Stanford CS224v Course
Conversational Virtual Assistants with Deep Learning

Lecture 5

Minimizing Hallucination in Retrieval-Augmented Generation

Monica Lam & Sina Semnani

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia
Sina J. Semnani, Violet Z. Yao*, Heidi C. Zhang*, Monica S. Lam
In Proceedings of Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023.
Winner of the Wikimedia's Research of the Year Award, 2024.

Stanford University

Retrieval Augmented Generation (RAG)

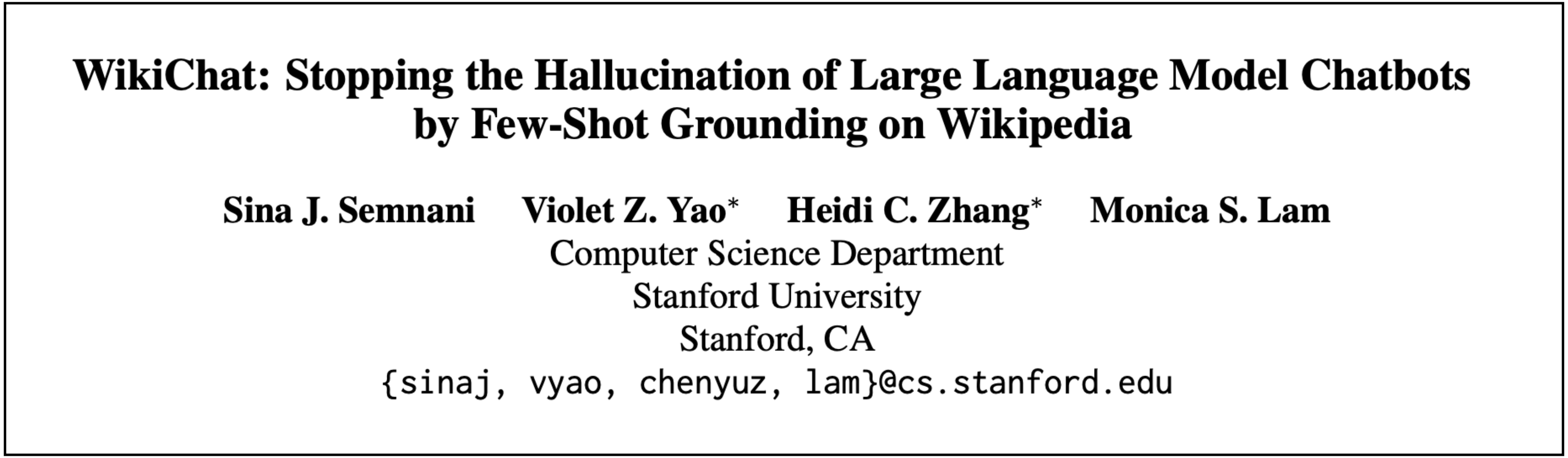
In 2022-2023

Can't we just ask LLMs?

No, no, no, you need to RAG.
You have only tested on things you know!
LLMs know what you know!

OK, it may hallucinate now,
but can't we just train a bigger model with more data.

LLMs do not know the long tail, new real-time info,
or private info!



WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia

Sina J. Semnani **Violet Z. Yao*** **Heidi C. Zhang*** **Monica S. Lam**
Computer Science Department
Stanford University
Stanford, CA
`{sinaj, vyao, chenyuz, lam}@cs.stanford.edu`

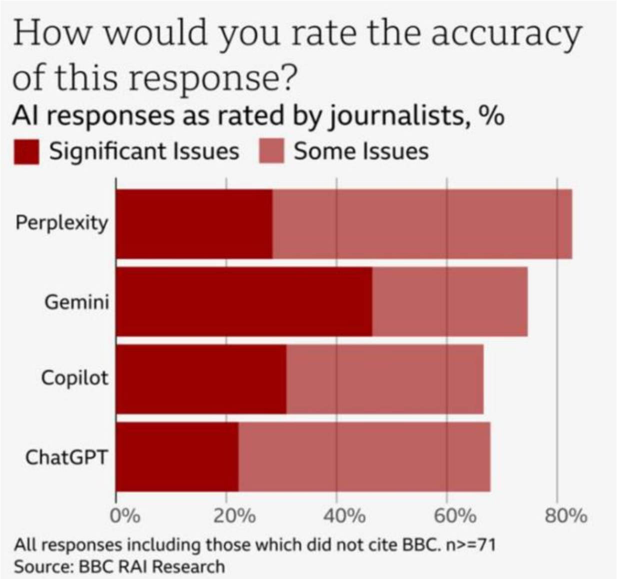
- Conversational Q&A in open domain with 97% accuracy
- In Proceedings of Association for Computational Linguistics: EMNLP, Singapore, **December 2023**.
- Winner of the Wikimedia's Research of the Year Award, 2024

RAG in Commercial Products (2024)

RAG available in : ChatGPT, Co-Pilot, Gemini, Perplexity

BBC reported in **January 2025**:

- 51% of AI answers to news problems have significant issues.
- 19% of AI answers which cited BBC content introduced factual errors.
- 13% of the quotes sourced from BBC articles were either altered or didn't actually exist in that article.



How would you rate the accuracy of this response?

AI responses as rated by journalists, %

All responses including those which did not cite BBC. n>=71
Source: BBC RAI Research

Representation of BBC News content in AI Assistants. 2025.

LECTURE GOAL

THE INSIDE STORY OF HALLUCINATION-FREE RAG

PARADIGM SHIFT IN NLP RESEARCH

LLM-ENGINEERING IN GENERAL
(USEFUL FOR YOUR PROJECTS)

Lecture Outline

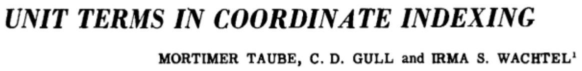
- Background of Knowledge Assistants Technology
- Metrics for Knowledge Assistants
- Methodology Paradigm Shift
- Design of Wikichat
- Evaluation

Lecture Outline

- **Background of Knowledge Assistants Technology**
- **Metrics for Knowledge Assistants**
- Methodology Paradigm Shift
- Design of Wikichat
- Evaluation

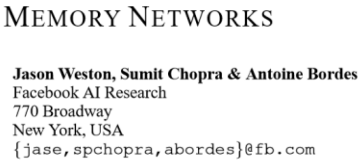
Many Papers on Knowledge Assistants

Key papers



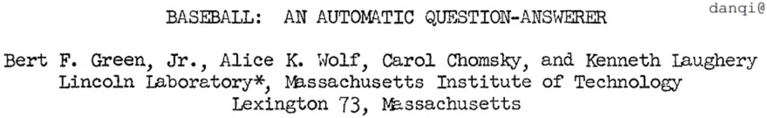
UNIT TERMS IN COORDINATE INDEXING

MORTIMER TAUBE, C. D. GULL and IRMA S. WACHTEL[1]



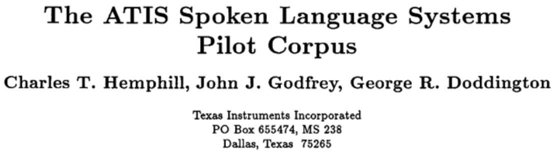
MEMORY NETWORKS

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



BASEBALL: AN AUTOMATIC QUESTION-ANSWERER

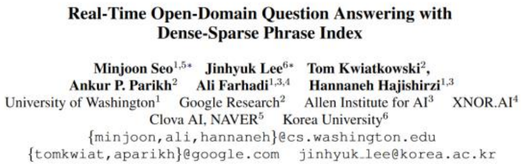
Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery
Lincoln Laboratory*, Massachusetts Institute of Technology
Lexington 73, Massachusetts



The ATIS Spoken Language Systems Pilot Corpus

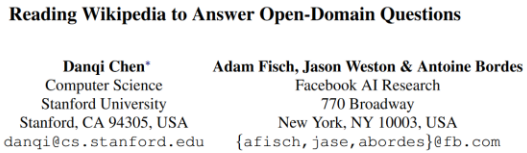
Charles T. Hemphill, John J. Godfrey, George R. Doddington

Texas Instruments Incorporated
PO Box 655474, MS 238
Dallas, Texas 75265



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

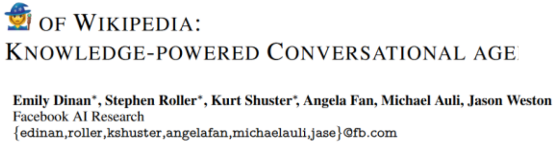
Minjoon Seo[1,5*] Jinhyuk Lee[6*] Tom Kwiatkowski[2], Ankur P. Parikh[2] Ali Farhadi[1,3,4] Hannaneh Hajishirzi[1,3]
University of Washington[1] Google Research[2] Allen Institute for AI[3] XNOR.AI[4]
Clova AI, NAVER[5] Korea University[6]
{minjoon,ali,hannaneh}@cs.washington.edu
{tomkwiat,aparikh}@google.com jinhyuk_lee@korea.ac.kr



Reading Wikipedia to Answer Open-Domain Questions

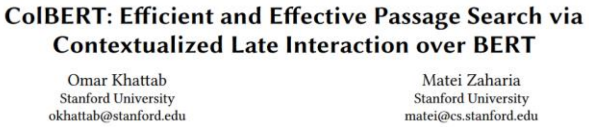
Danqi Chen[*]
Computer Science
Stanford University
Stanford, CA 94305, USA
danqi@cs.stanford.edu

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 AGE

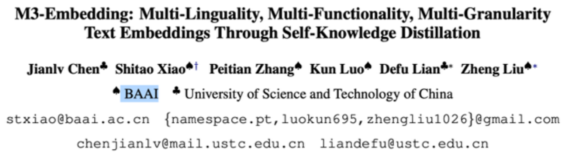
Emily Dinan[*], Stephen Roller[*], Kurt Shuster[*], Angela Fan, Michael Auli, Jason Weston
Facebook AI Research
{edinan,roller,kshuster,angelafan,michaelauli,jase}@fb.com



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

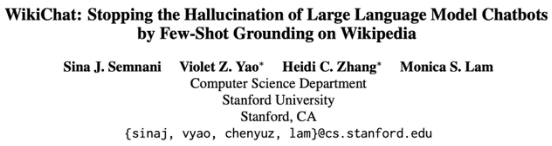
Omar Khattab
Stanford University
okhattab@stanford.edu

