

LLMs for I/Os

A Functionalities and Applications Masterclass



AON

BGSU®

SHL.

Introductions



Karim Badr

Senior Research Scientist



Matthew Bennett

Research Scientist



Bao Ho

PhD Student

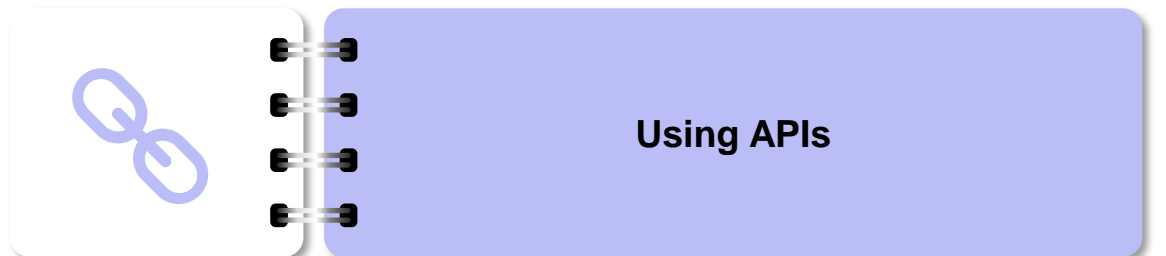
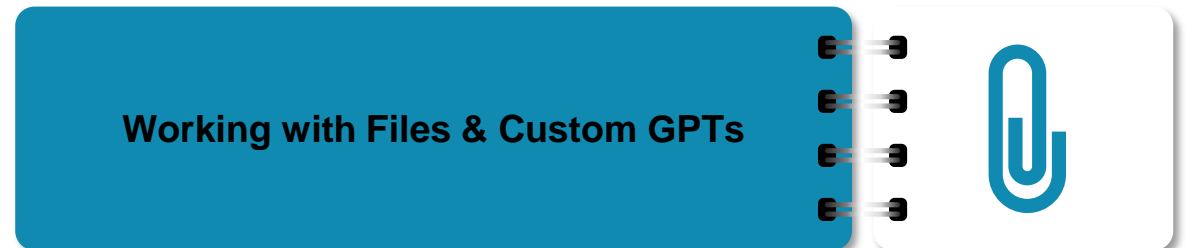
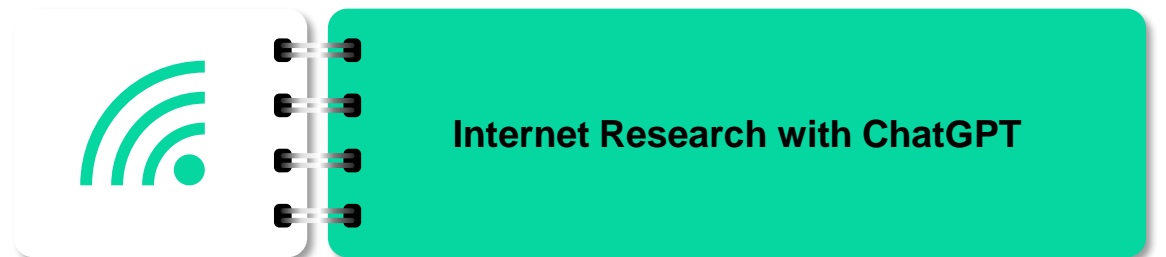
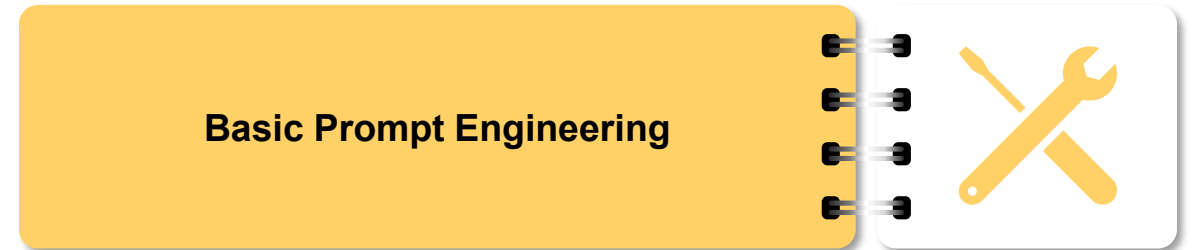
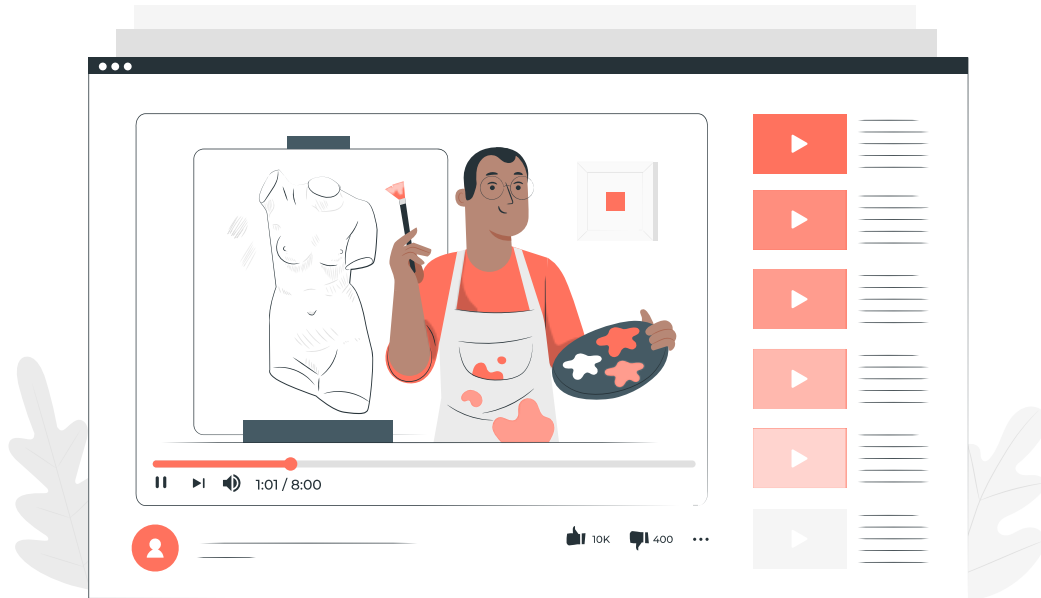


Nino Sziedell

Global Analytics Consultant



Agenda



ChatGPT & LLMs

An Overview

4 April 2025



Could ChatGPT Have Written This?

