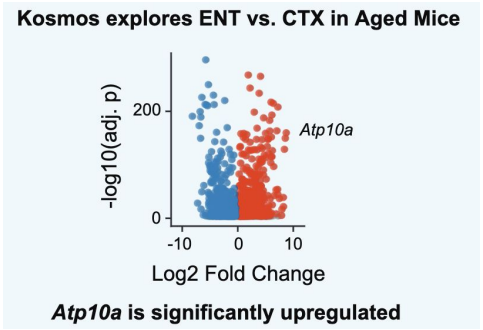
ResearchQA

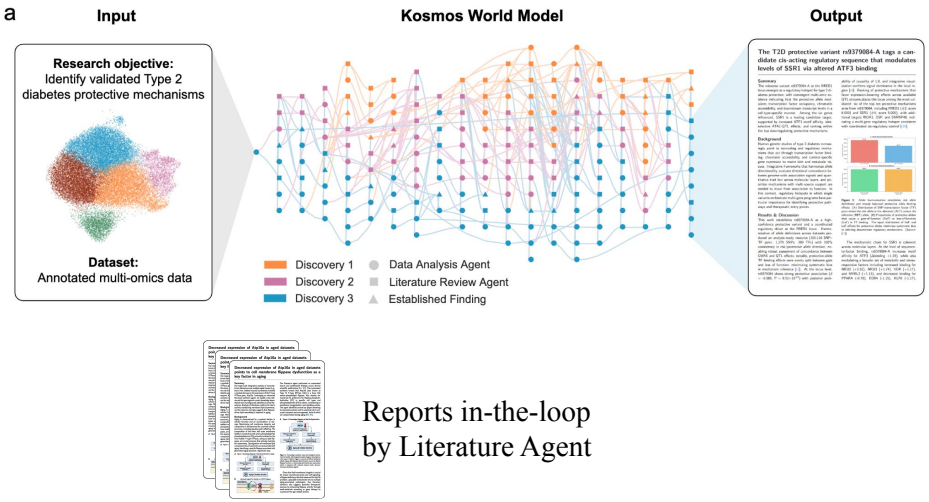
Evaluating Deep Research at Scale

Li S. Yifei*, Allen Chang*, Chaitanya Malaviya, Mark Yatskar

Publication: TACL 2026 (will present in ACL 2026)

Deep Research Agent Can Help Scientific Discovery





<u>Novel Discovery by Kosmos</u>: a new aging mechanism where "flippase collapse" exposes "eat-me" signals on aging neurons, causing immune cells (microglia) to mistakenly destroy them

<u>Deep Research Helps</u>: Kosmos found the Atp10a gene decreasing in aged datasets, then immediately used *literature agent* to identify it as a "flippase" enzyme (in a long-form report format)

Kosmos: An AI Scientist for Autonomous Discovery
Edison Scientific Inc.. 2025.

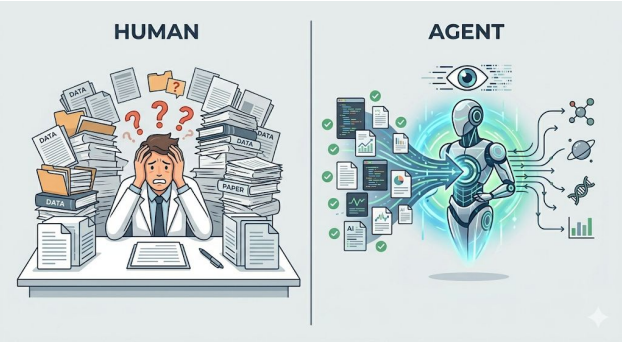
Deep Research Agent Can Help Scientific Discovery

Massive Scale and Speed

- Unmatched Volume: In a single 12-hour run, agents can read 1,500 full-length papers [Kosmos]
- Extreme Efficiency: Compresses the equivalent of over 6 months of expert human research into a single automated session [Kosmos]

Overcoming the "Breadth and Depth Conundrum"

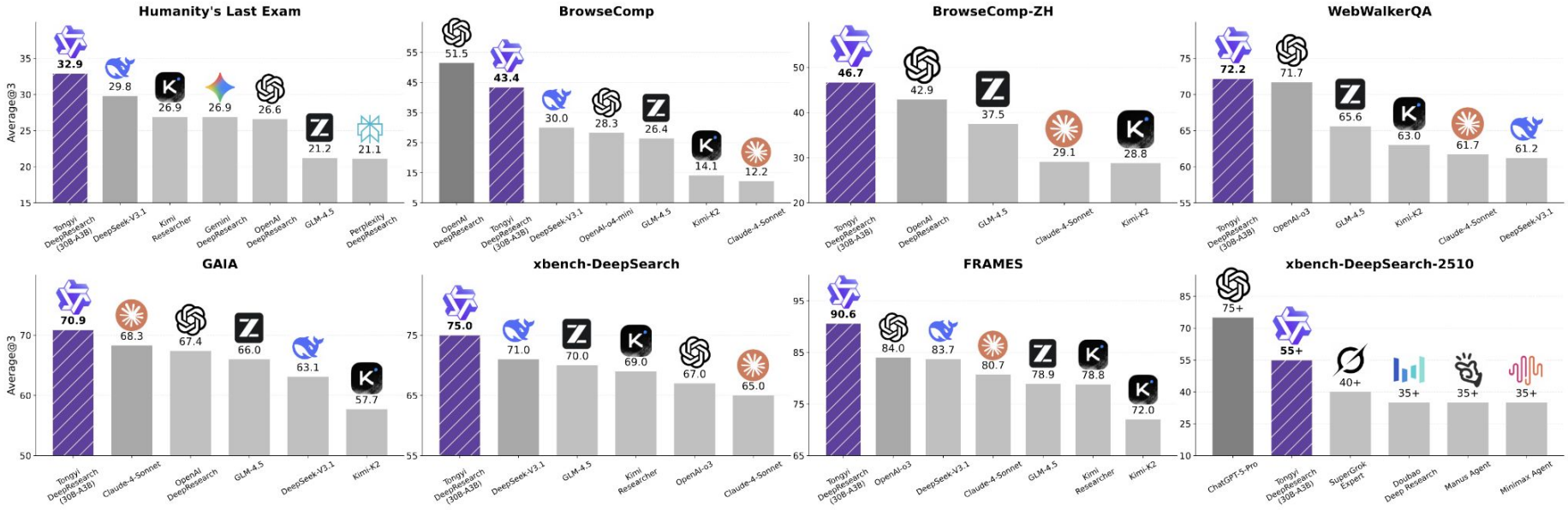
- Modern science demands intense specialization, making it hard for humans to maintain broad, cross-disciplinary insight
- Agents bridge this gap by instantly synthesizing knowledge across diverse, disconnected fields



Towards an AI Co-Scientist. Google. Nature 2026
Robin. Edison Scientific Inc.. Nature 2026
Kosmos. Edison Scientific Inc.. 2025
The Virtual Lab of AI Agents. Stanford. Nature 2025
The AI Scientist. Sakana AI. 2024

Evaluating deep research in report generation and scholarly domains is necessary, but current benchmarks are problematic.

The Lack of Long-Form Deep Research Benchmark



Tongyi Deep Research Technical Report
Tongyi Deep Research Team. 2025.

The Lack of Long-Form Deep Research Benchmark

[Query] Who was the judge of Bártfa (now Bardejov) in 1461?

[Answer] George Stenczel

Humanity's Last Exam
ScaleAI and HLE Team. Nature 2026.

[Query] I am searching for the pseudonym of a writer and biographer who authored numerous books, including their autobiography. In 1980, they also wrote a biography of their father. The writer fell in love with the brother of a philosopher who was the eighth child in their family. The writer was divorced and remarried in the 1940s.

[Answer] Esther Wyndham

BrowseComp: a benchmark for browsing agents
Jason Wei, et al. 2025.

