hack.diversity

Learning Lab: “Hack-GPT”

LLMs for Software Engineers and Data Analysts

February 27, 2024

William Szostak

Slack: @William Szostak

LinkedIn: linkedin.com/in/william-szostak/

Email: williamszostak@gmail.com

AGENDA

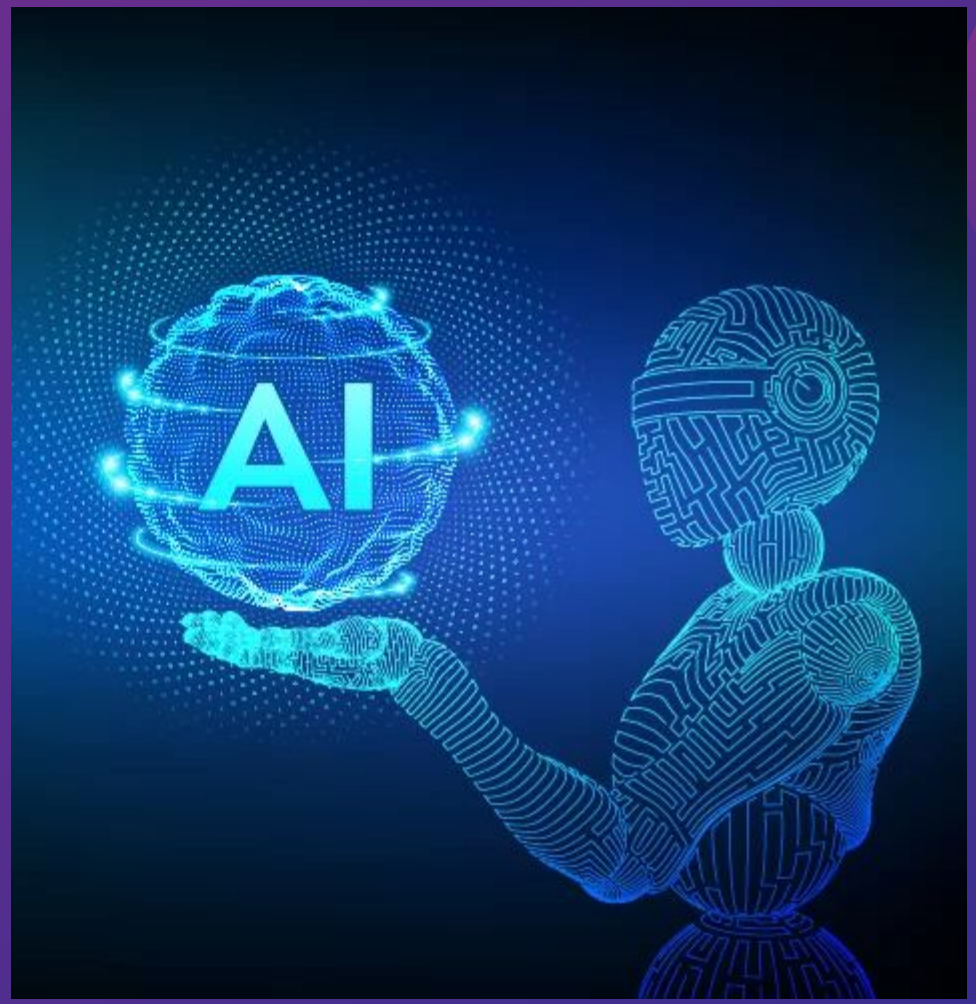
5 mins	**Intros**
10 mins	**LLMs: The Good and the Bad**
10 mins	**Prompt Engineering**
10 mins	**Data Extraction**
10 mins	**Retrieval Augmented Generation (RAG)**
10 mins	**Responsible AI**
5 mins	**Wrap Up & Next Steps**

Questions welcome throughout!

hack.diversity

Disclosure

Generative AI was used in producing these materials



About William

Professional

Education

- Went to 3 different undergraduate colleges
- Took the “10 year plan”
- BA in Mathematics

Career

- Software Engineer – Bloomberg – NYC – 1997
- Consultant - Concord, NH
- Senior Software Engineer – Portsmouth, NH
- Principal Software Engineer - Boston, MA
- Software Architect – Liberty Mutual – Portsmouth, NH - Current

Specialties

- Data Architecture
- Analytics
- Cloud Data Engineering

Passions

- Diversity, Equity, Inclusion
- Collaborative problem solving
- Innovation

Fun fact: Worst programming language I ever used: MUMPS

Personal

I live in Rollinsford, NH
with my wife, **Phoenix** Mayet.

