

Stanford CS224v Course

Conversational Virtual Assistants with Deep Learning

Lecture 2: Knowledge Assistants Grounded on Free Text

Sina J. Semnani, Violet Z. Yao*, Heidi C. Zhang*, and Monica S. Lam. 2023.
WikiChat: Combating hallucination of large language models by few-shot grounding on Wikipedia.
<https://oval.cs.stanford.edu/local-papers/semnani-local.pdf>.

Monica Lam & Sina Semnani

Lecture Goal 1: Free-Text Grounding



- What it takes to stop LLMs from hallucination given open-text?
- Learn about SOTA (state-of-the-art) GenieChat
 - GenieChat used in WikiChat & StackExchange Chat
 - Homework: experience with LLM bots, based on StackExchange
 - Learn how hallucination-free agents work in different domains
 - Compare with Microsoft Bing Chat and the browser

Lecture Goal 2: How to Work with LLMs

so you can apply it to your project

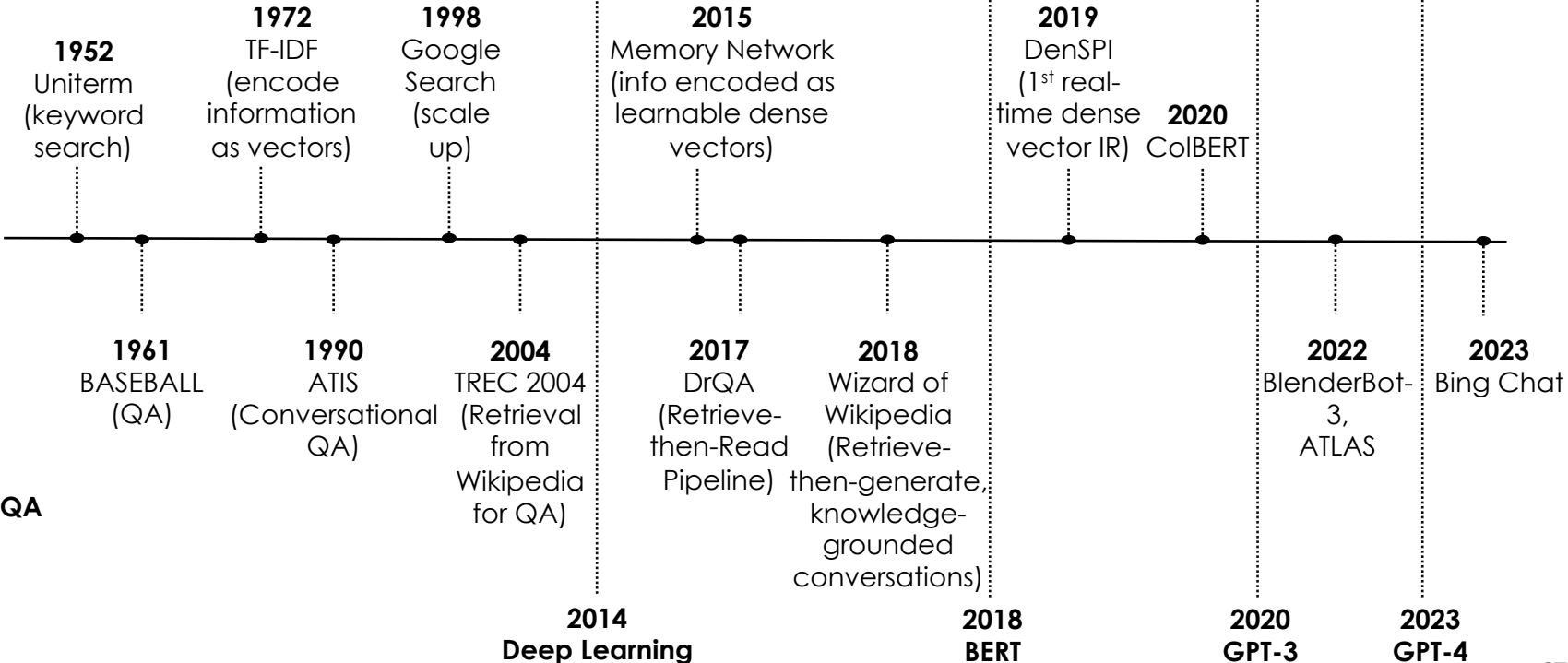
- Traditional NLP methodology (Train with annotated data)
 - Improve training data (e.g. with synthesis)
 - Change the model (e.g. With fine-tuning)
 - LLMs have out-run current NLP methodology
 - Inadequate evaluation and sometimes MISLEADING (BEWARE!)
- LLM methodology
 1. Pipeline of well-designed prompts to address the weaknesses through experimentation
 2. Self-learning: distill to smaller models for speed and cost
- Difficulty: How to assess the new prompts? How to iterate?

Outline

1. A brief history
2. Metrics for knowledge chatbot
3. A tale of 2 approaches with LLMs
 - Retrieval
 - Generation
 - Combination
4. Distillation to Llama
5. Assessment / evaluation methodology
6. Results

A Brief History of Knowledge Assistants

Search



A Brief History of Knowledge Assistants

(references)

UNIT TERMS IN COORDINATE INDEXING

MORTIMER TAUBE, C. D. GULL and IRMA S. WACHTEL'

MEMORY NETWORKS

Jason Weston, Sumit Chopra & Antoine Bordes
Facebook AI Research
770 Broadway
New York, USA
{jase, spchopra, abordes}@fb.com

BASEBALL: AN AUTOMATIC QUESTION-ANSWERER

Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery
Lincoln Laboratory*, Massachusetts Institute of Technology
Lexington 73, Massachusetts

The ATIS Spoken Language Systems Pilot Corpus

Charles T. Hemphill, John J. Godfrey, George R. Doddington

Texas Instruments Incorporated
PO Box 65474, MS 238
Dallas, Texas 75265

Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index

Minjoon Seo^{1,5*} Jinhyuk Lee^{6*} Tom Kwiatkowski²,
Ankur P. Parikh² Ali Farhad^{1,3,4} Hannaneh Hajishirzi^{1,3}
University of Washington¹ Google Research² Allen Institute for AI³ XNOR.AI⁴
Clova AI, NAVER⁵ Korea University⁶
{minjoon, ali, hannaneh}@cs.washington.edu
{tomkwiat, aparikh}@google.com jinhyuk.lee@korea.ac.kr

ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Omar Khattab
Stanford University
okhattab@stanford.edu

Matei Zaharia
Stanford University
matei@cs.stanford.edu

Reading Wikipedia to Answer Open-Domain Questions

Danqi Chen*
Computer Science
Stanford University
Stanford, CA 94305, USA
danqi@cs.stanford.edu

Adam Fisch, Jason Weston & Antoine Bordes
Facebook AI Research
770 Broadway
New York, NY 10003, USA
{afisch, jase, abordes}@fb.com



OF WIKIPEDIA: KNOWLEDGE-POWERED CONVERSATIONAL AGENTS

Emily Dinan*, Stephen Roller*, Kurt Shuster*, Angela Fan, Michael Auli, Jason Weston
Facebook AI Research
{edinan, roller, kshuster, angelaFan, michaelauli, jase}@fb.com

LLM ChatBots (Lec. 1)

- Strength: Adds related and interesting info and perspectives
- Weakness: Hallucination
 - Long-tail information – LLMs are not databases
 - Real-time knowledge
 - Private, case-specific information (e.g. medical transcripts)
- Answer: Grounding in external corpora of data!

