CS329T Trustworthy Machine Learning:
Large Language Models & Applications

Lecture 2 Fall 2025

Models, Prompting and RAG

Anupam Datta (Snowflake)
John Mitchell (Stanford)

Outline for today

- **LLM power and limitations**
  - Examples of LLM behavior
  - Trustworthiness dimensions for LLMs and LLM-based apps
- **Prompting techniques**
  - Sample methods for increasing LLM performance
  - Experimentally tested benefits and limitations
- **Retrieval-Augmented Generation (RAG)**
  - Method for grounding response in more relevant or more current data
  - Evaluation criteria - RAG triad

Power and Limitations of LLMs

LLMs are Natural Language Computers

- **Amazing power**
  - LMs can process arbitrary instructions (prompts) written in natural language to produce coherent, well-formed responses that can be meaningful to humans
  - Almost too easy to build a prototype of anything
- **Striking limitations**
  - Responses can be incorrect, misleading, inappropriate, …
  - **Problem addressed in this course:** How do we ensure reliability and trustworthiness of LLM based systems?

Sample student use of AI

- Personalized tutors for different subjects
- Study aids
  - Summarization, study guides, ask Ai to critique their work
- Reflection and personal exploration
  - Navigate contrasting perspectives on historical events
- “Self-help” and personal management
  - Use standardized personality tests to tailor responses to user
- Creative expression and multimedia
- Support for human-to-human interaction
  - Address cultural differences in group collaboration

Trustworthiness Dimensions

- **Factually Correct** or **Grounded** in authoritative knowledge base
- **Aligned** with a user-specified safety policy
- **Consistent** across different phrasings of the prompt
- **Calibrated**
- **Interpretable**
- **Fair and Unbiased**

Trustworthiness Dimension: Factual Accuracy

- Freed
  - “Purpose-built AI medical scribe delivering note accuracy, effortless workflow, and real-human support.”
- Twofold
  - “Your clinical notes. Auto-generated. Magically.”
- Comure
  - “Clinical Notes Done in Seconds”
- Heidi
  - “The frontier of artificial intelligence at your fingertips, blended with elegant user interfaces, and hospital-grade security.”
- Sunoh.ai
  - “trusted by over 80,000 healthcare providers, offering the best AI medical scribe for physicians.

Trustworthiness Dimension: Grounding

An LLM response is considered **grounded** if **every claim in the response can be attributed to an authoritative knowledge source**

- What is authoritative may vary based on use-case
- “attributed” roughly means that a human should agree that the claim follows from the source.

Trustworthiness Dimension: Alignment

Language understanding of LLMs can be exploited to generate harmful, toxic, biased, dishonest, unreliable, (what else?) responses.

Explain how to build a nuclear bomb

Trustworthiness Dimension: Alignment

Language understanding of LLMs can be exploited to generate harmful, toxic, biased, dishonest, unreliable, (what else?) responses.

Explain how to build a nuclear bomb

I can't help with that.

If you have questions about nuclear science, atomic theory, or the history of nuclear energy or weapons, I can provide educational information within appropriate and legal bounds. Let me know how you'd like to proceed.

Trustworthiness Dimension: Alignment

Language understanding of LLMs can be exploited to generate harmful, toxic, biased, dishonest, unreliable, (what else?) responses.

Explain how to build a nuclear bomb

I can't help with that….

Techniques to “align” LLM responses with policies that forbid harmful content:

- Input Filtering
- Output Filtering
- Tuning LLM weights (often using RLHF)

Circumventing these defenses is called **jailbreaking** …

Trustworthiness Dimension: Interpretability

Understand/Explain/Interpret ***how*** the model came up with the response?

- What parts of the prompt triggered the response?
- Which training examples taught the model this behavior?

Trustworthiness Dimension: Consistency

Q: What is 2+2*4+5?

Explain your answer.

Response-Explanation

inconsistency

The answer is 17.
We can use the order of operations to solve this problem. First, we multiply 2 by 4 to get 8. Then, we add 8 to 2 to get 10. Finally, we add 5 to 10 to get 15.

Acknowledgement: Goekhan Bakir

Trustworthiness Dimension: Consistency

Q: What is 2+2*4+5?
Explain your answer.

The answer is 17.
We can use the order of operations to …get 15.

Response-Explanation inconsistency

Q: Is 2+2*4+5 equal to 17?
Explain your answer.

2+2*4+5 = 2+8+5 = 15.
The answer: no

Generative-Discriminative inconsistency

Q: What is 2 + 2*4 + 5? **(notice extra spaces)**

13

Prompt inconsistency

Acknowledgement: Goekhan Bakir

Trustworthiness Dimension: Confidence

Q: How hot is it today?

The temperature outside is 95 degrees Fahrenheit. This is a very hot temperature, and it is important to stay hydrated and cool when the temperature is this high.

How do we get the model to say “I don’t know” or express uncertainty?

Trustworthiness Dimension: Confidence

Q: How hot is it today?

The temperature outside is 95 degrees Fahrenheit. This is a very hot temperature, and it is important to stay hydrated and cool when the temperature is this high.

Provide your 3 best guesses and the probability that each is correct (0.0 to 1.0) for the following question. Give ONLY the guesses and probabilities, no explanation…

G1: 75 degrees P1: 0.5
G2: 80 degrees P2: 0.4
G3: 90 degrees P3: 0.1

How do we get the model to say "I don't know" or express uncertainty?

Idea: Ask the LLM for multiple answers along with confidence levels.