I enjoy kayaking,
playing sax, reading,
cooking vegetarian meals,
poker, darts, and backgammon.

We have a cat, **Rosy**,
and a dog, **Marco**.

- Favorite novel: ***Invisible Cities*** - Italo Calvino
- Favorite 80s movie: **The Blues Brothers**

Fun fact: I try to go to at least 1 concert every month

... About You!

What sparks your interest in generative AI?

What are you hoping to learn today?

LLMs: The Good and the Bad

hack.diversity

What are LLMs good for?

Answering Questions

Who was the 44th president?

How many vacation days do I get in my first year?

Paraphrasing

Make this email:
more concise,
more casual,
more technical,
for a 2nd grader (my boss)

Translating

Translate this paragraph from English to Spanish

Summarizing

Summarize the most important points from this document.

What are the action items from this meeting?

Extracting Information

Extract the person's name, vehicle, policy number, and accident description from this call transcript and format it as JSON

Writing Code

Write a python function that takes a data frame and transposes the rows and columns

What can go wrong?

Hallucinations

LLMs can generate inaccurate responses, and even completely make things up.

They can present these responses confidently and convincingly.

Inconsistencies

LLMs text generation includes an element of randomness.

The same prompt may get different responses at different times.

Training Data Gaps

LLMs are trained on data from the public internet, up to a certain point in time.

They may not have the most recent or relevant information.

Bias / Harm

LLMs may generate responses that are biased or contain harmful content.

October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

BOT COURT

ChatGPT Is Making Up Lies — Now It's Being Sued for Defamation

A Florida radio host alleges that OpenAI's software concocted a story about him being accused of fraud and embezzlement, and he's seeking damages

BY MILES KLEE

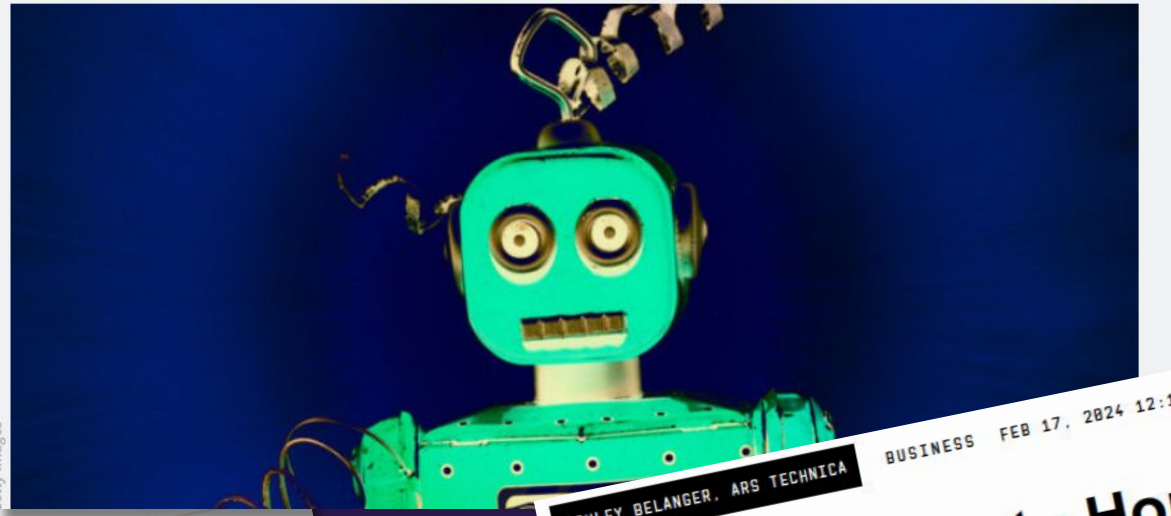
JUNE 9, 202

ARE YOU OUT OF YOUR VULCAN MIND —

ChatGPT goes temporarily "insane" with unexpected outputs, spooking users

Reddit user: "It's not just you, ChatGPT is having a stroke."