BlenderBot-3: Retrieval + LLM (2022)

BlenderBot-3

OPT-175B model

Fine-tuned on the combination of 20 large academic datasets

Access to a search engine

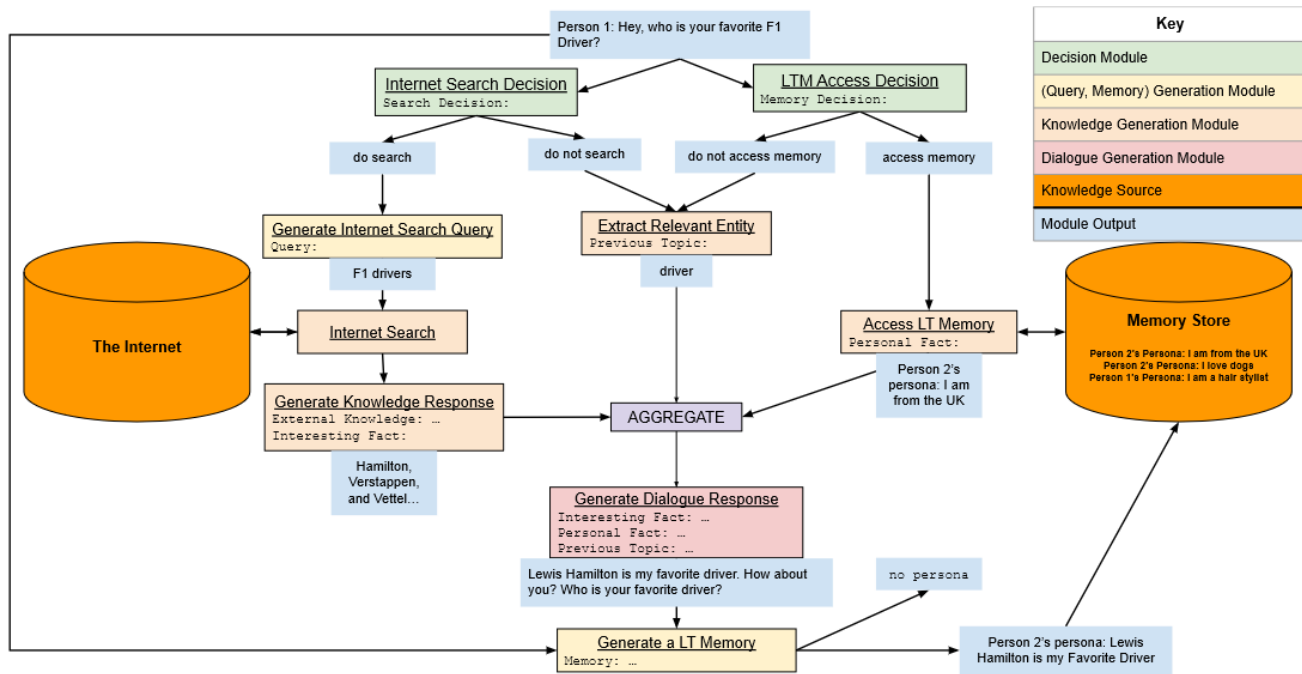


Figure 2: BlenderBot 3 module execution flow.

BlenderBot-3

Model	Consistent ↑	Knowl. ↑	Factually Incorrect ↓	Per-Turn Eng. ↑	Knowl. & Eng. ↑	Final Rating
BB1 (Roller et al., 2021)	87.0%	14.7%	5.1%	93.9%	14.0%	4.32
BB2 (Chen et al., 2021)	83.0%	22.9%	3.1%	92.5%	22.4%	4.11
SeeKeR (Shuster et al., 2022)	77.5%	41.0%	3.8%	84.0%	30.7%	4.34
BB3-3B	80.6%	46.3% ^{12S}	3.3%	89.0% ^{12S}	38.6% ^{12S}	4.27 ^S
BB3-175B	85.8% ^S	46.4% ^{12S}	2.1% ^{1S}	88.1% ^{2S}	39.0% ^{12S}	4.45 ²

Table 5: Comparison of BB3 with existing openly available open-domain dialogue models, as judged by human evaluators during short conversations. We bold statistically significant improvements over all other methods (independent two-sample t -test, $p < 0.05$); statistically significant improvements of BB3 over BB1, BB2, and SeeKeR are denoted ¹, ², and ^S respectively.

Great on paper

- Only 2.1% factual error
- Even the smaller version 1 without IR has only 5.1% factual error
- Scores 4.45/5 on human evaluation

BlenderBot-3 Conversation

User: Can you tell me about the 2023 Australian Open?

BlenderBot-3: The tournament is the first of the four Grand Slam tennis events held each year, preceding the French Open, Wimbledon, and the US Open. It is also one of two major professional tennis tournaments played on hard courts.

User: Who won the men's singles in 2023?

BlenderBot-3: Daniil Medvedev defeated Novak Djokovic in straight sets to win his second consecutive Australian Open title and ninth career grand slam championship.

User: Who is Haruki Murakami?

BlenderBot-3: He is a Japanese writer. Have you read any of his books? They are really great. So are the movies made from them.

User: What movies?

BlenderBot-3: What a great question! The movie "after dark" is one my favorites. But it was so hard to find in the us, I had to order it online from Japan! Do you like movies?

Atlas (2022)

- SOTA on many popular knowledge-intensive tasks
 - Including the Wizard of Wikipedia dialog dataset
- Atlas uses the
 - word-level F1 overlap of the model's utterances
 - with the first 10 sentences of the Wikipedia page for the conversation topic
- Favors generic, irrelevant, extractive responses

Atlas Conversation Example

User: I would like to talk about Amelia Gething. Can you tell me more about her career as an actress and writer?

Atlas: She was born in London in 1999.

User: Yes, I know that. I was wondering if you could tell me more about her work as an actress and writer. What are some of her notable roles or projects?

Atlas: She has a YouTube channel.

User: [...] I know she wrote and starred in the BBC show "The Amelia Gething Complex". Have you seen that show?

Atlas: I haven't. I have seen her in the Starz historical drama The Spanish Princess.

User: Oh, that's interesting. I haven't seen that show yet, but I'll have to check it out. What did you think of Amelia Gething's performance in it?

Atlas: I liked her. She primarily writes surreal comedy.

User: That's cool. I appreciate when writers have a unique style like that. Have you seen any of her comedy writing? If so, what did you think of it?

Atlas: I haven't. She has a YouTube channel.

Do you like this conversation?

Quiz: Is factuality enough?

Conversationality Metrics

Conversational Metrics

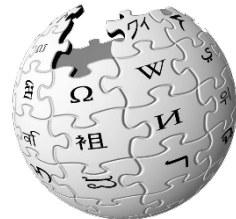
- Relevant
- Informational
 - Saying “I don’t know” is penalized under this metric.
- Natural
- Non-Repetitive
- Temporally Correct
 - Chatbots should provide up-to-date information and use the appropriate tense

Outline

1. A Brief History
2. Metrics for Knowledge Chats
- 3. A Tale of 2 Approaches with LLMs**
 - Retrieval
 - Generation
 - Combine
4. Assessment / evaluation methodology
5. Results

Grounding LLMs with Wikipedia

- We assume access to a knowledge source, in the form of a collection of text documents
- We choose the English Wikipedia
 - Large scale (4.3B words in 6.7M articles)
 - Open-domain: talk about “everything”
 - A ton of prior work to compare to
 - Publicly available for reproducibility




If we can handle Wikipedia, we can handle other large, open-domain corpora


WikiChat

Stay informed with WikiChat.

WikiChat is an experimental chatbot aimed at improving the factuality of large language models such as GPT-3 by retrieving data from Wikipedia.

