
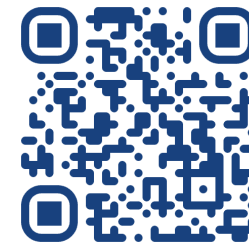




Stanford
University



Watch the talk on
Youtube 



Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models

NAACL 2024



Yijia Shao



Yucheng Jiang



Theodore Kanell



Peter Xu



Omar Khattab



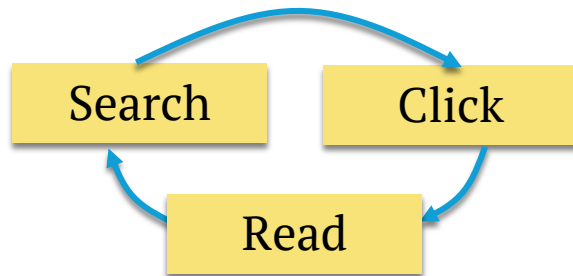
Monica Lam

shaoyj@stanford.edu

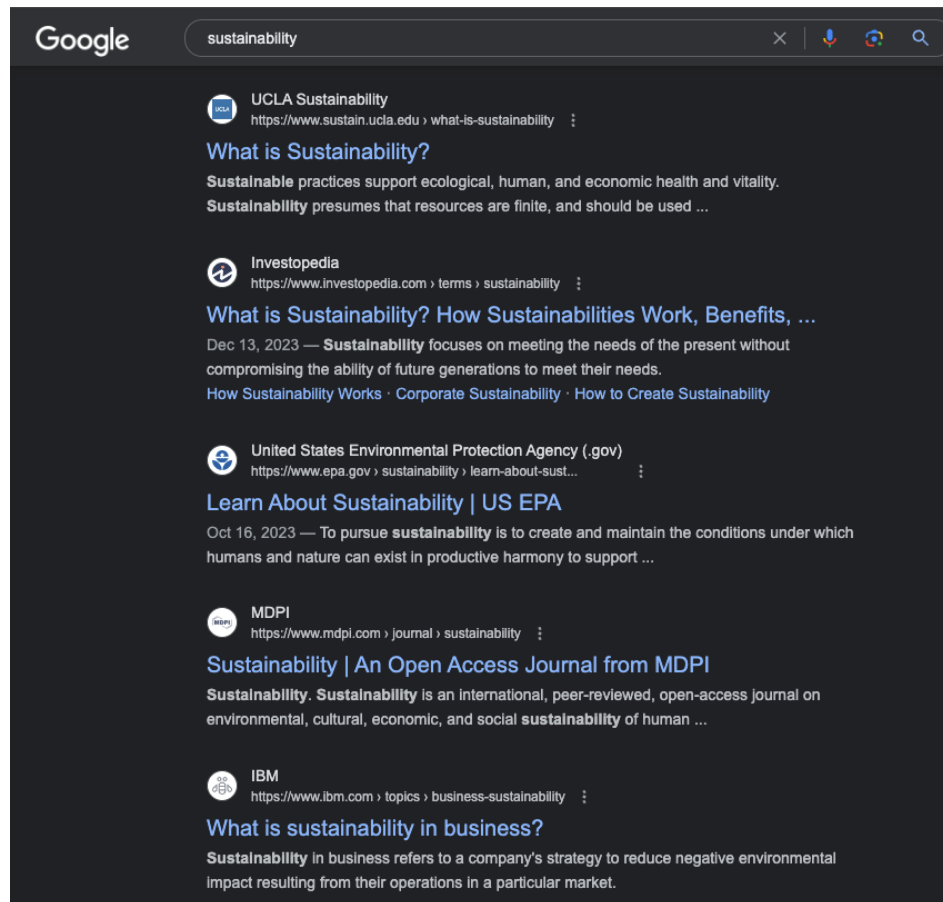
There Is More Raw Information Than One Can Consume



I want to learn about “Sustainability” in-depth.

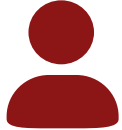


Inefficient!