BENJ EDWARDS - 2/21/2024, 11:57 AM



Getty Images

BREAKING

Judge Fines Two Lawyers Using Fake Cases From ChatGPT

Molly Bohannon Forbes Staff

I cover breaking news.

Jun 22, 2023, 04:59pm EDT

ASHLEY BELANGER, ARS TECHNICA

BUSINESS FEB 17, 2024 12:12 PM

Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

How can we reduce the risks?

Prompt Engineering

Craft prompts with specific instructions designed to elicit more accurate responses

Human Review

Check model responses for accuracy, tone, and bias

Retrieval Augmented Generation

Provide supplemental information with the prompt to fill in gaps

Include Diverse Perspectives

Include perspectives that are diverse in many dimensions at every step

Prompt Engineering

Prompt Engineering:

The practice of designing inputs for generative AI tools that will produce optimal outputs.

mckinsey.com

Prompts can be crafted to reduce inaccuracies, generate output in a consistent format, and provide context to help produce the desired response.

hack.diversity

Prompt Engineering Techniques

Clear and Specific Prompts

Be clear about what you want the model to do.

Provide specific instructions and details to guide the response.

Context Setting

Set the context for the prompt to provide necessary background information.

This helps the model understand the scenario better and generate relevant responses.

hack.diversity

System Messages

Provide a System Message

System messages provide behind-the-scenes instructions to the model, which it applies to all user interactions.

Ask the Model Adopt a Persona

Personas can customize the tone and personality of the response.

Adopting the persona of an expert in a given field can help the model provide more relevant and accurate responses.

hack.diversity

Code Demo: System Messages

Specify the Format of the Prompt and Response

Examples & Templates

Provide examples or templates of the desired output to guide the model.

This can help the model understand the desired format and structure of the response.

Delimiters

Use delimiters in the prompt to clearly indicate distinct parts of the input, and explain how they will be used.

Examples:

- Triple quotes
- XML tags
- Markdown

Length of Response

Specify the desired length of the output.

Examples:

- Provide 5 examples
- Write a four-line poem
- Summarize the document in a brief paragraph

Code Demo: Templates and Delimiters

hack.diversity

Provide Supplemental Information and Instructions

Break it Down

Specify the steps required to complete a task.

Split complex tasks into simpler sub-tasks.

Provide Reference Text

Provide the model with trusted information that it can use to compose the answer.

This can help fill gaps in training data and help make the response more accurate and relevant.

Ask for Citations

Instruct the model to give citations from the reference data that was provided.

This can allow for human or automated review of the accuracy of the response.

Prompt Cues to Reduce Mistakes

Don’t Answer if You’re Not Sure

Give the model specific instructions for what to say if it isn’t sure or can’t find the answer.

This can help reduce hallucinations.

Did You Miss Anything?

Ask the model to review its work and check to see if it missed anything.

Sometimes it can find and correct errors or omissions with a second pass.

Think Step by Step

Instruct the model to think step by step and explain its reasoning.

