Generative AI and LLMs in Search and Recommendation engines

Andrei Lopatenko, PhD
VP AI and Engineering

What is this deck about

In this presentation, we will explore the benefits of generative AI for search and recommender engines.

We'll start with an overview and then delve deeper into specific approaches that have proven valuable in the industry, such as FreshLLMs/FreshPrompts and LLMRank.

Additionally, we will discuss LLM evaluation methods that are particularly useful for search and recommender engines.

What is this deck about

Strategic Focus

This discussion extends beyond immediate business metrics such as conversion rates, transaction volumes, and cross-merchandising value—which, as evidenced by industry practices, are comparatively straightforward to enhance.

Instead, our focus is on cultivating long-term value. For customers, this means enabling better, more impactful decisions that save time and money, and enhance value post-purchase. For the company, it involves fostering sustained engagement with customers, ensuring a lasting relationship that extends well beyond initial transactions.

Current State of AI in Search

Generative AI has demonstrated tremendous value for search, as evidenced by recent launches from OpenAI, Google, and Microsoft. In this presentation, we will explore the fundamental technologies behind these advancements and discuss how generative AI can enhance the search and recommender engines within your company.

In this presentation, we will explore the potential contributions of Large Language Models (LLMs) and Generative AI to enhance value for customers looking to make purchases, bookings, or rentals. We'll focus on how these technologies can not only refine the search process but also fundamentally improve the decision-making experience for users.

About Andrei Lopatenko

- I bring over two decades of expertise in developing search and recommendation engines that have served billions of users at major corporations including Google, Apple (specifically Maps, App Store, iTunes), Walmart, eBay, and Zillow.
- Since 2008, I have been involved with language models, expanding my expertise to the BERT family in 2019 and more recently to large language models (LLMs).
- PhD in Computer Science (The University of Manchester, UK), 1600 citations to my publications, 29 patents
- And please, check my LLM Evaluation Compendium https://github.com/alopatenko/LLMEvaluation

About Srijan Kumar

- Co-founder and CEO of Lighthouz AI, a multiagent AI platform for Gen AI development and evaluations
- Assistant Professor at Georgia Tech working on AI, NLP, Graph Neural Networks, Recommender Systems
- 60+ publications, 5500+ citations
- 10+ years in AI/ML
- Previously worked at Stanford, Google, Univ. Maryland

Search

Andrei Lopatenko | April 30, 2024

Supporting and anticipating the evolution of customer needs to drive future purchases.

Customer Support

Item Support , QA, Documentation ,

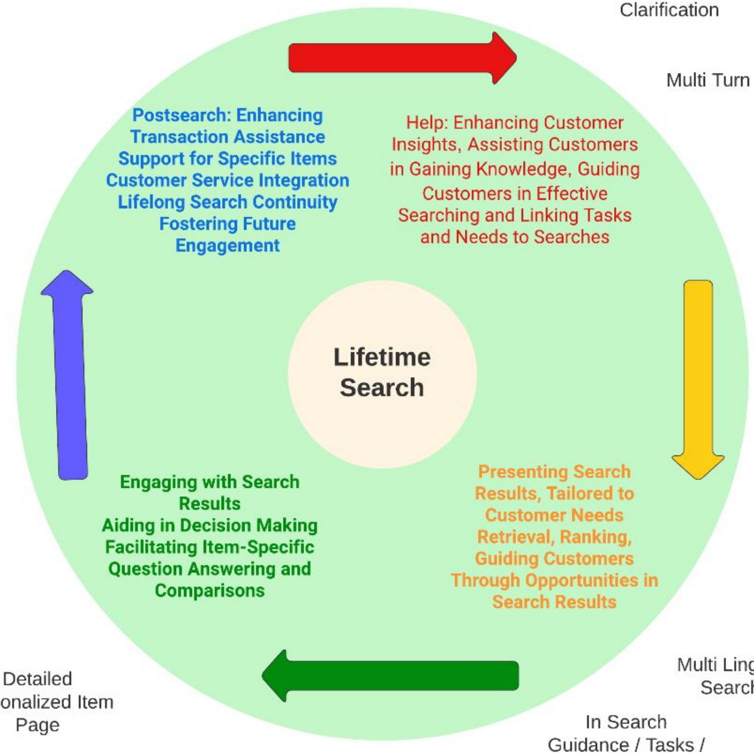
Transaction Support

Item Question Answering ,

Comparison, Similarity, Complimentarity Adviser

Expert Decision Making Helper

Detailed Personalized Item Page



Lifetime Search

Help: Enhancing Customer Insights, Assisting Customers in Gaining Knowledge, Guiding Customers in Effective Searching and Linking Tasks and Needs to Searches

Presenting Search Results, Tailored to Customer Needs Retrieval, Ranking, Guiding Customers Through Opportunities in Search Results

Engaging with Search Results Aiding in Decision Making Facilitating Item-Specific Question Answering and Comparisons

Postsearch: Enhancing Transaction Assistance Support for Specific Items Customer Service Integration Lifelong Search Continuity Fostering Future Engagement

Proactive engagement with users to know their intents

Expert Search Personas

Search Clarification

In Domain Question Answering

Multi Turn Search

Task Solver

Sequential User Modeling

Multi Modal Search

Retriebal and Ranking (RR): Embedding Representations and new RR models

Generated Search Page

Generated Whole Page

Multi Lingual Search

In Search Recommendation

In Search Guidance / Tasks / Alternatives

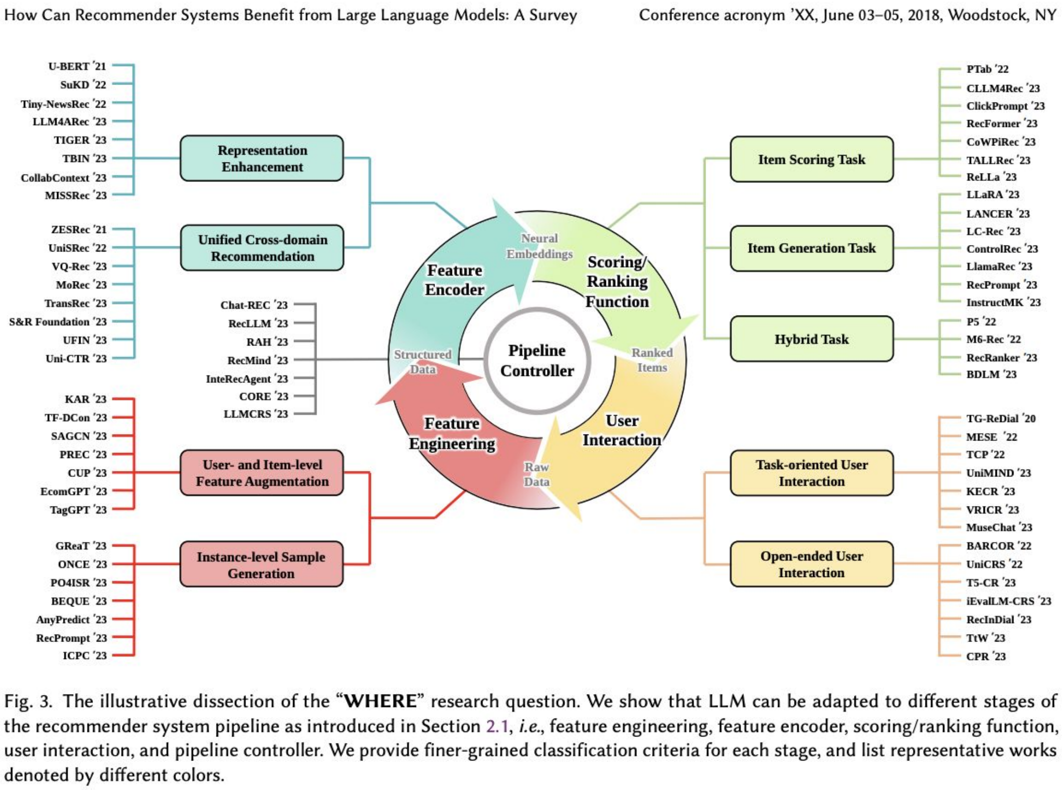
Multi Turn Memory

Generative AI in Search

1. We prioritize tasks to deliver substantial customer and business value and have been identified as important to address across various search engines.
2. We filter tasks that are now feasible to be solved using generative AI, or where generative AI can significantly enhance the solution.
3.
4.

