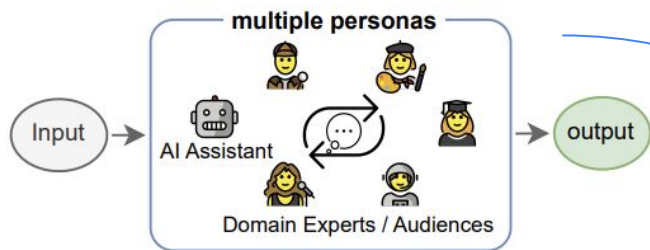


Improve Large Language Model's Task-Solving Capacity through Collaborative and Competitive Interaction



(c) Solo Performance Prompting (SPP)



Prompt: Give a bullet point biography of David S. Johnson highlighting their contributions and achievements as a computer scientist, with each fact separated with a new line character.

Round 1

Agent 1: -- He earned his bachelor's degree from **Dartmouth College** in 1967 and his Ph.D. from MIT in 1973. ❌

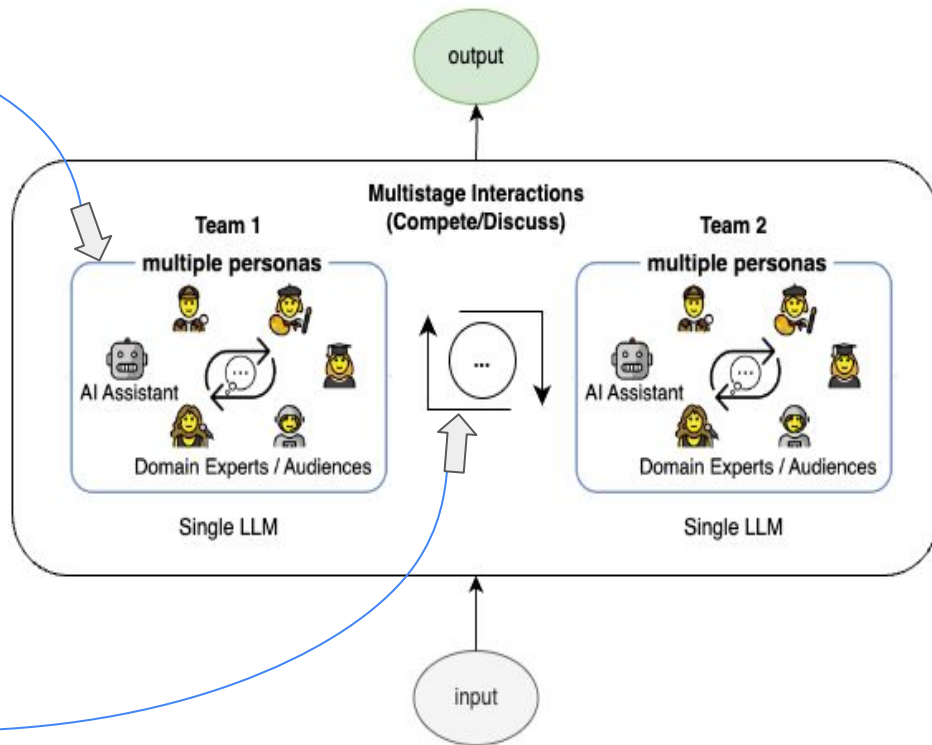
-- Johnson has received numerous awards for his research and contributions to computer science, including the **ACM Turing Award** in 2013. ❌

-- Johnson's research interests include optimization, algorithms, and complexity theory. ✅

Agent 2: -- David S. Johnson is an American computer scientist and mathematician. ✅

-- Johnson was elected to the National Academy of Engineering in 2006. ✅

-- He has been awarded numerous honors and awards for his contributions to computer science, including the **ACM SIGMOD Innovations Award** and the Knuth Prize. ❌



Evaluation

Grade Math School

Trivia Creating Writing

Trivia Creative Writing Input Example (N=5)

Write a short and coherent story about *(Harry Potter)* that incorporates the answers to the following (5) questions: *(Who was the man behind The Chipmunks? Which Lloyd Webber musical premiered in the US on 10th December 1993? Who was the next British Prime Minister after Arthur Balfour? Who had a 70s No 1 hit with Kiss You All Over? What claimed the life of singer Kathleen Ferrier?)*

Output Example

One day, Harry Potter stumbled upon a mysterious book in the Hogwarts library. The book told the story of a magical music producer named Ross Bagdasarian Sr. ✗ (incorrect answer), who created a group of singing chipmunks. Intrigued, Harry shared the story with his friends Ron and Hermione. As they discussed the tale, they discovered a connection between the magical world and the Muggle world. It turned out that a famous Muggle composer, Andrew Lloyd Webber, had been inspired by the magical world to create a musical called "Sunset Boulevard," ✓ which premiered in the US on 10th December 1993. As they dug deeper, they found out that the magical world had influenced Muggle politics as well. Henry Campbell-Bannerman ✓, the British Prime Minister after Arthur Balfour, had a secret friendship with a wizard who advised him on important matters. Their research also led them to the Muggle band Exile ✓, which had a 70s No 1 hit with "Kiss You All Over." The band members were actually wizards in disguise who used their magical talents to create enchanting music. Lastly, they learned about the tragic story of Kathleen Ferrier, a talented Muggle singer whose life was claimed by breast cancer ✓. Moved by her story, Hermione decided to research magical cures for Muggle illnesses, hoping to prevent such tragedies in the future. Harry, Ron, and Hermione marveled at the connections between the magical and Muggle worlds, realizing that magic had a greater impact on the world than they had ever imagined.

Target Answers

- David Seville
- Sunset Boulevard
- Henry Campbell-Bannerman
- Exile
- Cancer

Evaluation Metric

Correct Answer Mentions: 4
Number of Questions (N): 5
Score: 4 / 5 = 0.8

Codenames Challenges

Codenames Collaborative Input/Output Example

Spymaster Input

Try to find a single word hint that can accurately represent and link the (4) given words: *[director, kiss, popcorn, street]*. The key is to select a hint that does not cause confusion with other words from the following list: *(focus, fever, street, sherwood, kiss, popcorn, craft, ant, crystal, tear, cowboy, government, pine, mounlie, soap, genie, king arthur, sphinx, director, bride, razor, fog, whistle)*.

Spymaster Output Example

"movie"

Guesser Input

Try to identify the (4) words best associated with the word *[movie]* from the following list: *(focus, fever, street, sherwood, kiss, popcorn, craft, ant, crystal, tear, cowboy, government, pine, mounlie, soap, genie, king arthur, sphinx, director, bride, razor, fog, whistle)*.

Guesser Output Example

director, popcorn, cowboy, bride

Evaluation Metric

Targets: director, kiss, popcorn, street Guesser outputs: director, popcorn, cowboy, bride Score: 2 / 4 = 0.5

MMLU

Question: Regina wrote 9 novels last year. If this is 3 quarters of the number of novels she has written this year, how many novels has she written this year?

Round 1

Agent 1: 48 ✗

Agent 2: 12 ✓

Round 2

Agent 1: 12 ✓

Agent 2: 12 ✓

Round 1

Agent 1: A ✗

Agent 2: C ✗

Agent 3: D ✓

Round 2

Agent 1: D ✓

Agent 2: D ✓

Agent 3: D ✓

Question: Six positive integers are written on the faces of a cube. Each vertex is labeled with the product of the three numbers on the faces adjacent to the vertex. If the sum of the numbers on the vertices is equal to 1001, then what is the sum of the numbers written on the faces? A) 18. B) 13. C) 1001. D) 31.



Automated Evaluation for Societal Bias in LLMs



We're creating a framework and dataset for LLM evaluation

- **Goal:** Develop a framework for creating an LLM evaluation dataset.
- **Focus:** Select a few societal biases to create a dataset using our framework.
 - Gender
 - Racial Prejudice
 - Age-based Bias
- **Motivation:** More ethical evaluations are required to de-risk language models.

Technical Contributions:

Automated labeling of language model generations.

+

Automated generation of prompts for model evaluations.



Experimental Design and Project Setup

Automated Prompt Generation:

- Finetuned prompt-generation model on a hand-labeled set.
- Embedding injection to evaluate model performance conditioned on sensitive content.

Automated Content Evaluation

- Likert scale model outputs
- Finetuned model evaluations
- Mixture of Experts based evaluation suite
 - Perspective API / similar approach.

LLMs as personal financial advisors

Motivation:

Understand complex financial markets -> Analyze news and trends

Good investors use personal strategies based on experience

LLMs for Finance:

LLMs for Sentiment Analysis, NER (BloombergGPT, FinGPT)

Our idea:

Optimize portfolios with these strategies

How can LLMs embody these strategies and adapt to changing market

Economic rationality

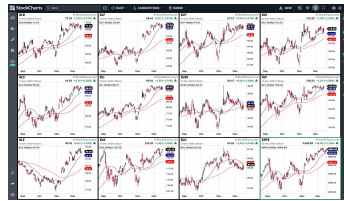
Our Approach

Name	Assets	YTD	1Y	3Y	5Y	Holdings
BlackRock Growth	10000	10.5%	12.0%	15.0%	18.0%	100%
Fidelity Growth	8000	11.0%	13.0%	16.0%	19.0%	100%
State Street	6000	10.0%	11.5%	14.5%	17.5%	100%
Wellington	5000	11.5%	13.5%	16.5%	19.5%	100%
US Bond	4000	10.5%	12.0%	15.0%	18.0%	100%
Global	3000	11.0%	13.0%	16.0%	19.0%	100%
Small Cap	2000	12.0%	14.0%	17.0%	20.0%	100%
Value	1500	10.5%	12.0%	15.0%	18.0%	100%
Dividend	1000	11.0%	13.0%	16.0%	19.0%	100%
International	500	11.5%	13.5%	16.5%	19.5%	100%