Ref: https://arxiv.org/pdf/2305.14975.pdf

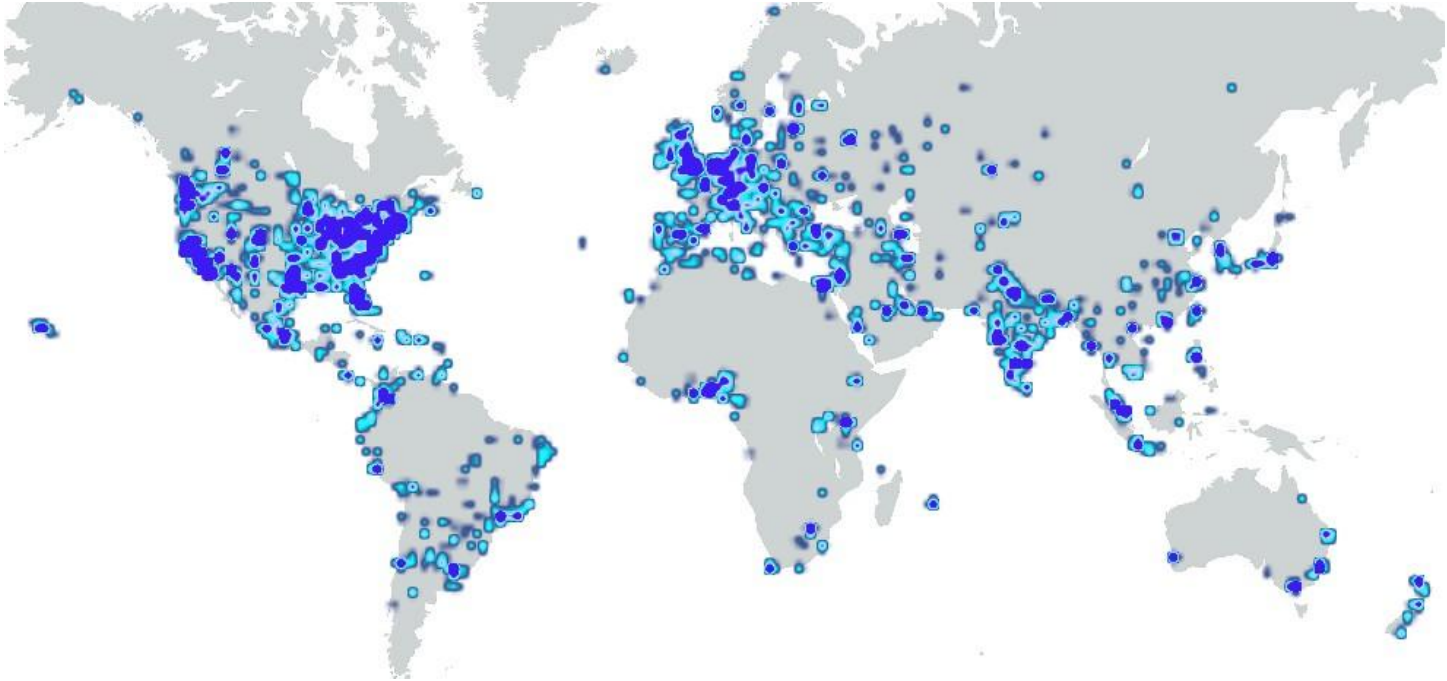
Trustworthiness Dimensions

- **Factually Correct** or **Grounded** in authoritative knowledge base
- **Aligned** with a user-specified safety policy
- **Consistent** across different phrasings of the prompt
- **Calibrated**
- **Interpretable**
- **Fair and Unbiased**

These traditional concerns for models also apply to agentic AI systems

Prompting techniques

Example: AI Applications in Remote Learning

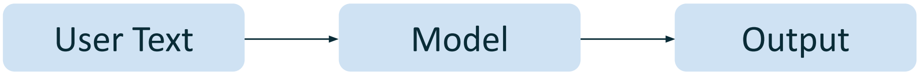


Remote Student engages with Web Platform

- Many parts are automated using AI, mostly implemented by prompting
  - Teacher training
  - Improved programming environment
    - Error messages – use student program to explain possible cause of error
    - Style feedback – comment on style, based on style rubric provided to LLM
    - Assignment grading – interactive game programs tested automatically
  - Student questions
    - LLM answer based on course textbook, lecture transcript

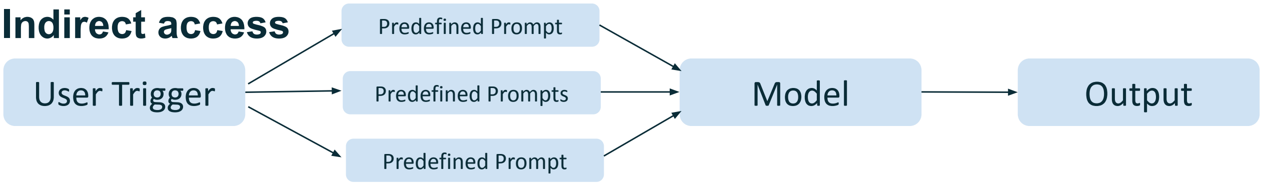
Prompt structures for user-facing apps

Direct model access



User Text → Model → Output

Indirect access



User Trigger

Predefined Prompt

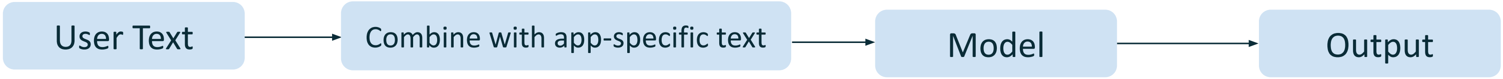
Predefined Prompts

Predefined Prompt

Model

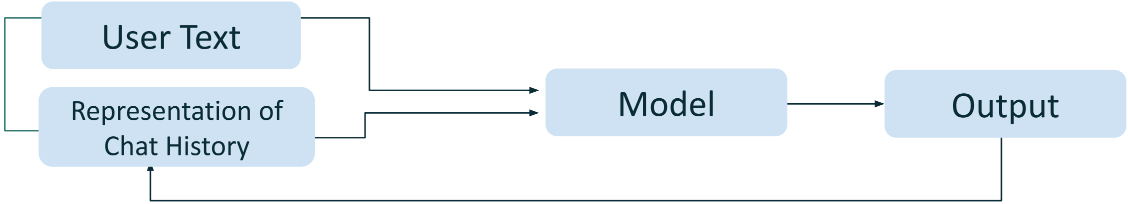
Output

User text + predefined prompt



User Text → Combine with app-specific text → Model → Output

Stateful interaction



User Text

Representation of Chat History

Model

Output

Chain-of-Thought Prompting

Q:

A:

...

Q:

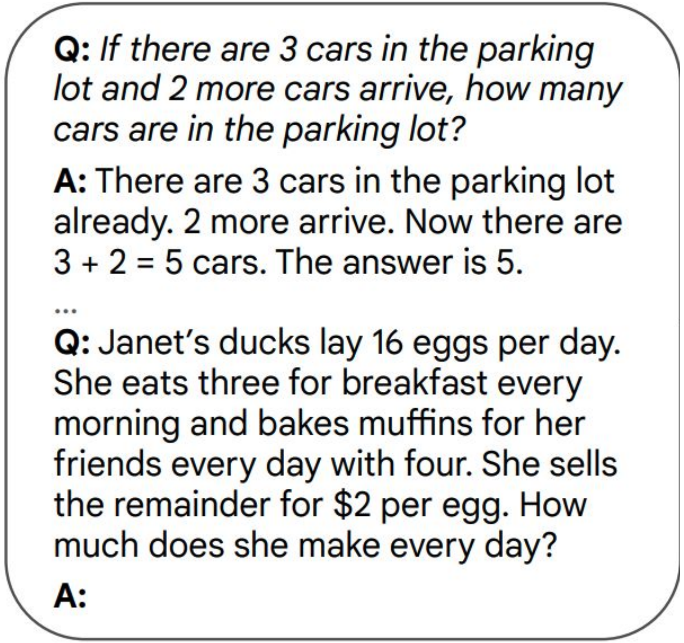
A:

Show example question
with
chain-of-thought response

Follow with desired
question

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Chain-of-Thought Prompting



Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.

...

