Knowledge Curation

Stanford CS224V Course
Conversational Virtual Assistants with Deep Learning

Monica Lam, Yucheng Jiang

Announcement

1. Student intro form due **today**

2. HW1 released today (Due: October 2nd)

    - Get familiar with tool stack of knowledge curation pipeline
    - Have hands-on experience designing LLM-enpowered system

Lecture Plan

1. **Knowledge Curation (STORM)**

   - Information Retrieval & RAG at a Glance

   - Evaluation? Evaluation!

2. **Bring Human into the Loop (Co-STORM)**

3. **HW1 Overview**

Information Retrieval at a Glance

When we have an information-seeking need,

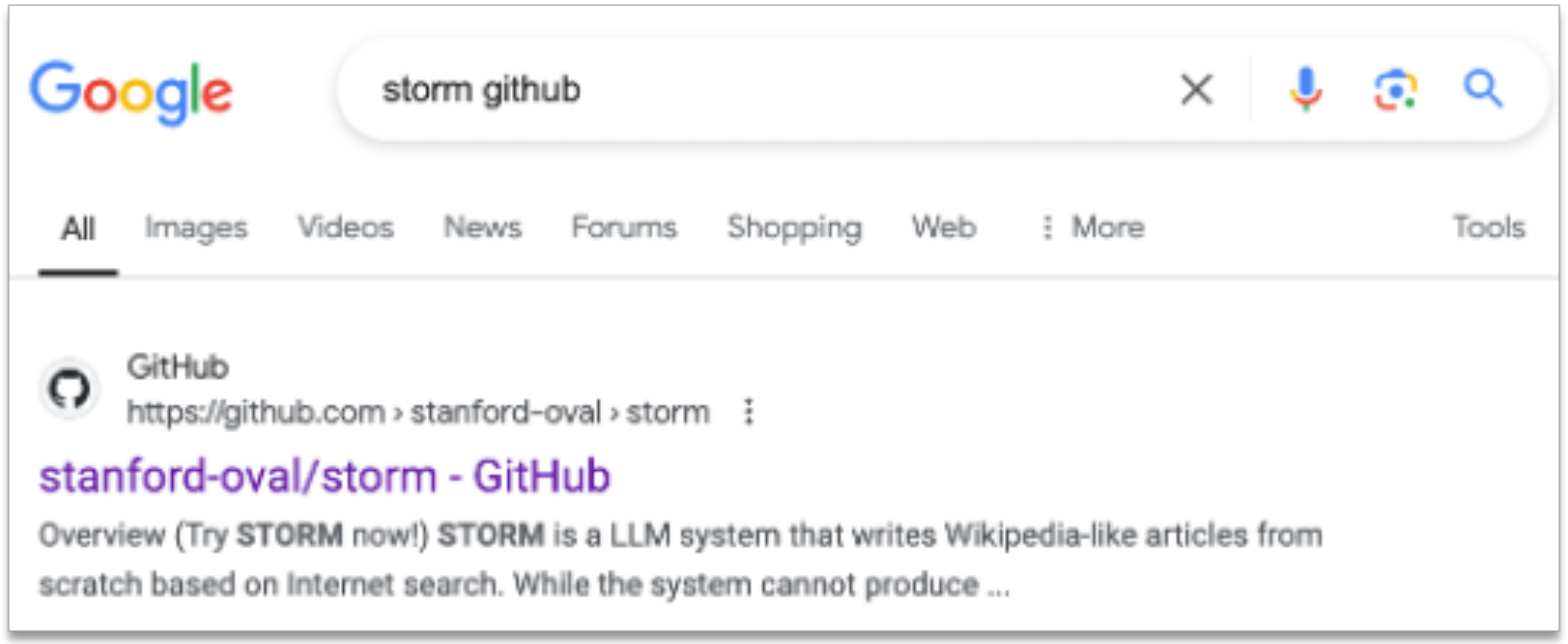
Query → **Retriever** → Ranked Documents

↑

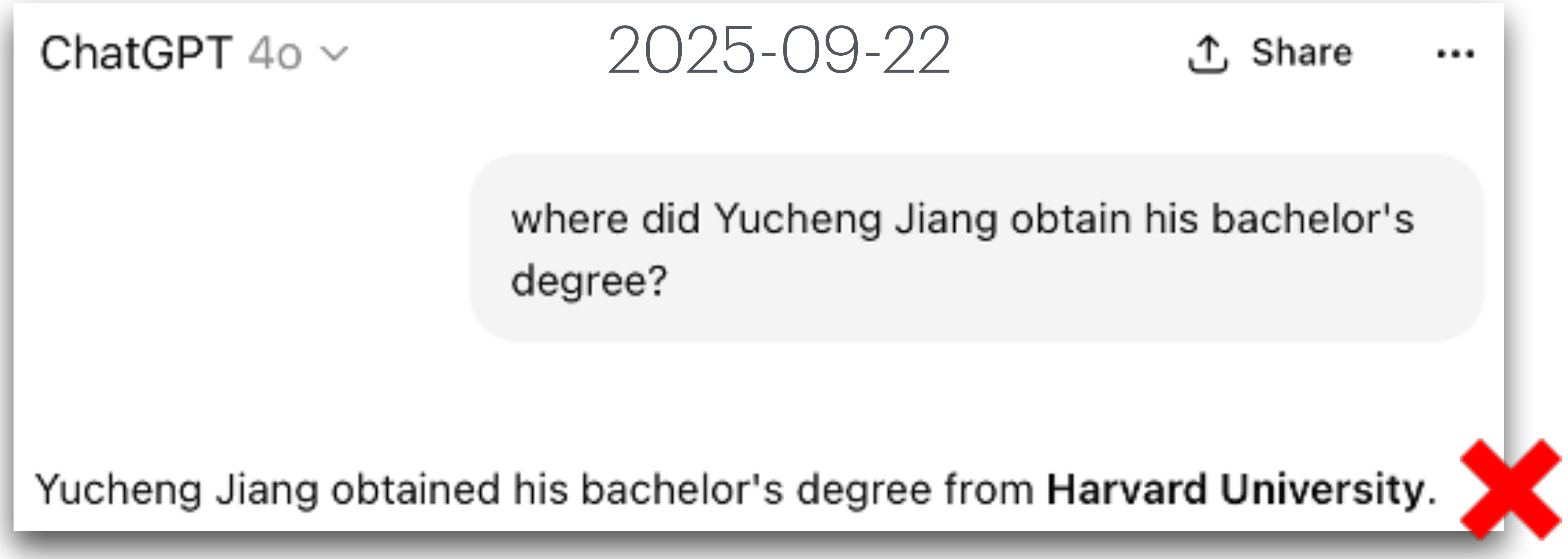
The Internet

Public / Private Corpora

Structure & Unstructured Database



Information Retrieval at a Glance



The major issue of using LLMs for knowledge tasks: **Hallucination**

A response that is not faithful to the facts of the world.

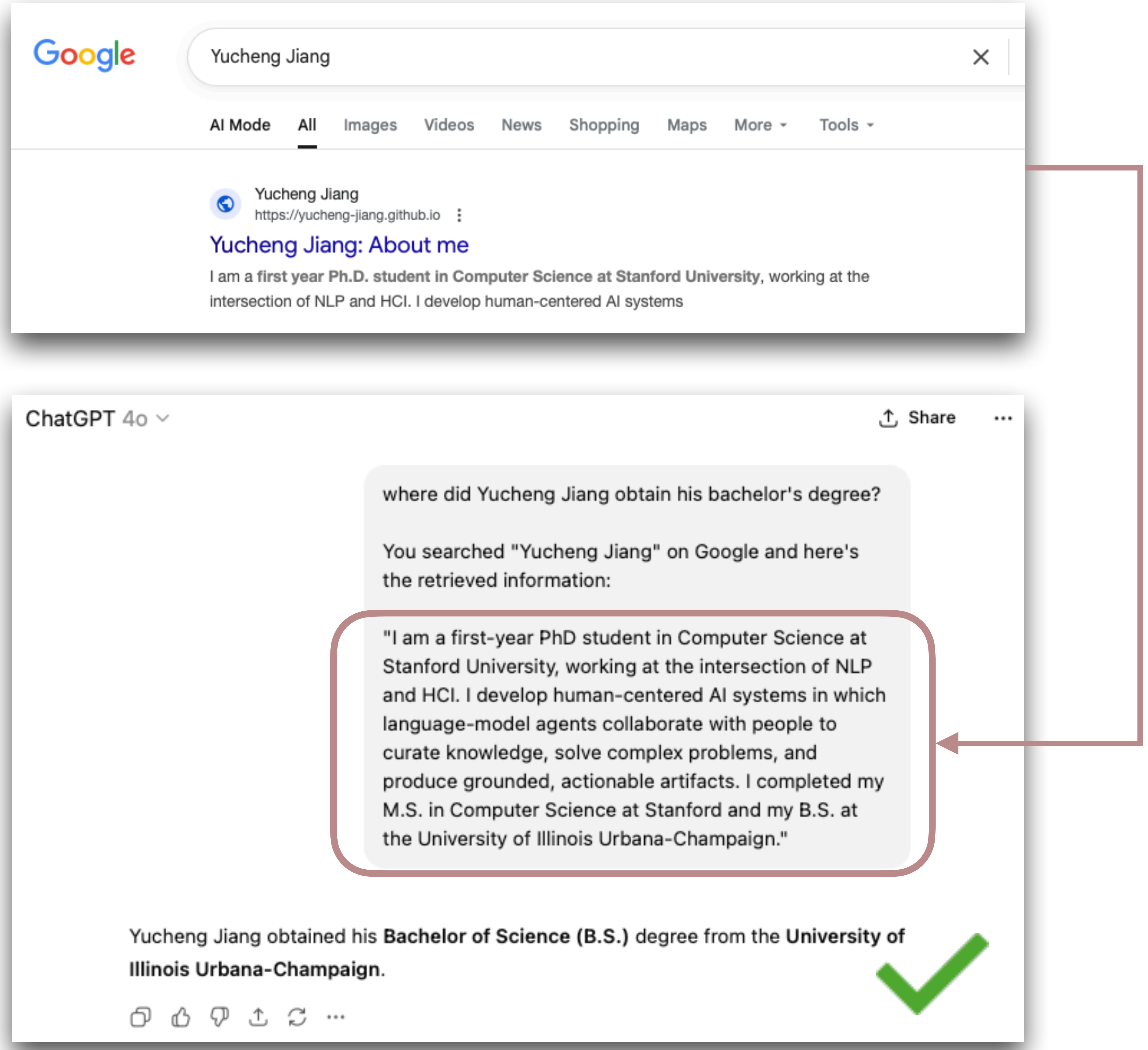
- Long-tail information
- Knowledge cutoffs
- Private data

Retrieval Augmented Generation

Step 1: **Retrieve**

Step 2: **Augment**

Step 3: **Generate**



Retrieval Augmented Generation

Retrievers

Data sources

HW1 The Internet

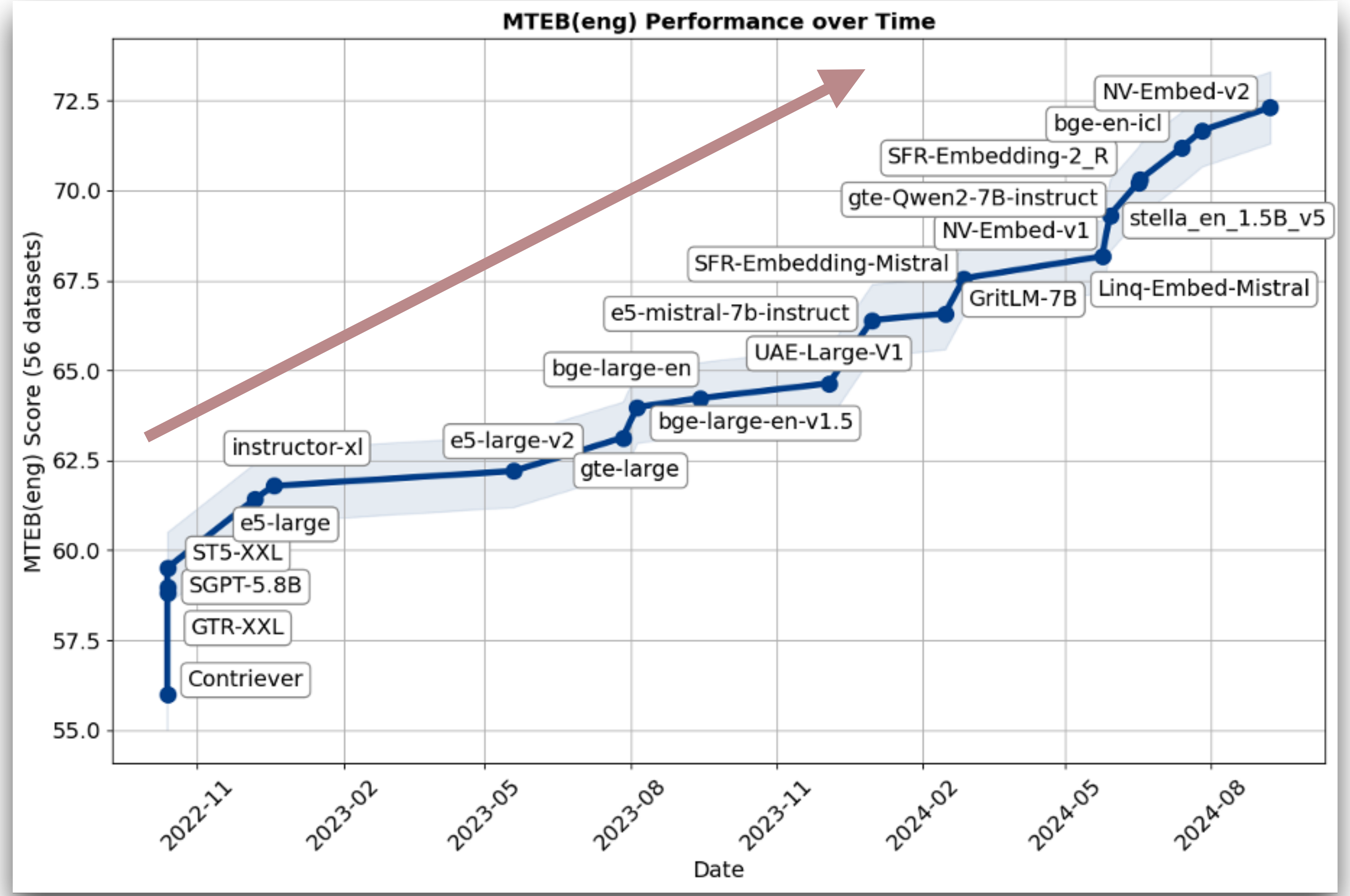
10/22, 10/27 Lecture Structured and Unstructured database

10/29 Lecture Public/Private corpora

11/12 Lecture Hand-written documents

→ Indexing → Chunking → Reranking

Retrieval Augmented Generation



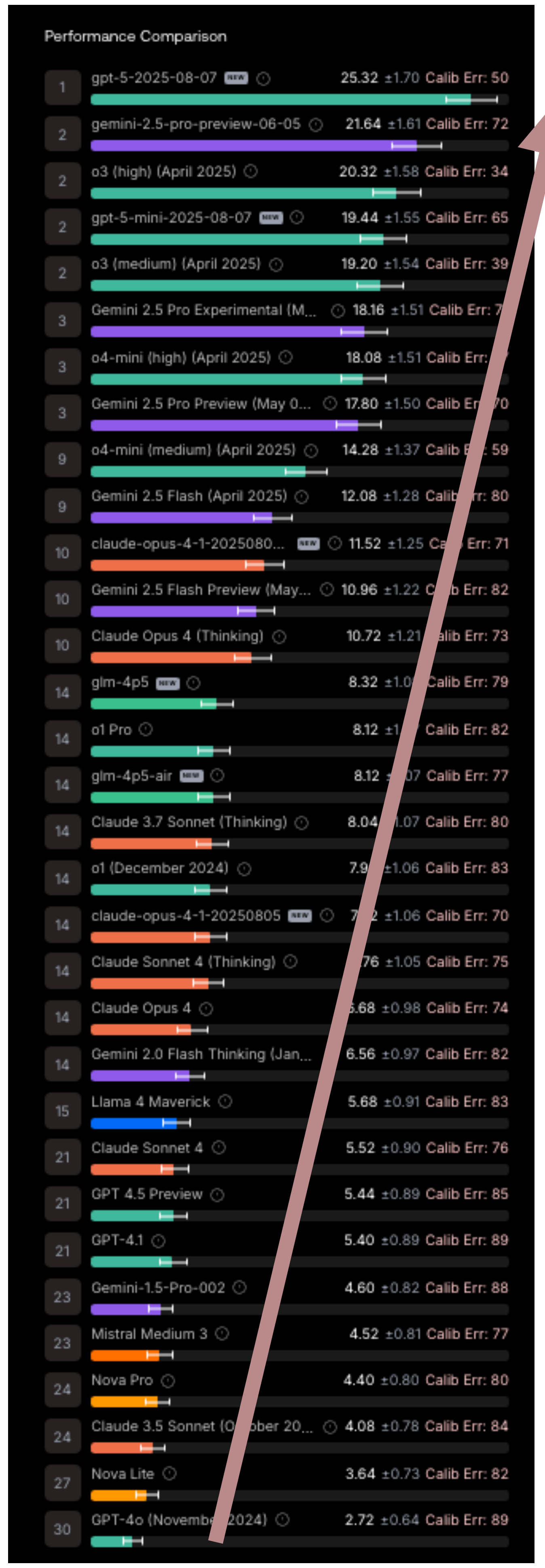
MTEB: Massive Text Embedding Benchmark, Niklas et al. 2023 (This illustration is contributed by Niklas Muennighoff.)

