Learn Hands-on Skills to Minimize Hallucinations in AI-Generated Content

Ruopeng An, Ph.D., MPP, FACE, FAAHB
Associate Professor
Faculty Lead in Public Health Sciences, Brown School
Faculty Fellow in AI Innovations for Education, Office of the Provost
Brown School & Division of Computational & Data Sciences

Washington University in St. Louis

Outline

- What is hallucination in LLMs?
- Why do LLMs hallucinate?
- What are the ways to address hallucination?
  - Due diligence in AI utilization
  - Prompt engineering
  - Fact-checking and verification
  - Customized GPTs
  - Retrieval augmented generation (RAG)

Introduction to Hallucination in LLMs

- Hallucination refers to the generation of information or data that is not grounded in the input provided or factual evidence.

- It is akin to creating plausible-sounding but incorrect or unverifiable content.

- Understanding and addressing hallucinations is crucial for reliable AI applications.

What Does Hallucination Look Like?

- Non-factual Statements: False information presented as true.
- Overgeneralizations: Broad conclusions drawn from narrow data.
- Misinterpretations: Skewed output unrelated to input context.

Real-World Examples of Hallucination

- Incorrect Facts: "The Eiffel Tower is the tallest building in Europe."
- Invented Quotes: "I have a dream" attributed to the wrong person.
- Wrong Attribution: Crediting Newton with the invention of the lightbulb.

Visualizing Hallucinations

- Visible output is just the tip.
- Underlying inaccuracies create the bulk of hallucinations.

Underlying Causes of Hallucinations in LLMs

- Training Data Issues: Biased, inaccurate, or insufficient datasets.
- Static Knowledge: Models are trained on data up to a certain point.
- Lack of True Understanding: Responses are generated based on likelihood rather than understanding.
- Overfitting: Models may memorize rather than understand data.

Underlying Causes of Hallucinations in LLMs (continued)

- Lack of World Knowledge: LLMs lack an inherent understanding of the world and cannot verify facts against real-world data or events. They operate on textual data alone.

- Complexity of Human Language: The complexities of human language, including idioms, metaphors, and implicit meanings, can be challenging for LLMs to interpret correctly.

- Contextual Limitations: The limited context window of LLMs (the number of words or tokens they can consider at one time) can lead to misunderstandings or oversimplified interpretations of complex topics.

The Imperative of Due Diligence in AI

Why due diligence is non-negotiable?

- Human Oversight: Essential to catch and correct AI errors.
- Responsibility: Machines lack moral and legal responsibility.

The Fallibility of Machines

AI is not infallible!

- Error-Prone: Just like humans, machines make mistakes.
- Trust: Blind trust in AI can lead to the propagation of errors.

The Risks of Overreliance on AI

- Misinformation: Incorrect AI outputs can spread quickly if unchecked.

- Ethical Implications: Overreliance on AI may lead to ethical lapses and accountability issues.

Collaborative Role of Humans in AI

- Partnership: Humans and AI should work together, leveraging the strengths of each.

- Quality Control: Human expertise is crucial for validating AI-generated content.

- “Human-in-the-loop”

Case Study of Due Diligence Impact

- Example: A review process that caught and corrected an AI's inaccurate legal advice.

- Lesson: Vigilance in review prevented potential legal ramifications.

Introduction to Prompt Engineering

- Definition: Crafting input to guide AI toward the desired output.

- Goal: Minimize hallucination by precise communication.

Zero-Shot Prompting: The Foundation

- Definition: Providing a task without example, relying on AI's pre-training.
- Example: "List the behavioral risk factors for cardiovascular disease."

One-Shot Prompting

- Definition: Providing one example to establish context.

- Example: "Behavioral risk factors for diseases are important. For example, smoking is a behavioral risk factor for lung cancer. Now, list the behavioral risk factors for cardiovascular disease."

Few-Shot Prompting

- Definition: Presenting several examples to define a pattern.
- Example: "Here are a few behavioral risk factors for different diseases: smoking for lung cancer, excessive sun exposure for skin cancer, and high sugar intake for diabetes. Based on these, list the behavioral risk factors for cardiovascular disease."

One-Shot Prompting - Tailored Example

- Approach: Provide one precise instance, then ask for similar instances.

- Example: "A behavioral risk factor for cardiovascular disease is a sedentary lifestyle. Identify more such risk factors."

Few-Shot Prompting - Tailored Example

- Approach: Supply multiple specific examples to shape AI response.

- Example: "Behavioral risk factors for various diseases include sedentary lifestyle for cardiovascular disease, frequent alcohol consumption for liver disease, and poor diet for type 2 diabetes. Continue this list for cardiovascular disease."

What is Context Injection?

- Definition: A method of providing additional background information in prompts to guide AI understanding.

- Mechanism: Embedding extra background information within prompts to help narrow down AI responses.

Importance of Context Injection

- Enhances Clarity: Helps AI to discern the intent of the prompt more accurately.

- Limits Errors: Specific context can reduce the likelihood of irrelevant or incorrect outputs.

Baseline Prompt Without Context

- Example Prompt: "List the risk factors for cardiovascular disease.”
- Note: This basic prompt lacks detail and may yield a general response.