The Lack of Long-Form Deep Research Benchmark

[Query] Who was the judge of Bártfa (now Bardejov) in 1461?

[Answer] George Stenczel

In part due to the fact that
there is more than one correct long-form answer
(thus, making the evaluation subjective and hard-to-construct)

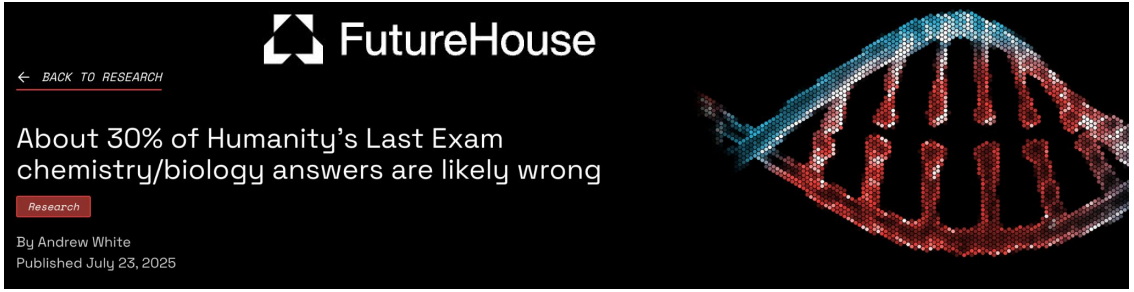
manity's Last Exam
E Team. Nature 2026.

[Query] ... of a writer and biographer who authored numerous books, including their autobiography. In 1980, they also wrote a biography of their father. The writer fell in love with the brother of a philosopher who was the eighth child in their family. The writer was divorced and remarried in the 1940s.

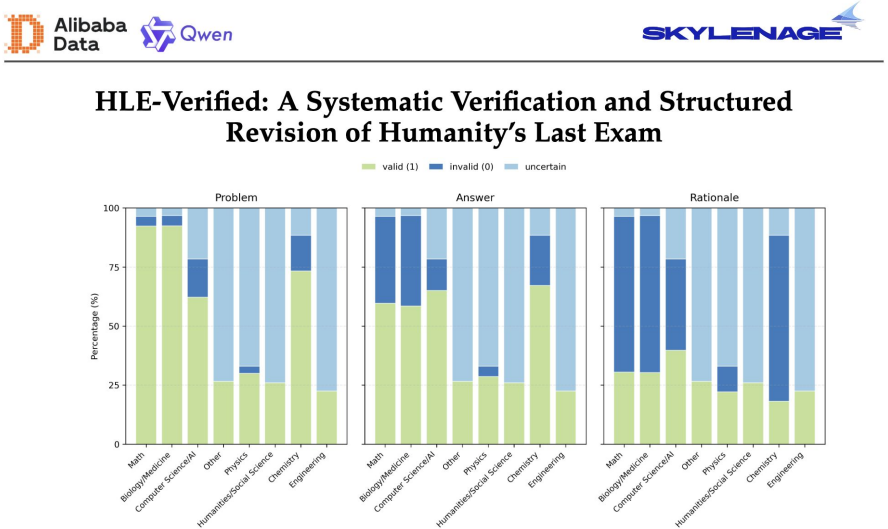
[Answer] Esther Wyndham

BrowseComp: a benchmark for browsing agents
Jason Wei, et al. 2025.

Problematic Deep Research Benchmark

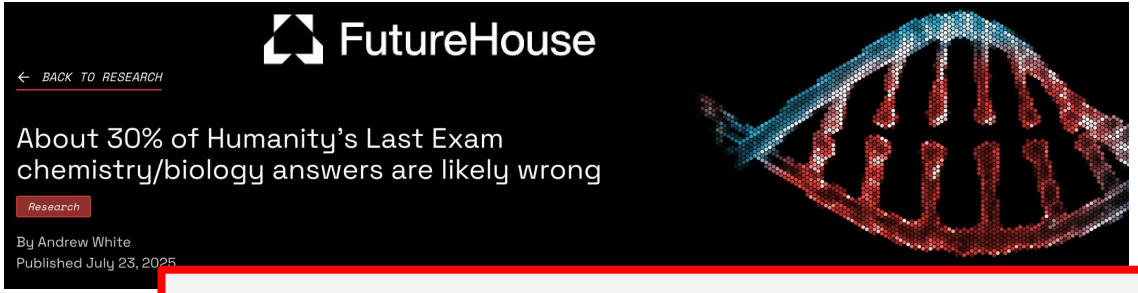


About 30% of Humanity's Last Exam chemistry/biology answers are likely wrong
FutureHouse (Blog). 2025.



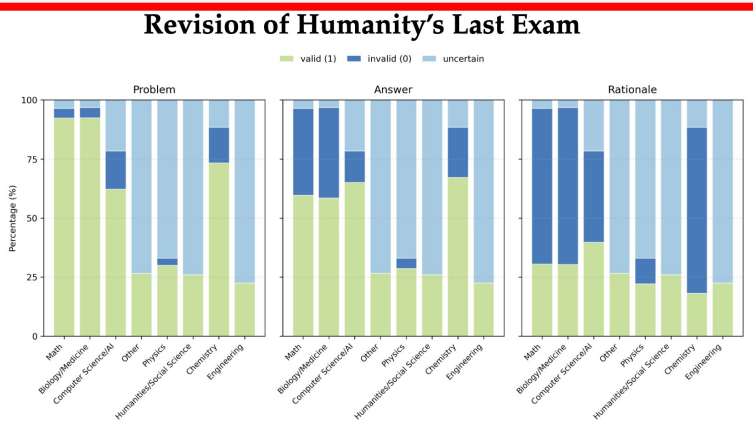
HLE-Verified: A Systematic Verification and Structured Revision of Humanity's Last Exam
Alibaba & Qwen Team. 2026.

Problematic Deep Research Benchmark



Due to the limited, unscalable access to expert supervision
(in HLE, the errors stem from a poorly-designed annotation pipeline.
The team cannot give human verifiers enough time)

ity's Last Exam
are likely wrong
use (Blog). 2025.



HLE-Verified: A Systematic Verification and Structured Revision of Humanity's Last Exam
Alibaba & Qwen Team. 2026.

The Difficulty of Deep Research Evaluation

Difficulty I. More than one correct long-form answer

Difficulty II. The limited, unscalable access to expert supervision

Research Question. Can we address these two difficulties at scale?