We have better embedding models and infrastructure for Information Retrieval over time.

Retrieval Augmented Generation

Humanity's Last Exam:

2,500 challenging questions across over a hundred subjects, at the frontier of human knowledge



GPT-5 (2025.08) score: 25.32

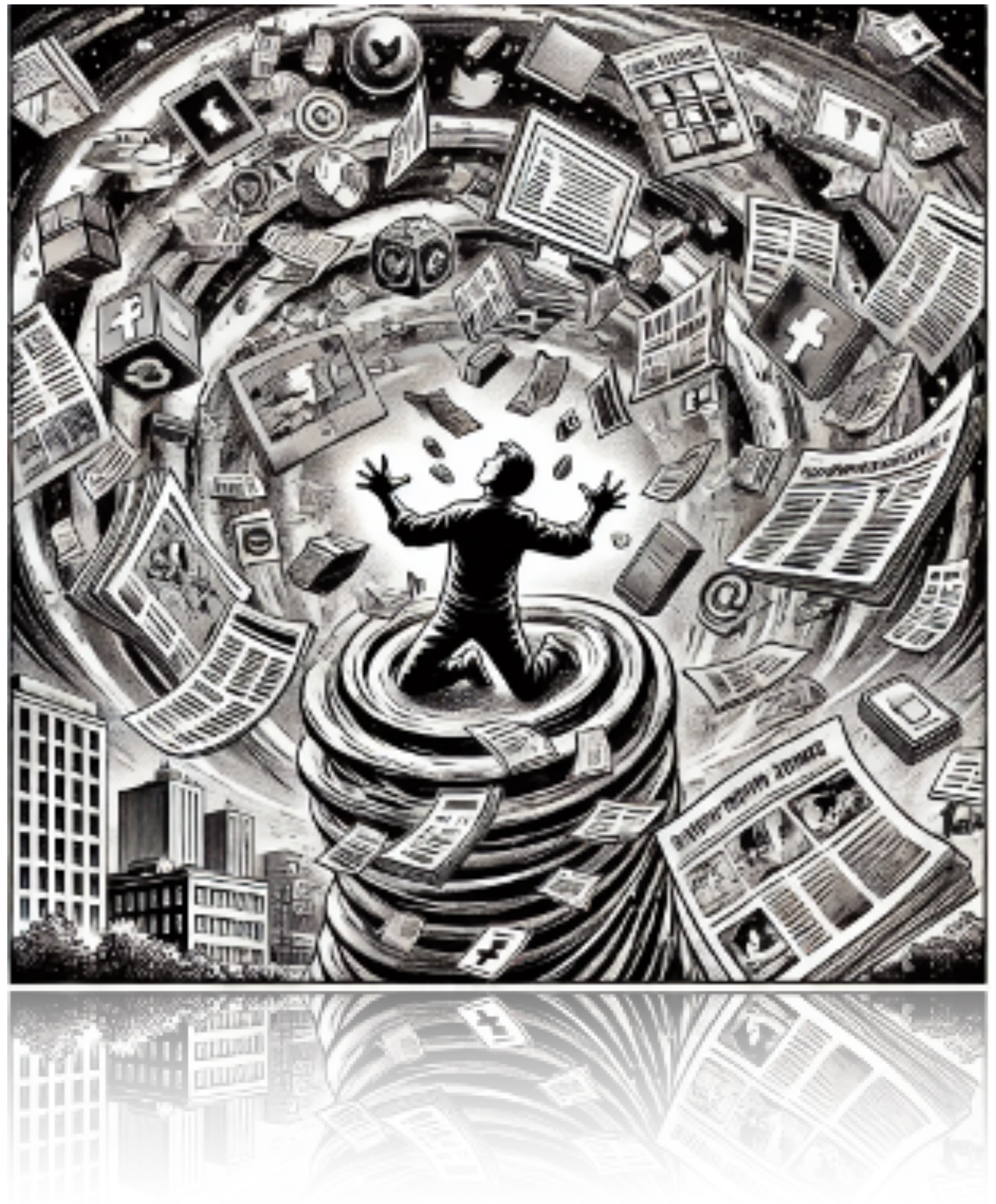
O4-mini (2025.04) score: 14.28

GPT-4o (2024.11) score: 2.72

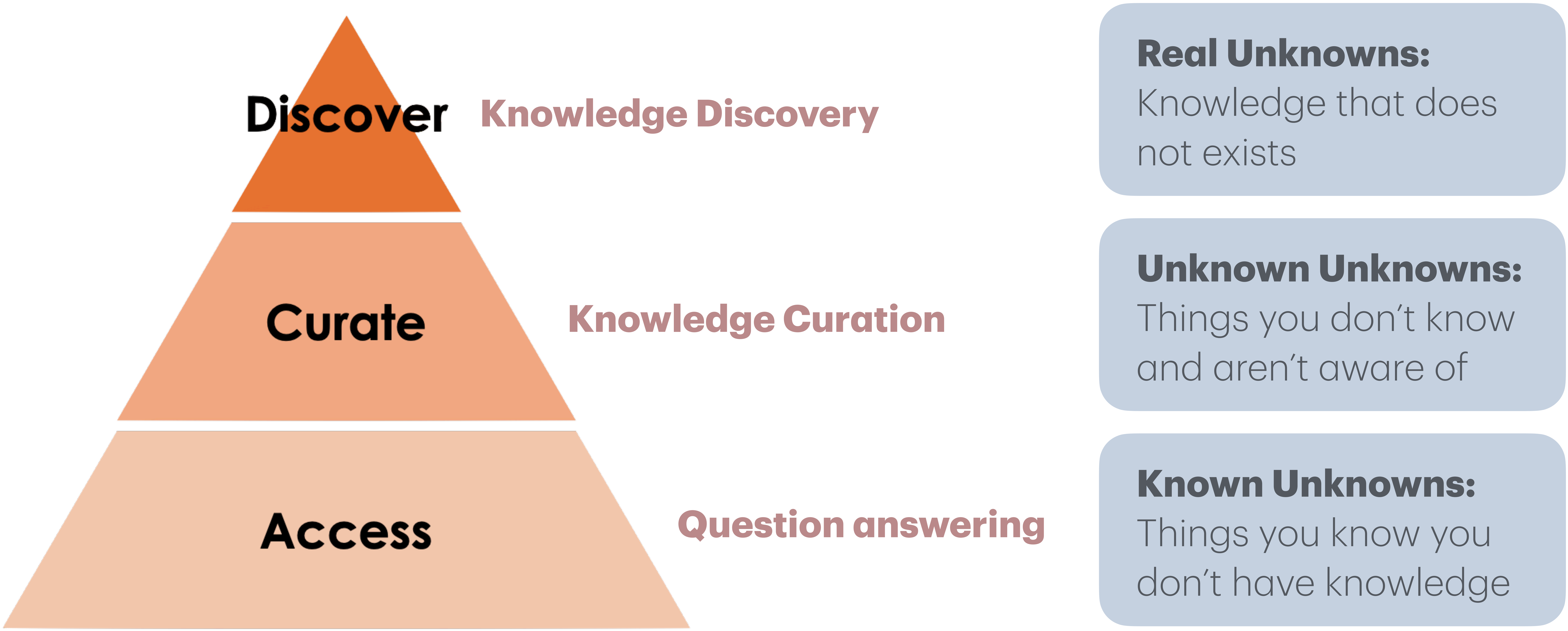
We have stronger, more intelligent language models.

Meta question

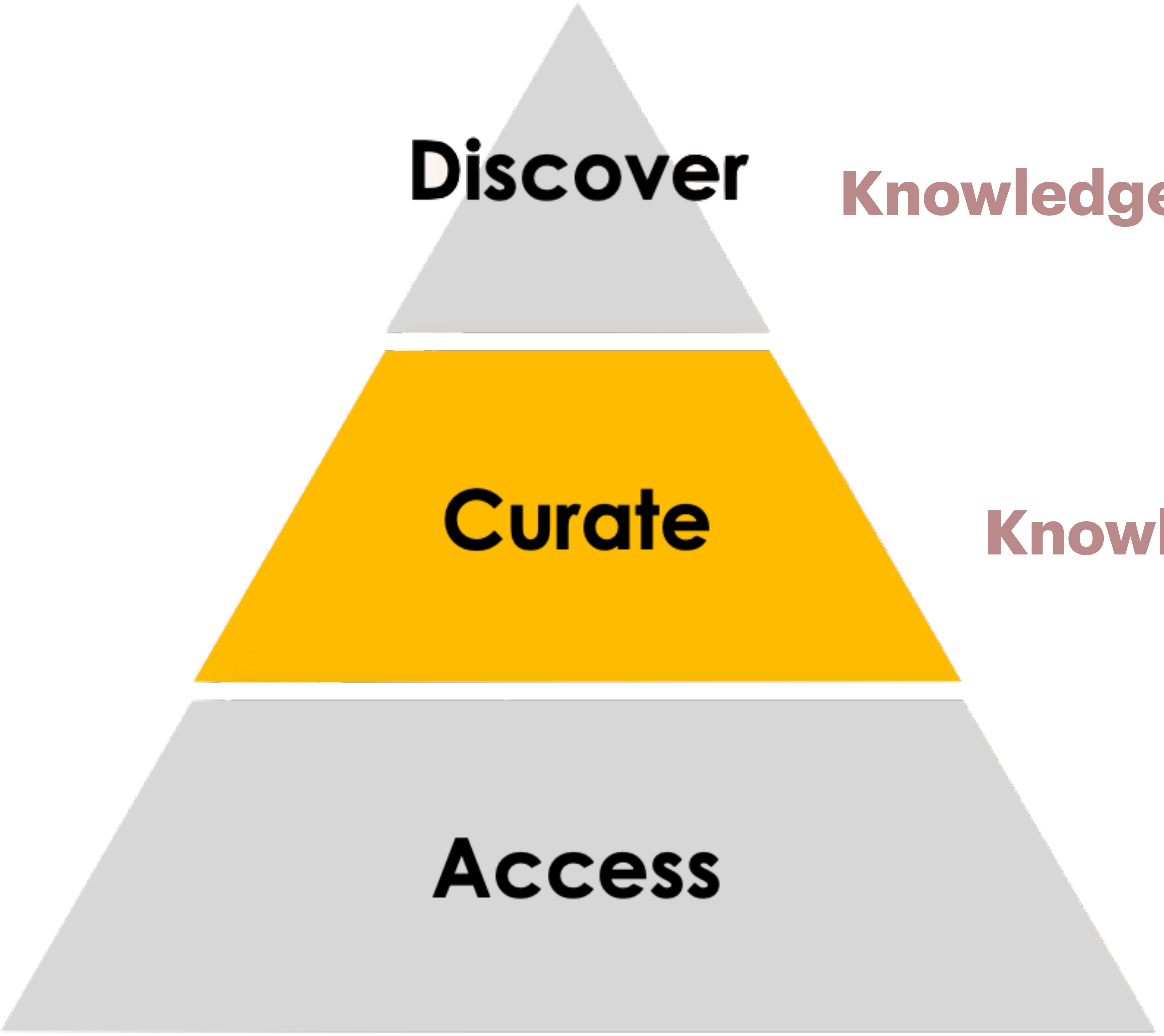
Are people's information needs satisfied?



The illustration is co-created with DALL-E.



The illustration is contributed by Yijia Shao



Knowledge Discovery

Knowledge Curation

Question answering

Real Unknowns: Knowledge that does not exists

Unknown Unknowns: Things you don’t know and aren’t aware of

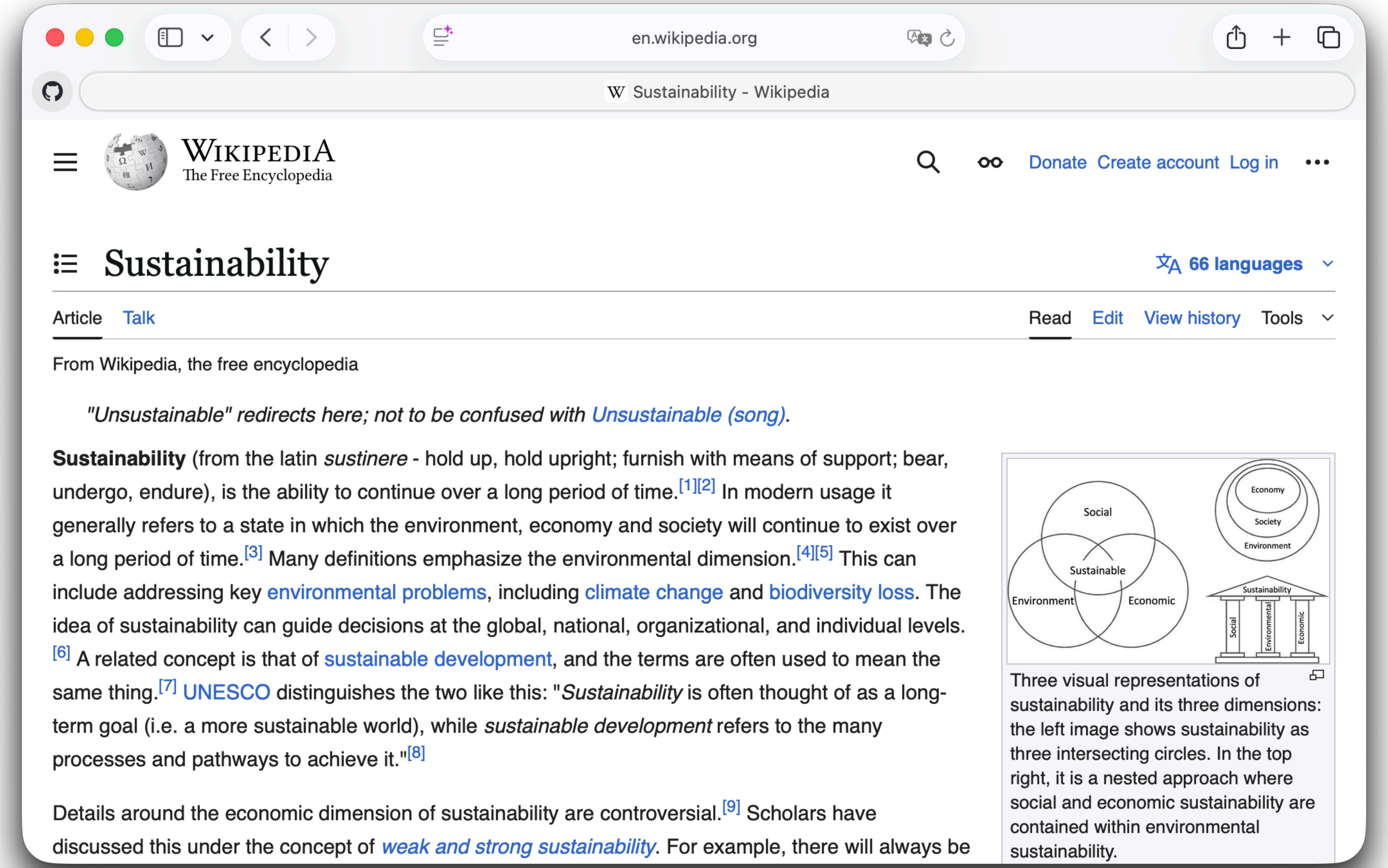
Known Unknowns: Things you know you don’t have knowledge