Text

Gen AI for Recommender Systems



How Can Recommender Systems Benefit from Large Language Models: A Survey Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

Fig. 3. The illustrative dissection of the "**WHERE**" research question. We show that LLM can be adapted to different stages of the recommender system pipeline as introduced in Section 2.1, *i.e.*, feature engineering, feature encoder, scoring/ranking function, user interaction, and pipeline controller. We provide finer-grained classification criteria for each stage, and list representative works denoted by different colors.

Fig 3 From

https://arxiv.org/abs/2306.05817

Search & LLMs

New paradigms

- Generating vs listing
- Conversational search (multi-turn)
- Search integrated into the whole customer journey
- Expert knowledge, question answering and decision making
- In-search task solving

Model improvements for better experience but within the existing paradigm (silent heroes)

- Better retrieval and ranking and query understanding
- Multi modality
- Deeper understanding of customers and items and queries
- Better navigation elements (facets, navigational panels)
- Evaluation

FreshLLMs / FreshPrompt

Google / OpenAI / Univ of Mass. Amherst Oct 2023

https://arxiv.org/abs/2310.03214

Perplexity AI launched the system inspired by FreshLLMs one month after

https://www.perplexity.ai/hub/blog/introducing-pplx-online-llms

Experiments were conducted in April 2023

Gemini, Perplexity.AI, You.com, Contextual.AI are inspired by this approach

FreshLLMs / FreshPrompt

There are multiple aspects of quality of answer for search (including search based on LLMs): Factuality, Authority, Helpfulness, Freshness etc

This approach is focused on Freshness

Different type of queries require different freshness of the answer: in FreshLLMs paper, the authors talk about never changing, slow changing, fast changing and false premise queries.

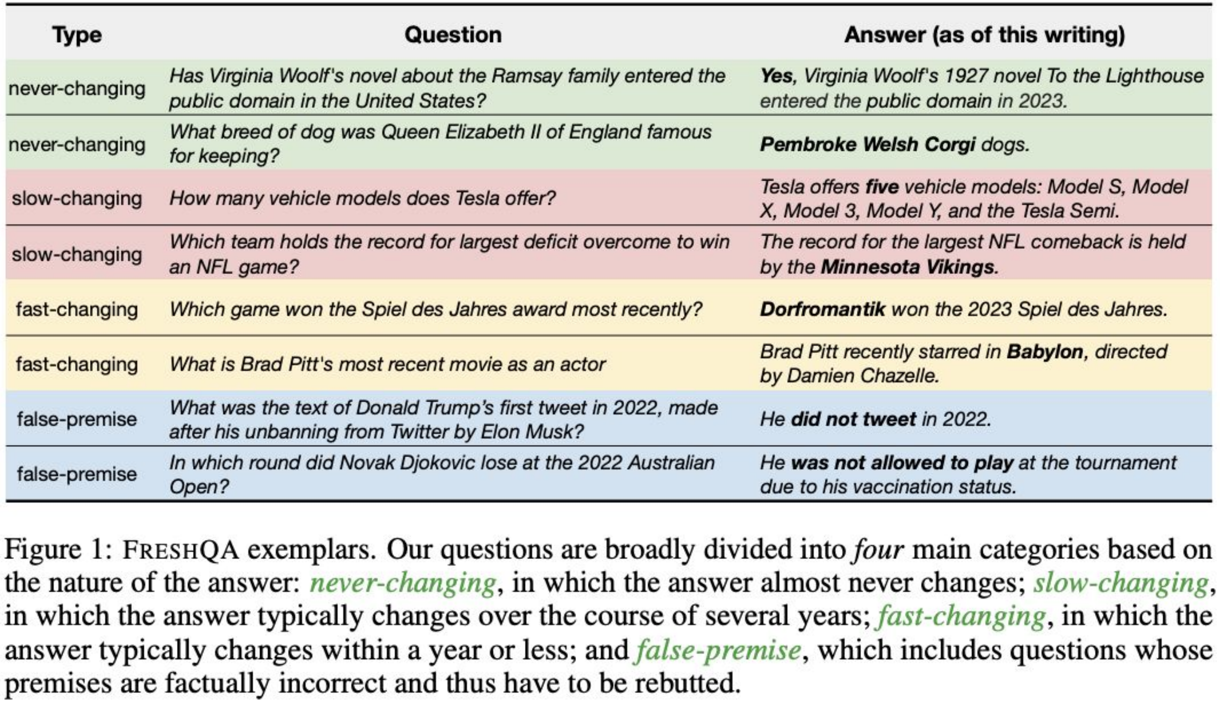
It’s an over simplified approach, there are more types of temporal behavior in search, but this model helps to solve many freshness problems

Let’s call it Freshness V0

FreshLLMs / FreshPrompt

From

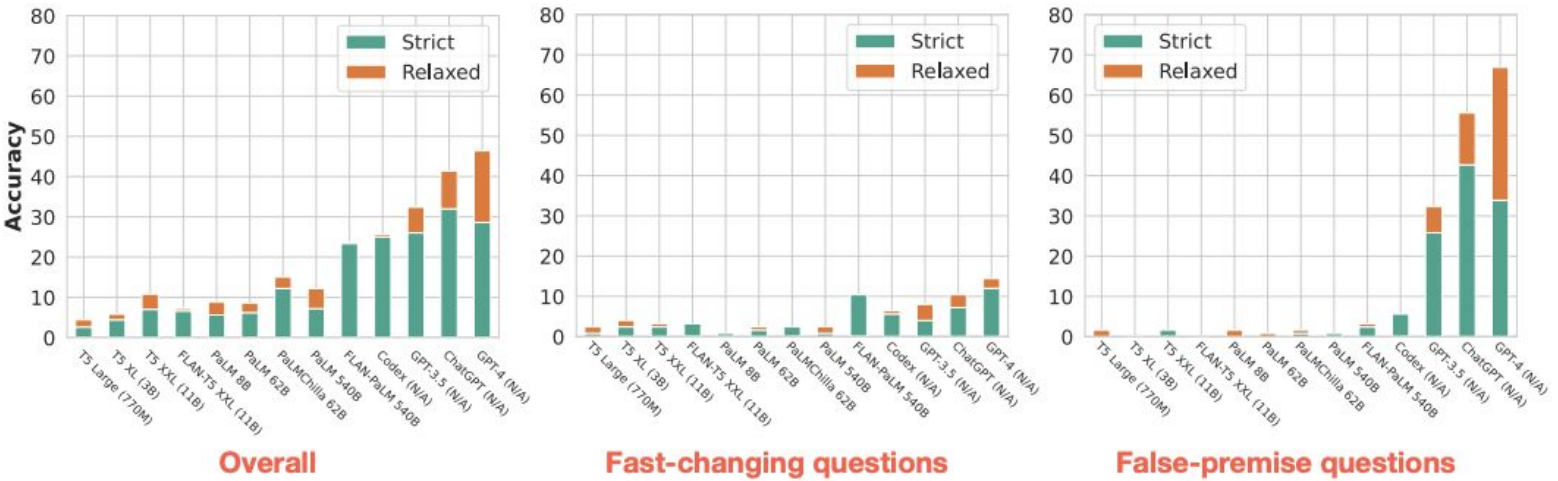
FreshLLMs



Type	Question	Answer (as of this writing)
never-changing	*Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?*	***Yes**, Virginia Woolf's 1927 novel To the Lighthouse entered the public domain in 2023.*
never-changing	*What breed of dog was Queen Elizabeth II of England famous for keeping?*	***Pembroke Welsh Corgi** dogs.*
slow-changing	*How many vehicle models does Tesla offer?*	*Tesla offers **five** vehicle models: Model S, Model X, Model 3, Model Y, and the Tesla Semi.*
slow-changing	*Which team holds the record for largest deficit overcome to win an NFL game?*	*The record for the largest NFL comeback is held by the **Minnesota Vikings**.*
fast-changing	*Which game won the Spiel des Jahres award most recently?*	***Dorfromantik** won the 2023 Spiel des Jahres.*
fast-changing	*What is Brad Pitt's most recent movie as an actor*	*Brad Pitt recently starred in **Babylon**, directed by Damien Chazelle.*
false-premise	*What was the text of Donald Trump's first tweet in 2022, made after his unbanning from Twitter by Elon Musk?*	*He **did not tweet** in 2022.*
false-premise	*In which round did Novak Djokovic lose at the 2022 Australian Open?*	*He **was not allowed to play** at the tournament due to his vaccination status.*

Figure 1: FRESHQA exemplars. Our questions are broadly divided into *four* main categories based on the nature of the answer: *never-changing*, in which the answer almost never changes; *slow-changing*, in which the answer typically changes over the course of several years; *fast-changing*, in which the answer typically changes within a year or less; and *false-premise*, which includes questions whose premises are factually incorrect and thus have to be rebutted.