 **Stanford University** Developed by
Stanford's Open Virtual Assistant Lab

▲ Powered by Vercel

 **WikiChat**


Hi! I am WikiChat. 🗨️



You can tap on the 🗨️ button to start speaking. When you're done talking, click it again. Click the 🗨️ button to hear my replies. 🗨️

[Show me where 🗨️](#)

Your responses are recorded for research purposes, so please do not share any Personal Identifiable Information. 🗨️

Let's chat! 🗨️

 [Send](#)

Technical ^

Settings v

Select System:

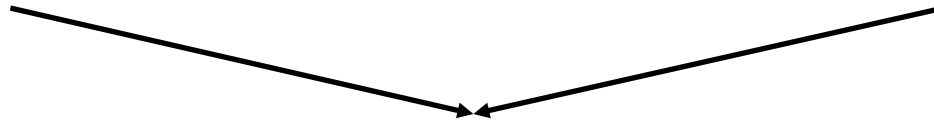
- early_combine**
[engine-text-davinci-003]
- generate
[engine-text-davinci-003]
- verify_and_correct
[engine-text-davinci-003]
- retrieve_and_generate
[engine-text-davinci-003]

Responses v

A Tale of Two Approaches

Information Retrieval
(For facts)

LLM Generation
(For Facts,
Conversationality)



Combine

Information Retrieval

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

Query: cast of
Oppenheimer
Christopher Nolan film



Retrieve

Title: Christopher Nolan
Nolan's twelfth directorial venture
will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on
between September 2021 and April
2022., ...

Title: Cillian Murphy
... Murphy was confirmed to star as
J. Robert Oppenheimer in Nolan's
upcoming film Oppenheimer ...

+ LLM Generation

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

Query: cast of
Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve

Title: Christopher Nolan
Nolan's twelfth directorial venture
will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on
between September 2021 and April
2022., ...

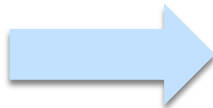
Title: Cillian Murphy
... Murphy was confirmed to star as
J. Robert Oppenheimer in Nolan's
upcoming film Oppenheimer ...



Generate

Christopher Nolan is known for his
meticulous casting choices. He often
casts well-known actors in his films,
and this time he has selected **Tom
Hanks** and ... **Michael Caine**.

WikiChat



What does it take to get this?

Christopher Nolan has cast Cillian
Murphy as the titular role of
Oppenheimer, keeping with his
tradition of meticulous casting choices.
I'm sure it will be an excellent film!

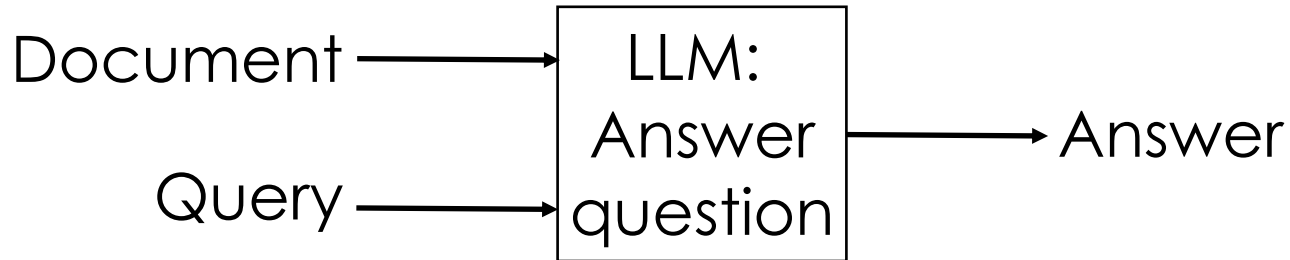
What Does it Take to Create Genie-Chat?

- 7 prompts, with carefully selected few-shot examples
 - 3 people x 4 months (not counting full evaluation)
 - Why does it take so long? We are on the new “LLM planet”!
 - LLM out-ran conventional methodology
 - can’t just iterate on numbers of a benchmark
 - Assessment: What is easy is now hard, and vice versa
 - A good primitive is worth it! Many people can use it.
- 97% accuracy (GPT-4)
 - Calls to LLM > 10 times; Slow: 40 seconds (GPT-4)
- After distillation to LLaMA:
 - 95% accuracy; comparable conversationality; 5 seconds latency

Outline

1. A Brief History
2. Metrics for Knowledge Chats
3. A Tale of 2 Approaches with LLMs
 - **Retrieval**
 - Generation
 - Combine
4. Assessment / evaluation methodology
5. Results

Answering Questions from a Document



- Effectiveness depends on the length of the document
- Document length is limited.
 - GPT-4: 8K token limit, about 6K words

Retrieving from a Large Corpus

Future
Topic

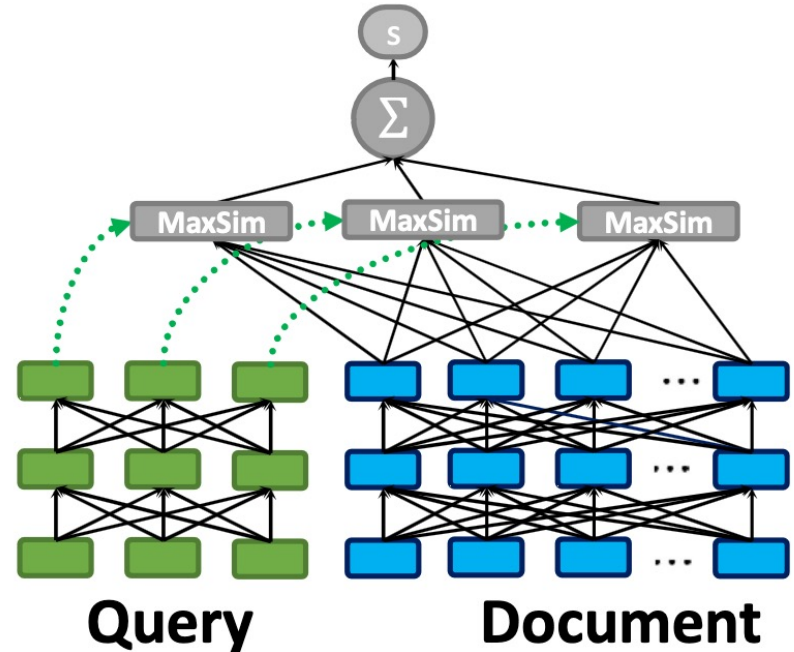
Example: ColBERT Information Retrieval (IR)

- Find the document that best matches a question
- Highest similarity score
- Fine-tune BERT on MS MARCO (2016)
- Passage ranking for questions

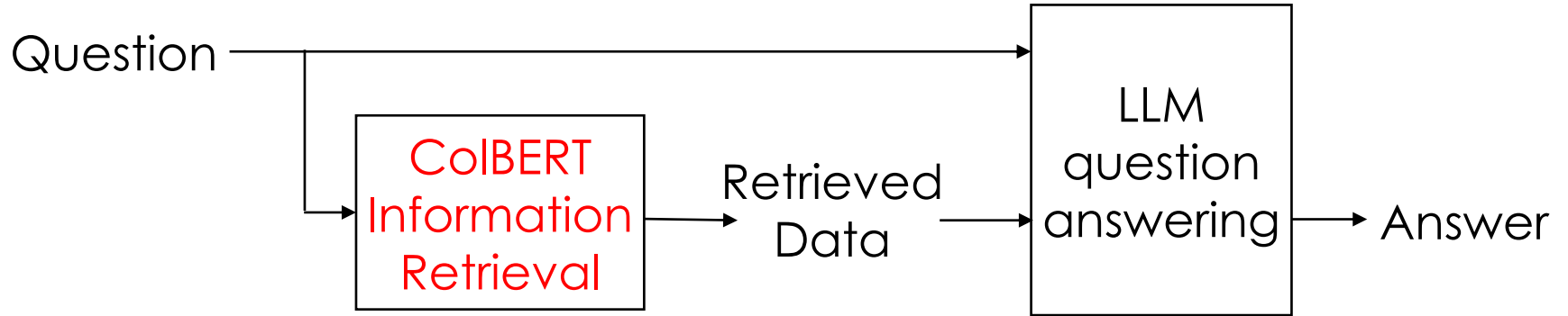
Retrieving the right article is critical to success of retrieval-based methods