The illustration is contributed by Yijia Shao

Knowledge Curation

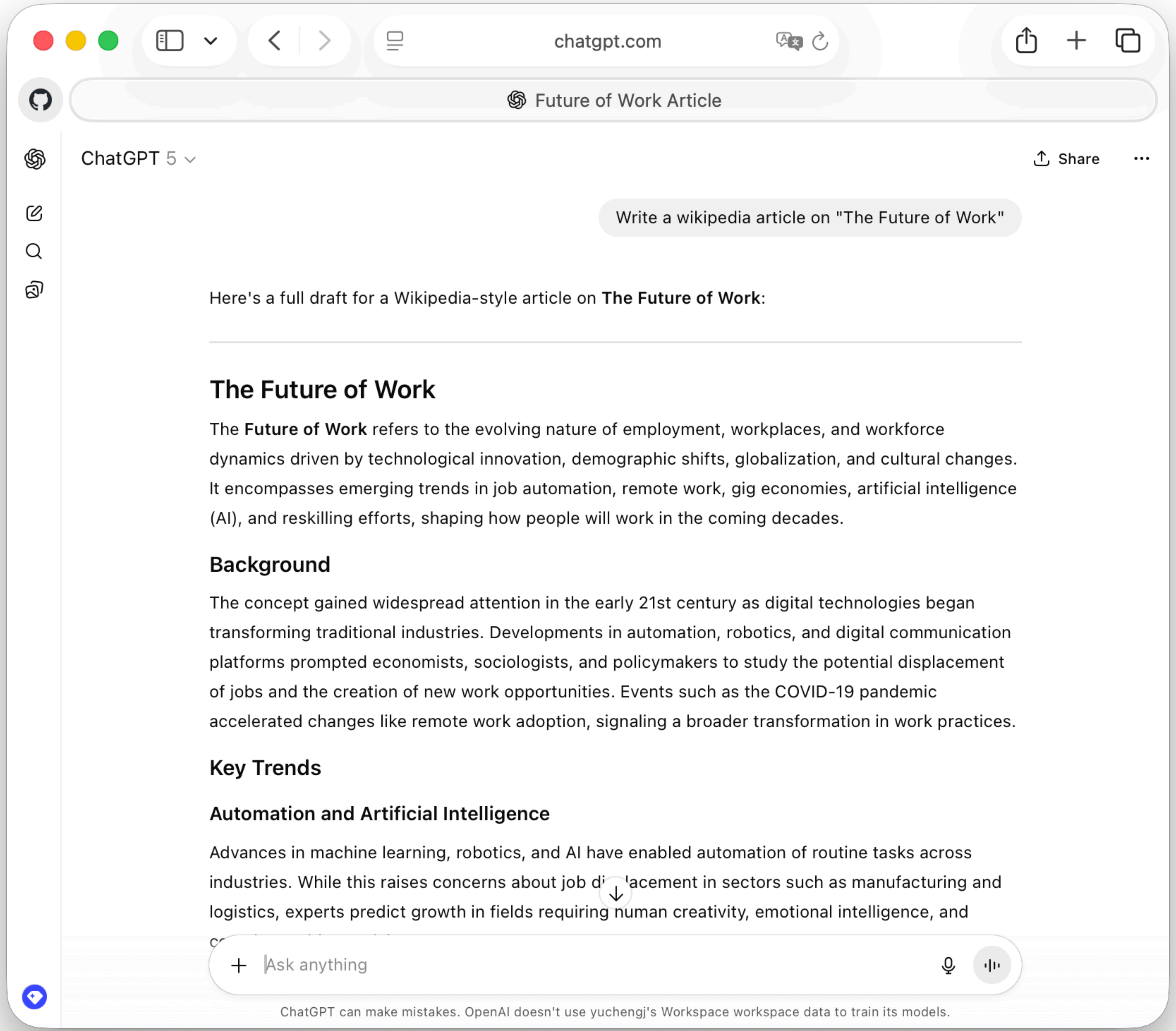
Wikipedia is a good example of knowledge curation.

- Comprehensive
- Organized
- Reliable
- Verifiable



Knowledge Curation

Generating wikipedia-like article is non-trivial



- Lack of details
- Hard to verify

Knowledge Curation

Early stage long form generation methods (1/2) - Training Neural Models

Given the ordered paragraphs $\{p^i_{R_{i(j)}}\}$

Encode, concatenate, and truncate

$$\text{text}_i = T(a_i) \,||\, \{p^i_{R_{i(j)}}\}$$

$$\text{tokenize}(\text{text}_i) = x_i = (x_i^1, x_i^2, \ldots, x_i^{n_i})$$

$$m_i^L = (x_i^1, \ldots, x_i^{min(L,n_i)})$$

Train an abstractive model W that learns to write articles, $a_i = W(m_i^L)$

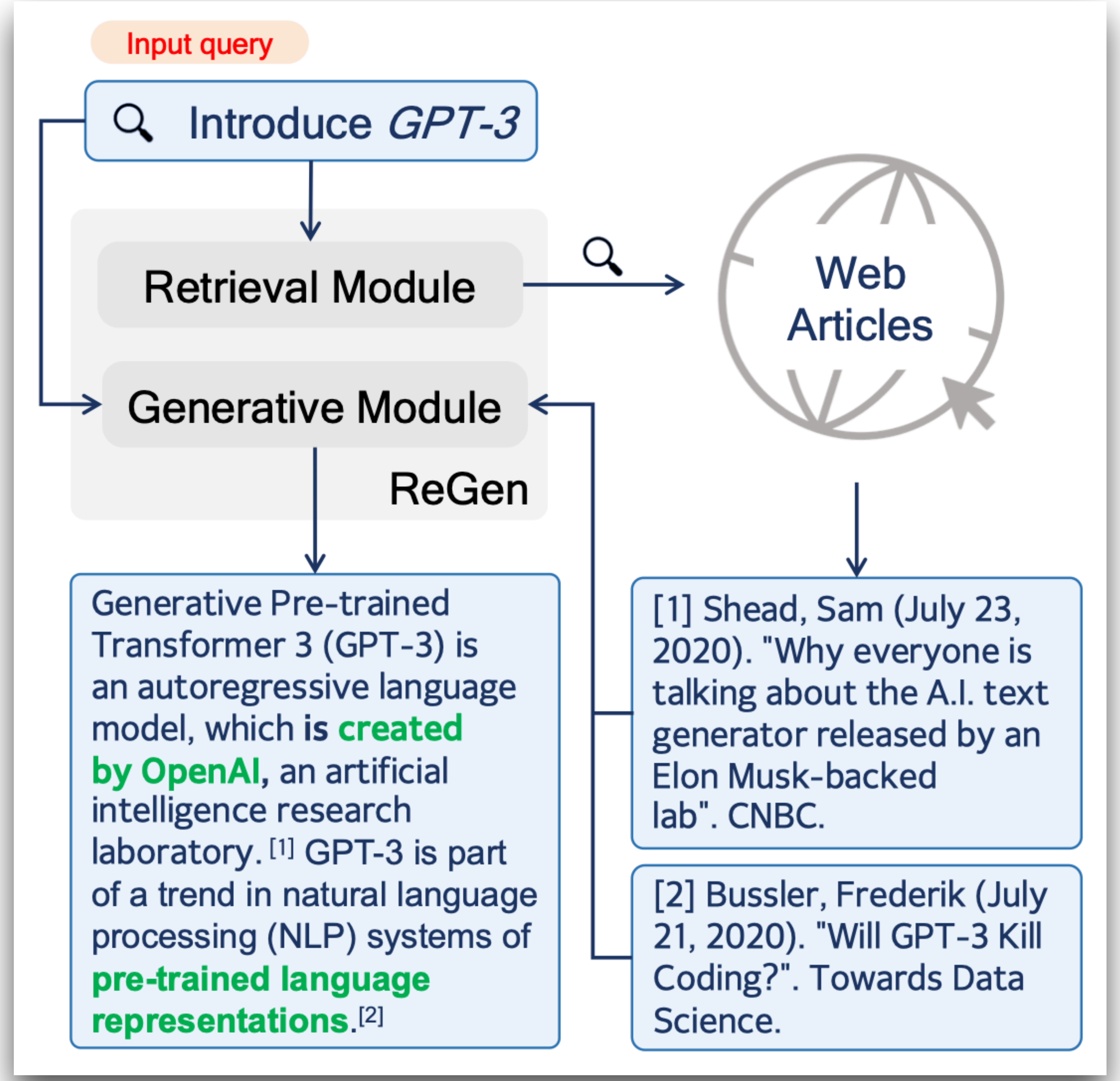
Early prior works usually **assumes** the references are given.

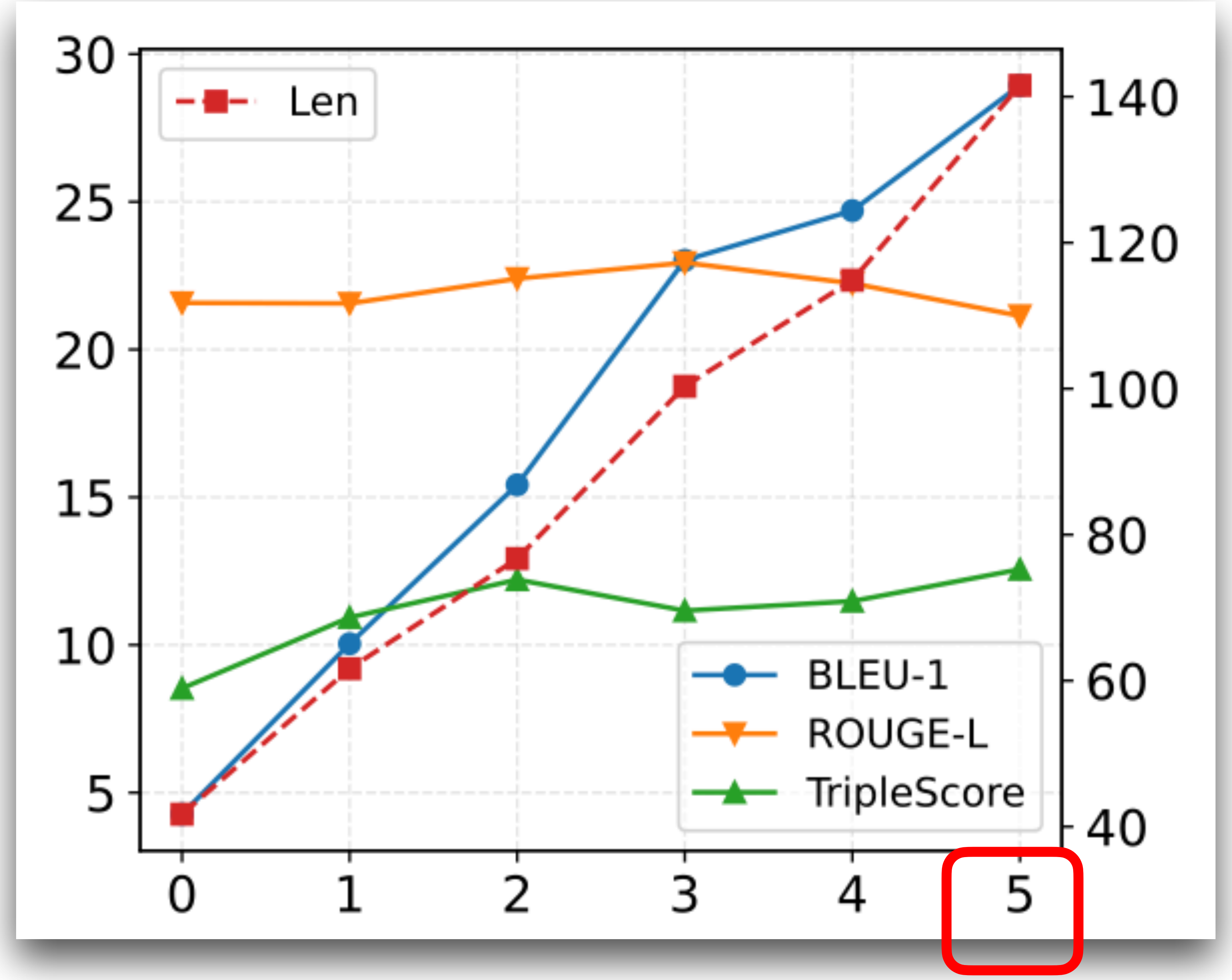
However, collecting references requires literature research which is non-trivial

Generating Wikipedia by Summarizing Long sequences, Liu et al., 2018

Knowledge Curation

Early stage long form generation methods (2/2) - Prompting an LLM





Limited length, only a few citations

WebBrain: Learning to Generate Factually Correct Articles for Queries by Grounding on Large Web Corpus. Qian et al., 2023

Knowledge Curation

STORM: Assist in writing Wikipedia-like articles from scratch with LLMs

2024/02: **STORM** - First open source knowledge curation system - Beginning of Deep Research