Matei Zaharia
Stanford University
matei@cs.stanford.edu



M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation

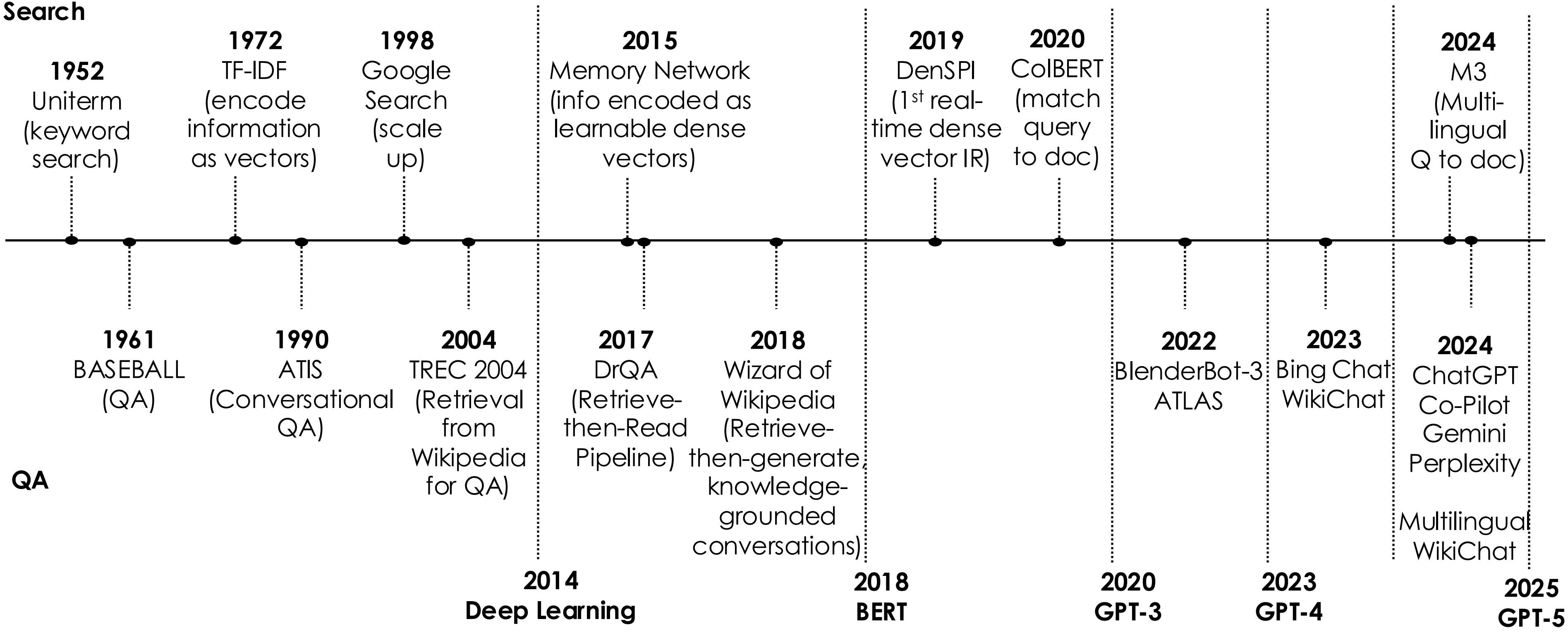
Jianlv Chen[♣] Shitao Xiao[♠†] Peitian Zhang[♠] Kun Luo[♠] Defu Lian[♣*] Zheng Liu[♠*]
[♠] BAAI [♣] University of Science and Technology of China
stxiao@baai.ac.cn {namespace.pt,luokun695,zhengliu1026}@gmail.com
chenjianlv@mail.ustc.edu.cn liandefu@ustc.edu.cn



WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia

Sina J. Semnani Violet Z. Yao[*] Heidi C. Zhang[*] Monica S. Lam
Computer Science Department
Stanford University
Stanford, CA
{sinaj, vyao, chenyuz, lam}@cs.stanford.edu

A Brief History of Knowledge Assistants



Search

1952
Uniterm
(keyword
search)

1972
TF-IDF
(encode
information
as vectors)

1998
Google
Search
(scale
up)

2015
Memory Network
(info encoded as
learnable dense
vectors)

2019
DenSPI
(1st real-
time dense
vector IR)

2020
ColBERT
(match
query
to doc)

2024
M3
(Multi-
lingual
Q to doc)

QA

1961
BASEBALL
(QA)

1990
ATIS
(Conversational
QA)

2004
TREC 2004
(Retrieval
from
Wikipedia
for QA)

2017
DrQA
(Retrieve-
then-Read
Pipeline)

2018
Wizard of
Wikipedia
(Retrieve-
then-generate,
knowledge-
grounded
conversations)

2022
BlenderBot-3
ATLAS

2023
Bing Chat
WikiChat

2024
ChatGPT
Co-Pilot
Gemini
Perplexity

Multilingual
WikiChat

2014
Deep Learning

2018
BERT

2020
GPT-3

2023
GPT-4

2025
GPT-5

BlenderBot-3: Retrieval + LLM (Meta, 2022)

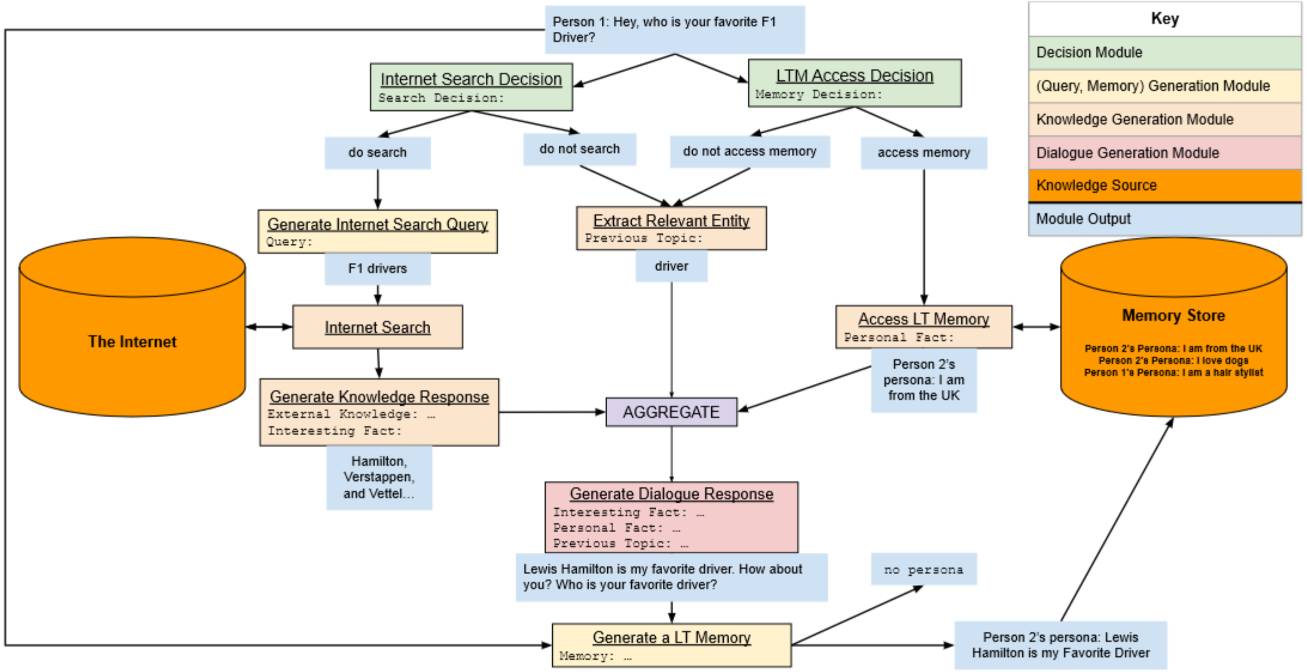
BlenderBot-3

OPT-175B model

Fine-tuned on the combination of 20 large academic datasets

Access to a search engine

Deployed



Person 1: Hey, who is your favorite F1 Driver?

Internet Search Decision
`Search Decision:`

LTM Access Decision
`Memory Decision:`

do search

do not search

do not access memory

access memory

Generate Internet Search Query
`Query:`

F1 drivers

Extract Relevant Entity
`Previous Topic:`

driver

The Internet

Internet Search

Access LT Memory
`Personal Fact:`

Person 2's persona: I am from the UK

Memory Store

Person 2's Persona: I am from the UK
Person 2's Persona: I love dogs
Person 1's Persona: I am a hair stylist

Generate Knowledge Response
`External Knowledge: …`
`Interesting Fact:`

Hamilton, Verstappen, and Vettel…

AGGREGATE

Generate Dialogue Response
`Interesting Fact: …`
`Personal Fact: …`
`Previous Topic: …`

Lewis Hamilton is my favorite driver. How about you? Who is your favorite driver?

`no persona`

Generate a LT Memory
`Memory: …`

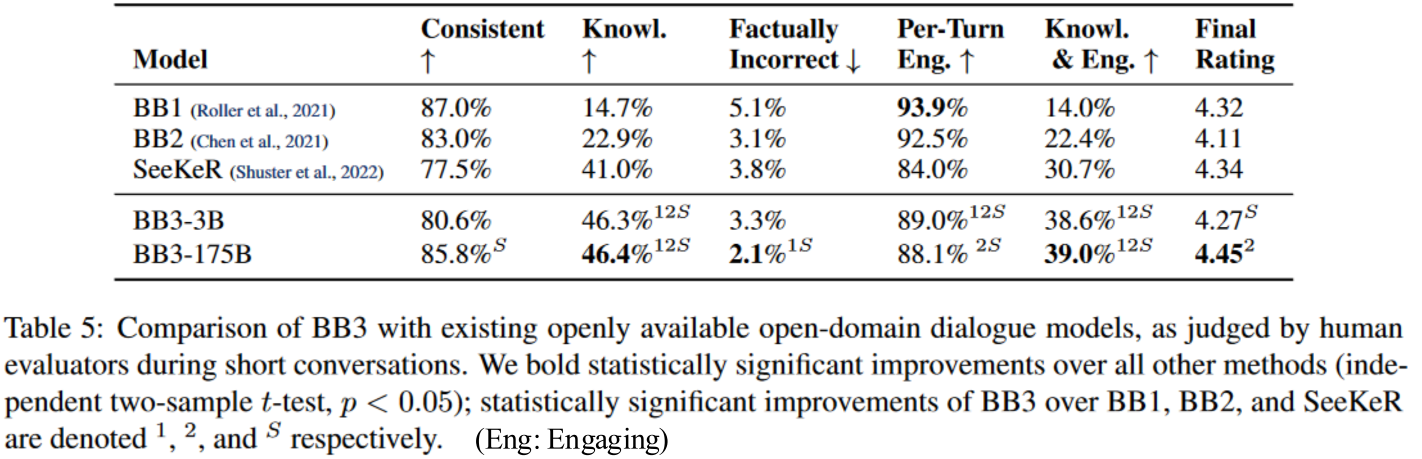
Person 2's persona: Lewis Hamilton is my Favorite Driver

Key
Decision Module
(Query, Memory) Generation Module
Knowledge Generation Module
Dialogue Generation Module
Knowledge Source
Module Output

Figure 2: BlenderBot 3 module execution flow.

BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage, Shuster et al, arXiv 2022

Evaluation on 100 Crowdsource Conversations



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. (Eng: Engaging)

Evaluation

- 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: Danill 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?

FACT CHECK
Novak Djokovic claimed the men's singles title in 2023, his tenth Australian Open title and 22nd major title overall, tying Rafael Nadal's all-time record.

“After Dark” is a book

Key Takeaway

- Blenderbot-3 is very conversational
  - Different from a Q and A bot
- Blenderbot-3 has high evaluation from crowdworkers but has poor accuracy in our tests!
  - Quiz: Why?

