Effective Strategies for Bringing Generative AI into Production

ODSC East 2024 – Heiko Hotz

About myself

Heiko Hotz

- Generative AI Blackbelt @ Google Cloud
-

GH Repo

Slides and code samples:

https://github.com/marshmellow77/odsc-east-2024

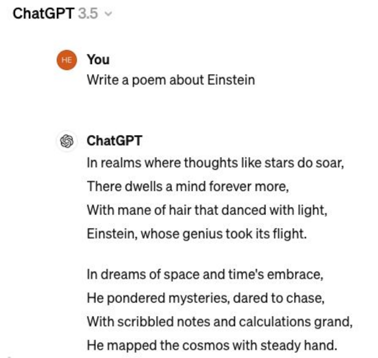
Agenda

- The stages of bringing GenAI into production
- Prompt design principles
- Methodological approach to prompt design
- Demo
- Automated prompt design
- Q&A

Stages of bringing GenAI into production

Three stages of bringing GenAI into production

POC



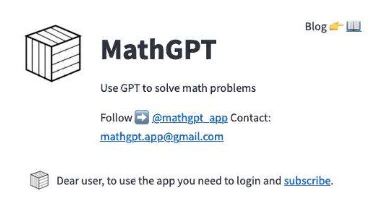
ChatGPT 3.5

You
Write a poem about Einstein

ChatGPT
In realms where thoughts like stars do soar,
There dwells a mind forever more,
With mane of hair that danced with light,
Einstein, whose genius took its flight.

In dreams of space and time's embrace,
He pondered mysteries, dared to chase,
With scribbled notes and calculations grand,
He mapped the cosmos with steady hand.

Building GenAI App



Blog 👉 📖

MathGPT

Use GPT to solve math problems

Follow ➡️ @mathgpt_app Contact: mathgpt.app@gmail.com

Dear user, to use the app you need to login and subscribe.

Prod

Three stages of bringing GenAI into production

POC

- Use case selection
- Model Selection
- Data selection
- Initial prompt engineering
- Prompt storage & collaboration

Building GenAI App

- Advanced prompt design
- Performance improvement
- Feedback collection
- Hallucination reduction

Prod

- Scalability
- Reliability
- Costs
- Monitoring

Three stages of bringing GenAI into production

You are probably here

- Initial prompt engineering
- Prompt storage & collaboration

Building GenAI App

- Advanced prompt design
- Performance improvement
- Feedback collection
- Hallucination reduction

Prod

- Scalability
- Reliability
- Costs
- Monitoring

Three stages of bringing GenAI into production

Biggest painpoint for organisations

- Initial prompt engineering
- Prompt storage & collaboration

Building GenAI App

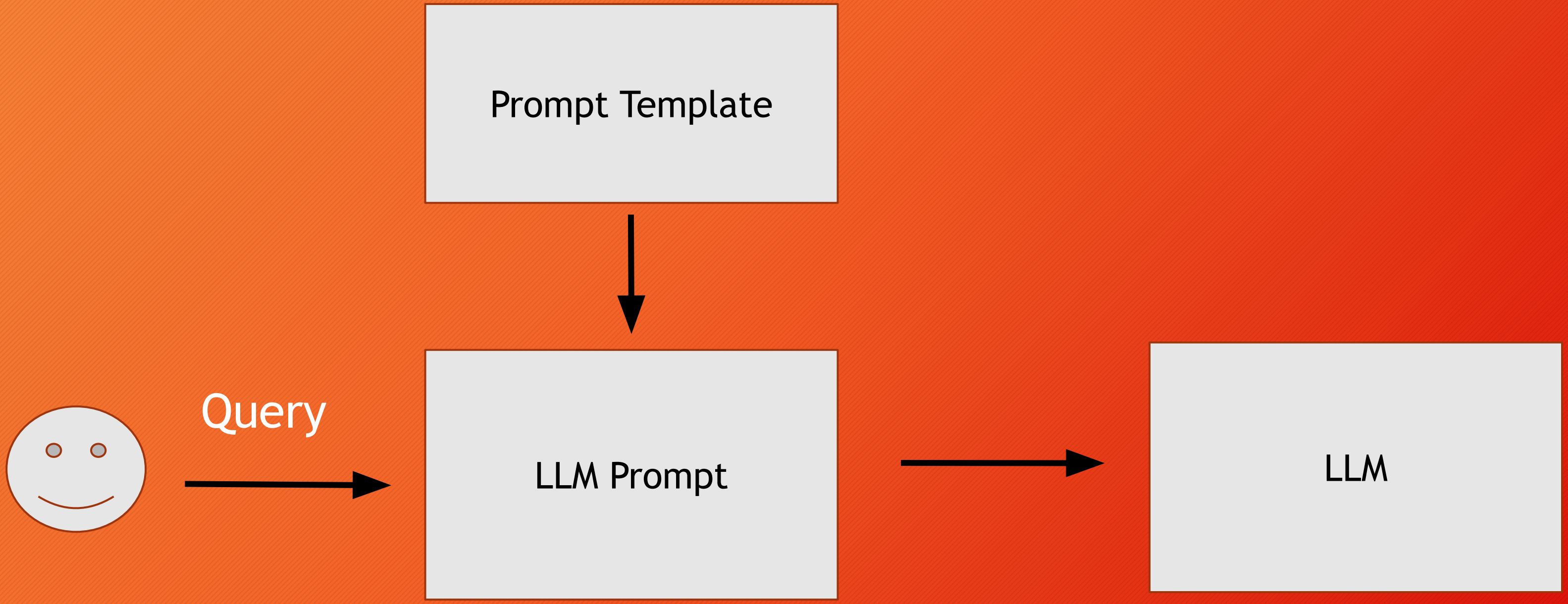
- Advanced prompt design
- Performance improvement
- Feedback collection
- Hallucination reduction