2024/12: Gemini Deep Research

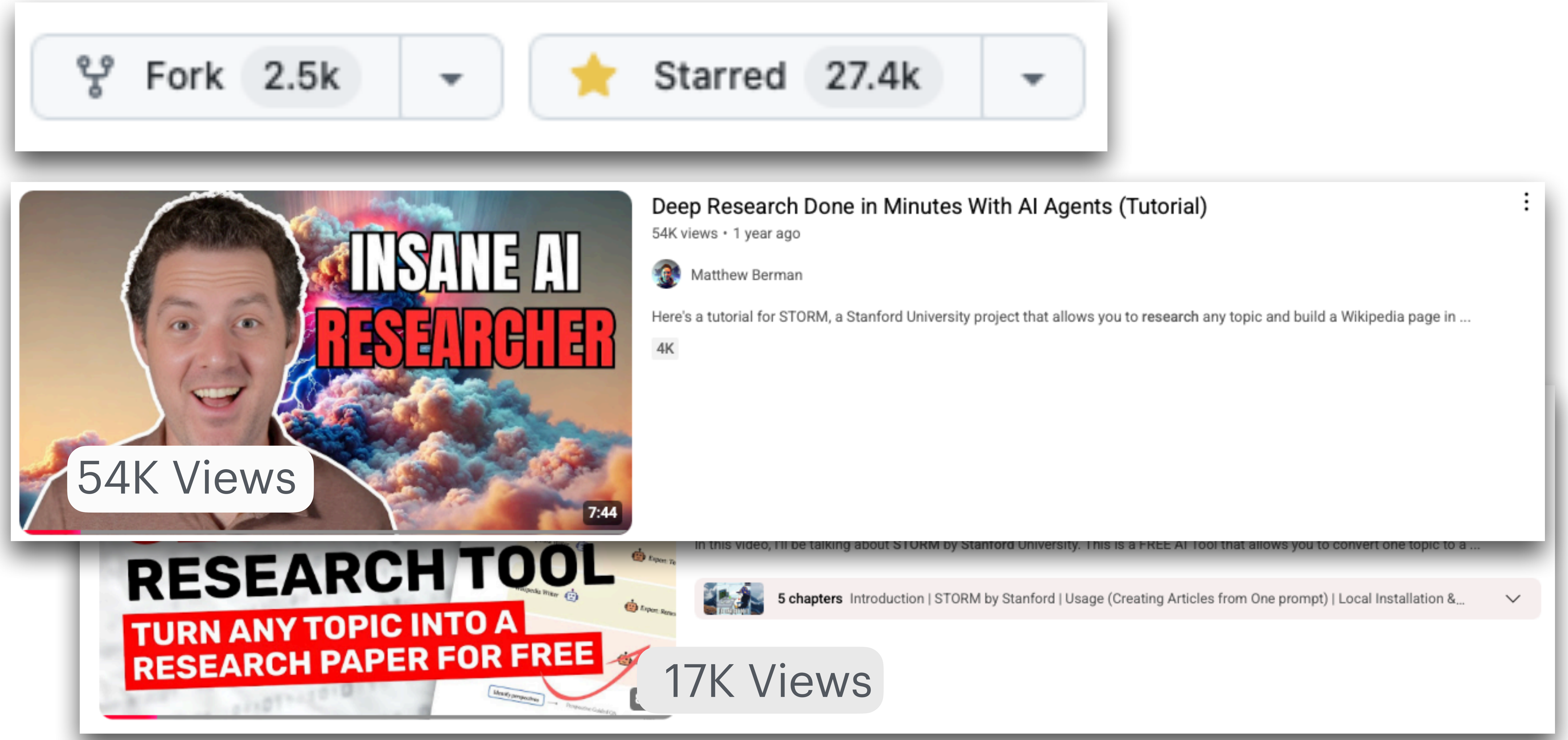
2025/02: OpenAI Deep Research, Perplexity Deep Research

...

Knowledge Curation

STORM: Assist in writing Wikipedia-like articles from scratch with LLMs

STORM has aroused interest across various communities



Shao, Yijia, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. "Assisting in writing Wikipedia-like articles from scratch with large language models.", In NAACL 2024

Knowledge Curation

STORM: How to generate grounded articles with good breadth & depth

Key Idea: Mimic Human Writing Process

How do humans write?

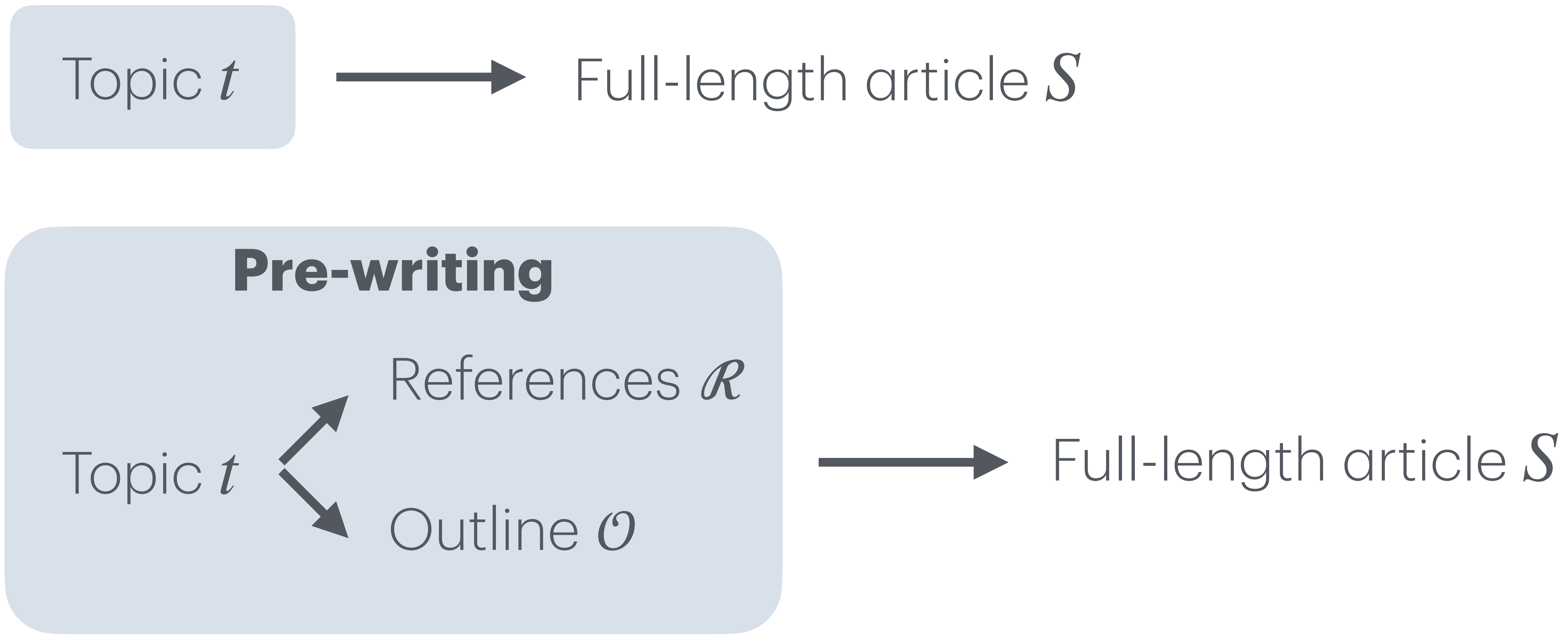
Rohman, 1965: **Pre-Writing** the Stage of Discovery in the Writing Process

How do humans do literature search?

Booth et al., 2003: The Craft of Research “Chapter II: **Asking Questions**, Finding Answers”

Knowledge Curation

STORM: Pre-writing



The pre-writing task:

Give a topic t, the pre-writing task is to find a set of references $\mathcal{R}$, and create an outline $\mathcal{O}$, which is defined as a list of multi-level section headings, to organize $\mathcal{R}$.

Knowledge Curation

STORM: Literature research via question asking

Topic: 2022 Winter Olympics Opening Ceremony

Prompt: Ask 30 questions about the given topic

1. **When** was the opening ceremony held?
2. **Where** was the opening ceremony held?
3. **How many** countries participated in the opening ceremony?

Direct prompting results in questions that lack breadth and depth

We cannot simply reply on "brute force" or inference-time scaling

Knowledge Curation

STORM: Literature research via perspective guided QA

STORM uses **perspective** as a latent variable to control the breadth of the search.

Topic: 2022 Winter Olympics Opening Ceremony

Survey related topics:
wiki/2020_summer_olympics
wiki/2018_winter_olympics

Identify perspectives:
(e.g. Economist: this editor will bring in the economic perspective, focusing on topics such as national macro economic effects...)

Knowledge Curation

STORM: Simulating conversations to allow follow-up questions

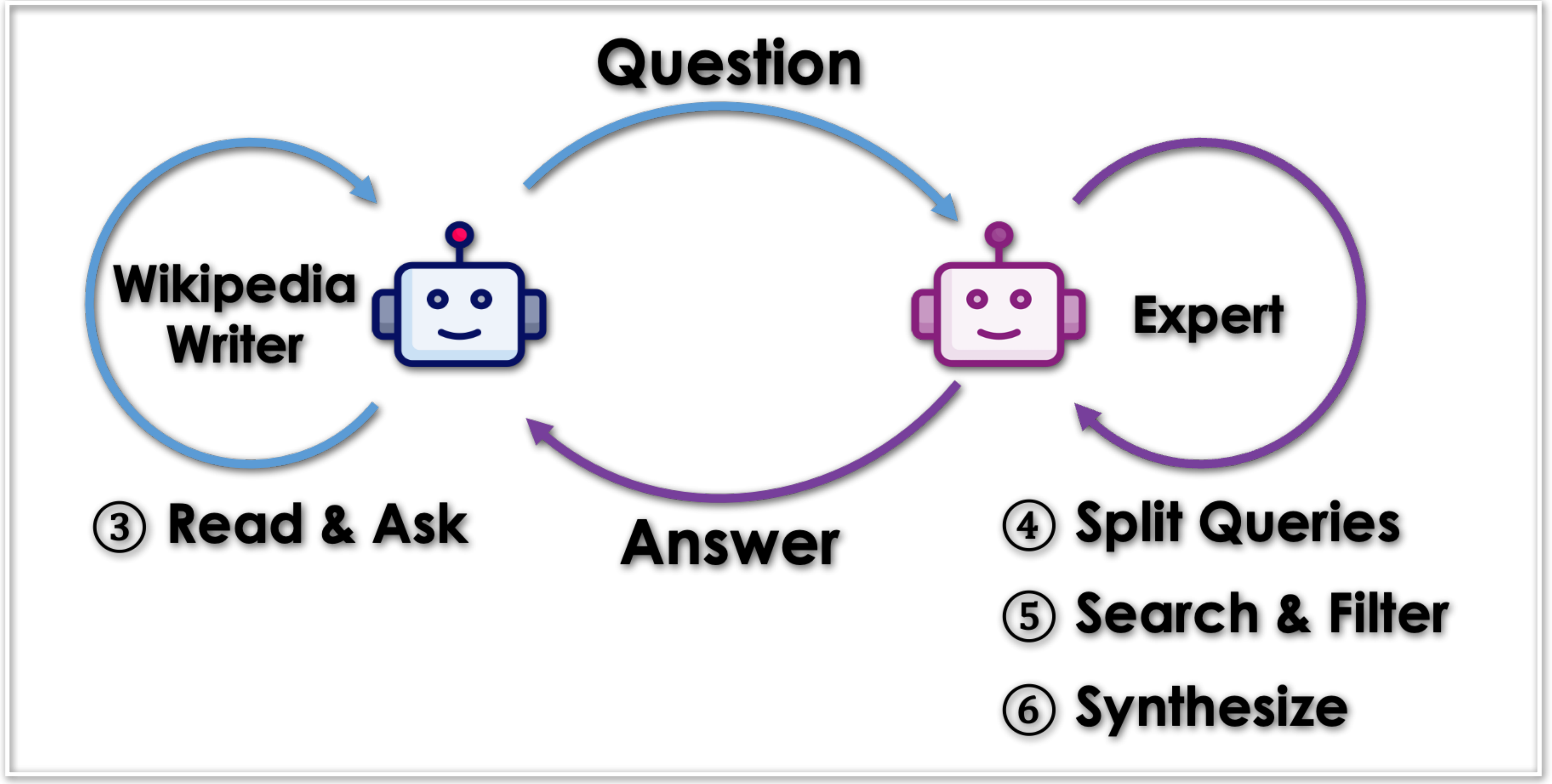
Some in-**depth** questions arise only after reading the information gathered in previous rounds.

Topic: 2022 Winter Olympics Opening Ceremony

Q: Can you provide me with a list of the participating countries in the 2022 Winter Olympics opening ceremony?

A: The 2022 Winter Olympics featured a diverse group of countries... Athletes from over 90 countries will enter the stadium **in a specific order.**

Q: **How is the order** of participating countries in the 2022 Winter Olympics opening ceremony **determined**?



Knowledge Curation

STORM: Conducting meaningful evaluations

What should we evaluate? and how?

Do we have ground truth / golden answer?

Besides the final report, what else should we evaluate?

Knowledge Curation

STORM: Automatic evaluation - Outline quality

Introduce **outline coverage metrics** as a proxy of the pre-writing stage quality for **fast prototyping**

Given the human-written wikipedia article on topic t

1. **Heading soft recall**

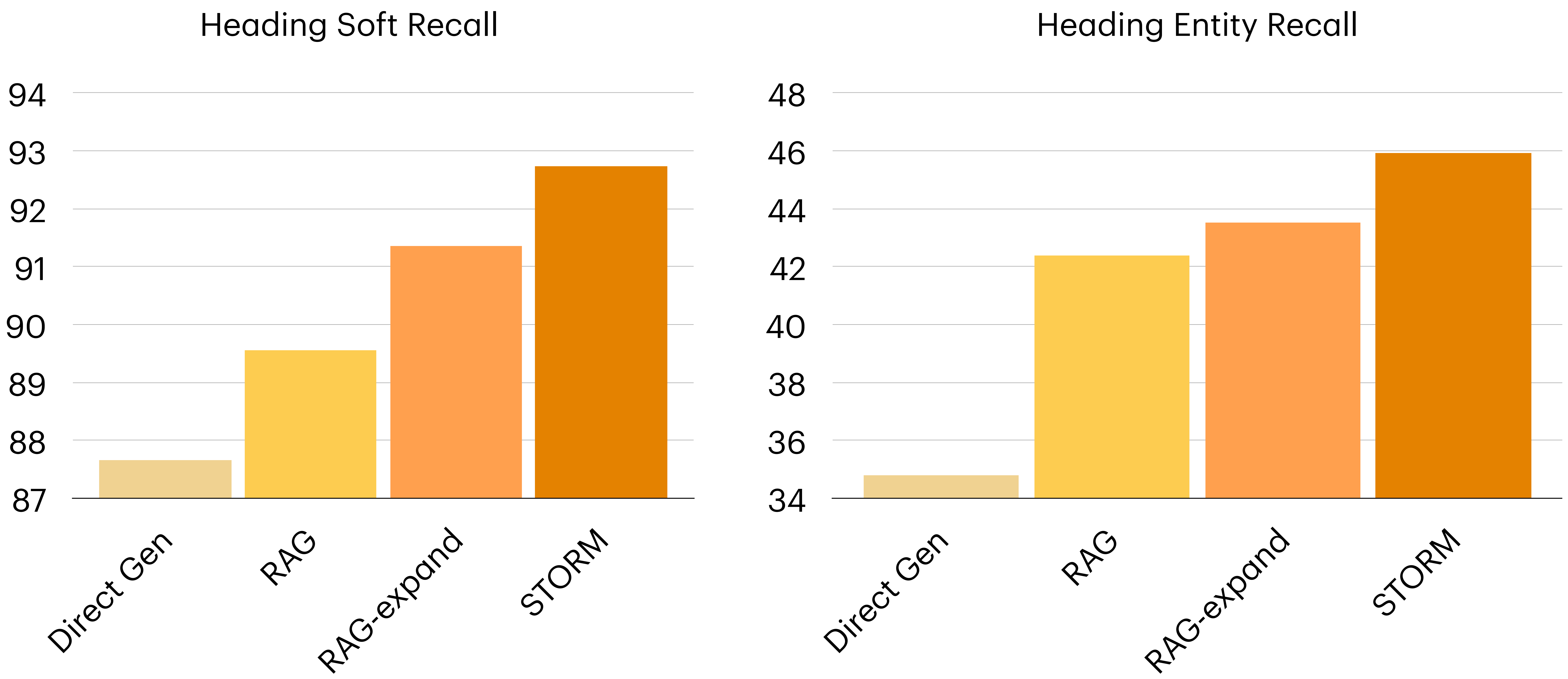
   Compare embedding of headings in $\mathcal{O}$ and the human-written article

2. **Heading entity recall**

   The percentage of named entities in the human-written article covered by $\mathcal{O}$