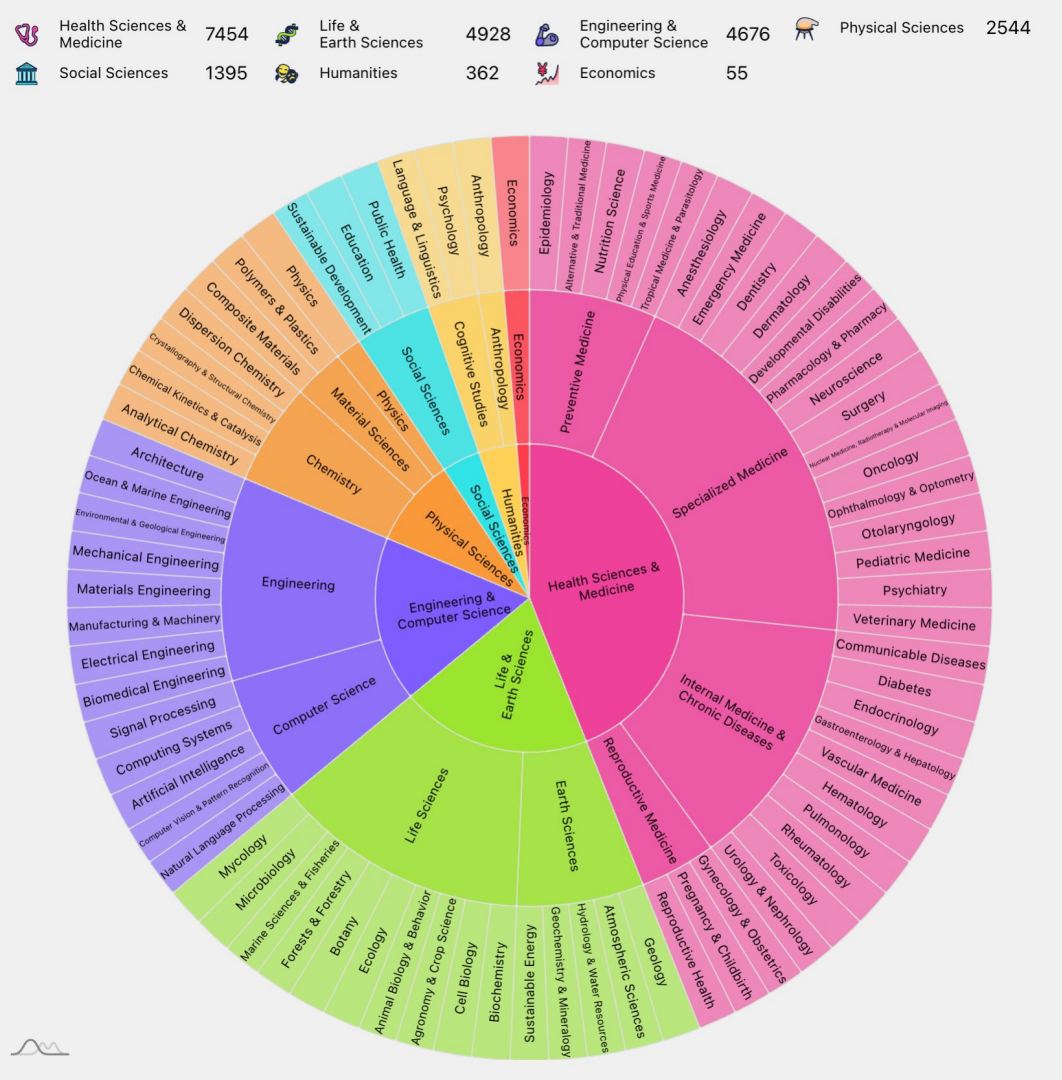
ResearchQA Dataset

Outline

1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. LLM Eval

ResearchQA

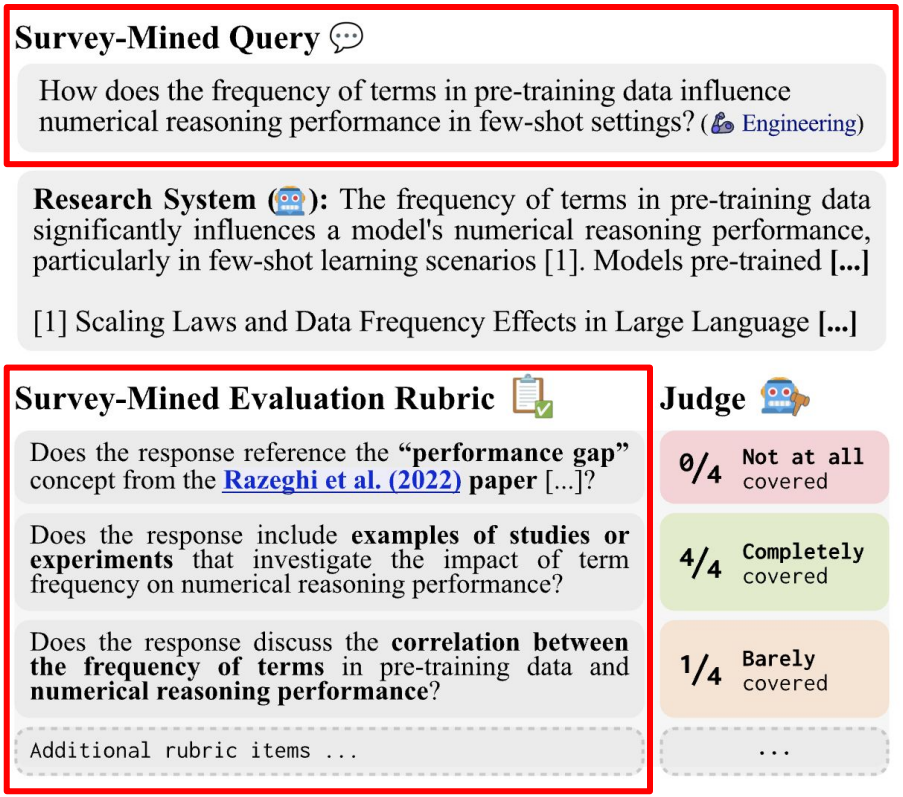
ResearchQA is a long-form scholarly question–answering benchmark consisting of 21K queries and 160K rubric items, spanning 75 research fields.



A Query-Rubric Sample and its Application

For each scholarly query, it provides the corresponding rubric items.

Given this query, a LLM research system (e.g. openai deep research) generates a long-form answer, and an automatic evaluator aggregates the LLM judge results for all the rubric items.

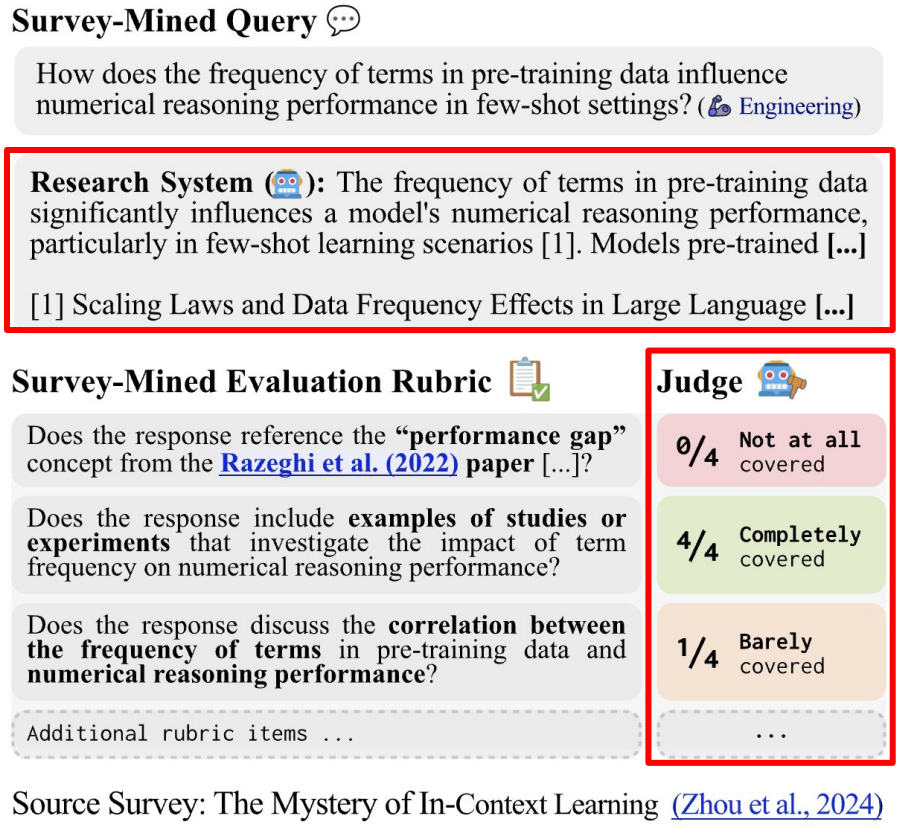


Source Survey: The Mystery of In-Context Learning (Zhou et al., 2024)

A Query-Rubric Sample and its Application

For each scholarly query, it provides the corresponding rubric items.

Given this query, a LLM research system (e.g. openai deep research) generates a long-form answer, and an automatic evaluator aggregates the LLM judge results for all the rubric items.



