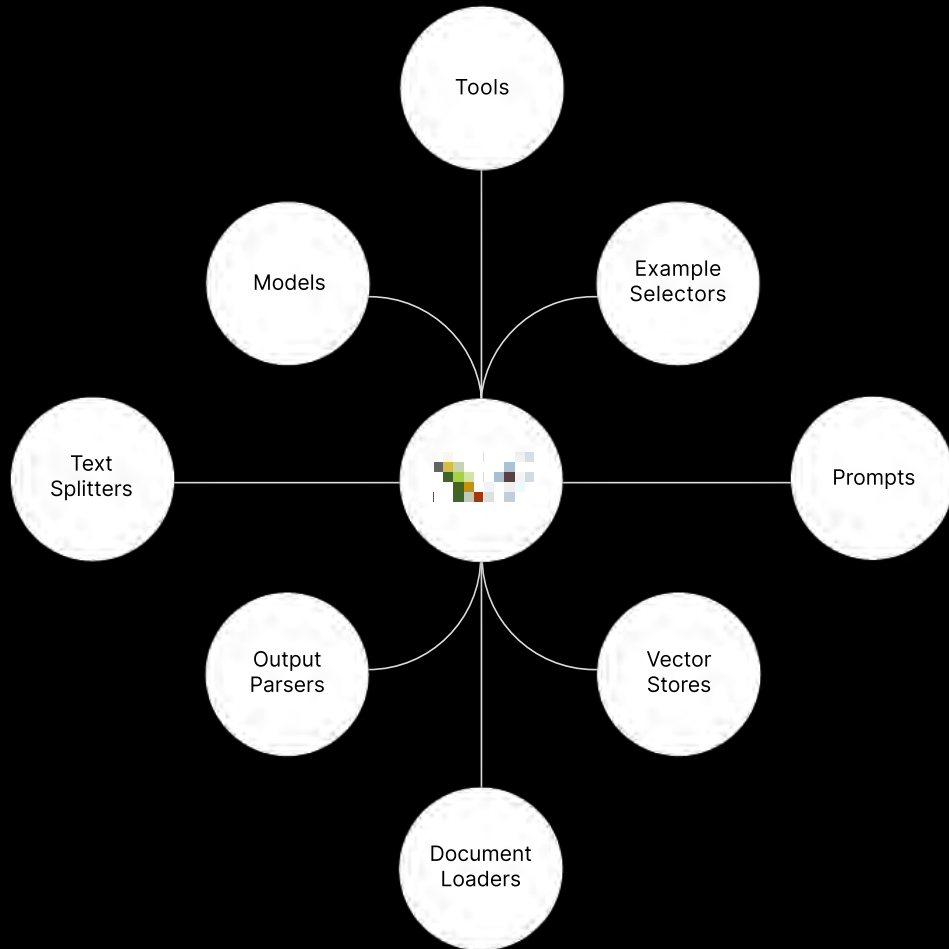


Output

# What about DX for the prompts LLM apps

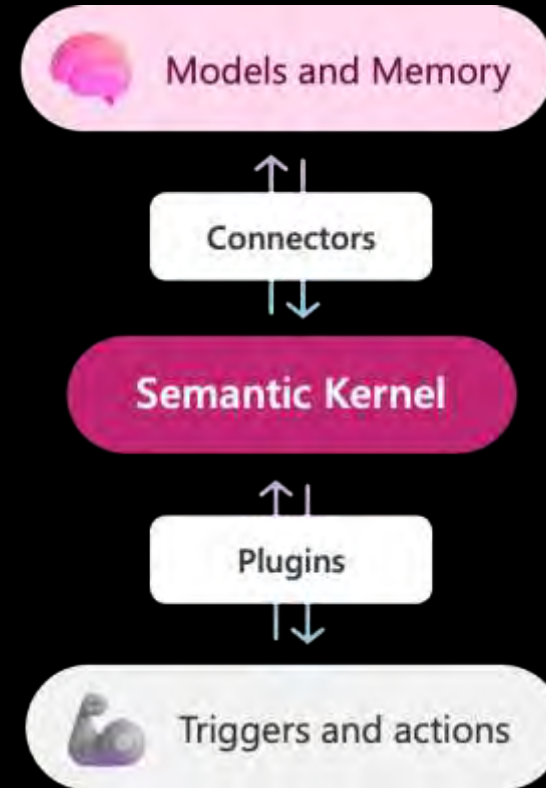
- Prototyping
- Experimenting
- Iterating
- Deploying
- Optimizing
- Monitoring