Have you had to write anything this past year and wondered if ChatGPT could do it faster?



Overview Agenda

1

Understand ChatGPT & LLMs

2

Review recent developments in the commercial LLM service space

3

Look ahead to what's coming next

What is ChatGPT?

- Commercial, publicly available artificial intelligence (AI) service
- Launched November 2022 by OpenAI
 - Reached 100M+ users in 2.5 months
- Generates any written content based on simple inputs
 - Source of knowledge
 - General chatting
 - Help with work – writing reports, lesson plans, etc.
 - Coding

How Does ChatGPT Work?

- Generative AI – Generates content (output) based on a user prompt (input)

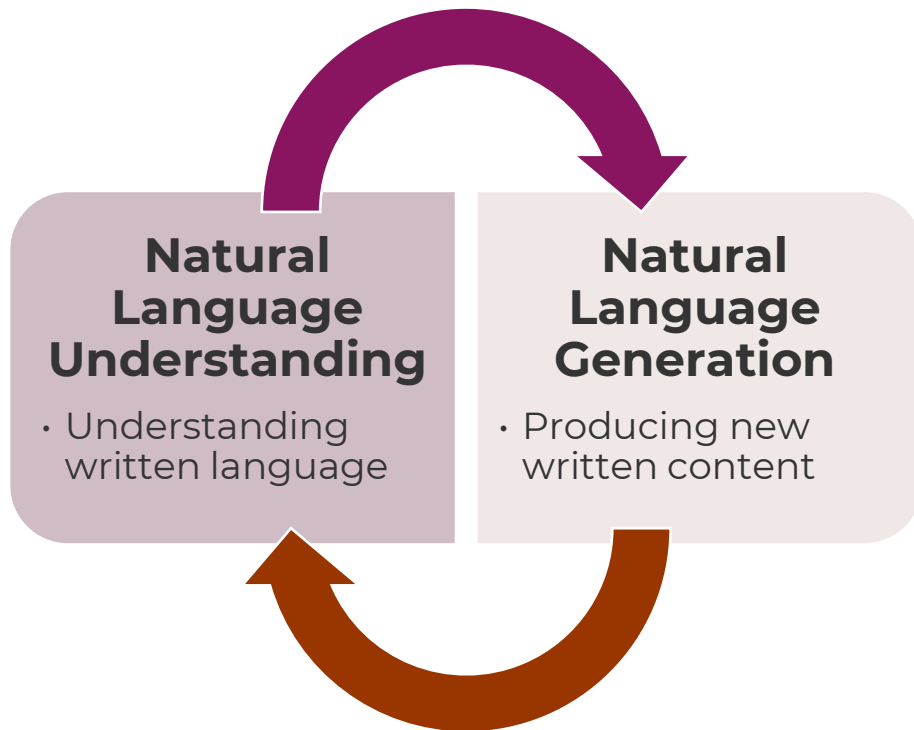


- Large Language Model (LLM) – Generative AI model that uses text as input & produces text as output



*GPT = Generative Pretrained Transformer

Natural Language Processing (NLP)



- **Text Generation and Comprehension:** Crafting coherent and contextually relevant text based on input prompts.
- **Question Answering and Reasoning:** Providing informative responses and engaging in complex problem-solving.
- **Language Translation:** Converting text between languages while maintaining nuance and context.
- **Summarization:** Condensing lengthy documents into concise summaries without losing essential information.

It's Not Just ChatGPT

Model	Company
ChatGPT	OpenAI
Claude	Anthropic
Gemini (formerly Bard)	Google
LLAMA	Meta
DeepSeek	DeepSeek



LLM Services are Evolving Quickly!


Factor	Then (2022)	Now (2025)
Context Window Limit (Tokens)	4,096	(Up to) 128,000
Input Modality	Text only	Text, images, voice
Features	Limited plugins & tools	Custom GPTs, advanced API features, agents


What's on the horizon?

- AI Agents (more autonomy & control)
 - E.g., Operator (OpenAI)
- Democratization of AI
 - Lower barriers to entry
- Ethical & Regulatory Considerations
 - Environmental, social impact concerns

Discussion Topics


Basic Prompt Engineering






Internet Research with ChatGPT

Working with Files & Custom GPTs





Using APIs

Working with Prompts



What is a prompt and what is prompt engineering?

A **prompt** in context of AI and Large Language Models is a piece of information used to **communicate** between different **entities** of a **conversation/chat**.

