

Effective Strategies for Bringing Generative AI into Production

ODSC East 2024 - Heiko Hotz

About myself



Heiko Hotz

- Generative AI Blackbelt @ Google Cloud
- Working with organisations who want to leverage Generative AI

GH Repo

Slides:

<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

 **MathGPT** Blog  

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

 **Customer Agents**



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 painpoints for organisations

- Initial prompt engineering
- Prompt storage & collaboration

Building GenAI App

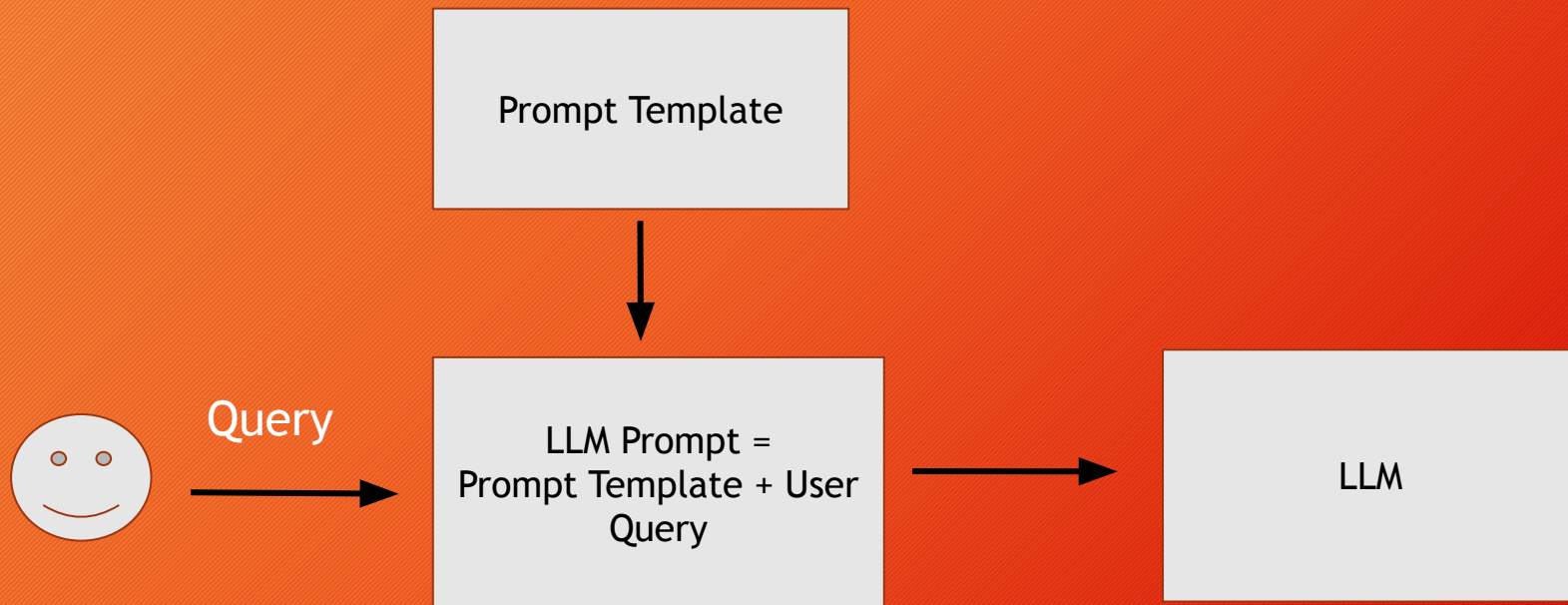
- Advanced prompt design
- Performance improvement
- Feedback collection
- Hallucination reduction

Prod

- Scalability
- Reliability
- Costs
- Monitoring

Prompt Design Principles

Prompt Templates



General Prompt Design Principles

Component	Prompt
Persona + Goal (Vision + Mission)	You are a seasoned travel blogger. Your goal is to provide recommendations to travellers.
Context	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.
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	Be conversational and friendly!
Format	Format the 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}

General Prompt Design Principles

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

<EXAMPLES>

{example 1}

{example 2}

</EXAMPLES>

<DOCUMENT>{context}**</DOCUMENT>**

<QUERY>{query}**</QUERY>**

<INSTRUCTIONS>

Your response should include a 2-step cohesive answer

</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

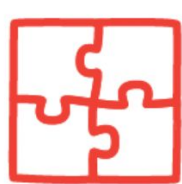
PROBLEM STATEMENT:

- We want to create a math tutor powered by an LLM
- Sometimes our LLM gets math questions wrong
- We want to teach our LLM to reason better by providing examples
- Coming up with these examples is manual & tedious
- Can we use LLMs to create the examples for us?

DEMO

Automated Prompt Design

General Prompt Design Principles

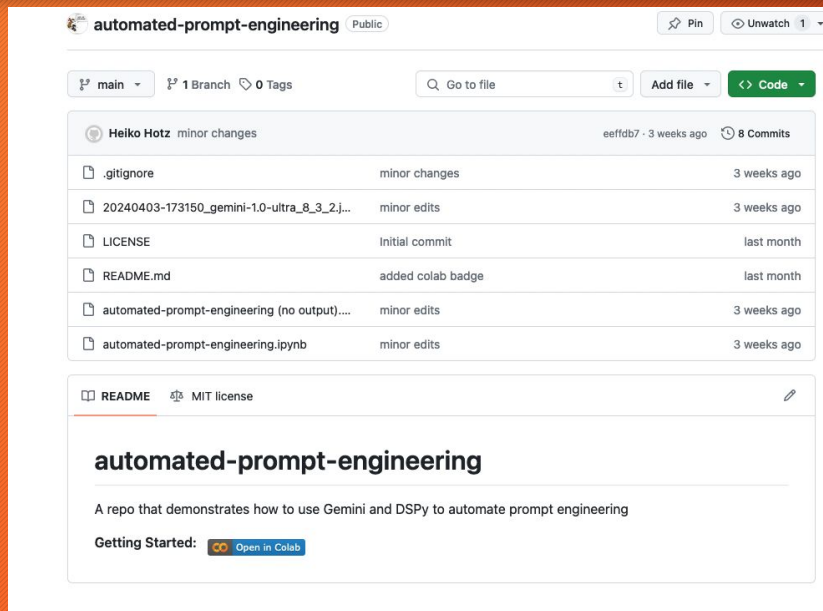


DSPy

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



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

Questions?

Connect with me on LinkedIn:

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