The Role of Context Injection

- Clarifies Intent: Informs the model about the specific angle or domain of interest.
- Narrows Focus: Directs the AI to consider specific conditions or populations.
- Improves Relevance: Encourages more accurate and detailed responses aligned with the user's needs.

Context Injection in Action

- Detailed Context: "Cardiovascular diseases remain the leading cause of death globally. Despite advancements in medical technology, behavioral risk factors play a significant role in the prevalence of these conditions. Considering the latest research linking lifestyle and heart health, particularly in urban environments where sedentary behavior is more common..."

- Injected Prompt: "...please provide an updated list of behavioral risk factors for cardiovascular disease that are most relevant in urban settings."

Integrating Verifications into Prompt Design

- Citation Integration: Instruct LLMs to include citations in their responses. This practice encourages the model to generate outputs anchored in verifiable sources.

- Keyword-Based Verification: Include specific keywords in prompts that direct LLMs to confirm the relevance and accuracy of the sources they reference.

- Direct Link Testing: Provide URLs in responses, enabling users to test the existence and relevance of cited sources via direct links.

Benefits of Verification Mechanisms

- Trust: Builds user confidence in AI reliability.
- Independence: Empowers users to verify information themselves.
- Engagement: Encourages critical thinking and engagement with AI outputs.

Citation Integration in LLM Prompts

- Purpose: Guides LLMs to base responses on cited sources.
- Method: The prompt includes an instruction to cite sources.
- Example: ”Please elucidate the impact of soda tax on sugar-sweetened beverage consumption. Please provide complete citations in AMA format at the end of your response."

Key-word Based Verification

- Strategy: Embed specific, verifiable keywords related to reputable sources.
- Outcome: AI cross-references keywords with its knowledge to enhance reliability.
- Example: "Include keywords such as 'A1C levels,' 'insulin therapy,' and 'carbohydrate counting' when listing diabetes treatment protocols."

Direct Web Link Testing

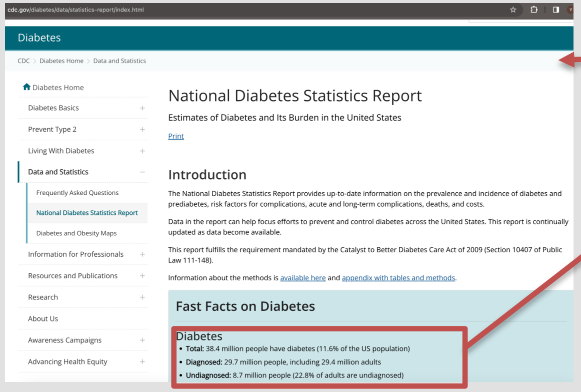
- Advantage: Allows users to independently verify the AI's responses.
- Approach: Include URLs in AI-generated content for source checking.
- Example: "In the discussion of heart disease prevention, provide URLs to the American Heart Association's published resources."

Case study: Verifying Health Data through Enhanced Prompt Engineering

- Case: Query an LLM about the latest statistics on diabetes prevalence in the United States.
  - Traditional Prompt: “What are the current statistics on diabetes prevalence in the United States?”
  - Traditional Response: “As of the latest data, approximately 10.5% of the U.S. population has diabetes.” (Note: This response could be outdated or incorrect depending on the LLM’s training data cut-off.)
- Enhanced Prompt:

You
What are the current statistics on diabetes prevalence in the United States? Please include citations from CDC or recent medical studies published in 2024, and provide a direct link to the source.

- The enhanced response not only provides the latest statistics but also cites a specific, authoritative source. The direct link allows users to verify the information independently.


cdc.gov/diabetes/data/statistics-report/index.html

Diabetes

CDC > Diabetes Home > Data and Statistics

Diabetes Home
Diabetes Basics +
Prevent Type 2 +
Living With Diabetes +
Data and Statistics –
Frequently Asked Questions
National Diabetes Statistics Report
Diabetes and Obesity Maps
Information for Professionals +
Resources and Publications +
Research +
About Us
Awareness Campaigns +
Advancing Health Equity +

National Diabetes Statistics Report

Estimates of Diabetes and Its Burden in the United States

Print

Introduction

The National Diabetes Statistics Report provides up-to-date information on the prevalence and incidence of diabetes and prediabetes, risk factors for complications, acute and long-term complications, deaths, and costs.

Data in the report can help focus efforts to prevent and control diabetes across the United States. This report is continually updated as data become available.

This report fulfills the requirement mandated by the Catalyst to Better Diabetes Care Act of 2009 (Section 10407 of Public Law 111-148).

Information about the methods is available here and appendix with tables and methods.