Knowledge Curation

STORM: Automatic evaluation - Outline quality



Knowledge Curation

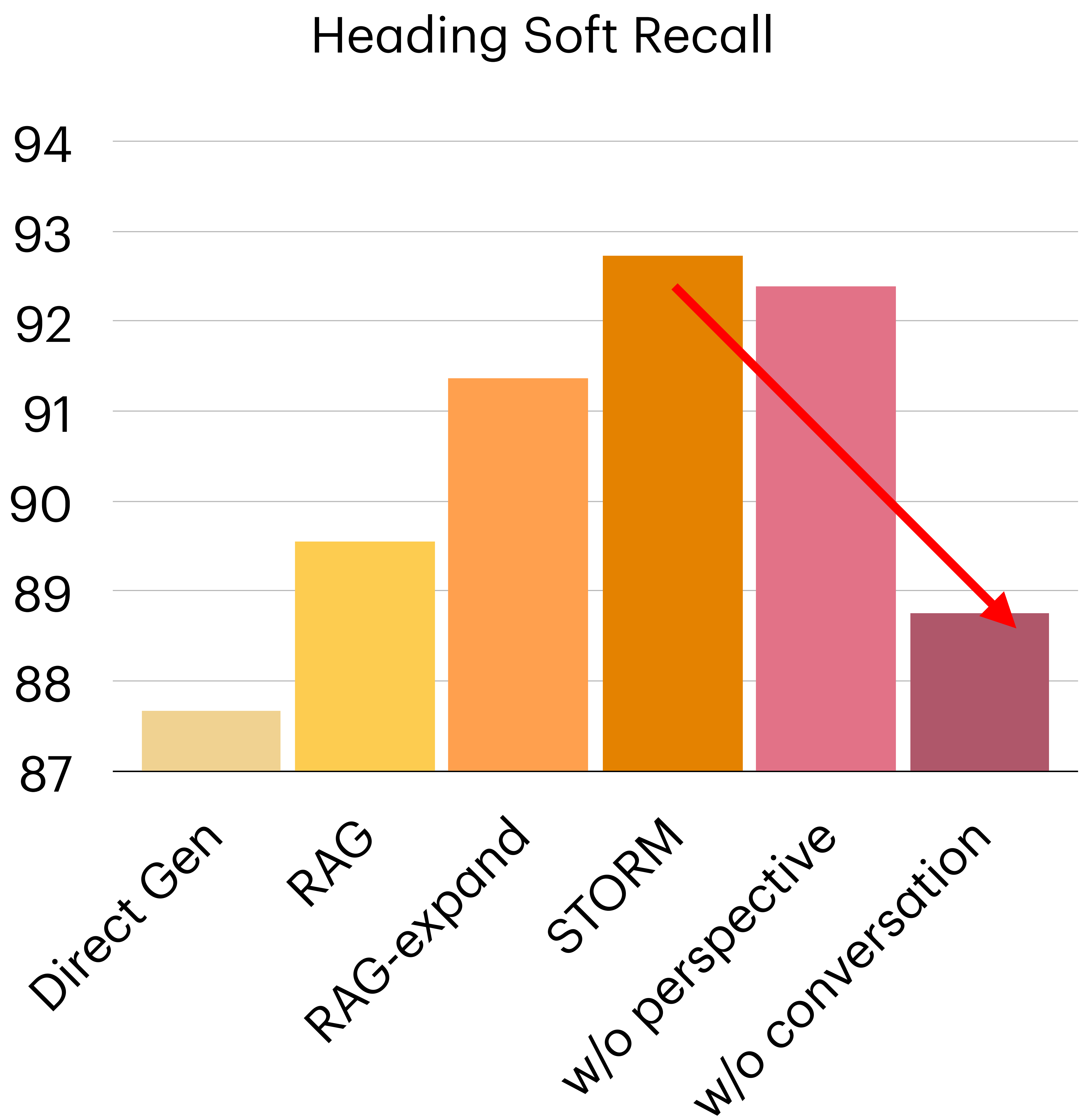
STORM: Automatic evaluation - Outline quality

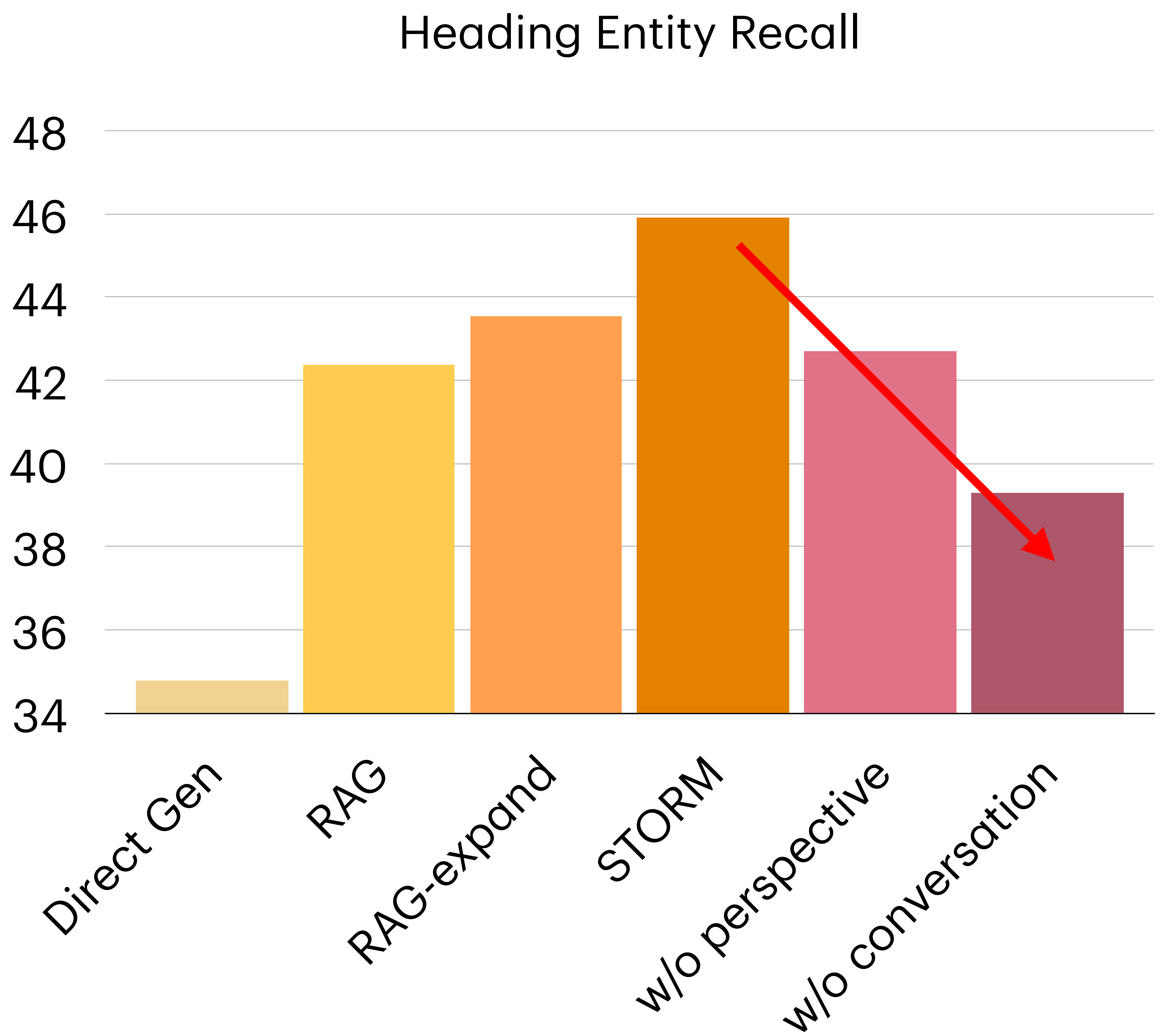
Ablation study is important!

Ablation study help us to understand how different parts of a system contribute to its overall performance

Knowledge Curation

STORM: Automatic evaluation - Outline quality





Knowledge Curation

STORM: Human evaluation - Wikipedia editor evaluation

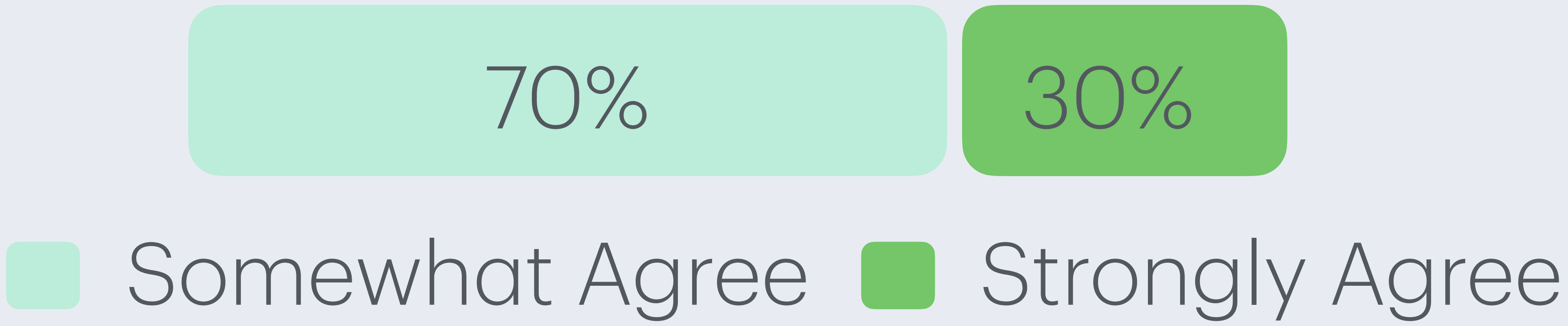
Careful human evaluation is necessary to evaluate LM-empowered systems.

Knowledge Curation

STORM: Human evaluation - Wikipedia editor evaluation

Experienced Wikipedia editors favor articles produced by STORM.

"I (Wikipedia Editor) think it can be specifically helpful for my pre-writing stage."



≥ 4 Rate (1-7 Scale)	Interest Level	Organization	Relevance	Coverage	Verifiability
oRAG	57.5%	45.0%	62.5%	57.5%	67.5%
STORM	**70.0%**	**70.0%**	**65.0%**	**67.5%**	67.5%

Knowledge Curation

STORM: Human evaluation - In the wild evaluation

UIUX design is critical for larger scale human evaluation in the wild

https://storm.genie.stanford.edu

820,000 Users

1,400,000 Articles

2,700,000 Browsing

250,000 Feedbacks

Knowledge Curation

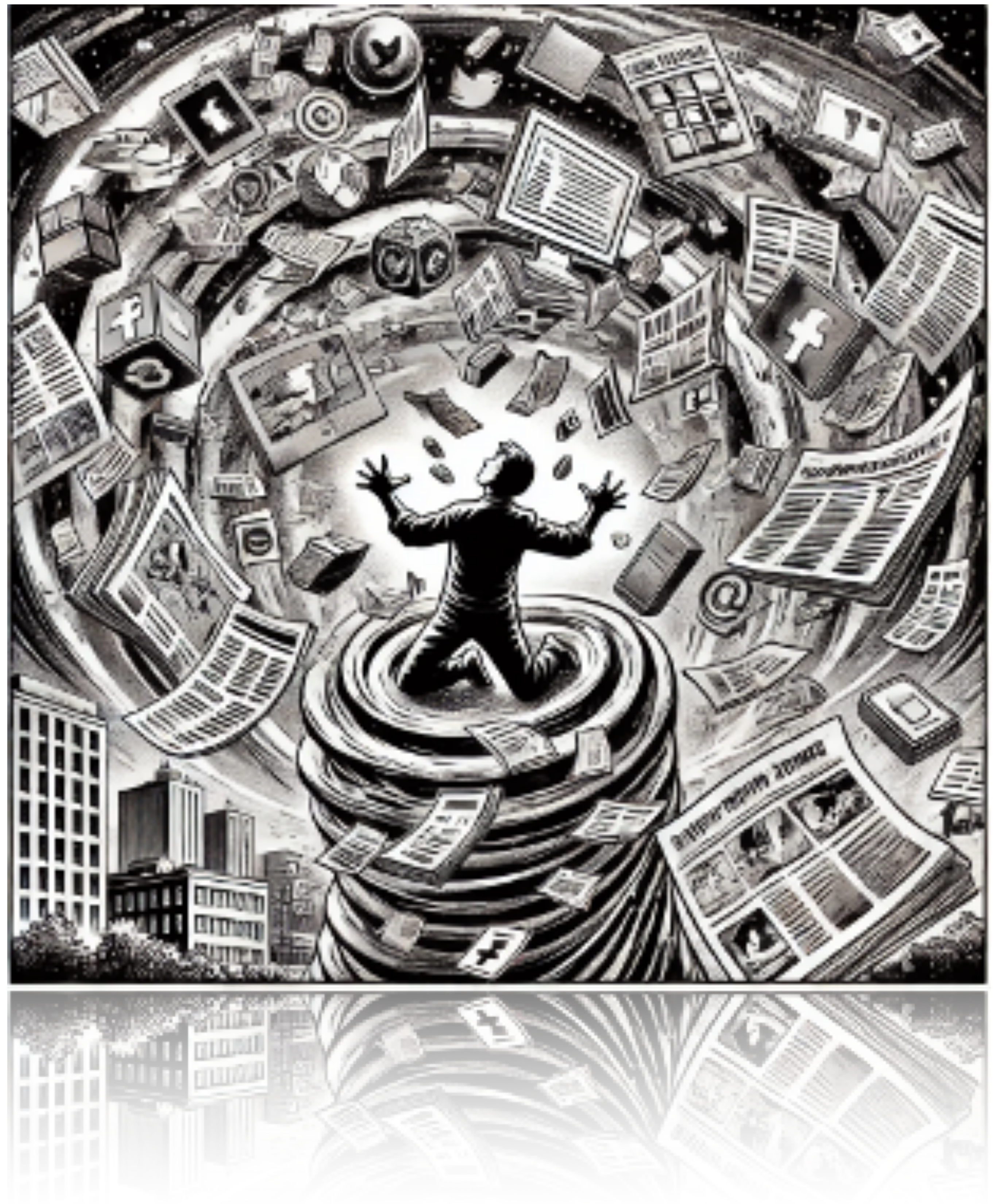
STORM: Human evaluation - In the wild evaluation

People have used STORM across a diverse range of topics & use cases

Agriculture	Fitness	Design	Management
Computer Science	Medicare	Gaming	Business
Environment science	Law	Music	Animal science
Biology	News	Food	Transportation
Physics	Politics	Travel	Emergency management
Geology	Cultural study	Education	...

Revisit: Meta question

Are people's information needs satisfied?