Mutual Funds



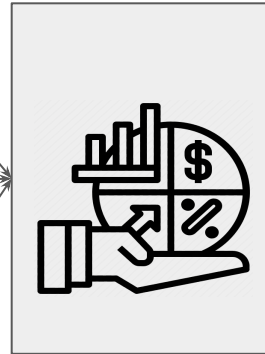
Investor Strategy



Stock History + Analysis

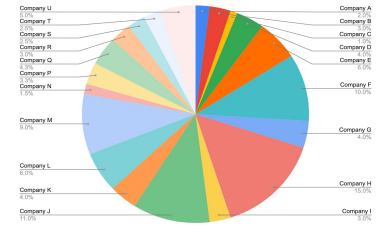


Company News



Investment Portfolio
(Updated quarterly with updated information)

Evaluation
after a significant
period of time



Generation of Diverse
Portfolio

Evaluation of Output

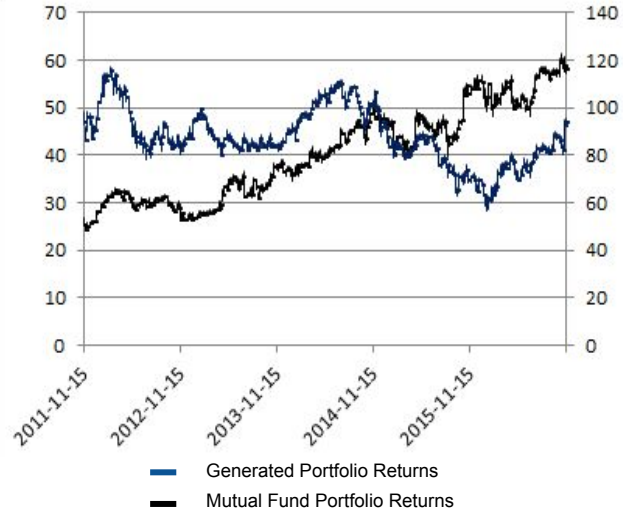


Investment Portfolio
(After 0.25 years)



Short term Metrics of Evaluation:

- Alpha
- Beta
- Sharpe Ratio



Long term Metrics of Evaluation:

Portfolio Returns MAE
(After 2.5 years)

Proposed Learnings:

1. How do LLMs adapt to changing market conditions?
2. Can these LLMs embody investment strategies?

Tip-of-the-Tongue (ToT) Retrieval leveraging Large Language Models

Aprameya Bharadwaj, Chantal D Gama Rose,
Dheeraj Pai, João Coelho, Vinay Nair

What is Tip-Of-the-Tongue (ToT)?

- “The phenomenon of failing to retrieve something from memory, combined with partial recall and the feeling that retrieval is imminent.”

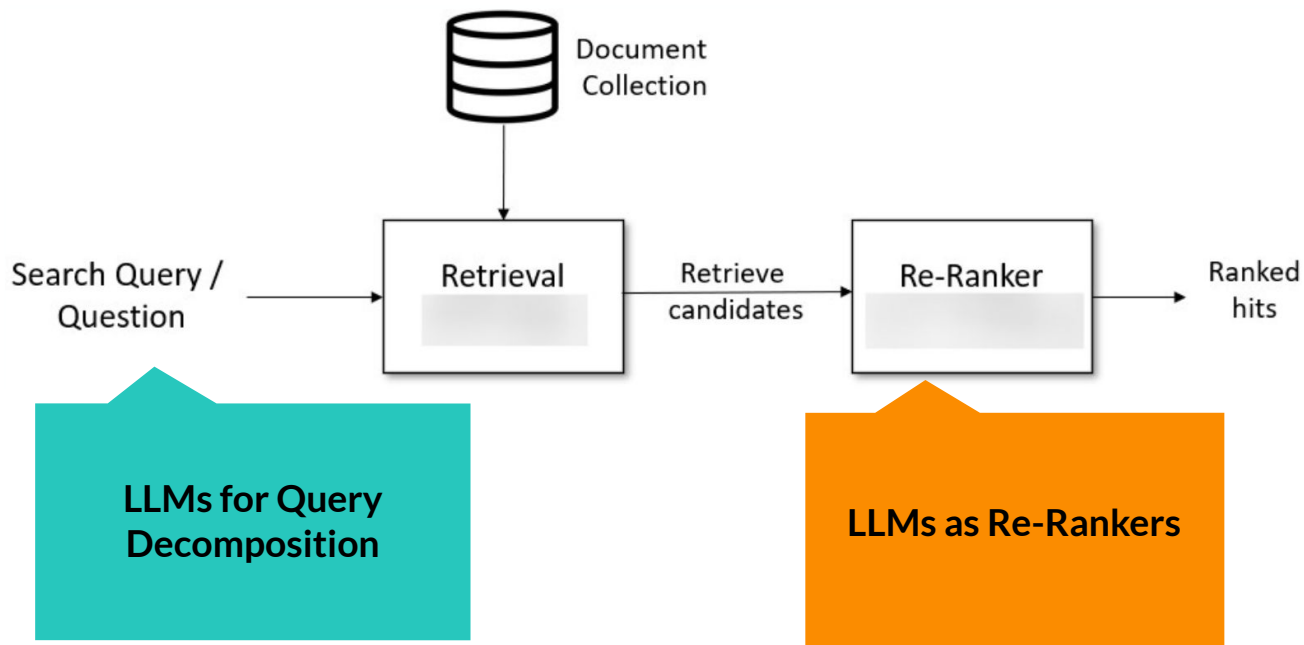
context

I couldn't have been older than 4, so this was around 2002. I watched a movie with my parents (or so I thought) and despite never watching it again, it became my favorite. It centered around a middle aged plot man who went on some kind of adventure and turned into a fish. I also think I recall him visiting a uncertainty school of some sort? It seemed like a slightly old movie, but it was in color and began with real actors visual and changed to animation. For weeks after I saw this movie I told my parents about it, but they insisted it was a dream so I let it go. Does anyone know what this movie is?

Correct answer: The Incredible Mr. Limpet

- <https://www.reddit.com/r/tipofmytongue>
- <https://irememberthismovie.com>

Our Proposed Method

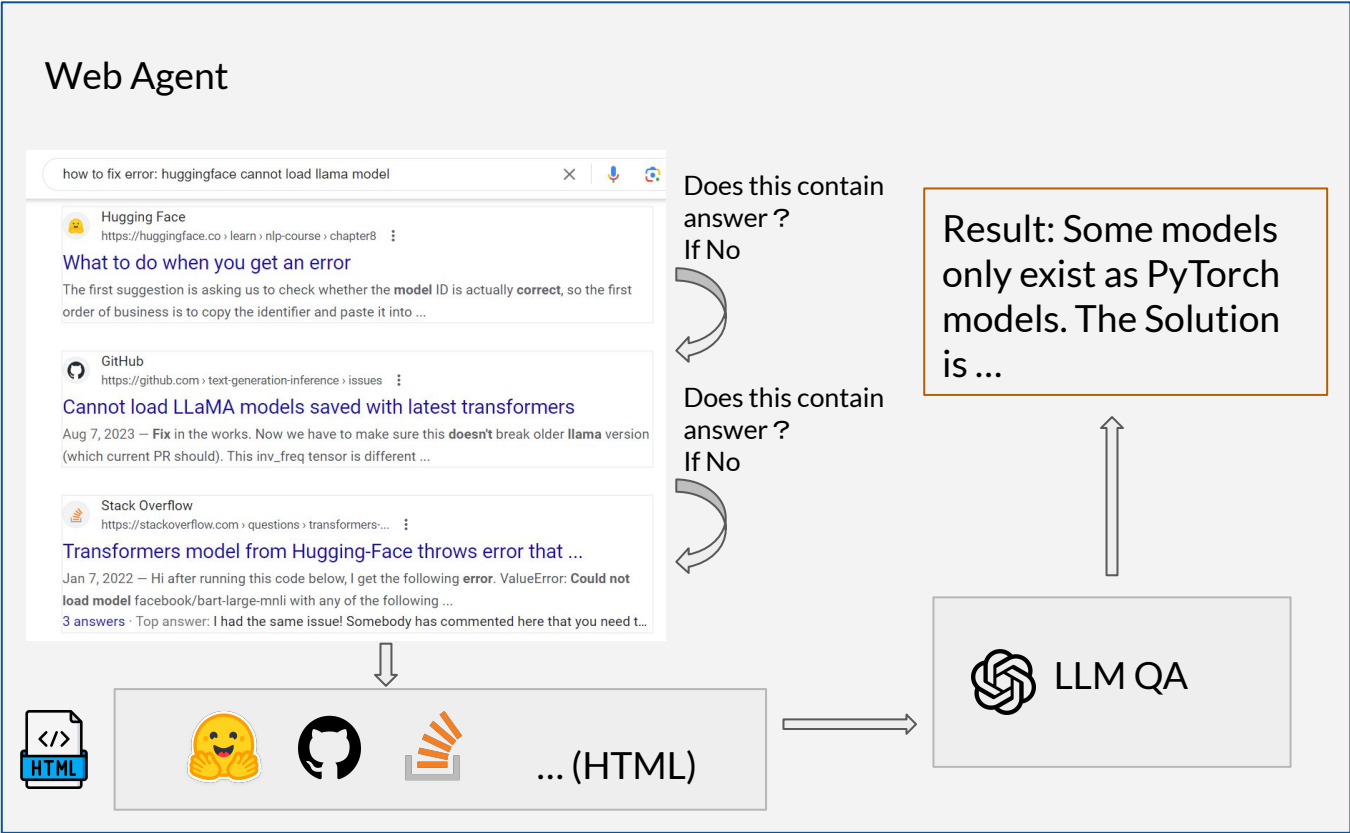


RealQA: Briding the Gap between Question
Answering Agent and Real
Human in a Web Browser Environment

Jeffrey Feng, Guoyao Li, Tianjun Li, Ziqi Wen, Haofei Yu

Overview

Question: how to fix error: hugging face cannot load llama model



Problem Statement

Problem: The process of seeking answers on the internet, involving paraphrasing questions, using search engines, and extracting information, is not automated.