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: Few-shot Learning with Retrieval Augmented Language Models, Izacard et al, arXiv 2022

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

Lecture Outline

- Background of Knowledge Assistants Technology
- Metrics for Knowledge Assistants
- **Methodology Paradigm Shift**
- Design of Wikichat
- Evaluation

Need for a Paradigm Shift in NLP Techniques

- The classic NLP development methodology:
  - Create a dataset (train, eval, test)

    "Annotations in existing NLP datasets are full of mistakes!
    LLM-generated results are better!"

    "There is no single golden answer. How do we compare?"
  - Train

    "Fine-tuning on a problem-specific dataset
    can make LLMs worse (and it is very expensive).

WikiChat Engineering Took 3 Students 4 Months!

Evaluation Took Another 2-3 months!

LLM Engineering

Systematically create well-tested step-by-step algorithms (computational thinking)

1. Design the metrics

2. Start with the most general instruction

3. Debug until you can't break it

- Test with "critical" examples
- Root-cause analysis of LLM errors
- Add/refine the steps
- Add rationale to the prompt; add critical examples as an aid

4. Evaluate

- Tests must cover the hardest cases
- Test with LLM-simulated chats with hard, varied personalities
- Evaluate with an LLM
  - If LLM disagrees with human
    - Use human evaluation
- Test with real users

Good news: Debugging with critical examples achieves high Wikichat quality.

The key: Look at the errors!

Lecture Outline

- Background of Knowledge Assistants Technology
- Metrics for Knowledge Assistants
- Methodology Paradigm Shift
- **Design of Wikichat**
- Evaluation

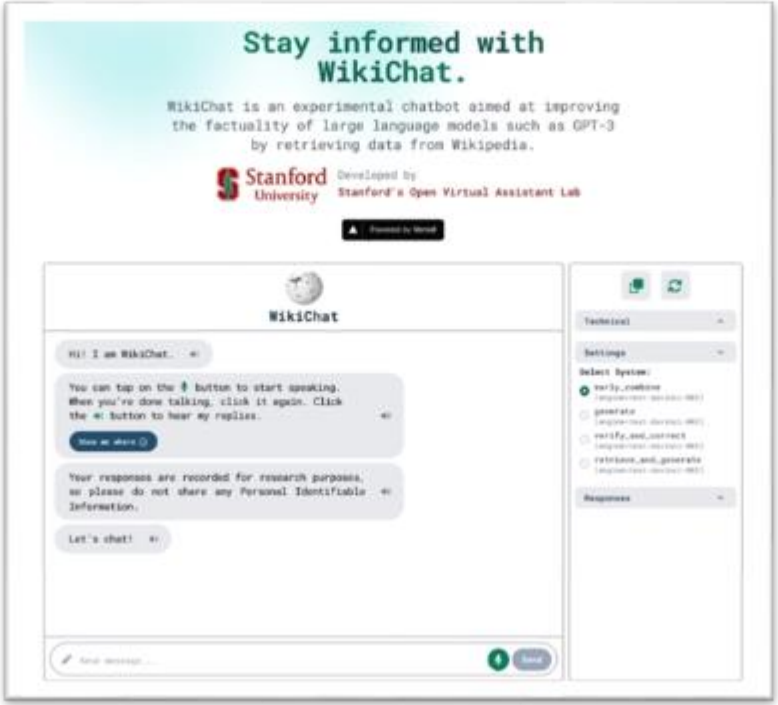
Grounding LLMs with Wikipedia

- 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

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.

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

Let's chat!

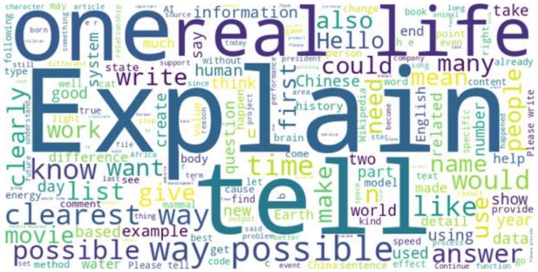
Select System:
● early_combine
○ generate
○ verify_and_correct
○ retrieve_and_generate

WikiChat Demo

https://wikichat.genie.stanford.edu/

35K+ questions so far, in 10+ languages

Words from conversations translated into English



How small could a fusion reactor be?

Tell me about the Mamba architecture.

1901年的蘇格蘭兒童幾歲要上學？
有義務教育嗎？

Cosa bisogna sapere se si vuole scalare il monte Fuji？

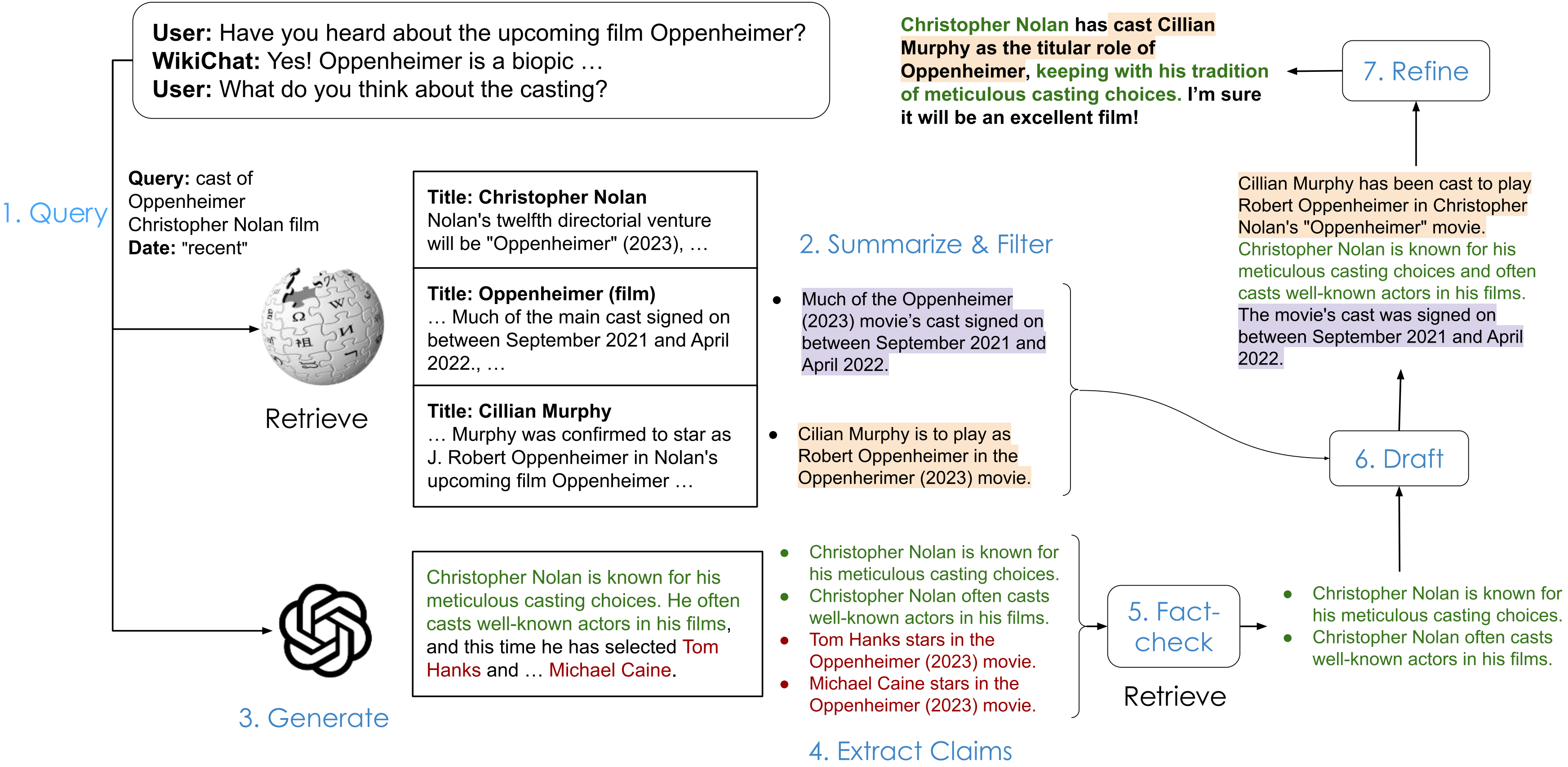
Formalizing our Human Cognitive Process

Information Retrieval
(For facts)

LLM Generation
(For Facts,
Conversationality)

Combine

The 7-Stage WikiChat Pipeline



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

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

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.

4. Extract Claims

- 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.

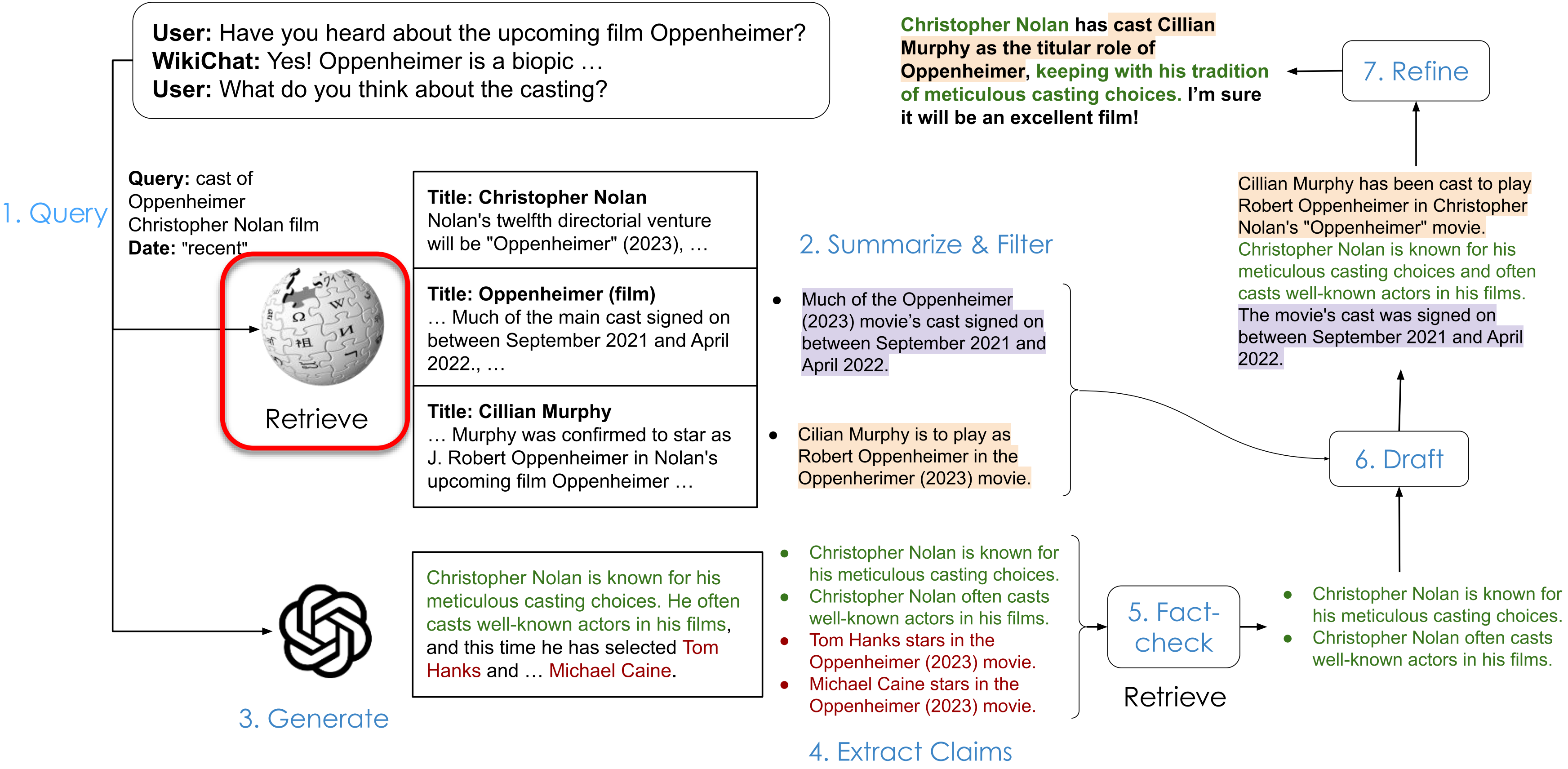
6. Draft

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.