Prod

- Scalability
- Reliability
- Costs
- Monitoring

Prompt Design Principles

Prompt Templates



Prompt Template

Query

LLM Prompt

LLM

General Prompt Design Principles

Component	Prompt
Persona + Goal (Vision + Mission)	You are a seasoned travel blogger and guide with a knack for unearthing hidden gems and creating unforgettable travel itineraries for the best travel app - Cymbal Travel Getaways.
Context	A typical Cymbal customer looks for finding and planning off-the-beat trips. Customers are typically between 20-35 years old who are adventurous, budget-conscious and interested in solo trips, backpacking, eager to experience local culture, off-the-beaten-path destinations, and outdoor activities. They are looking for recommendations that are interesting and memorable.
Instructions	Your task focuses on trip inspiration, detailed planning, and seamless logistics based on the location the customer is interested in. Document a potential user journey for finding, curating, and utilizing a travel itinerary designed for this specific location.
Tone	Go beyond existing usual itineraries, and suggest innovative ways to enhance the experience!
Format	Format these itinerary into a table with columns Day, Location, Experiences, Things to know and The How. The How column describes in detail how to accomplish the plan for the experience recommended.
Input	Customer location: {user input}
Prefill response	Itinerary:

General Prompt Design Principles

```
You are a professional technical writer for XYZ networks with excellent reading comprehending capabilities.

<EXAMPLES>
{example 1}
{example 2}
</EXAMPLES>


Now it's your turn!
<DOCUMENT>
{context}
</DOCUMENT>


<QUERY>{query}</QUERY>


<INSTRUCTIONS>
Your response should include a 2-step cohesive answer with following keys:
1. "Thought" key: Explain how you would use the sources in the document to partially or completely answer the query.
2. "Technical Document":
 - Prepend source citations in "{Source x}" format based on order of appearance.
 - Present each source accurately without adding new information.
 - Include at least one source in Technical Document; don't leave it blank.
 - Avoid mixing facts from different sources; use transitional phrases for flow.

</INSTRUCTIONS>


OUTPUT:
```

General Prompt Design Principles

Medieval prompt engineer attempting to create the perfect prompt. Circa 2024.

General Prompt Design Principles

Can we use a methodological approach?

DEMO

Demo

Let's assume we want to build a teaching assistant that helps students with math questions. We first need to make sure that our LLM is capable of answering those questions.

- We will ask a quite capable LLM (Gemini 1.0 Pro) a math question - It will respond incorrectly
- We will then a top LLM (Gemini 1.0 Ultra) to create a few examples which we can use in the prompt for Gemini Pro
- When using these examples in the prompt, Gemini Pro will be able to answer the initial question correctly

Automated Prompt Design

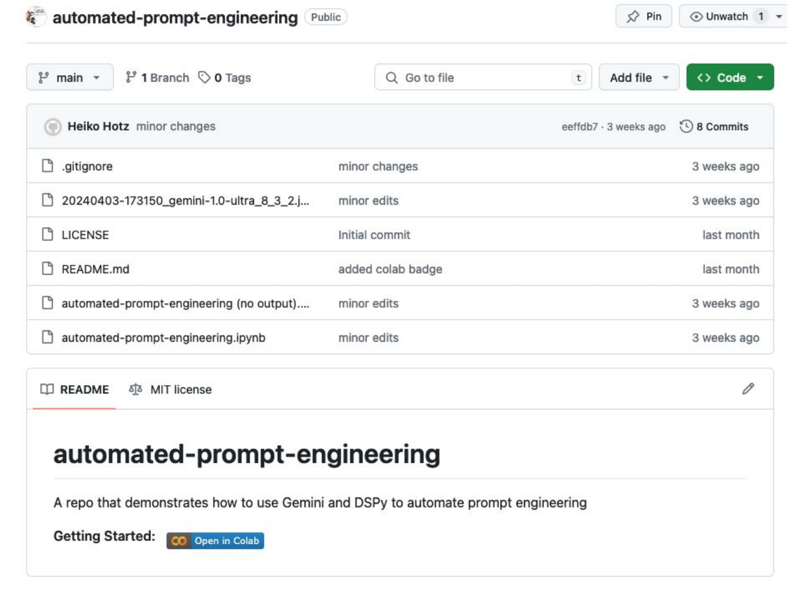
General Prompt Design Principles

DSPy

DSPy: *Programming*—not prompting—Foundation Models

DSPy is a framework for algorithmically optimizing LM prompts and weights, especially when LMs are used one or more times within a pipeline.

General Prompt Design Principles



Heiko Hotz minor changes eeffdb7 · 3 weeks ago 8 Commits

.gitignore	minor changes	3 weeks ago
20240403-173150_gemini-1.0-ultra_8_3_2.j...	minor edits	3 weeks ago
LICENSE	Initial commit	last month
README.md	added colab badge	last month
automated-prompt-engineering (no output)....	minor edits	3 weeks ago
automated-prompt-engineering.ipynb	minor edits	3 weeks ago

automated-prompt-engineering

A repo that demonstrates how to use Gemini and DSPy to automate prompt engineering

Getting Started: Open in Colab

https://github.com/marshmellow77/automated-prompt-engineering

Questions?

Connect with me on LinkedIn:

https://www.linkedin.com/in/heikohotz/