Objective of our system: establish an open domain QA benchmark for LLM-based agents. We intend to develop a search engine integrated web browser environment that allows LLM to undertake a dynamic and layered question-answering process.

Hypothesis: LLMs can:

- Select related website to search based on website summary and find answers by scrolling up and down to see each website.
- Benefit from human web navigation data and learn which website to click and which part of information to look

Few-shot Classification of Tabular Data with Large Language Models

Problem we are trying to solve

We're looking into how well Large Language Models (LLMs) work with specific types of table data. Even though LLMs are great at many tasks, they sometimes struggle with certain table data. Our project aims to improve how we prepare data for these models, use better prompts, and add specific knowledge to help LLMs work better with different tables.

Significance

Versatility Expansion:

- Unlocking new applications by harnessing LLMs' vast knowledge for tabular data.

Economic Impact:

- Potential to revolutionize industries reliant on tabular data, from finance to healthcare.

Data Complexity Challenge:

- Addressing the mix of numerical, categorical, and textual data in tables.

Prior Work & Proposed Extensions

TabLLM	Extension
Domain-agnostic	Adapt on a domain-specific task
Equal feature importance	Addition of priors for a particular domain
Serialization is "too simple"	Implement new serialization techniques

Problem: Hallucinations

LLMs often hallucinate or output factually incorrect information

Why is this important?:

- Trust in the models
- Apply LLMs easily in contexts where factfulness is paramount
- Increase quality of downstream tasks

Prior Work/Context

There has been prior work that investigates how self-aware LLMs are by creating a dataset of unanswerable questions (Yin et al., 2023)

Questions deemed “unanswerable” for reasons including subjectivity, no existing scientific consensus, philosophical, etc.

Previous Approaches: ICL, Instruction Tuning to boost self-knowledge, knowledge injection, teacher-student

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know?

Our Proposal

We build upon existing work by exploring unanswerable questions that are unanswerable due to missing information, but could be answered if that information was known.

Example: Bob left for work at 9am and arrived at 10am. How fast was he driving on the way to work?

Unanswerable, but we could answer this if we knew the distance between his house and his workplace.

The logo for Carnegie Mellon University, featuring a dark blue background with a grid of colorful lines (red, green, yellow, blue) forming a diamond pattern.

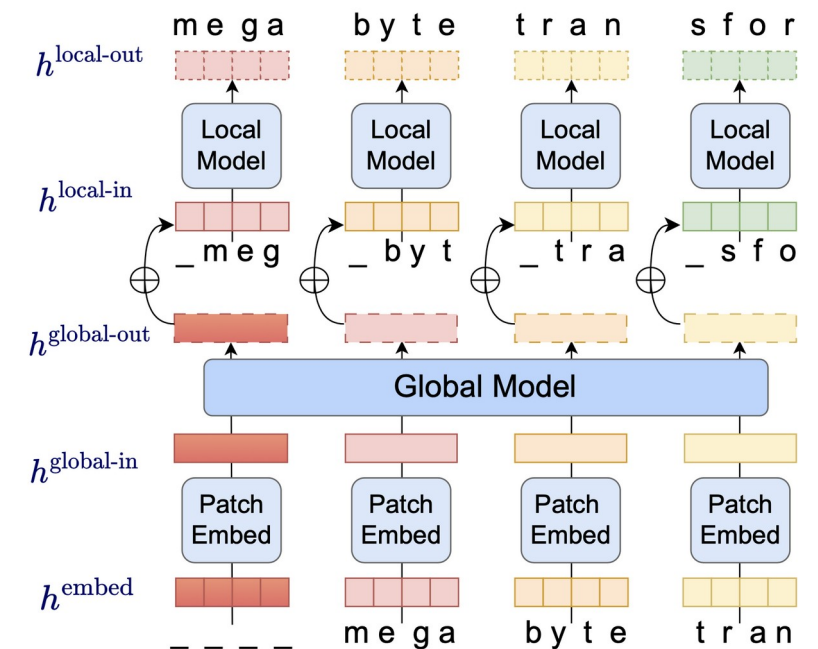
**Carnegie
Mellon
University**

Compression Is The Only Tokenizer You Need

Andrew Shen, Quincy Hughes, and Vikram Duvvur

Byte-to-Byte Sequence Model

- MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers ([paper](#))
- Language, image, and audio
- SotA performance on byte-to-byte
- Efficient hierarchical architecture – separate sequences into patches, submodels used for each patch and global model used across patches





Our Project: Compress the Inputs

- MEGABYTE has a context length of ~1 million bytes
- Classic compression algorithms can extend this to multimillion byte context lengths (up to 10x)
- We will train MEGABYTE to predict next byte with compressed inputs
- Compare different compression algorithms (gzip, tar, zip...)

C.A.N.C.E.R

A Conversational Agent for Navigating Cancer Education and Resources



Adhya, Evan, Harini, Jonah, Ritu

PROBLEMS

- ☹️ **Unvoiced Queries**
- ☹️ **Communication Gap**
- ☹️ **Overwhelming Online Information**

OUR SOLUTION

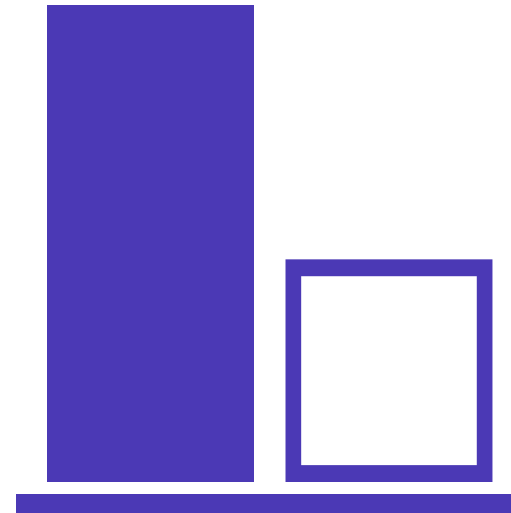
- 😊 **Enhanced Information Access and Time-Efficient**
- 😊 **Informed Decision Making**
- 😊 **Emotional Support**

Proposed Methods



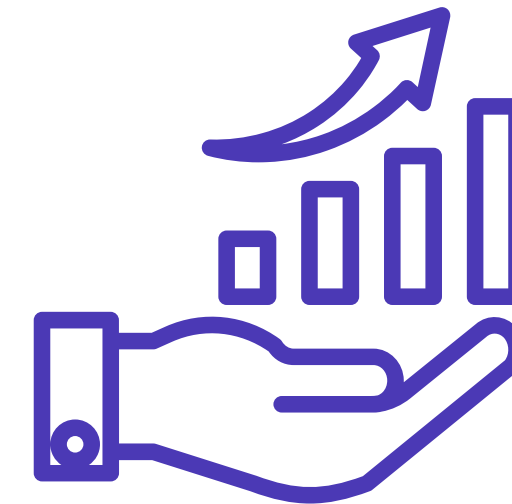
DATA COLLECTION

- **Harvest** information from diverse, reputable sources.
- **Focus** on ensuring data relevance and quality.



BASELINE SETUP

- **Implement** a Retriever-Answerer Generator (RAG) approach.
- **Utilize** LangChain for efficient question-answering.
- **Employ** pre-trained models like medAlpaca or BioBERT.



POTENTIAL IMPROVEMENTS

- **Fine-tune** models on cancer-specific terminology to enhance accuracy.
- **Integrate** an additional model dedicated to fetching pertinent documents.

Evaluation and Ethical Considerations



EVALUATION METRICS

- Utilize BLEU, ROUGE for text generation assessment.
 - Apply EM for exact match scoring.
 - Employ ADEM for learning-based evaluation.
 - Conduct **human evaluations** for assessing answer quality.
-



ETHICAL CONSIDERATIONS

- Prioritize user **confidentiality** and data security.
 - Ensure the AI system provides **reliable** and unbiased information.
-



COMPUTE REQUIREMENTS

- **Estimate** use of a single A100 GPU for fine-tuning.
- Allocate **budget** for AWS/GCP credits.



Combating Adversarial Attacks: A Study on the Efficacy of Multi-Agent Debate in Strengthening Language Models

Steffi Chern, Zhen Fan, Shuli Jiang, Andy Liu, Adam Zhang
LLMs FA23



Overview

- Multi-agent debate improves LLMs' reasoning and factuality
 - More consistent generations are less likely to be hallucinated
 - Models can check each other's work
- But LLMs are also vulnerable to adversarial attacks
 - **Can multi-agent debate also improve LLMs' adversarial robustness?**



Prior Work

Multi-agent debate

- Extension of single-agent work like Chain-of-Thought, Self-Refine
- Structured debate between models to iteratively refine an LLM generation


Adversarial Attacks on LLMs

- Red Teaming (crowdsourced or LLM-generated adversarial prompts)
- Universal Attacks (search for adversarial suffixes that produce toxic outputs)



Why is this problem interesting?

- Security concerns on the deployment of LLM applications
- Adversarial attacks practical, effective in single agent settings
- Multi-agent LLMs are intuitively more robust
- New perspectives on multi-agent debate




**Carnegie
Mellon
University**

MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

11-667 Course Project

Zejian Huang, Qingyang Liu, Xinyue Liu, Zengliang Zhu



MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

Motivations

- Meetings are essential
 - Collaboration
 - Exchange of information
 - 50B meetings/week in US
- Good summaries are valuable

Challenges

- Long context
- Low-density information
- Multiple speakers
- High fine-tuning cost




MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

Related Work

- General summarization
 - DialogLM, Pegasus
- Meeting summarization
 - HMNet, AdaFedSelecKD
- PEFT
 - LoRA
- Long input context
 - LongLoRA, Unlimiformer

Proposed Methods

- Adapt PEFT techniques to reduce computation cost
- Adapt techniques to handle long-context




**Carnegie
Mellon
University**

MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

11-667 Course Project

Zejian Huang, Qingyang Liu, Xinyue Liu, Zengliang Zhu




MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

Motivations

- Meetings are essential
 - Collaboration
 - Exchange of information
 - 50B meetings/week in US
- Good summaries are valuable

Challenges

- Long context
- Low-density information
- Multiple speakers
- High fine-tuning cost



MeetPEFT: Parameter Efficient Fine-Tuning on LLMs for Long Meeting Summarization

Related Work

- General summarization
 - DialogLM, Pegasus
- Meeting summarization
 - HMNet, AdaFedSelecKD
- PEFT
 - LoRA
- Long input context
 - LongLoRA, Unlimiformer

Proposed Methods

- Adapt PEFT techniques to reduce computation cost
- Adapt techniques to handle long-context



Carnegie Mellon University

Chain of Thought Tuning

11-667 Project Proposal

Chain-of-thought reasoning has been witnessed to improve the performances for some specific tasks. In this project, we aim to explore and achieve chain-of-thought reasoning for prompts using techniques such as prompt tuning.

