

Stanford CS224v Course

Conversational Virtual Assistants with Deep Learning

Lecture 2

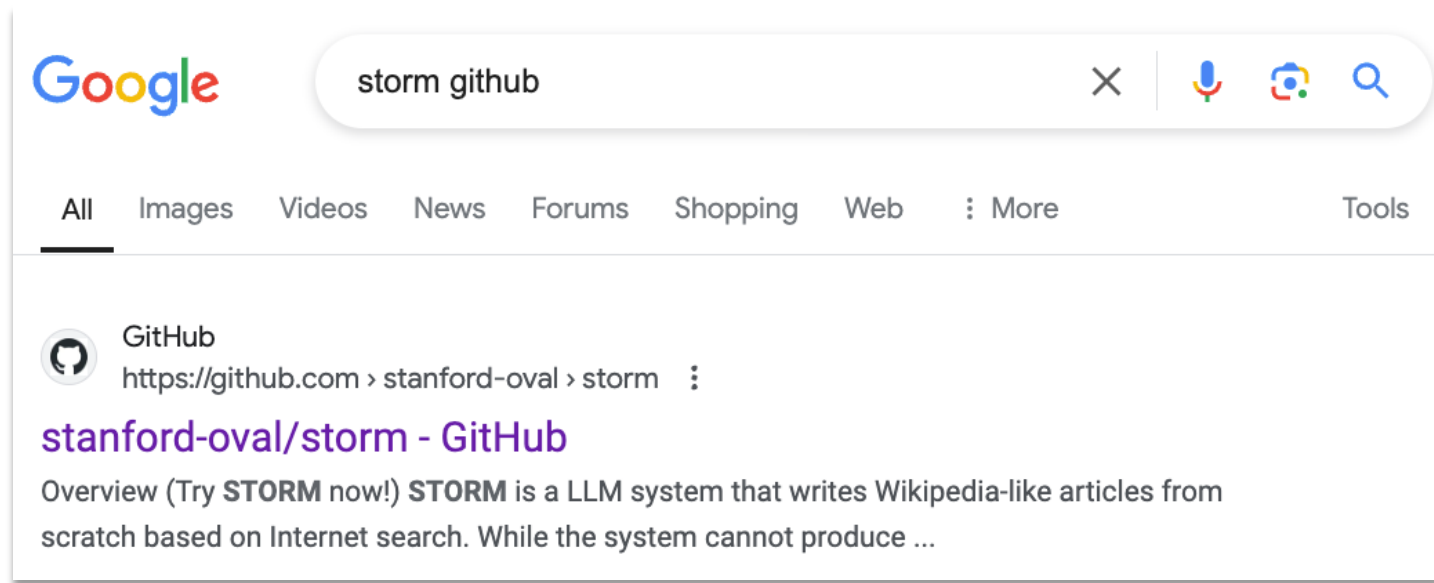
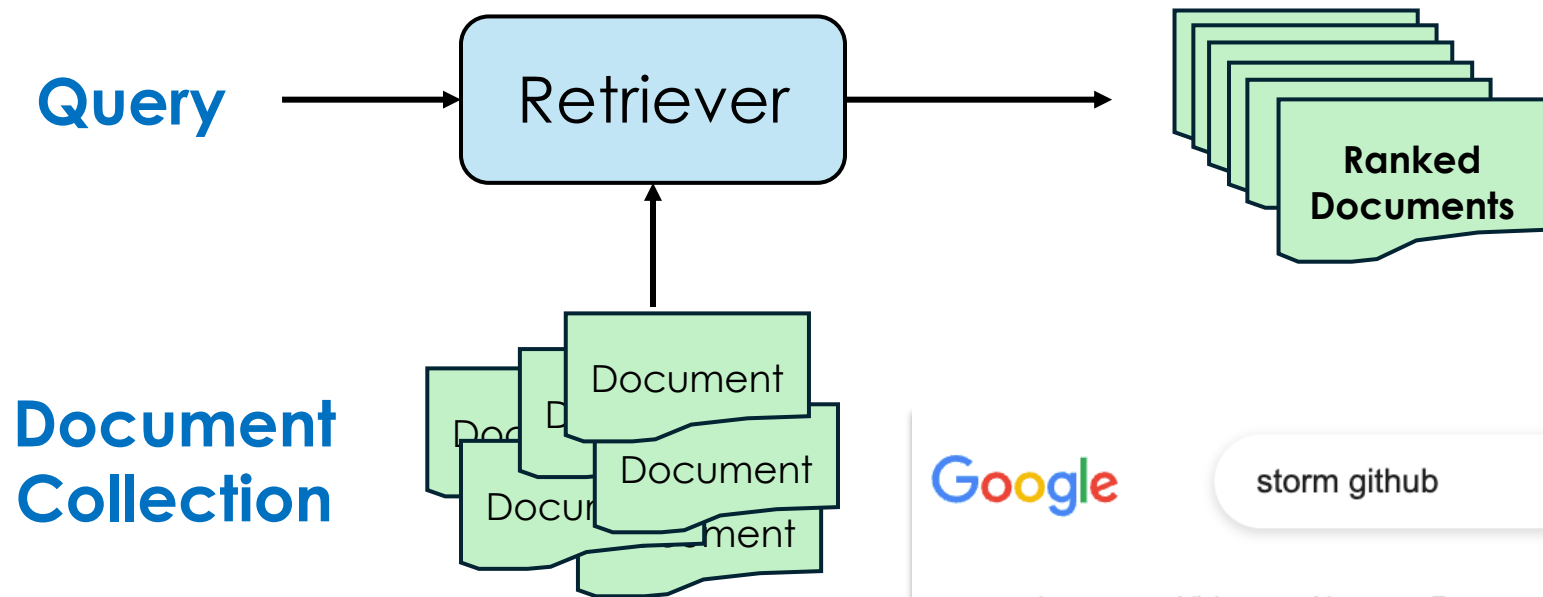
Knowledge Curation

Yijia Shao, Yucheng Jiang, Monica Lam

Lecture Plan

- 1. Information Retrieval & RAG at a Glance**
2. From Retrieving Information to Knowledge Curation
3. Building LM-Empowered Systems
4. Evaluation? Evaluation!
5. Bring Human into the Loop

Information Retrieval at a Glance





ChatGPT 4 ▾ On 9/15/2024



Where did Yijia Shao obtain her bachelor's degree?



Yijia Shao obtained her bachelor's degree from National Taiwan University.



A major issue of using language models for knowledge tasks:

Hallucination

- Long-tail information
- Knowledge cutoffs
- Private data

Retrieval-Augmented Generation (RAG)

RAG is Widely Used in Question Answering



Yijia Shao



Stanford University

<https://cs.stanford.edu> > people > shaoyj

About | Yijia Shao's Personal Website

I am a 1st-year PhD student at Stanford NLP advised by Diyi Yang. I had the pleasure to work with Monica S. Lam and Michael Bernstein during the rotation ...

Step 1: Retrieve

Where did Yijia Shao obtain her bachelor's degree?

You search "Yijia Shao" on Google and here is the retrieved information:

Source: <https://cs.stanford.edu/people/shaoyj/>

Snippet: I am a 1st-year PhD student at Stanford NLP advised by Diyi Yang. I had the pleasure to work with Monica S. Lam and Michael Bernstein during the rotation program. Previously, I was an undergraduate student at Yuanpei College in Peking University where I got into ML and NLP research by working with Bing Liu. In summer 2022, I had a research internship in UCLA hosted by Nanyun Peng. Before that, I have worked as a research intern in Microsoft Research Asia (blog spotlight in Chinese) and an engineering intern in Tensorflow Lite team at Google, Beijing.

Step 2: Read & Answer

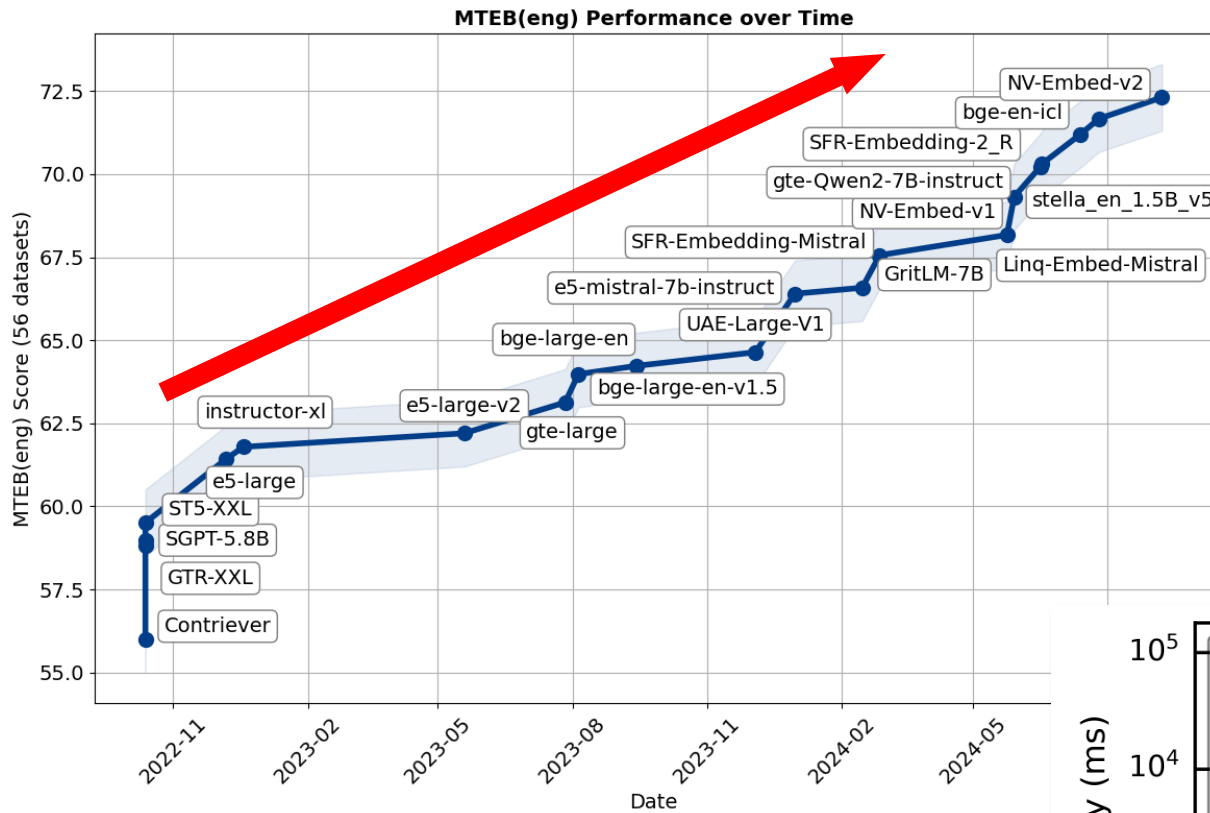
Recommended Reading:

- [Chen et al., 2017](#)
- [Lewis et al., 2021](#)



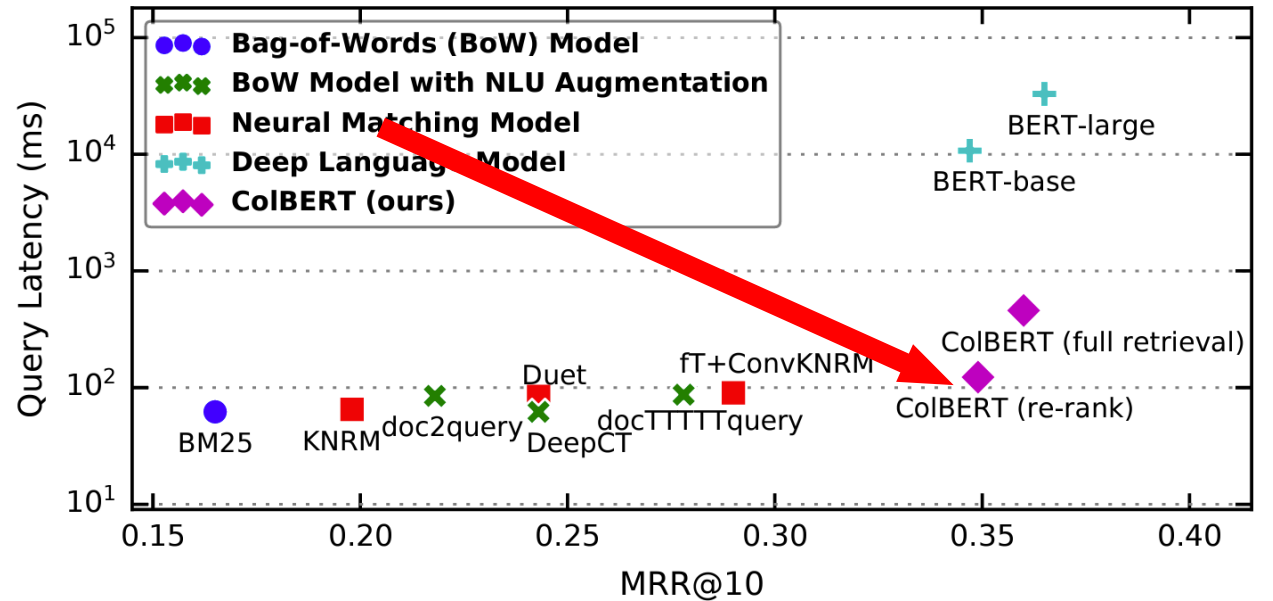
Yijia Shao obtained her bachelor's degree from Yuanpei College at Peking University.