# LangChain



<https://github.com/langchain-ai/langchain>

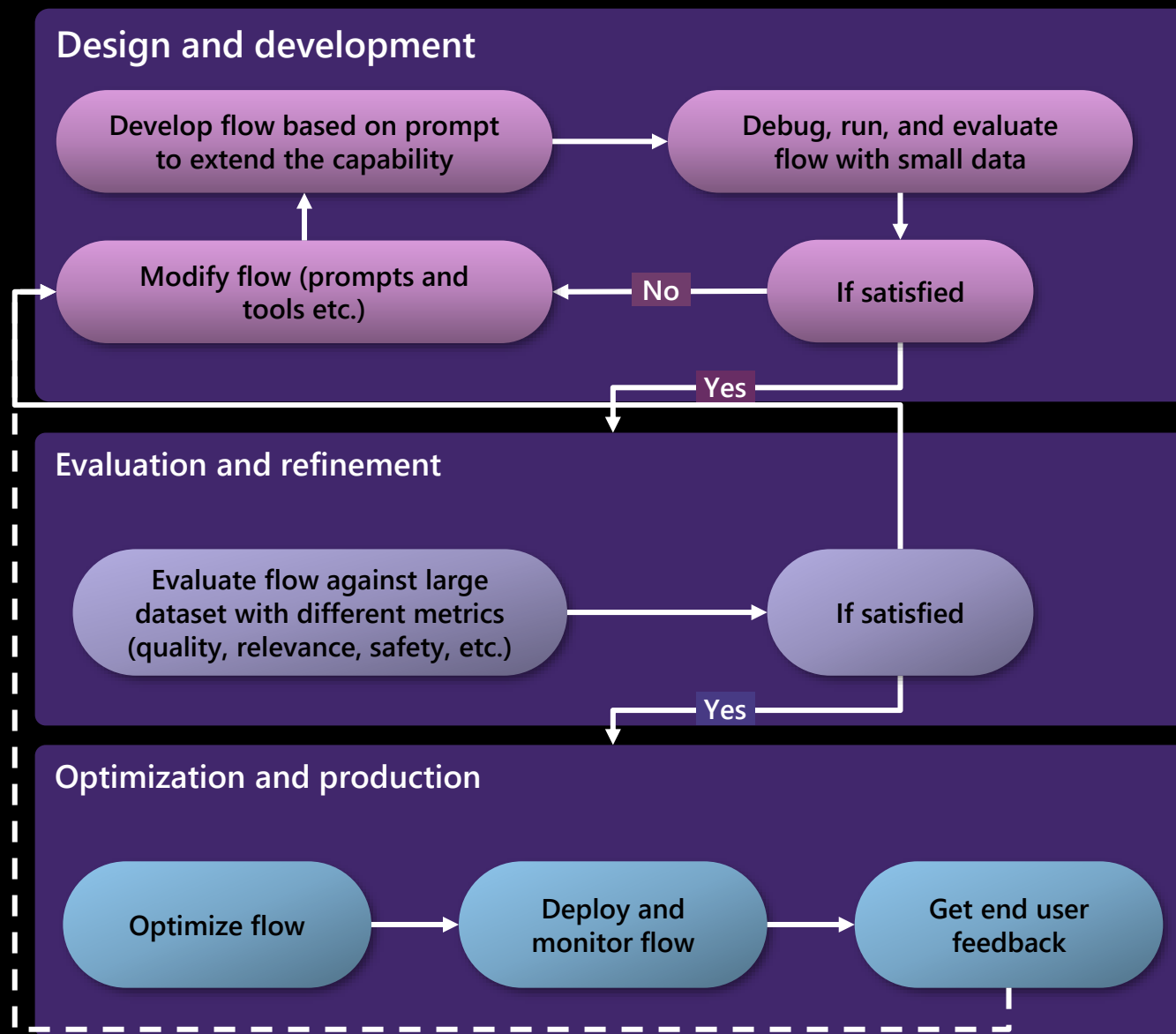
# Semantic Kernel



<https://github.com/microsoft/semantic-kernel>

# Operationalize LLM app development

- Prompt engineering: authoring, tracking history, variants
- CI/CD-ready
- Iterative experimentation
- Versioning and reproducibility
- Deployment and optimization
- Code-first