Source Survey: The Mystery of In-Context Learning (Zhou et al., 2024)

Rubric Can Evaluate Various Correct Long-Form Answers at Once

- Defines correctness by properties, not a single gold answer
  - Credits multiple valid reasoning paths
- Atomic items → independently checkable, lower judge variance
  - Natural partial credit (per-item scoring)
- Interpretable and actionable
  - Better for human understanding and downstream processing

Outline

1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. LLM Eval

Limitations of Previous Scholarly Benchmarks

Limited **size**: unreliable evaluation under high variance
Limited **domain**: mainly engineering domain (easiest to recruit colleagues for annotations)

Scholarly QA Benchmark	# Scholarly Queries	# Scholarly Fields & Research Domains	Abstractive Eval. Format	Multi-Doc Reasoning	Evaluation Rubrics	Auto-Generated
QASPER (1)	5.0K	1 (NLP ← 💪)	✗ Extractive	✗	✗	✗
QASA (2)	1.8K	1 (AI ← 💪)	✗ Extractive	✗	✗	✗
PubMedQA (3)	1.0K	— (🧬)	✗ Yes/No	✗	✗	✓
SciQA (4)	2.5K	1 (CS ← 💪)	✗ Extractive	✗	✗	✓
KIWI (5)	0.2K	1 (NLP ← 💪)	✓ Abstractive	✓	✗	✗
SciDQA (6)	2.9K	1 (AI ← 💪)	✗ Extractive	✓	✗	✓
SciQAG (7)	188.0K	20 (C&MS ← 🦾)	✗ Extractive	✗	✗	✓
ScholarQABench (8)	3.0K	— (🧬💪🦾)	✓ Abstractive	✓	✓	✗
SciArena (9)	8.2K	— (🧬💲💪🦾🏛️🎭📈)	✓ Abstractive	✓	✗	✗
RESEARCHQA (Ours)	21.4K	75 (🧬💲💪🦾🏛️🎭📈)	✓ Abstractive	✓	✓	✓

Limitations of Previous Scholarly Benchmarks

Limited **difficulty**: mostly short-form, single-document, and require nothing beyond reading comprehension

Scholarly QA Benchmark	# Scholarly Queries	# Scholarly Fields & Research Domains	Abstractive Eval. Format	Multi-Doc Reasoning	Evaluation Rubrics	Auto-Generated
QASPER (1)	5.0K	1 (NLP ← 💪)	✗ Extractive	✗	✗	✗
QASA (2)	1.8K	1 (AI ← 💪)	✗ Extractive	✗	✗	✗
PubMedQA (3)	1.0K	— (🩺)	✗ Yes/No	✗	✗	✓
SciQA (4)	2.5K	1 (CS ← 💪)	✗ Extractive	✗	✗	✓
KIWI (5)	0.2K	1 (NLP ← 💪)	✓ Abstractive	✓	✗	✗
SciDQA (6)	2.9K	1 (AI ← 💪)	✗ Extractive	✓	✗	✓
SciQAG (7)	188.0K	20 (C&MS ← 🔬)	✗ Extractive	✗	✗	✓
ScholarQABench (8)	3.0K	— (🩺💪🔬)	✓ Abstractive	✓	✓	✗
SciArena (9)	8.2K	— (🩺🧬💪🔬🏛️🎭📈)	✓ Abstractive	✓	✗	✗
RESEARCHQA (Ours)	21.4K	75 (🩺🧬💪🔬🏛️🎭📈)	✓ Abstractive	✓	✓	✓

Limitations of Previous Scholarly Benchmarks

Limited **evaluation structure**: lack of query-specific rubrics for fine-grained, interpretable evaluation

Scholarly QA Benchmark	# Scholarly Queries	# Scholarly Fields & Research Domains	Abstractive Eval. Format	Multi-Doc Reasoning	Evaluation Rubrics	Auto-Generated
QASPER (1)	5.0K	1 (NLP ← 💪)	✗ Extractive	✗	✗	✗
QASA (2)	1.8K	1 (AI ← 💪)	✗ Extractive	✗	✗	✗
PubMedQA (3)	1.0K	— (🫀)	✗ Yes/No	✗	✗	✓
SciQA (4)	2.5K	1 (CS ← 💪)	✗ Extractive	✗	✗	✓
KIWI (5)	0.2K	1 (NLP ← 💪)	✓ Abstractive	✓	✗	✗
SciDQA (6)	2.9K	1 (AI ← 💪)	✗ Extractive	✓	✗	✓
SciQAG (7)	188.0K	20 (C&MS ← 🔬)	✗ Extractive	✗	✗	✓
ScholarQABench (8)	3.0K	— (🫀💪🔬)	✓ Abstractive	✓	✓	✗
SciArena (9)	8.2K	— (🫀🧬💪🔬🏛️🎭📈)	✓ Abstractive	✓	✗	✗
RESEARCHQA (Ours)	21.4K	75 (🫀🧬💪🔬🏛️🎭📈)	✓ Abstractive	✓	✓	✓

Limitations of Previous Scholarly Benchmarks

Limited **scalability**: hard to grow, hard to stay live to latest updates

Scholarly QA Benchmark	# Scholarly Queries	# Scholarly Fields & Research Domains	Abstractive Eval. Format	Multi-Doc Reasoning	Evaluation Rubrics	Auto-Generated
QASPER (1)	5.0K	1 (NLP ← 💪)	✗ Extractive	✗	✗	✗
QASA (2)	1.8K	1 (AI ← 💪)	✗ Extractive	✗	✗	✗
PubMedQA (3)	1.0K	— (🧬)	✗ Yes/No	✗	✗	✓
SciQA (4)	2.5K	1 (CS ← 💪)	✗ Extractive	✗	✗	✓
KIWI (5)	0.2K	1 (NLP ← 💪)	✓ Abstractive	✓	✗	✗
SciDQA (6)	2.9K	1 (AI ← 💪)	✗ Extractive	✓	✗	✓
SciQAG (7)	188.0K	20 (C&MS ← 🔭)	✗ Extractive	✗	✗	✓
ScholarQABench (8)	3.0K	— (🧬💪🔭)	✓ Abstractive	✓	✓	✗
SciArena (9)	8.2K	— (🧬💰💪🔭🏛️🎭📈)	✓ Abstractive	✓	✗	✗
RESEARCHQA (Ours)	21.4K	75 (🧬💰💪🔭🏛️🎭📈)	✓ Abstractive	✓	✓	✓

Our Benchmark (ResearchQA)

- Very large size; Across 75 domains; and Scalable
- Tackle long-form answer of difficult query
- Provide actionable rubric items specialized to each query

Scholarly QA Benchmark	# Scholarly Queries	# Scholarly Fields & Research Domains	Abstractive Eval. Format	Multi-Doc Reasoning	Evaluation Rubrics	Auto-Generated
QASPER (1)	5.0K	1 (NLP ← 💪)	✗ Extractive	✗	✗	✗
QASA (2)	1.8K	1 (AI ← 💪)	✗ Extractive	✗	✗	✗
PubMedQA (3)	1.0K	— (🫀)	✗ Yes/No	✗	✗	✓
SciQA (4)	2.5K	1 (CS ← 💪)	✗ Extractive	✗	✗	✓
KIWI (5)	0.2K	1 (NLP ← 💪)	✓ Abstractive	✓	✗	✗
SciDQA (6)	2.9K	1 (AI ← 💪)	✗ Extractive	✓	✗	✓
SciQAG (7)	188.0K	20 (C&MS ← 🦐)	✗ Extractive	✗	✗	✓
ScholarQABench (8)	3.0K	— (🫀💪🦐)	✓ Abstractive	✓	✓	✗
SciArena (9)	8.2K	— (🫀🧬💪🦐🏛️🎭📈)	✓ Abstractive	✓	✗	✗
RESEARCHQA (Ours)	21.4K	75 (🫀🧬💪🦐🏛️🎭📈)	✓ Abstractive	✓	✓	✓

Our Benchmark (ResearchQA)

How?