7. Refine

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!

The 7-Stage WikiChat Pipeline



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

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

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.

4. Extract Claims

- 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.

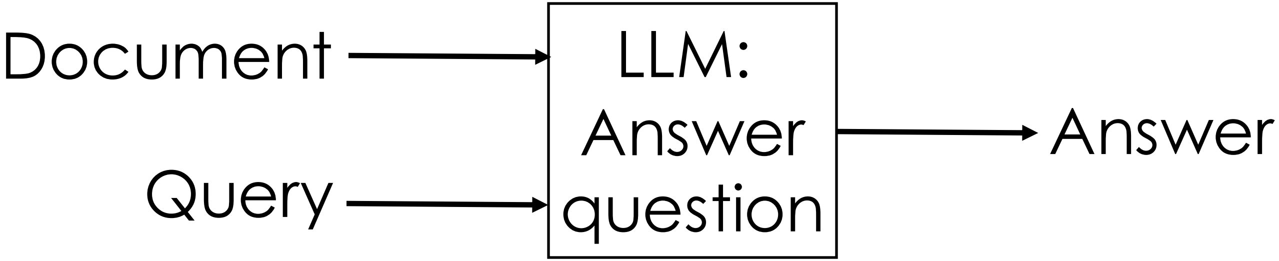
6. Draft

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.

7. Refine

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!

Answering Questions from a Document



Document

Query

LLM: Answer question

Answer

- Effectiveness depends on the length of the document
- Document length is limited.
    - GPT-4: 128K token limit, about 96K words

Retrieving from a Large Corpus

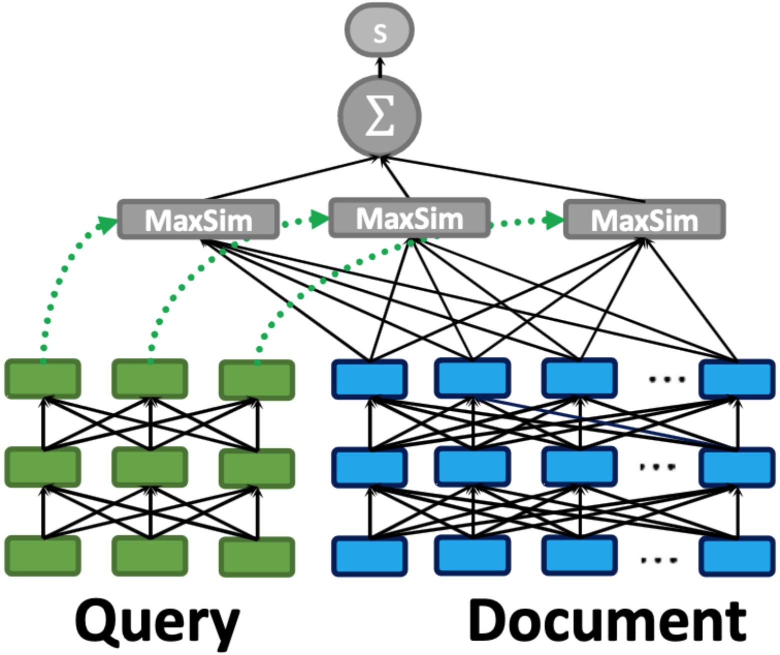
Future Topic

ColBERT Information Retrieval (IR)

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

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

Similarity score between Query and Document



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

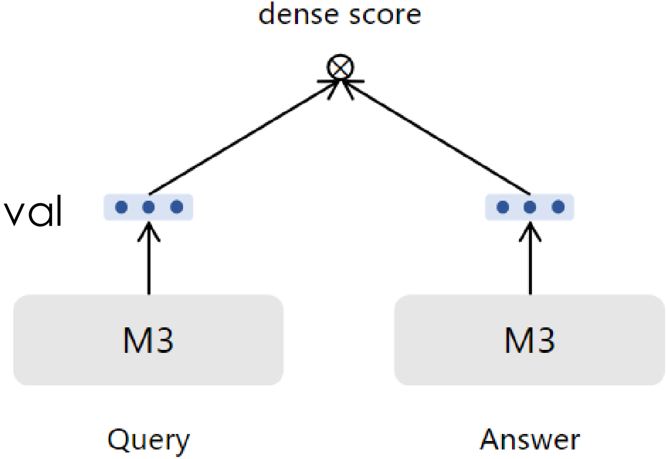
M3 Embedding (Multi-lingual Documents)

Future Topic

Find the document that best matches a question

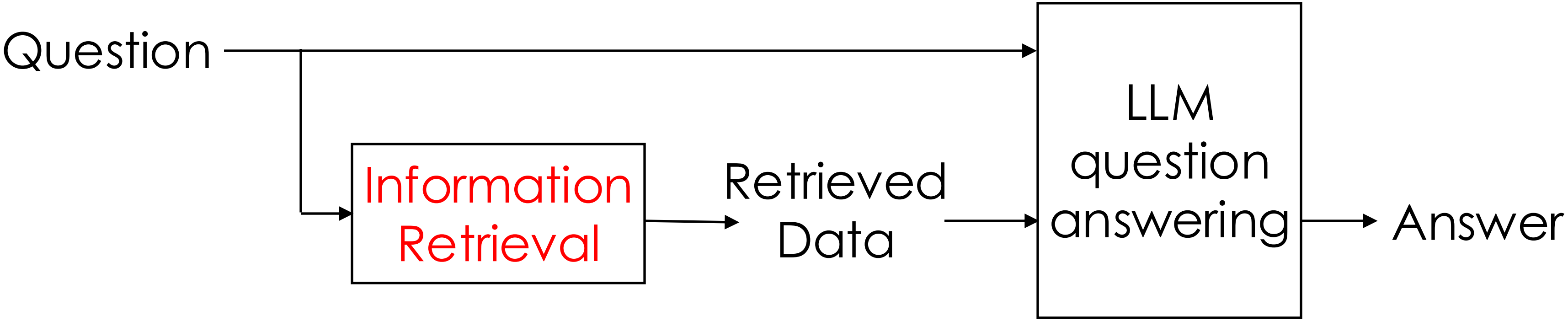
- Fine-tune XLM-R on 17 multilingual passage ranking datasets
  - And some synthetic data too
- Multilingual: > 100 languages
- Multifunctional: dense, multi-vector and sparse retrieval
- Multigranularity: short sentences to long documents
- Wikichat updated to use M3: multi-lingual!

Similarity score between Query and Document



M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, Chen et al, Findings of ACL August 2024

Basic RAG Pipeline



Question

Information Retrieval

Retrieved Data

LLM question answering

Answer

Retrieval must return the right document for accuracy

Quiz: What’s Wrong with this Conversation?

Time of the conversation: March 2023;
Retrieved data1:
‘House of the Dragon’ won the Golden Globe Best Television Series in January 2023.

Retrieved data2:
‘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: Context of Retrieved Information vs. Now

- LLMs are trained with all past knowledge all at once
- Does not have a good model of the passage of time
- We need to tell it “the current time”