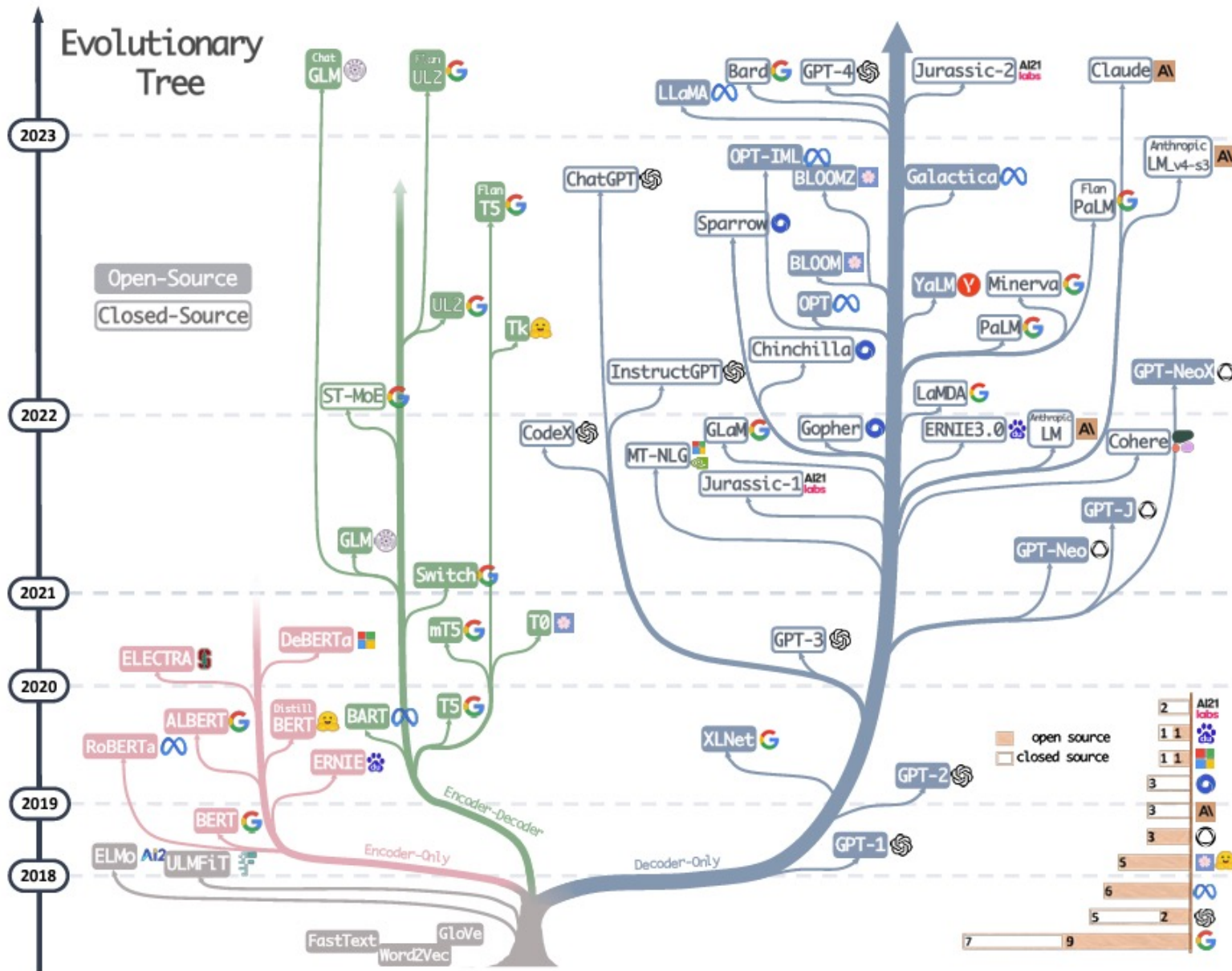


MTEB: Massive Text Embedding Benchmark, Niklas et al. 2023
(This illustration is contributed by Niklas Muennighoff.)

We have better embedding models and infrastructure for Information Retrieval.



ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, Khattab et al. 2020



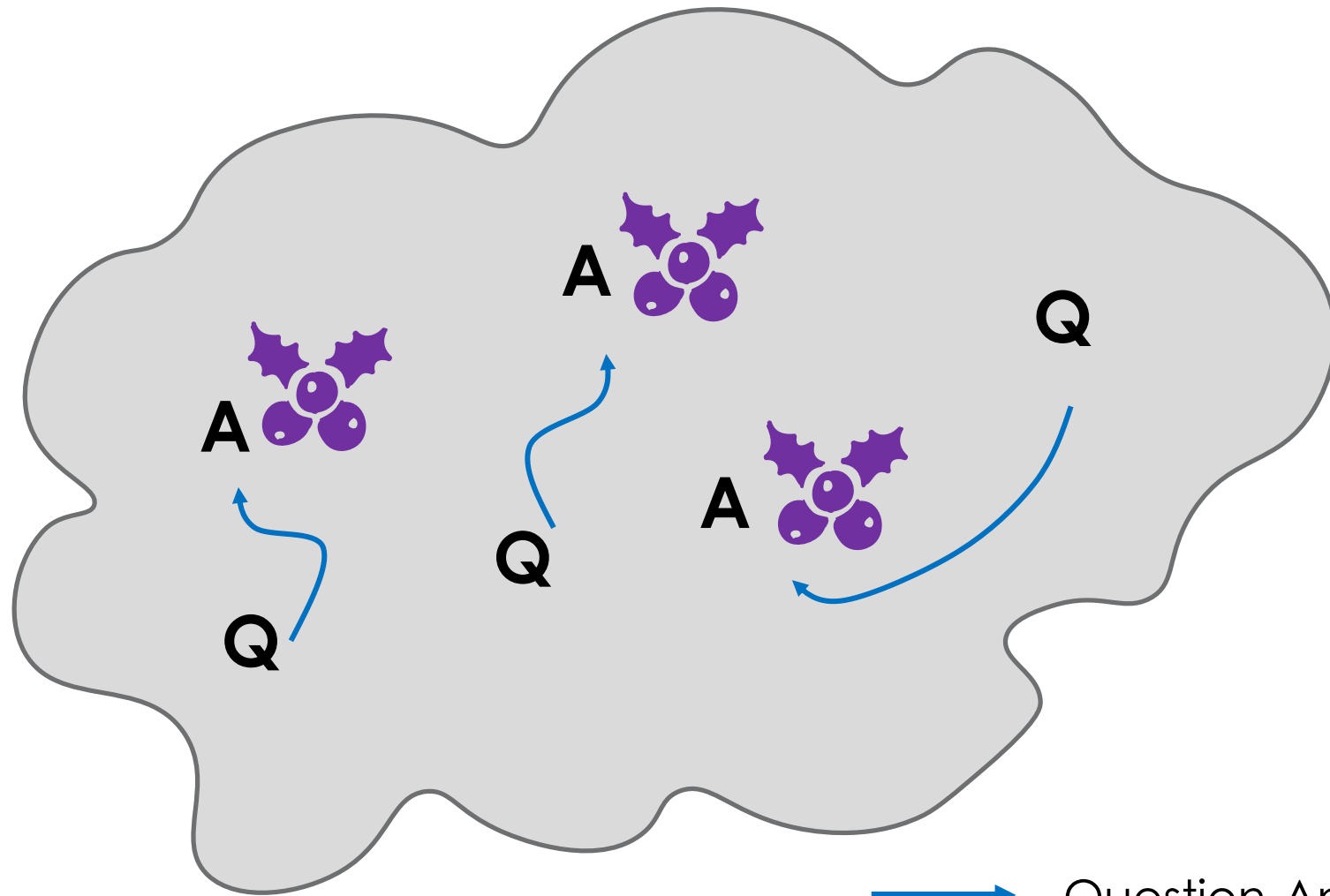
We have stronger language models for generation.

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, Yang et al., 2023

Meta Question:

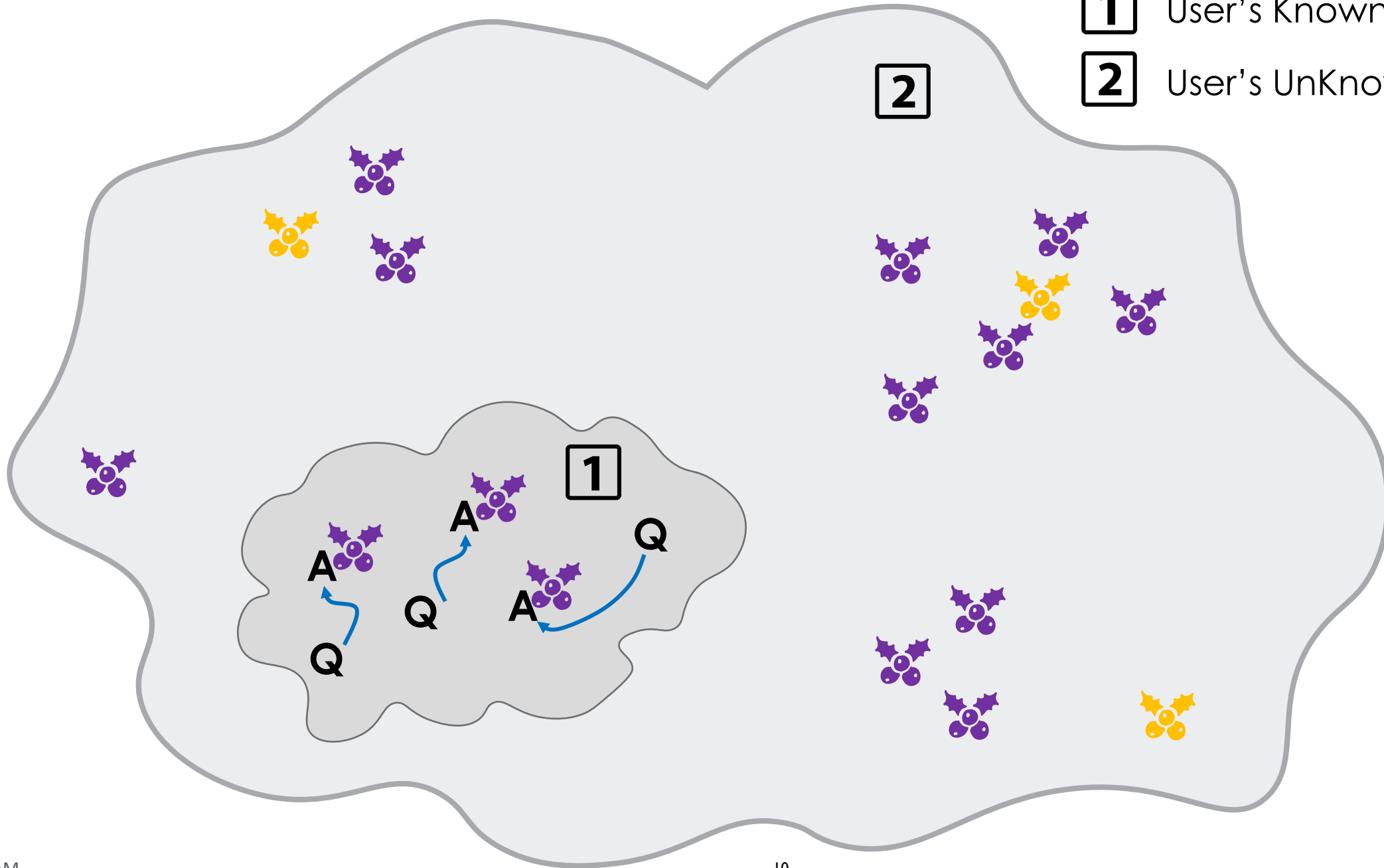
Are people's information needs satisfied?