FreshLLMs/ Plain LLM Question Answering



Overall

Fast-changing questions

False-premise questions

Figure 2: Accuracy of different LLMs on FreshQA under Relaxed and Strict (no hallucination) evaluations. Models benchmarked on the same date of April 26, 2023. *All* models (regardless of model size) struggle on questions that involve *fast-changing* knowledge and *false premises*.

FreshPrompt

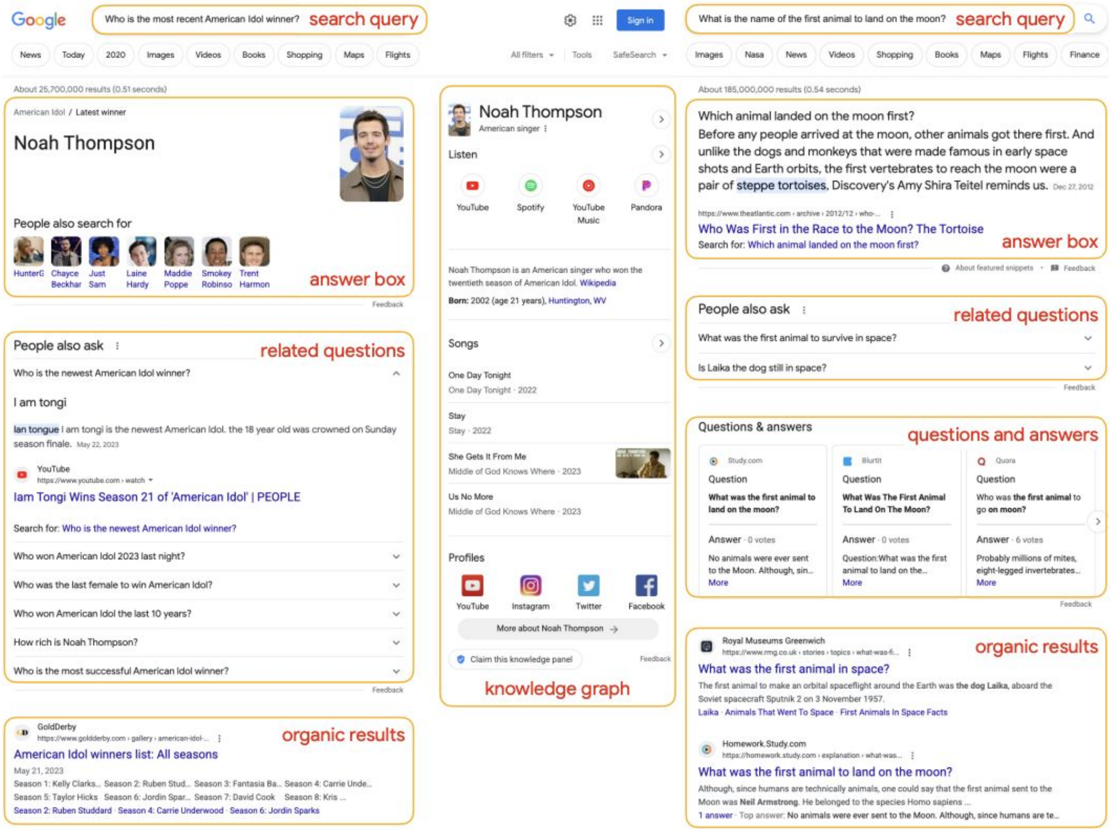
Algorithm: (high level)

1 use search engine to retrieve relevant information

2 use LLM to extract the answer

1 in their case, retrieve all information boxes from Google, get all results as (source, date, title,snippet, highlighted words)

At this stage, there are several aggregation subroutines such as sort the responses from the oldest to newest (by date) to help the model to get more fresh results



Who is the most recent American Idol winner?

About 25,700,000 results (0.51 seconds)

American Idol / Latest winner

Noah Thompson

People also search for

HunterG · Chayce Beckhar · Just Sam · Laine Hardy · Maddie Poppe · Smokey Robinso · Trent Harmon

People also ask

Who is the newest American Idol winner?

I am tongi

Ian tongue I am tongi is the newest American Idol. the 18 year old was crowned on Sunday season finale. May 22, 2023

YouTube
https://www.youtube.com › watch

Iam Tongi Wins Season 21 of 'American Idol' | PEOPLE

Search for: Who is the newest American Idol winner?

Who won American Idol 2023 last night?

Who was the last female to win American Idol?

Who won American Idol the last 10 years?

How rich is Noah Thompson?

Who is the most successful American Idol winner?

GoldDerby
https://www.goldderby.com › gallery › american-idol-...

American Idol winners list: All seasons

May 21, 2023

Season 1: Kelly Clarks... Season 2: Ruben Stud... Season 3: Fantasia Ba... Season 4: Carrie Unde... Season 5: Taylor Hicks Season 6: Jordin Spar... Season 7: David Cook Season 8: Kris ...

Season 2: Ruben Studdard · Season 4: Carrie Underwood · Season 6: Jordin Sparks

Noah Thompson
American singer

Listen

YouTube · Spotify · YouTube Music · Pandora

Noah Thompson is an American singer who won the twentieth season of American Idol. Wikipedia

Born: 2002 (age 21 years), Huntington, WV

Songs

One Day Tonight
One Day Tonight · 2022

Stay
Stay · 2022

She Gets It From Me
Middle of God Knows Where · 2023

Us No More
Middle of God Knows Where · 2023

Profiles

YouTube · Instagram · Twitter · Facebook

More about Noah Thompson →

Claim this knowledge panel

What is the name of the first animal to land on the moon?

About 185,000,000 results (0.54 seconds)

Which animal landed on the moon first?

Before any people arrived at the moon, other animals got there first. And unlike the dogs and monkeys that were made famous in early space shots and Earth orbits, the first vertebrates to reach the moon were a pair of steppe tortoises, Discovery's Amy Shira Teitel reminds us. Dec 27, 2012

https://www.theatlantic.com › archive › 2012/12 › who-...

Who Was First in the Race to the Moon? The Tortoise

Search for: Which animal landed on the moon first?

About featured snippets · Feedback

People also ask

What was the first animal to survive in space?

Is Laika the dog still in space?

Questions & answers

Study.com
Question
What was the first animal to land on the moon?
Answer · 0 votes
No animals were ever sent to the Moon. Although, sin...
More

Blurtit
Question
What Was The First Animal To Land On The Moon?
Answer · 0 votes
Question:What was the first animal to land on the...
More

Quora
Question
Who was **the first animal** to go **on moon?**
Answer · 6 votes
Probably millions of mites, eight-legged invertebrates...
More

Royal Museums Greenwich
https://www.rmg.co.uk › stories › topics › what-was-fi...

What was the first animal in space?

The first animal to make an orbital spaceflight around the Earth was **the dog Laika**, aboard the Soviet spacecraft Sputnik 2 on 3 November 1957.