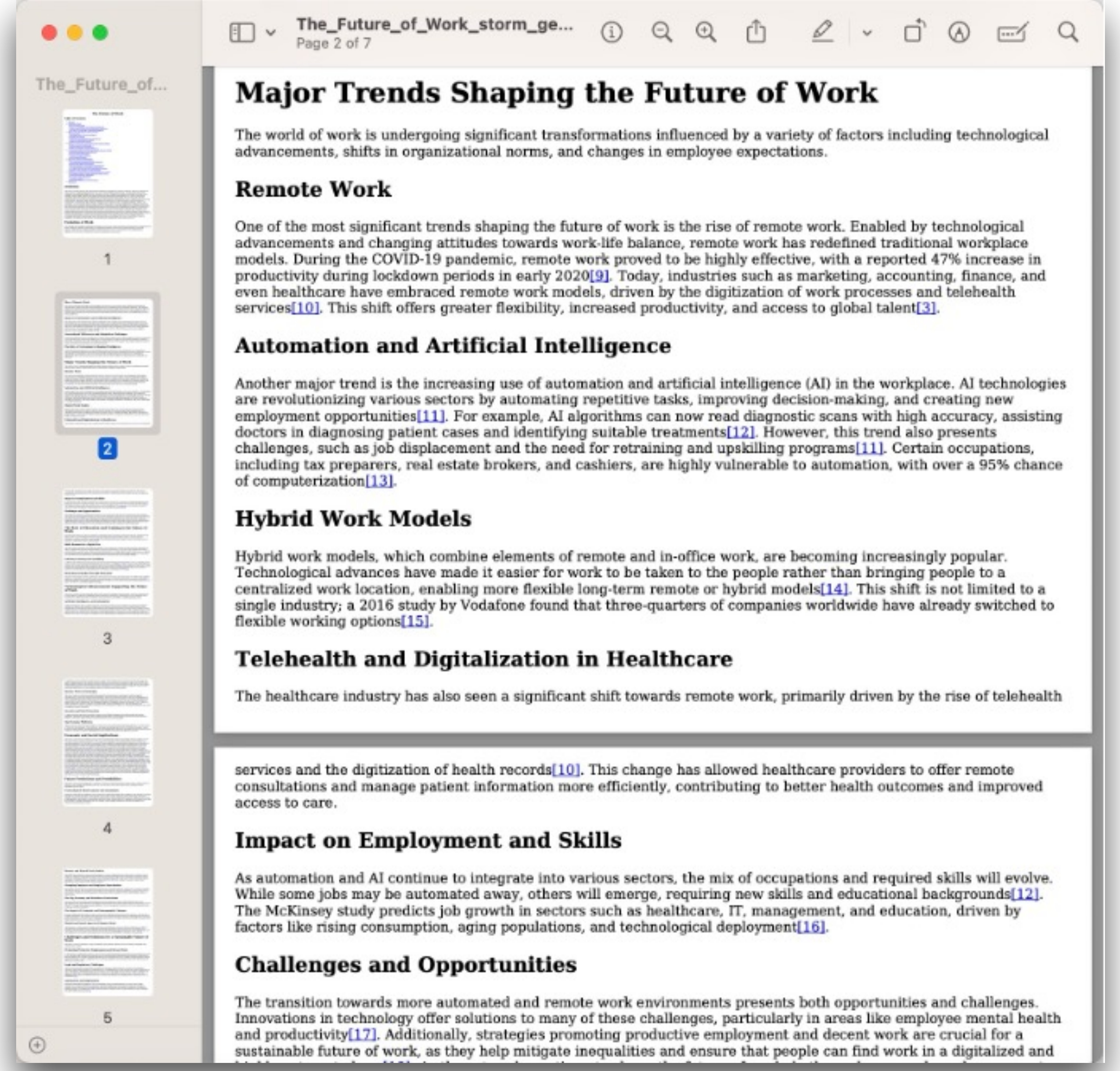


The illustration is co-created with DALL-E.

Revisit: Meta question

Are people's information needs satisfied?

STORM


The_Future_of_Work_storm_ge...
Page 2 of 7

The_Future_of...

Major Trends Shaping the Future of Work

The world of work is undergoing significant transformations influenced by a variety of factors including technological advancements, shifts in organizational norms, and changes in employee expectations.

Remote Work

One of the most significant trends shaping the future of work is the rise of remote work. Enabled by technological advancements and changing attitudes towards work-life balance, remote work has redefined traditional workplace models. During the COVID-19 pandemic, remote work proved to be highly effective, with a reported 47% increase in productivity during lockdown periods in early 2020[9]. Today, industries such as marketing, accounting, finance, and even healthcare have embraced remote work models, driven by the digitization of work processes and telehealth services[10]. This shift offers greater flexibility, increased productivity, and access to global talent[3].

Automation and Artificial Intelligence

Another major trend is the increasing use of automation and artificial intelligence (AI) in the workplace. AI technologies are revolutionizing various sectors by automating repetitive tasks, improving decision-making, and creating new employment opportunities[11]. For example, AI algorithms can now read diagnostic scans with high accuracy, assisting doctors in diagnosing patient cases and identifying suitable treatments[12]. However, this trend also presents challenges, such as job displacement and the need for retraining and upskilling programs[11]. Certain occupations, including tax preparers, real estate brokers, and cashiers, are highly vulnerable to automation, with over a 95% chance of computerization[13].

Hybrid Work Models

Hybrid work models, which combine elements of remote and in-office work, are becoming increasingly popular. Technological advances have made it easier for work to be taken to the people rather than bringing people to a centralized work location, enabling more flexible long-term remote or hybrid models[14]. This shift is not limited to a single industry; a 2016 study by Vodafone found that three-quarters of companies worldwide have already switched to flexible working options[15].

Telehealth and Digitalization in Healthcare

The healthcare industry has also seen a significant shift towards remote work, primarily driven by the rise of telehealth services and the digitization of health records[10]. This change has allowed healthcare providers to offer remote consultations and manage patient information more efficiently, contributing to better health outcomes and improved access to care.

Impact on Employment and Skills

As automation and AI continue to integrate into various sectors, the mix of occupations and required skills will evolve. While some jobs may be automated away, others will emerge, requiring new skills and educational backgrounds[12]. The McKinsey study predicts job growth in sectors such as healthcare, IT, management, and education, driven by factors like rising consumption, aging populations, and technological deployment[16].

Challenges and Opportunities

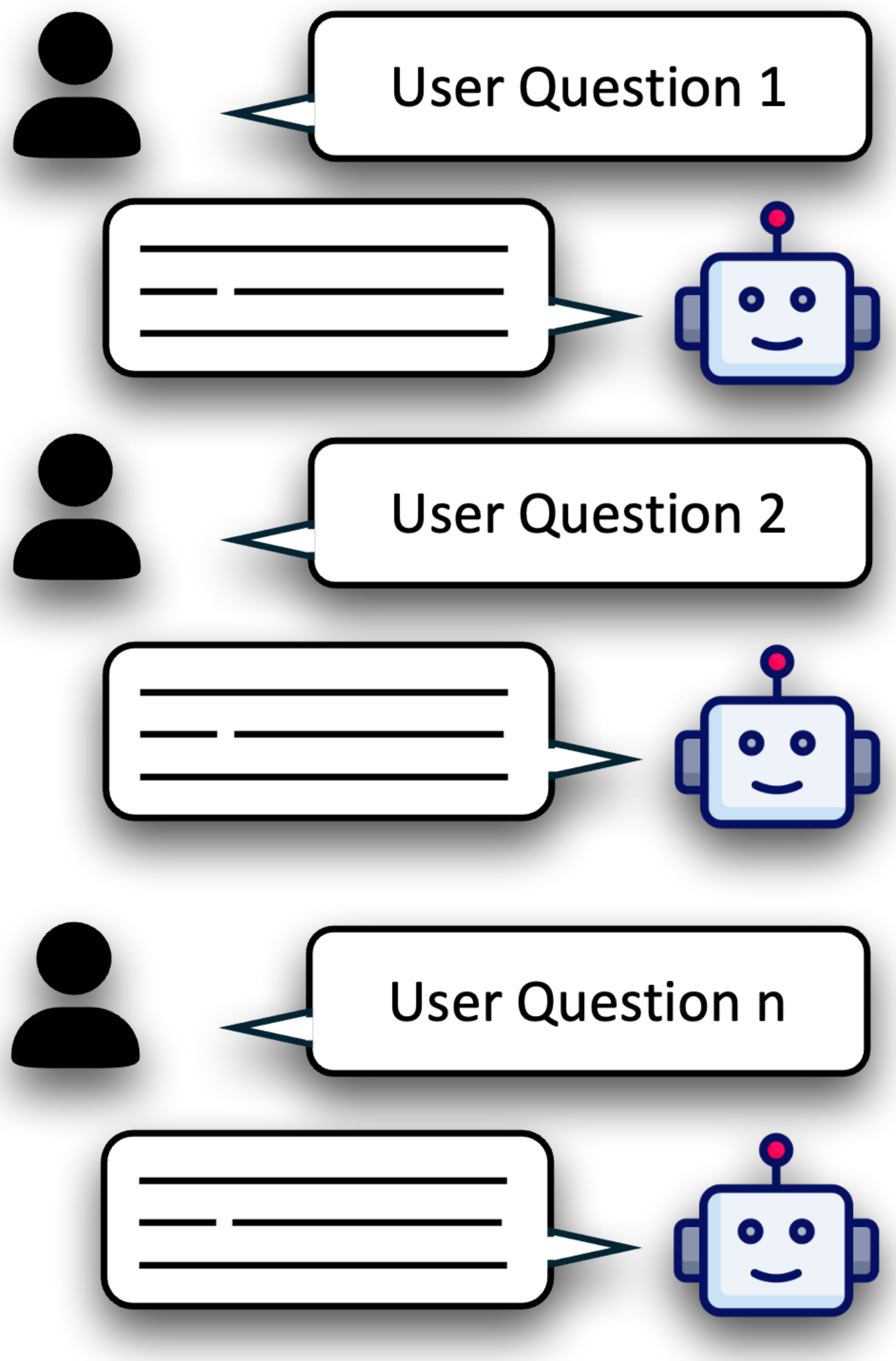
The transition towards more automated and remote work environments presents both opportunities and challenges. Innovations in technology offer solutions to many of these challenges, particularly in areas like employee mental health and productivity[17]. Additionally, strategies promoting productive employment and decent work are crucial for a sustainable future of work, as they help mitigate inequalities and ensure that people can find work in a digitalized and

My thoughts evolve, so I want to update my queries.

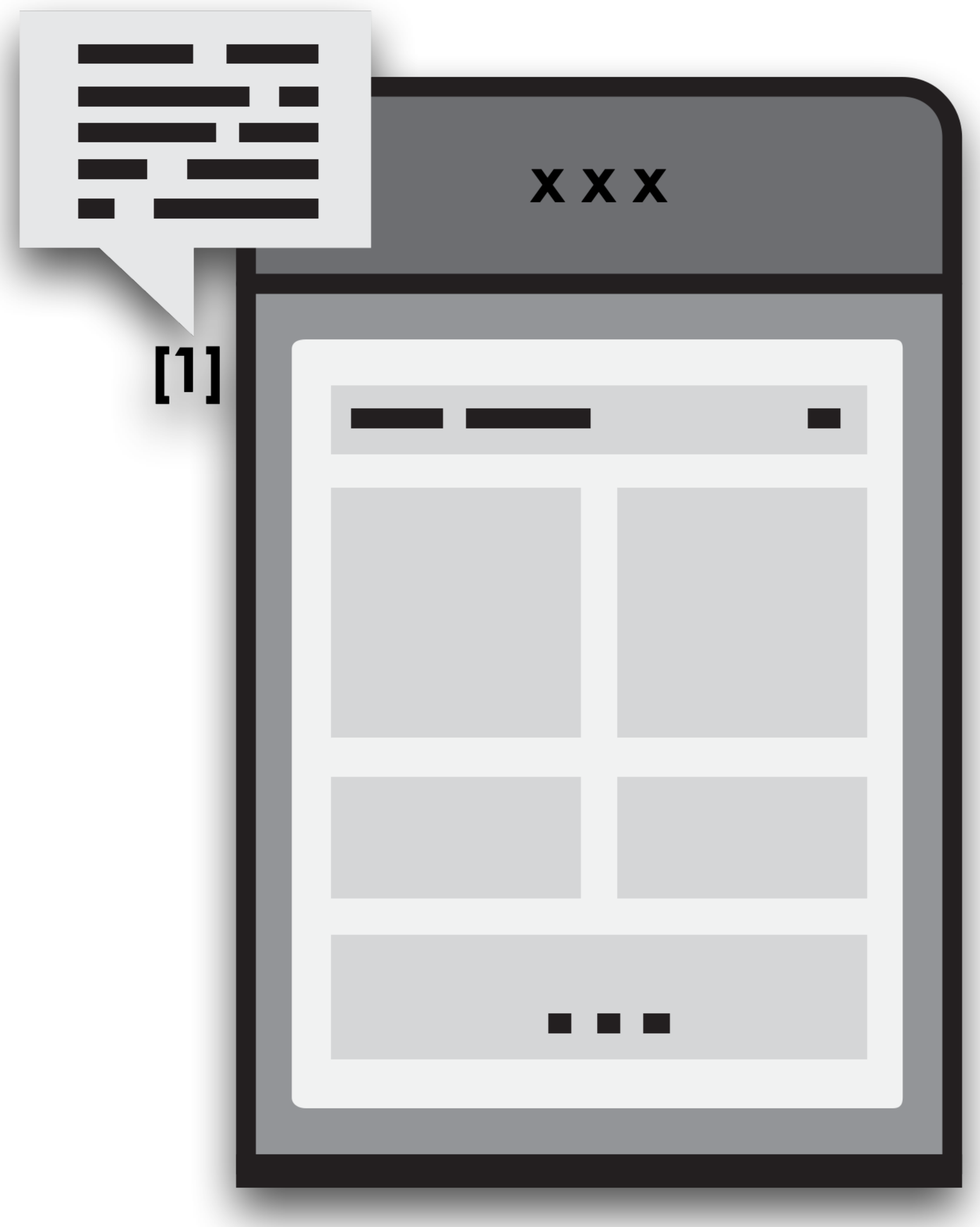
I am inspired by this link and hope to learn more about it.

I know this topic is also relevant. Can you include it?

Human-AI interaction/collaboration

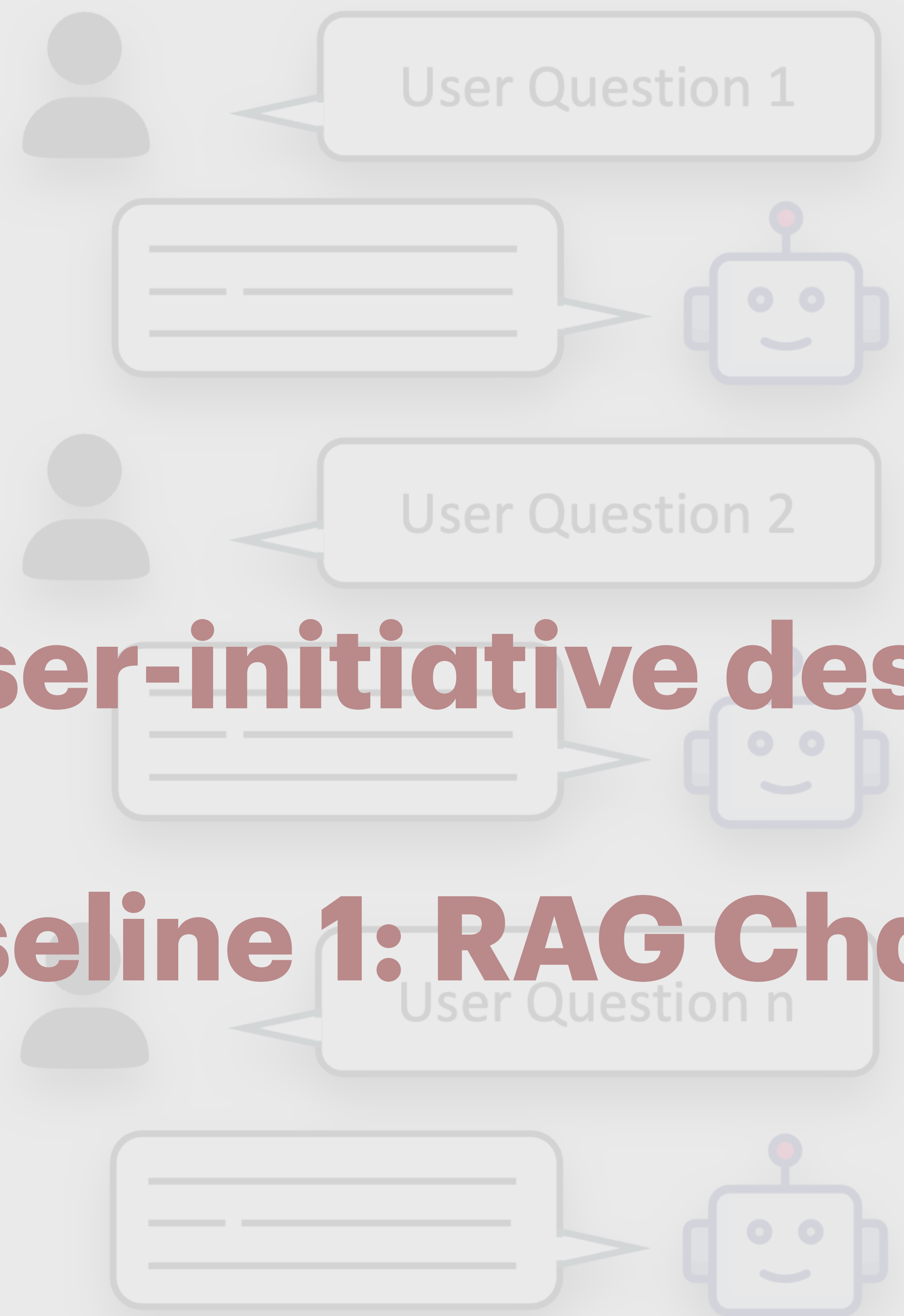


Convert STORM into a hallucination-free question answering system



After the long report is generated, allow the user to edit or ask questions.