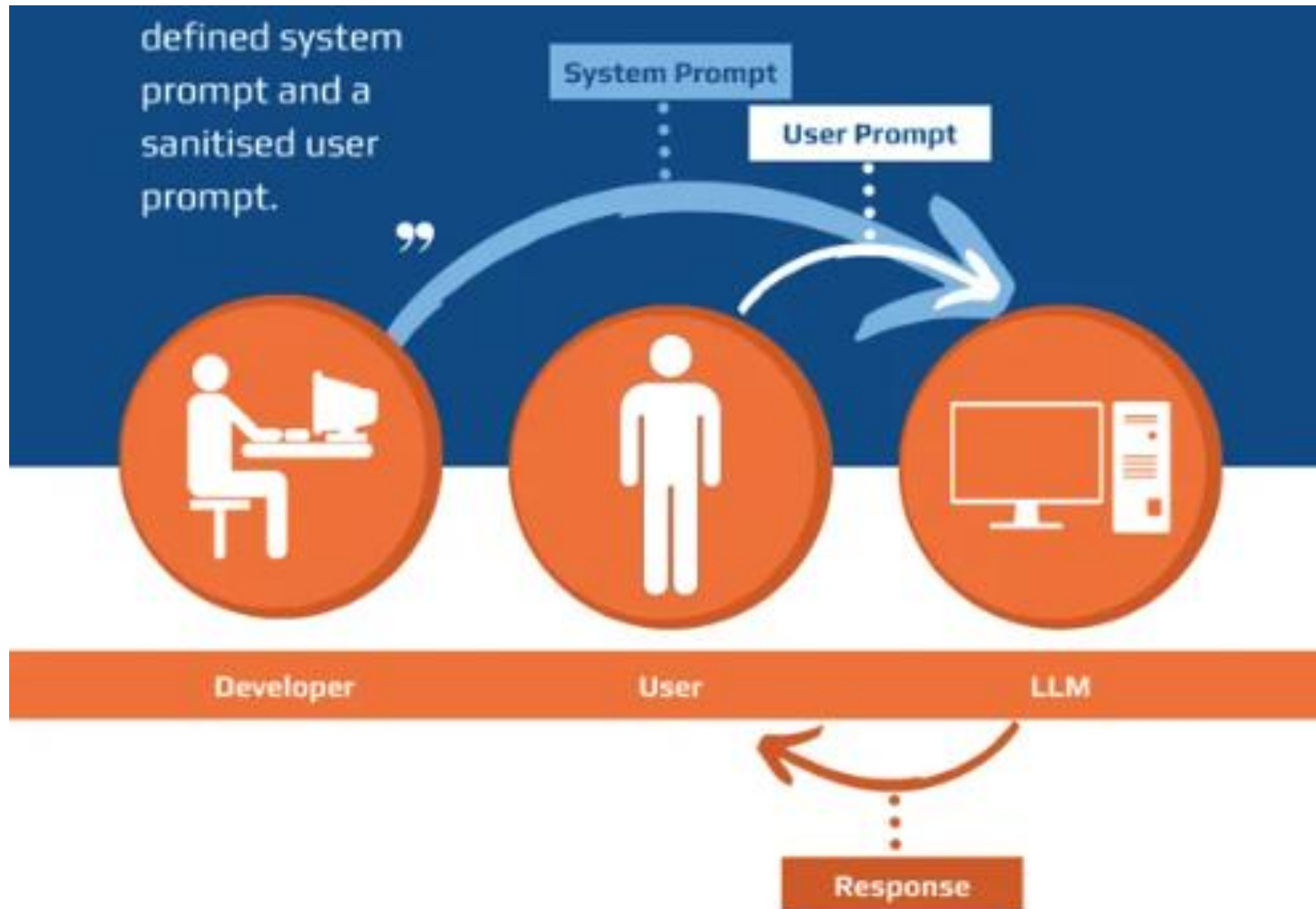
- **Agent** (LLM)
- **User** (Person interacting with the LLM)
- **Developer** (Person building the LLM)

Prompts can occur in different ways. **Most common is a natural language text prompt** but in some cases prompts are images, videos or other file/document types.

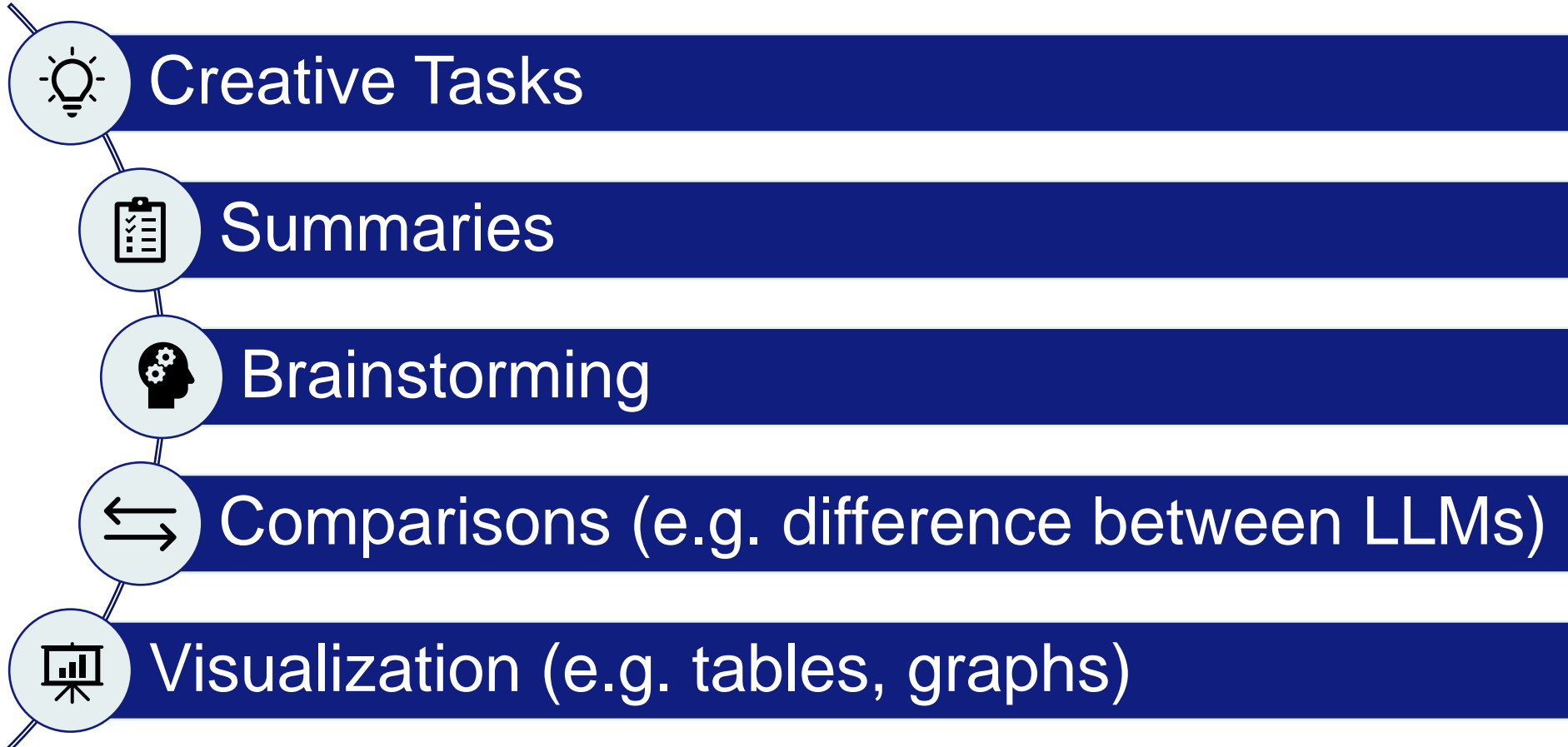
Prompt engineering describes the process where the user structures or crafts an instruction (user prompt) for the AI model in order to **produce the best possible response**.