Research expertise is abundant and already captured in survey papers

We *distill* survey articles into query-rubric data for evaluation => thus scalable!

Recap: Now we solved the two difficulties:

Difficulty I. More than one correct long-form answer **- resolved by rubric**

Recap: Now we solved the two difficulties:

Difficulty I. More than one correct long-form answer **- resolved by rubric**

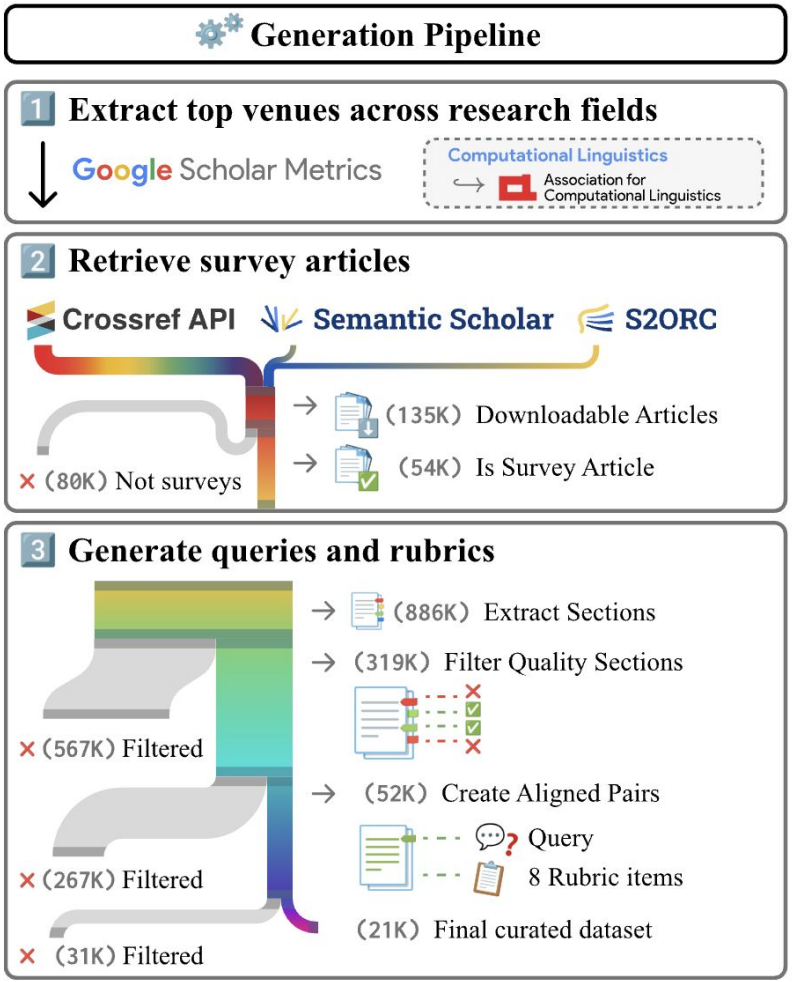
Difficulty II. The limited, unscalable access to expert supervision **- resolved by survey paper distillation**

Recap: Now we solved the two difficulties:

Difficulty I. More than one correct long-form answer **- resolved by rubric**

Difficulty II. The limited, unscalable access to expert supervision **- resolved by survey paper distillation**

But how does this work in detail?

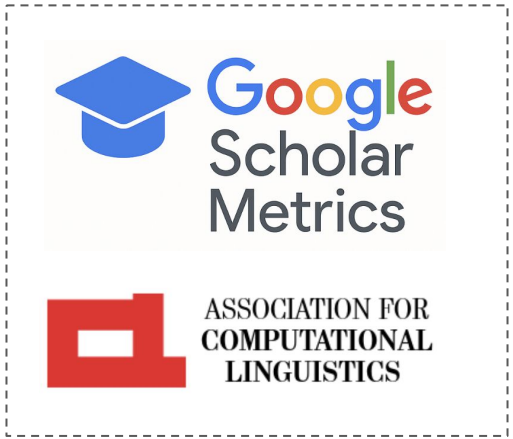


Outline

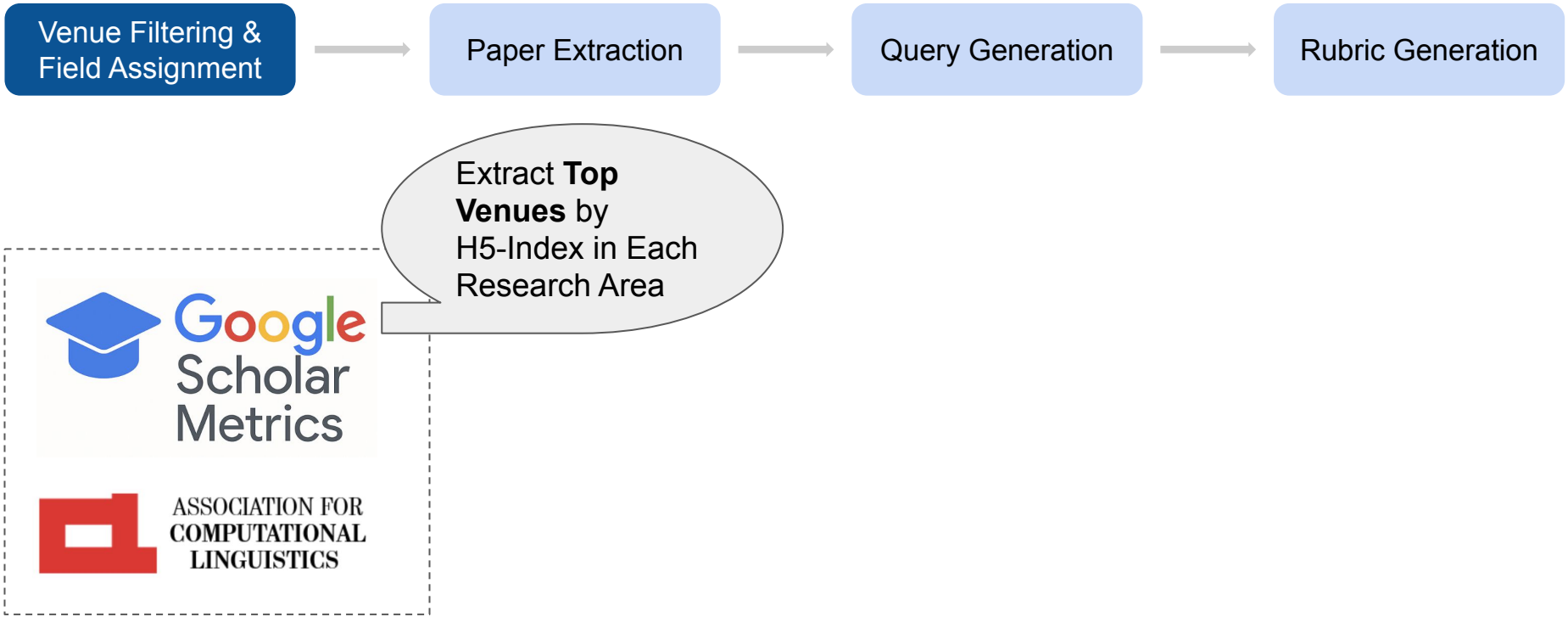
1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. LLM Eval

ResearchQA Data Synthesis Pipeline: Step 1/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → Rubric Generation