Fast Facts on Diabetes

Diabetes
- **Total:** 38.4 million people have diabetes (11.6% of the US population)
- **Diagnosed:** 29.7 million people, including 29.4 million adults
- **Undiagnosed:** 8.7 million people (22.8% of adults are undiagnosed)

Diabetes
- **Total:** 38.4 million people have diabetes (11.6% of the US population)
- **Diagnosed:** 29.7 million people, including 29.4 million adults
- **Undiagnosed:** 8.7 million people (22.8% of adults are undiagnosed)

- Enhanced Response:

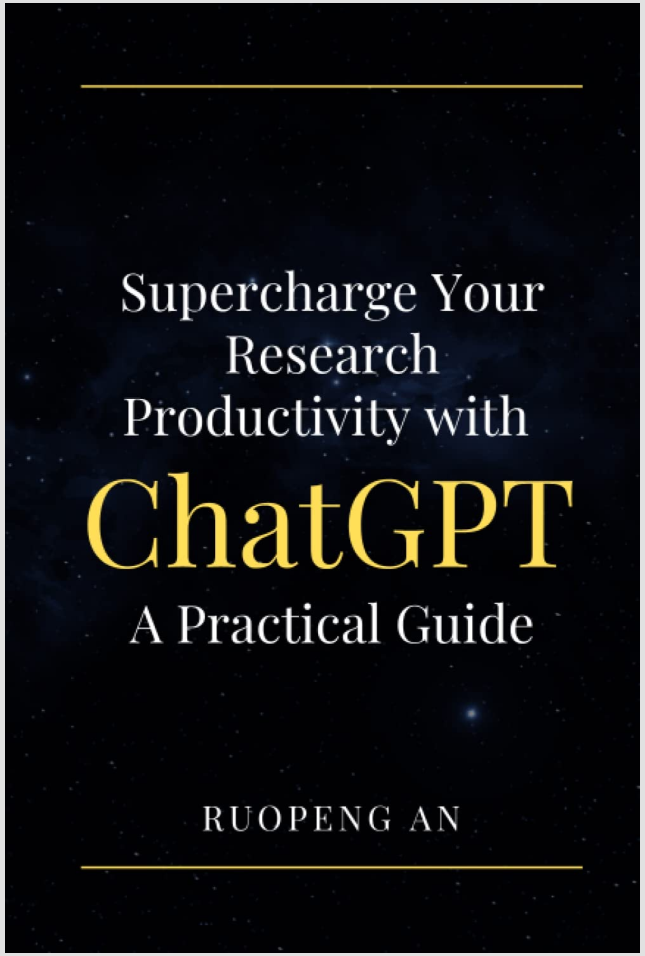
ChatGPT
As of 2024, the prevalence of diabetes in the United States continues to be a significant public health concern. According to the latest data from the Centers for Disease Control and Prevention (CDC), approximately 11.6% of the U.S. population, or 38.4 million Americans, have diabetes. This figure includes both diagnosed and undiagnosed cases, with 29.7 million people diagnosed and an estimated 8.7 million undiagnosed (CDC).

You can access detailed information and additional data from the National Diabetes Statistics Report directly from the CDC at CDC's Diabetes Statistics Report. This report includes comprehensive statistics on the prevalence and incidence of diabetes across various age, racial, and ethnic groups, as well as data on complications and comorbidities associated with diabetes.

Demo from My Book

100 real-world examples covering 10 phases of research:

1. Identifying research topics and framing questions through an in-depth discussion with ChatGPT
2. Formulating and refining hypotheses based on the chosen research question
3. Undertaking literature reviews, covering all steps of a systematic review protocol
4. Selecting adequate research design and corresponding methodology
5. Developing valid, reliable, and efficient research tools
6. Handling every aspect of data collection, management, and ethics
7. Interpreting and analyzing both quantitative and qualitative data
8. Writing and refining research papers and reports
9. Addressing peer review comments
10. Disseminating study findings through mass and social media platforms