A **user prompt** contains any of the following **elements**:

- Instruction
- Context
- Input Data
- Output Indicator



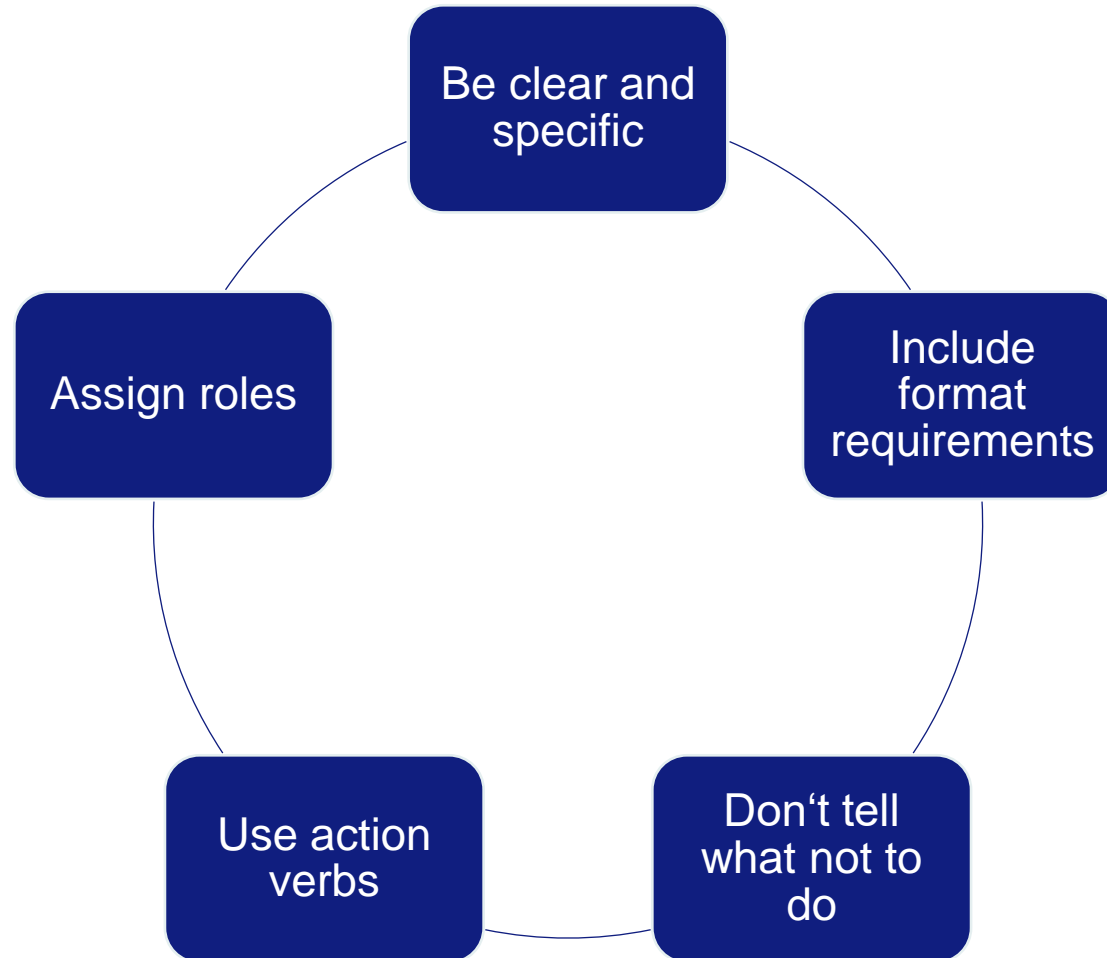
What are prompts used for?



Prompt Engineering Techniques



How to build a prompt – Basics



What makes a good leader?

vs.

*I am a recruiter in a large organization.
List the 3 most effective predictors of
job performance during the
recruitment process. Summarize each
predictor in one sentence.*

Zero-shot vs. One-shot vs. Few-shot

0

Classify the Feedback sentence into neutral, negative or positive: "I think the test was okay"

The feedback sentence "I think the test was okay" can be classified as **neutral**.

1

Waipu = positive

Classify the Feedback sentence "I would rate my Overall Assessment Experience as waipu" into neutral, positive or negative

Given that "waipu" means positive, the feedback sentence "I would rate my Overall Assessment Experience as waipu" can be classified as **positive**.

≥ 2

There was not enough time - negative
I Had Mixed Feelings - neutral
It was a pleasure finishing this - positive
It was fun - positive
The process was horrible - negative
I did Not have enough time -

The phrase "I did not have enough time" would fall under **negative**, as it expresses frustration or dissatisfaction due to time constraints.