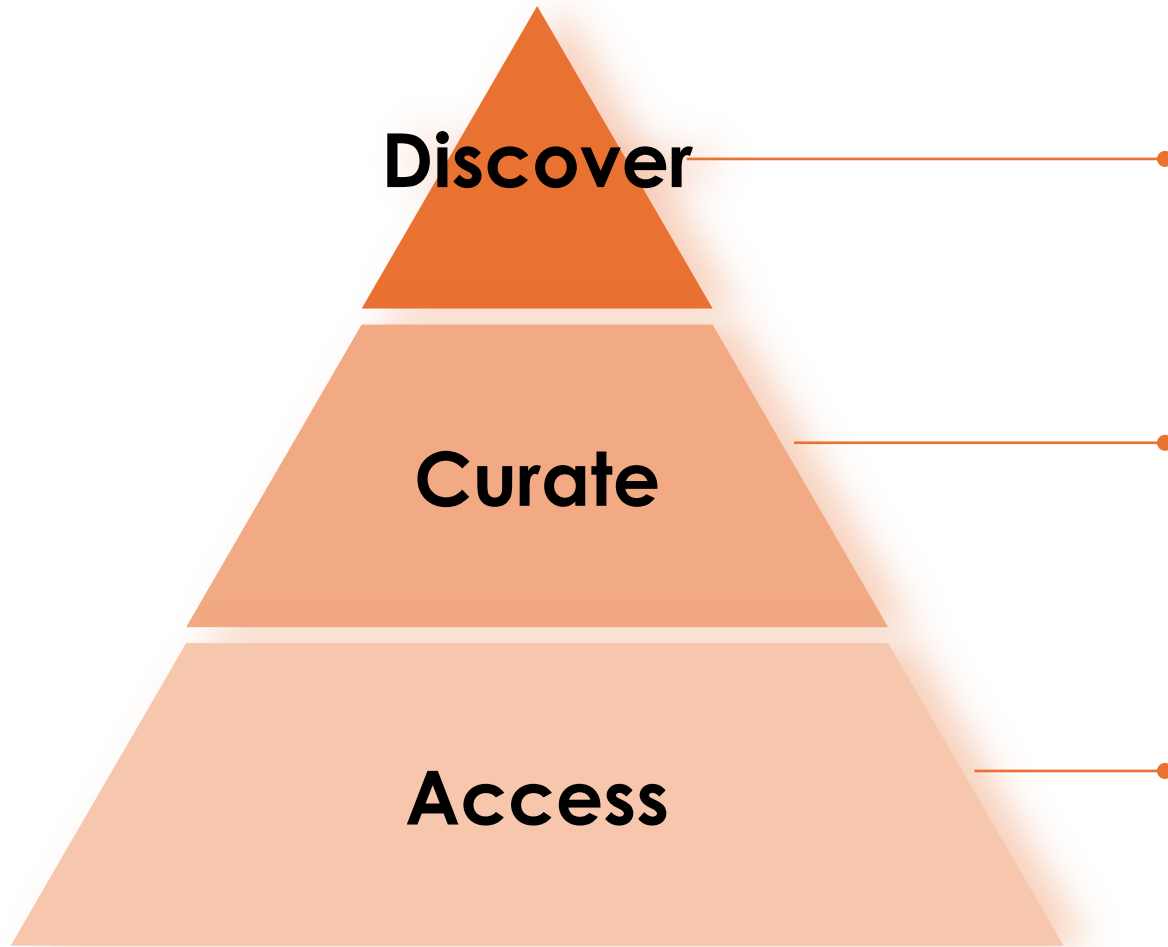


→ Question-Answering System

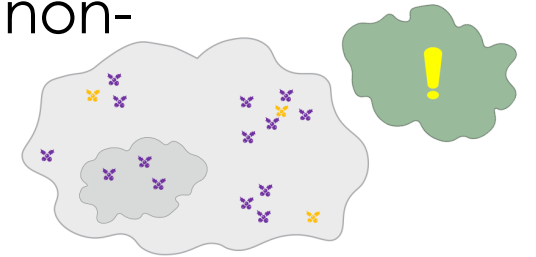
- 1** User's Known Unknowns
- 2** User's UnKnown Unknowns



2. From Retrieving Information to Knowledge Curation



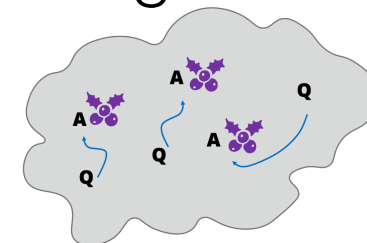
Discover previously non-existent information



Find, organize, and present relevant and valuable content



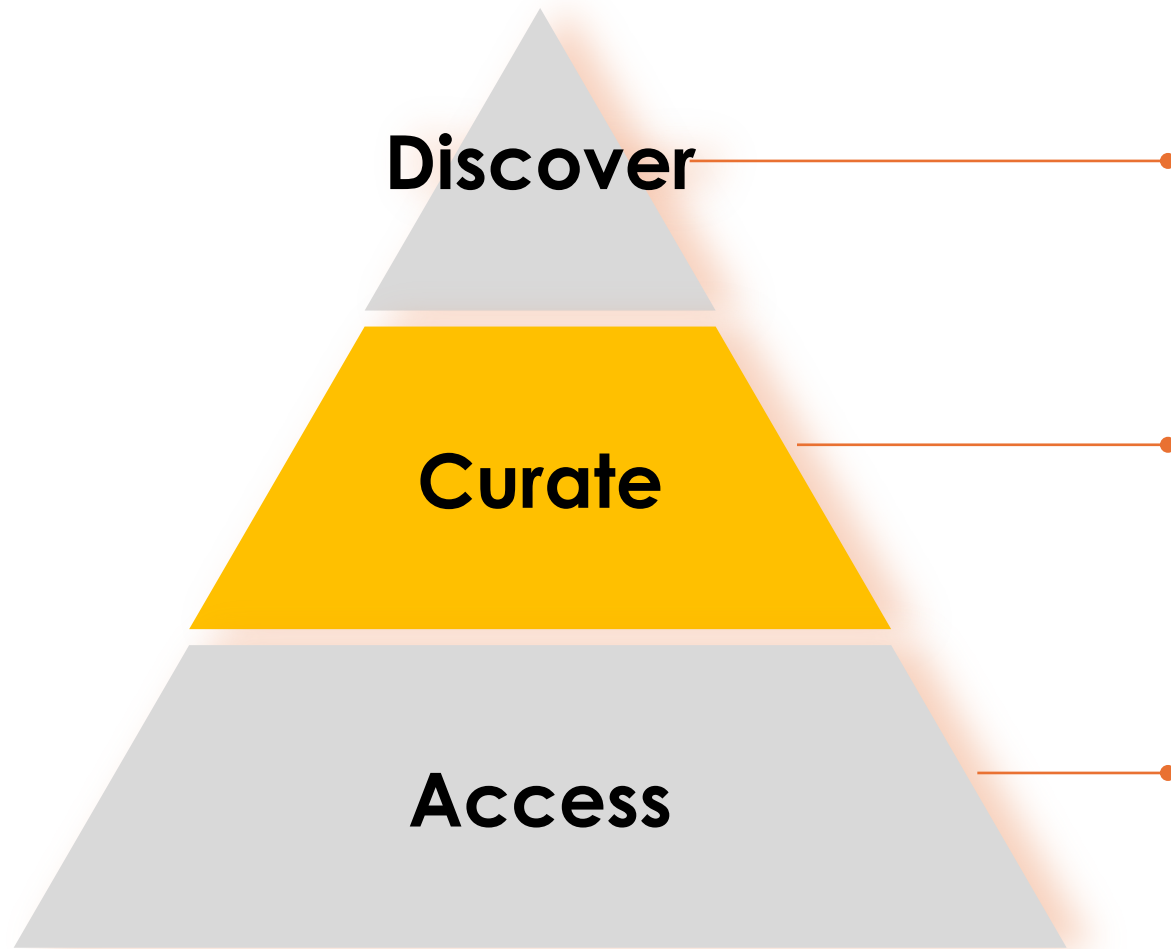
Information Retrieval, Question Answering



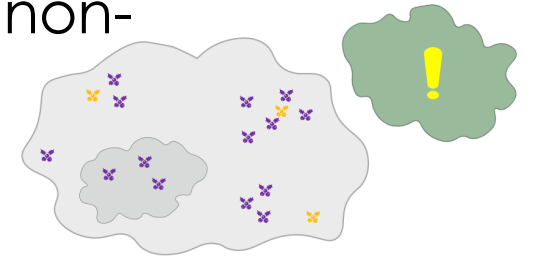
Lecture Plan

1. Information Retrieval & RAG at a Glance
- 2. From Retrieving Information to Knowledge Curation**
3. Building LM-Empowered Systems
4. Evaluation? Evaluation!
5. Bring Human into the Loop

2. From Retrieving Information to Knowledge Curation



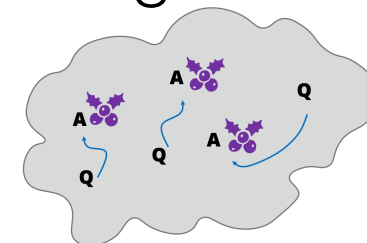
Discover previously non-existent information

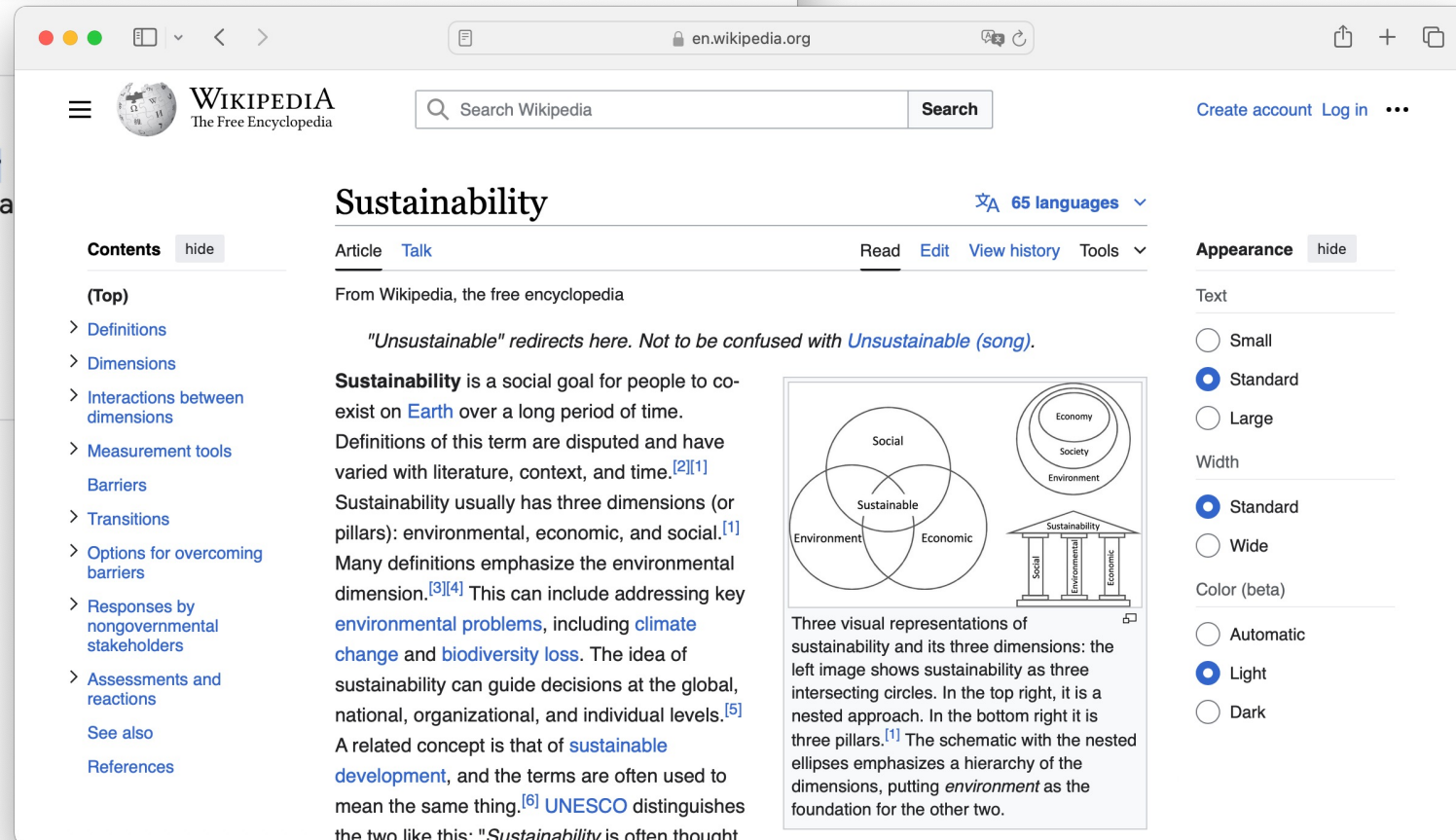
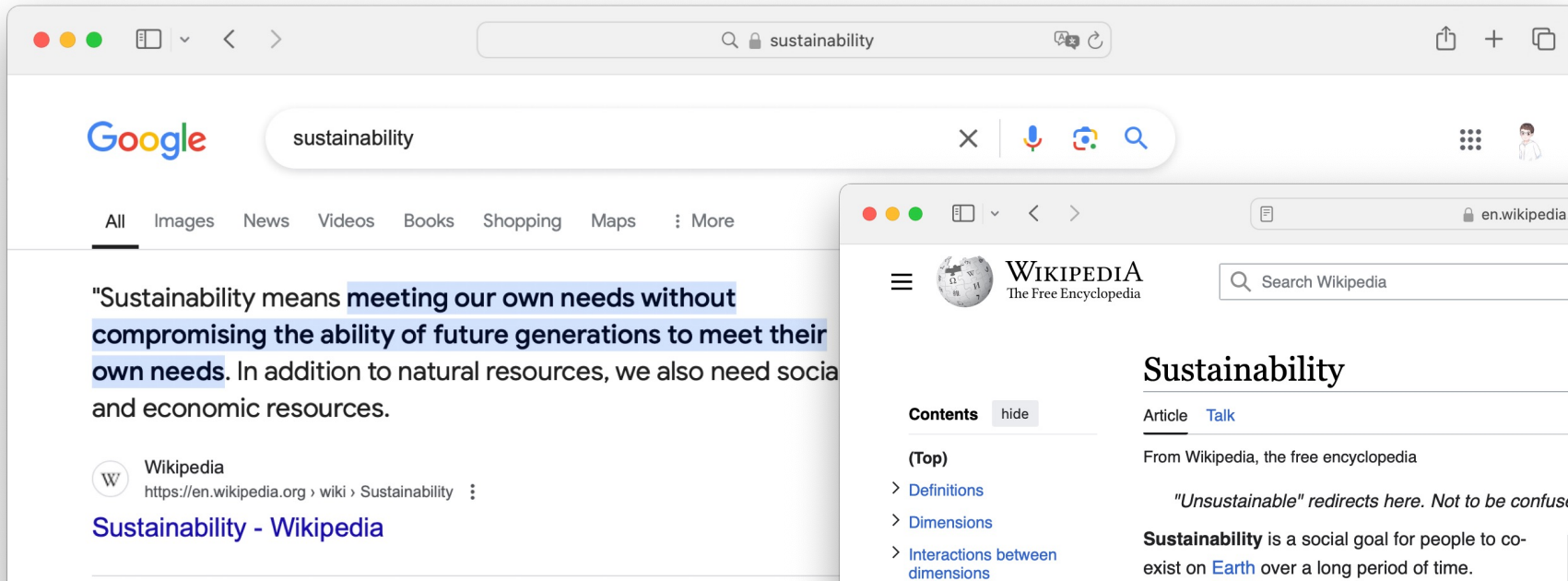


Find, organize, and present relevant and valuable content



Information Retrieval, Question Answering





Wikipedia is a good example of knowledge curation.