Amazon: https://a.co/d/6Y7C5G0

Third-Party Fact-Checking Tools

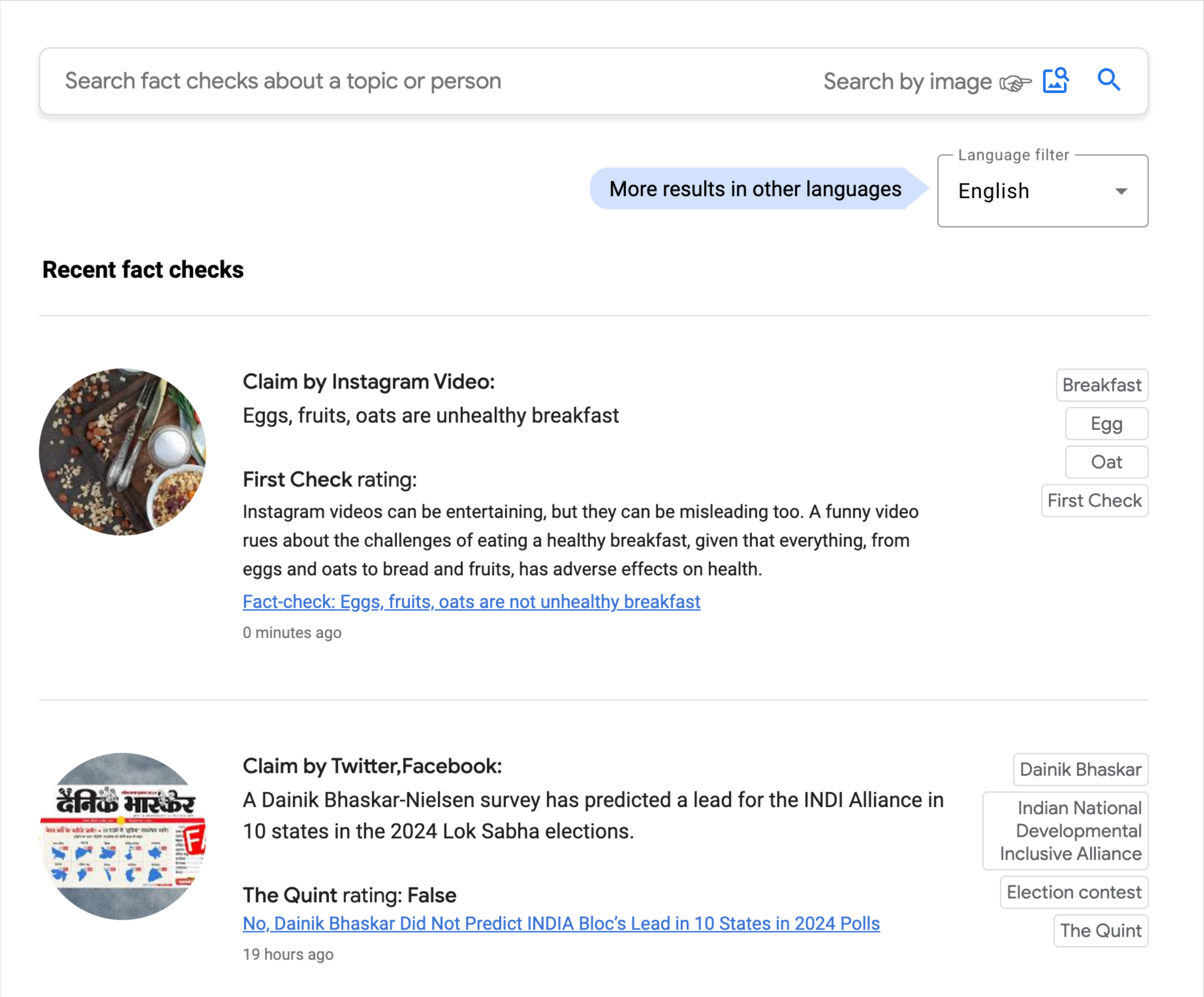
- Overview: External tools validate AI-generated content.
- Integration: LLMs can be programmed to query databases or use APIs of trusted fact-checking services.

Leveraging Fact-Checking Tools Against Hallucinations

Google Fact Check Tools API
https://developers.google.com/fact-check/tools/api

- This API allows you to query the claims that Google's Search engine has associated with fact-checking articles. It helps validate the facts that LLMs produce by comparing them with verified sources.

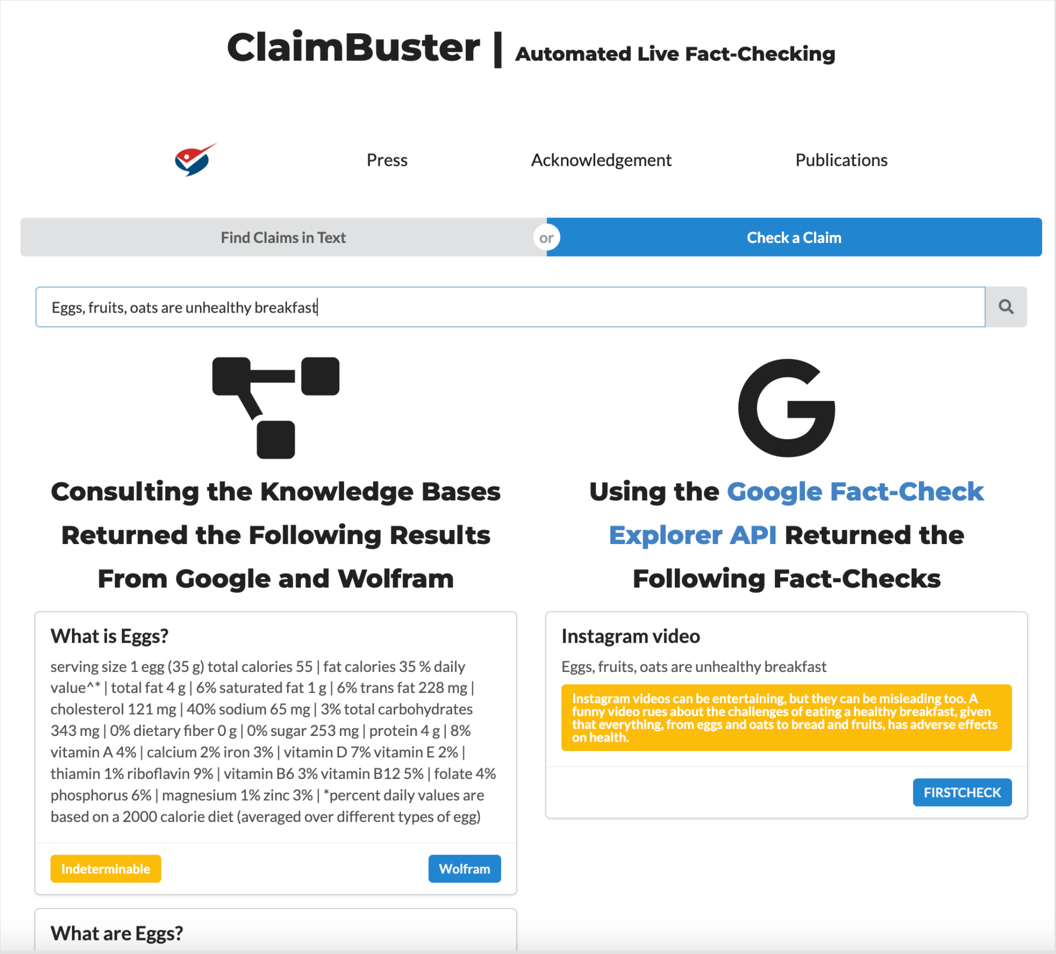
- Use Case: Ideal for real-time verification of content generated by LLMs to ensure accuracy and authenticity. https://toolbox.google.com/factcheck/explorer/search/list:recent;hl=en