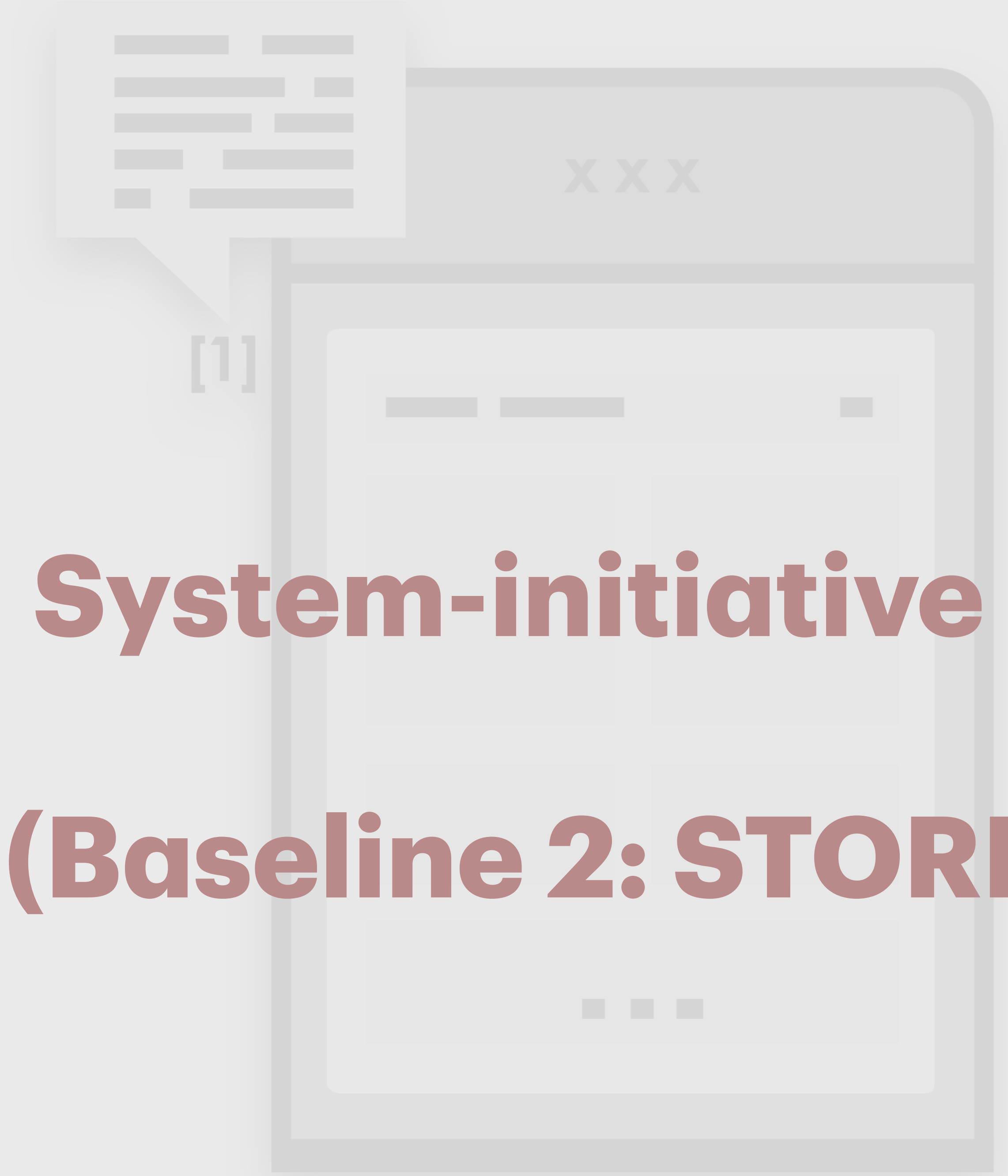
Human-AI interaction/collaboration



User-initiative design

(Baseline 1: RAG Chatbot)

Convert STORM into a hallucination-free question answering system



System-initiative design

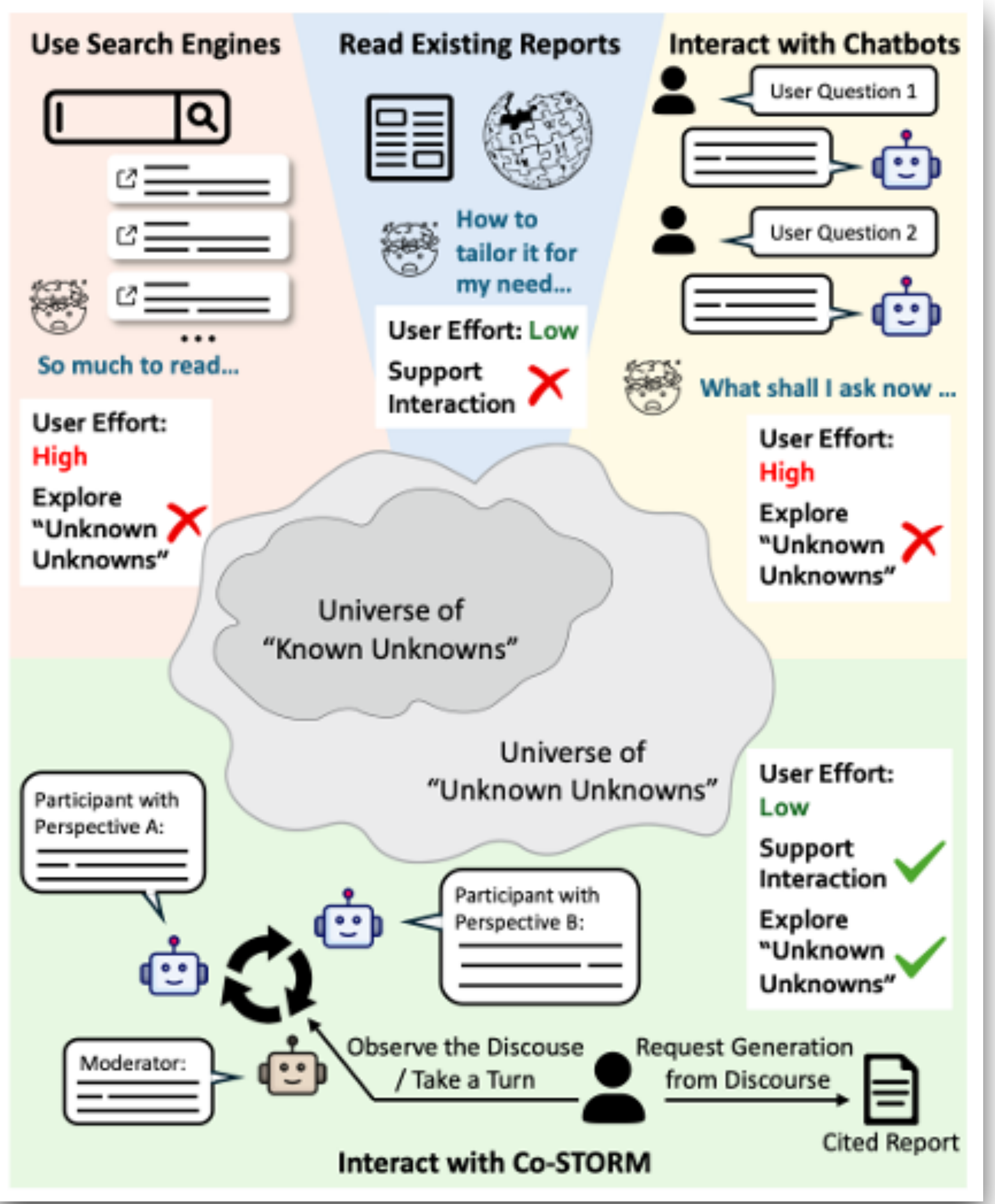
(Baseline 2: STORM + QA)

After the long report is generated, allow the user to edit or ask questions.

Human-AI interaction/collaboration

CoSTORM:
Engaged human learning through Participation in LM agent conversations

Jiang, Yucheng, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. "Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations.", In EMNLP 2024



Human-AI interaction/collaboration

CoSTORM: Human learning, unknown unknowns discovery

Key Idea: Mimic Human Learning Process

How do children/students learn?

Nussbaum, 2008: **Collaborative discourse** and collaborative argumentation is important for promoting students' deep-level understanding of contents.

How do humans retain information?

Buzan, 1974: Using **mind map** for note taking to help recall and critical thinking.

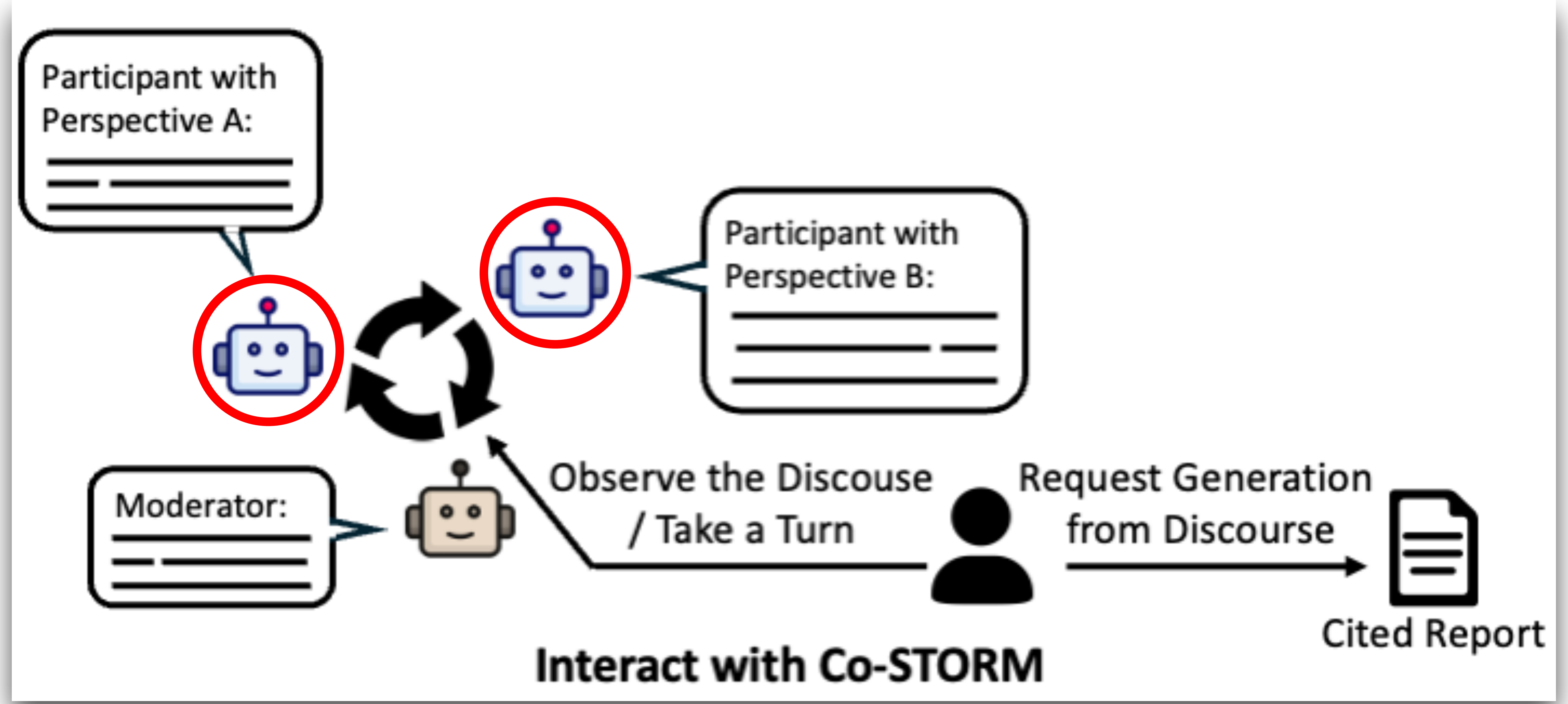
Human-AI interaction/collaboration

CoSTORM: Collaborative Discourse Protocol Design

Agents form a roundtable, **answering** and **asking** questions grounded on external sources

The user can jump in at any time to steer the discourse and inject questions and opinions.

Maintains a dynamic, hierarchical mind map so users can easily follow and engage.



Input Optional Design

Human-AI interaction/collaboration

CoSTORM: Collaborative Discourse Protocol Design

However, the agent almost always choose **question answering**, causing the conversation to focus on a narrow topic, which can result in **overly niche content**.

How do human ask follow up questions during information seeking?