- Comprehensive
- Organized
- Reliable
- Verifiable

Generating Wikipedia-like Articles is Challenging

ChatGPT o1-preview ▾

Write a Wikipedia article on "The Future of Work".



Thought for 8 seconds ▾

The Future of Work

The **future of work** encompasses the evolving landscape of employment, influenced by technological advancements, economic shifts, and changing societal values. Key factors shaping this future include automation, artificial intelligence (AI), remote working, and the gig economy. Understanding these trends is crucial for individuals, businesses, and policymakers to navigate the transforming job market effectively.

Technological Advancements

Automation and Artificial Intelligence

Automation and AI are redefining various industries by streamlining processes and enhancing efficiency. Routine tasks in manufacturing, administration, and even service sectors are increasingly performed by machines and algorithms. While this shift raises concerns about job displacement, it also creates opportunities for new roles centered around technology management, data analysis, and AI ethics.

Digitalization and Remote Work

The rise of digital communication tools has made remote work more feasible and widespread. Events like the COVID-19 pandemic accelerated the adoption of work-from-home arrangements, demonstrating that many jobs can be effectively performed outside traditional office settings. This shift offers flexibility but also presents challenges in maintaining work-life balance, team cohesion, and cybersecurity.



Hard to verify



Lack of details

Generating Wikipedia-like Articles is Challenging

Given the ordered paragraphs $\{p_{R_i(j)}^i\}$.

Encode, concatenate, and truncate:

$$text_i = T(a_i) \parallel \{p_{R_i(j)}^i\}$$

$$tokenize(text_i) = x_i = (x_i^1, x_i^2, \dots, x_i^{n_i})$$

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

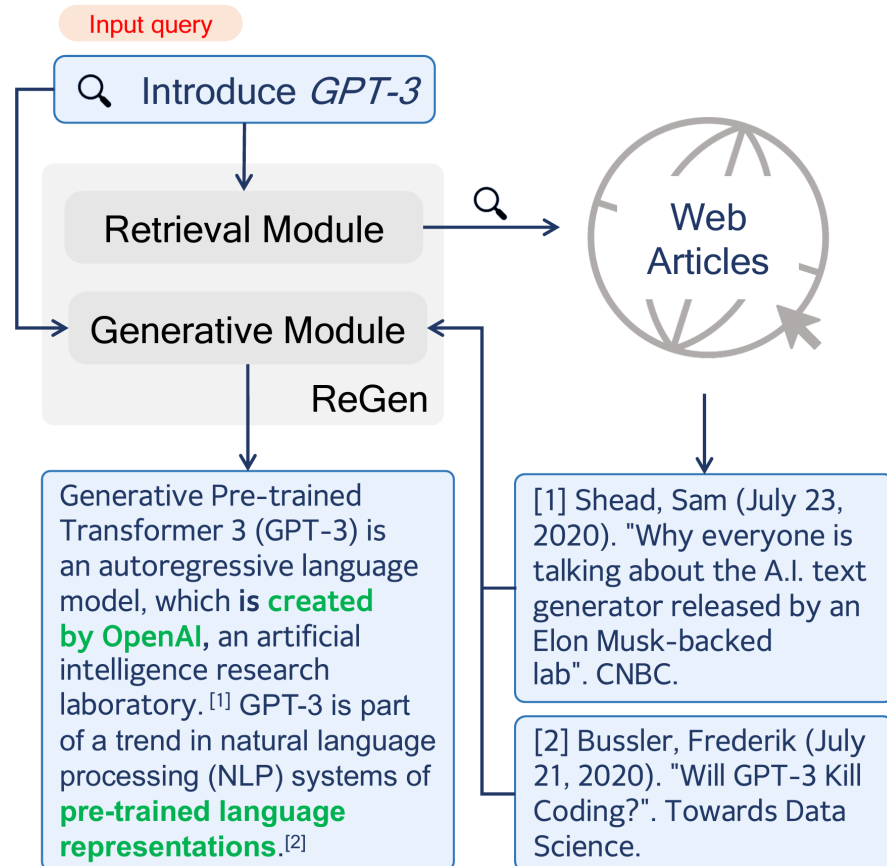
Train an abstractive model W that learns to write articles, $a_i = W(m_i^L)$.

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

Existing works usually assume the references are given.

Collecting references requires literature research which is non-trivial.

Generating Wikipedia-like Articles is Challenging



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

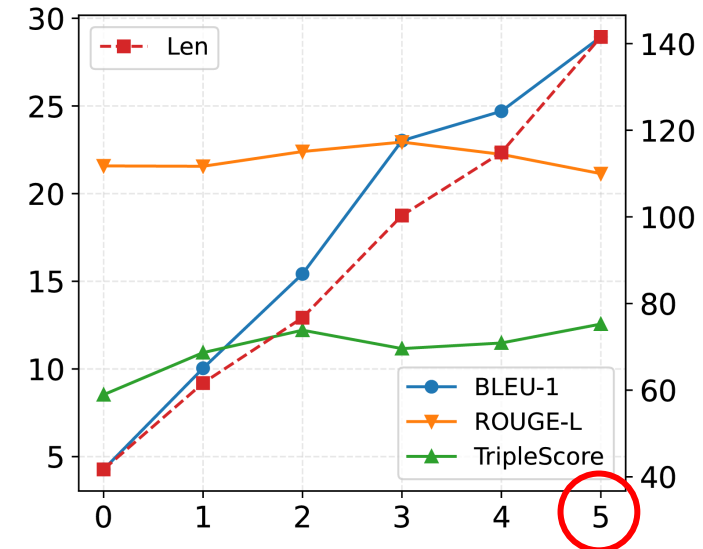


Figure 2: Performance of ReGen (Large) with different numbers of retrieved references.

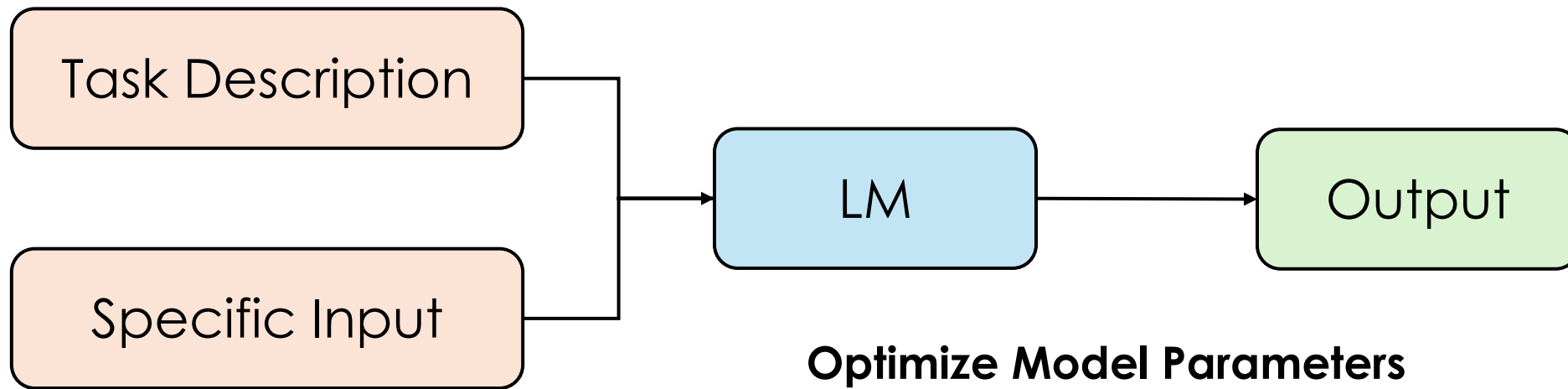
Existing works usually study generating a single paragraph.

Lecture Plan

1. Information Retrieval & RAG at a Glance
2. From Retrieving Information to Knowledge Curation
- 3. Building LM-Empowered Systems**
4. Evaluation? Evaluation!
5. Bring Human into the Loop

3. Building LM-Empowered Systems

There is a paradigm shift in how to solve a task.

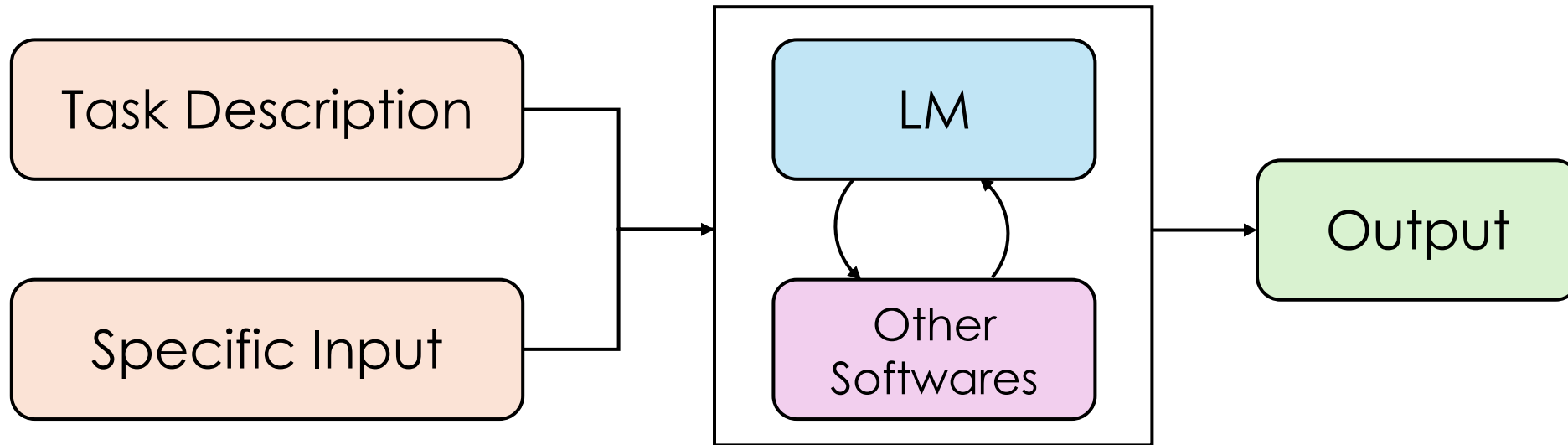


Optimize Input
(Prompt Engineering)

Optimize Model Parameters
(Continual Pre-training/Post-training, Fine-tuning)

3. Building LM-Empowered Systems

There is a paradigm shift in how to solve a task.



LM-Empowered System

The Future of Work



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



■ Somewhat Agree ■ Strongly Agree

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

Artificial intelligence (AI) in the workplace. AI technologies tasks, improving decision-making, and creating new an now read diagnostic scans with high accuracy, assisting treatments[12]. However, this trend also presents ning and upskilling programs[11]. Certain occupations, re highly vulnerable to automation, with over a 95% chance

in-office work, are becoming increasingly popular. aken to the people rather than bringing people to a remote or hybrid models[14]. This shift is not limited to a quarters of companies worldwide have already switched to

Healthcare

ards remote work, primarily driven by the rise of telehealth

ge has allowed healthcare providers to offer remote ty, contributing to better health outcomes and improved

ors, the mix of occupations and required skills will evolve, e, requiring new skills and educational backgrounds[12]. ealthcare, IT, management, and education, driven by nological deployment[16].

vironments presents both opportunities and challenges. allenges, particularly in areas like employee mental health luctive employment and decent work are crucial for a s and ensure that people can find work in a digitalized and

Table of contents

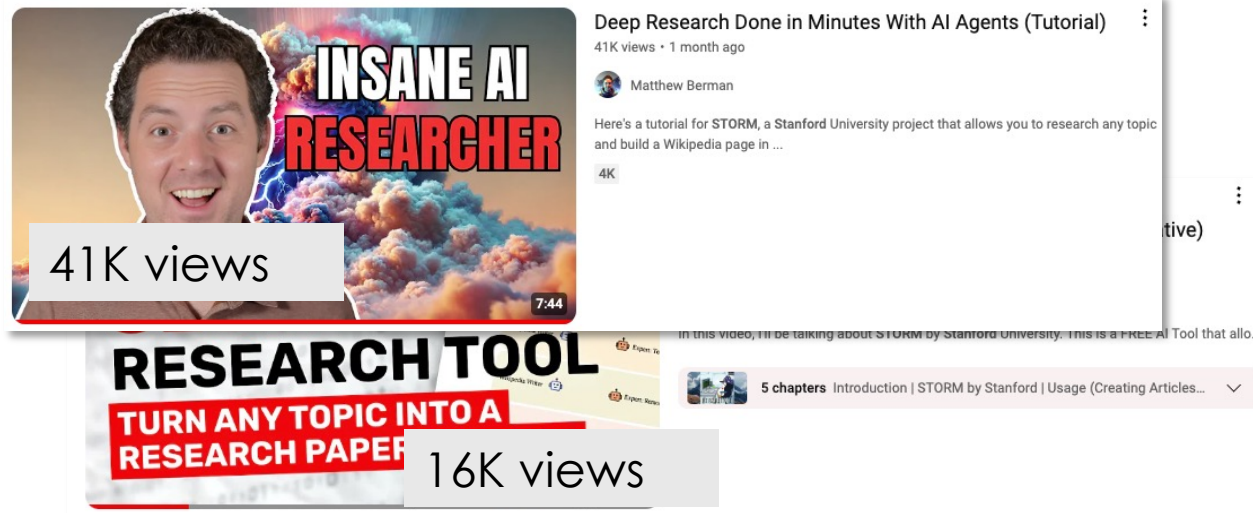
- Evolution of Work
 - Rise of Remote Work
 - Impact of Automation and Artificial Intelligence
 - Generational Differences and Adaptation Challenges
 - The Role of Technology in Shaping Workplaces
- Major Trends Shaping the Future of Work
 - Remote Work
 - Automation and Artificial Intelligence
 - Hybrid Work Models
 - Telehealth and Digitalization in Healthcare
 - Impact on Employment and Skills
 - Challenges and Opportunities
- The Role of Education and Training in the Future of Work
 - Skills Demand in a Digital Era
 - Lifelong Learning and Upskilling
 - Reducing Inequality Through Education
- Technological Advancements Supporting the Future of Work
 - Artificial Intelligence and Automation
 - Remote Work Technologies
 - Security and Data Protection
 - Gig Economy Platforms

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

STORM has
aroused interest
across various
communities.