Leveraging Fact-Checking Tools Against Hallucinations

ClaimBuster API
https://idir.uta.edu/claimbuster/api/

- Developed by the University of Texas at Arlington, this tool is designed to automatically detect check-worthy factual claims. It's particularly useful for identifying and verifying factual statements produced by LLMs.
- Use Case: Can be integrated into LLM pipelines to automatically flag and verify claims that need fact-checking. https://idir.uta.edu/claimbuster/factchecker/

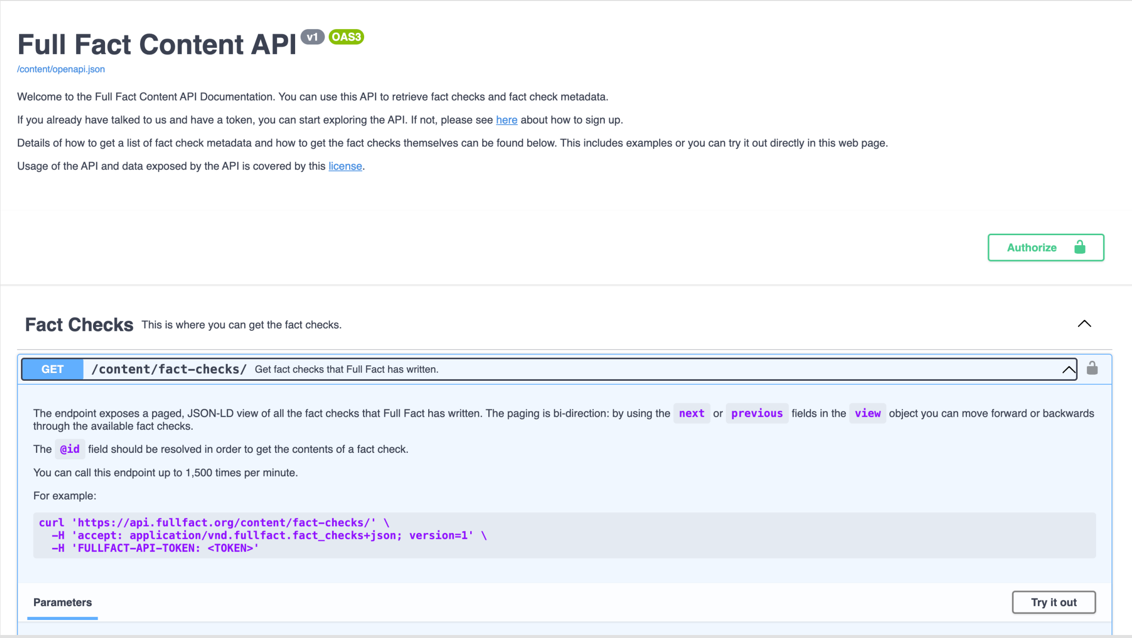


Leveraging Fact-Checking Tools Against Hallucinations

Full Fact API
https://api.fullfact.org/content/documentation

- Full Fact offers tools that support automated fact-checking. While primarily UK-focused, it provides access to a broad range of fact-checked data that can be used to verify LLM outputs.

- Use Case: Useful for LLMs generating content related to topics commonly discussed in public discourse, as it can cross-reference and verify these against established facts.



Leveraging Fact-Checking Tools Against Hallucinations

Bing-Copilot

https://www.bing.com/chat?q=Bing+AI&FORM=hpcodx

Try the following prompt yourself:

“Does egg consumption increase blood pressure?”

Leveraging Fact-Checking Tools Against Hallucinations

You.com

https://you.com/

Try the following prompt yourself:

"Consuming bananas reduces one's risk of COVID-19 infection."

Customizing GPTs for Grounded Outputs

- Customization: Tailoring GPTs to specific datasets for informed responses.

- Grounding: Ensuring outputs reference the content within uploaded documents.

Steps to Customize GPTs via ChatGPT

- Step 1: Define your topic or domain of interest.
- Step 2: Gather relevant and reliable documents.
- Step 3: Upload documents to the ChatGPT interface as references.
- Step 4: Craft prompts that reference the uploaded material.
- Step 5: Interact and provide feedback to refine the model’s responses.

Ensuring Accuracy in Customized GPTs

- Verification: Regularly check the model's references against uploaded content.
- Feedback Loop: Use ChatGPT's feedback feature to continuously improve the output quality.

Practical Application of Custom GPTs

- Usage Example: Reference a database of scientific articles to answer research-related queries.

- Output Example: "Based on the documents provided on cardiovascular health, summarize the latest findings on risk factors."

Case Study: Customized GPT for Education

- Scenario: A GPT customized to assist with homework based on a curriculum's textbooks.