Solution: Provide the Time/place Context

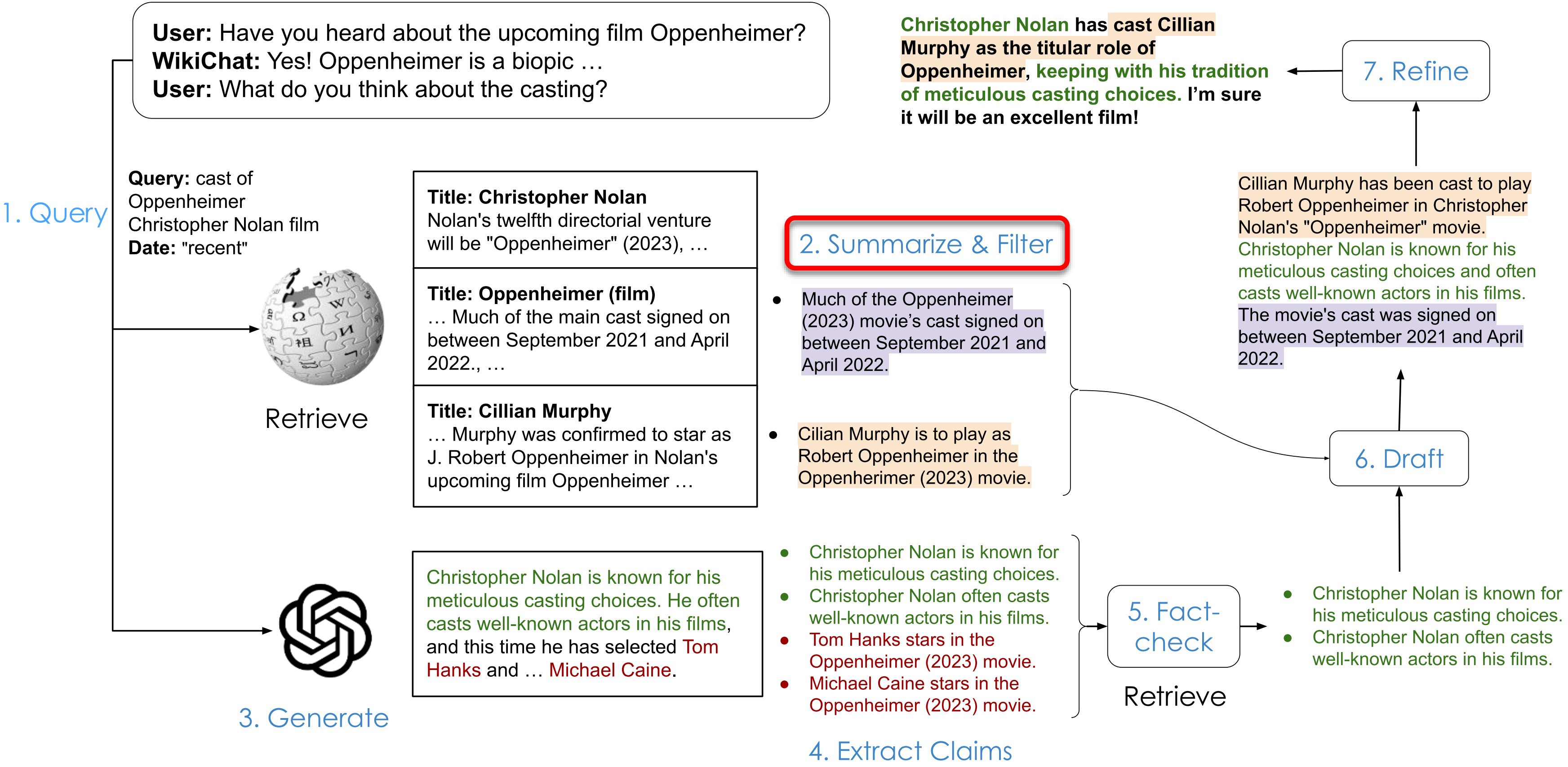
1. Query Prompt

```
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!

The 7-Stage WikiChat Pipeline



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

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

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.

4. Extract Claims

- 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.

7. Refine

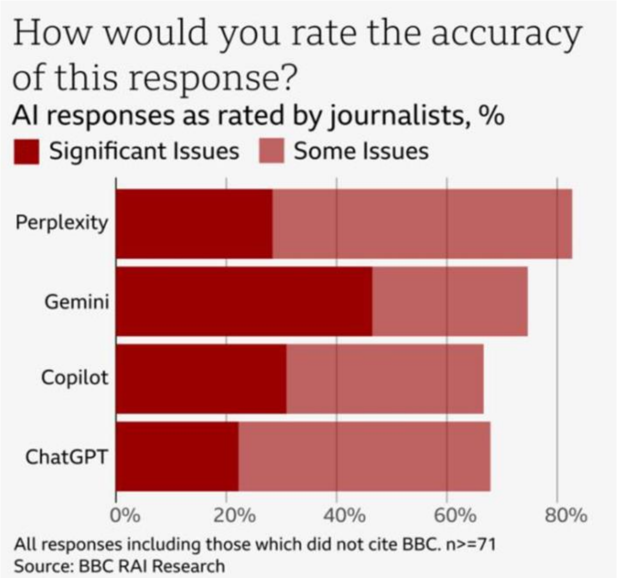
Christopher Nolan has cast Cillian Murphy as the titular role of Oppenheimer, keeping with his tradition of meticulous casting choices. I'm sure it will be an excellent film!

Recall: Commercial RAG (2024)

RAG available in : ChatGPT, Co-Pilot, Gemini, Perplexity

BBC reported in Jan 2025:

- 51% of AI answers to news problems have significant issues.
- 19% of AI answers which cited BBC content introduced factual errors.
- 13% of the quotes sourced from BBC articles were either altered or didn't actually exist in that article.



How would you rate the accuracy of this response?

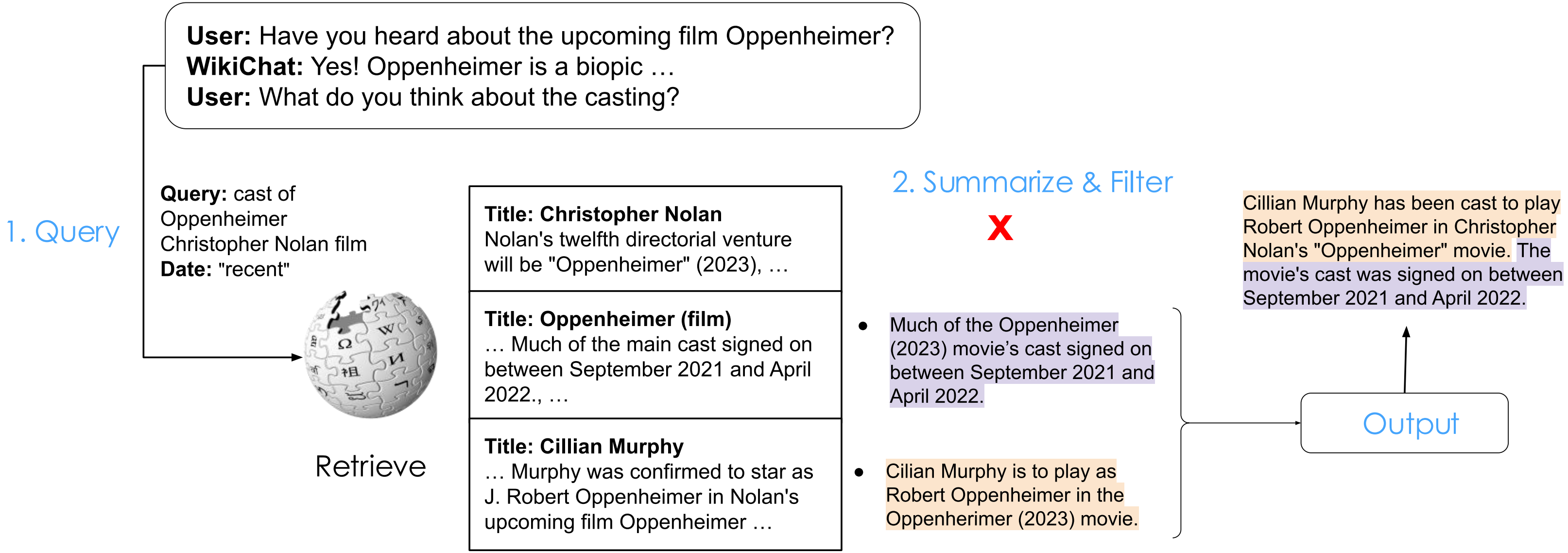
AI responses as rated by journalists, %

All responses including those which did not cite BBC. n>=71
Source: BBC RAI Research

Representation of BBC News content in AI Assistants. 2025.

STAGE 2: ANSWERING WITH RETRIEVED INFO

Solution
+ 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.
- Cilian Murphy is to play as Robert Oppenheimer in the Oppenherimer (2023) movie.

Output

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.

STAGE 2: ANSWERING WITH RETRIEVED INFO

Prompt 2: Summarize & Filter

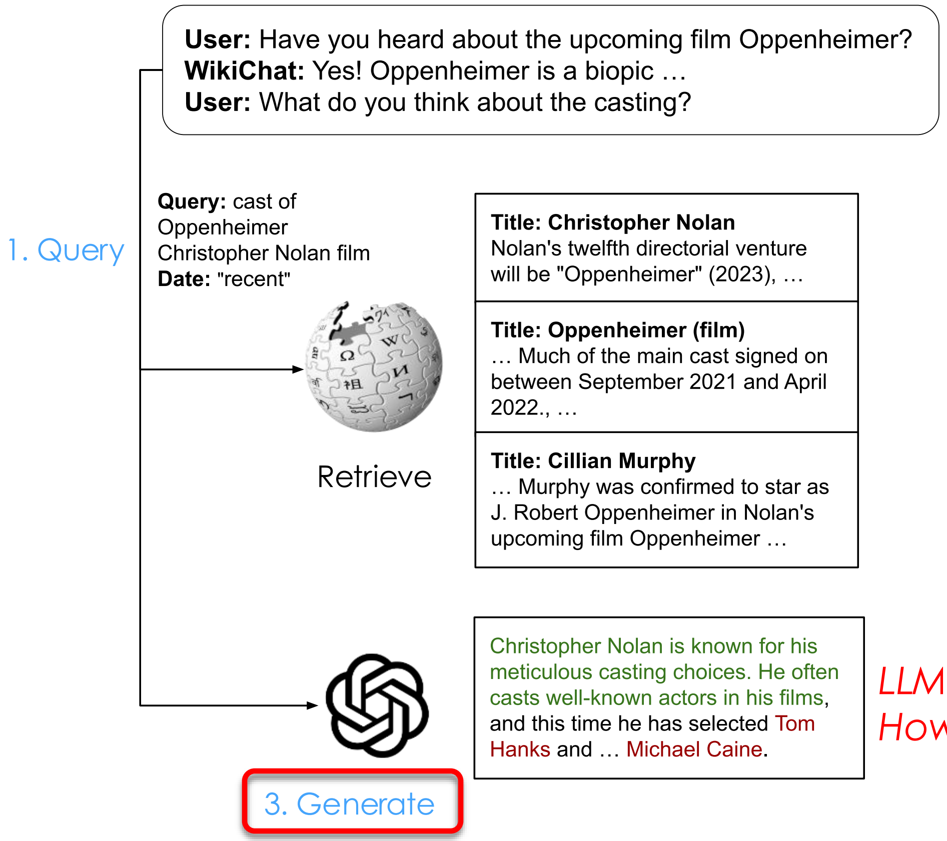
```
You Google different search queries and then extract the relevant parts of the
    articles you find. Today's date is {{ today }}.

Query: ”worst earthquake ever”
Title: January 1934 earthquake in India and Nepal
Article: The 1934 Nepal\u2013India earthquake or 1934 Bihar\u2013Nepal 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 2:13\u00a0pm 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 ), and from Kathmandu in the north
    to Munger in the south (a distance of nearly ).”