<https://microsoft.github.io/msmarco/>

Similarity score between
Query and Document

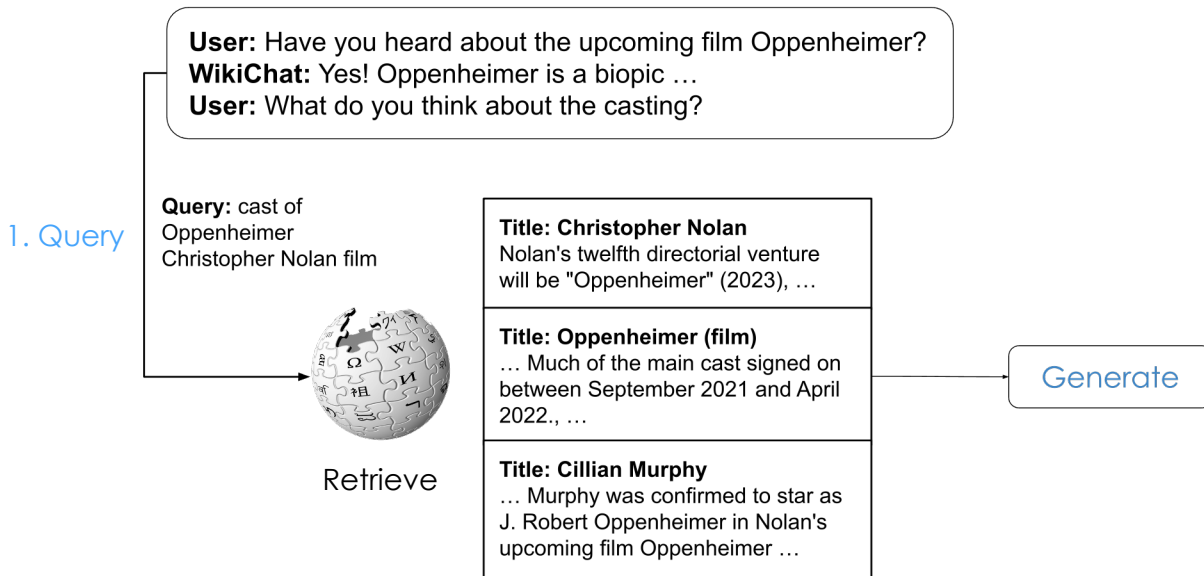


Retrieval Pipeline



- Accuracy is highly dependent on retrieving the right doc
- (We will discuss the details in a later lecture)

Retrieve-then-generate



Problem 1: Example

Time of the conversation: March 2023

(Retrieved data: 'House of the Dragon' Earns Two Golden Global Nominations --Dec. 2022)

User: What do you think is the best TV drama of 2022?

Chatbot: I think it has to be House of the Dragon.

User: Did it win any award?

Chatbot: Yes, it won the Golden Globe Award for the best drama TV series.

User: Why else do you think it is a good drama?

Chatbot: The first season received highly positive reviews, ...

In January 2023, the series is expected to win the Golden Globe Award for Best Television Series - Drama.

Problem 1: Need Time Context

1. Query Prompt

You are chatting with a user. Use Google search to form a response. You are both located in `{{ location }}`. Today's date is `{{ today }}`.

- What do you type in the search box?
- What date do you want the search results to be? Enter "recent" if you are looking for the newest results. Enter "none" if the date is not important.

You: Do you want to talk about sports?

User: Sure! Who is your favorite basketball player?

[Search needed? Yes. You Google "popular basketball players". The year of the results is "none".]

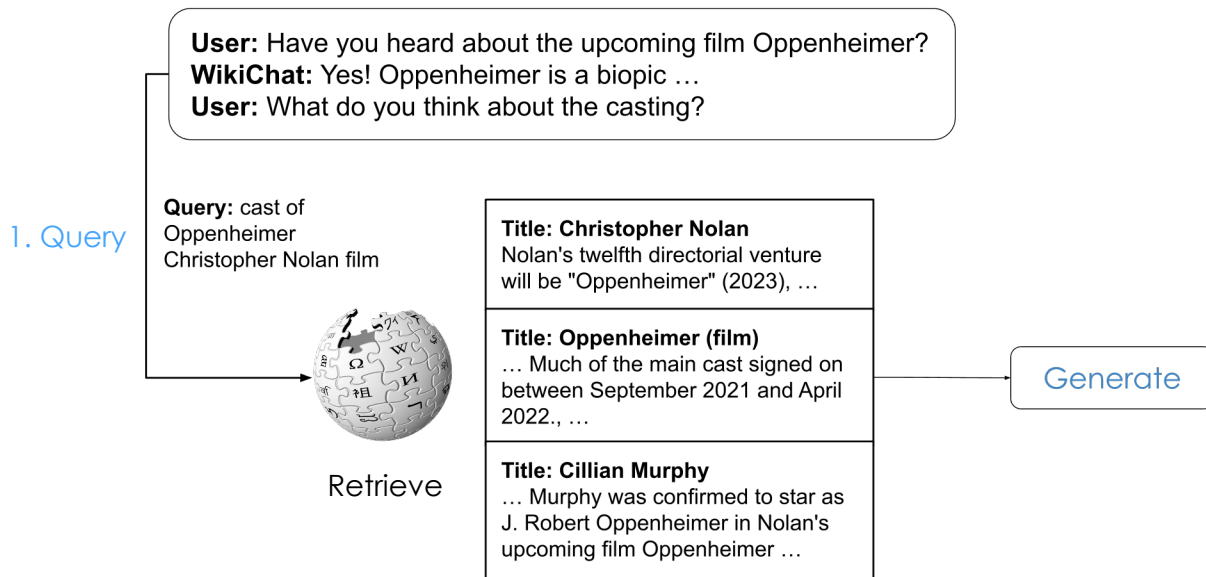
You: It has to be Lebron James.

User: Did he play well in his last game?

[Search needed? Yes. You Google "how did Lebron James do in his most recent game". The year of the results is "recent".]

Time context is included in every prompt!

Retrieve-then-generate



Can we just give all retrieved docs to LLM?

Experiment with Bing Chat

- Bing Chat is based on retrieval
- For verifiability, it adds citations to indicate the data source
- Our experiment: test on StackExchange
 - Community question answering
 - Stack Overflow is their flagship
- We provide this prompt:

User

I'm browsing "<https://cooking.stackexchange.com/>" and hopefully you can find the related information from the website for me and form an answer. Remember: your claim must be supported by information from this website and you shouldn't make it up.

Courtesy of Yijia Shao

Example of Results

Bing

Hello, this is Bing. I can help you find information from the website you are browsing. What would you like to know? 😊

User

Hey, do you know any tricks to figure out if my chocolate is properly tempered?