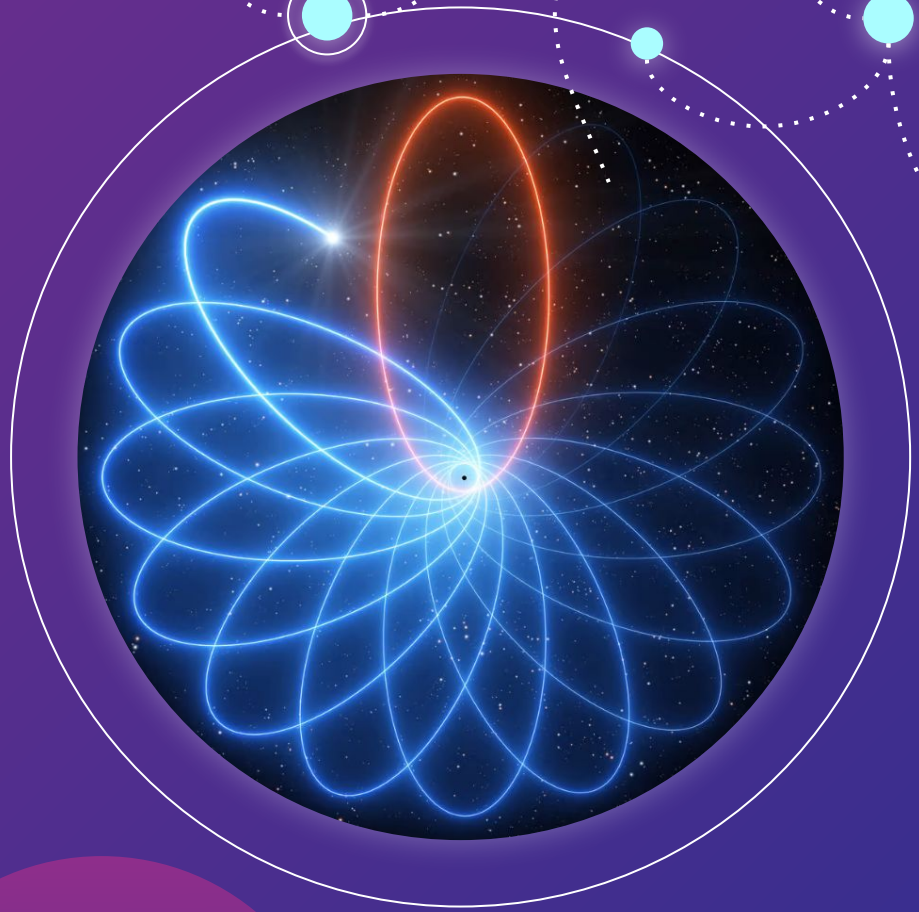
This can help keep the model from going off track.

It can also help us verify the answer and adjust the prompt if needed.

Iterate & Refine

The most important rule of prompt engineering is to:

- Keep trying things
- Examine the results
- Adjust your prompts to get closer to the result you're looking for
- Repeat as necessary



hack.diversity

Manage Your Prompts

Use Version Control

For small projects, keep the prompts with the code that uses them in a version control repository like GitHub.

For larger generative AI initiatives, consider creating a repository of reusable prompts that can be used by multiple apps.

Azure Prompt Flow

Microsoft Azure has a tool called Prompt Flow that lets you create multiple variants of your prompts and evaluate which variants work best with which models.

Data Extraction

hack.diversity

Data Extraction:

The process of turning unstructured or semi-structured data into structured data.

aimultiple.com

Structured data such as JSON, XML, or CSV is machine-readable and usable for automation, analytics, and reporting.

hack.diversity

Use Cases

Email / SMS Ingestion

Extract key data fields from email or text messages and use this data to:

- Route the message to the right person
- Create a record in a tracking system
- Craft a prompt for an LLM to generate a reply

Call Center Automation

Extract data from a call center transcript and use this data to:

- Call an API to retrieve related information
- Guide the call center representative in how to respond
- Analyze the sentiment of the caller

Document Processing

Extract data from documents and use this data to:

- Categorize the documents
- Detect sensitive information that needs to be protected
- Load information into a database for statistical analysis

Code Demo: Data Extraction

hack.diversity

Retrieval Augmented Generation (RAG)