# Prompt Flow for LLMOps!

## Create and iteratively develop flow

- Create executable flows that link LLMs, **prompts**, Python code and other tools together.
- Debug and iterate your flows, especially the interaction with LLMs with ease.

## Evaluate flow quality and performance

- Evaluate your flow's quality and performance with larger datasets.
- Integrate the testing and evaluation into your CI/CD system to ensure quality of your flow.

## Streamlined development cycle for production

- Deploy your flow to the serving platform you choose or integrate into your app's code base easily.
- Collaborate with your team by leveraging the cloud version of Prompt flow in Azure AI.

<https://github.com/microsoft/promptflow>



```
inputs:
  customer_info:
    type: string
  chat_history:
    type: dict
outputs:
  output:
    type: object
    reference: ${generate_code.output}
nodes:
- name: system_prompt_template
  type: prompt
  source:
    type: jinja2
    path: system_prompt_template.jinja2
  inputs: {}
- name: user_prompt_template
  type: prompt
  source:
    type: jinja2
    path: user_prompt_template.jinja2
  inputs: {}
- name: extract_customer_intent
  type: python
  source:
    type: code
    path: extract_customer_intent.py
    function: extract_customer_intent
  inputs:
    customer_info: ${inputs.customer_info}
    chat_history: ${inputs.chat_history}
    system_prompt_template: ${system_prompt_template.output}
    user_prompt_template: ${user_prompt_template.output}
env: requirements.txt
```

Code-first!

Flow

+ LLM

+ Prompt

+ Python

- More tools

📁 Save

🔌

Raw file mode ...

## 🛠️ modify\_query\_with\_history ✎ 📄 llm

☰ Show variants

✎ Generate variants

▶

🗑️

▶

↓

🔄

↻

Connection enterprompt-aoai-connection ▾

Api chat ▾

deployment\_name gpt-35-turbo-16k ▾

temperature 0

stop

max\_tokens 1000

&gt; Advanced

&gt; Function calling

▽ Prompt ⓘ Referring to: modify\_query\_with\_history.jinja2

```
1 system:
2 * Given the following conversation history and the users next question, rephrase the question to be
3 If the conversation is irrelevant or empty, just restate the original question.
4 Do not add more details than necessary to the question.
5
6 chat history:
7 {% for item in chat_history %}
8 user:
9 {{ item.inputs.chat_input }}
10 assistant:
11 {{ item.outputs.output }}
12 {% endfor %}
13
14 Follow up Input: {{ chat_input }}
15 Standalone Question:
```