Chain-of-Thought (CoT)

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Two LLM types

Non-reasoning (e.g. GPT-4)

Prompted reasoning

Ideal for simpler tasks

Short answers

Fast and relatively cheap

→ Zero-/one-/few-shot and chain of thought

Reasoning (o1 and o3-mini)

Built-in reasoning

Ideal for complex tasks

Longer and more technical answers

Slower and more expensive

→ Different prompting technique

Reasoning models – o1 and o3-mini

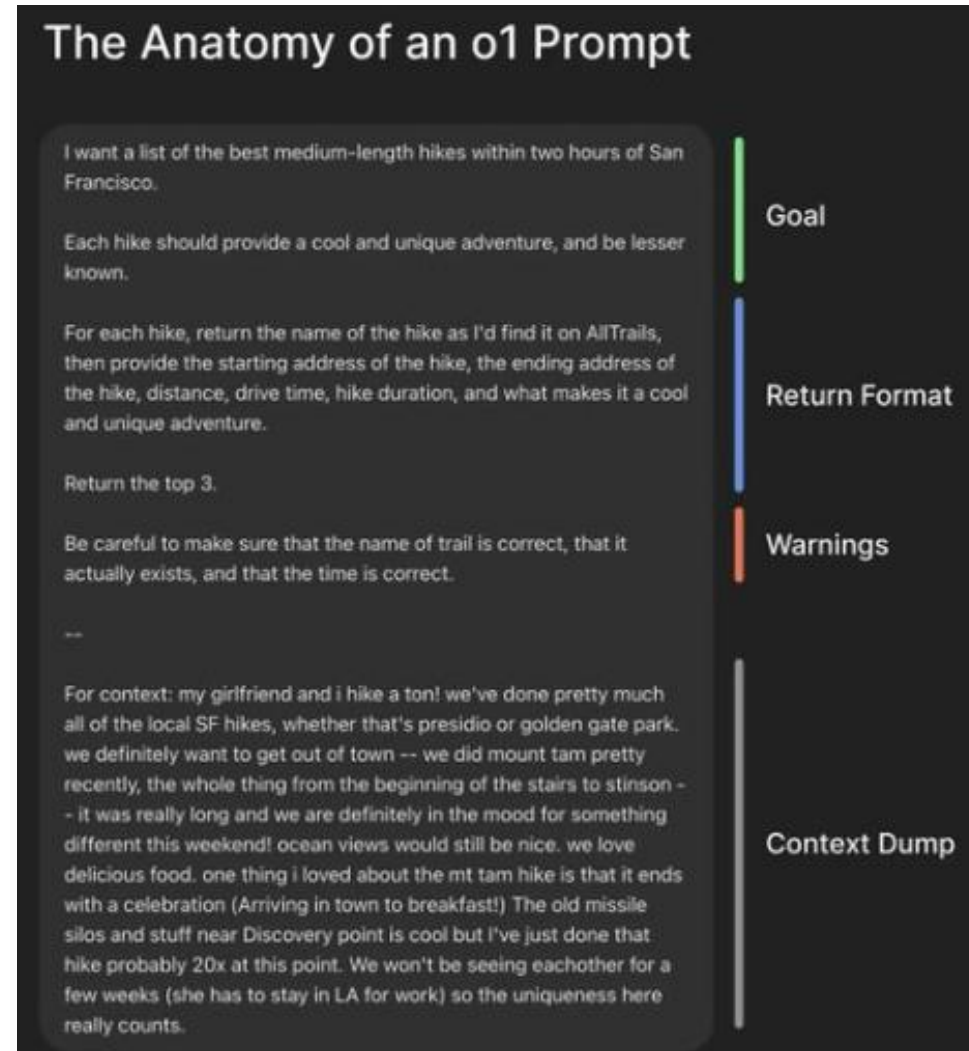
Built-in CoT reasoning

→ **Conventional prompting methods not recommended**

→ o1 Prompt

Idea is to guide the AI through a clear sequence of logical segments with a **single well-defined prompt**

- Set **clear context**
- Give **specific instructions**
- Define the **output format**
- Include **warnings** if needed
- Set **role or tone** if needed



Advanced Prompting



Prompt Templates

Translate user input and parameters into instructions for a language model

You are a customer support agent. This is the refund policy:

{refund_policy}

Please respond to the user's question:

{question}

Prompt Template

+

```
{  
  "refund_policy": "no refunds  
    under any circumstances",  
  "question": "can I get a refund  
    for this hat?"  
}
```

Input Variables

=

You are a customer support agent. This is the refund policy:

no refunds under any circumstances

Please respond to the user's question:

Can I get a refund for this hat?

Prompt

How far can we go?

Other types of
prompts (e.g.
video, voice,
links)