Laika · Animals That Went To Space · First Animals In Space Facts

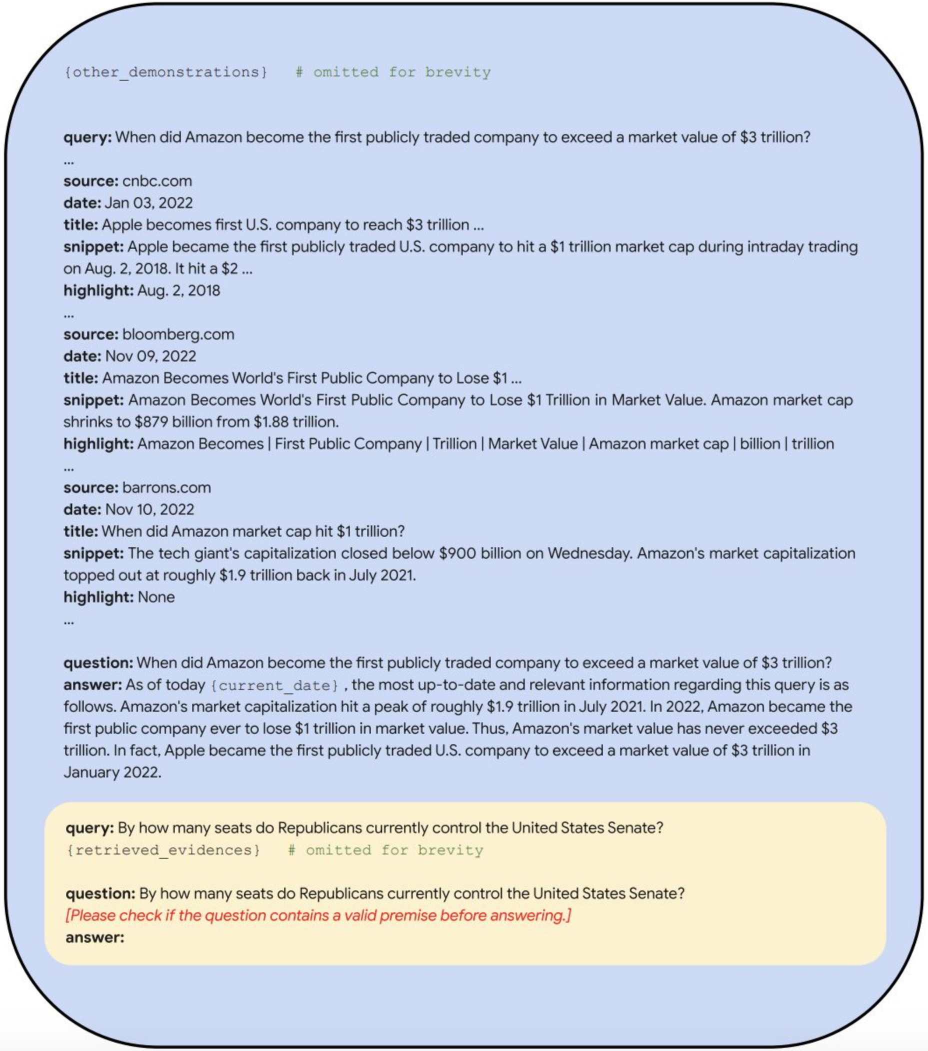
Homework.Study.com
https://homework.study.com › explanation › what-was...

What was the first animal to land on the moon?

Although, since humans are technically animals, one could say that the first animal sent to the Moon was **Neil Armstrong**. He belonged to the species Homo sapiens ...

1 answer · Top answer: No animals were ever sent to the Moon. Although, since humans are te...

Figure 9: GOOGLE SEARCH produces different types of search results for given a query, including the *answer box*, *organic results*, and other useful information, such as the *knowledge graph*, *questions and answers* from crowdsourced QA platforms, and *related questions* that search users also ask. Each of these results contains an associated *text snippet* along with other information, such as *source webpage*, *date*, *title*, and *highlighted words*.



`{other_demonstrations}   # omitted for brevity`

query: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
...
source: cnbc.com
date: Jan 03, 2022
title: Apple becomes first U.S. company to reach $3 trillion ...
snippet: Apple became the first publicly traded U.S. company to hit a $1 trillion market cap during intraday trading on Aug. 2, 2018. It hit a $2 ...
highlight: Aug. 2, 2018
...
source: bloomberg.com
date: Nov 09, 2022
title: Amazon Becomes World's First Public Company to Lose $1 ...
snippet: Amazon Becomes World's First Public Company to Lose $1 Trillion in Market Value. Amazon market cap shrinks to $879 billion from $1.88 trillion.
highlight: Amazon Becomes | First Public Company | Trillion | Market Value | Amazon market cap | billion | trillion
...
source: barrons.com
date: Nov 10, 2022
title: When did Amazon market cap hit $1 trillion?
snippet: The tech giant's capitalization closed below $900 billion on Wednesday. Amazon's market capitalization topped out at roughly $1.9 trillion back in July 2021.
highlight: None
...

question: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
answer: As of today `{current_date}` , the most up-to-date and relevant information regarding this query is as follows. Amazon's market capitalization hit a peak of roughly $1.9 trillion in July 2021. In 2022, Amazon became the first public company ever to lose $1 trillion in market value. Thus, Amazon's market value has never exceeded $3 trillion. In fact, Apple became the first publicly traded U.S. company to exceed a market value of $3 trillion in January 2022.

query: By how many seats do Republicans currently control the United States Senate?
`{retrieved_evidences}   # omitted for brevity`

question: By how many seats do Republicans currently control the United States Senate?
[Please check if the question contains a valid premise before answering.]
answer:

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
comparison against baselines												
Google Search (N/A)	UTD	39.6	48.9	32.0	46.4	68.3	67.4	37.9	55.6	32.4	11.3	9.7
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + Self-Ask (N/A)	UTD	41.6	51.1	36.8	43.2	73.0	73.8	37.4	52.2	48.1	12.9	17.2
GPT-3.5 + FreshPrompt	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
PPLX.AI (N/A)	UTD	52.2	57.2	38.4	53.6	79.4	73.0	47.7	63.8	40.7	37.1	38.7
GPT-4 (N/A)	2021^{+}	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 + Self-Ask (N/A)	UTD	47.8	47.1	39.2	46.4	55.6	51.8	44.3	43.7	55.6	50.0	61.3
GPT-4 + FreshPrompt	UTD	**75.6**	**77.1**	**59.2**	**77.6**	**94.4**	**88.7**	**70.2**	**81.3**	**66.7**	**71.0**	**77.4**
sensitivity and ablation studies												
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + FreshPrompt	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
w/ premise check	UTD	35.2	27.1	14.4	28.0	38.9	36.2	21.7	31.0	17.6	59.7	67.7
GPT-4 (N/A)	2021^{+}	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 w/ snippets only & search order	UTD	74.0	75.5	56.8	75.2	94.4	87.9	68.1	79.9	64.8	69.4	77.4
GPT-4 w/ snippets only & time order	UTD	74.8	75.5	58.4	74.4	93.7	87.9	68.1	79.9	64.8	72.6	**82.8**
GPT-4 w/ snippets only & random order	UTD	72.4	73.7	56.8	69.6	94.4	87.9	65.1	78.4	62.0	68.5	76.3
GPT-4 + FreshPrompt	UTD	75.6	77.1	59.2	77.6	94.4	**88.7**	70.2	81.3	66.7	71.0	77.4
w/ premise check	UTD	75.0	74.2	56.8	76.0	89.7	85.1	67.7	79.5	61.1	**77.4**	79.6
w/o answer box	UTD	74.2	74.7	57.6	74.4	92.1	**88.7**	66.4	79.1	63.9	72.6	78.5
w/o answer box & relevant info	UTD	72.4	72.9	54.4	71.2	92.9	87.2	64.3	78.0	60.2	71.0	78.5
w/ 1 evidence	UTD	61.4	60.9	40.0	55.2	87.3	79.4	49.8	66.8	46.3	62.9	75.3
w/ 5 evidences	UTD	70.6	72.1	56.0	69.6	90.5	81.6	66.4	78.0	57.4	66.1	73.1
w/ 15 evidences	UTD	**77.6**	**78.5**	**60.8**	**78.4**	**96.0**	**88.7**	**72.3**	**81.7**	**70.4**	75.0	80.6
w/ 15 demonstrations	UTD	74.6	75.5	56.8	76.0	93.7	87.9	68.1	79.9	64.8	71.8	76.3
w/ long demonstration answers	UTD	73.0	72.6	55.2	71.2	91.3	83.7	66.0	77.6	60.2	74.2	81.7