Search Engines /
IR Systems

There Is More Raw Information Than One Can Consume

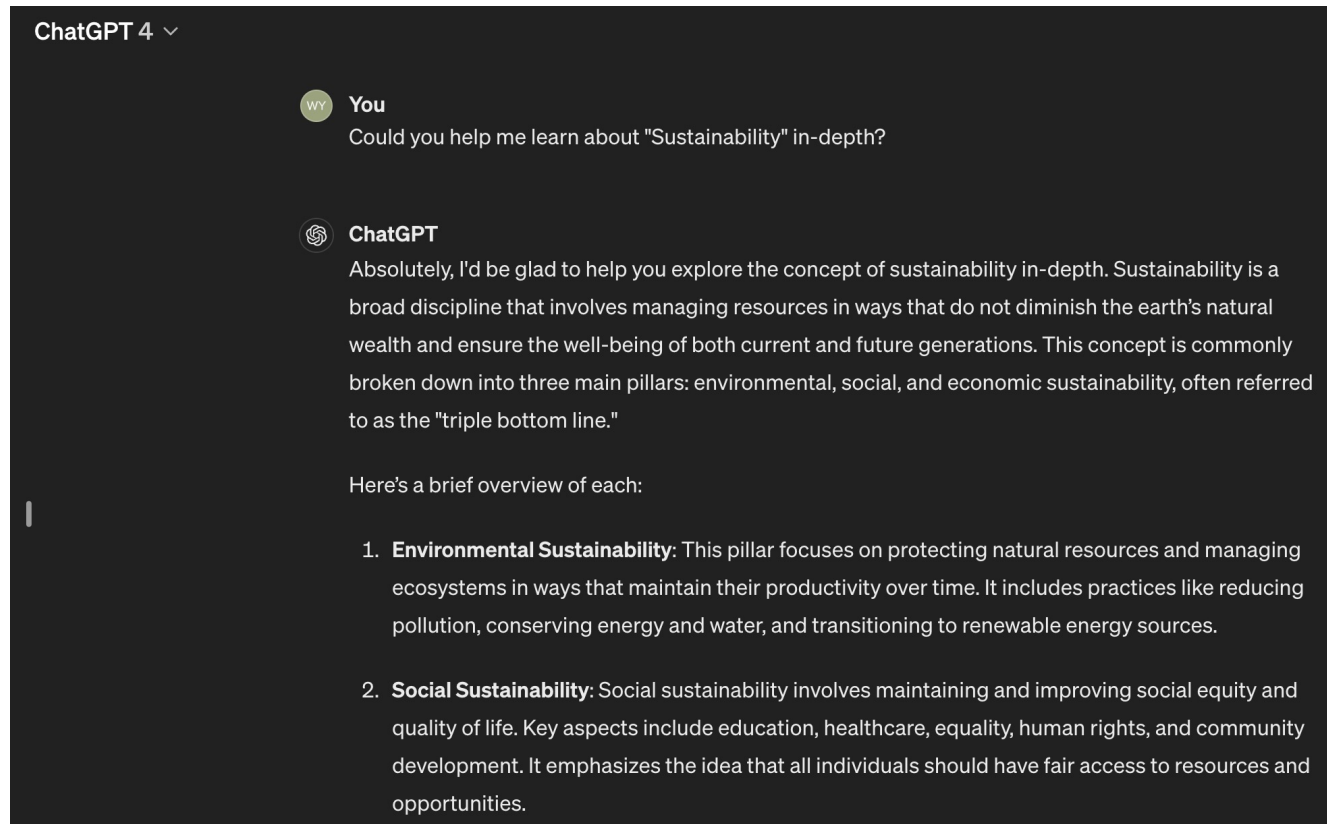


I want to learn about “Sustainability” in-depth.



Chatting with Large
Language Models (LLMs)

**Hard to
verify!**



There Is More Raw Information Than One Can Consume



I want to learn about “Sustainability” in-depth.

**Limited depth
and breadth!**

Help me learn about "Sustainability".

📖 Sources

Learn About Sustainability | US EPA
epa • 1

45 Sustainability Resources You Need to...
purdueglobal • 2

12 Ways to Live More Sustainably
biologicaldiversity • 3

View 2 more

☰ Answer

Sustainability is based on the principle that everything we need for our survival and well-being depends, either directly or indirectly, on our natural environment. It aims to create and maintain the conditions under which humans and nature can thrive, ensuring fulfillment of social, economic, and other requirements for both present and future generations ¹. Some key aspects of sustainability include:

- **Population:** Understanding how human populations have evolved over time and their impact on the environment ⁵.
- **Ecosystems:** Recognizing the importance of maintaining ecosystems and their services for human well-being ⁵.
- **Global Change:** Examining the effects of human activities on global systems, such as climate change and biodiversity loss ⁵.
- **Energy:** Exploring sustainable energy sources and their role in reducing greenhouse gas emissions ⁵.
- **Agriculture:** Investigating sustainable agricultural practices that can provide food security while minimizing environmental impacts ⁵.
- **Water:** Addressing water scarcity and the need for efficient water management ³.



Commercial Generative
Search Engines

**It's hard to ask
questions about
something we
don't know.**

There Is More Raw Information Than One Can Consume



I want to learn about “Sustainability” in-depth.

The screenshot shows a Google search interface with the query 'sustainability'. The top result is from Wikipedia, titled 'Sustainability', with the URL 'https://en.wikipedia.org/wiki/Sustainability'. Below the title is a yellow smiley face icon. The snippet reads: '"Sustainability" means meeting our own needs without compromising the ability of future generations to...'. To the right of the snippet is a 'Contents' sidebar with links to 'History', 'Definitions', 'Dimensions', 'Interactions between dimensions', 'Measurement tools', 'Barriers', 'Transitions', 'Options for overcoming barriers', 'Responses by nongovernmental stakeholders', 'Critiques', 'See also', and 'References'. The main article content is titled 'Sustainability' and includes a note that 'Unsustainable' redirects here. The text discusses the social goal of sustainability, its three dimensions (environmental, economic, and social), and various visual representations of the concept.



Discover a well-organized article/report!

There Is More Raw Information Than One Can Consume



I want to learn about “Sustainability of Large Language Models” in-depth.



Search results

Q Sustainability of Large Language Models

Advanced search:

Search in:

**There are thousands of topics on
en.wikipedia.org/wiki/Wikipedia:Requested_articles!**

The page “*Sustainability of Large Language Models*” does not exist. You can [create a draft and submit it for review](#) or [request that a redirect be created](#), but consider checking the search results below to see whether the topic is already covered.

View (previous 20 | [next 20](#)) (20 | [50](#) | [100](#) | [250](#) | [500](#))

This Work: Study How To Generate Wikipedia-like Articles From Scratch

Generating grounded, long-form articles in open domains is challenging.

	Domain
Sauper and Barzilay (2009)	Two
Fan and Gardent (2022)	One

Different domains emphasize different aspects.

This Work: Study How To Generate Wikipedia-like Articles From Scratch

Generating grounded, long-form articles in open domains is challenging.

	Domain	Given Refs?
Sauper and Barzilay (2009)	Two	No
Fan and Gardent (2022)	One	No
Liu et al. (2018)	All	Yes
Balepur et al. (2023)	One	Yes

Different domains emphasize different aspects.

Collecting references requires literature research.

This Work: Study How To Generate Wikipedia-like Articles From Scratch

Generating grounded, long-form articles in open domains is challenging.

	Domain	Given Refs?	Scope
Sauper and Barzilay (2009)	Two	No	Full article
Fan and Gardent (2022)	One	No	Full article
Liu et al. (2018)	All	Yes	One para.
Balepur et al. (2023)	One	Yes	One para.
Qian et al. (2023)	All	No	One para.

Different domains emphasize different aspects.

Collecting references requires literature research.

Generating long grounded articles is hard.

This Work: Study How To Generate Wikipedia-like Articles From Scratch

We tackle these challenges with **STORM**, an LLM-powered knowledge curation system.

	Domain	Given Refs?	Scope
Sauper and Barzilay (2009)	Two	No	Full article
Fan and Gardent (2022)	One	No	Full article
Liu et al. (2018)	All	Yes	One para.
Balepur et al. (2023)	One	Yes	One para.
Qian et al. (2023)	All	No	One para.
Our work	All	No	Full article

Different domains emphasize different aspects.

Collecting references requires literature research.

Generating long grounded articles is hard.

Learning From The Human Writing Process

How do humans write?

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

“It is crucial to the success of any writing that occurs later, and it is seldom given the attention it consequently deserves.”