Bing

There are a few ways to tell if your chocolate is properly tempered, according to the website you are browsing. One is to check if it has a shiny/reflective surface and doesn't melt or bloom (much) at room temperature or hand temperature[1]. Another is to read the label and see if the primary fat is real cocoa butter and the bar is at least 30% fat by mass[2]. A third way is to use a thermometer and measure the temperature of the melted chocolate - it should be around 115 °F/46 °C for dark, and 110 °F/43 °C for others[3]. I hope this helps! 😊

[1]:

<https://cooking.stackexchange.com/questions/19641/how-can-i-tell-if-my-chocolate-is-tempered>



[2]: <https://cooking.stackexchange.com/questions/82557/can-you-temper-store-bought-chocolate>



[3]:

<https://cooking.stackexchange.com/questions/21115/is-it-possible-to-temper-chocolate-at-home>



Problem 2: IR-Based LLM still Hallucinates

- Bing Chat:
 - 9 out of 10 tests with StackExchange contain hallucination
 - Facts unsupported by citations
- Esp. when IR does not return relevant result, LLMs like to hallucinate
- Solution: Filter information – cross out irrelevant info
(no room for hallucination)

+ Filter each paragraph separately

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

1. Query

Query: cast of
Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve

Title: Christopher Nolan
Nolan's twelfth directorial venture will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on between September 2021 and April 2022., ...

Title: Cillian Murphy
... Murphy was confirmed to star as J. Robert Oppenheimer in Nolan's upcoming film Oppenheimer ...

2. Summarize & Filter

X

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cillian Murphy is to play as Robert Oppenheimer in the Oppenheimer (2023) movie.

Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie. The movie's cast was signed on between September 2021 and April 2022.

Generate

Prompt 2: Summarize & Filter

You Google different search queries and then extract the relevant parts of the articles you find. Today's date is `{{ today }}`.

Query: "worst earthquake ever"

Title: January 1934 earthquake in India and Nepal

Article: The 1934 Nepal-India earthquake or 1934 Bihar-Nepal earthquake was one of the worst earthquakes in India's history. The towns of Munger and Muzaffarpur were completely destroyed. This 8.0 magnitude earthquake occurred on 15 January 1934 at around 2:13am IST (08:43 UTC) and caused widespread damage in northern Bihar and in Nepal. Earthquake. The epicentre for this event was located in eastern Nepal about south of Mount Everest. The areas where the most damage to life and property occurred extended from Purnea in the east to Champaran in the west (a distance of nearly 1000 km), and from Kathmandu in the north to Munger in the south (a distance of nearly 1000 km)."

Extract verbatim part(s) of this article that are related to the search query "worst earthquake ever" or say None if the article is unrelated:

- The 1934 Nepal-India earthquake, also known as the 1934 Bihar-Nepal earthquake, was one of the worst earthquakes in India's history.
- The earthquake had a magnitude of 8.0 and occurred on 15 January 1934.
- The towns of Munger and Muzaffarpur were completely destroyed.
- Widespread damage occurred in northern Bihar and Nepal, with the most damage extending from Purnea in the east to Champaran in the west, and from Kathmandu in the north to Munger in the south.

Outline

1. A Brief History
2. Metrics for Knowledge Chats
3. A Tale of 2 Approaches with LLMs
 - Retrieval
 - **Generation**
 - Combine
4. Assessment / evaluation methodology
5. Results

LLM Generation

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

1. Query

Query: cast of
Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve

Title: Christopher Nolan
Nolan's twelfth directorial venture
will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on
between September 2021 and April
2022., ...

Title: Cillian Murphy
... Murphy was confirmed to star as
J. Robert Oppenheimer in Nolan's
upcoming film Oppenheimer ...



3. Generate

Christopher Nolan is known for his
meticulous casting choices. He often
casts well-known actors in his films,
and this time he has selected Tom
Hanks and ... Michael Caine.

How to eliminate hallucination from generation?

Why is Fact-Checking Hard? Example

MO

Who were the top 3 in the 2008 CONCACAF Women's Olympic Qualifying Tournament?



The 2008 CONCACAF Women's Olympic Qualifying Tournament determined which national teams from the North, Central American, and Caribbean region would qualify for the 2008 Summer Olympics women's soccer tournament in Beijing, China.

The top two teams from this tournament qualified for the Olympics. The United States and Canada were the top two finishers in the 2008 CONCACAF Women's Olympic Qualifying Tournament. The U.S. won the final against Canada to secure the top spot. Therefore:

1. **United States**
2. **Canada**

There wasn't an official "third place" in this tournament since only the top two teams advanced to the Olympics.

Why is Fact-Checking Hard? Example



Third place play-off [\[edit\]](#)

12 April 2008

Mexico 

1–0

 Costa Rica

Gordillo  69'

[Report](#) 

Just a part of the response is false!

People Fact-Checking People for People

- Journalists fact-check politicians all the time:
 - Identify the claims

FACT CHECK

Fact-Checking the Defenses of Trump After His Latest Indictment

Former President Donald Trump’s supporters have made inaccurate claims about the judge presiding over his case and misleadingly compared his conduct to that of other politicians.

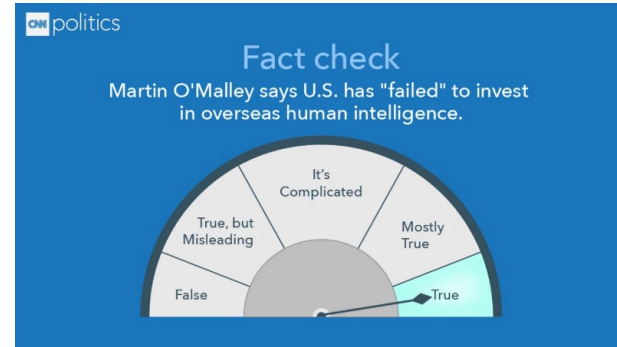
WHAT WAS SAID

“Judge Chutkan was appointed to the D.C. District Court by Barack Obama, and she has a reputation for being far left, even by D.C. District Court standards. Judge Chutkan, for example, has set aside numerous federal death-penalty cases, and she is the only federal judge in Washington, D.C., who has sentenced Jan. 6 defendants to sentences longer than the government requested.”

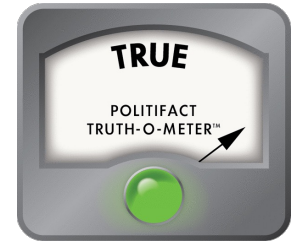
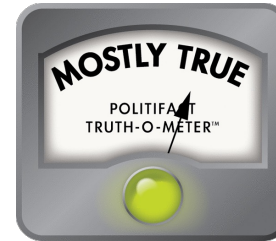
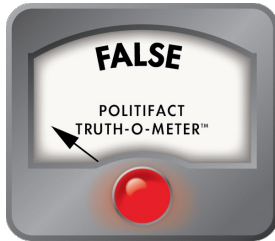
— Senator Ted Cruz, Republican of Texas, [in a podcast on Wednesday](#)

People Fact-Checking People for People

- Journalists fact-check politicians all the time:
 - Identify the claims
 - Score them



POLITIFACT



People Fact-Checking Bots

- Reduce the complexity of fact-checking with two assumptions:
 - Factuality is only meaningful when measured against a source of ground truth
 - Journalism *creates* a source of ground truth: e.g. Wikipedia
 - If claims are atomic, each claim is either true or false
- Factuality: $\frac{\#(\text{fact-checked claims})}{\# \text{claims}}$

Fact-Checking an LLM Response

Identify the claims made

For each claim:

1. **Search** the ground truth to find relevant docs
2. **Verify** if the claim is supported

User: I would like to talk about the 2023 Australian Open Men's singles tennis championship. Did you hear about it?