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

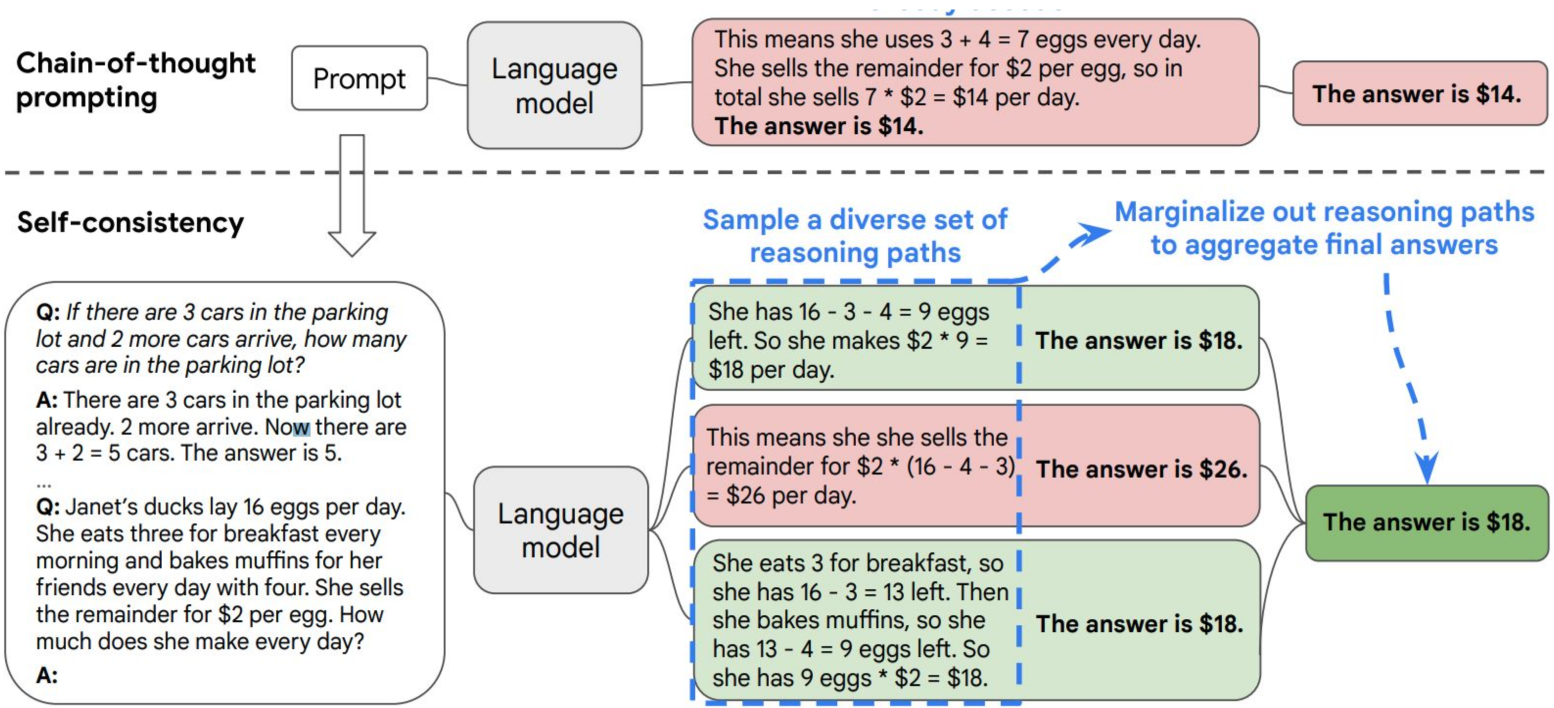
A:

Show example question
with
chain-of-thought response

Follow with desired
question

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Self-Consistency Reasoning



Chain-of-thought prompting

Prompt

Language model

This means she uses 3 + 4 = 7 eggs every day. She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day. **The answer is $14.**

The answer is $14.

Self-consistency

Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.

...

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

A:

Language model

Sample a diverse set of reasoning paths

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. **The answer is $18.**

This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. **The answer is $26.**

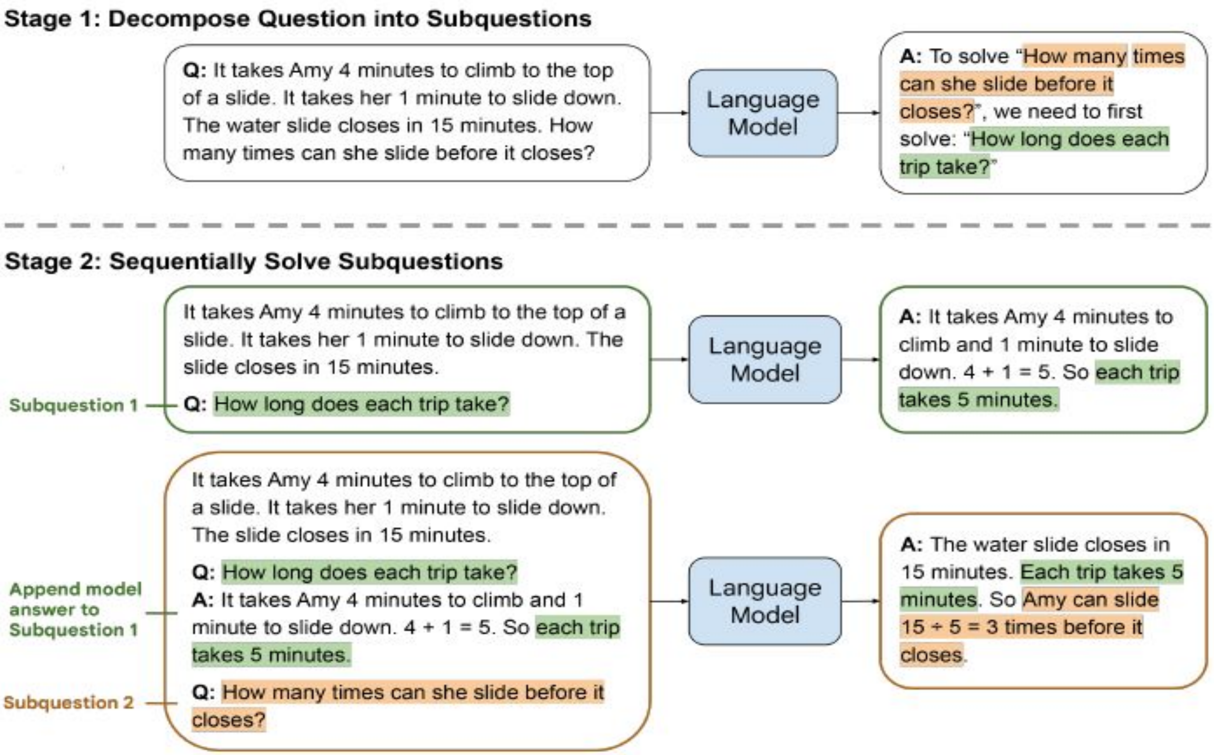
She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. **The answer is $18.**

Marginalize out reasoning paths to aggregate final answers

The answer is $18.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Subquestion Reasoning



Stage 1: Decompose Question into Subquestions

Q: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

A: To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

Stage 2: Sequentially Solve Subquestions

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down. 4 + 1 = 5. So each trip takes 5 minutes.

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

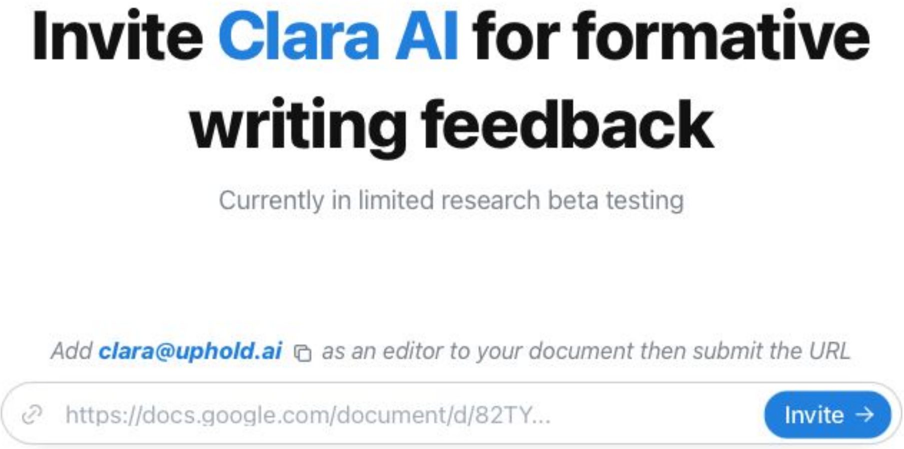
A: It takes Amy 4 minutes to climb and 1 minute to slide down. 4 + 1 = 5. So each trip takes 5 minutes.