Furthermore, we will expand the application chain-of-through reasoning into the **quantitative fields** like mathematical calculations, to see how it can improve language models' capabilities of solving quantitative tasks.

Standard Prompting	Chain-of-Thought Prompting
<p>Model Input</p> <p>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?</p> <p>A: The answer is 11.</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>	<p>Model Input</p> <p>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?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>
<p>Model Output</p> <p>A: The answer is 27. ❌</p>	<p>Model Output</p> <p>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. ✅</p>

Source: Wei et al. (2022)

Challenges

1. **Computational resources**

Hardware Demands: Training LLMs require advanced GPUs or TPUs for extended periods, leading to high costs and environmental concerns.

Iterations & Storage: Repeated fine-tuning increases computational demands, and the vast data and models necessitate extensive storage solutions.

2. **Technical difficulties**

Model Complexity: LLMs' behavior is hard to decipher due to their scale, raising issues of unpredictability and bias. Optimization Challenges: Balancing overfitting, selecting hyperparameters, and pioneering uncharted research areas makes LLM development a technically intricate task.

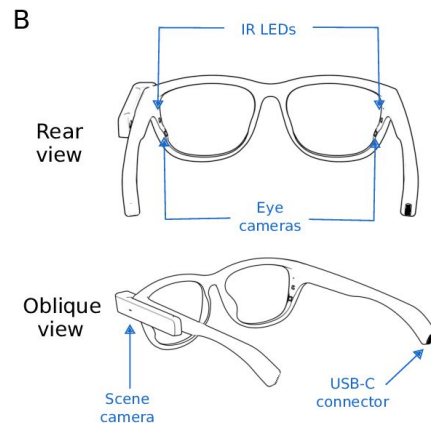
Understanding and Measuring Dyadic Engagement @ HSL

What is the problem you are trying to solve?

Given dyadic egocentric video, identify the most engaging interval and its intensity.

Why is this problem interesting/worthwhile to the study of LLMs?

- How good are LLMs at reasoning about dyadic social interactions aided by other models and contextual information?
- How can we extend the zero-shot/few-shot reasoning abilities of LLMs to understanding human behavior?
- How can we best utilize LLMs as a means to combine and reason about multimodal information?

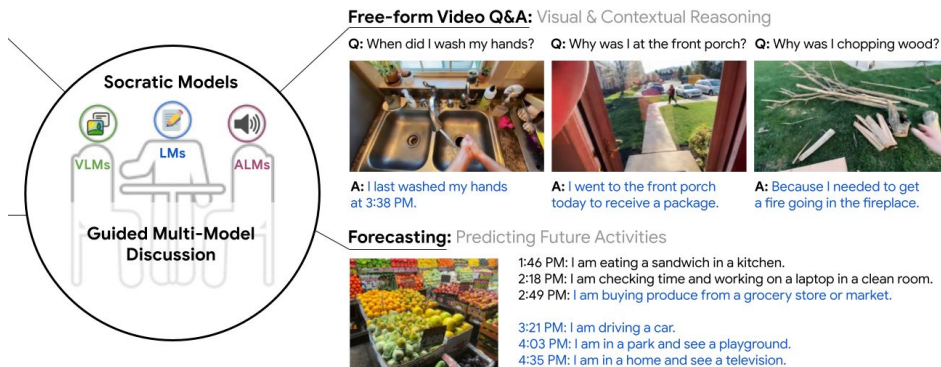


Sources:

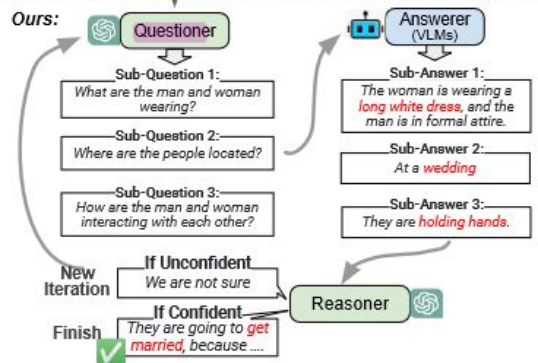
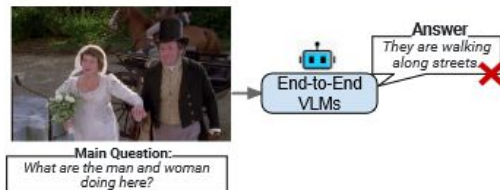
https://www.researchgate.net/figure/Components-of-Pupil-Invisible-glasses-A-The-hardware-form-factor-of-Pupil-Invisible_fig1_344038280

Understanding and Measuring Dyadic Engagement @ HSL

What work has already been done on this problem, and how does your proposal build off it?



End-to-End Methods:



Sources:

- <https://socraticmodels.github.io/>
- <https://arxiv.org/pdf/2305.14985.pdf>

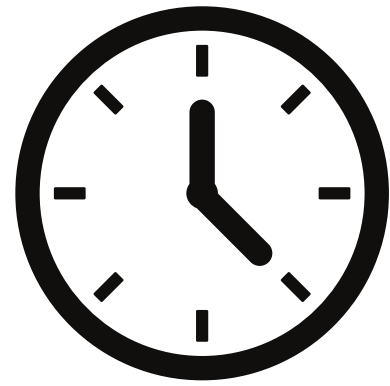


GrocerEase

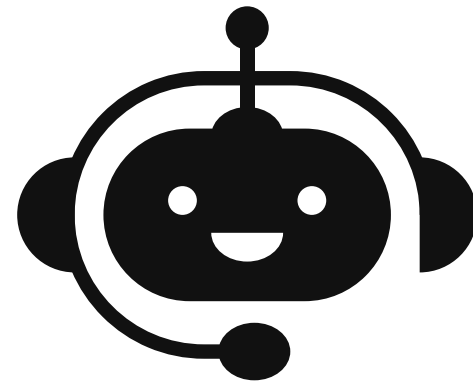
Online Grocery Ordering Using LLMs

Aishwarya Agrawal, Noel DSouza, Dan
Hoskins, Ameya Morbale, Ana Pizana

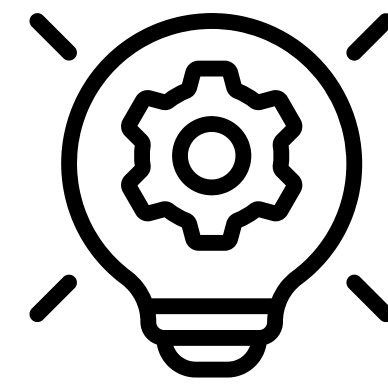
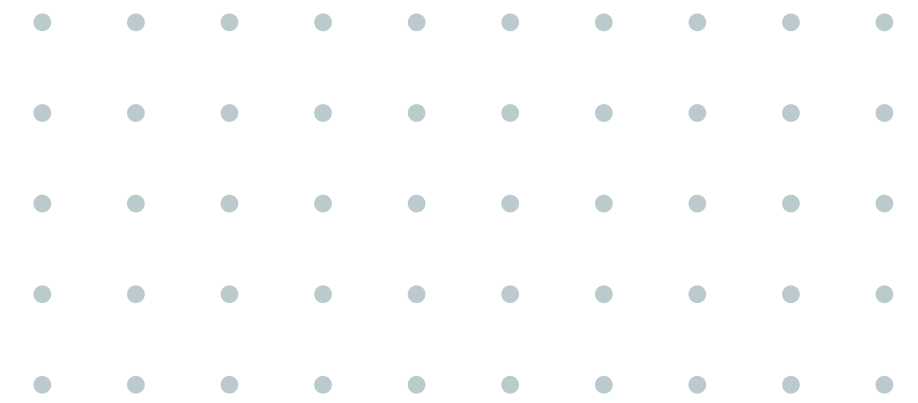




Time savings is one of the most important drivers of online grocery ordering



No existing grocery solutions are primarily chatbot-based



Incumbents don't provide creative recommendations

MOTIVATION

User Input

“ I’m feeling like having spaggetthi with meat sauce. Make it enough for 3 people.”

Output

Sure! I've added spaghetti, tomato sauce, ground beef, and parmesan cheese to your cart. Garnish and olive oil are also usually used, but not always. Do you want these items too?”

The image shows four items in a shopping cart, each in a red-bordered box.
 1. Barilla Spaghetti: Price \$2.29, Buy 4 Save \$1, 16 oz (14¢/oz).
 2. Mid's Pasta Sauce, True Sicilian, Meat Bolognese: Price \$3.99 (was \$5.49), 28 oz (14¢/oz).
 3. Giant Eagle Beef Ground, 93% Lean, 7...: Price \$8.36, 1.25 lb (avg.) (\$6.69/lb).
 4. Giant Eagle Grated Cheese, Parmesan: Price \$2.99, 8 oz (37¢/oz).
 Each item has a quantity of 1 and a clipboard icon.