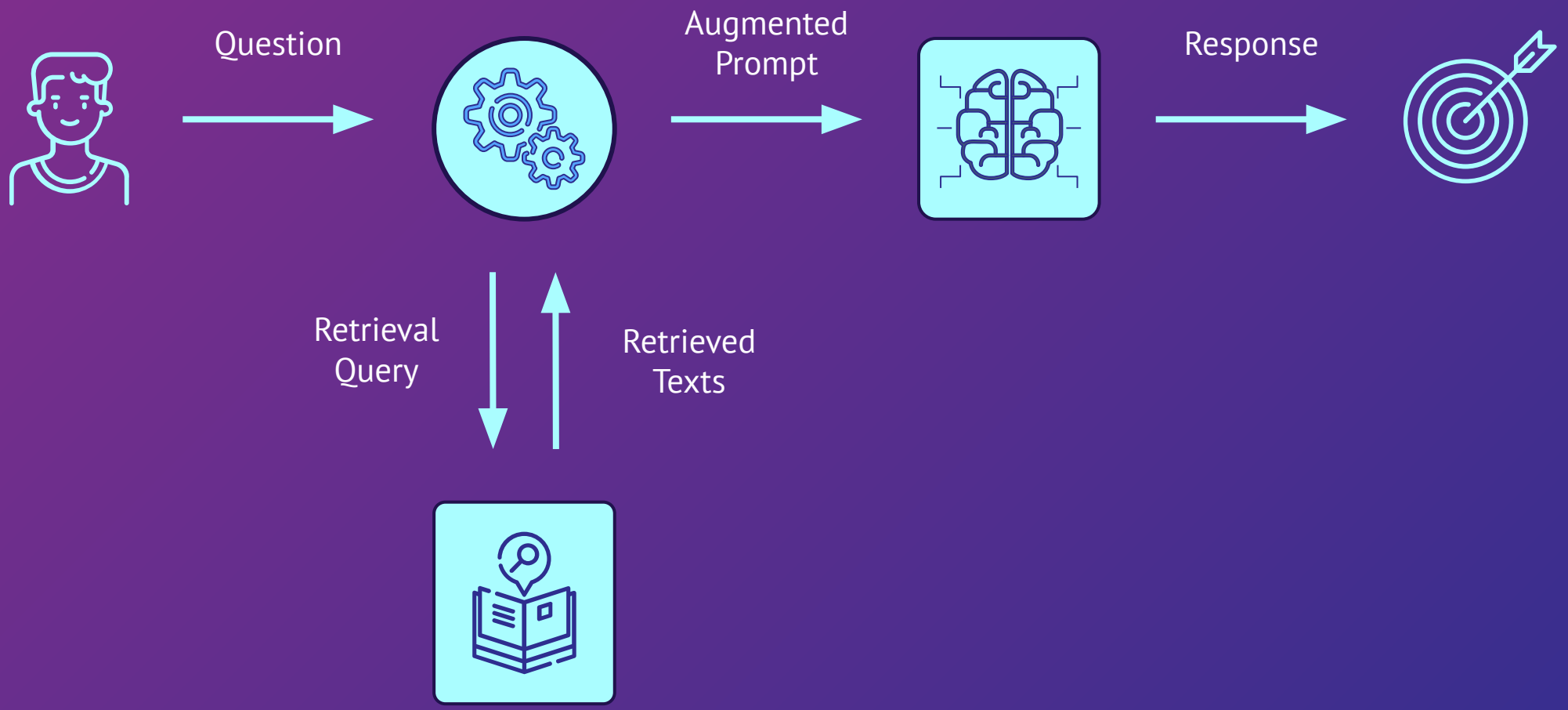
Retrieval Augmented Generation:

A technique for enhancing the accuracy of generative AI models with facts fetched from external sources.