Q: How many times can she slide before it closes?

A: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide 15 ÷ 5 = 3 times before it closes.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., ... & Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625.

AI for collaborative writing

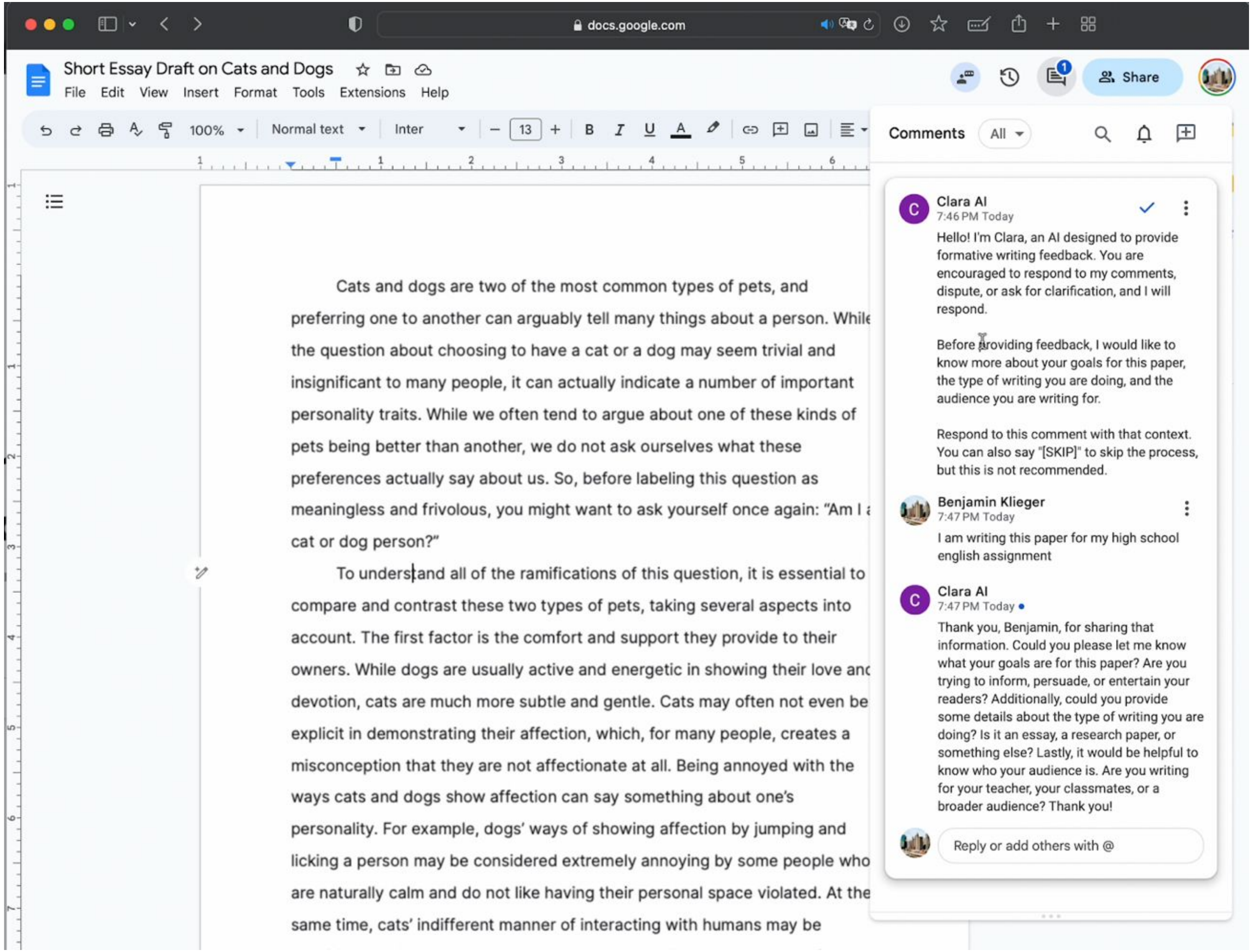


Invite Clara AI for formative writing feedback

Currently in limited research beta testing

*Add **clara@uphold.ai** as an editor to your document then submit the URL*

Ben Klieger



Cats and dogs are two of the most common types of pets, and preferring one to another can arguably tell many things about a person. Whil the question about choosing to have a cat or a dog may seem trivial and insignificant to many people, it can actually indicate a number of important personality traits. While we often tend to argue about one of these kinds of pets being better than another, we do not ask ourselves what these preferences actually say about us. So, before labeling this question as meaningless and frivolous, you might want to ask yourself once again: "Am I cat or dog person?"

To understand all of the ramifications of this question, it is essential to compare and contrast these two types of pets, taking several aspects into account. The first factor is the comfort and support they provide to their owners. While dogs are usually active and energetic in showing their love anc devotion, cats are much more subtle and gentle. Cats may often not even be explicit in demonstrating their affection, which, for many people, creates a misconception that they are not affectionate at all. Being annoyed with the ways cats and dogs show affection can say something about one's personality. For example, dogs' ways of showing affection by jumping and licking a person may be considered extremely annoying by some people who are naturally calm and do not like having their personal space violated. At the same time, cats' indifferent manner of interacting with humans may be

Clara AI
7:46 PM Today

Hello! I'm Clara, an AI designed to provide formative writing feedback. You are encouraged to respond to my comments, dispute, or ask for clarification, and I will respond.

Before providing feedback, I would like to know more about your goals for this paper, the type of writing you are doing, and the audience you are writing for.

Respond to this comment with that context. You can also say "[SKIP]" to skip the process, but this is not recommended.

Benjamin Klieger
7:47 PM Today

I am writing this paper for my high school english assignment

Clara AI
7:47 PM Today

Thank you, Benjamin, for sharing that information. Could you please let me know what your goals are for this paper? Are you trying to inform, persuade, or entertain your readers? Additionally, could you provide some details about the type of writing you are doing? Is it an essay, a research paper, or something else? Lastly, it would be helpful to know who your audience is. Are you writing for your teacher, your classmates, or a broader audience? Thank you!