How do humans discover new information?

- Booth et al., 2003: The Craft of Research

“Chapter II: Asking Questions, Finding Answers”

Our System

The Future of Work



STORM:
Synthesis of Topic **Outlines**
through **Retrieval** and
Multi-perspective **Question Asking**

The screenshot shows a web browser window with a document titled "The_Future_of_Work_storm_ge...". The document is on page 2 of 7. The main heading is "Major Trends Shaping the Future of Work". Below this, there is a section titled "Remote Work" and another titled "Automation and Artificial Intelligence". A table of contents is visible on the left side of the page, listing various topics related to the future of work.

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

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 training and upskilling programs[11]. Certain occupations, re highly vulnerable to automation, with over a 95% chance

in-office work, are becoming increasingly popular. taken 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

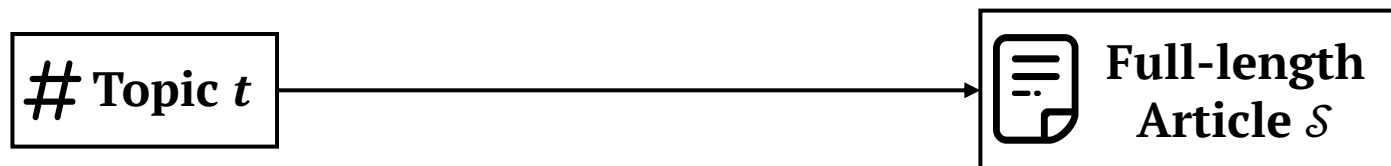
ards remote work, primarily driven by the rise of telehealth

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

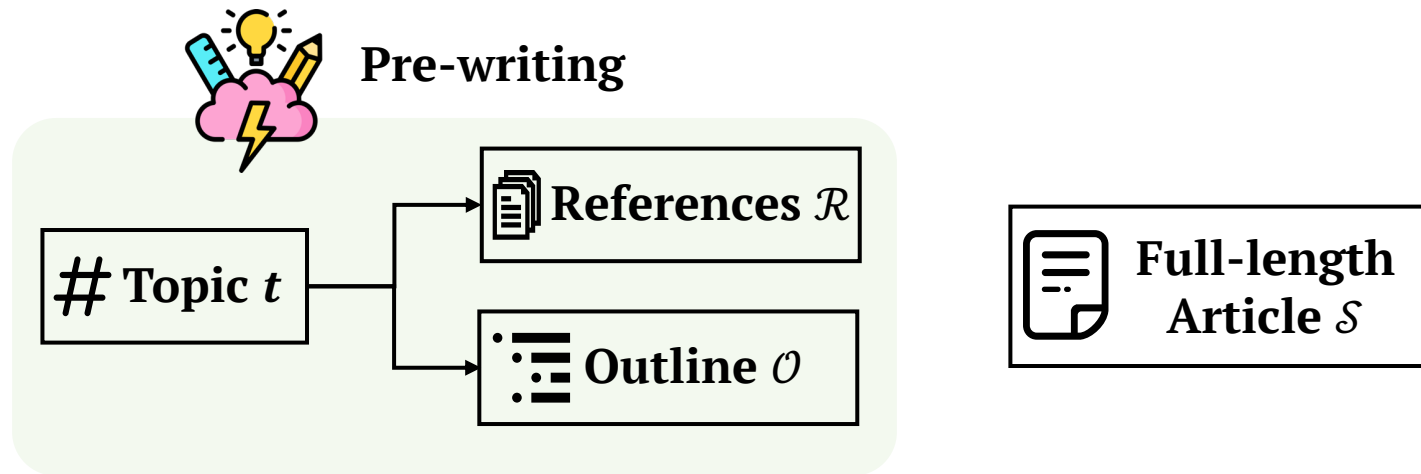
ors, the mix of occupations and required skills will evolve, e, requiring new skills and educational backgrounds[12]. healthcare, IT, management, and education, driven by nological deployment[16].

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

Breaking Down The Challenging Task



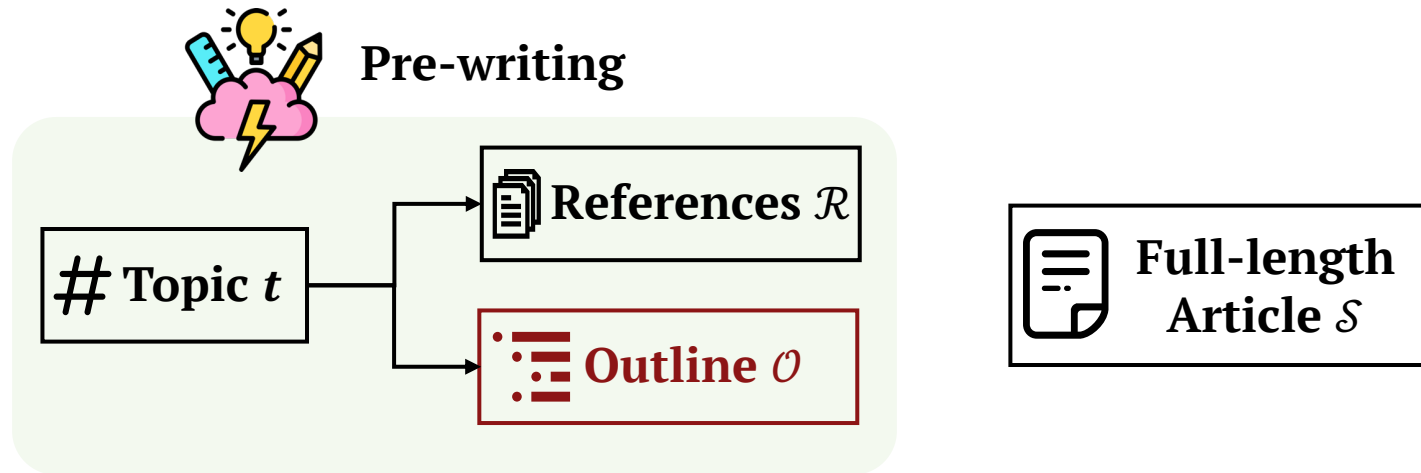
Breaking Down The Challenging Task



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} .