ResearchQA Data Synthesis Pipeline: Step 1/4



ResearchQA Data Synthesis Pipeline: Step 1/4

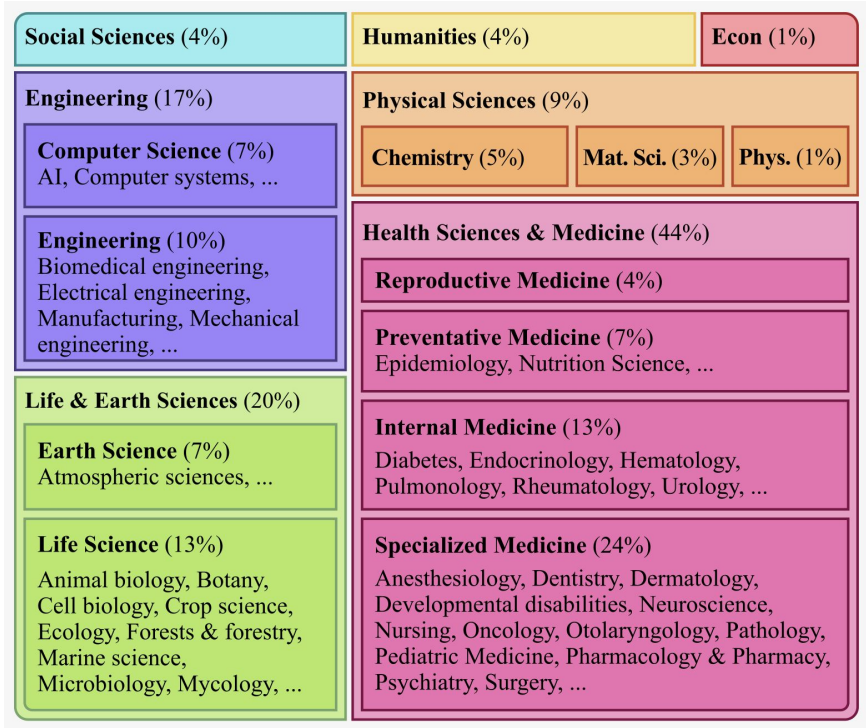
Venue Filtering & Field Assignment → Paper Extraction → Query Generation → Rubric Generation

Google Scholar Metrics

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

Extract **Top Venues** by H5-Index in Each Research Area

Assign **Research Fields** for Each Top Venue by Rules



ResearchQA Data Synthesis Pipeline: Step 2/4

Corpus Sources

ResearchQA Data Synthesis Pipeline: Step 2/4

Venue Filtering & Field Assignment → **Paper Extraction** → Query Generation → Rubric Generation

Crossref API
Semantic Scholar
S2ORC

Corpus Sources

Survey Paper Filtering

- Published in a **Top Venue**
- **Openly Accessible**
- Identified as a **Survey Paper**
  (via keyword match and LLM classification of title)

ResearchQA Data Synthesis Pipeline: Step 2/4

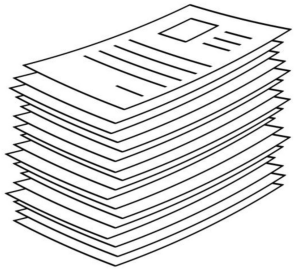
Venue Filtering & Field Assignment → **Paper Extraction** → Query Generation → Rubric Generation

Crossref API
Semantic Scholar
S2ORC

Corpus Sources

Survey Paper Filtering

- Published in a **Top Venue**
- **Openly Accessible**
- Identified as a **Survey Paper**
  (via keyword match and LLM classification of title)



Final Survey Corpus
(54K papers)

ResearchQA Data Synthesis Pipeline: Step 3/4

Input a Survey Paper's **Section**

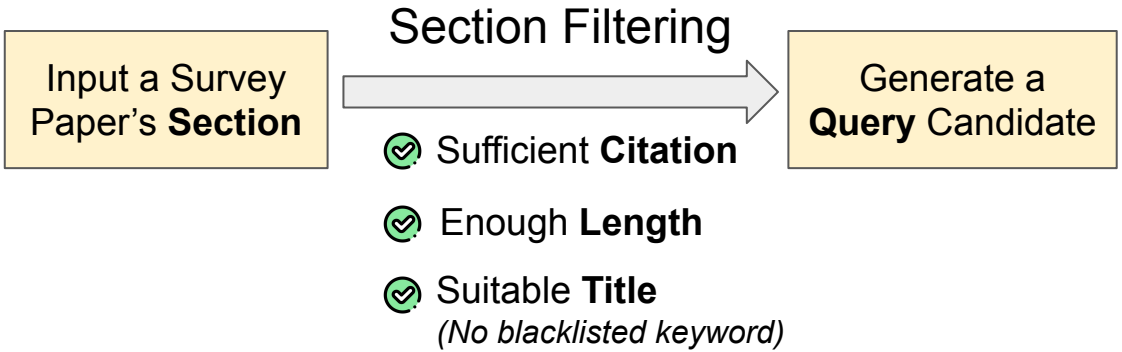
ResearchQA Data Synthesis Pipeline: Step 3/4

Input a Survey Paper's **Section**

Section Filtering

- Sufficient **Citation**
- Enough **Length**
- Suitable **Title**
  (No blacklisted keyword)

ResearchQA Data Synthesis Pipeline: Step 3/4



ResearchQA Data Synthesis Pipeline: Step 3/4

Input a Survey Paper's **Section**

Section Filtering

- Sufficient **Citation**
- Enough **Length**
- Suitable **Title**
  (No blacklisted keyword)

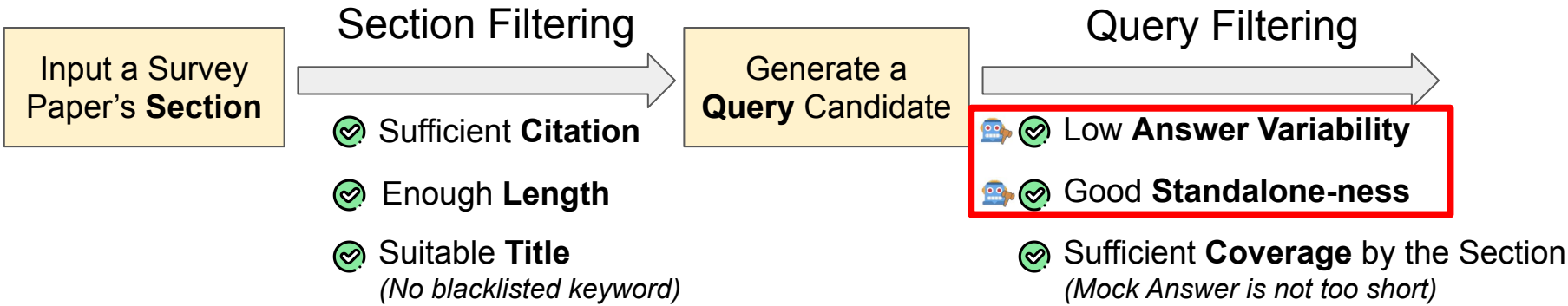
Generate a **Query** Candidate

Query Filtering

- Low **Answer Variability**
- Good **Standalone-ness**
- Sufficient **Coverage** by the Section
  (Mock Answer is not too short)

I am a LLM Judge

ResearchQA Data Synthesis Pipeline: Step 3/4



(1) Answer Variability: different experts are likely to provide similar answers to a query
(2) Standalone-ness: understandable by experts without extra decontextualization

I am a LLM Judge

ResearchQA Data Synthesis Pipeline: Step 3/4

Venue Filtering & Field Assignment → Paper Extraction → **Query Generation** → Rubric Generation

Input a Survey Paper's **Section**

Section Filtering

- Sufficient **Citation**
- Enough **Length**
- Suitable **Title**
  (No blacklisted keyword)