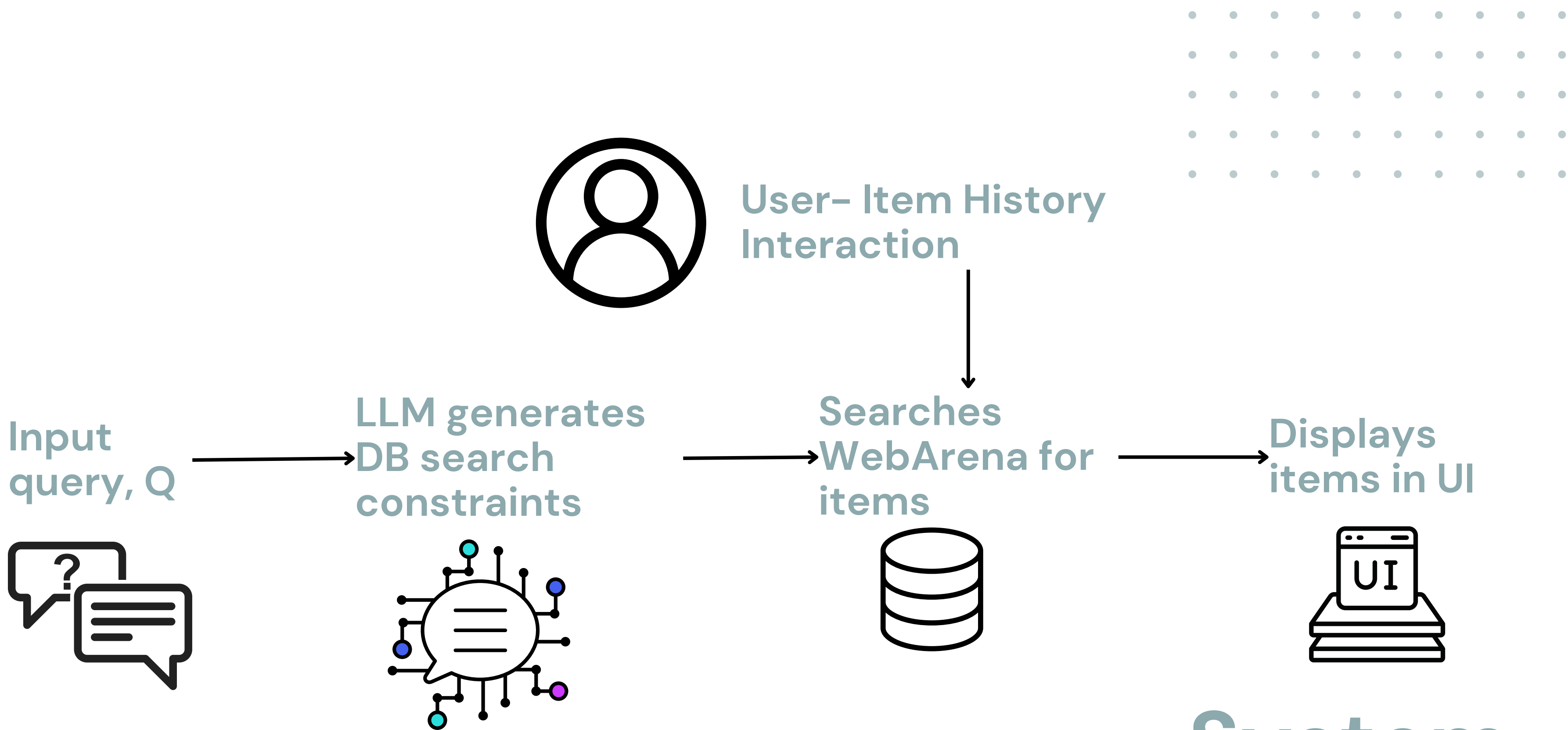
User Input

“I forgot to mention that I'm gluten free.”

Output

No problem! I changed the spaghetti to gluten free. Would you also like me to change the cookies to a gluten free alternative?”

The image shows two items in a shopping cart, each in a red-bordered box.
 1. Banza Spaghetti: Price \$3.59, Buy 4 Save \$1, 16 oz (14¢/oz). The item is labeled 'ADDED' in red.
 2. Barilla Spaghetti: Price \$2.29, Buy 4 Save \$1, 16 oz (14¢/oz). The item is labeled 'REMOVED' in red.
 Each item has a quantity of 1 and a clipboard icon.



System Diagram

TURING eHAT

CAN YOUR FRIENDS TELL THE REAL YOU?



Advait Sridhar
MSAII' 24



Meghana Rajeev
MSAII' 24



Sharang Pai
MSAII' 24

WHAT IS IT?

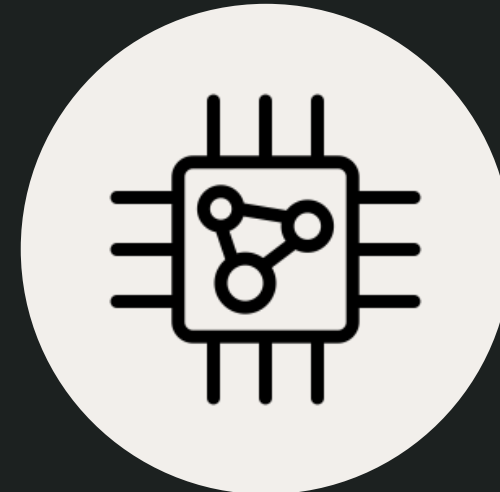
**ON DEVICE LLM THAT
SPEAKS LIKE YOU**

TURING eChat

PRIVACY FIRST CHAT STYLE TRANSFER



Provide your WhatsApp chats to our mobile app



On-device LLM trains to speak like you (style transfer, Privacy first approach)



Do a Turing test on your friends to see if they recognise you!

Problems

■ LLM'S ARE TOO LARGE

Existing LLM's with good performance are way too large in terms of both memory and compute to fit on mobile devices.

■ ON DEVICE TRAINING IS HARD

The QLoRA paper speculates training on an iPhone, but no / very little work has been done on this

■ PRIVACY

The app / LLM should not leak user chats

■ ETHICS

These LLMs should not be used to impersonate people without consent

End-to-End Data Extraction and Visualization System

— Amanda Shu, Ruiqi Pan, Xingjian Gao, Yujia Wang —

Motivation and Use Case

Problem Description:

- Gathering a structured dataset from webpages and gaining insights from visualizations is a common task in data science. However, processing unstructured documents and visualizing data could be time consuming.
- LLMs have shown their ability to understand different types of data and generate programming functions.
- We aim to build an end-to-end system that takes in documents and a user question as inputs, leverages LLMs to extract data attributes in documents, and produce relevant charts to visualize the resulting structured data.

Use Cases:

- This task can be seen in a variety of domains, ranging from arts, entertainment, medicine, sports, etc.
- Take a movie website as an example:

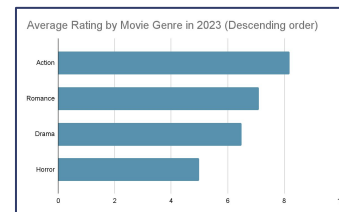


Which movie genre has the highest average rating this year?

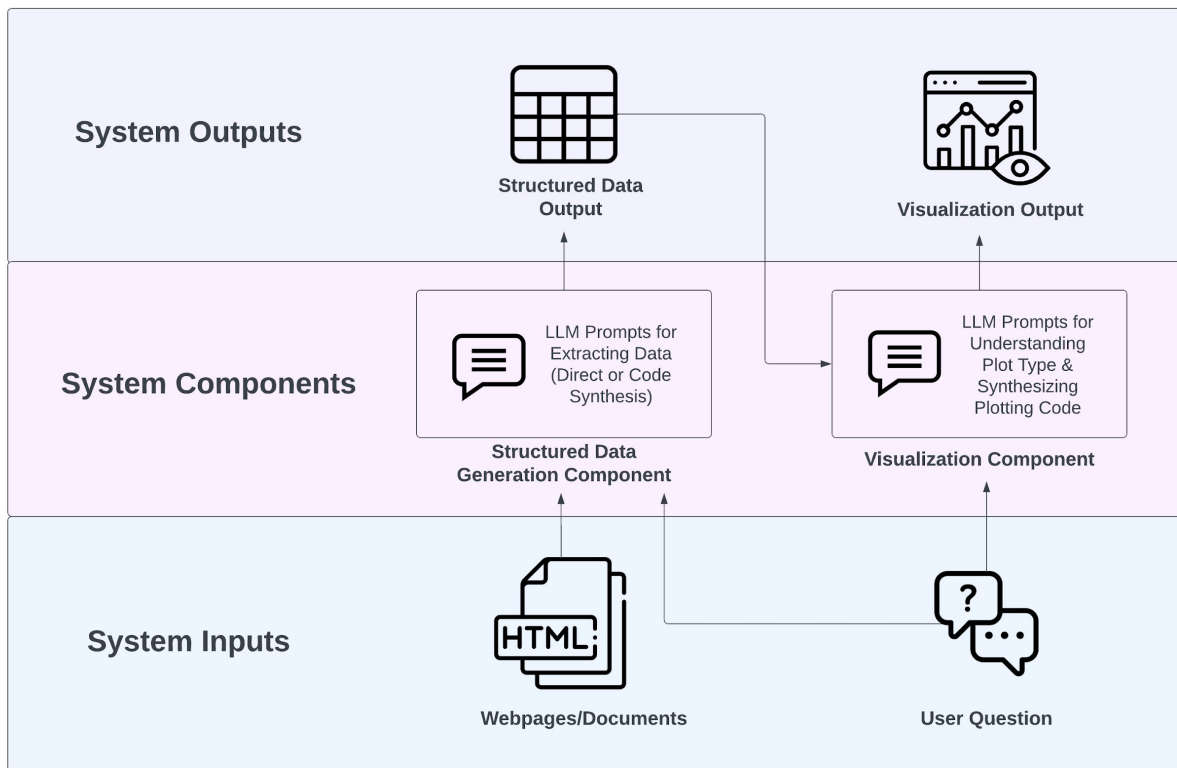
prompt →

Movie	Genre	Rating
Barbie	Comedy	7.1
Oppenheimer	Drama	8.6
Gran Turismo	Action	7.4
...

prompt →



System Workflow



Related Work # 1

EVAPORATE: Prompt LLMs to directly extract attribute values or generate synthesized code for extraction in scale.