Breaking Down The Challenging Task

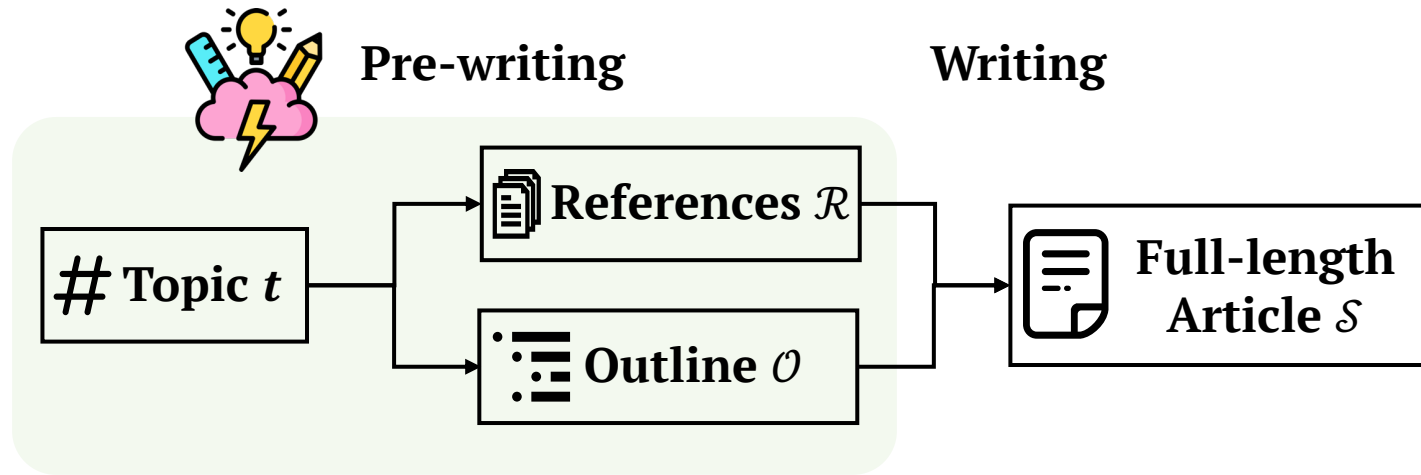


Intruduce outline coverage metrics as a proxy of the pre-writing stage quality.

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} .

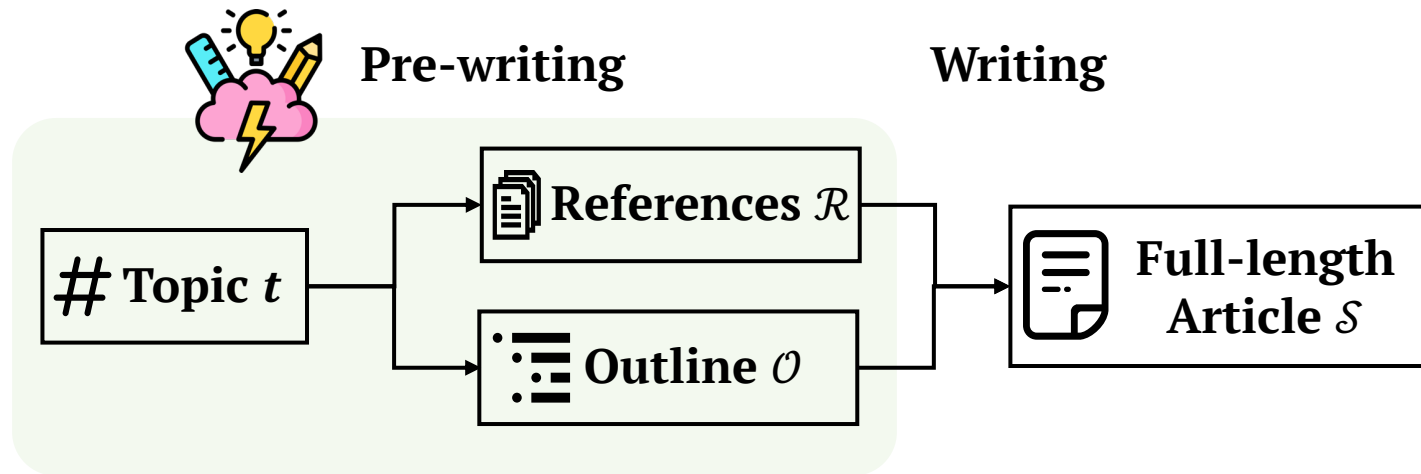
Breaking Down The Challenging Task



The writing task:

Generate a full-length article \mathcal{S} based on $t, \mathcal{R}, \mathcal{O}$.

Focusing On The Pre-writing Stage



Literature Research via **Question Asking**

Directly prompting LLMs to generate questions often results in questions that lack breadth and depth.

Research Question:

How to instruct LLM to ask good questions to research t ?

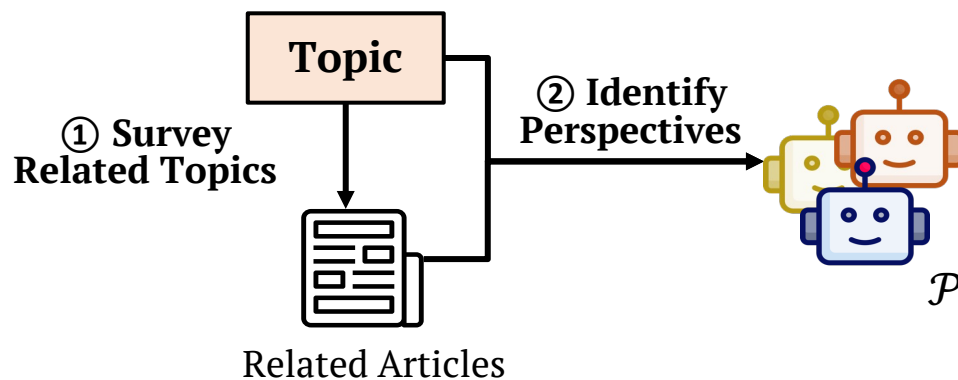
Perspective-Guided Question Asking

Inspiration: Freeman's stakeholder theory in business

- Different stakeholders come with different backgrounds.
 - Adding perspectives provides LLM with prior knowledge.
- They care about different aspects of the same topic.
 - Adding perspectives steers LLM generations to different directions.