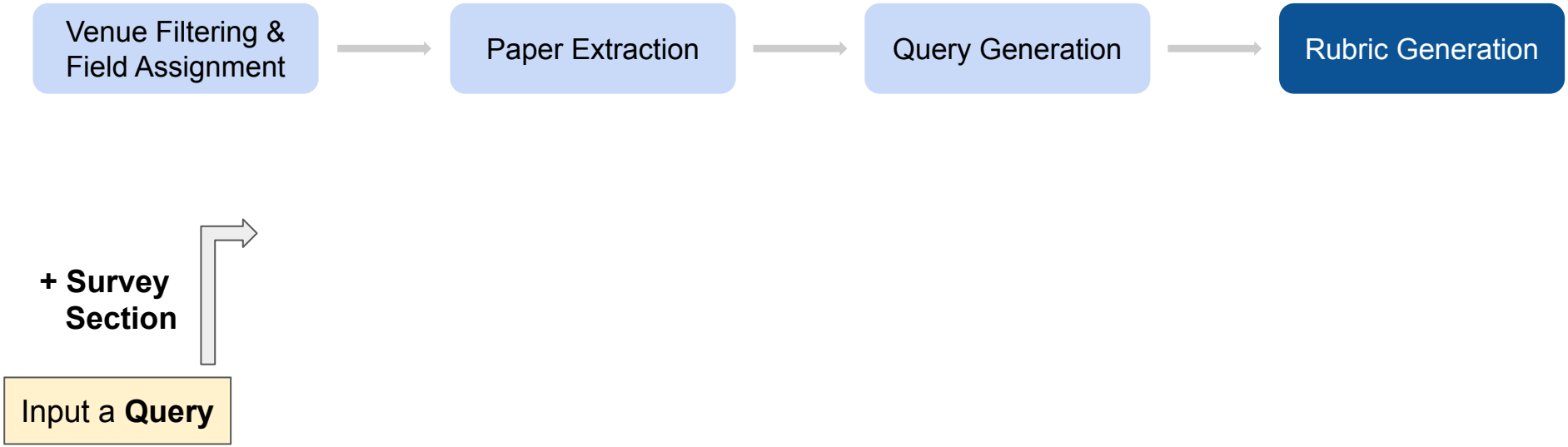
Generate a **Query** Candidate

Query Filtering

- Low **Answer Variability**
- Good **Standalone-ness**
- Sufficient **Coverage** by the Section
  (Mock Answer is not too short)

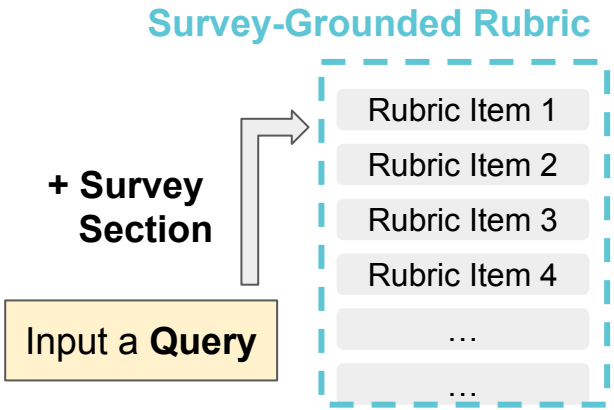
I am a LLM Judge

ResearchQA Data Synthesis Pipeline: Step 4/4



ResearchQA Data Synthesis Pipeline: Step 4/4

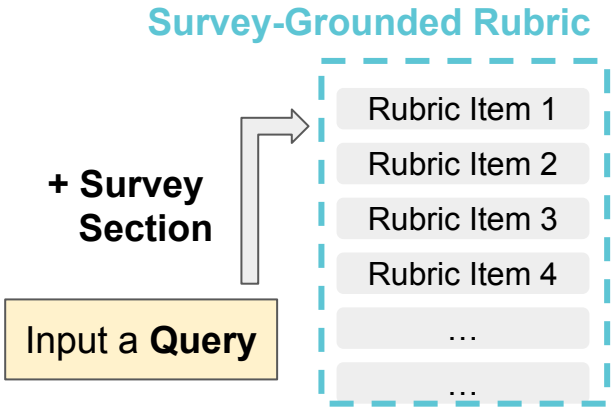
Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**



Rubric is generated from *two* sources:
(1) **Survey-grounded context**

ResearchQA Data Synthesis Pipeline: Step 4/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**

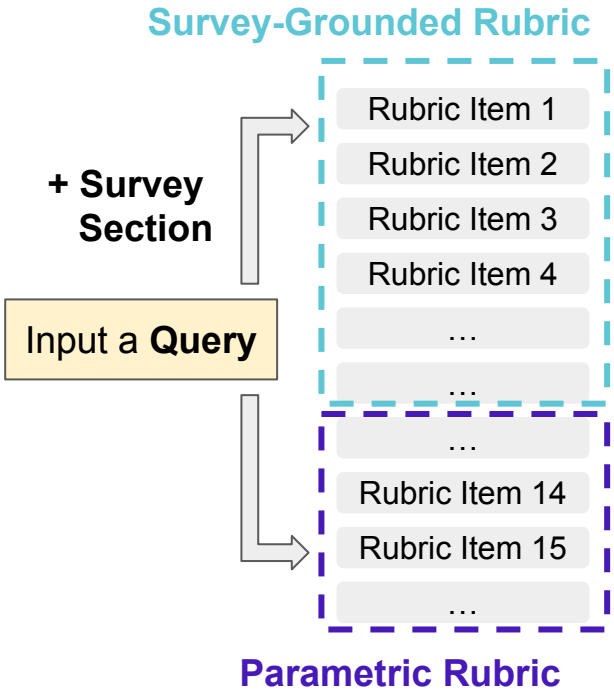


Rubric is generated from *two* sources:
(1) **Survey-grounded context**

However, while survey-grounded rubric contain specific contents, it may not have enough generic contents

ResearchQA Data Synthesis Pipeline: Step 4/4

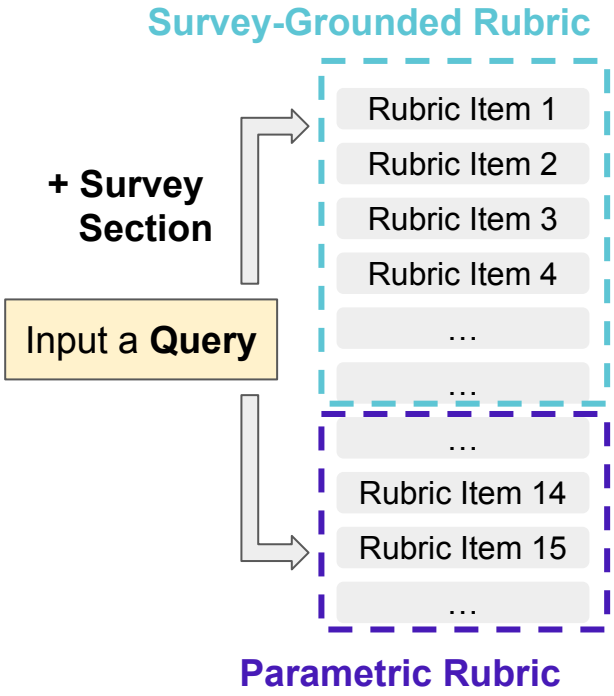
Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**



Rubric is generated from *two* sources:
(1) **Survey-grounded context**
(2) **Parametric memory**

ResearchQA Data Synthesis Pipeline: Step 4/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → Rubric Generation