Extract verbatum 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.
```

STAGE 3: LLM GENERATION

Problem 3: LLM's Generation May Not be Factual



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 …

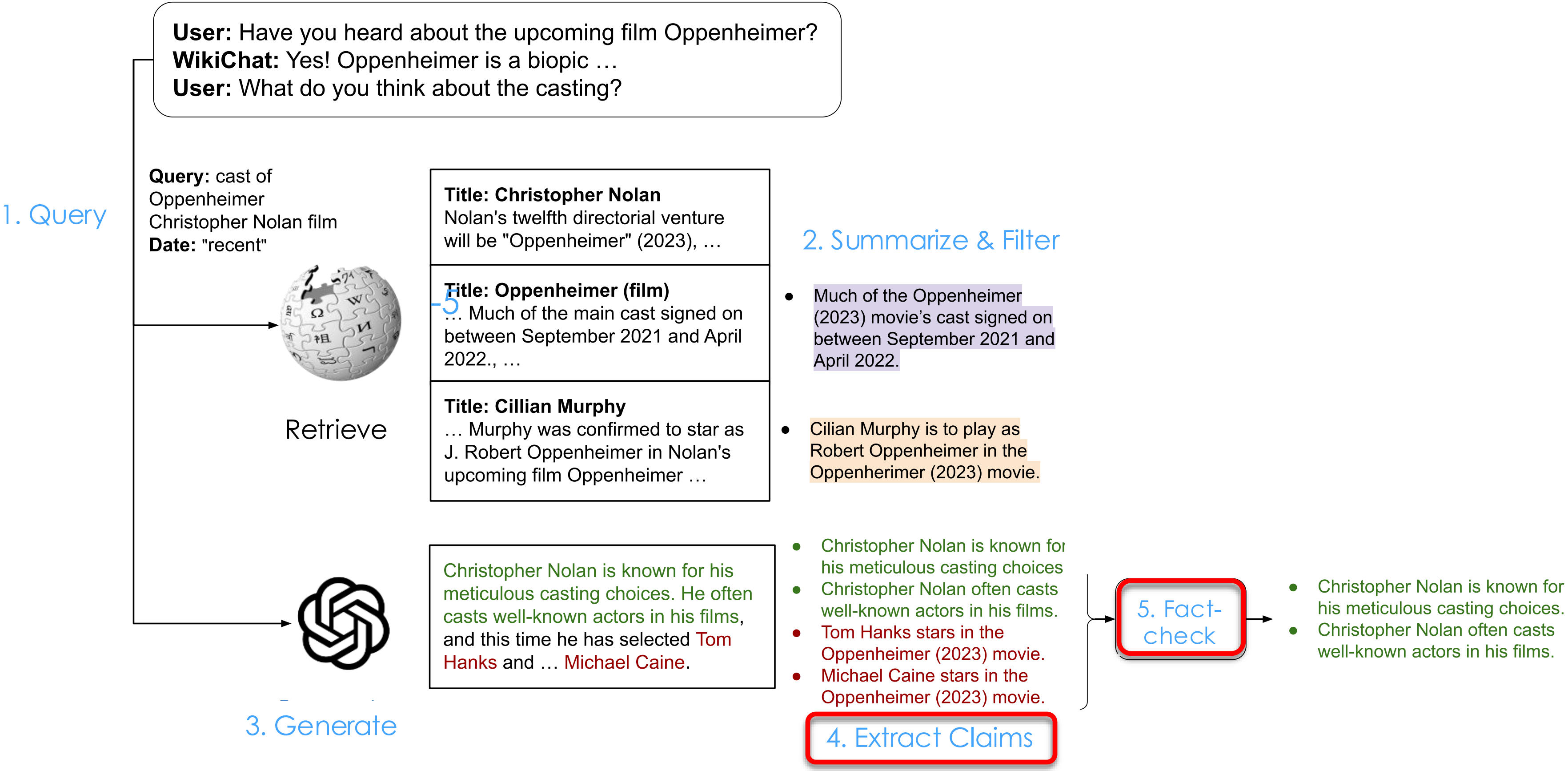
Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected Tom Hanks and … Michael Caine.

3. Generate

LLM may hallucinate.
How to fact-check generation?

STAGE 4-5: FACT-CHECKING LLM GENERATION

+ Extract Claims + 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

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

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

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.

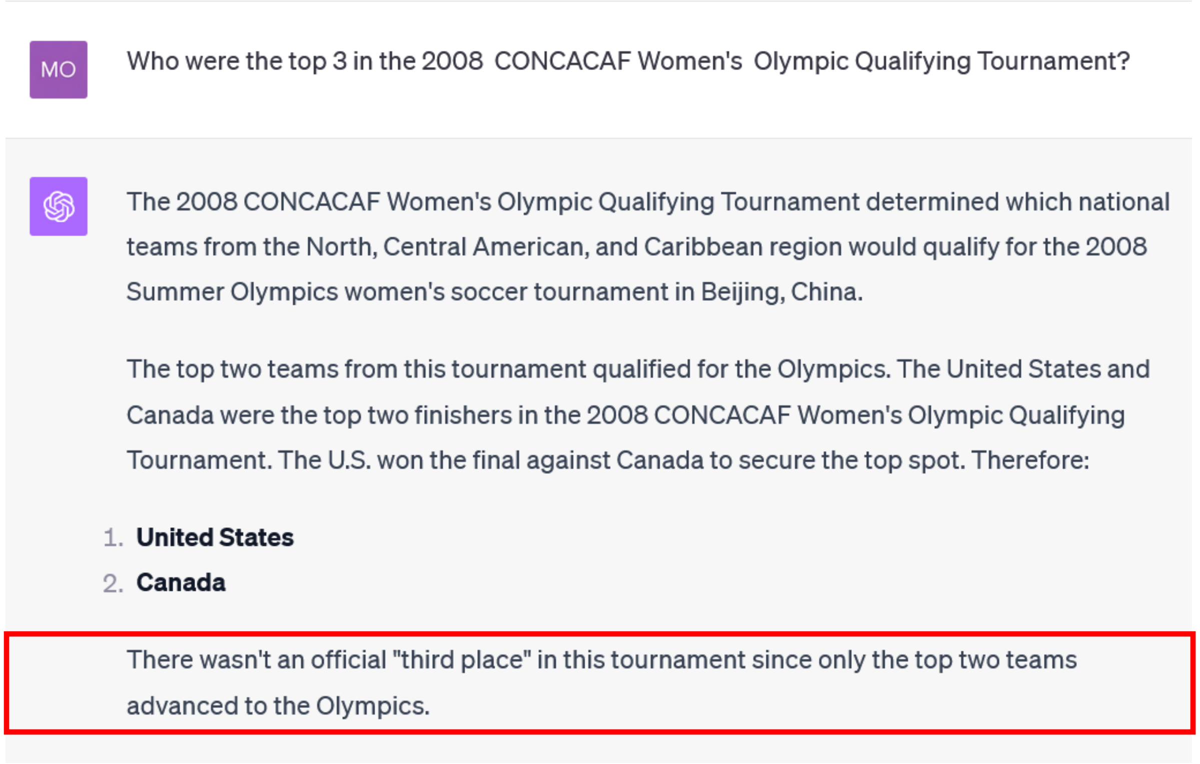
- Christopher Nolan is known for his meticulous casting choices
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

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.

Why is Fact-Checking Hard? Example



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

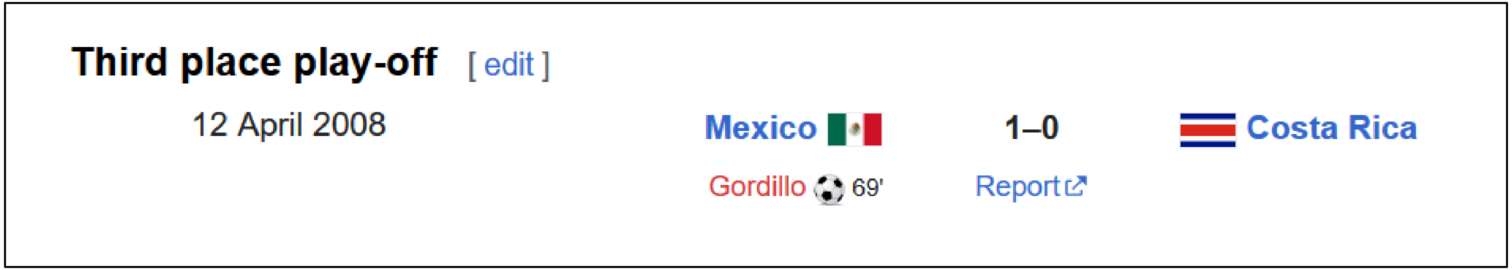
The 2008 CONCACAF Women's Olympic Qualifying Tournament determined which national teams from the North, Central American, and Caribbean region would qualify for the 2008 Summer Olympics women's soccer tournament in Beijing, China.

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

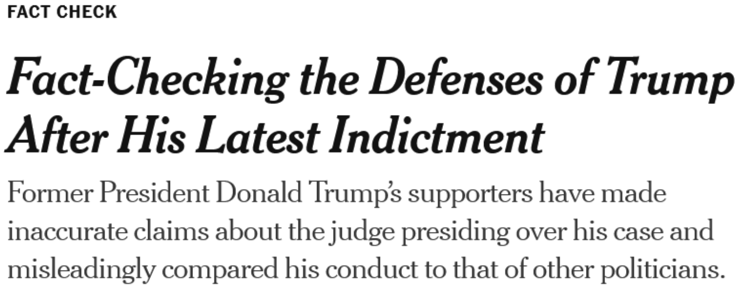
Just a part of the response is false!

Correct: only top 2 teams qualified
But there was indeed a third place match!

STAGE 4-5: FACT-CHECKING LLM GENERATION

People Fact-Checking People for People

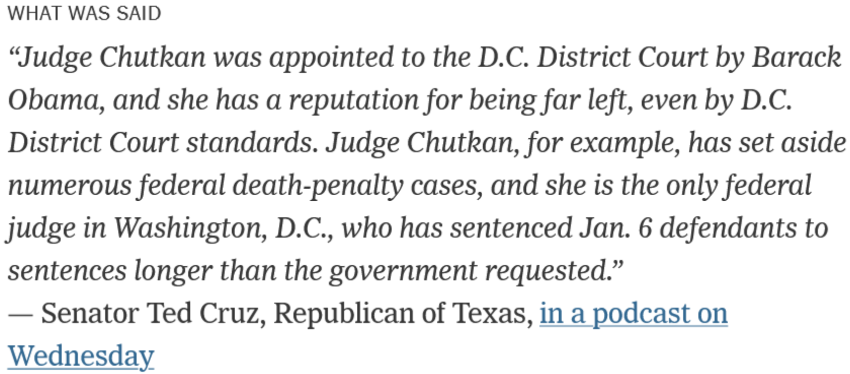
- Journalists fact-check politicians all the time:
  - Identify the claims



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."

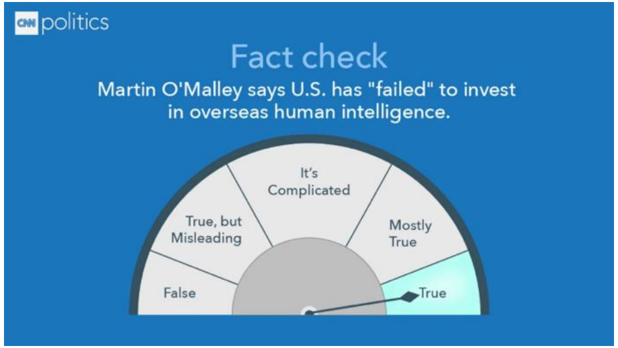
— Senator Ted Cruz, Republican of Texas, in a podcast on Wednesday

https://www.nytimes.com/2023/08/03/us/politics/trump-indictment-fact-check.html

STAGE 4-5: FACT-CHECKING LLM GENERATION

People Fact-Checking People for People

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

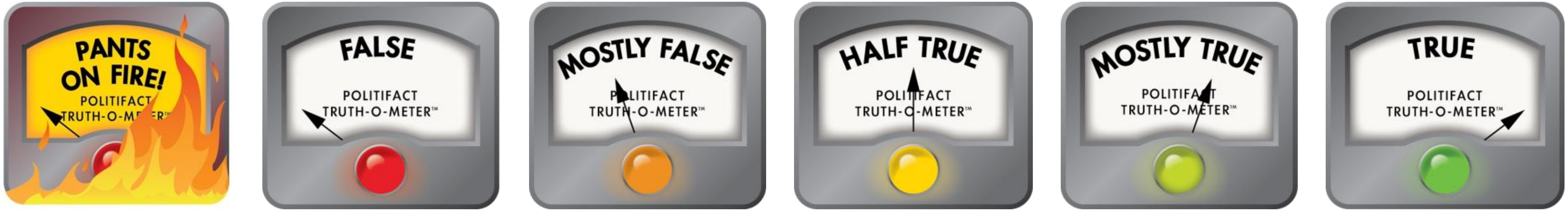


CNN politics

Fact check

Martin O'Malley says U.S. has "failed" to invest in overseas human intelligence.

POLITIFACT



PANTS ON FIRE!
POLITIFACT TRUTH-O-METER™
FALSE
POLITIFACT TRUTH-O-METER™
MOSTLY FALSE
POLITIFACT TRUTH-O-METER™
HALF TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY TRUE
POLITIFACT TRUTH-O-METER™
TRUE
POLITIFACT TRUTH-O-METER™

People Fact-Checking Bots

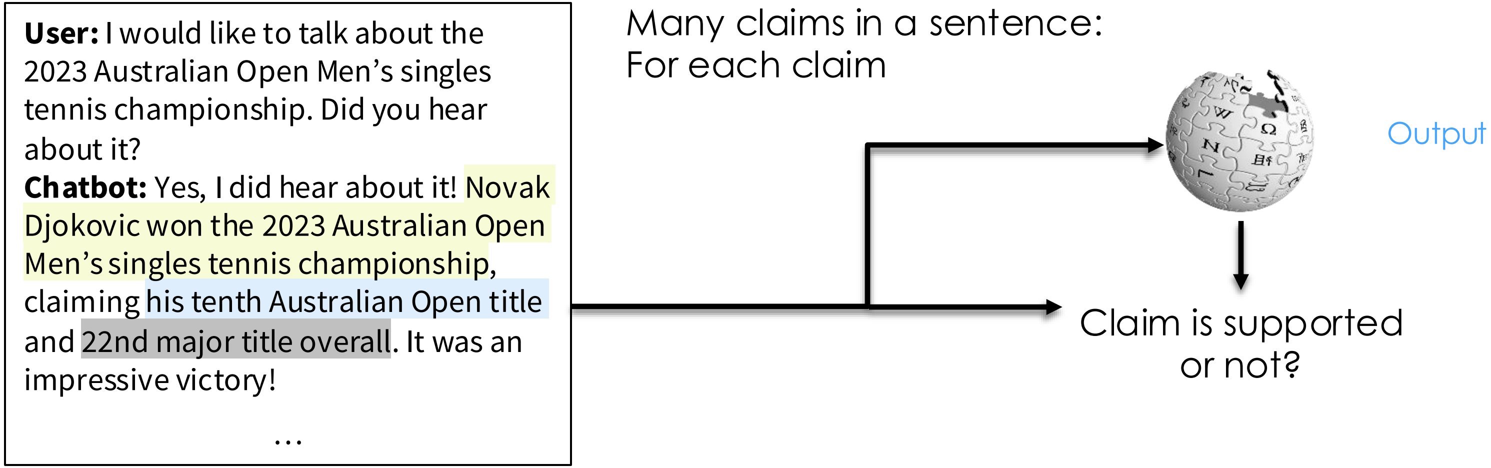
- Reduce the complexity of fact-checking with two assumptions:
  - Factuality is only meaningful
    when measured against a source of ground truth
    - Wikipedia
    - Newspapers trust their own prior publications
  - If claims are atomic, each claim is either true or false
- Factuality: $\dfrac{\#(\text{fact-checked claims})}{\#\text{claims}}$

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!

...

Many claims in a sentence:
For each claim

Performance on Fact-Checking

- 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!

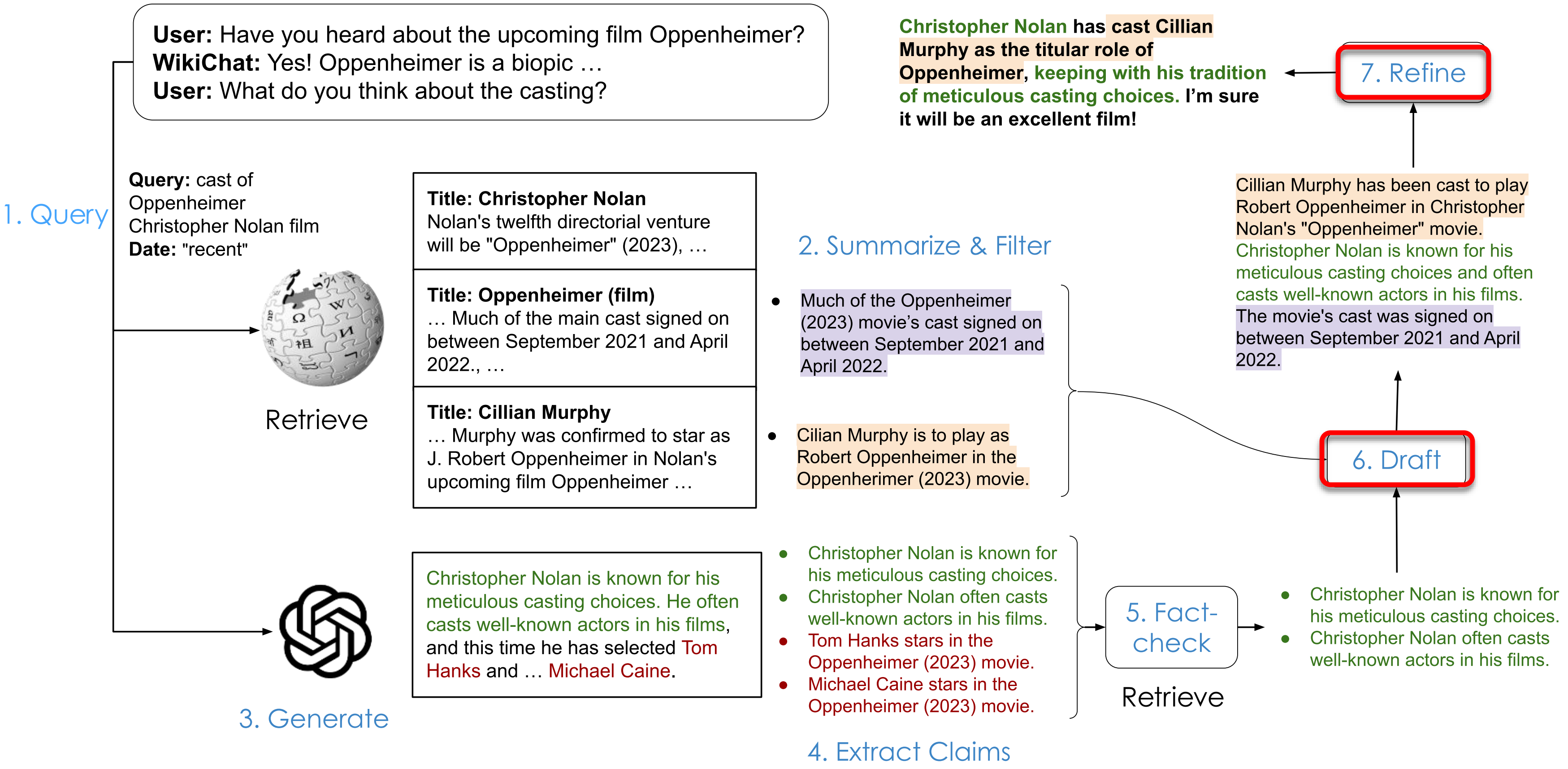
Must evaluate by hand

Automatic Evaluation of Attribution by Large Language Models, Yue et al, arXiv 2023

FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, Min et al, arXiv 2023

Stage 6-7: Combining Retrieval with Generation

+ 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 …

2. Summarize & Filter

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

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.

4. Extract Claims

- 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.

7. Refine