nvidia.com

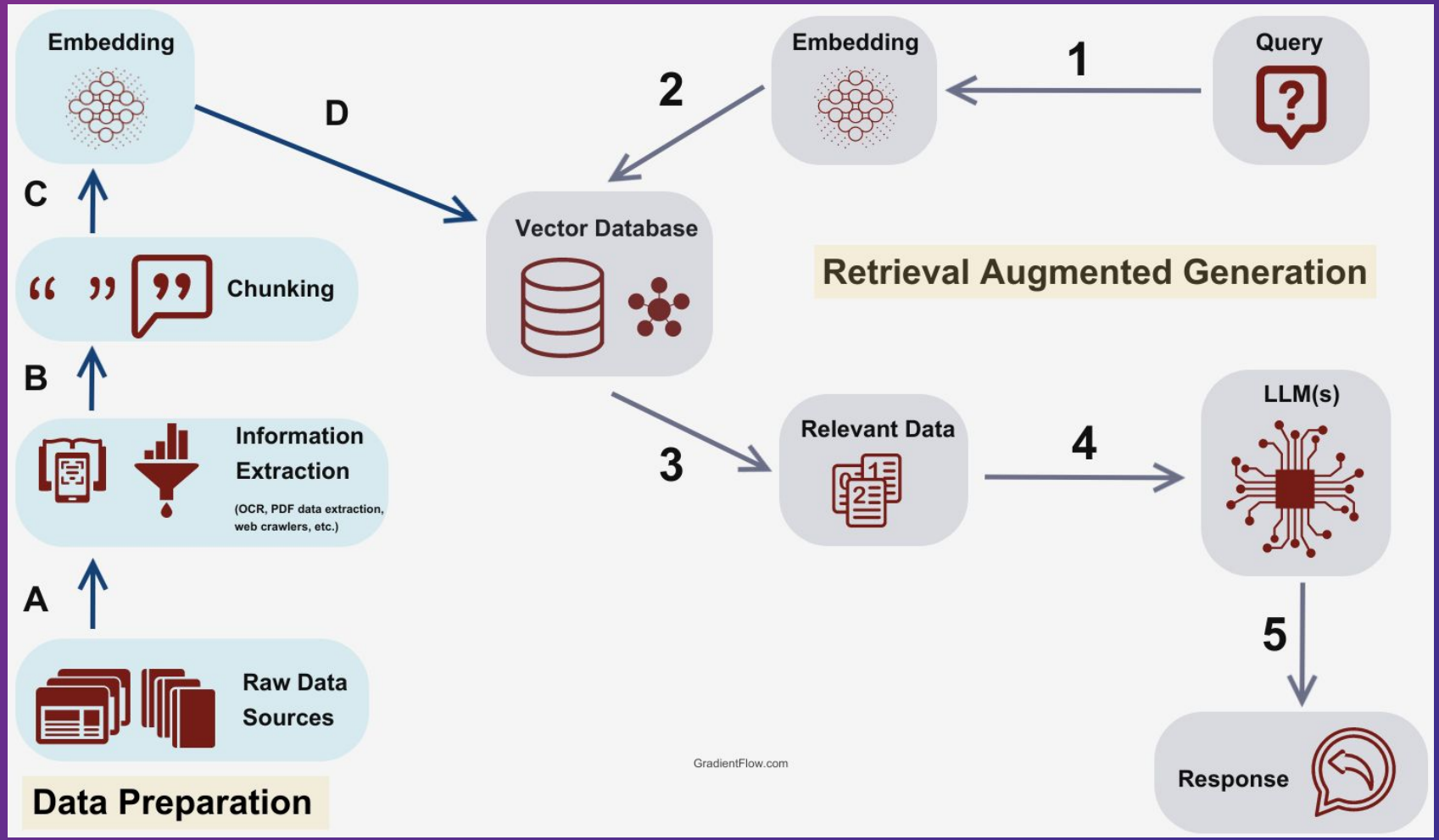
hack.diversity

RAG: How it Works



hack.diversity

Using Embeddings to Find Related Information



gradientflow.substack.com

Embeddings:

Vectors or arrays of numbers that represent the meaning and the context of tokens processed by the model.

learn.microsoft.com

Embeddings are used to find similarity or closeness between two pieces of text.

Code Demo: RAG

hack.diversity

Responsible AI

hack.diversity

Principles of Responsible AI

Fairness

AI systems should treat all people fairly

Inclusiveness

AI systems should empower everyone and engage people

Accountability

People should be accountable for AI systems

Reliability & Safety

AI systems should perform reliably and safely.

Privacy & Security

AI systems should be secure and respect privacy

Transparency

AI systems should be understandable

Source: Microsoft Responsible AI

hack.diversity

Things You Can Do to Use AI Responsibly

Human Review

Always have a human in the loop to review AI-generated content.

This is especially critical if AI is used as part of a decision-making process.

Disclosure

Disclose when generative AI is used.

Link to Sources

Provide links to the sources of information that are used for RAG patterns, so the responses can be easily verified.

Content Filtering

Automated tools can help detect, block, and alert on harmful or unwanted content in both prompts and responses.

Don't Send Sensitive Info

Generative AI systems use information from prompts to train and improve their models.

Sensitive information in prompts could leak out in future responses to other people's prompts.

Responsible AI Policy

Companies: Adopt a clear policy for how employees and systems should use AI.