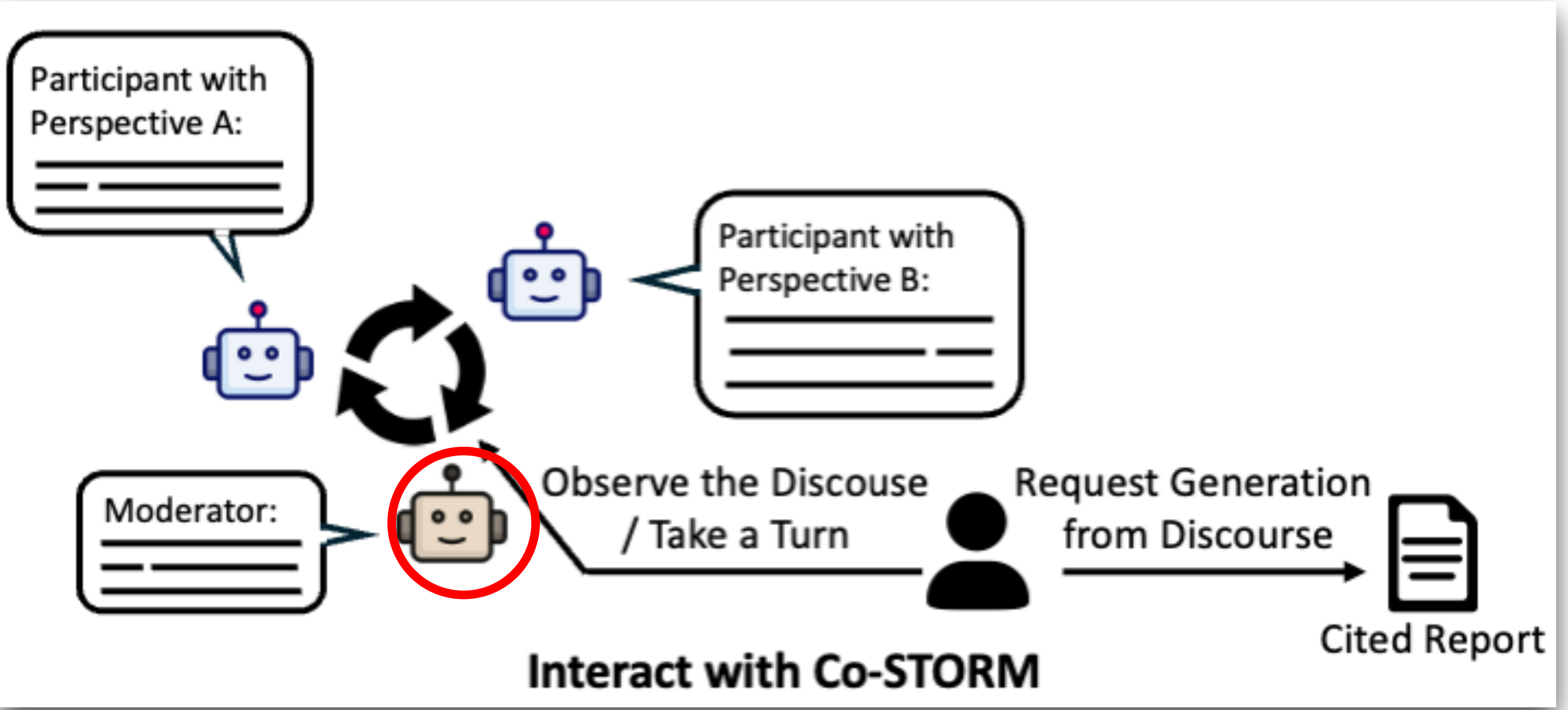
Serendipidity: We may discover topics not directly related but particularly interesting

For example, when we search for “improving software engineering practices”, we might stumble upon an article about “the cognitive psychology behind team decision-making”.

Human-AI interaction/collaboration

CoSTORM: Collaborative Discourse Protocol Design

Solution: Ask thought provoking questions via **Moderator** role



Human-AI interaction/collaboration

CoSTORM: Collaborative Discourse Protocol Design

Solution: Ask thought provoking questions via **Moderator** role

Step 1: Extract and rerank unused information throughout discourse history

$$cos(i, t)^{\alpha}[1 - cos(i, q)]^{1-\alpha}$$

(where i, q, t are embeddings of the information, question, and topic)

Step 2: Generate thought-provoking questions & polish the utterance

Human-AI interaction/collaboration

CoSTORM: Conducting meaningful evaluations

What should we evaluate? and how?

Do we have ground truth / golden answer?

Besides the final report, what else should we evaluate?

Human-AI interaction/collaboration

CoSTORM: Automatic evaluation - Discourse quality

	Question-Answering Turn Quality		
	Consistency	Engagement	# Unique URLs
RAG Chatbot	4.37	4.13	2.94
STORM + QA	4.34	4.11	2.89
Co-STORM	**4.40**†	**4.33**†	**6.04**†
w/o Multi-Expert	**4.40**	4.32	5.91
w/o Moderator	4.39	4.28	5.67

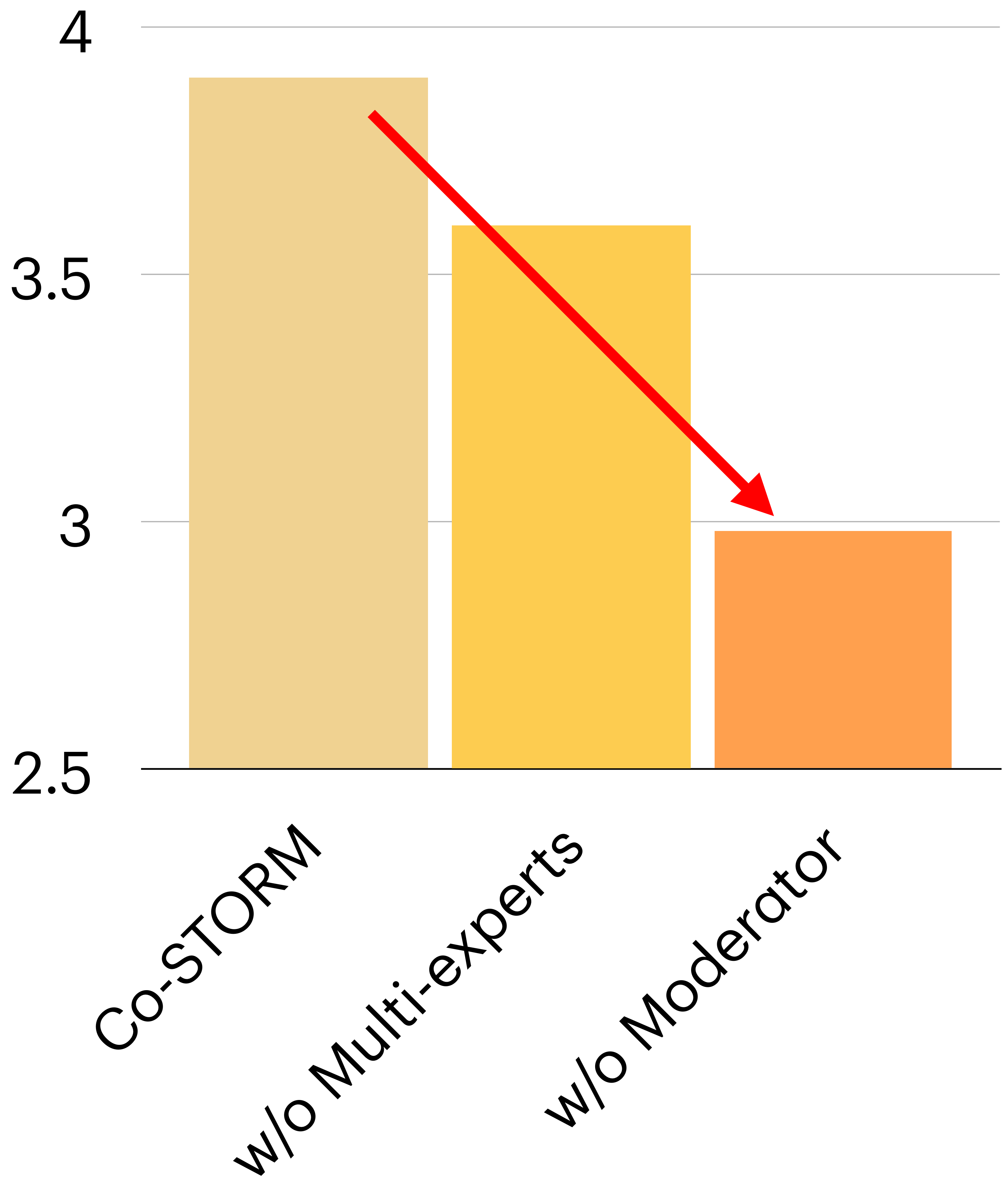
Human-AI interaction/collaboration

CoSTORM: Automatic evaluation - Final report quality

	Report Quality				
	Relevance	Breadth	Depth	Novelty	Info Diversity
RAG Chatbot	3.57	3.50	3.26	2.44	0.595
STORM + QA	3.61	3.61	3.43	2.50	0.592
Co-STORM	**3.78**	**3.79**	**3.77†**	**3.05†**	**0.602**
w/o Multi-Expert	3.73	3.75	**3.77**	2.93	0.589
w/o Moderator	3.56	3.69	3.41	2.89	0.577

Human-AI interaction/collaboration

CoSTORM: Automatic evaluation - Ablation study



	# User	# expert agent	# moderator
Co-STORM	1	N	1
w/o multi-experts	1	**1**	1
w/o multi-agent	1	N	**0**

Having just one expert and one moderator can already provide most of the benefits

Human-AI interaction/collaboration

CoSTORM: Human Evaluation

	Co-STORM v.s. Search Engine				Co-STORM v.s. RAG Chatbot			
	Search Engine	Co-STORM	Win % (Lose %)	p-value	RAG Chatbot	Co-STORM	Win % (Lose %)	p-value
Relevance	3.90	**4.00**	30% (30%)	0.758	3.89	**4.22**	33% (0%)	0.081
Breadth	3.60	**4.10**	50% (10%)	0.096	3.11	**4.22**	67% (0%)	0.013
Depth	3.10	**4.00**	60% (10%)	0.081	3.11	**4.00**	56% (33%)	0.069
Serendipity	2.70	**3.90**	70% (10%)	0.030	2.78	**3.78**	67% (0%)	0.009

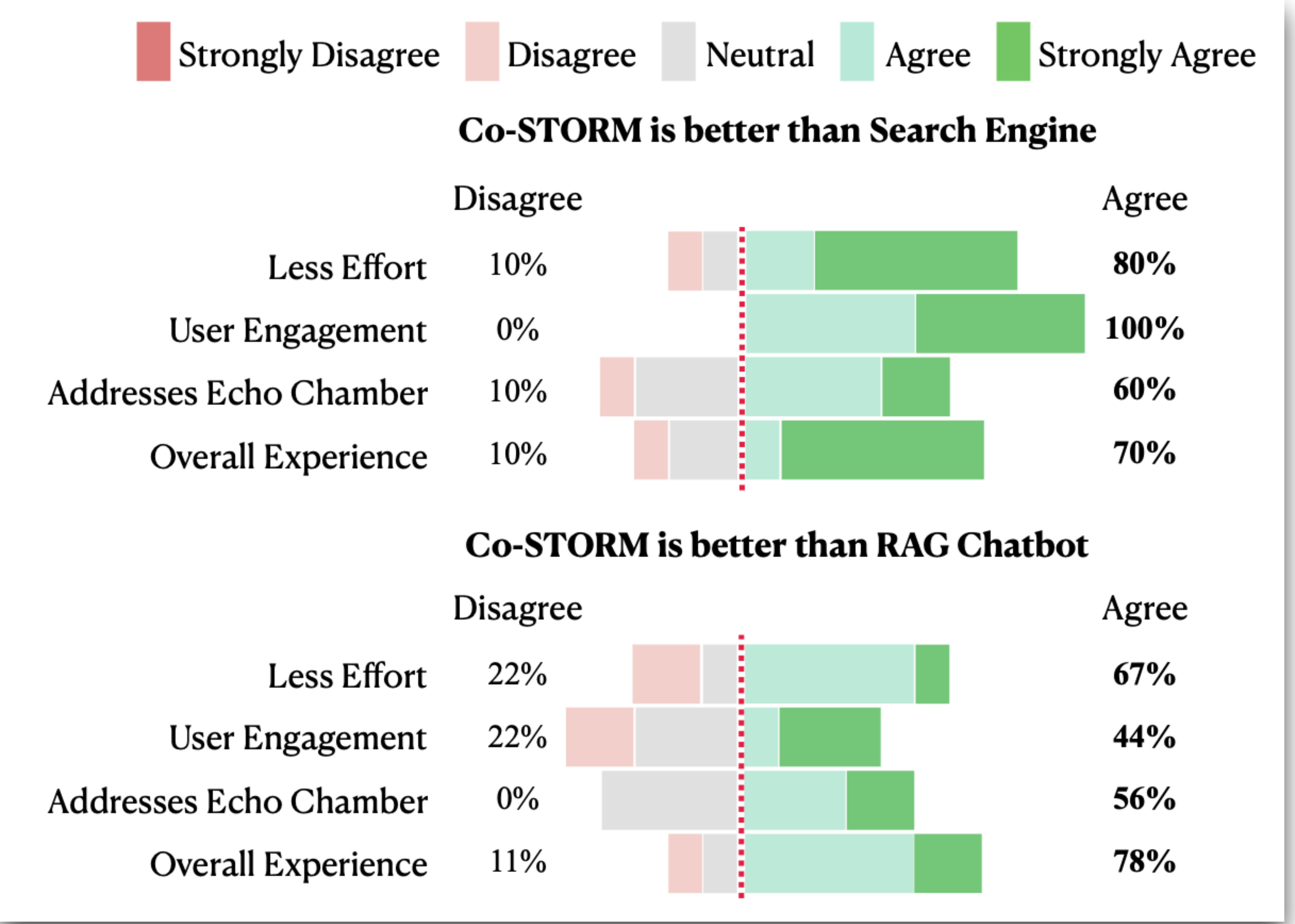
Table 4: Human ratings on different aspects of the information-seeking experience with Co-STORM and Search Engine (n=10) and with Co-STORM and RAG Chatbot (n=9)[6]. The ratings are given on a scale from 1 to 5 with 3 as "Average". We report the win rate of Co-STORM in pairwise comparison and the p-value in a paired t-test.

Human-AI interaction/collaboration

CoSTORM: Automatic evaluation - Human Evaluation

Co-STORM allows for almost full automation and much better understanding as it brings up topics that the user may not even think of.

“Co-STORM is so much less mentally taxing for me to use”



DataSTORM and HW 1

HW1 Overview

STORM and other deep research systems focus on **literature search (or literature review of research)** and summarizing existing information.

In HW1 we will go further—conducting **original research** to produce an investigative journalism article on a real-time world conflict.

DataSTORM and HW 1

HW1 Overview - Provided building blocks

1. **Internet based literature search (similar to STORM)**

2. **Database exploration agent (DataSTORM)**
   Given a topic and an initial set of research questions, it interacts with the database, autonomously generating and refining questions, retrieving answers, and returning a curated set of interesting search results throughout the process.

DataSTORM and HW 1

HW1 Algorithm overview

Algorithm 1: HW1

Input : Topic t

Output: Report R

Existing building blocks

```
B ← RunLiteratureSearch(t) ;                          // Background report on topic
Q_s ← GenerateSeedQuestions(B) ;                      // Initial research questions
D ← DatabaseExploration(t, Q_s) ;                     // Curated retrieved results
T ← GenerateTheses(D) ;                               // Set of proposed theses
τ* ← SelectBestThesis(T) ;                            // Select most promising thesis
S ← RunLiteratureSearch(τ*) ;                         // Evidence supporting selected thesis
R ← ConsolidateFindings(B, D, τ*, S) ;                // Comprehensive final report
return R;
```

DataSTORM and HW 1

HW1: Your tasks

Implement these small building blocks

Algorithm 1: HW1

Input : Topic t
Output: Report R

$B \leftarrow$ RunLiteratureSearch(t) ; // Background report on topic
$Q_s \leftarrow$ $\boxed{\text{GenerateSeedQuestions}(B)\ ;}$ // Initial research questions
$D \leftarrow$ DatabaseExploration(t, Q_s) ; // Curated retrieved results
$T \leftarrow$ $\boxed{\text{GenerateTheses}(D)\ ;}$ // Set of proposed theses
$\tau^* \leftarrow$ $\boxed{\text{SelectBestThesis}(T)\ ;}$ // Select most promising thesis
$S \leftarrow$ RunLiteratureSearch(τ^*) ; // Evidence supporting selected thesis
$R \leftarrow$ $\boxed{\text{ConsolidateFindings}(B, D, \tau^*, S)\ ;}$ // Comprehensive final report
return R;

Takeaways

Build LM-empowered systems.
An emerging paradigm in the era of foundation models.

Crafting LM pipelines resembles how we observe human workflows.
STORM resembles how human write.
Co-STORM resembles collaborative discourse in education.

Conduct user study in addition to automatic evaluation.
STORM invites 20 Wikipedia editors during paper writing.
Co-STORM invites 20 users in the wild during paper writing
STORM & Co-STORM deployed in the wild, tested by over 800,000 users.

Questions

Feel free to reach out to **yuchengj@stanford.edu** for more questions/thoughts.