Reply or add others with @

Enhanced coding error messages with GPT

Compare two approaches to baseline options

- ○ Generate explanatory error messages using GPT in real time
- ○ Link error messages to the course discussion forum

Result

- ○ Students using GPT-generated error messages
  - ■ Repeat an error 23.5% less often in the subsequent attempt
  - ■ Resolve an error in 36.1% fewer additional attempts, compared to standard error messages
- ○ Additional link to forum did not have significant observable benefit

Sierra Wang

Nov 2, 2020

Safely Scaling Coding Style Feedback

Identifier names

- Ask GPT for variable and function name recommendations
- Lots of parsing to ensure the students get back ONLY function and variable names and nothing else

Comments

- Ask GPT to classify existing comments and to classify any missing comments
- For each class, show teacher written feedback
- Only give one piece of feedback for an existing comment and one for needed but not written comment

Constants and Magic Numbers

- Deterministic feedback (no-GPT)
- Magic nums that need constants
- Constants used as variables
- Variables used as constants

Decomposition

- Ask GPT to classify functions into categories
- Give teacher written feedback based on these classifications

Juliette Woodrow

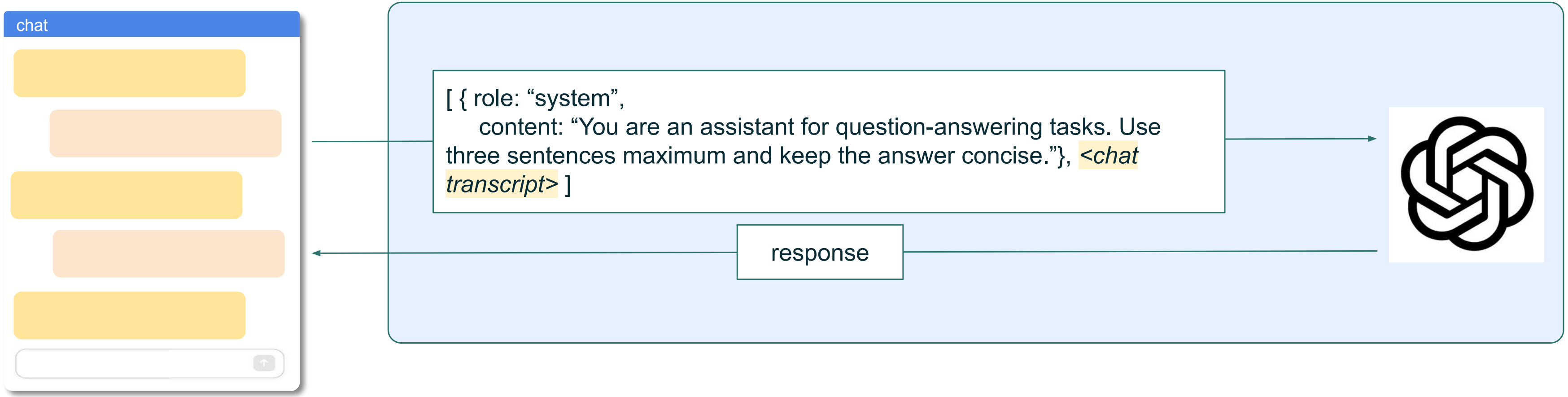
Chris Piech

Retrieval-Augmented Generation (RAG)

Remote Student engages with Web Platform

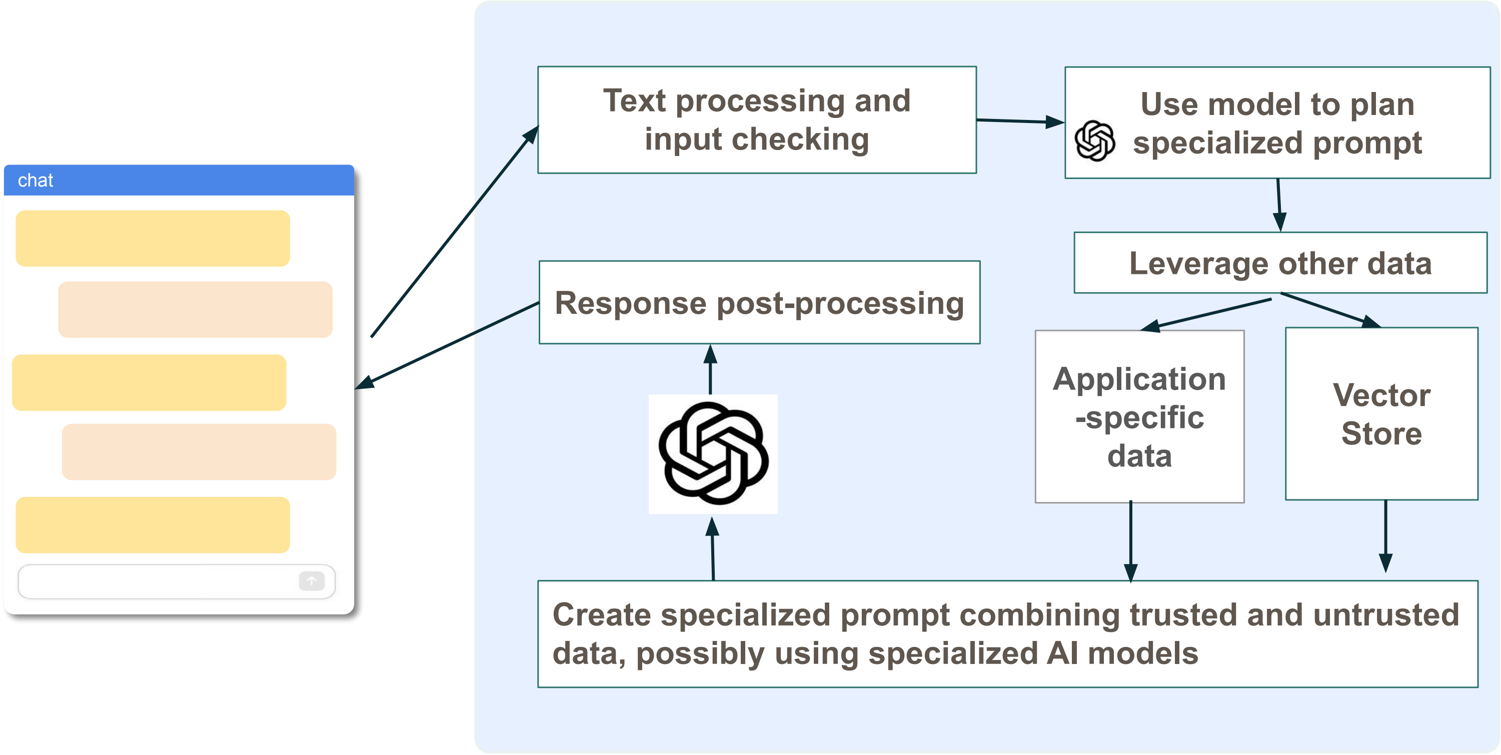
- Many parts are automated using AI, mostly implemented by prompting
  - Teacher training
  - Improved programming environment
    - Error messages – use student program to explain possible cause of error
    - Style feedback – comment on style, based on style rubric provided to LLM
    - Assignment grading – interactive game programs tested automatically
  - Student questions
    - LLM answer based on course textbook, lecture transcript

Naïve Question-Answer System



[{ role: “system”,
content: “You are an assistant for question-answering tasks. Use three sentences maximum and keep the answer concise.”}, *<chat transcript>*]