- Process: Uploading textbooks and crafting prompts for subject-specific educational assistance.

Customized GPTs Demo

Try the following prompt on my customized GPTs:

“Please design a multiple-choice question to test my knowledge about random sampling based on the statistics textbook I uploaded.”

Build Your Own AI Tutor with GPTs

Dr. An's YouTube Channel:

Introduction to Retrieval Augmented Generation (RAG)

- Definition: RAG combines retrieval of information from a database with the generative capabilities of an LLM.

- Process: The model retrieves relevant documents and uses that information to inform its responses.

How RAG Reduces Hallucinations

- Information Retrieval: RAG first searches a dataset to find relevant context.

- Response Generation: Then, it synthesizes the information into coherent, grounded responses.

The Mechanism of RAG

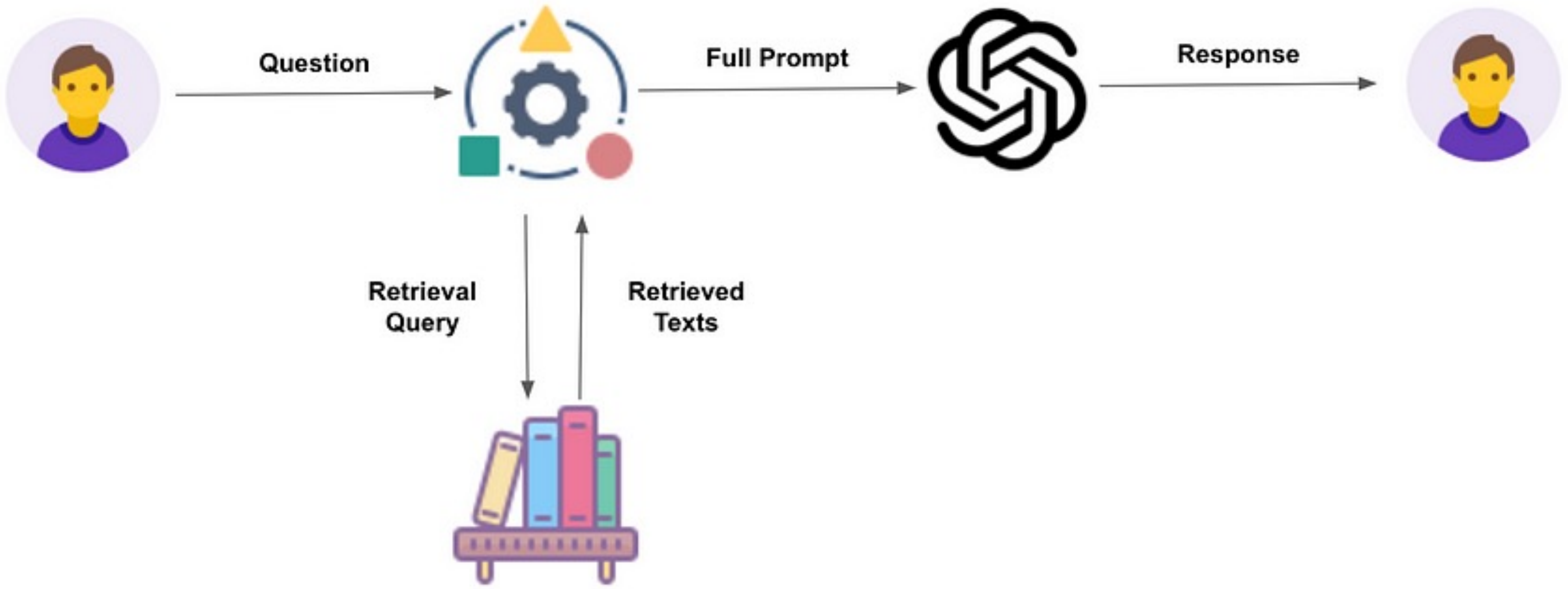
- Step 1: Query processing to understand the user's request.
- Step 2: Document retrieval from a large corpus of data.
- Step 3: Content generation that incorporates retrieved data to ensure factual correctness.

Benefits of Using RAG

- Relevancy: Outputs are directly related to source material, increasing accuracy.

- Trust: Grounding responses in documented sources builds user trust.