<https://github.com/stanford-oval/storm>

Fork 963 Starred 10.2k



Since published in NAACL this year, our paper has already led to multiple follow-up works:

- Knowledge Navigator: LLM-guided Browsing Framework for Exploratory Search in Scientific Literature
- Eliciting Problem Specifications via Large Language Models
- ResearchArena: Benchmarking LLMs' Ability to Collect and Organize Information as Research Agents
- Filling Gaps in Wikipedia: Leveraging Data-to-Text Generation to Improve Encyclopedic Coverage of Underrepresented Groups
- Modeling and Enhancing Human Knowledge Navigation
-

How can STORM generate grounded articles with good breadth and 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?

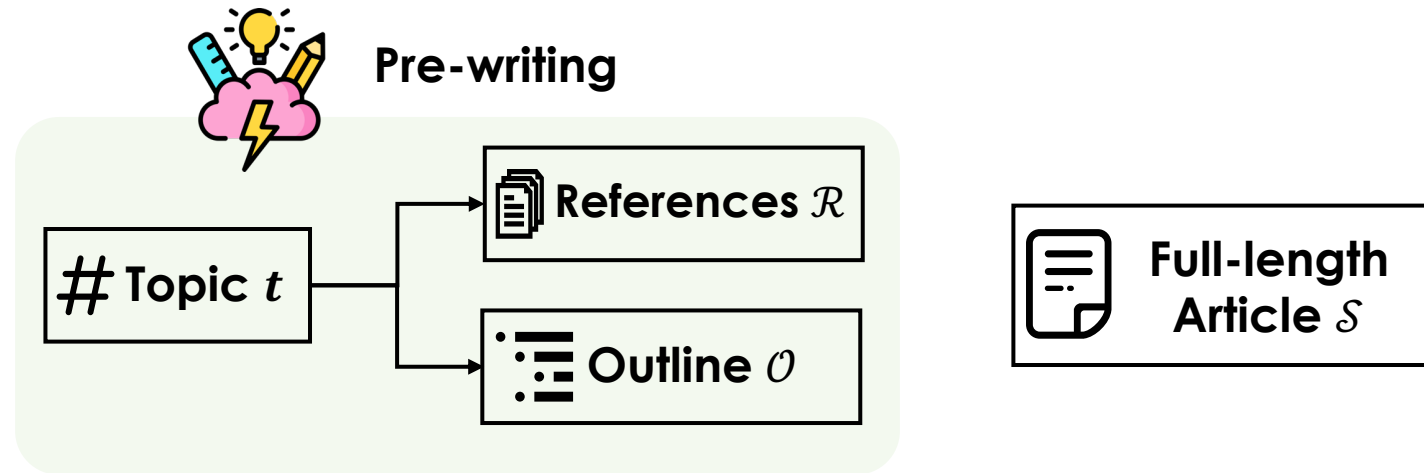
- Booth et al., 2003: The Craft of Research “Chapter II: **Asking Questions, Finding Answers**”

Topic t



 Full-length
Article \mathcal{S}

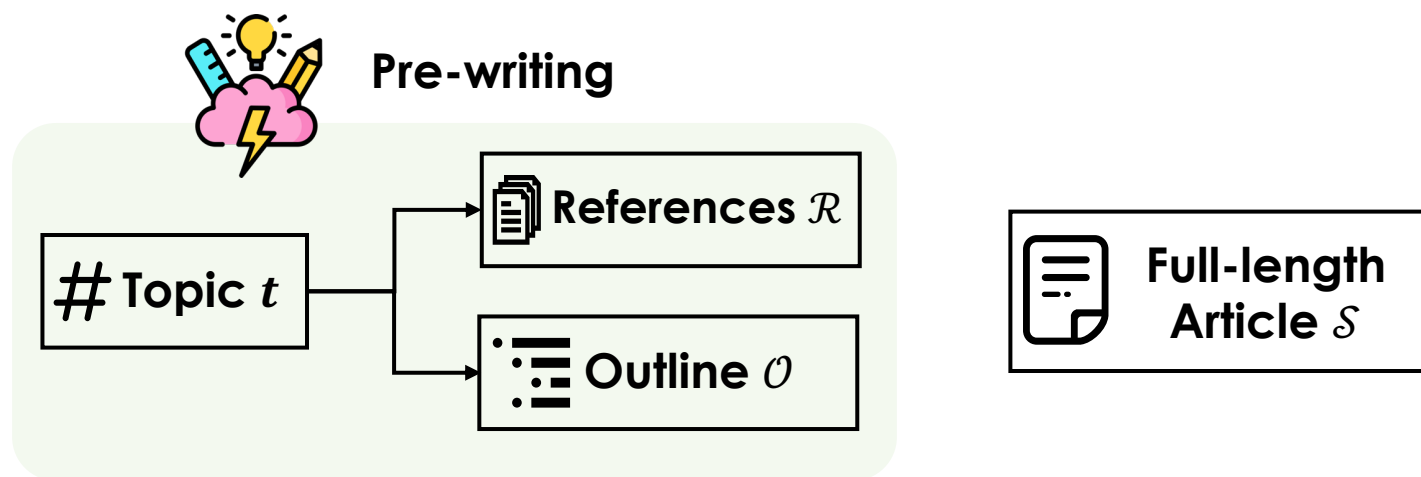
Introducing the Pre-writing Challenge



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} .

Introducing the Pre-writing Challenge



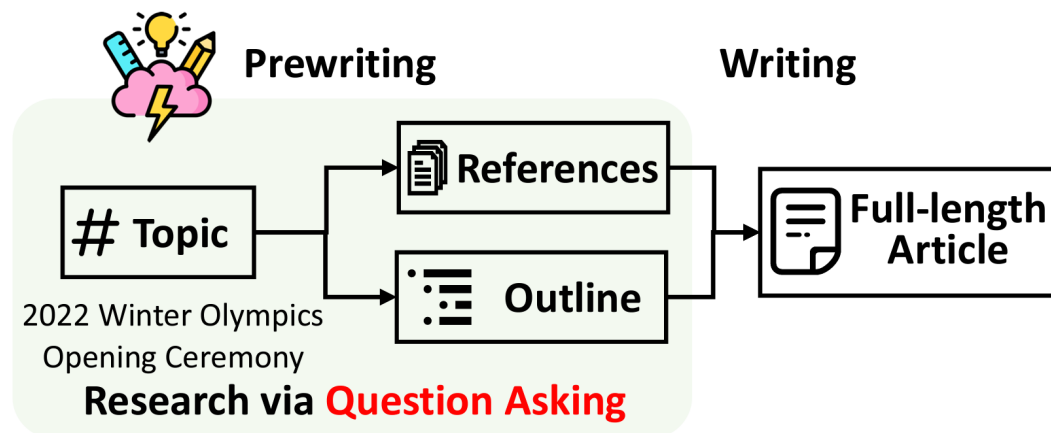
The pre-writing task:

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} .

Concept Induction

Multi-hop Search

Our Idea: Literature Research via **Question Asking**



Direct prompting results in questions that lack breadth and depth.

We can't simply rely on "brute force" or inference-time scaling.

(A) Direct Prompting

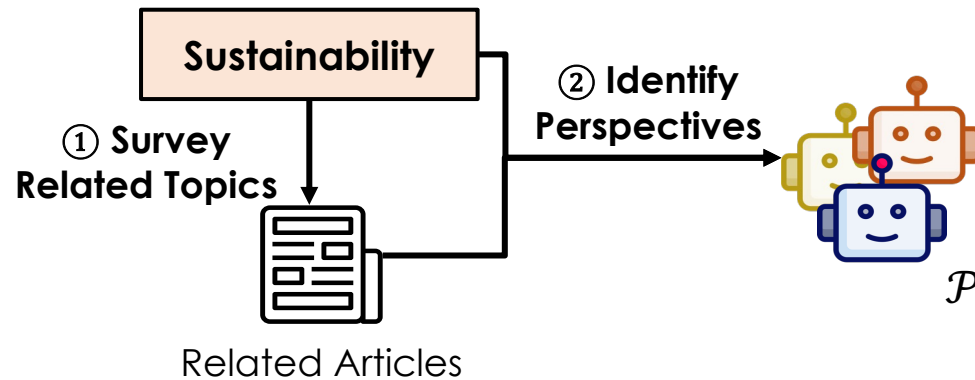


Prompt: Ask 30 questions about the given topic.

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

Perspective-Guided Question Asking

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



E.g., **Social scientist** – This editor will bring in a social perspective, focusing on topics such as social sustainability, ...

E.g., [wiki/Sustainable_development](https://en.wikipedia.org/wiki/Sustainable_development);
[wiki/Corporate_social_responsibility](https://en.wikipedia.org/wiki/Corporate_social_responsibility)

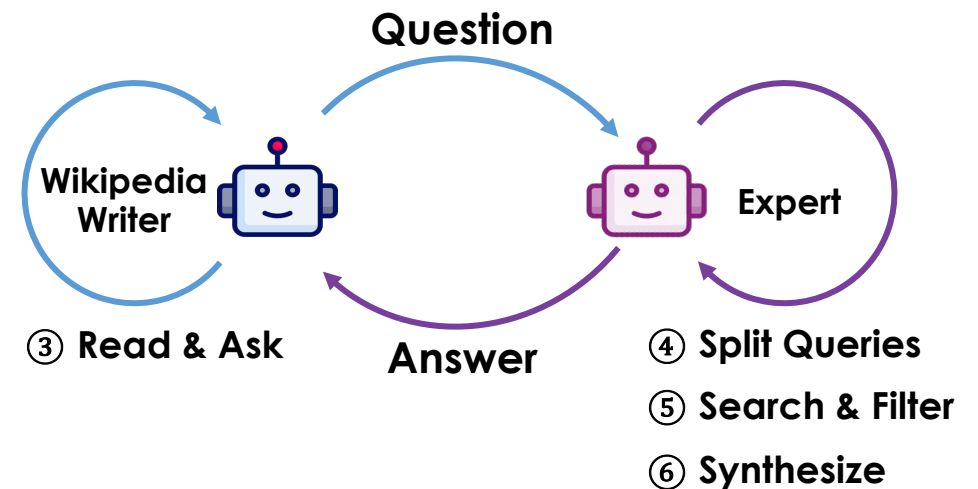
Simulating Conversations to Allow Follow-up Questions

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

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 participating in the opening ceremony. These included ... 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?**





Live demo!

<http://storm.genie.stanford.edu>



Lecture Plan

1. Information Retrieval & RAG at a Glance
2. From Retrieving Information to Knowledge Curation
3. Building LM-Empowered Systems
- 4. Evaluation? Evaluation!**
5. Bring Human into the Loop

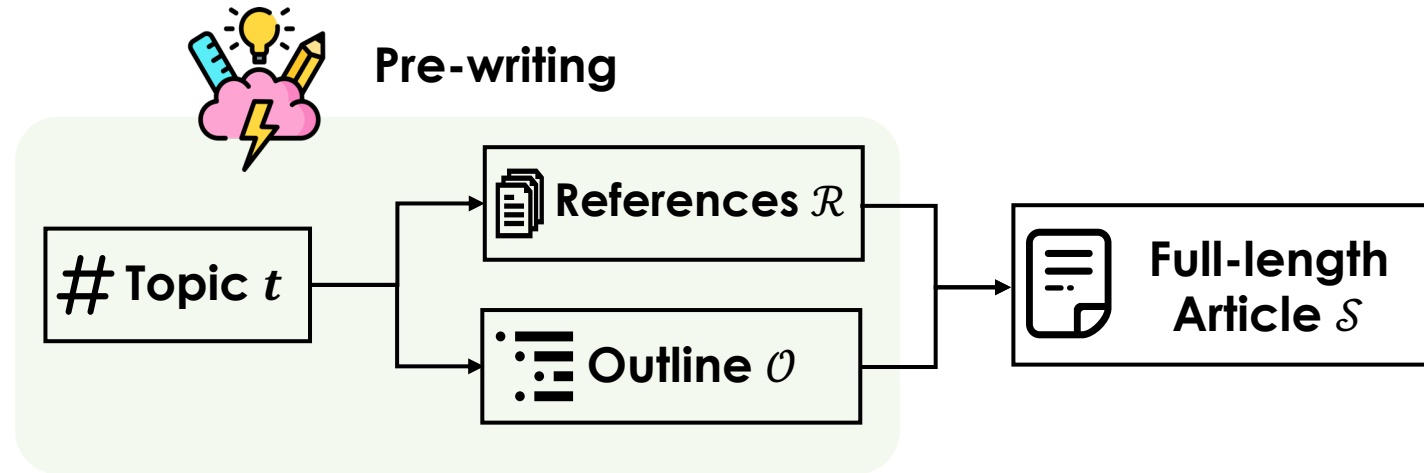
4. Conducting Meaningful Evaluation

Question: What to evaluate? And how?