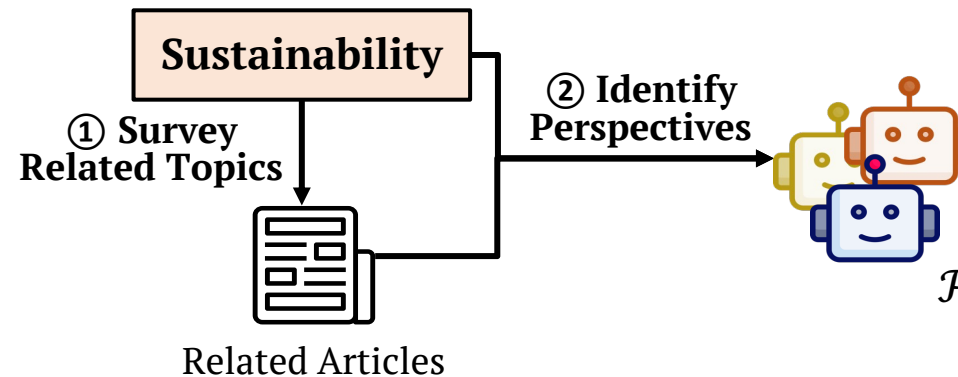
Perspective-Guided Question Asking

STORM reads Wikipedia articles about related topics before searching any information about t .



Perspective-Guided Question Asking

STORM reads Wikipedia articles about related topics before searching any information about t .



E.g., [wiki/Sustainable_development](#);
[wiki/Corporate_social_responsibility](#)

E.g., Social scientist – This editor will bring in a social perspective, focusing on topics such as social sustainability, cultural sustainability, and the role of nongovernmental stakeholders in promoting sustainability.

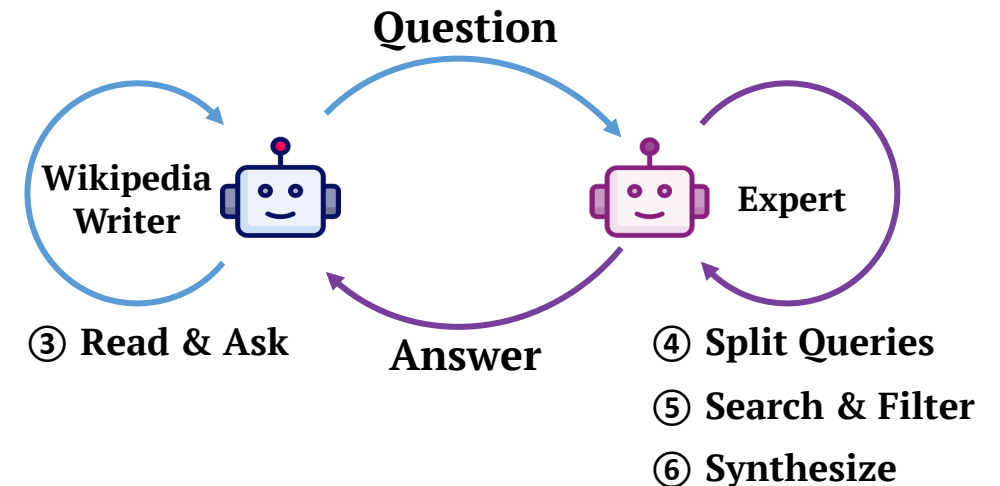
Simulating Conversations To Allow Follow-up Questions

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

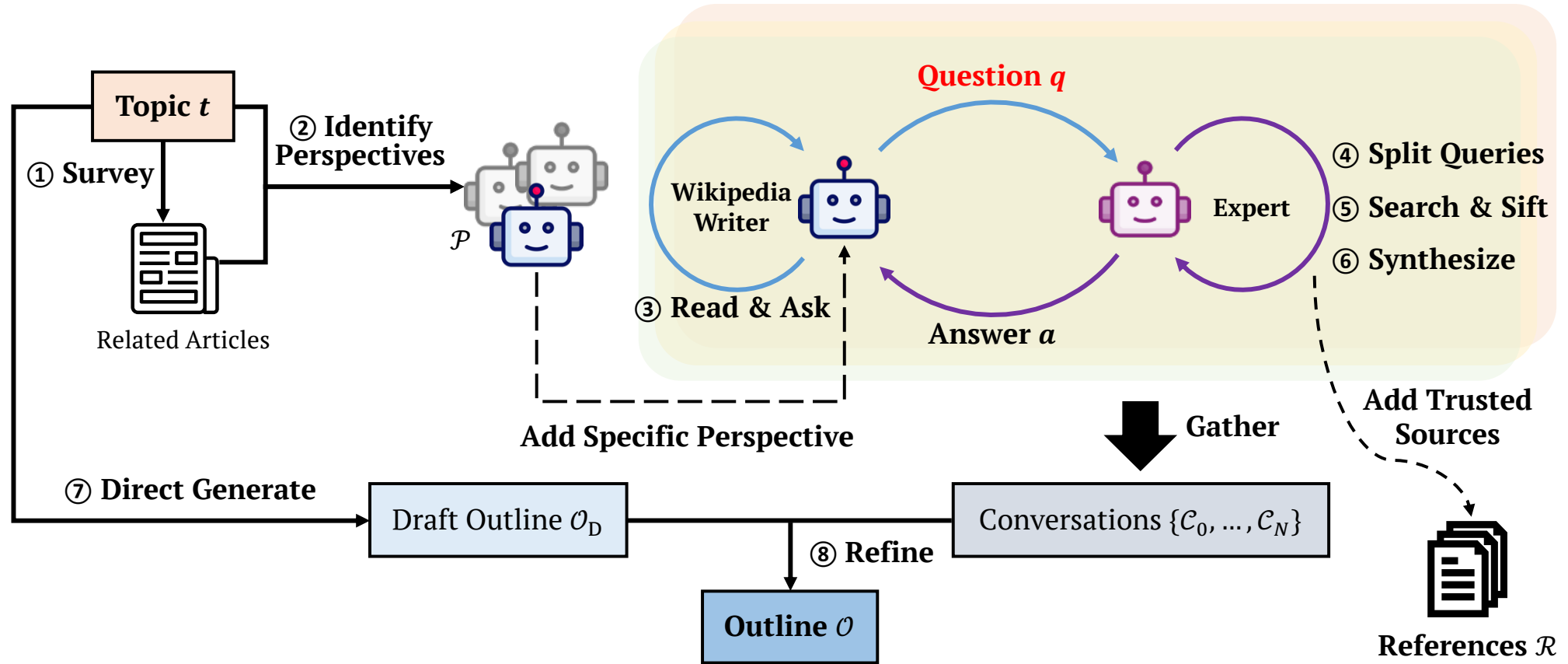
Q: Can you provide me with a list of the participating countries in the 2022 Winter Olympics opening ceremony?