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!

Problem: LLMs Still May Hallucinate

- If we simply ask LLMs to answer based on the combination
  - May still hallucinate!
- Hardest thing for LLMs to say: “I don’t know”
  - It likes to hallucinate
- When no information remains in a topic
  - Draft is prompted to say: ”Sorry, I don’t know”

Summary of GenieChat (7 Prompts)

Traditional (Factuality)

1. Formulate query from input
   - Retrieve documents (M3)
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 (M3)

6. Draft
7. Refine

# LLM calls: 5 + *n* + c
n: # documents retrieved based on user queries
c: # claims generated

RECAP

LLM Engineering

Systematically create well-tested step-by-step algorithms (computational thinking)

1. Design the metrics

2. Start with the most general instruction

3. Debug until you can't break it

- Test with "critical" examples
- Root-cause analysis of LLM errors
- Add/refine the steps
- Add rationale to the prompt;
  add critical examples as an aid

4. Evaluate

- Tests must cover the hardest cases
- Test with LLM-simulated chats
  with hard, varied personalities
- Evaluate with an LLM
  - If LLM disagrees with human
    - Use human evaluation
- Test with real users

Good news: Debugging with critical examples achieves high Wikichat quality.

The key: Look at the errors!

Lecture Outline

- Background of Knowledge Assistants Technology
- Metrics for Knowledge Assistants
- Methodology Paradigm Shift
- Design of Wikichat
- **Evaluation**

Old Data Sets Do Not Work

No simple numeric benchmark to iterate on

- Evaluation conversations in the past 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

How to Eval WikiChat?

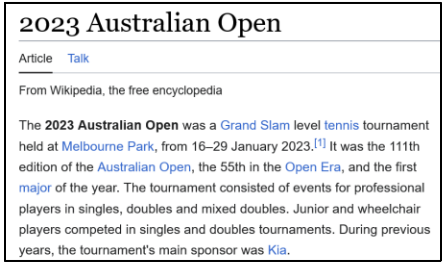
- How to evaluate tail questions?
  - Humans do not know tail topics!
- How to evaluate LLM *conversations*?
  - How to compare the results between systems since their dialogue diverges after the first turn!
  - Cannot get a numeric result to measure success
- How to minimize human effort, cost, and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

Solution: LLMs come to the rescue!

Evaluation Summary

Conversation Generation

Head/Tail/Recent

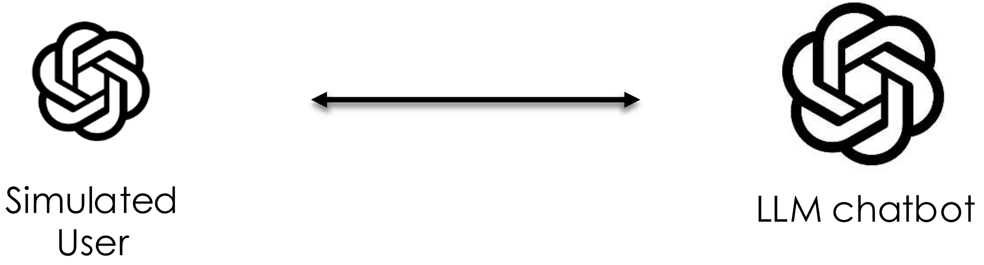


2023 Australian Open

Article Talk

From Wikipedia, the free encyclopedia

The **2023 Australian Open** was a Grand Slam level tennis tournament held at Melbourne Park, from 16–29 January 2023.[1] It was the 111th edition of the Australian Open, the 55th in the Open Era, and the first major of the year. The tournament consisted of events for professional players in singles, doubles and mixed doubles. Junior and wheelchair players competed in singles and doubles tournaments. During previous years, the tournament's main sponsor was Kia.



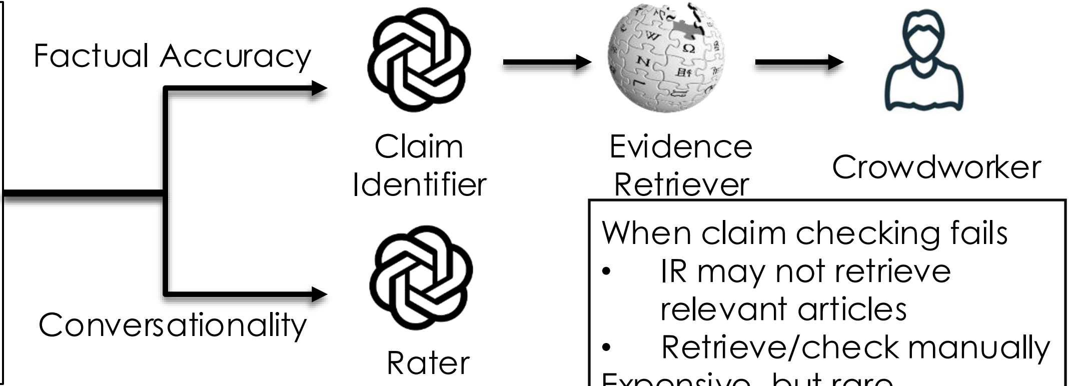
Simulated User

LLM chatbot

Evaluation

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!



Factual Accuracy

Claim Identifier

Evidence Retriever

Crowdworker

Conversationality

Rater

When claim checking fails

- IR may not retrieve relevant articles
- Retrieve/check manually

Expensive, but rare

Implementation of WikiChat (v1)

- 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

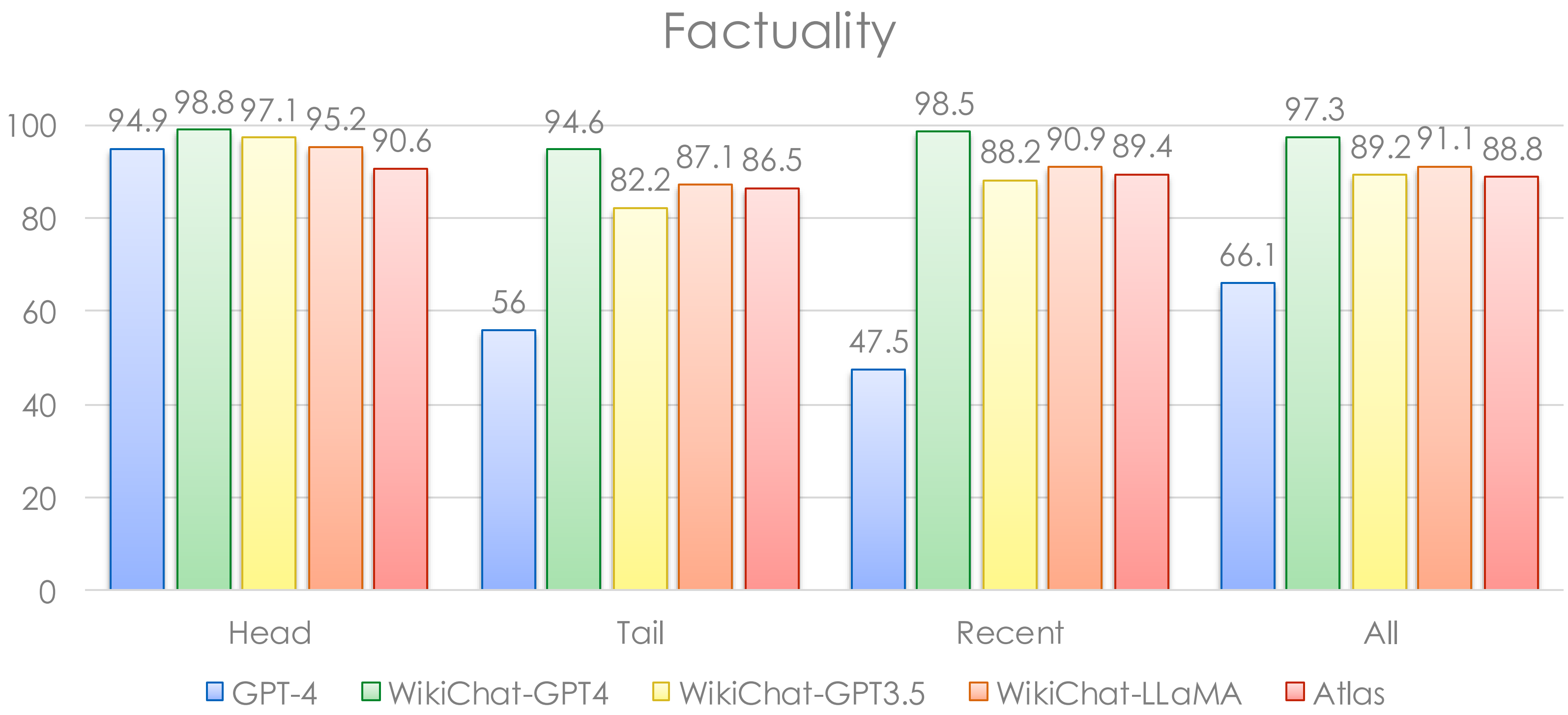
ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction, Santhanam et al, NAACL 2022

Experiment

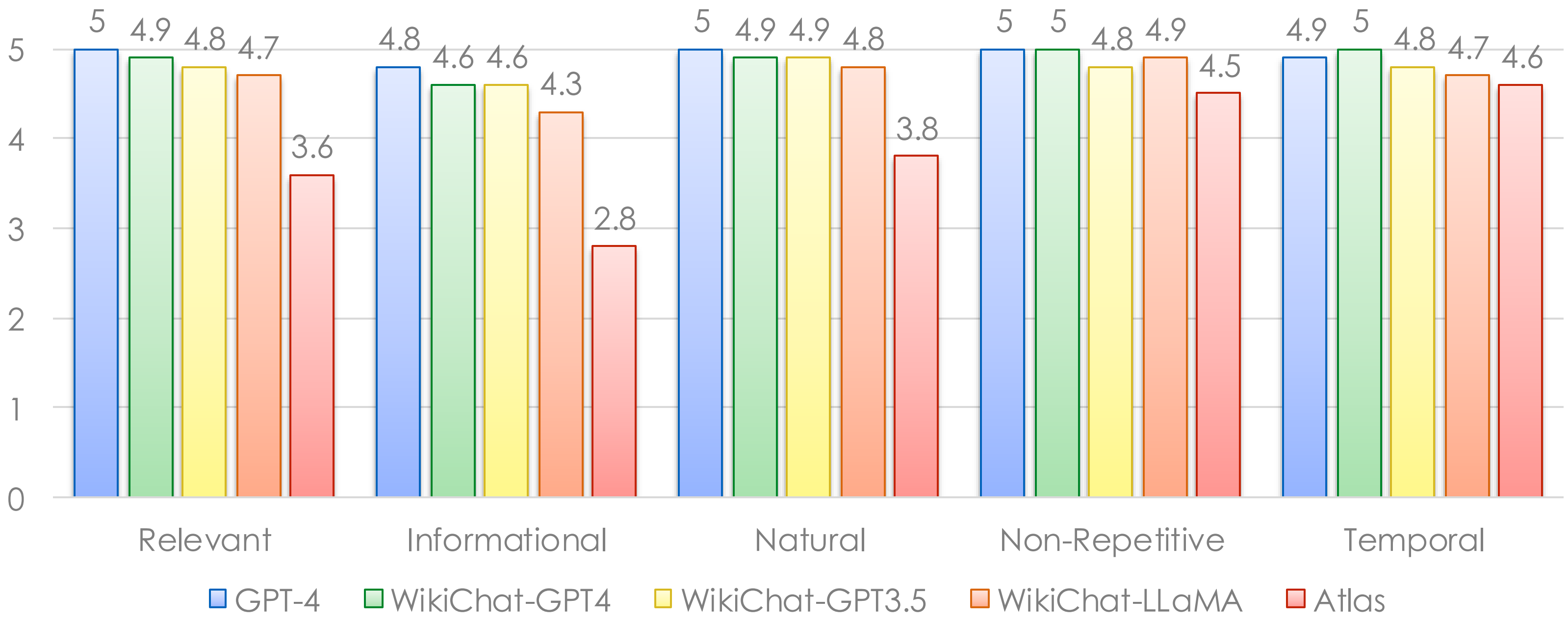
- Baselines
  - Fine-tuned models
    - Atlas: state-of-the-art on many knowledge-grounded tasks
    - 3B parameters
  - LLMs
    - Text-davinci-003 or GPT-4, prompted to be a chatbot

- Each evaluated for 600 dialog turns