Do we have ground truth / golden answer?

Besides final article, what else to evaluate?

Automatic Evaluation – Outline Quality



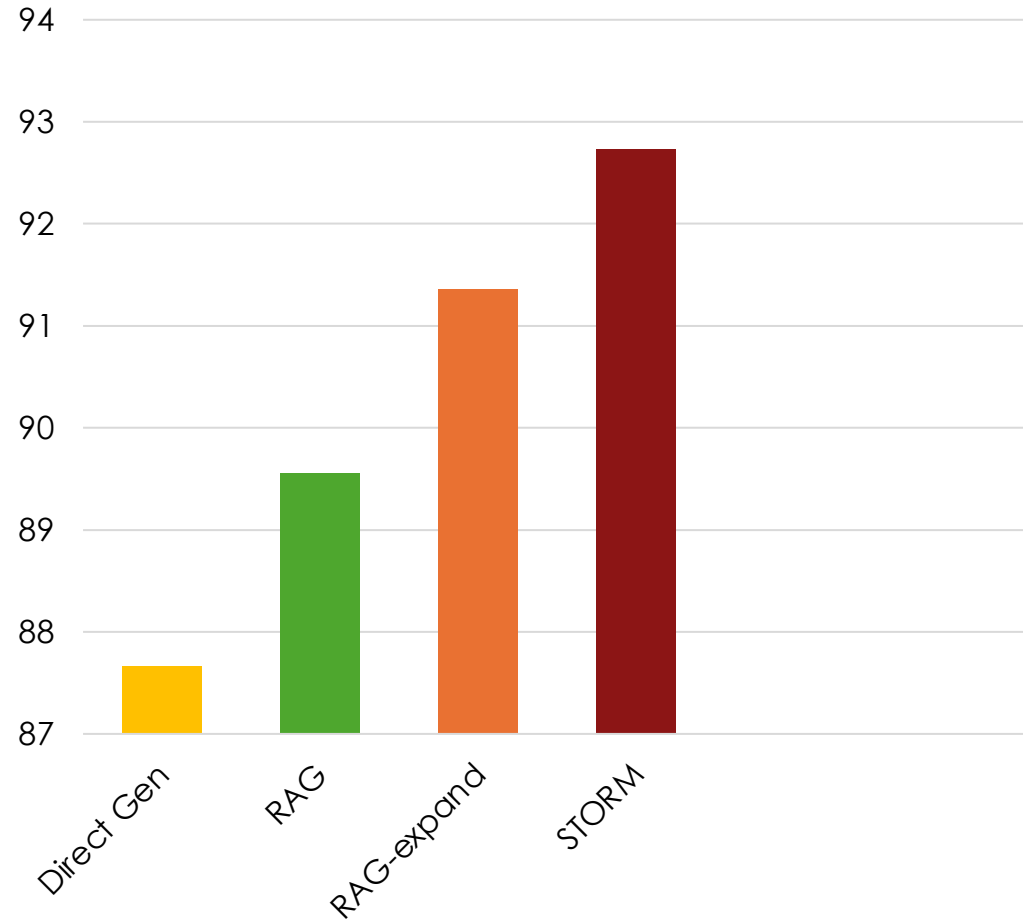
Intruduce **outline coverage metrics** as a proxy of the pre-writing stage quality for fast prototyping.

Given a human-written Wikipedia article on topic t ,

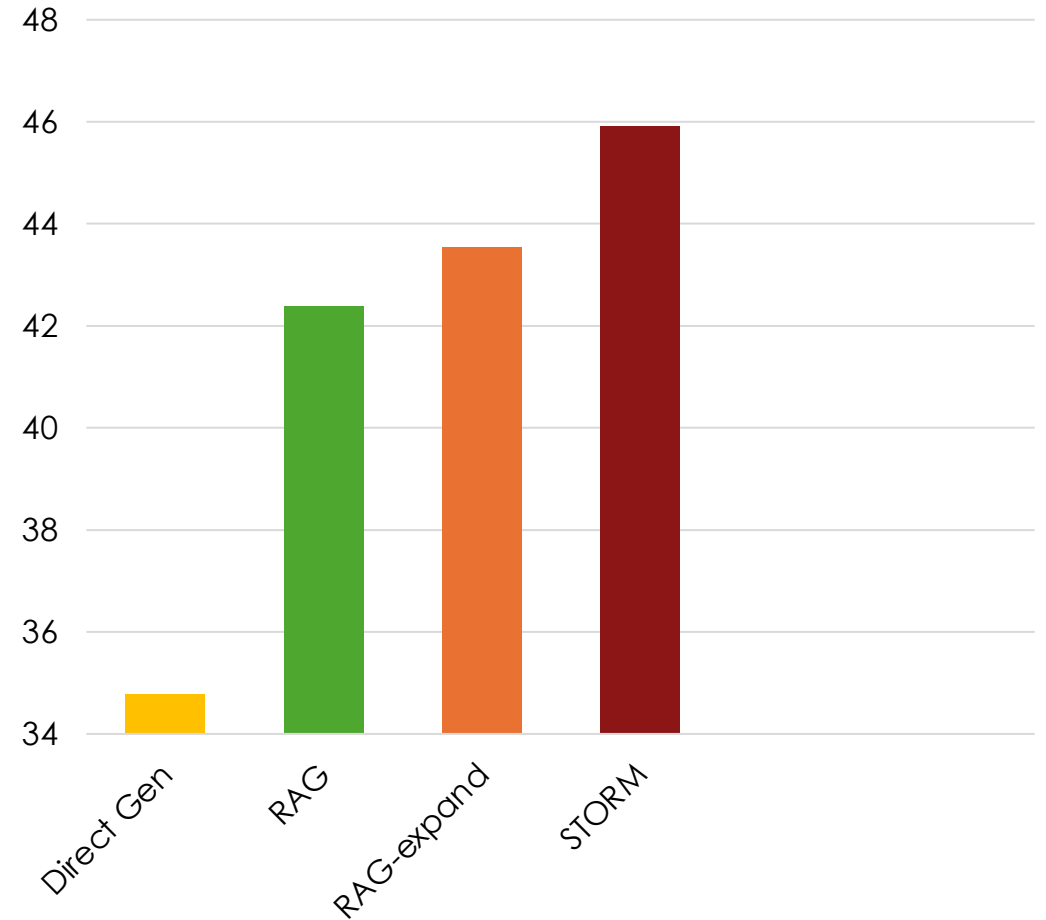
- Heading soft recall
Compare the sentence-BERT embeddings of headings in \mathcal{O} and the human-written article.
- Heading entity recall
The percentage of named entities in the human-written article headings covered by \mathcal{O} .

Outline Quality

Heading Soft Recall



Heading Entity Recall

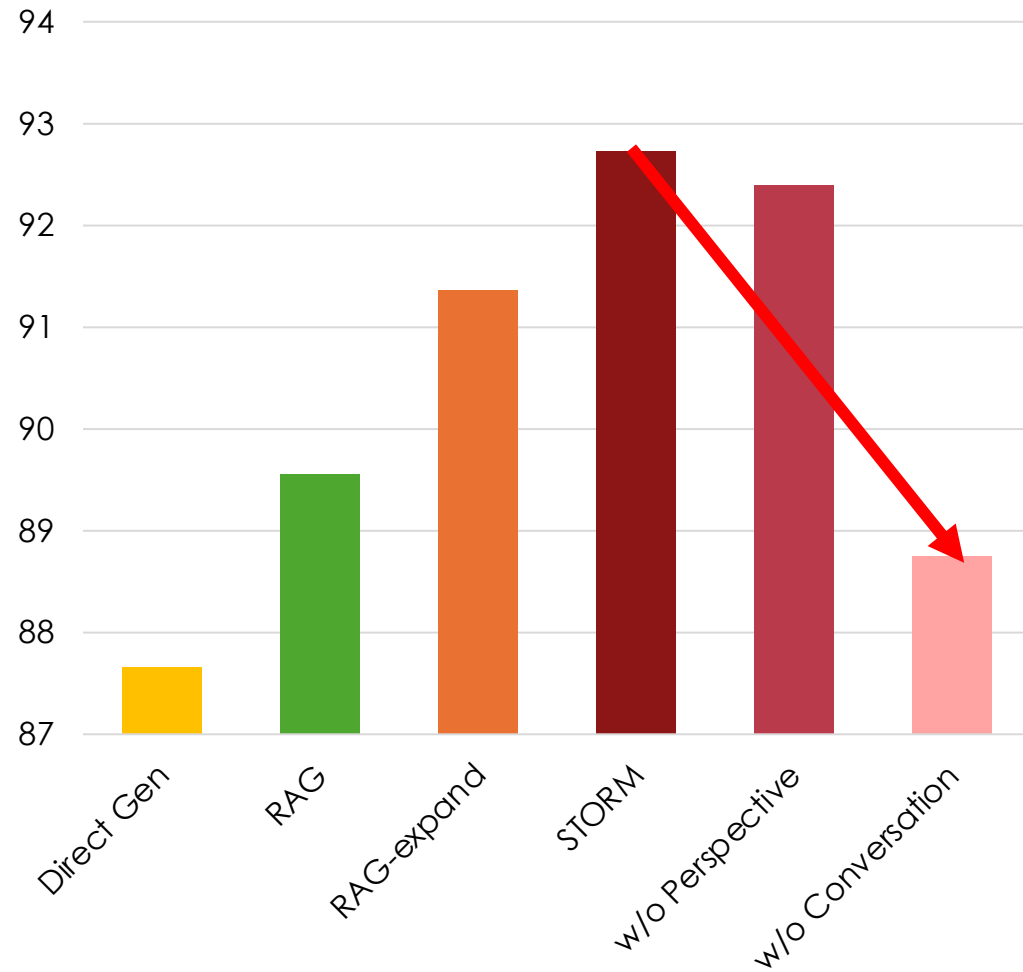


Ablation studies!

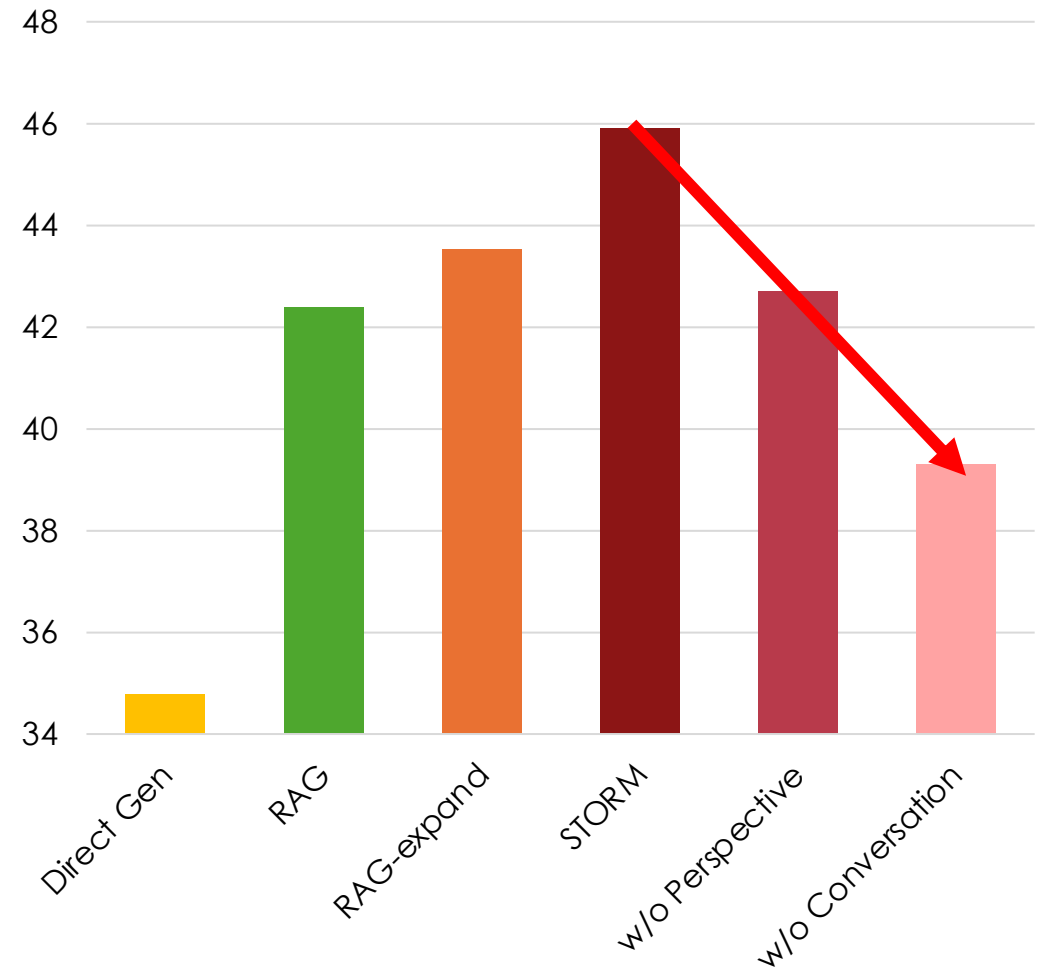
Takeaway: Ablation study help us to understand how different parts of a system contribute to its overall performance

Outline Quality

Heading Soft Recall



Heading Entity Recall



Automatic Evaluation – Article Quality

STORM outperforms baselines across **ALL** automatic metrics.

| | Comparsion with Human-written Articles | | | Rubric Grading | | | |
|-------------------|--|--------------|--------------------------|-------------------------|--------------|-------------------------|-------------------------|
| | ROUGE-1 | ROUGE-L | Entity Recall | Interest Level | Organization | Relevance | Coverage |
| Direct Gen | 25.62 | 12.63 | 5.08 | 2.87 | 4.60 | 3.10 | 4.16 |
| RAG | 28.52 | 13.18 | 7.57 | 3.14 | 4.22 | 3.05 | 4.08 |
| oRAG | 44.26 | 16.51 | 12.57 | 3.90 | 4.79 | 4.09 | 4.70 |
| STORM | 45.82 | 16.70 | 14.10[†] | 3.99[†] | 4.82 | 4.45[†] | 4.88[†] |
| w/o Outline Stage | 26.77 | 12.77 | 7.39 | 3.33 | 4.87 | 3.35 | 4.37 |

Table 2: Results of automatic article quality evaluation. [†] denotes significant differences ($p < 0.05$) from a paired t -test between STORM and the best baseline, *i.e.*, oRAG. The rubric grading uses a 1-5 scale.