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?



Putting Things Together



Watch the demo video 



Putting Things Together

✕

★ Featured Articles

📁 My Articles

🔧 Create New Article

logout

🔧 Search a New Topic 🔧

Enter the topic you want to learn in depth

the symbiosis of Generative AI and human

Press Enter to submit form

Search

23

Automatic Evaluation

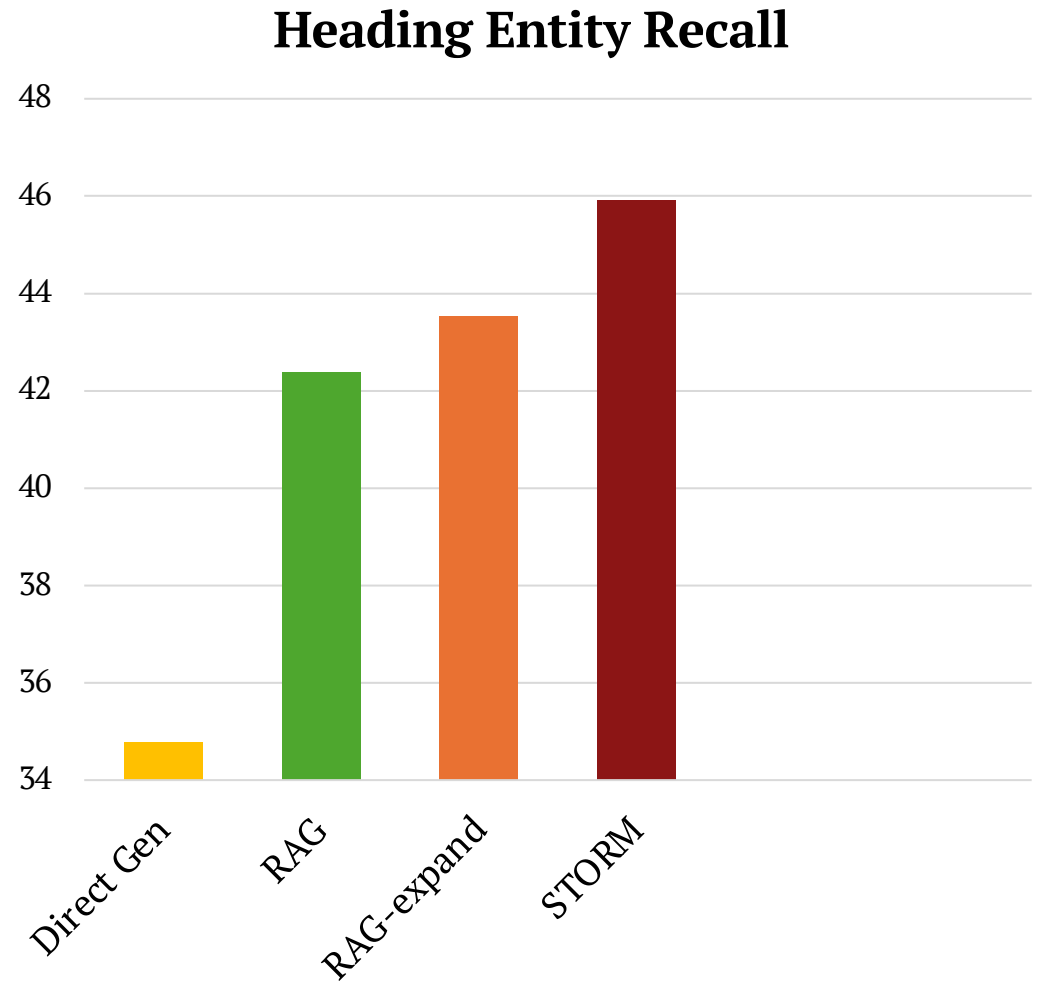
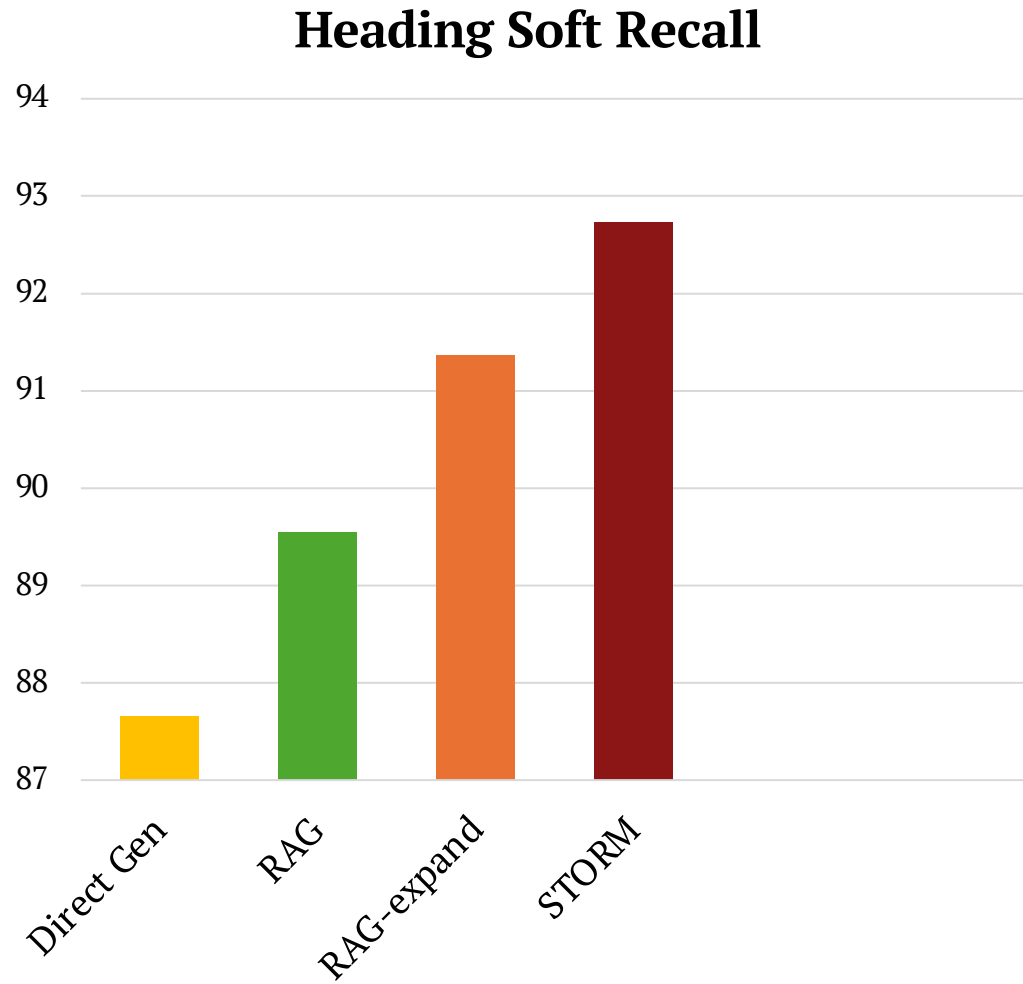
FreshWiki Dataset (n=100)