Results – Factual Accuracy



Results - Conversationality



Design Analysis: Do we need both IR and LLM?

Information Retrieval and LLM are color-coded

User: Can we talk about Sting, the musician?
WikiChat: Sure! Sting is an English musician and actor, best known as the frontman, songwriter and bassist for the new wave rock band the Police. He has released several solo albums and has won numerous awards, including 17 Grammy Awards and a Golden Globe.

User: […] What do you think makes Sting's music so unique?
WikiChat: Sting's music is often very personal and introspective, and he often uses his music to express his own thoughts and feelings. He also has a great ability to blend different genres and create unique sounds.

User: Do you think his birthplace has influenced his music?

WikiChat: Yes, Sting has been heavily influenced by his birthplace of Newcastle upon Tyne, England. As a child, he was inspired by the Queen waving at him from a Rolls-Royce, which diverted him from the shipyard prospect towards a more glamorous life. He was also influenced by the music he heard in local nightclubs he visited such as Club A'Gogo.

Analyzing the Design

- 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)

Quiz:

Are We Done Analyzing WikiChat?

Real User Evaluation

- Real user evaluation was seldom performed before LLM because of poor accuracy
- Most important metric!

User study: User reads the first sentence of a new Wikipedia page

Model	User Rating (out of 5)	Factuality
GPT-4	3.4	42.9%
WikiChat using GPT-4	3.8	97.9%

GPT-4: Users are not even aware that over half of the statements are false

Extra Notes

- **Speed**
  - We showed distilling to Llama could speed up the process
  - OpenAI O1-mini is fast enough without fine-tuning
- **Multi-lingual Knowledge**
  - Requires multi-lingual entity resolution
    e.g. Hsiang-Tsung Kung and 孔祥重 is the same person (Wikidata)

Conclusion

- WikiChat
  - Combines the best of both worlds (LLM generation and IR)
  - Is as conversational as LLMs
  - Is far more factual than all baselines – useful for all tasks

Is this Applicable to Persuasion Bots?

- **Is this applicable to persuasive tasks?**
  - **Yes**, otherwise it hallucinates success stories! "There is a girl called Maria …."
- **Is this sufficient?**
  - **No!** all the success stories are removed! Not very persuasive!
  - Solution: Search with the intent of the fake story:
    - "Tell an impact story about an individual"
- **Is the evaluation method for conversationality applicable?**
  - **No!** LLM evaluation on persuasiveness is not correlated with human evaluation
  - Solution: Manual evaluation

Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval
Kazuaki Furumai, Roberto Legaspi, Julio Vizcarra, Yudai Yamazaki, Yasutaka Nishimura, Sina J. Semnani, Kazushi Ikeda, Weiyan Shi, Monica S. Lam
In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Miami, Florida, November 12 –16, 2024.

RECAP

LLM Engineering

Systematically create well-tested step-by-step algorithms (computational thinking)

1. Design the metrics

2. Start with the most general instruction

3. Debug until you can't break it

- Test with "critical" examples
- Root-cause analysis of LLM errors
- Add/refine the steps
- Add rationale to the prompt;
  add critical examples as an aid

4. Evaluate

- Tests must cover the hardest cases
- Test with LLM-simulated chats
  with hard, varied personalities
- Evaluate with an LLM
  - If LLM disagrees with human
    - Use human evaluation
- Test with real users

Good news: Debugging with critical examples achieves high Wikichat quality.

The key: Look at the errors!