Various other
prompt
engineering
techniques

Working with
files

Interacting with
API

Internet Searches & Research with ChatGPT

Internet Searches

4 April 2025

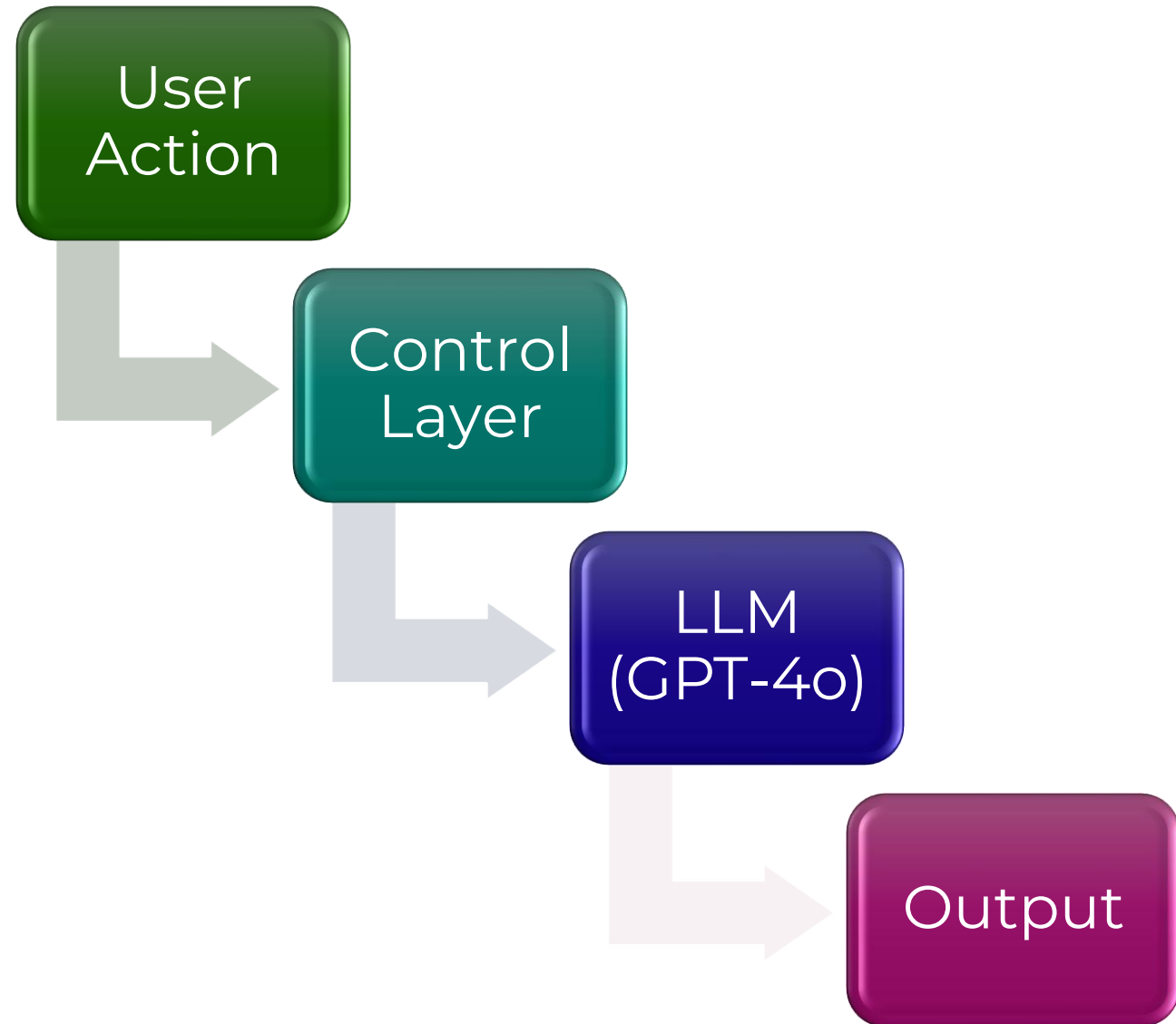


Internet Searches with ChatGPT

- LLMs must be trained on a corpus of content that has a cutoff date
 - By default, unable to answer questions about recent events
 - Knowledge is limited to whatever is in the content bank
 - Even when writing about content in the content bank, hallucinations may cause errors when answering questions
- When given the right tools, LLMs can access search engines & read websites to obtain information to use in a response

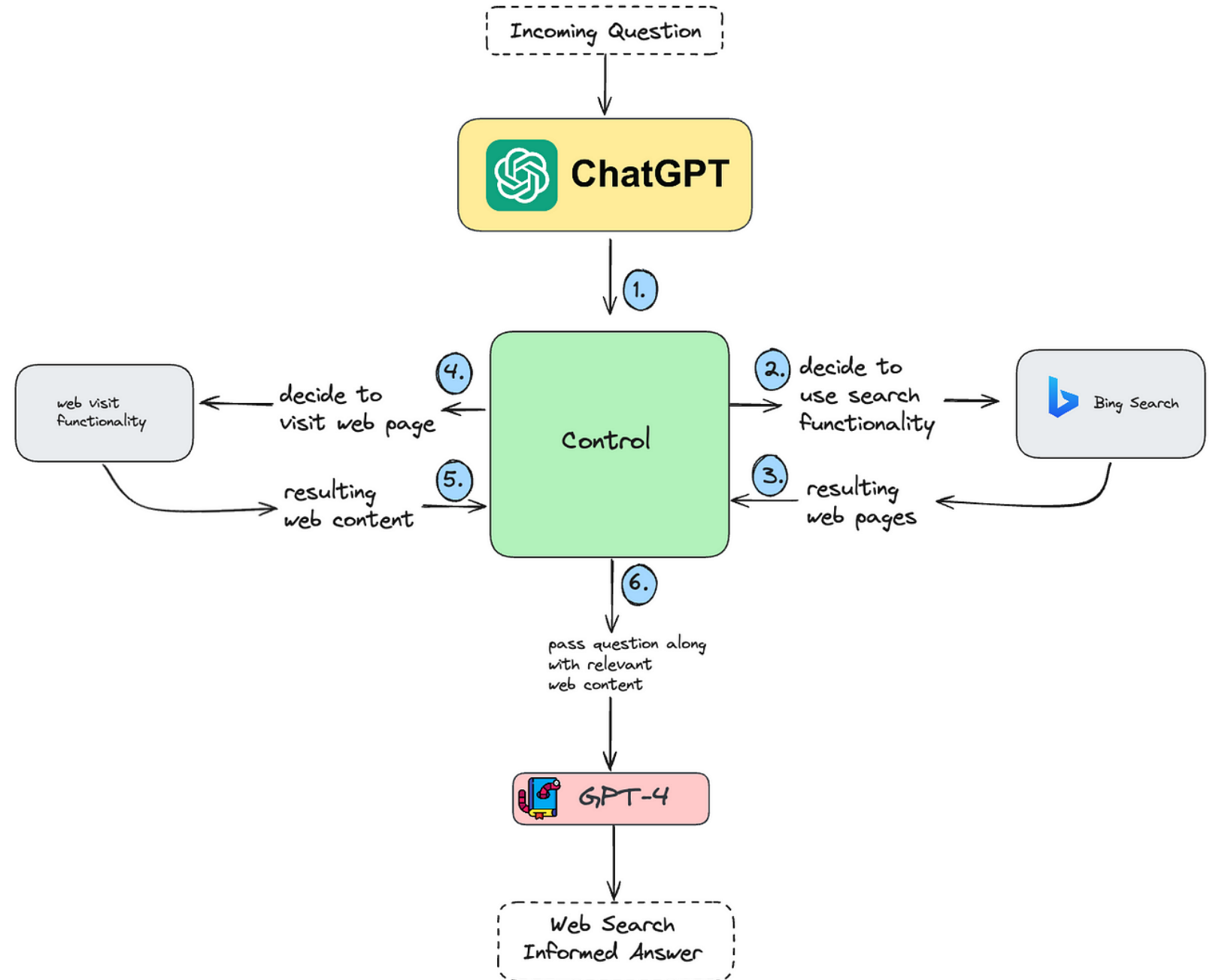
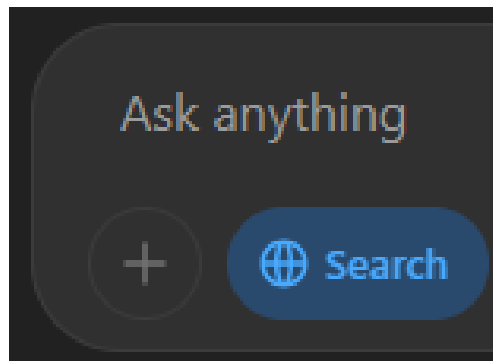
ChatGPT as a System