- We seek out recent, high-quality Wikipedia articles that were created (or very heavily edited) after the training cutoff of the LLMs we test.
- Our process can be repeated at future dates when new LLMs emerge.

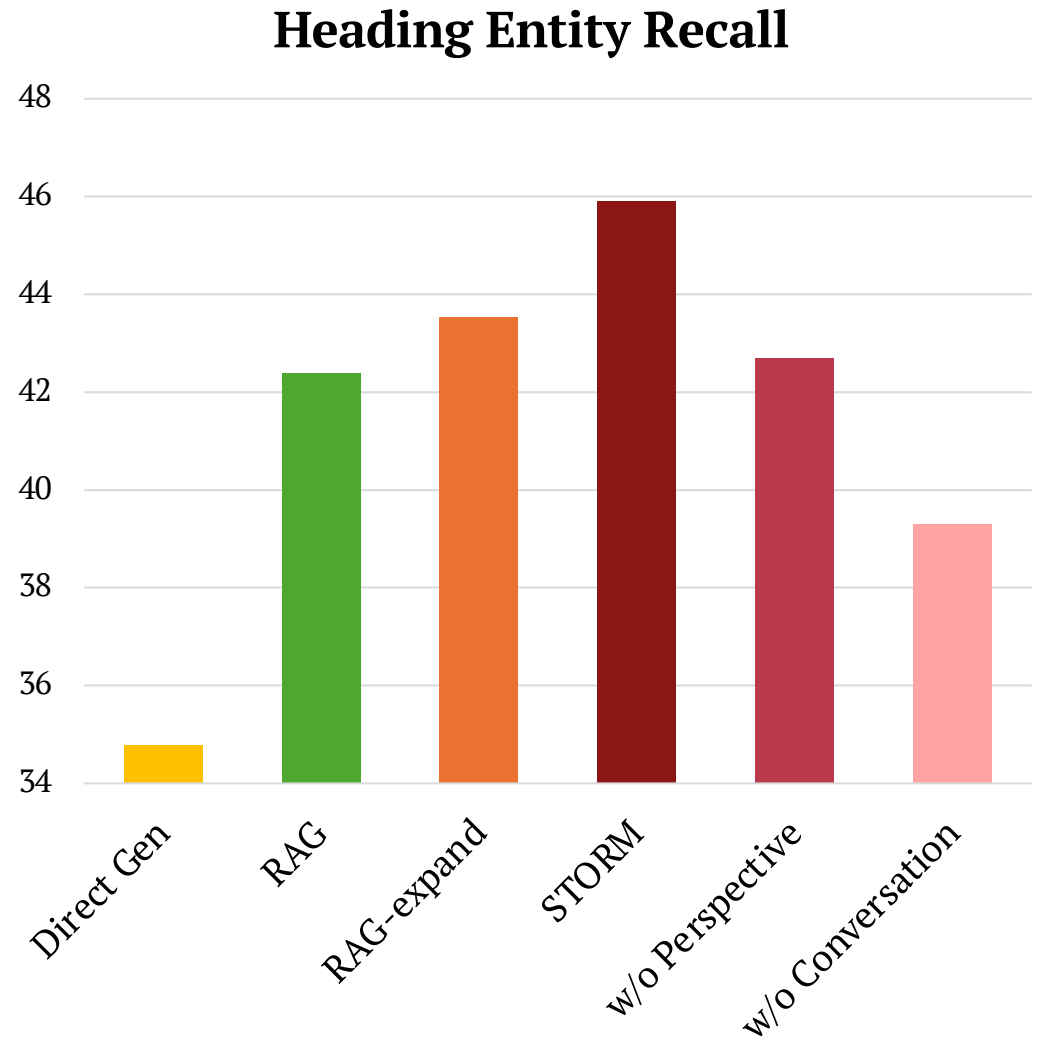
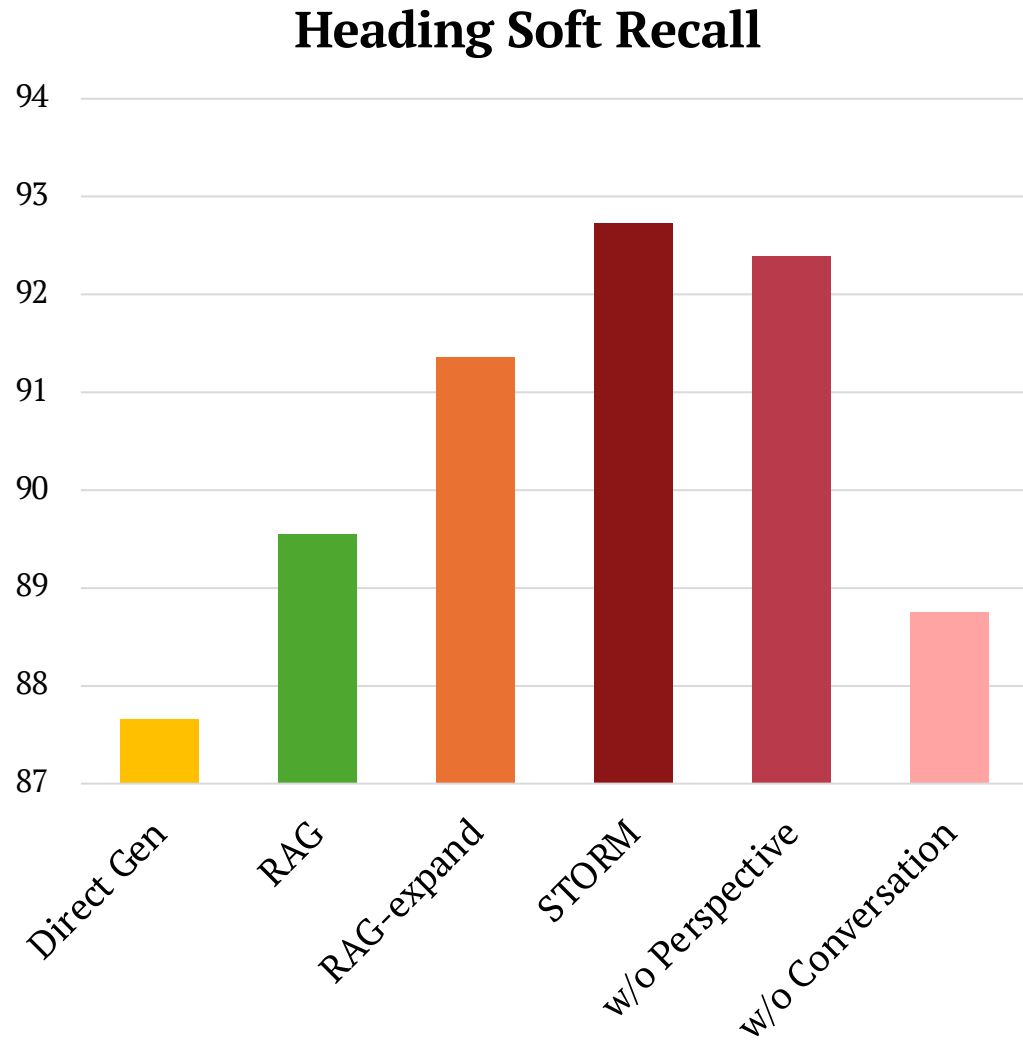
Two-stage Evaluation