A High-level Overview of RAG Pipeline



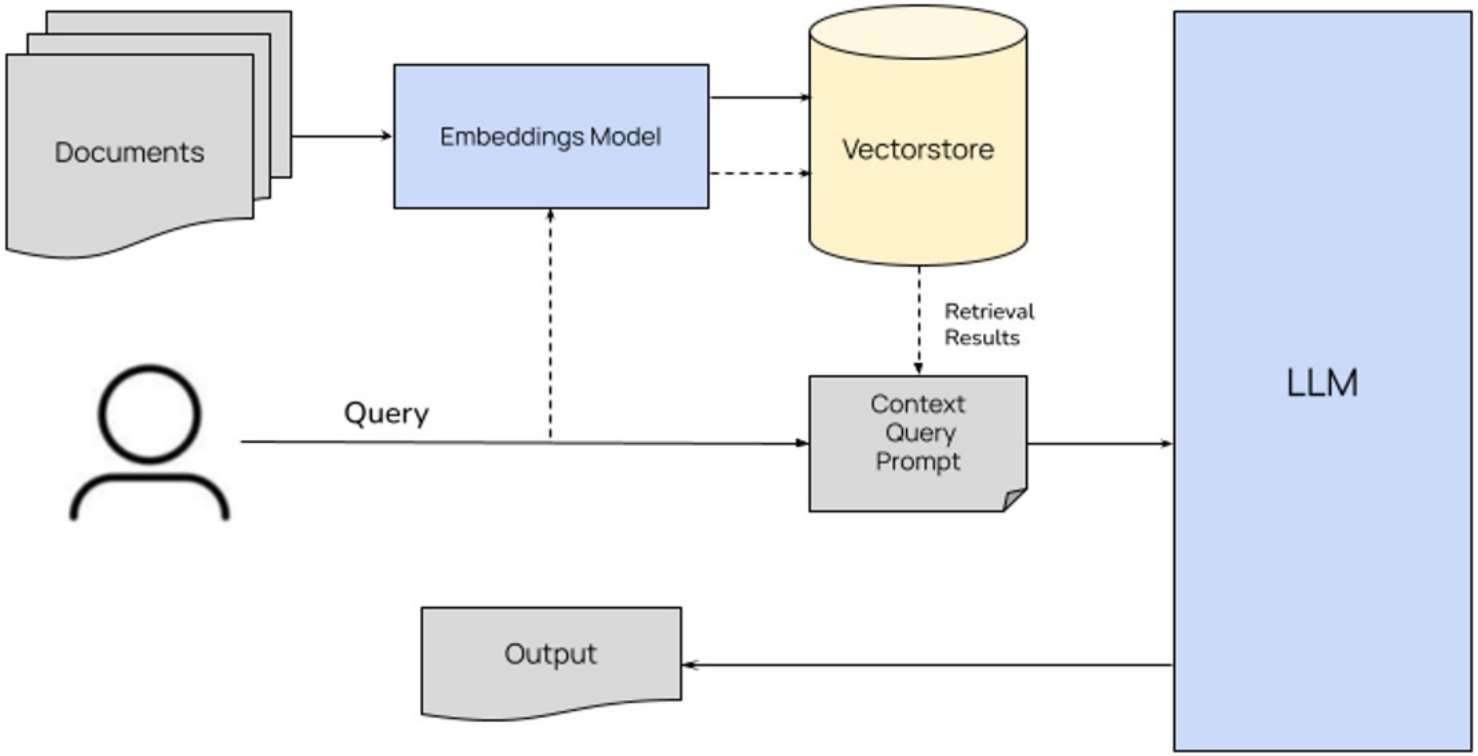
COVID-19 Fake Claims on Social and Mass Media

- Widespread disinformation on COVID-19 over social media and news.

- Generative AI (e.g., “deep fakes”) accelerates such dissemination.

- LLMs suffer from hallucination problems and are less helpful in fake claim detection.

Real-time Detect & Debunk COVID-19 Fake Claims



- Vector Database: 120,000 peer-reviewed articles on COVID-19
- Embedding Model: OpenAI
- LLM: GPT-4

Performance: LLM vs. RAG

- 500 real-world claims (50% true & 50% false):
  - GPT-4 without RAG: 85.8% accuracy
  - GPT-4 with RAG: 94.6% accuracy

- 500 AI-generated claims (50% true & 50% false):
  - GPT-4 without RAG: 97.3% accuracy
  - GPT-4 with RAG: 98.6% accuracy

Example Claim: Houseflies Can Transmit COVID-19

Response:

The claim that houseflies can transmit the COVID-19 virus is FALSE. The references given discuss the presence of SARS-CoV-2 on various surfaces, potential fecal-oral transmission, and the virus' presence in gastrointestinal cells. However, there is no evidence provided in these documents that directly indicates houseflies can transmit COVID-19.

Titles and authors of the context referred:

1. Nekoei S, Khamesipour F, Benchimol M, Bueno-Marí R, Ommi D. SARS-CoV-2 Transmission by Arthropod Vectors: A Scoping Review. Biomed Res Int. 2022;2022:4329423.
2. Dolinaj V, Grujić J, Križanović D, Potkonjak A, Pape T, Banović P. The Price of Hospital Reshaping: Nasal Myiasis Caused by Flesh Fly (Diptera: Sarcophagidae) in Reallocated COVID-19 Intensive Care Unit. Healthcare (Basel). 2023;11(11):1533.

Recap

- What is hallucination in LLMs?
- Why do LLMs hallucinate?
- What are the ways to address hallucination?
  - Due diligence in AI utilization
  - Prompt engineering
  - Fact-checking and verification
  - Customized GPTs
  - Retrieval augmented generation (RAG)

Conclusions

- Hallucinations are inherent to LLMs but there are many practical ways we can substantially mitigate their impact.

- Don’t be scared away from AI by its hallucinations. On the contrary, try to form a human-machine team/partnership.

- “Cautious optimism”: Hope for the best, but not solely rely on the best outcome, striking a balance between optimism and realism.

- Learn AI ASAP and AMAP!

Interested in Learn More About AI?