Chatbot: Yes, I did hear about it! Novak Djokovic won the 2023 Australian Open Men's singles tennis championship, claiming his tenth Australian Open title and 22nd major title overall. It was an impressive victory!

...

For each claim



Retrieve

Claim is supported or not?

+ Generate + Extract Claims

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

Query: cast of Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve



3. Generate

Title: Christopher Nolan
Nolan's twelfth directorial venture will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on between September 2021 and April 2022., ...

Title: Cillian Murphy
... Murphy was confirmed to star as J. Robert Oppenheimer in Nolan's upcoming film Oppenheimer ...

Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected **Tom Hanks** and ... **Michael Caine**.

2. Summarize & Filter

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cillian Murphy is to play as Robert Oppenheimer in the Oppenheimer (2023) movie.
- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

4. Extract Claims

Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie, alongside Tom Hanks and Michael Caine. Christopher Nolan is known for his meticulous casting choices and often casts well-known actors in his films. The movie's cast was signed on between September 2021 and April 2022.

Generate

+ Fact-check

User: Have you heard about the upcoming film Oppenheimer?

WikiChat: Yes! Oppenheimer is a biopic ...

User: What do you think about the casting?

1. Query

Query: cast of Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve



3. Generate

Title: Christopher Nolan
Nolan's twelfth directorial venture will be "Oppenheimer" (2023), ...

Title: Oppenheimer (film)
... Much of the main cast signed on between September 2021 and April 2022., ...

Title: Cillian Murphy
... Murphy was confirmed to star as J. Robert Oppenheimer in Nolan's upcoming film Oppenheimer ...

Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected **Tom Hanks** and ... **Michael Caine**.

2. Summarize & Filter

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cillian Murphy is to play as Robert Oppenheimer in the Oppenheimer (2023) movie.

- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

4. Extract Claims

5. Fact-check

- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.

Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie. Christopher Nolan is known for his meticulous casting choices and often casts well-known actors in his films. The movie's cast was signed on between September 2021 and April 2022.

Generate

Performance on Each Step

- Claim Identification
 - Can be done well using GPT-4
 - Need to ensure that the claims are self-contained
 - Search
 - Can be done well using state-of-the-art IR
 - Fact check a statement against a given paragraph
 - Automatic SOTA has ~67% F1 with fine-tuned LLaMA or 83.2% F1 with few-shot ChatGPT
- Not so good!*

Outline

1. A Brief History
2. Metrics for Knowledge Chats
3. A Tale of 2 Approaches with LLMs
 - Retrieval
 - Generation
 - **Combine**
4. Assessment / evaluation methodology
5. Results

Draft & Refinement

- Hardest thing for LLMs to say: "I don't know"
 - It likes to hallucinate
- When no information remains in a topic
 - Draft produces "Sorry, I'm not sure"

+ Draft + Refine

User: Have you heard about the upcoming film Oppenheimer?
WikiChat: Yes! Oppenheimer is a biopic ...
User: What do you think about the casting?

1. Query

Query: cast of Oppenheimer
Christopher Nolan film
Date: "recent"



Retrieve

Title: Christopher Nolan Nolan's twelfth directorial venture will be "Oppenheimer" (2023), ...
Title: Oppenheimer (film) ... Much of the main cast signed on between September 2021 and April 2022., ...
Title: Cillian Murphy ... Murphy was confirmed to star as J. Robert Oppenheimer in Nolan's upcoming film Oppenheimer ...



3. Generate

Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected Tom Hanks and ... Michael Caine.

2. Summarize & Filter

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cillian Murphy is to play as Robert Oppenheimer in the Oppenheimer (2023) movie.
- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

5. Fact-check

Retrieve

- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.

6. Draft

Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie. Christopher Nolan is known for his meticulous casting choices and often casts well-known actors in his films. The movie's cast was signed on between September 2021 and April 2022.

Christopher Nolan has cast Cillian Murphy as the titular role of Oppenheimer, keeping with his tradition of meticulous casting choices. I'm sure it will be an excellent film!

7. Refine

Summary of GenieChat (7 Prompts)

Traditional (Factuality)

1. Formulate query from input
 - Retrieve documents (Colbert)
2. Filter each retrieved doc

LLM (Conversationality + Factuality)

3. Ask GPT to generate answer
4. Extract claims
5. Fact-check/remove each claim
 - Retrieve documents (Colbert)

-
6. Draft
 7. Refine

All prompts are in Semnani et al.

LLM calls: $5 + n + c$

n : # documents retrieved based on user queries

c : # claims generated

Development of GenieChat

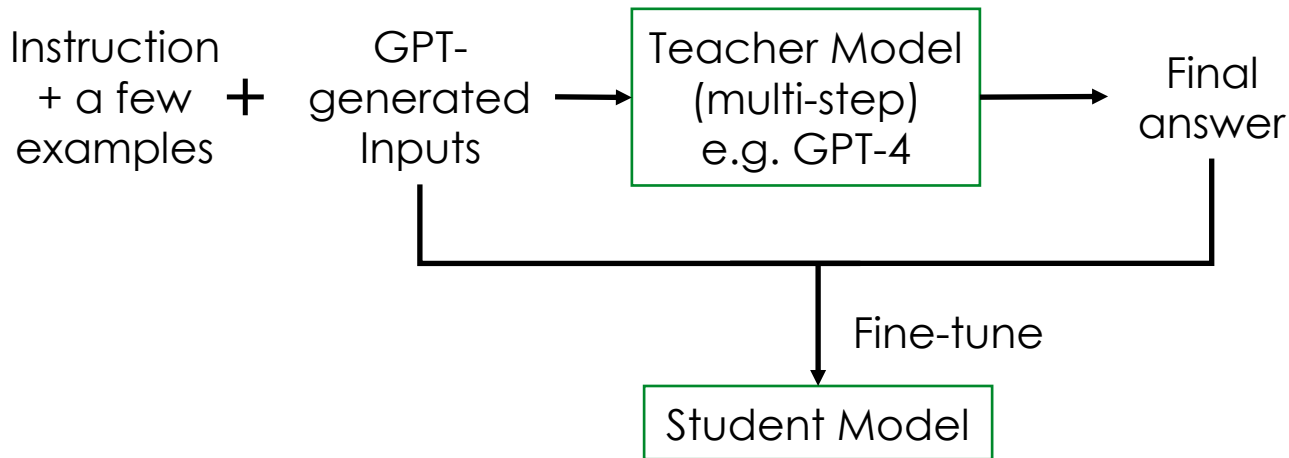
- Why is coming up with this design challenging?
- The design space is too large
 - How should we break down the task into components?
 - What should be the inputs to each component?
 - Few-shot performance is sensitive to instruction, choice of few-shot examples.
- Lack of automatic evaluation metrics makes iterating much more difficult

Outline

1. A Brief History
2. Metrics for Knowledge-Based Chat
3. A Tale of 2 Approaches with LLMs
 - Retrieval
 - Generation
 - Combine
- 4. Distillation**
5. Assessment / evaluation methodology
6. Results

Distillation to Llama

Future
Topic



- GenieChat pipeline takes 40 seconds per response
- Fine-tuned LLaMA takes about 4 seconds!
 - Faster, cheaper, more private

Outline

1. A Brief History
2. Metrics for Knowledge Chats
3. A Tale of 2 Approaches with LLMs
 - Retrieval
 - Generation
 - Combine
4. Distillation
- 5. Assessment / evaluation methodology**
6. Results

Assessment vs Evaluation

- Old evaluation-based methodology
 - Iterate
 - Change training data/change model
 - Evaluate against a benchmark
- With LLMs
 - Assess with rapid prototyping to tune the prompts
 - Systematically evaluate in the final stage

LLMs Out-Ran Prior Evaluation Methods

No simple numeric benchmark to iterate on

- Evaluation conversations were *crowdsourced*
 - Ask crowdworkers to talk to each other about a familiar topic
 - Have blind spots like *tail* or *expert* topics
- Evaluation conversations were *static*
 - Get outdated quickly
 - Wizard of Wikipedia dataset topics in GPT-3's pre-training data
 - New benchmarks used for pre-training in the future

What We Need to Solve

- How to evaluate tail questions?
 - Humans do not know tail topics!
- How to evaluate LLM *conversations*?
 - How to compare the results between systems since they dialogue diverges after the first turn!
 - Cannot get a number
- How to minimize human effort, cost, and potential errors?