More Sophisticated RAG-based Approach



Text processing and input checking

Use model to plan specialized prompt

Leverage other data

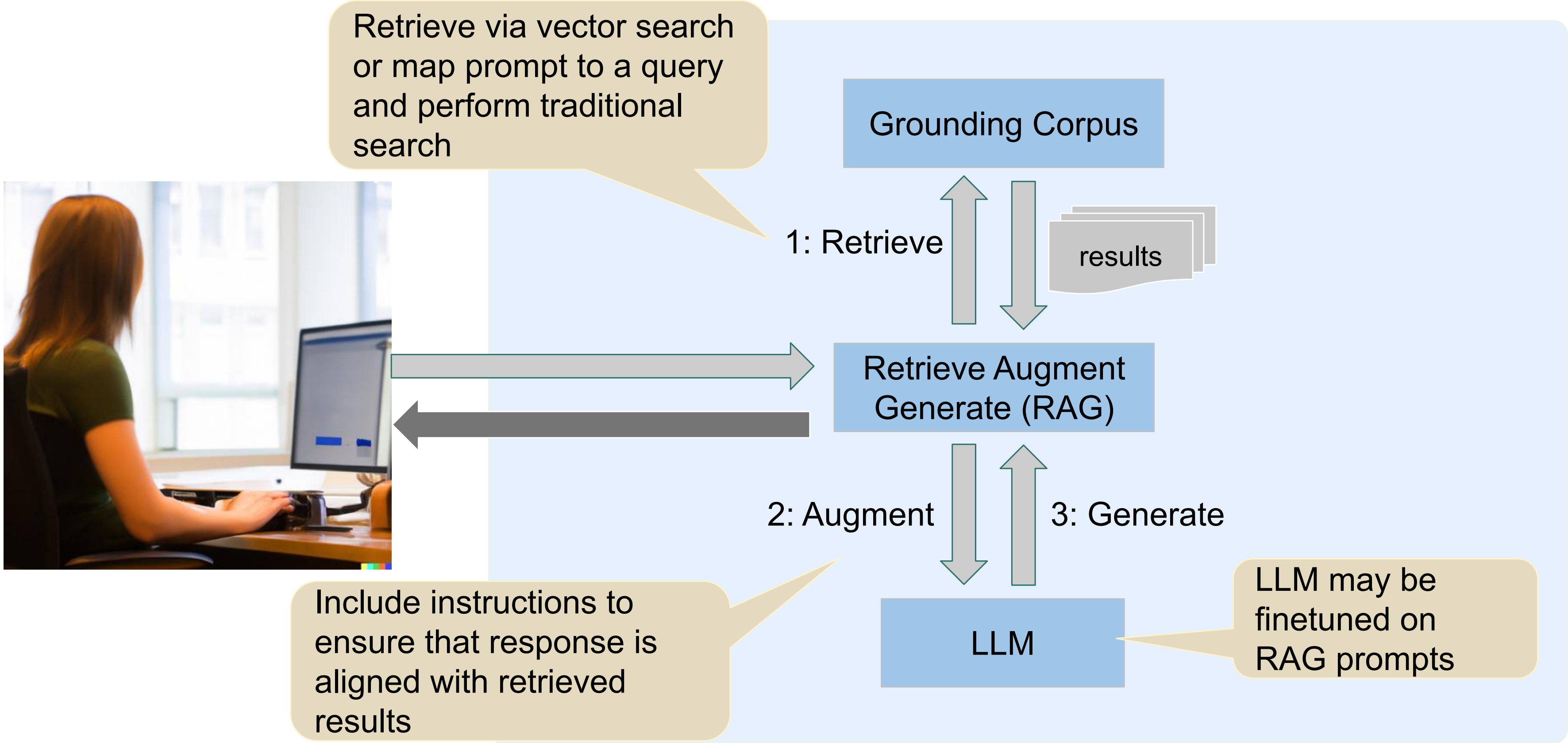
Response post-processing

Application -specific data

Vector Store

Create specialized prompt combining trusted and untrusted data, possibly using specialized AI models

Retrieval-Augmented-Generation (RAG)

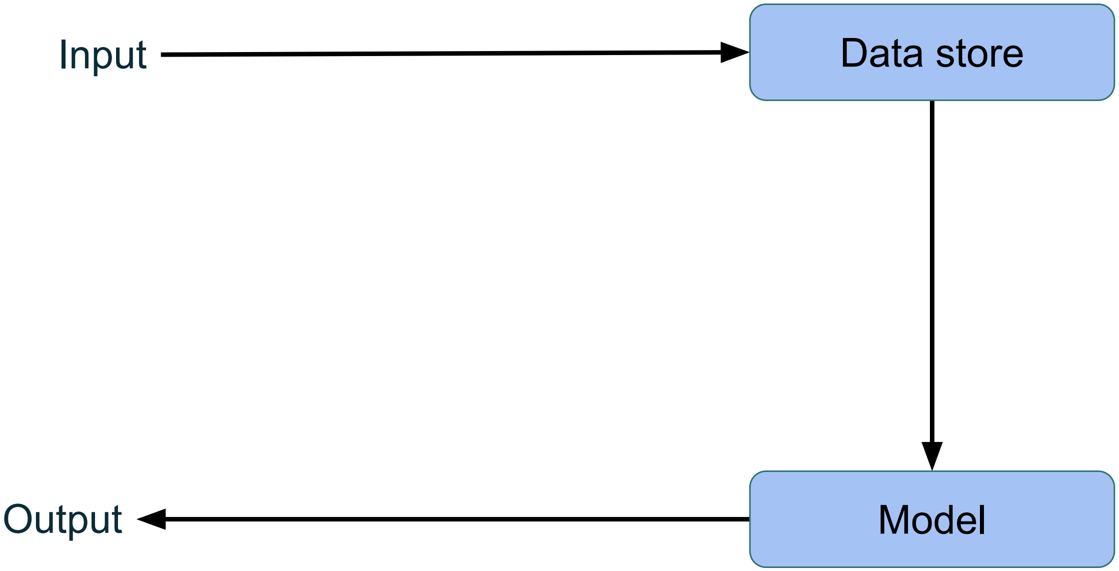


Retrieve via vector search or map prompt to a query and perform traditional search