Artificial Intelligence Applications for Health Data Advanced Learning Certificate Program at WashU

To learn more about the certificate program and participants’ testimonials:

https://aicademe.publish.library.wustl.edu/advanced-learning-certificate/

Program Schedule

- Online-only Program via Zoom
- 15 weeks, 3 hours per week
- Weekly project-based assignment
- Early Sep – mid Dec, 2024
- Contact Prof. Ruopeng An: ruopeng@wustl.edu

Program Content (Weeks 1 and 2)

- An overview of AI
- Learn how to code in Python
- Use NumPy and Pandas to do data wrangling
- Use Matplotlib to do data visualization

Program Content (Weeks 3-7: Machine learning)

- End-to-end ML project using Scikit-Learn
- Classification tasks
- Regression tasks
- Model training and validation
- Support vector machines
- Decision trees
- Ensemble methods (e.g., random forest, XGBoost)
- Dimensionality reduction
- Unsupervised learning

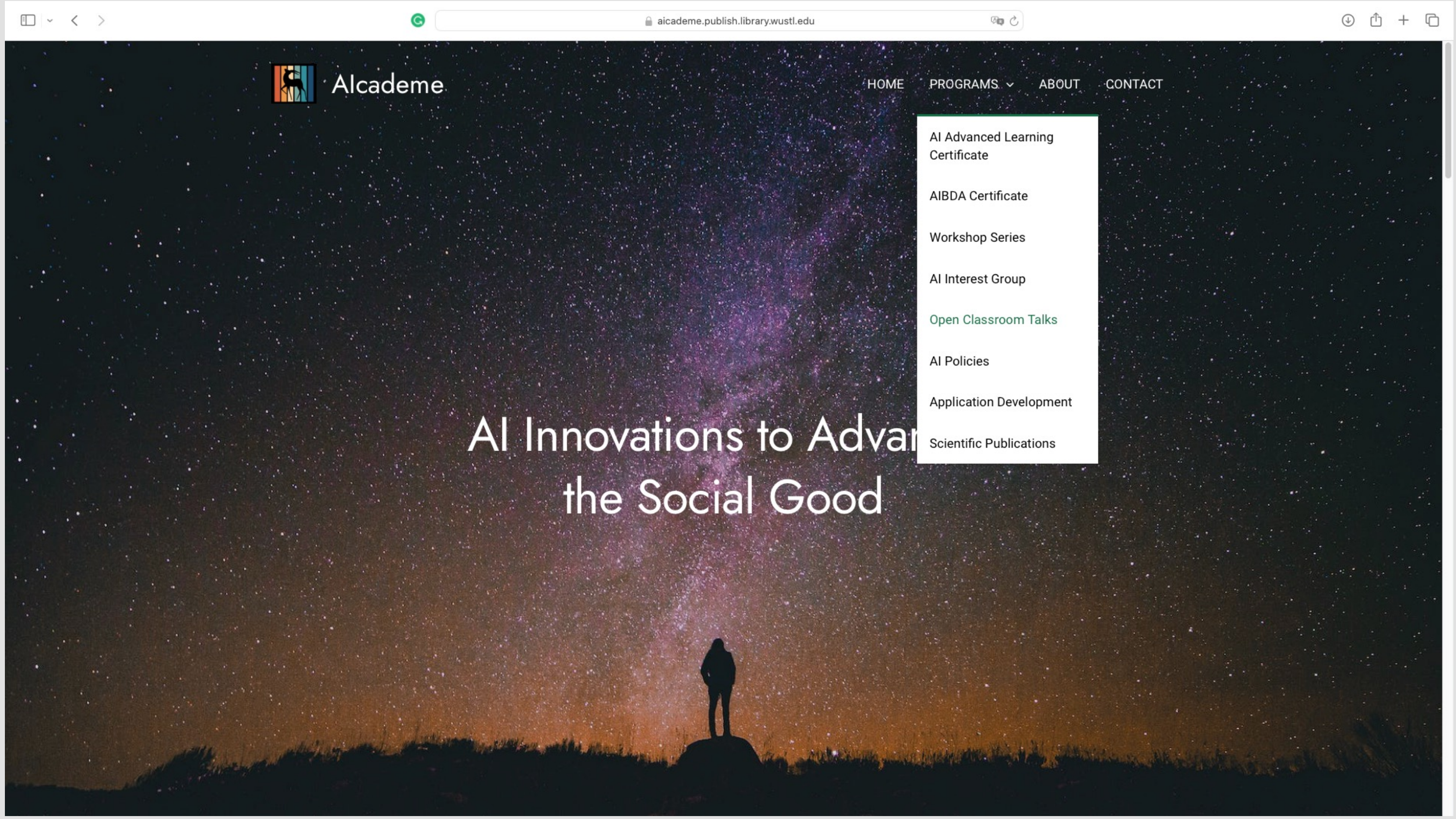
Program Content (Weeks 8-15: Deep learning)

- Neural network basics
- Computer vision: image classification, object detection, and image segmentation
- Time series forecasting
- Recommender system
- Natural language processing: sentiment analysis, text summarization, question-answering, chatbot, translation
- Generative deep learning for image and text generation
- Synthetic data generation

Course Format

- Seamless conjunction of lecture presentations and hands-on lab sessions
- Google Colab Jupyter notebook with free GPU-computing
- All code provided as templates for AI model prototyping
- Weekly assignment focusing on applying AI models to address real-world problems

Today’s Recordings & Slides on My Website



https://aicademe.publish.library.wustl.edu