▽ Inputs

🔍 Validate and back input

Name

Type

Value

chat\_history

string

\${inputs.chat\_history}

Files

↗

Graph

↗

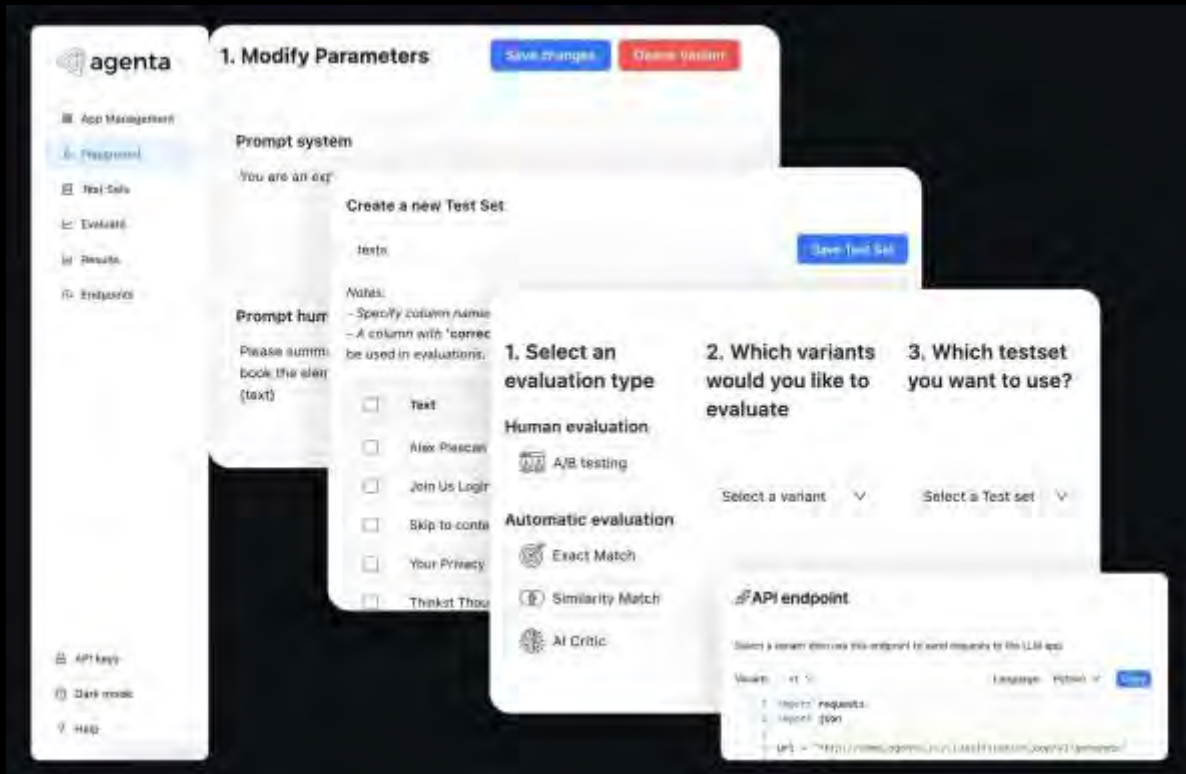
inputs


🛠️ modify\_query\_with\_history  
✔ Completed📊 embed\_the\_question  
✔ Completed📊 search\_question\_from\_ind...  
✔ Completed📊 generate\_prompt\_context  
✔ Completed


💡 Prompt\_variants





# Agenta: e2e LLM developer platform



 **Experiment and compare prompts** on any LLM workflow (chain-of-prompts, RAG, LLM agents...)

 Evaluate your application with pre-existing or custom evaluators

 Collaborate with product teams for **prompt engineering** and evaluation

 Deploy your application in one-click in the UI, through CLI, or through GitHub workflows

Works with any LLM framework and any LLM provider

<https://github.com/Agenta-AI/agenta>

# Learning resources

- [Learn Prompting](#) Community
- [Prompting Guide](#) by DAIR.AI
- [Prompt Engineering for the developers](#) on Deep Learning
- [Prompt Examples](#) on OpenAI
- [Introduction to Prompt Engineering](#) on Microsoft Learn

## Future of Prompt Engineering?

- Separate job title or essential skill?
- Simpler (with tooling, more intuitive models) or more complex (multi-modality, vectors, etc.)?
- Democratizing (+ job title inflation) or gating?
- Linguists or technologists (or domain experts able to formulate the problem)?
- Competing with “LLM Prompt Engineers”?



# Prompt Engineering Conference

The world's first conference dedicated to the only way to communicate with the most powerful AI tools currently available

**OCTOBER 30, ONLINE, FREE**

GET A FREE TICKET

REGISTER AS SPEAKER

SEE VIDEOS FROM 2023

PROMPT ENGINEERING IS A CUTTING-EDGE  
DISCIPLINE THAT FOCUSES ON  
HARNESSING THE FULL POTENTIAL OF  
LANGUAGE MODELS BY STRATEGICALLY  
DESIGNING INPUT INSTRUCTIONS  
("PROMPTS")

**v2.0**



<https://promptengineering.rocks>

Thank you! I kindly prompt you:



Let's stay in touch: connect with me on LinkedIn using this QR-code to ask questions about Prompt Engineering, Generative AI, AI-assisted development. And for getting this slidedeck immediately :)