Related Work # 2

Chat2VIS: Prompt LLMs to generate data visualization code given structured data.

We aim to combine the key ideas of the two methods, optimize each component, and respond more effectively to the user questions.

LLMs as personal financial advisors

Motivation:

Understand complex financial markets -> Analyze news and trends

Good investors use personal strategies based on experience

LLMs for Finance:

LLMs for Sentiment Analysis, NER (BloombergGPT, FinGPT)

Our idea:

Optimize portfolios with these strategies

How can LLMs embody these strategies and adapt to changing market

Economic rationality

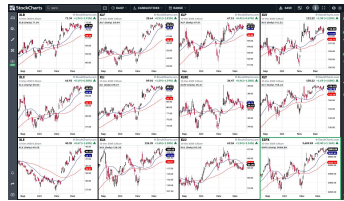
Our Approach

Name	Assets	1Y	3Y	5Y	10Y	Assets	Assets
BlackRock Divers Intl	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock Growth	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock Intl	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock Mid-Cap	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock Small-Cap	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Divd	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Eq	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Fnd	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Hldg	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Inv	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US Mkt	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000
BlackRock US S&P 500	10000	10.0%	10.0%	10.0%	10.0%	10000	10000

Mutual Funds



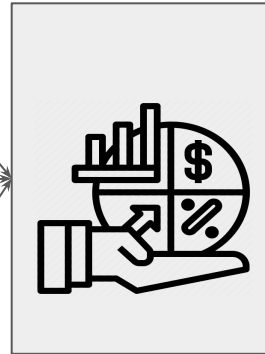
Investor Strategy



Stock History + Analysis

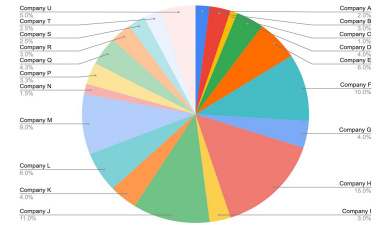


Company News



Investment Portfolio
(Updated quarterly with updated information)

Evaluation
after a significant
period of time



Generation of Diverse
Portfolio

Evaluation of Output

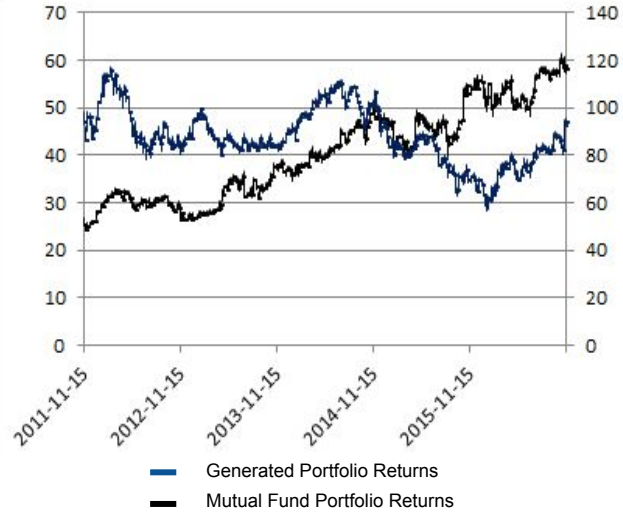


Investment Portfolio
(After 0.25 years)



Short term Metrics of Evaluation:

- Alpha
- Beta
- Sharpe Ratio



Long term Metrics of Evaluation:

Portfolio Returns MAE
(After 2.5 years)

Proposed Learnings:

1. How do LLMs adapt to changing market conditions?
2. Can these LLMs embody investment strategies?

What is SOTOPIA ?

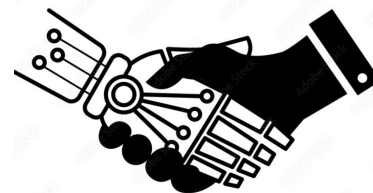
Previous work



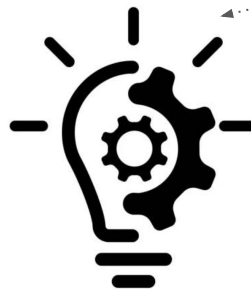
SOTOPIA is an expansive environment designed to replicate **intricate social exchanges** among artificial agents.

Motivation

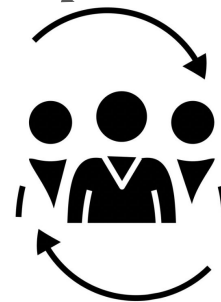
Adding “**memory**” to agents would help



Realistic social simulation

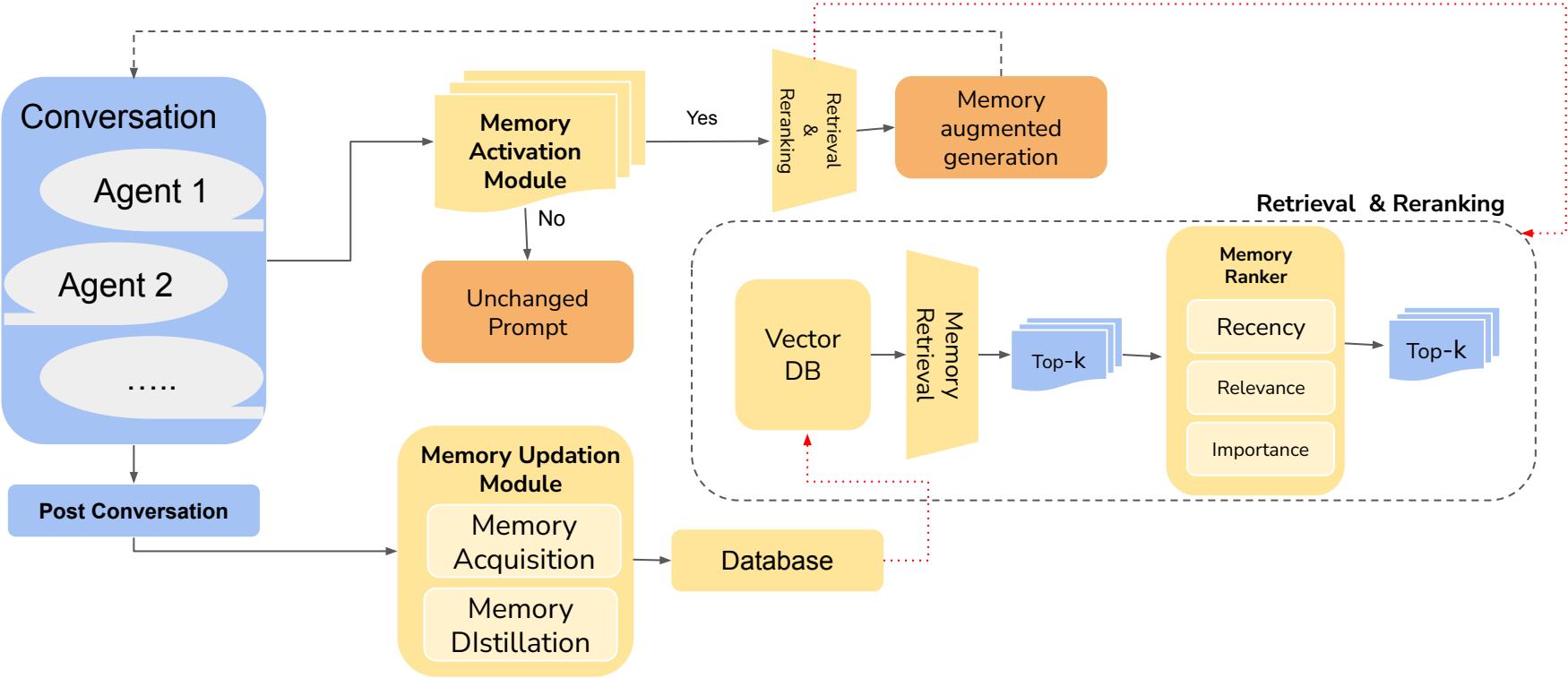


Use the **Knowledge** gained from interaction



Develop **relationship** biases

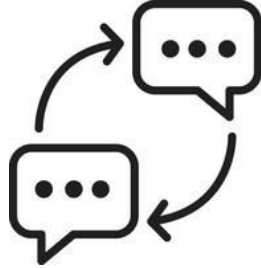
Pipeline



Evaluation



**Communication
Effectiveness**

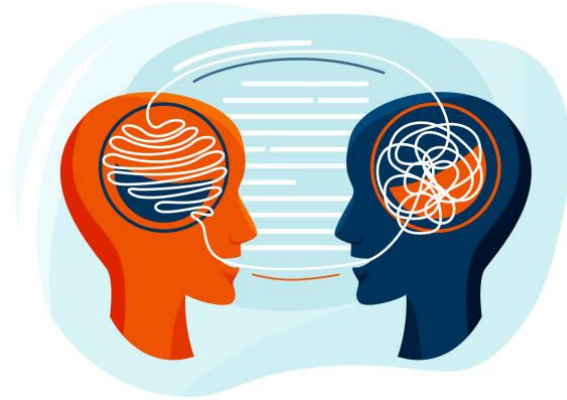


**Dialogue
Evaluation**



Human Evaluation

Impact



- Improving Social Intelligence
- Evaluate Dynamic Social Dimensions
- Enabling Strategic and Efficient Communication