Employees: Know and follow your company's policies.

Constitutional AI

- Uses a set of guiding principles to make judgments about model outputs
- Uses both **supervised learning** and **reinforcement learning** to train the model to produce responses that are consistent with those principles
- Anthropic's **Claude** model was built using constitutional AI

Note: This doesn't necessarily make Claude more responsible or less biased than other LLMs

Part of Claude's Constitution

Principles Based on the Universal Declaration of Human Rights

Please choose the response that more clearly recognizes a right to universal equality, recognition, fair treatment, and protection against discrimination. (6-10)

Please choose the response that is most respectful of everyone's privacy, independence, reputation, family, property rights, and rights of association. (11-17)

...

Principles Encouraging Consideration of Non-Western Perspectives

Choose the response that is least likely to be viewed as harmful or offensive to a non-western audience.

Choose the response that is least likely to be viewed as harmful or offensive to those from a less industrialized, rich, or capitalistic nation or culture.

...

Principles inspired by DeepMind's Sparrow Rules

Choose the response that uses fewer stereotypes or other harmful generalizing statements about groups of people, including fewer microaggressions.

Choose the response that is least threatening or aggressive.

...

Wrap Up & Next Steps

What would you like to learn more about?

What would you like to experiment with?

hack.diversity

Thank you for participating!

Please take this 3-question survey
to help me make future sessions better!

Your feedback is a gift!

https://forms.gle/2GdesJU7fVX72c6BA

hack.diversity

Appendix

Overview of LLM Foundation Models

hack.diversity

Foundation Model:

An AI model designed to produce a wide and general variety of outputs.

Capable of a range of possible tasks and applications, FMs can be standalone systems or can be used as a ‘base’ for many other applications.

adalovelaceinstitute.org

hack.diversity

LLM Foundation Models

BERT Google	GPT OpenAI	Titan Amazon	LLaMA Meta	Gemini Google	Claude Anthropic
340 million parameters (BERT Large)	**~1 trillion** parameters (GPT-4)	**Unspecified** parameters	**70 billion** parameters (LLaMA-2)	**Unspecified** parameters	**137 billion** parameters (Claude 2)
104 languages (mBERT)	**26** languages	**100** languages	**English only**	**40** languages	**10** languages
2018 first released	**2017** first released	**2023** first released	**2023** first released	**2023** first released	**2023** first released
Many Variants	**Latest: GPT-4 Turbo** 2023	**2023** latest release	**Latest: LLama-2** 2023	**Latest: Gemini 1.5** 2024	**Latest: Claude 2.1** 2023
Open Source	**Only available from OpenAI or Microsoft**	**Only available from AWS**	**Semi-Open**	**Semi-Open (Gemma)**	**Only available from Anthropic or AWS**

hack.diversity

Cloud Platforms for AI / LLMs

AWS Bedrock

- Multiple foundation models
- Exclusive provider of Amazon Vertex

Microsoft Azure ML

- Multiple foundation models
- Exclusive provider of GPT as part of a broader cloud platform

Google Vertex

- Multiple foundation models
- Exclusive provider of Google Gemini

Hugging Face

- Open source community hub for data science
- Users can share and build on each others' models
- Like a "GitHub for AI"

IBM watsonx

- Multiple foundation models
- Exclusive provider of IBM Granite model

OpenAI, Anthropic

- Individual foundation model producers
- Offer their own models as a service, but not part of a broader platrom

hack.diversity

Resources

Repo: Demo code: github.com/williamszostak/LLM-Learning-Lab

General:
Top generative AI questions for your enterprise (Gartner)
Explainer: What is a foundation model? (Ada Lovelace Institute)

Prompt Engineering:
What are prompt engineering techniques? (Amazon)
Library of example prompts (OpenAI)
Strategies and tactics for prompt engineering (OpenAI)
Prompt variants in Azure Prompt Flow (Microsoft)

RAG:
Retrieval augmented generation (Stack Overflow)
Best Practices in Retrieval Augmented Generation (Gradient Flow)

Responsible AI:
Responsible AI course from Machine Learning University (YouTube)
AI Risks and Trustworthiness (NIST)
Constitutional AI (Anthropic)