- Control Layer:
Executive layer that guides the flow of processing user prompts & actions



Internet Search: Interaction Modes

- User-Guided Browsing
- Fully AI-Driven Search



Example: Searching for Recent Events

1

What can I help with?

What are the most recent developments in performance management practices among Fortune 500 companies?

+ Search Deep research ...

↑

2

What are the most recent developments in performance management practices among Fortune 500 companies?



Fortune 500 companies are continually evolving their performance management practices to enhance employee engagement, productivity, and alignment with organizational goals. Recent developments include:

Continuous Feedback and Regular Check-Ins

Many organizations are moving away from traditional annual performance reviews toward continuous feedback mechanisms. For instance, Adobe implemented a "Check-In" system focusing on ongoing discussions about goals and development, resulting in a 30% reduction in voluntary turnover.

Engagedly +1 People Managing People


Integration of Technology and AI

3

     Sources


Citations



 Engagedly


Top 8 Companies That Redefined Their Performance Management ...

April 20, 2022 — Accenture's new performance management model is designed to represent the...

 People Managing People


Companies with the Best Performance Management Practices

February 12, 2025 — Key Insight: Adobe's Check-In model has significantly improved manager-employee...

 business.com


The Future of Performance Management Trends - Business.com

October 21, 2024 — The introduction of new human resources (HR) technology, combined with an evolve...

 ThriveSparrow

Performance Management Statistics: What 2025 Holds for HR Leaders

January 30, 2025 — The evidence proves compelling - more than 85% of Fortune 500 companies now...

 WSJ

The Most Hated Way of Firing Someone Is More Popular Than Ever. It's the Age of the PIP.

November 29, 2024 — The Most Hated Way of Firing Someone Is More Popular Than Ever. It's the Age of...

Considerations for Internet Searches & Research

Best Practices

Treat ChatGPT as a knowledgeable and eager, albeit inexperienced and sometimes careless, assistant

- Clear, specific prompts
 - ✗ “Tell me about remote work”
 - ✓ “What does recent research say about the impact of remote work on employee productivity and satisfaction?”
- Save conversations & ask for links to be printed
 - ✓ After the search is concluded: “For all sources cited, print out the web page name and URL of each source without markdown formatting.”
- Steer towards high-quality & reputable sources
 - ✓ Tell it which sources to focus on
 - ✓ Tell it which sources to avoid
- Avoid the black box as much as possible
 - ✓ Document when and how you use ChatGPT web searching for any work or research project

Considerations for Internet Searches & Research

Best Practices

- **Be critical**
 - “Do I really need to use an LLM here? Does the benefit outweigh the cost of not doing the actual search myself?”
 - ✓ Review all sources for accuracy of claims
 - ✓ Critically evaluate the claims being made
 - Best use case may be to identify & categorize sources, which you review manually

Considerations for Internet Searches & Research

Warnings & Pitfalls

- Hallucinations
 - ⚠ May confidently cite sources that do not exist, or connect information to the wrong source
 - ✓ Always review your sources
- Source credibility & bias
 - ⚠ Sources may not be credible or may be biased
 - ✓ Keep source count low
 - ✓ Use iterate-and-drill-deeper methods to ensure high-quality sources
- Incomplete or truncated information. (The LLM cannot read all webpage contents. Often, snippets of webpage contents are read.)
 - ⚠ May cause it to miss important details
- Ethical & Confidentiality concerns
 - ⚠ Unless data controls are used, anything you enter may be saved & used to train/refine the model
- Over-reliance & skill atrophy