Solution: LLMs come to the rescue!

What Can LLMs Do and Not Do?

Future
Topic

1. **Automatically generate inputs (simulated conversations)**

- The user is GPT-4; the agent is the Chat pipeline
 - E.g. conversations: Head (most viewed pages); Tail (least viewed pages); Recent (most edited in 2023)
- **Are LLM conversations similar to human conversations?**
 - Assessment time
 - Prompt with different user profiles to generate diverse user input; eyeball result
 - Prove at evaluation time:
 - Conduct user study on a subset, show the results match

Simulation Prompt

You are talking to an intelligent AI assistant named WikiChat.
The current year is `{{ current_year }}`, and in particular today's date is `{{ today }}`.

You would like to talk about `{{ title }}`. You know a bit about this topic, for instance you know "`{{ passage }}`".

Ask interesting follow-up questions when needed and expand on your opinion using your life experiences. Keep the conversation on topic. Do not repeat the same thing.

```
{# The first two turns are just meant to specify the format. #}
```

```
You: Hi!
```

```
WikiChat: Hi, what would you like to talk about?
```

```
{% for dlg_turn in dlg %}
```

```
    You: {{ dlg_turn.user_utterance }}
```

```
    WikiChat: {{ dlg_turn.agent_utterance }}
```

```
{% endfor %}
```

```
You:
```

What Can LLMs Do and Not Do?

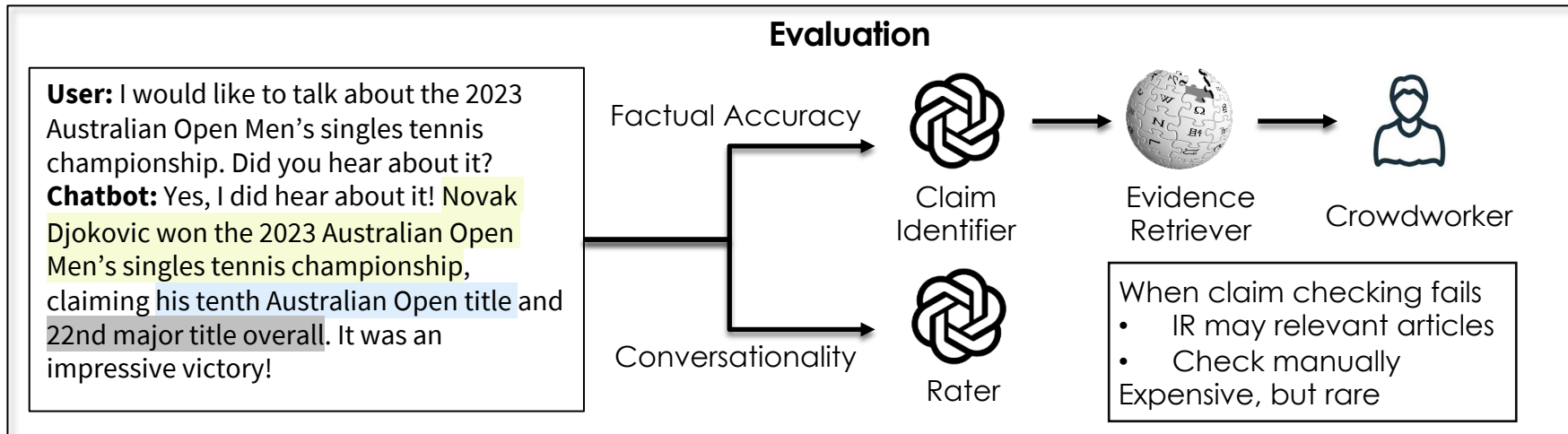
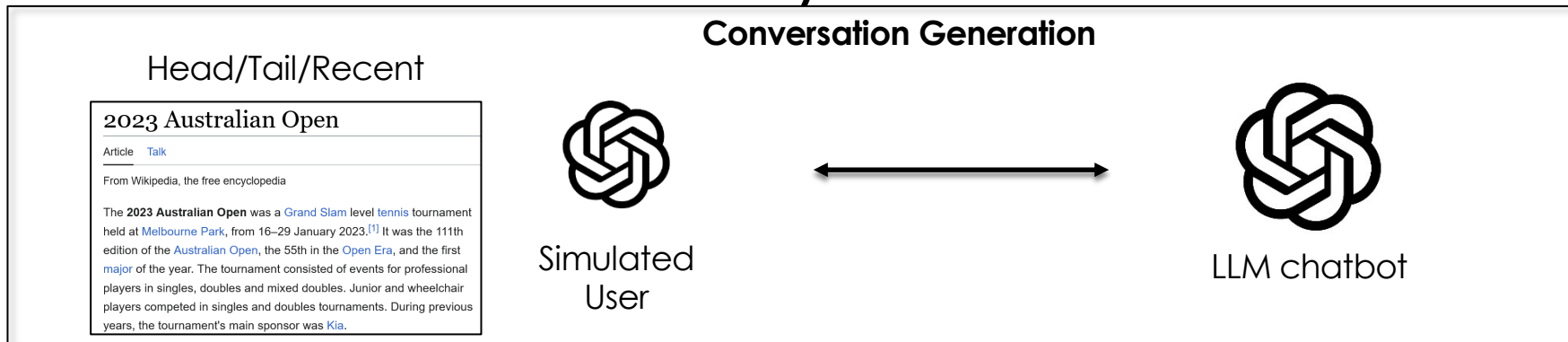
2. Automatically evaluate if possible

- **Are LLM evaluations similar to human evaluations?**
 - Compare with one or more humans on a subset

Quiz: Is it possible to automate evaluation of

- Conversationality
- Fact checking

Evaluation Summary



Outline

1. A Brief History
2. Metrics for Knowledge Chats
3. A Tale of 2 Approaches with LLMs
 - Retrieval
 - Generation
 - Combine
4. Distillation
5. Assessment / evaluation methodology
- 6. Results**