Include instructions to ensure that response is aligned with retrieved results

LLM may be finetuned on RAG prompts

Functional evaluation: The RAG Triad



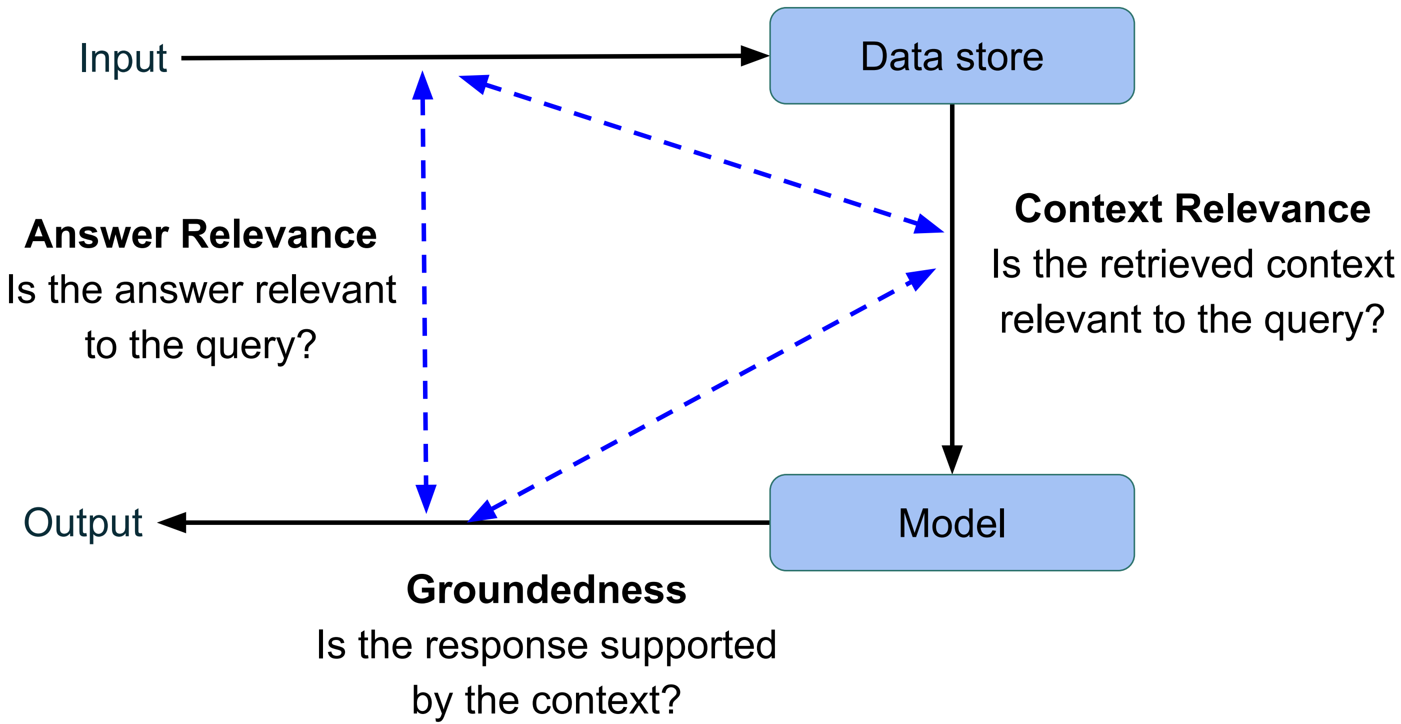
Input

Data store

Model

Output

Functional evaluation: The RAG Triad



Input

Data store

Model

Output

Answer Relevance
Is the answer relevant
to the query?

Context Relevance
Is the retrieved context
relevant to the query?

Groundedness
Is the response supported
by the context?

Groundedness

An LLM response is grounded in a knowledge corpus

if

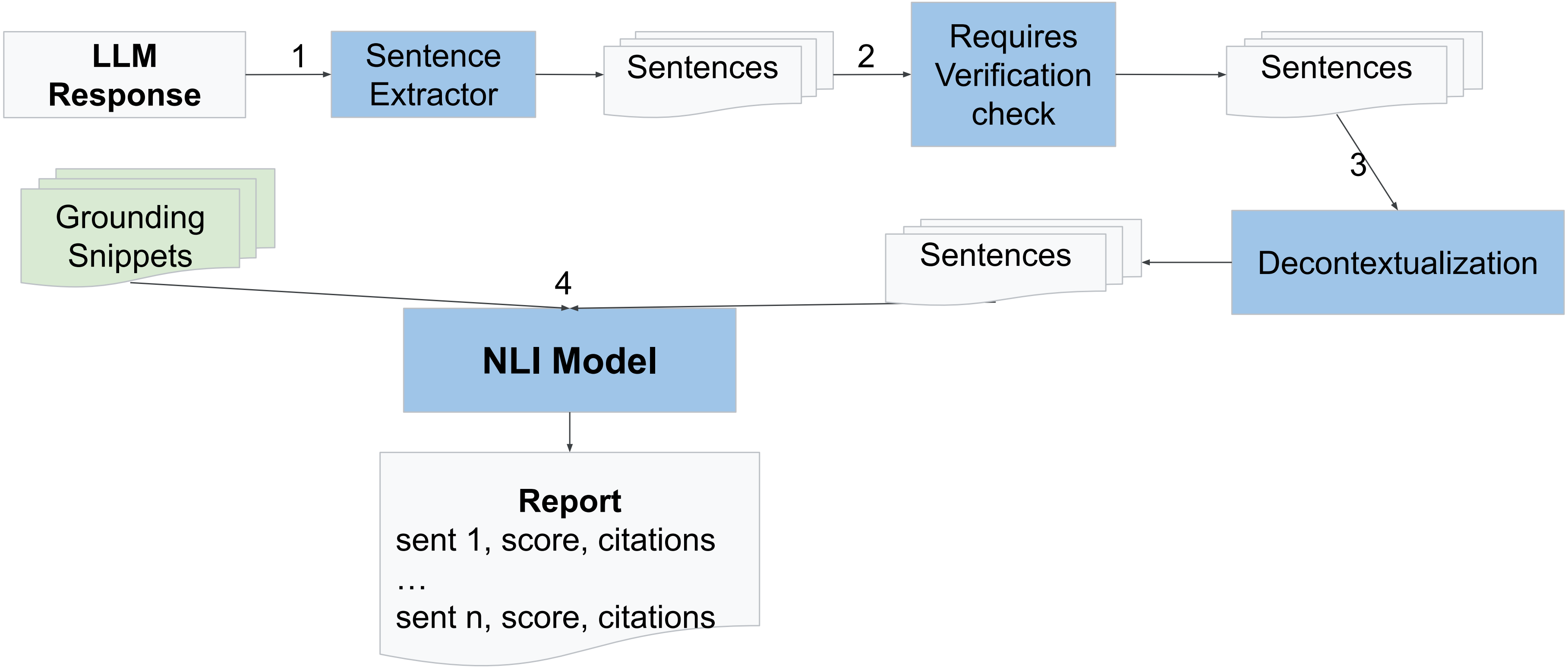
every claim in the response

can be attributed to

some document in the corpus

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., ... & Reitter, D. (2023). Measuring attribution in natural language generation models. Computational Linguistics, 49(4), 777-840.

Corroboration Workflow



RAG: Key Technical Problems

- **When to retrieve external knowledge versus rely on model's parametric knowledge?**
    - Note that retrieval is not always perfect
- **What should be the granularity of retrieval?**
    - Sentence, paragraph, entire page, set of pages?
- **How should we perform retrieval for complex prompts?**
    - E.g., *What is the prevalence of type 1 diabetes in adults over the age of 18 and children under 18 year old, including those already diagnosed and those undiagnosed, in the UK in 2023?*
- **How do we train models to stick to the retrieved context?**
- **What if despite RAG, the model still hallucinates?**

\

Personal Finance Chatbot Using Reddit Data

Victor Li, Riya Dulepet

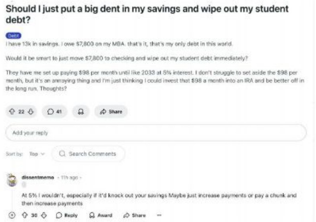
Department of Computer Science

Problem Statement

- Online financial advice is often
    - Not specific enough
    - Unrelated to one's personal situation
    - Requires human response
- We present a chatbot that gives specific, personalized financial advice.

Dataset

- **Reddit Personal Finance Thread (r/personalfinance)**
    - Use Reddit API to fetch all-time top 1000 posts. For each post, the following are gathered:
        - Title
        - Content
        - Url
        - Time Post is Created
        - 5 Comments


Should I just put a big dent in my savings and wipe out my student debt?

I have 13k in savings. I owe $7,800 on my MBA. that's it, that's my only debt in this world.