Human Evaluation - Wikipedia Editor Evaluation

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

Experienced Wikipedia editors favor articles produced by STORM.

| ≥ 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% |

In the wild Evaluation

75,000+ Users

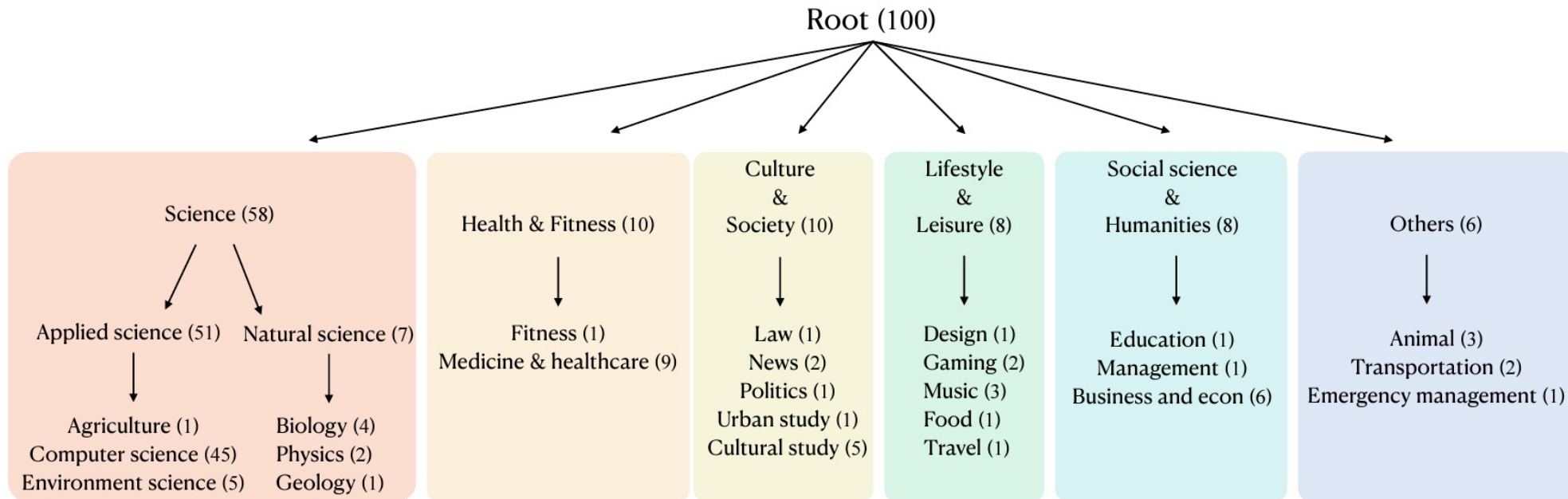
114,000+ Articles

211,000+ Browsing

20,000+ Feedbacks

In-the-wild Evaluation

People have used STORM across a diverse array of topics.



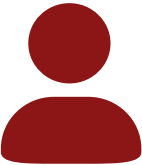
Revisit Our Meta Question:

Are **people**'s information needs satisfied?





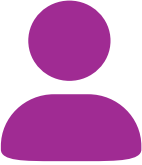
My thoughts evolve, so I want to update my queries.



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



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

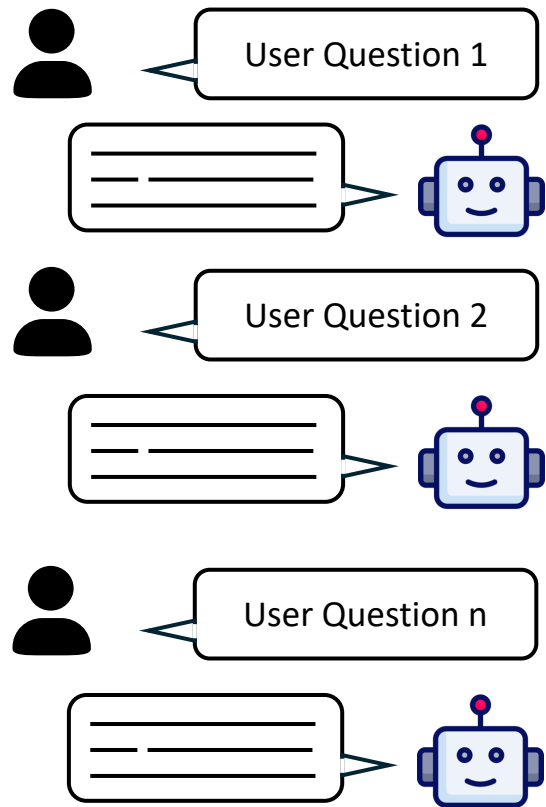


Lecture Plan

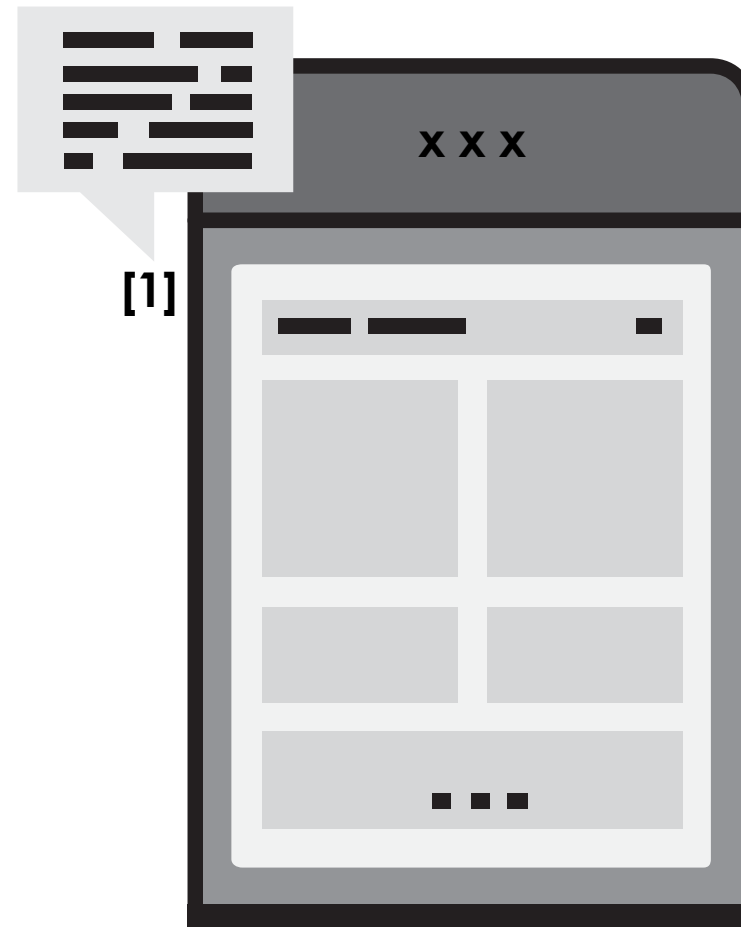
1. Information Retrieval & RAG at a Glance
2. From Retrieving Information to Knowledge Curation
3. Building LM-Empowered Systems
4. Evaluation? Evaluation!
- 5. Bring Human into the Loop**



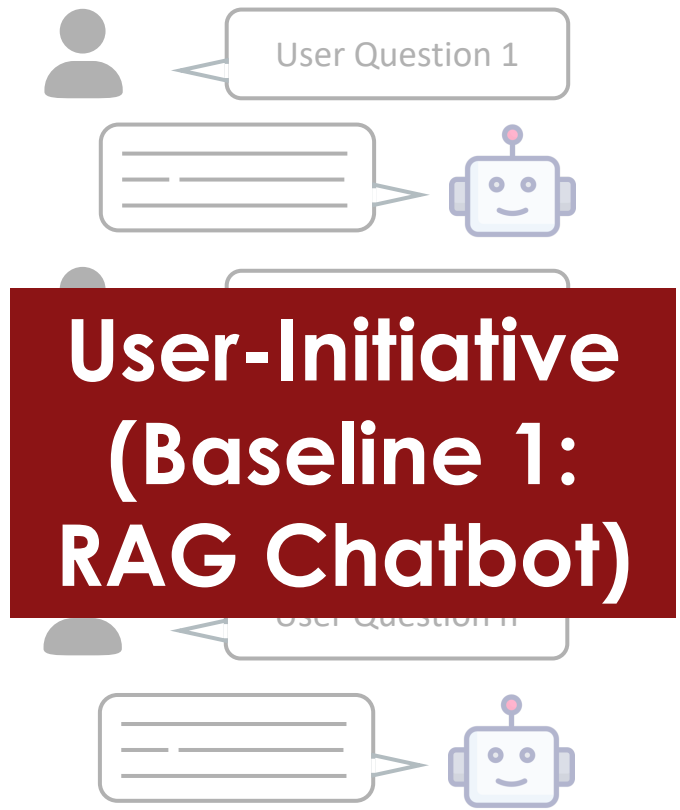
5. How to bring a human user into the loop?



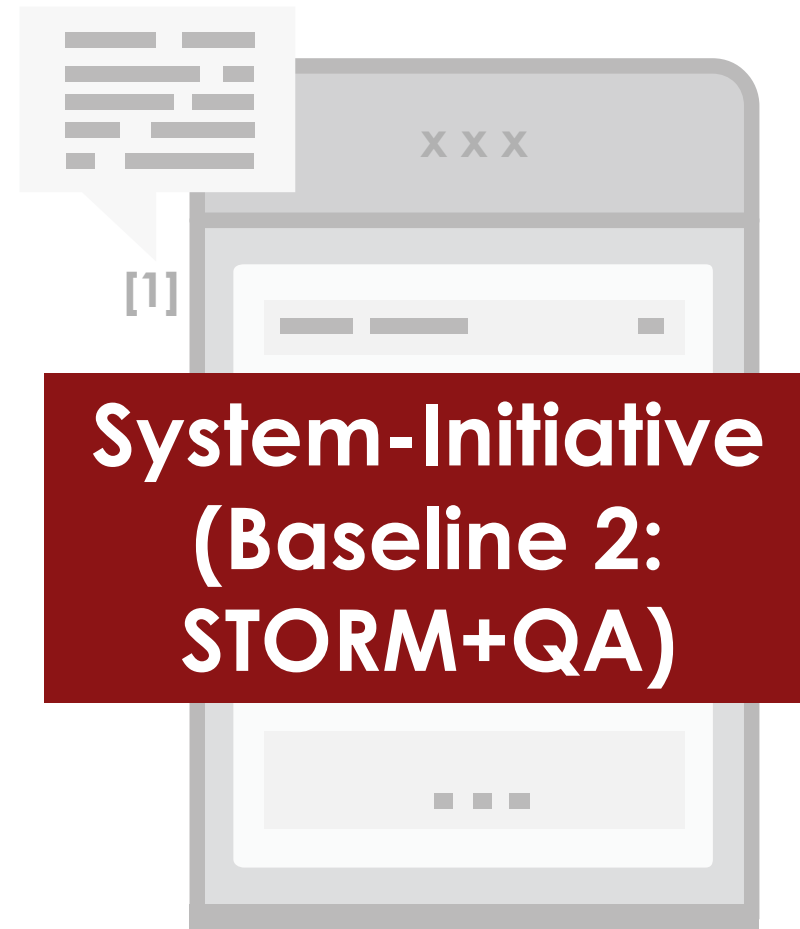
Convert STORM into a hallucination-free question answering system.



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



Convert STORM into a hallucination-free question answering system.