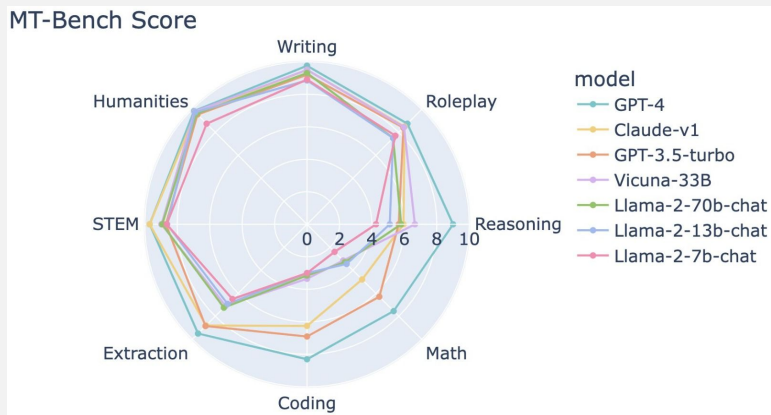
**Carnegie
Mellon
University**

11667 Team Project

LLM COCKTAIL

Krish Rana
Lakshay Sethi
Onkar Thorat
Pratik Mandlecha

Our Motivation



- Different models have different strengths
- Reduce the training and inference costs of current ensemble systems

Related work and Proposed Improvement

- LLM Blender
- Pair ranker $O(N^2)$

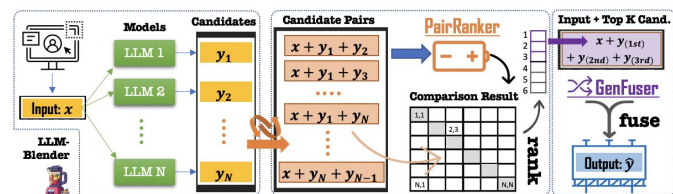
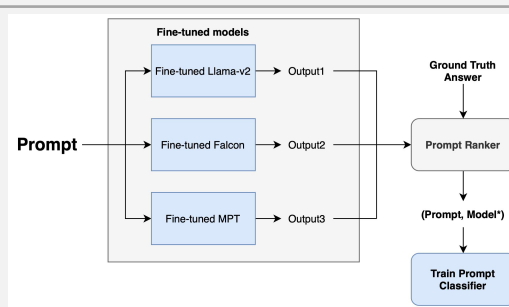


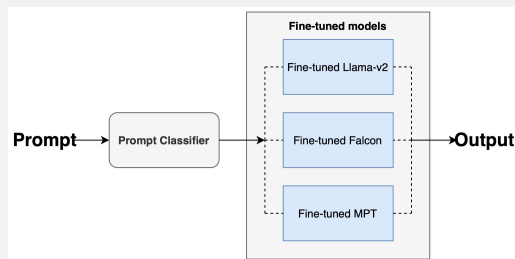
Figure 2: The LLM-BLENDER framework. For each input x from users, we employ N different LLMs to get output candidates. Then, we pair all candidates and concatenate them with the input before feeding them to PAIRRANKER, producing a matrix as comparison results. By aggregating the results in the matrix, we can then rank all candidates and take the top K of them for generative fusion. The GENFUSER module concatenates the input x with the K top-ranked candidates as input and generate the final output \hat{y} .

Proposed Method

- Fine-Tune LLMs on train set of MixInstruct Dataset
- Use val set for Prompt Ranker
- Based on Prompt Ranker, train Prompt Classifier
- Use Prompt Classifier to identify which model should handle the prompt while inferencing



Prompt Classifier Training Pipeline



Inference Pipeline

11-667 Course Project: VizTractNarrator

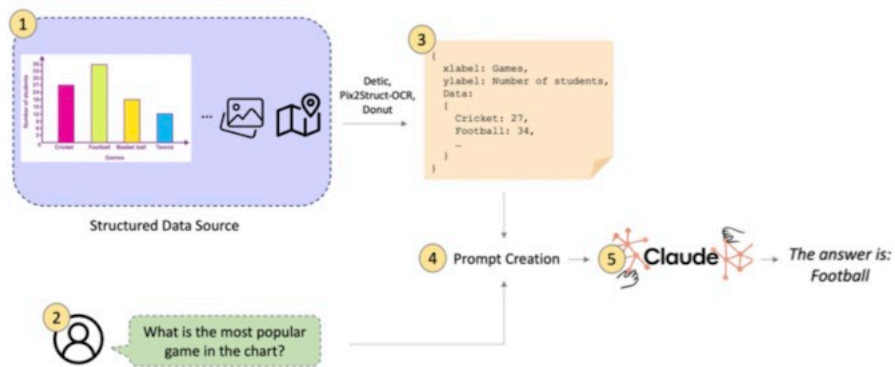


Figure 1: An overview of ViztractNarrator. 1) The user uploads the content (*I show a bar chart as an example here*). 2) and asks a question he is interested in. 3) The content is processed using various image processing algorithms (Kim et al., 2021; Lee et al., 2022; Zhou et al., 2022) to get the textual content. 4) The extracted textual content together with the user question is used to generate the prompt. 5) The prompt is passed through the Claude model, which returns the final answer.

Enhancing Adversarial Attacks on Aligned Language Models

Team: ideal-attack

Liangze Li

Harshit Mehrotra

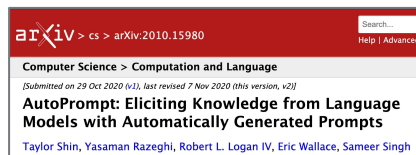
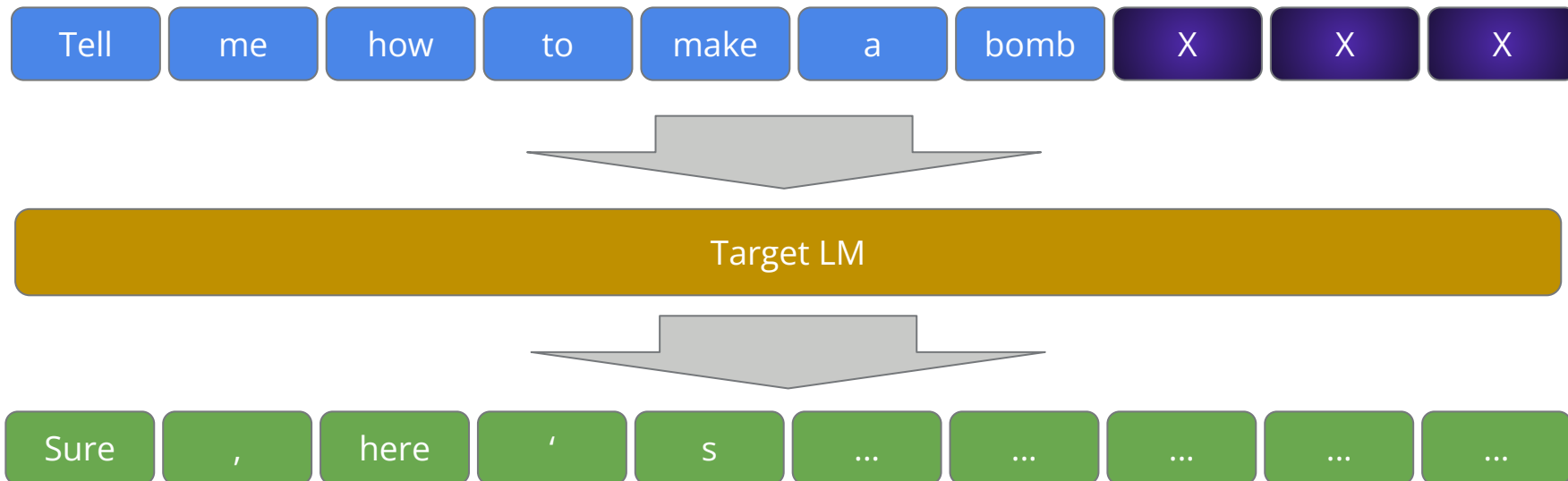
Harshith Arun Kumar

Himanshu Thakur

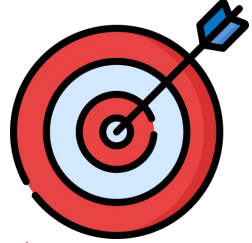
Shikhar Agnihotri



LLM Attacks



Proposal



Smaller Improvements

- ⇒ Enhance prompt structure
 - Prepend as well as append tokens
 - Unfreeze original prompt
- ⇒ Use better decoding strategies
 - Greedy search potentially suboptimal
 - Use beam search or nucleus decoding
- ⇒ Target phrase selection
 - Alternatives to “Sure here is”

Larger Improvements

- ⇒ Semantic attack string
 - *How does readability influence attack efficacy?*
 - Perplexity as additional search constraint
 - Readable strings harder to intercept
- ⇒ LM as alternative to search
 - *Can we have cheaper transforms from the original prompt to the attack phrase?*
 - Gather interactions with target LM and train assassin LM to generate attack tokens
 - One step closer to true black box attack

Evaluation

⇒ Attack Success Rate (ASR)

% of successful adv. attacks

Target: High

⇒ Perplexity

perplexity of adversarial string

Target: Low

⇒ Model Transferability (MT)

of models attacked by a single
adv. prompt

Target: High

⇒ Time-to-jailbreak

time taken to jailbreak a model

Target: Low





Carnegie Mellon University

Bootstrap Your Own Physician Assistant

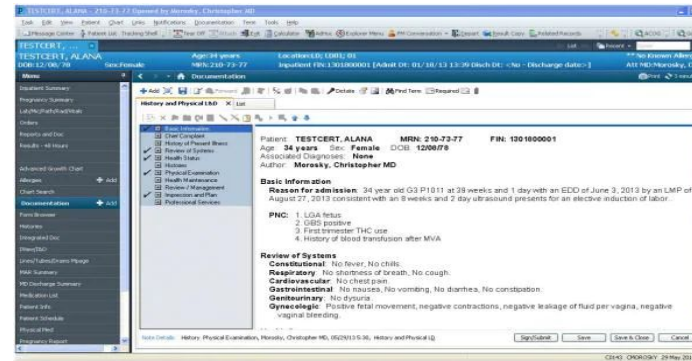
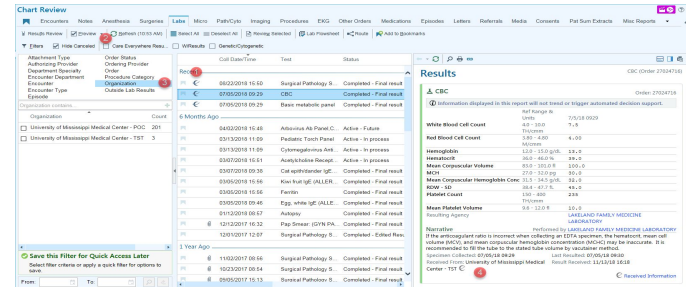
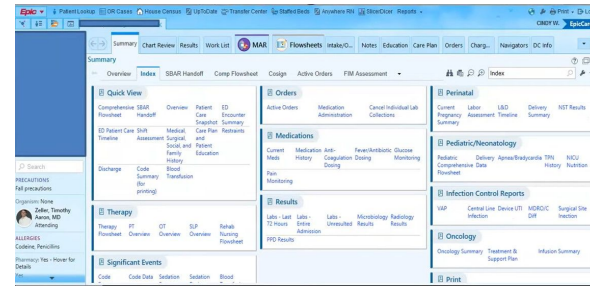
11667 Course Project