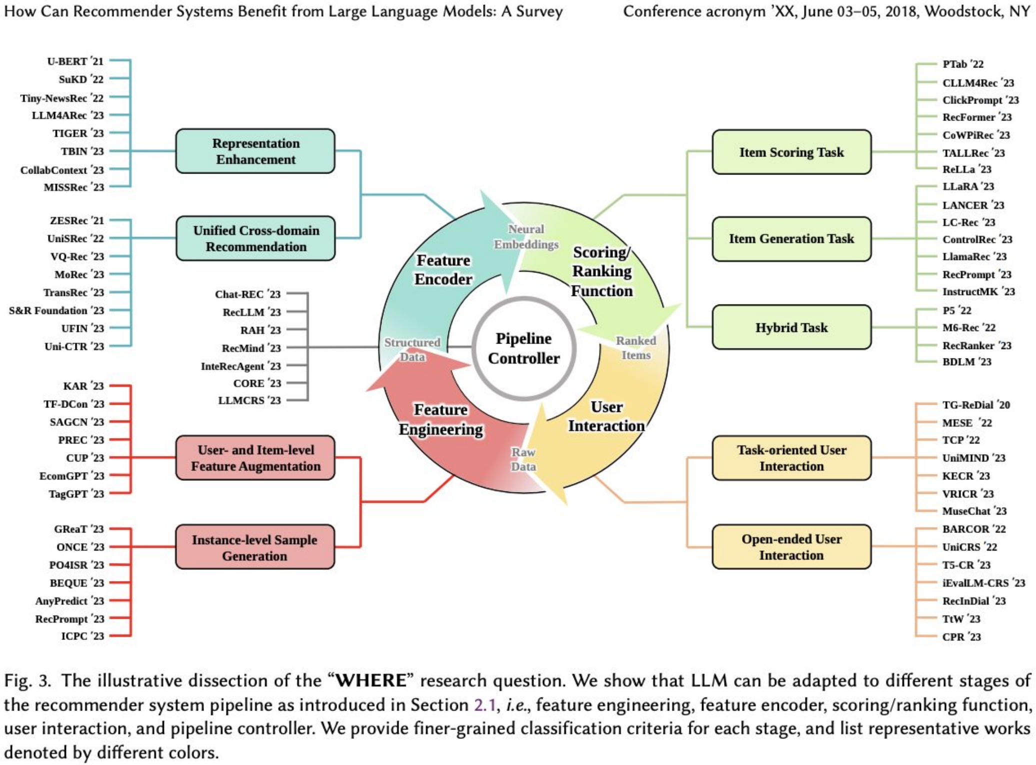
LLMs and Recommender systems

Can LLMs be used for Recommender engines?

Retrieval, ranking, user understanding (sequential), item understanding (using text and images), explanations of recommendations

Can LLMs leveraging large amount of information from the pre training serve as ranking models

LLMs and Recommender systems



How Can Recommender Systems Benefit from Large Language Models: A Survey Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

Fig. 3. The illustrative dissection of the "**WHERE**" research question. We show that LLM can be adapted to different stages of the recommender system pipeline as introduced in Section 2.1, *i.e.*, feature engineering, feature encoder, scoring/ranking function, user interaction, and pipeline controller. We provide finer-grained classification criteria for each stage, and list representative works denoted by different colors.

LLMRank

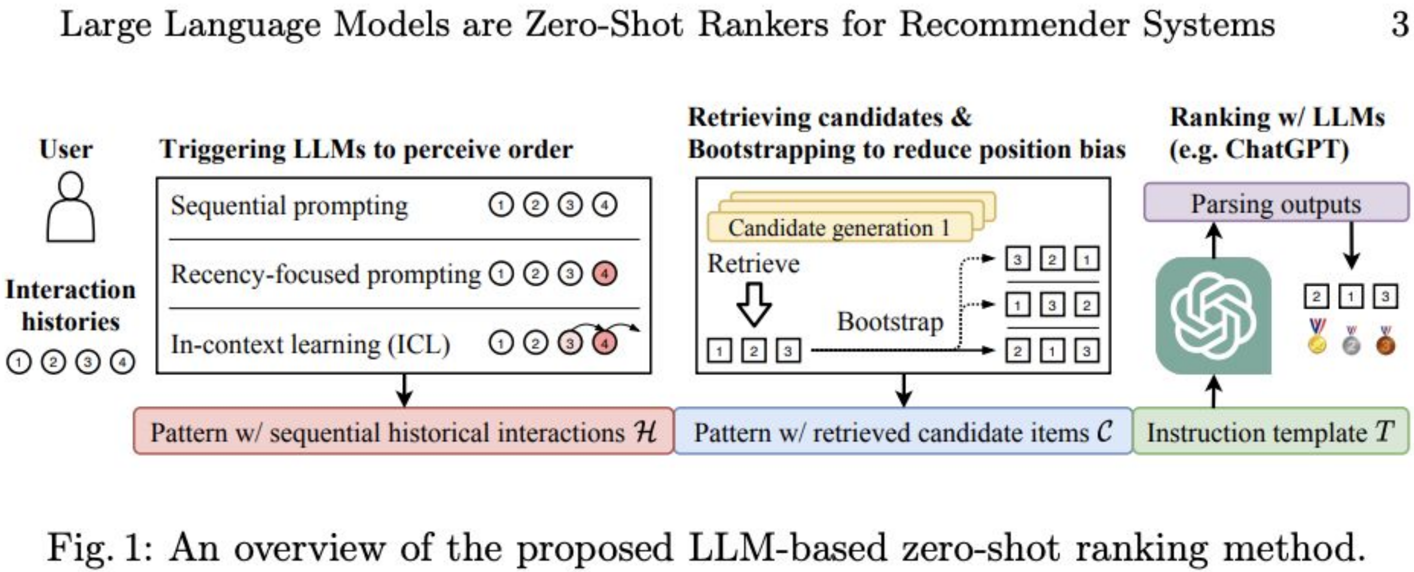
LLM as a zero-shot ranker

We assume that we have a set of candidate items to be recommended, the retrieval is a separate system

The recommendations are conditions on user history, previous interactions with items

There are significantly better rankers, the target here is to make efficient zero-shot ranker. Also, there are other useful features of this framework that will be discussed below

LLM Rank



Large Language Models are Zero-Shot Rankers for Recommender Systems 3

User
Interaction histories
1 2 3 4
Triggering LLMs to perceive order
Sequential prompting 1 2 3 4
Recency-focused prompting 1 2 3 4
In-context learning (ICL) 1 2 3 4
Retrieving candidates & Bootstrapping to reduce position bias
Candidate generation 1
Retrieve
1 2 3
Bootstrap
3 2 1
1 3 2
2 1 3
Ranking w/ LLMs (e.g. ChatGPT)
Parsing outputs
2 1 3
Pattern w/ sequential historical interactions $\mathcal{H}$
Pattern w/ retrieved candidate items $\mathcal{C}$
Instruction template T

Fig. 1: An overview of the proposed LLM-based zero-shot ranking method.

LLM Rank

Several key features

1 Sequential prompting: encode user history of engagement with items into a sequential prompt that list recent items

2 Recently focused prompting: emphasize the most recent interactions so LLM can understand what items are the most recent (LLM does not do well with the order in the sequential prompting)

3 In-Context Learning: using previous interactions, create examples to be learned from aas a Context

LLM Rank

The retrieval service returns small number of items (up to 20)