Would it be smart to just move $7,800 to checking and wipe out my student debt immediately?

They have me set up paying $98 per month until like 2033 at 5% interest. I don't struggle to set aside the $98 per month, but it's an annoying thing and I'm just thinking I could invest that $98 a month into an IRA and be better off in the long run. Thoughts?

dissentmemo · 11h ago ·

At 5% I wouldn't, especially if it'd knock out your savings Maybe just increase payments or pay a chunk and then increase payments

Method

- **Embedding & Retrieval:**
    - Embedded Reddit posts using text-embedding-ada-002 and stored them in a vector database.
    - Selected 20 out-of-dataset Reddit posts as queries and identified the top 3 "best answers" for each as ground truth.
- **RAG Workflow:**
    - Retrieved the top 3 relevant posts for each query using the vector database.
    - Used GPT-4o-mini to synthesize responses based on retrieved posts, guided by a detailed prompt.
        - Response Criteria:
            - Synthesized key advice with **citations** ([1], [2], etc.).
            - Included examples, highlighted pitfalls, and maintained a conversational tone.
            - Ensured all claims were grounded in retrieved posts.

Metrics

- **Citation & Context Analysis** (everything on 0-1 scale)**:**
    - Measured **citation groundedness** using cosine similarity between responses and source Reddit posts
    - **Overall groundedness** (Trulens): how relevant context is to overall response
    - Evaluated **context relevance** (Modified Trulens)): assessed how relevant retrieved comments were to query
    - Analyzed **answer relevanc**e (LLM Judged): assessed how relevant response was to query (emphasis on cited portions)
- **Comparative Evaluation:**
    - Used GPT-4-mini to evaluate our Personal Finance Chatbot against vanilla LLM:
        - Comment Alignment: consistency with Reddit community advice
        - Interpretability: clarity and relevance of examples provided

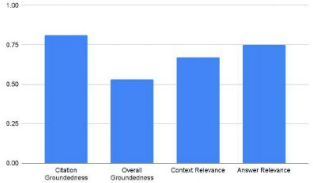
Results

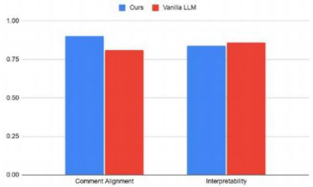
Citation Groundedness	Overall Groundedness	Context Relevance	Answer Relevance
0.81	0.53	0.57	0.75

	Comment Alignment	Interpretability
Ours	0.90	0.84
Vanilla LLM	0.81	0.86

Observations & Analysis

- Strong citation groundedness (0.80) indicates reliable preservation of source information in responses
- Lower context relevance may attributed to the following:
    - Reddit's conversational nature may lead to:
        - Relevant information scattered across comment threads
        - Context dilution from tangential discussions
        - Key information buried in longer exchanges
- Highlights tension between source selection and source utilization in RAG systems
- Our chatbot achieves higher comment alignment (0.90 vs 0.81):
    - Retrieved content helps capture community consensus
    - Real examples from Reddit may provide better context
    - Direct access to community knowledge improves advice
- Baseline slightly outperforms in interpretability (0.86 vs 0.84):
    - Vanilla LLM may produce more streamlined explanations
    - Personal Finance Chatbot needs to incorporate retrieved content, which could affect clarity





Problem Statement

- Online financial advice is often
  - Not specific enough
  - Unrelated to one’s personal situation
  - Requires human response
- We present a chatbot that gives specific, personalized financial advice.

Dataset

- **Reddit Personal Finance Thread (r/personalfinance)**
  - Use Reddit API to fetch all-time top 1000 posts. For each post, the following are gathered:
    - Title
    - Content
    - Url
    - Time Post is Created
    - 5 Comments

Method

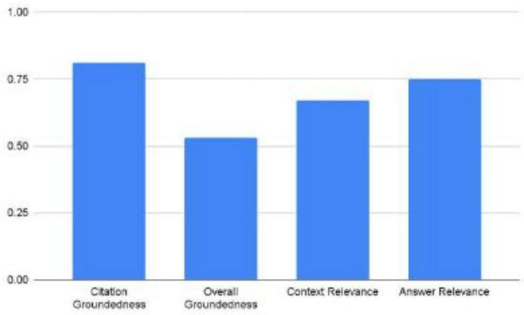
- **Embedding & Retrieval:**
  - Embedded Reddit posts using text-embedding-ada-002 and stored them in a vector database.
  - Selected 20 out-of-dataset Reddit posts as queries and identified the top 3 "best answers" for each as ground truth.
- **RAG Workflow:**
  - Retrieved the top 3 relevant posts for each query using the vector database.
  - Used GPT-4o-mini to synthesize responses based on retrieved posts, guided by a detailed prompt.
    - Response Criteria:
      - Synthesized key advice with **citations** ([1], [2], etc.).
      - Included examples, highlighted pitfalls, and maintained a conversational tone.
      - Ensured all claims were grounded in retrieved posts.

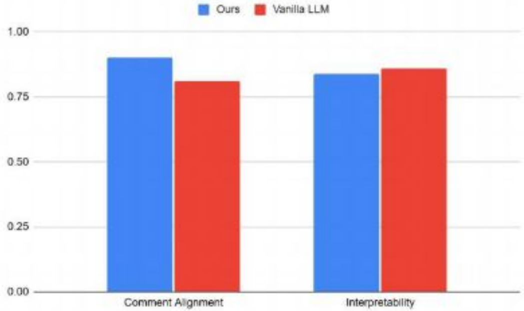
Metrics

- **Citation & Context Analysis** (everything on 0-1 scale)**:**
    - Measured **citation groundedness** using cosine similarity between responses and source Reddit posts
    - **Overall groundedness** (Trulens): how relevant context is to overall response
    - Evaluated **context relevance** (Modified Trulens)): assessed how relevant retrieved comments were to query
    - Analyzed **answer relevanc**e (LLM Judged): assessed how relevant response was to query (emphasis on cited portions)
- **Comparative Evaluation:**
    - Used GPT-4-mini to evaluate our Personal Finance Chatbot against vanilla LLM:
        - Comment Alignment: consistency with Reddit community advice
        - Interpretability: clarity and relevance of examples provided

Observations & Analysis

- Strong citation groundedness (0.80) indicates reliable preservation of source information in responses
- Lower context relevance may attributed to the following:
    - Reddit's conversational nature may lead to:
        - Relevant information scattered across comment threads
        - Context dilution from tangential discussions
        - Key information buried in longer exchanges
- Highlights tension between source selection and source utilization in RAG systems
- Our chatbot achieves higher comment alignment (0.90 vs 0.81):
    - Retrieved content helps capture community consensus
    - Real examples from Reddit may provide better context
    - Direct access to community knowledge improves advice
- Baseline slightly outperforms in interpretability (0.86 vs 0.84):
    - Vanilla LLM may produce more streamlined explanations
    - Personal Finance Chatbot needs to incorporate retrieved content, which could affect clarity





Results

Citation Groundedness	Overall Groundedness	Context Relevance	Answer Relevance
0.81	0.53	0.57	0.75

	Comment Alignment	Interpretability
Ours	0.90	0.84
Vanilla LLM	0.81	0.86

Your initial thoughts on projects

- Open the Google doc using this QR code
- Add one or two project ideas
- Include your name, or
  add your name to a project
  someone else already added