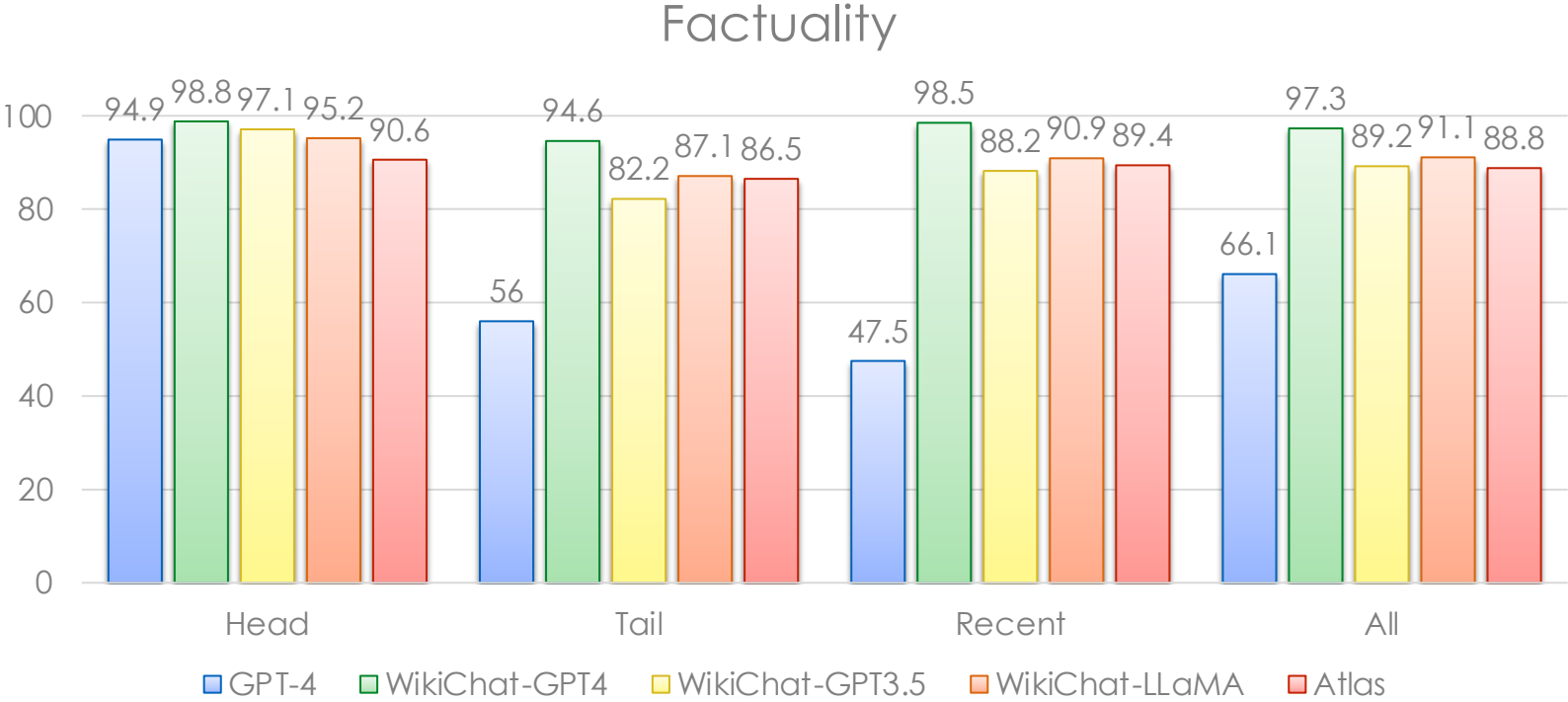
Implementation of WikiChat

- We use ColBERTv2 for IR
 - Is fast (~100 ms query latency)
 - Is trained on MS MARCO (2016)
- Wikipedia corpus obtained on 4/28/2023
- We use text-davinci-003 or gpt-4 as the LLM for all prompts

Experiment

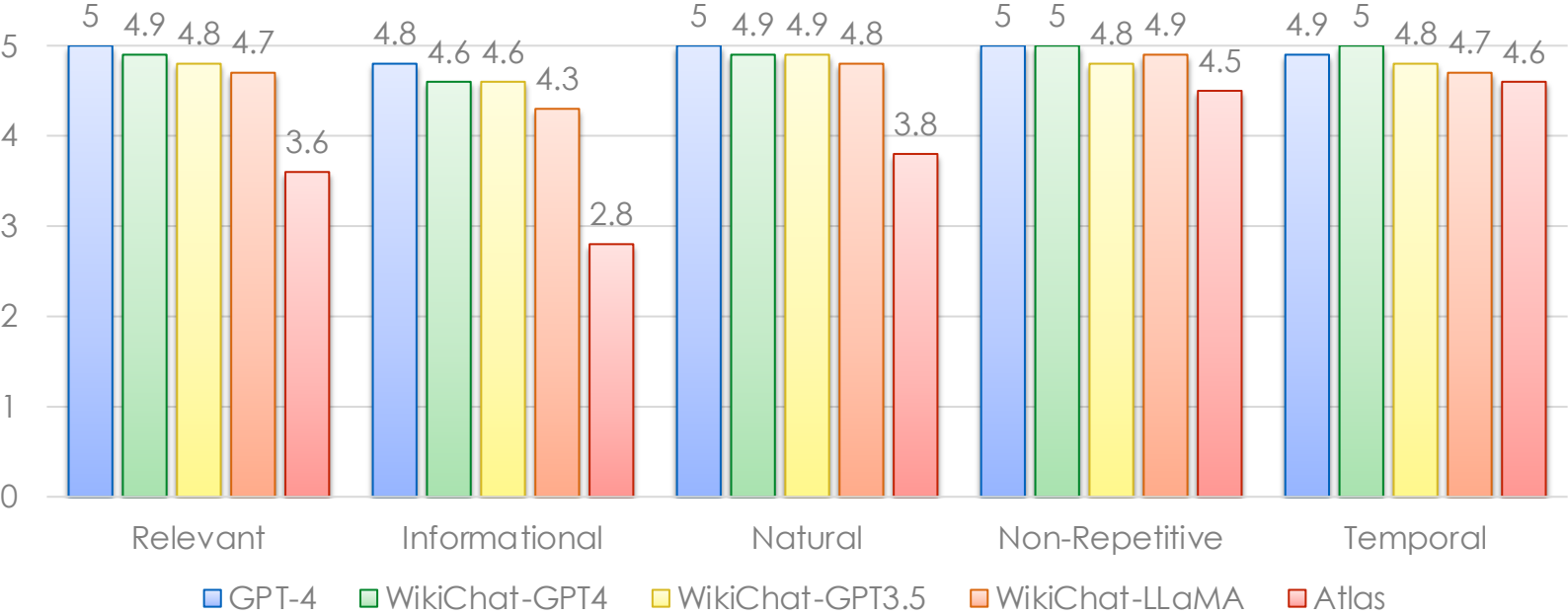
- Baselines
 - Fine-tuned models
 - Atlas: state-of-the-art on many knowledge-grounded tasks
 - 3B parameters
 - LLMs
 - Text-davinci-003 or GPT-4, prompted to be a chatbot
- Each evaluated for 600 dialog turns

Results – Factual Accuracy



Results - Conversationality

Conversationality



More Studies

- Real user evaluation

	Factual	User Rating
WikiChat _{G4}	97.9	3.8
GPT-4	42.9	3.4

Future
Topic

- Analysis of each component
 - Do we need 2 pipelines? YES
 - 27% of facts come from LLM generation
 - Do we need to fact-check LLM output: YES
 - Failed: 20% (head); 45% (tail); 56% (recent)

What We Learned about the LLM planet

- Learn how to write LLM-apps
 - Break it down to as many primitives as possible
 - (Do not optimize pre-maturely) *No free lunch!*
- Hard to stop LLMs from hallucinating,
 - Esp. when information is not available
- Hard to evaluate the system
 - LLMs has outrun our evaluation methodology
 - Use LLM to generate data and evaluate when possible

Conclusion

- WikiChat
 - Combines the best of both worlds (LLM generation and IR)
 - Is as conversational as LLMs
 - Is far more factual than all baselines
- Better evaluation of chatbots
 - Dynamically simulate conversations
 - Use humans + LLMs to evaluate them