Retrieved items are encoded sequentially in the prompt for ranking

Encode the ranking task in the ranking template

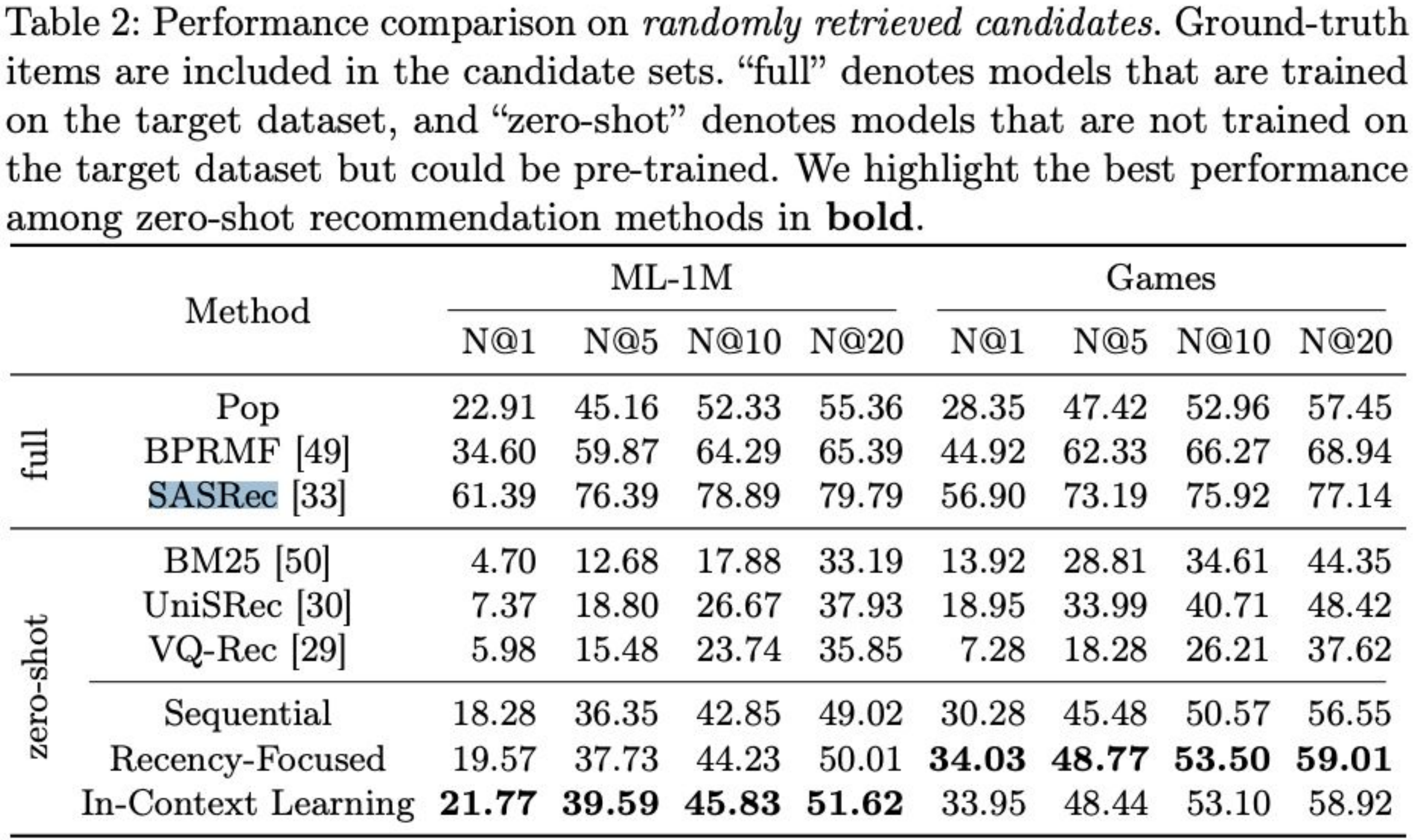


Table 2: Performance comparison on *randomly retrieved candidates*. Ground-truth items are included in the candidate sets. "full" denotes models that are trained on the target dataset, and "zero-shot" denotes models that are not trained on the target dataset but could be pre-trained. We highlight the best performance among zero-shot recommendation methods in **bold**.

	Method	ML-1M				Games			
		N@1	N@5	N@10	N@20	N@1	N@5	N@10	N@20
full	Pop	22.91	45.16	52.33	55.36	28.35	47.42	52.96	57.45
	BPRMF [49]	34.60	59.87	64.29	65.39	44.92	62.33	66.27	68.94
	SASRec [33]	61.39	76.39	78.89	79.79	56.90	73.19	75.92	77.14
zero-shot	BM25 [50]	4.70	12.68	17.88	33.19	13.92	28.81	34.61	44.35
	UniSRec [30]	7.37	18.80	26.67	37.93	18.95	33.99	40.71	48.42
	VQ-Rec [29]	5.98	15.48	23.74	35.85	7.28	18.28	26.21	37.62
	Sequential	18.28	36.35	42.85	49.02	30.28	45.48	50.57	56.55
	Recency-Focused	19.57	37.73	44.23	50.01	**34.03**	**48.77**	**53.50**	**59.01**
	In-Context Learning	**21.77**	**39.59**	**45.83**	**51.62**	33.95	48.44	53.10	58.92

LLM Rank

Low sensitivity to the order in the history (negative feature): mitigation emphasize the most recent items by prompting

Sensitivity to the order of the candidates (negative feature): mitigation: bootstrapping

Too long history effects recommendation quality negatively (negative feature): mitigation: cut the history to the most recent

LLM Rank Possible Extensions:

Recommendation of categories

Conditional recommendations (by attributes, users preferences)

Explanations of recommendations

LLM Evaluation

for

Search and Recommender Engines

LLM Evaluation for Search and Recommender engine

There are many types of evaluation needed. We will focus on few of them, that from our experience are the most frequently needed

Hallucinations

RAGAs

LLM Embeddings

Hallucination Evaluation

HaluEval https://aclanthology.org/2023.emnlp-main.397.pdf

3 tasks : question answering, knowledge grounded dialog, summarization

Responses from the LLM labeled by human annotators

The set is focused on understanding what hallucinations can be produced by LLM and the LLM is asked to produce to wrong answer through hallucination patterns (one pass instruction and conversation schema)

- four types of hallucination patterns for question answering (i.e., comprehension, factualness, specificity, and inference)
- three types of hallucination patterns for knowledge grounded dialogue (i.e., extrinsic-soft, extrinsic-hard, and extrinsic-grouped)
- three types of hallucination patterns for text summarization (i.e., factual, non-factual, and intrinsic)

Focus: Understanding Hallucination patterns and understanding if LLM can detect hallucinations

Hallucination Evaluation

Search-Augmented Factuality Evaluator (SAFE). An evaluation framework from Google DeepMind and Stanford

Available in open source

- Send a request, get an answer
- Break answer into atomic facts
- Check each fact with the verifier (in their case Google Search, and it can be your system)

https://arxiv.org/pdf/2403.18802.pdf

Hallucination Evaluation

Hallucination Leaderboard https://arxiv.org/abs/2404.05904

Variety of tasks: close-book open-domain qa, summarization, reading comprehension, instruction following, fact checking, hallucination identification

Evaluation of RAG as example of LLM App evaluation

RAG architectures will be one of the most frequent industrial pattern of LLM usage

- Correctness
- Comprehensiveness
- Readability
- Novelty/Actuality
- Quality of information
- Factual answering correctness
- Depth
- Other metrics

Similar to a traditional search engine evaluation as we evaluate a requested information but there is a substantial difference as we evaluate generate response rather than external documents

Traditional IR architecture: retrieval -> ranker, RAG architecture : retrieval -> generator, different type of evaluation

Evaluation of RAG

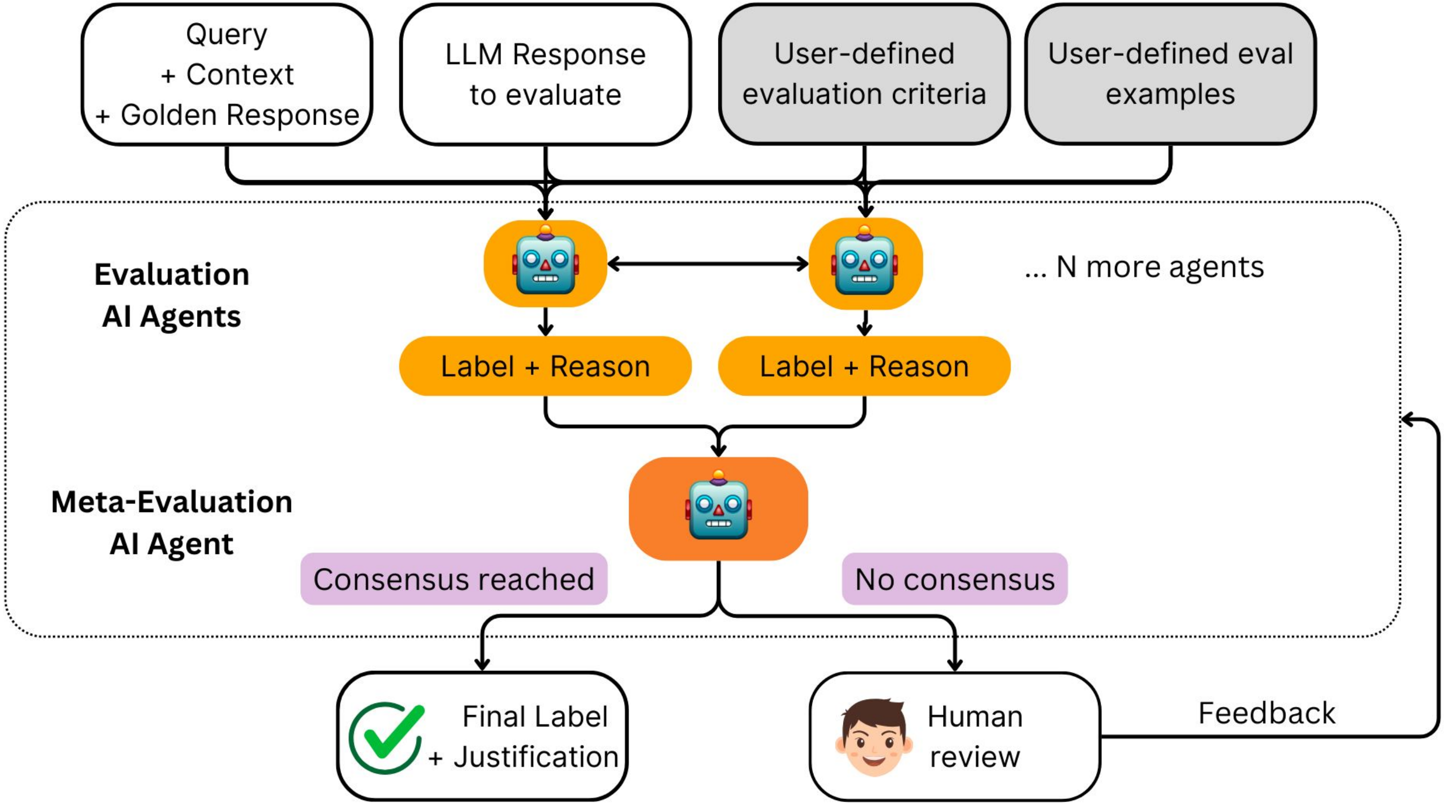
A problem, comparison of generated text (answer) with the reference answer. Semantic similarity problem