Considerations for Internet Searches & Research

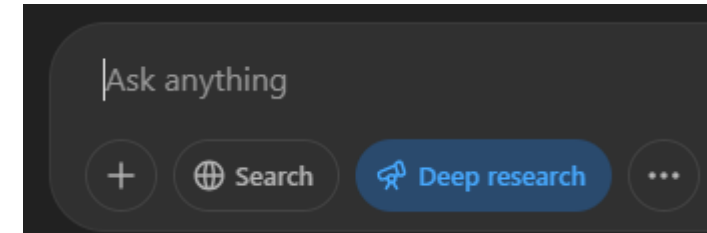
Warnings & Pitfalls

- No control over search flow
 - Changes to control layer or other parts of ChatGPT system may affect quality of web results over time
 - ⚠ Indexing system of search engine may change
 - ⚠ ChatGPT may not do a thorough search, and SEO practices may cause low-quality results to be prioritized
 - ⚠ Other changes, such as limitations to how many web pages to read per query

Deep Research

Like Research, but Deeper

- Released in early 2025 for all Plus and Teams users
 - Other services have similar functionality
- Key differences
 - Autonomous, multi-step approach
 - Multiple, sequential web searches based on a high-level direction
 - Directed mostly by the control layer
 - Designed to generate detailed reports with accurate citations
 - Long process – can take 5 – 30 minutes
- Best practices
 - Craft long, detailed prompts
 - Include background information with file uploads



Working with Files & Custom GPTs

4 April 2025



Working with Files & Custom GPTs

Research Assistant

Draft

Create

Configure

Hi! I'll help you build a new GPT. You can say something like, "make a creative who helps generate visuals for new products" or "make a software engineer who helps format my code."

What would you like to make?

Personnel Psychology - 2023 - Zhang ...

PDF

I would like to create a GPT that lets me query a research article I am reading for my class. I have attached the article. I want the GPT to be concise in its responses and to guide me to where I can look in the article to find the relevant information for my question.

I've set up your Research Assistant GPT! It will provide concise answers and direct you to the relevant sections of the article. Feel free to try it out in the preview, and let me know if you'd like any adjustments or refinements! How does this sound to you?

Ask anything

+

↑

?

Preview

Research Assistant

A research assistant that helps with concise answers and directs to relevant parts of a research article.

Can you explain the methodology in the article?

What is the main conclusion of the article?

Where does the article discuss the results?

Can you summarize the introduction section of the...

Ask anything

+

↑

?


Working with Files & Custom GPTs

Research Assistant

Draft

Create

Configure



Name

Research Assistant

Description

A research assistant that helps with concise answers and directs to relevant parts of a research article.

Instructions

This GPT is designed to assist with queries related to research articles. It provides concise answers and guides the user to the relevant section of the article for further details. The GPT should reference the article to answer questions, but its answers should not be overly detailed, instead directing the user to specific parts of the article where the information can be found.

Conversation starters

Can you explain the methodology in the article?

What is the main conclusion of the article?


Where does the article discuss the results?

Can you summarize the introduction section of the article?

Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

Preview



Research Assistant

A research assistant that helps with concise answers and directs to relevant parts of a research article.

Can you explain the methodology in the article?

What is the main conclusion of the article?

Where does the article discuss the results?

Can you summarize the introduction section of the...

Ask anything

Working with Files & Custom GPTs

Research Assistant

Share

What are the main findings of Study 1?

The main findings of Study 1 of this paper, titled "Are Fairness-Aware ML Algorithms Really Fair? Predictive Bias of Using ML in Personnel Selection," suggest that fairness-aware machine learning (ML) algorithms, which are designed to reduce adverse impact, may still introduce predictive bias. Specifically, the study shows that these algorithms may unintentionally bias the predictions against certain racial minority groups, even if the predictors themselves are unbiased. This bias arises from the nature of fairness constraints which, when applied to reduce adverse impact, can distort the predictive relationship between the model's score and the actual job performance criterion.

The study utilized mathematical analysis and simulation studies to demonstrate that fairness-aware algorithms often lead to situations where minority candidates, especially from racial groups, may be unfairly excluded from selection due to predictive bias. This effect is particularly pronounced when these fairness constraints are applied, as they attempt to optimize for both fairness and selection accuracy but often at the expense of fairness for minority groups.

Where can I find these findings in the article?

You can find the main findings of Study 1 in the article in the section titled "**Study 1: Are Fairness-Aware ML Algorithms Really Fair? Predictive Bias of Using ML in Personnel Selection**". Specifically, the discussion of these findings begins after the introduction and mathematical analysis in **Section 2.2 (Predictive Bias of Fairness-Aware ML Algorithms)**. [1.↓](#) section explains the core issue of predictive bias and its implications in personnel selection, including how fairness constraints in ML algorithms can...

Ask anything

New version of GPT available - Continue chatting to use the old version, or start a [new chat](#) for the latest version.

All About LLM API


Learn to interact with, deploy, and utilize LLMs via APIs and local solutions



Introduction

- **What You'll Learn Today:**
 - How to interact with popular LLM APIs
 - How to run your local LLM (it's easier than you think!)
 - All code demonstration will be done in **Google Colab**
- **APIs We'll Focus On:**
 - OpenAI
- **Local Deployment:**
 - Use vllm to run models efficiently on your own machine
 - Demo model: Distilled Deepseek

Interacting with LLM API

- What is an API?
 - API = *Application Programming Interface*
 - Enables developers to access models and tools from external platforms
- The OpenAI API 
 - A bridge to state-of-the-art models
 - Authentication: Use an **API key** to verify and manage requests
- Task Demo
 - Generating SJT items for skills and abilities from ONET

Running your local LLM

- **Why Run Locally?**
 - No rate limits or costs for API usage
 - More privacy and customization
 - Offline access
- **Hugging Face**
 - A central hub for thousands of open-source LLMs
 - Supports both full-precision and quantized models
- **Task Demo:** Generating SJT items for skills and abilities from ONET



Let's Do Some Coding!

Q&A Resources

- Q&A
- Slides available on **Whova**
- All materials (slides, code, and supplementary resources) found on **GitHub repo** here:

https://github.com/karimhbadr1/SIOP_2025_LLMS_for_IOs_Masterclass