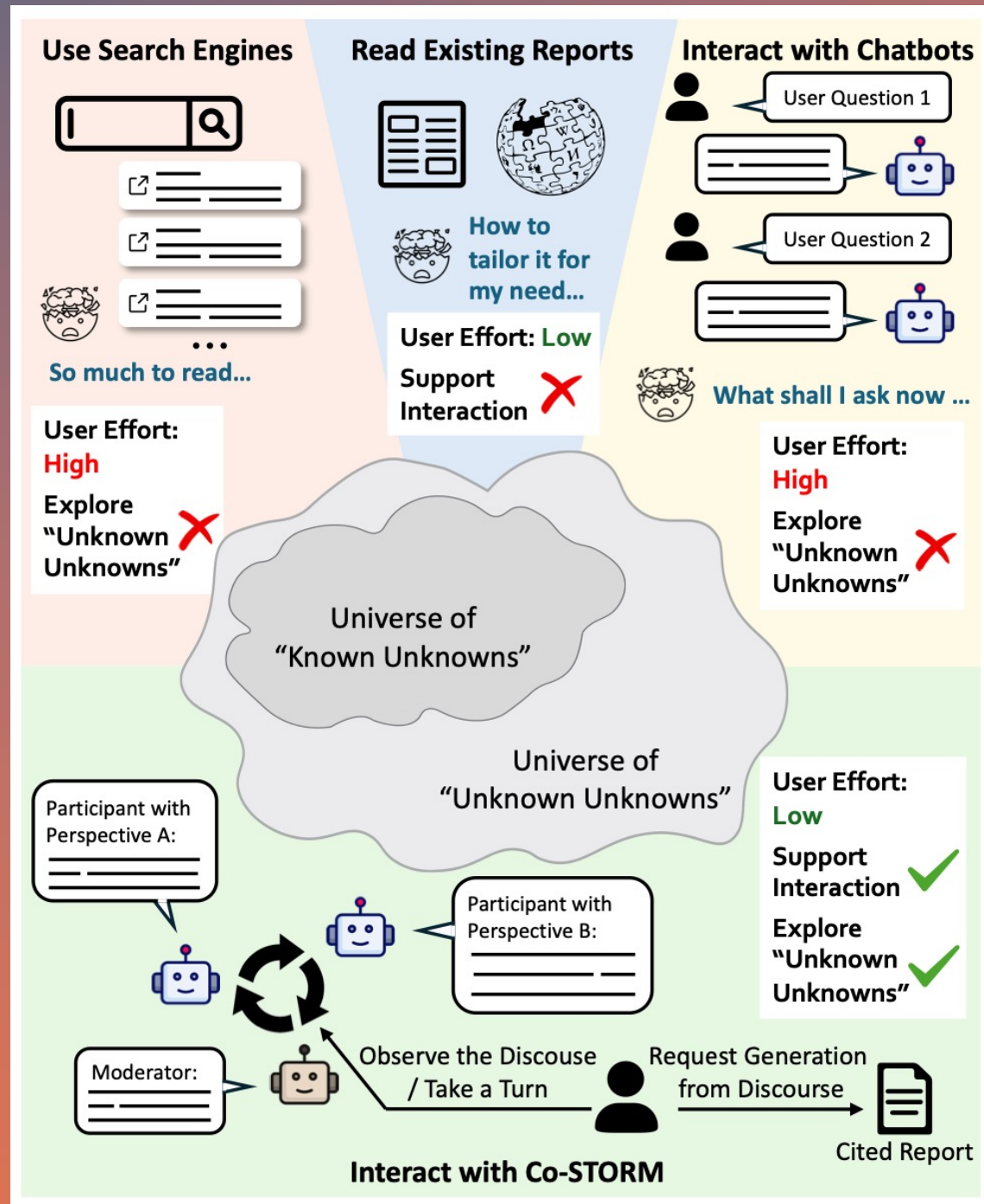
After the long report is generated, allow the user to edit or ask questions.

Mix-Initiative Systems

- Considering **uncertainty about a user's goals**.
- Considering the status of a **user's attention** in the timing of services.
- Allowing efficient **direct invocation and termination**.
- Providing **mechanisms for efficient agent-user collaboration** to refine results.
- Maintaining **working memory** of recent interactions.

Collaborative -STORM

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



How to surface **unknown unknowns**?

Key Idea: Let Users Observe & Involve in Expert Discussion

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.

Collaborative Discourse Protocol

Goal: Co-STORM allows users to learn by observing and participating occasionally in the discourse, emulating a common educational scenario.

- Simulates agent grounded in the search engine, **answering and asking questions**.
- 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.

Challenge: How to invoke thought-provoking questions?

Question: What types of utterance intents an LM agent could have?

- **Question Asking**
 - Posing an original question
 - making an information request
- **Question Answering**
 - Providing a potential answer
 - elaborating with further details

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

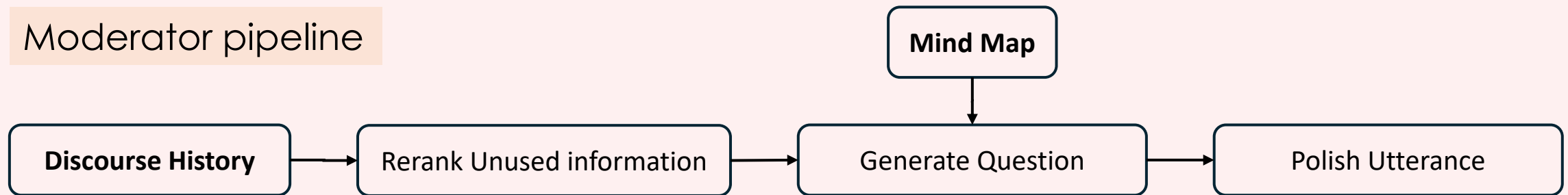
Recall: How do we ask follow up question during search?

Sometimes, we may find results **not directly related to the search query but relevant to the topic we are interested** 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”.

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

Moderator pipeline



Conducting Meaningful Evaluation

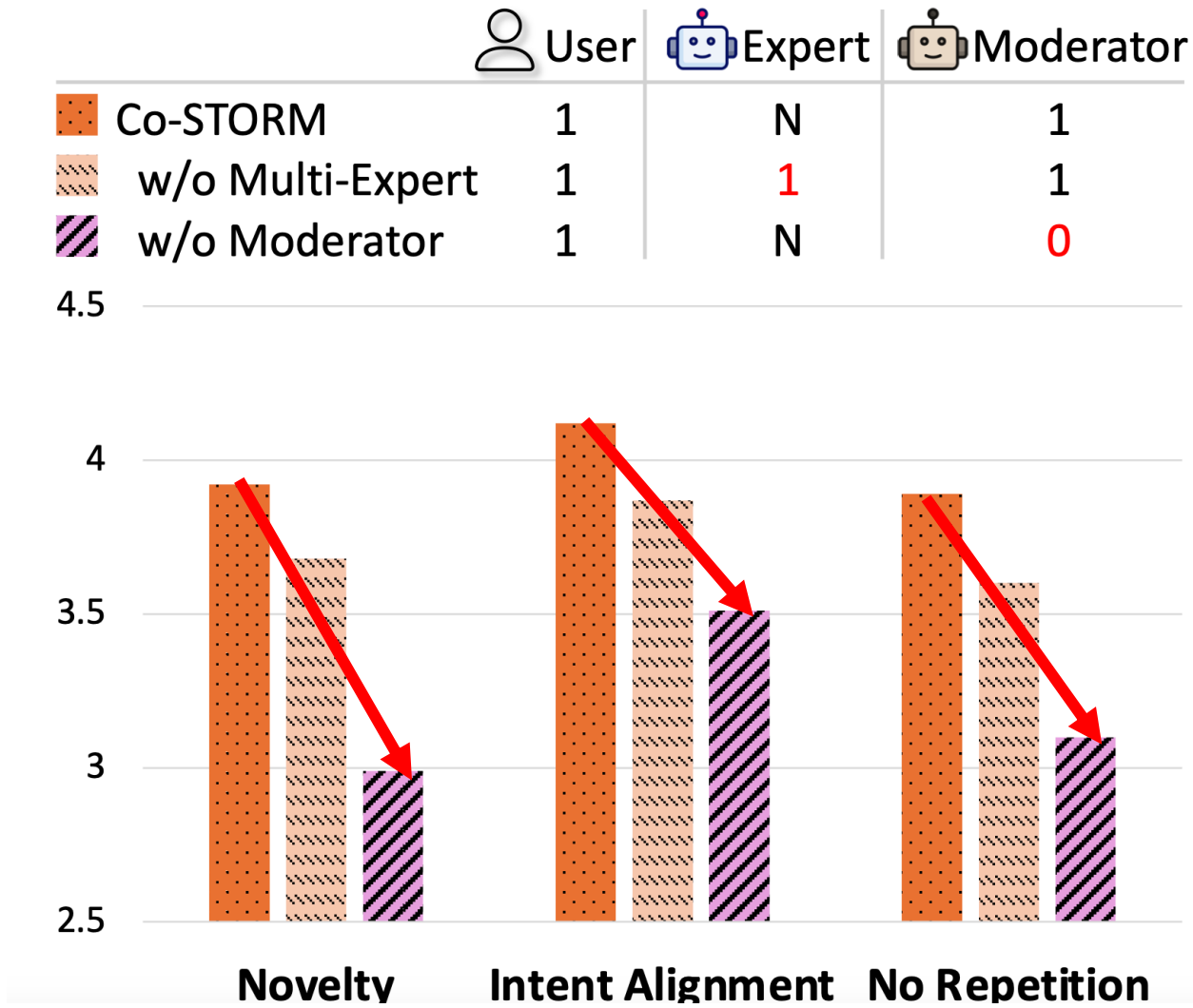
Question: What to evaluate? And how?

Do we have ground truth / golden answer?

Besides final article, what else to evaluate?

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

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



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 |

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 Evaluation

| | Co-STORM v.s. Search Engine | | | | Co-STORM v.s. RAG Chatbot | | | |
|-------------|-----------------------------|-------------|----------------|-----------------|---------------------------|-------------|----------------|-----------------|
| | Search Engine | Co-STORM | Win % (Lose %) | <i>p</i> -value | RAG Chatbot | Co-STORM | Win % (Lose %) | <i>p</i> -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.

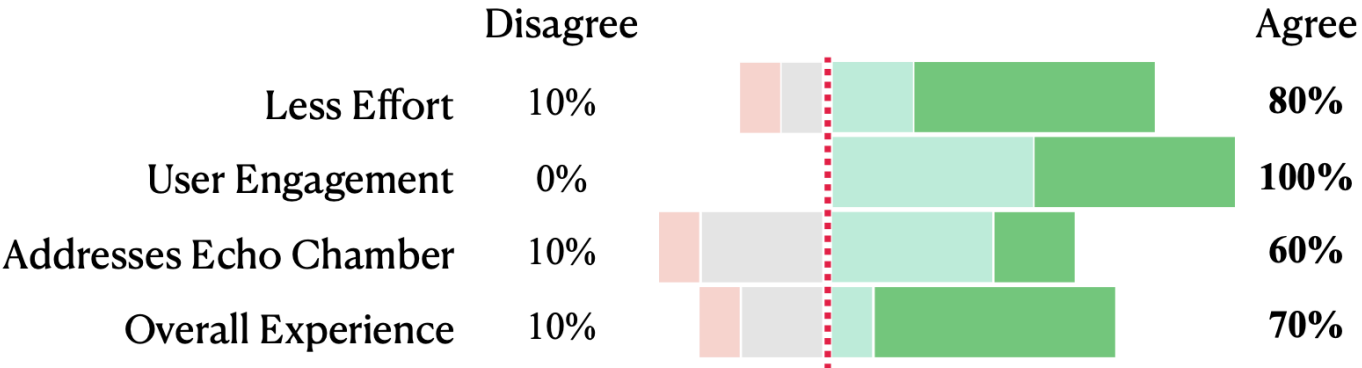
Human Evaluation

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

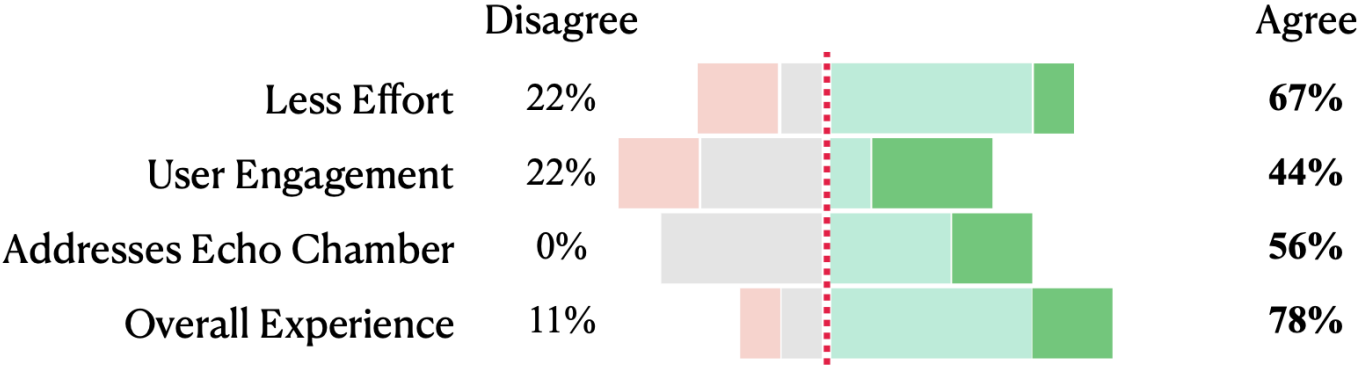
“Co-STORM is **so much less mentally taxing for me to use**”



Co-STORM is better than Search Engine



Co-STORM is better than RAG Chatbot



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.
 - STORM is then deployed in the wild, tested by over 50,000 users.
 - Co-STORM invites 20 users in the wild.
 - Co-STORM will be deployed in the wild as well!

Questions?

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

Assignment (Due Oct 2nd)

- **Use: Play with live research preview!**

You will have chance to play STORM and Co-STORM as a user

- **Know: Technical deep dive**

Dive deeper into core design choices and LLM system design

- **Inspire: Ideas to build better knowledge curation system**

Analyze strength and weakness of current work; think about what can be improve; what more use cases can be supported