Old lexical metrics (BLEU, ROUGE) are easy to compute but give little usable answers

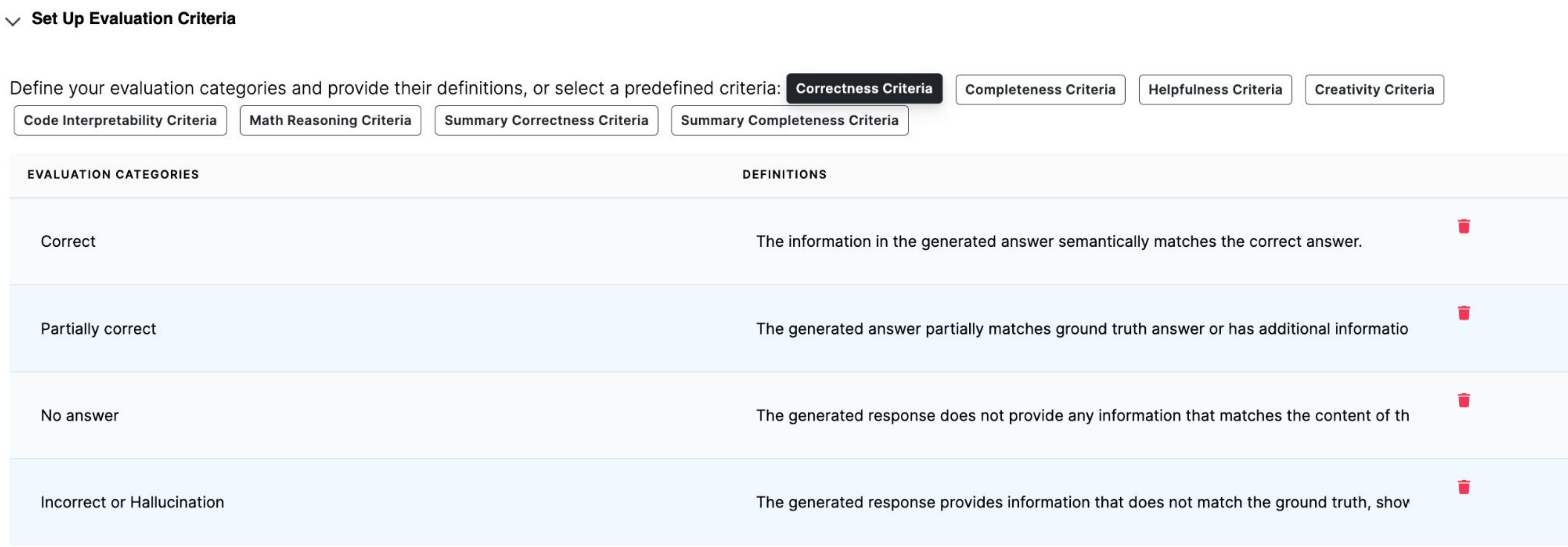
BERTScore, BARTScore, BLEURT and other text similarity functions

Calls to external LLMs

Lighthouz evaluations - via multiagent evaluators



Eval criteria: hallucinations / correct responses

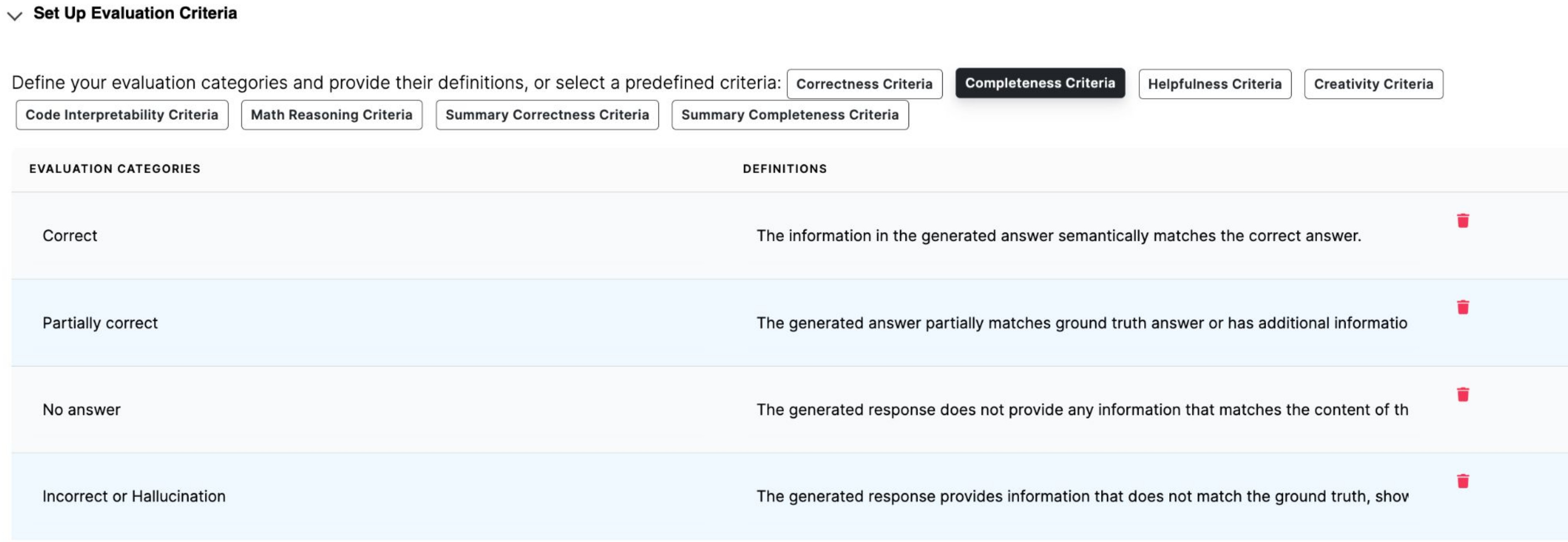


Set Up Evaluation Criteria

Define your evaluation categories and provide their definitions, or select a predefined criteria:

EVALUATION CATEGORIES	DEFINITIONS
Correct	The information in the generated answer semantically matches the correct answer.
Partially correct	The generated answer partially matches ground truth answer or has additional informatio
No answer	The generated response does not provide any information that matches the content of th
Incorrect or Hallucination	The generated response provides information that does not match the ground truth, shov

Eval criteria: completeness



Set Up Evaluation Criteria

Define your evaluation categories and provide their definitions, or select a predefined criteria:

EVALUATION CATEGORIES	DEFINITIONS
Correct	The information in the generated answer semantically matches the correct answer.
Partially correct	The generated answer partially matches ground truth answer or has additional informatio
No answer	The generated response does not provide any information that matches the content of th
Incorrect or Hallucination	The generated response provides information that does not match the ground truth, sho

Other evaluation criteria:

- Helpfulness
- Creativity
- Summary correctness
- Summary completeness
- Code interpretability
- Math reasoning

You can build any criteria you like!