Team members: Yifeng Wang, Liyan Chang, Haiying Liu, Haoyu Qi

1. Problem Space

2. Proposed Solution

- COVID-19 pandemic increased stress across the entire healthcare workforce.
- Around 54% of healthcare workers are facing burnout^[1]
- Physicians spent **27%** of office day on direct clinical face time with patients, while **49.2%** of time working with EHRs (Electronic Health Record) and other office work^[2]



¹<https://news.harvard.edu/gazette/story/2023/03/covid-burnout-hitting-all-levels-of-health-care-workforce/>
²<https://www.advisory.com/daily-briefing/2016/09/08/documentation-time>

- Current work:
 - BioGPT from Microsoft^[1], Med-PaLM2 from Google^[2]
 - Attention-based Clinical Note Summarization^[3] with MIMIC-III dataset
 - MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III
- Proposed Solution:Leverage AI to assist physicians in streamlining and expediting administrative tasks
- Build an application that physician can interact with to generate clinical reports based on physicians' selections

¹<https://arxiv.org/abs/2210.10341>

²<https://sites.research.google/med-palm/>

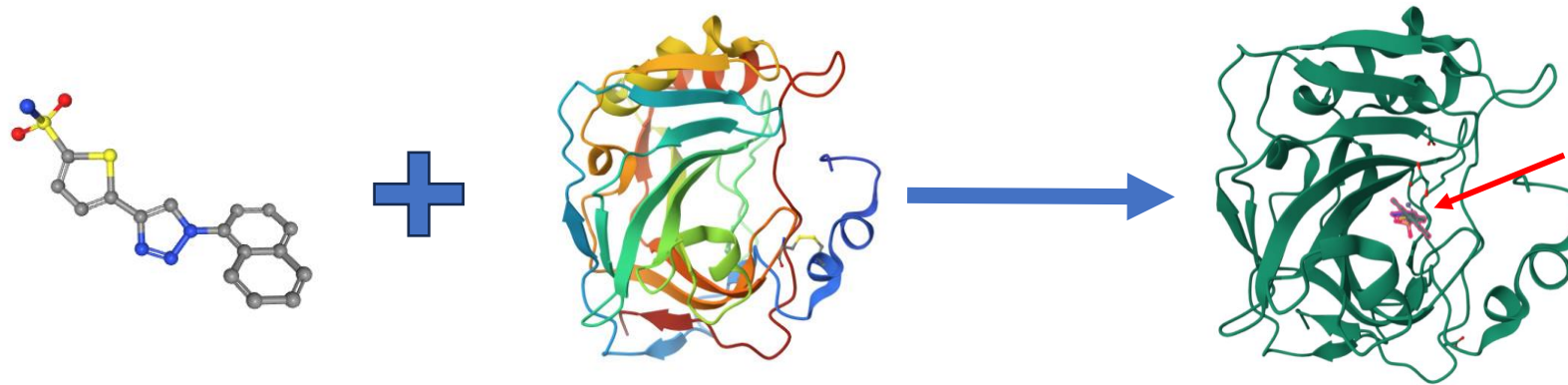
³<https://arxiv.org/pdf/2104.08942.pdf>

⁴<https://arxiv.org/pdf/1907.08322.pdf>

Thank You

Finding novel small molecule inhibitors of Carbonic anhydrase IX with fine-tuned LLM

Carbonic anhydrase IX (CA9) is an enzyme that found overexpressed in many types of human cancer cells including clear cell renal cell carcinoma, and thus be considered as a potential drug target.

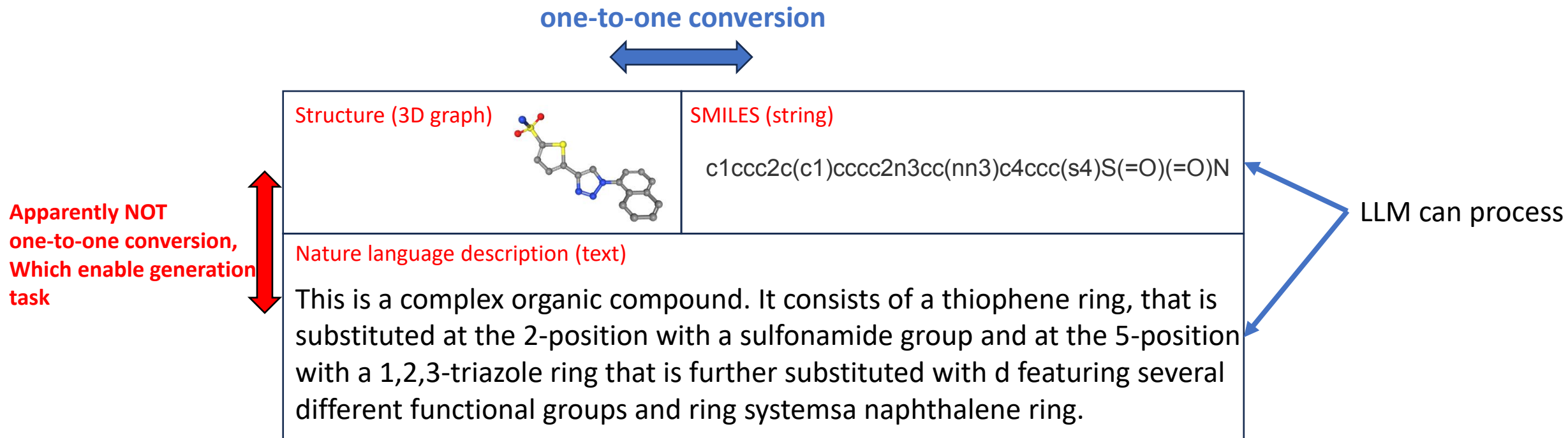


Inhibiting an enzyme with small molecules is like putting metal sheets into a paper shredder to choke it, but only metal sheets in certain shape can get into it

For CA9, we know molecules with sulfonamide group have high possibility to be good inhibitors

**Can we use LLM to find common features from known small molecule inhibitors of CA9?
Will LLM get the same conclusion as human experts? Or can LLM find new features?
Moreover, Can LLM generate and test novel small molecule inhibitors of CA9?**

Subtask1: Teach the LLM to translate across different molecule descriptions



An existing transformer-based model MolT5¹ did so, we only need to reproduce its ability by fine-tuning an LLM

For cost estimation: MolT5-large has ~800 million parameters

1. Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Subtask2: Fine-tuning with knowledge of inhibitor

Open accessible inhibitor datasets:

1. ~3K molecules with experimental Ki results¹
2. ~100K molecules with DEL screening experiment results²

Fine-tune and validate LLM by:

1. Feed LLM with subsets of known inhibitors, let LLM summarize common features from them
2. Let LLM generate novel structures based on features it found
3. Compare LLM-generated molecules with left-out known inhibitors

1. Shmilovich K, Chen B, Karaletsos T, et al. DEL-Dock: Molecular Docking-Enabled Modeling of DNA-Encoded Libraries[J]. *Journal of Chemical Information and Modeling*, 2023, 63(9): 2719-2727.

2. Gerry, C. J.; Wawer, M. J.; Clemons, P. A.; Schreiber, S. L. DNA barcoding a complete matrix of stereoisomeric small molecules. *Journal of the American Chemical Society* 2019, 141, 10225–10235.

Guided-Chain-of-Thought Prompting Improves Confidence Calibration of Large Language Models

Jinchuan Tian, Yiqing Xie, Zichun Yu, Xinran Zhao

CMU 11-667 Course Project

Background

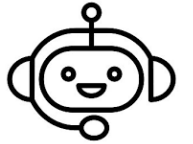
- **Research Question:** How to let model confidence reflect model performance.
- **Why Bother?** Obtaining calibrated performance is an important step towards treating LLMs as a responsive student that can do **self-evolvement**.
- **Literature:**
 - Previous work on **prompting** mainly studies the performance side, not the confidence calibration side
 - Previous work on **confidence** is not comprehensively compared nor grounded to context change
- **Ours:** how will changing prompt styles affects model confidence calibration? How our Guide-COT helps?



User

Can sunlight travel to the deepest part of the Black Sea? Answer with your confidence?

It can not, my confidence is 95%.



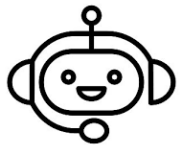
bot



User

What are the facts you rely on?
And what is confidence given the facts?

1. sunlight penetration ability ...
2. life in anoxic conditions ...
3. Black sea depth ...
My confidence is now 98%.



bot

Guided-COT

- **Intuition**: we wish models to output answers with respect to elicited internal knowledge as **Elicited Constraints**.
- **Sanity Check**: adding human knowledge and explicit steps help improve calibration.
- **Guided-COT**: let the model to elicit its known facts/sources/reasoning in the context to ground the confidence generated.
- **Practical Consideration**: How to incorporate external fact-checking pipeline in the future?

Prompting Method	ECE ↓
Default	30.3
CoT	29.6
Oracle-Steps	19.4
Oracle-Facts	20.6
Oracle-Steps&Facts	15.8

Table 1: Expected Calibration Error (ECE) of adding external human-annotated step questions and Facts.