- Evaluate the outline quality to assess the pre-writing stage.
heading soft recall, heading entity recall
- Evaluate the quality of the full-length article.
ROUGE scores, entity recall, rubric gradings with Prometheus model

Automatic Evaluation – Outline Quality



Automatic Evaluation – Outline Quality



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.

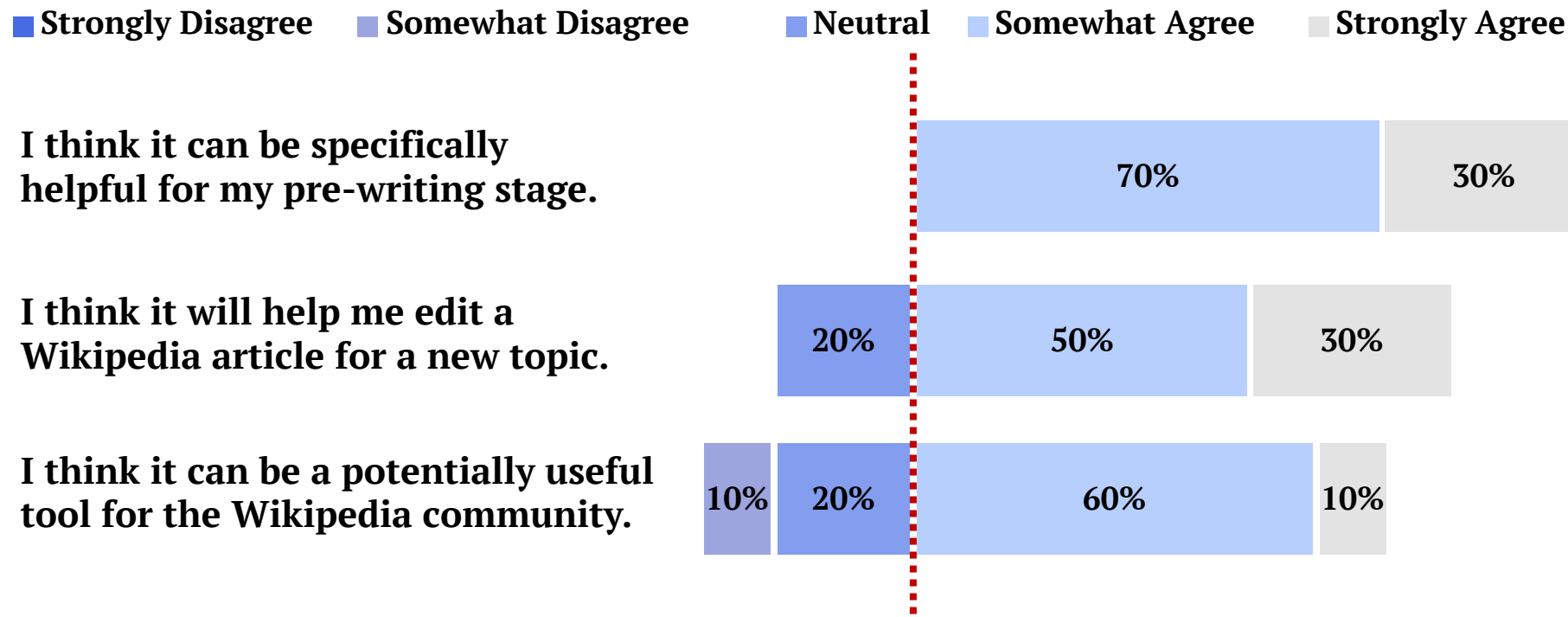
Wikipedia Editor Evaluation

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%

Wikipedia Editor Evaluation

All the participants found STORM useful for their pre-writing stage.



Thank you!

Paper



Code
(4k+ ★)
& Demo



Email: shaoyj@stanford.edu
Website: <https://cs.stanford.edu/~shaoyj/>