Agents at work for evaluations

Quick demo

RAG Evaluation 3 typical metrics

Context Relevance

Answer Relevance

Groundedness (precision)

(common across multiple frameworks RAGAs, ARES, RAG Triad of metrics)

but

Evaluation of RAG - RAGAs

Zero Shot LLM Evaluation

4 metrics:

- Faithfulness,
- Answer relevancy
- Context precision
- Context recall

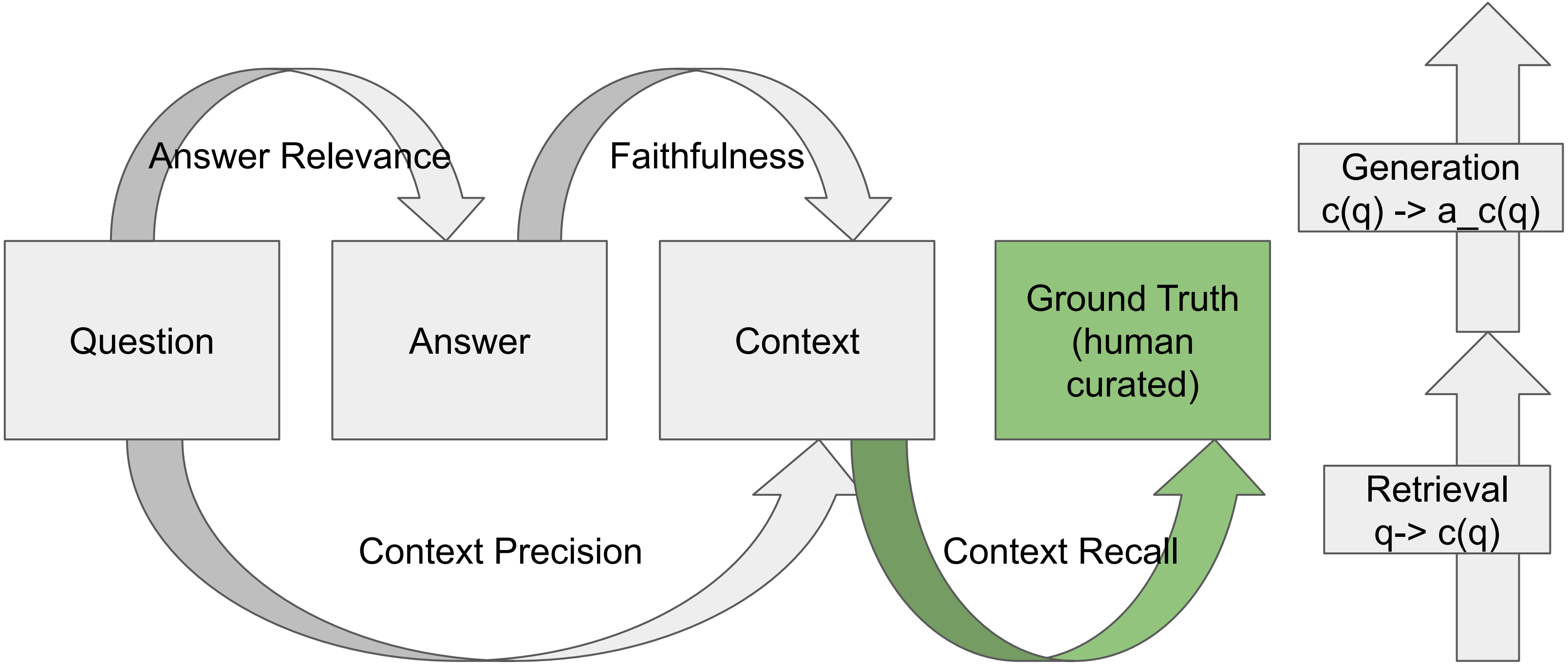
Important to enhance to what represents your RAG intents for your customers

https://arxiv.org/abs/2309.15217

https://docs.ragas.io/en/stable/

RAGAs framework integrated with llamaindex and LangChain

RAGAs


Answer Relevance
Faithfulness
Generation
c(q) -> a_c(q)
Question
Answer
Context
Ground Truth
(human
curated)
Retrieval
q-> c(q)
Context Precision
Context Recall

Evaluation of RAG - RAGAs

Faithfulness consistency of the answer with the context (but no query!) Two LLM calls, get context that was used to derive to answer, check if the statement supported by the context	**Context Relevance** how is the retrieved context “focused” on the answer , the amount of relevant info vs noise info , uses LLM to compute relevance of sentences / total number of retrieved sentences
Answer Relevancy is the answer relevant to the query , LLM call, get queries that may generate answer, verify if they are similar to the original query	**Context Recall** (ext, optional) if all relevant sentences are retrieved , assuming existence of ground_truth answer

RAGAs

Faithfulness : LLM prompt to decompose answer and context into statements F = supported statements / total statements	**Context Relevance**: LLM Prompt to decompose contexts into sentences and evaluate if the sentences are relevant to the question CR = sentences in the context / relevant sentences
Answer Relevance: LLM prompt to generate questions for the answer. For each question generate embedding. Compute the average semantic similarity score between original query and all generated queries AR = 1/n \sum sim(q, q_i)	**Context Recall:** CR = GT sentences attributed to recall / GT sentences

Evaluation of RAGs - RAGAs

Prompts should be well tuned, hard to move to another context, or LLM, requires a lot of work on tuning of prompts

Each metrics: faithfulness, answer relevancy, context relevance, context recall can be dependent on your domain/business. It requires tuning to measure that your business depends upon

Available in open source : https://docs.ragas.io/en/stable/ integrated with key RAG frameworks

More metrics in new version (Aspect Critique, Answer Correctness etc)

LLM Embedding evaluation

MTEB (Massive Text Embedding Benchmark) is a good example of the embedding evaluation benchmark

https://arxiv.org/pdf/2210.07316.pdf

https://huggingface.co/spaces/mteb/leaderboard

It’s relatively comprehensive - 8 core embedding tasks: Bitext mining, classification, clustering, pair classification, reranking, retrieval, semantic text similarity (sts), summarization and open source

LLM Embedding Evaluation

MTEB evaluation - Learning: no clear winner model across all tasks, the same most probably will be in your case, you ll find different model-winners for different tasks and may need to make your system multi-model

MTEB: Easy to plugin new models through a very simple API (mandatory requirement for model development)

MTEB: Easy to plugging new data set for existing tasks (mandatory requirement for model development, your items are very different from public dataset items)

LLM Embedding Evaluation

There are standard ‘traditional IR’ methods to evaluate embeddings such as recall @ k, precision @ k, ndcg @ k that are easy to implement on your own and create dataset representing your data

They are even supported by ML tools such as MLFlow LLM Evaluate

Important learning: find metrics that truly matches customer expectations (for ex NDCG is very popular but it’s old and it was built in different setting, one most probably needs to turn it to their system, to represent the way users interact with their system)

LLM Embedding Evaluation

Another critical part of LLM evaluation for embedding evaluation is software performance/operations evaluation of the model

Cost, latency, throughput

In many cases, embeddings must be generate for every item on every update (ex: 100M+ updates per day), or for every query (1000000+ qps with latency limits such as 50ms)

The number of model calls and the latency/throughput requirements are different for embedding tasks rather than other LLM tasks. Most embedding tasks are high load tasks

LLM Embedding evaluation

All traditional search evaluation data set requirement are still valid

Your evaluation set must represent users queries or documents (what you measure) with similar distribution (proper sampling for documents, query logs) , represent different topics, popular, tail queries, languages, types of queries (with proper metrics)

Queries and document are change over time, the evaluation set must reflect these changes

Take into account raters disagreement (typically high in retrieval and ranking and use techniques to diminish subjectivity (pairwise comparison, control of the number of raters etc))

LLM Embeddings

In most cases, core tasks are serving several tasks facing customers/business. For example, text similarity might be a part of discovery/recommendation engine (text similarity of items as a one of features for the similarity of items) or ranking (query similarity as if historical performance one of query is applicable as click signals for another query).

Evaluation is not only text similarity LLM output, but the whole end-to-end rank,recommendation etc output