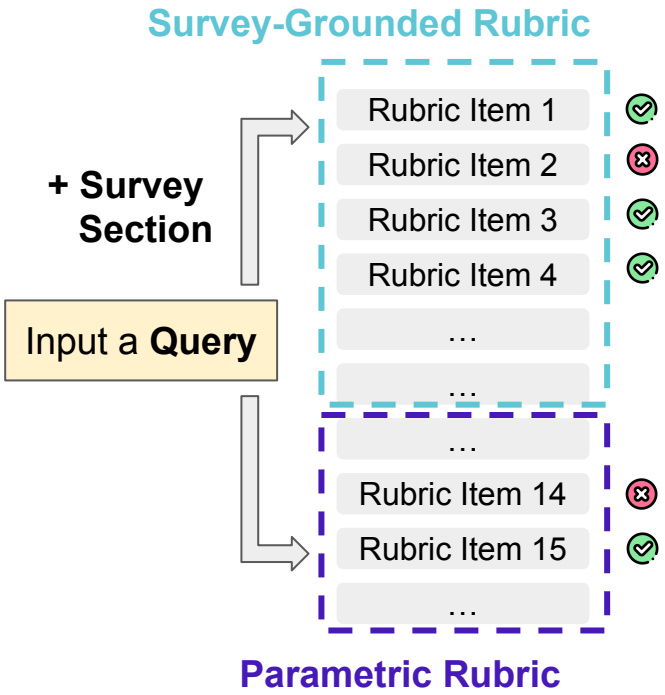


Rubric is generated from *two* sources:
(1) **Survey-grounded context**
(2) **Parametric memory**

Post-hoc filtering to avoid the hallucination of papers' name (if applicable)

ResearchQA Data Synthesis Pipeline: Step 4/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**

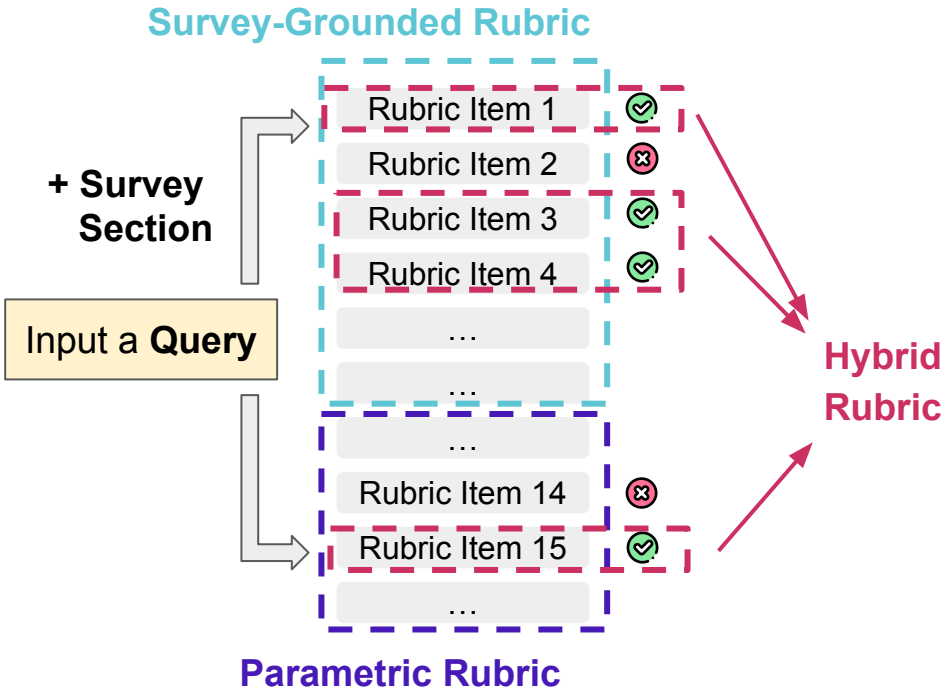


To get the best ones in these rubric items, they are then *deduplicated* and *reranked* by LLM …

- If survives from both deduplication and reranking
- Otherwise

ResearchQA Data Synthesis Pipeline: Step 4/4

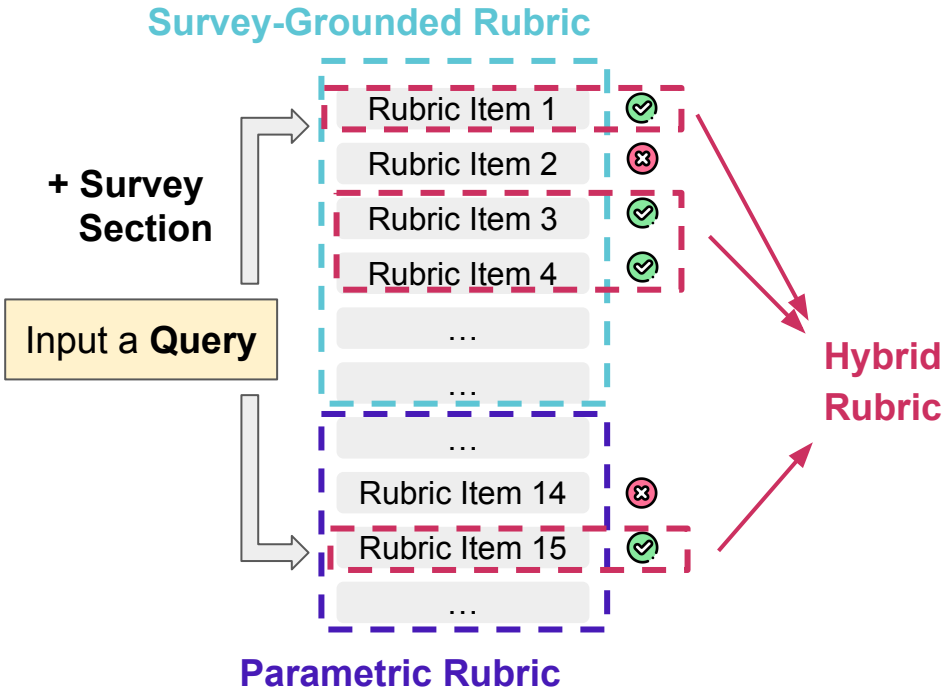
Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**



To get the best ones in these rubric items, they are then *deduplicated* and *reranked* by LLM, and construct **hybrid rubric**.

ResearchQA Data Synthesis Pipeline: Step 4/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**

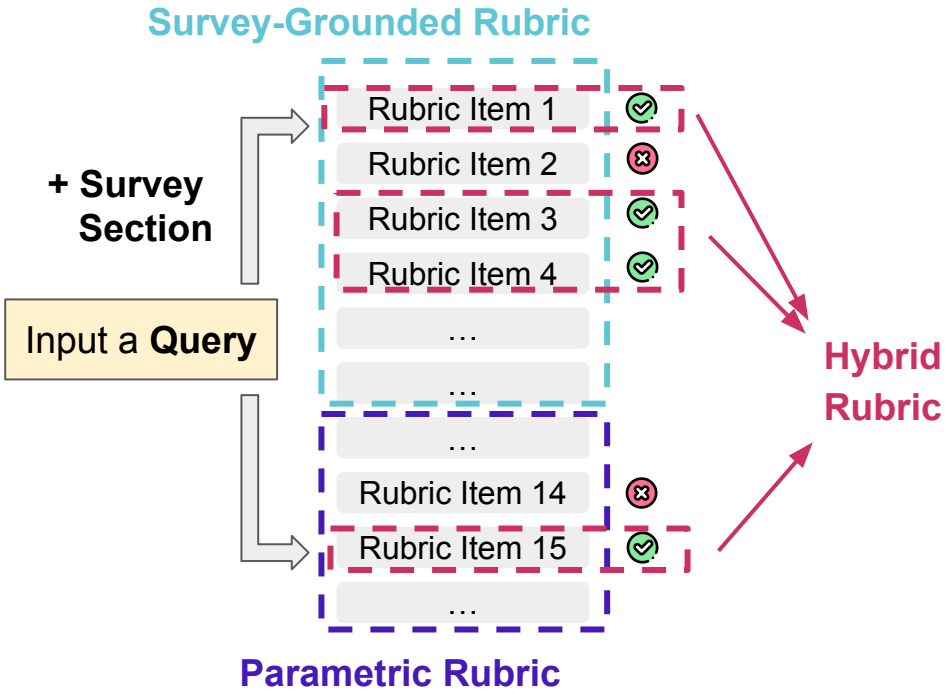


To get the best ones in these rubric items, they are then *deduplicated* and *reranked* by LLM, and construct **hybrid rubric**.

Now we can get the best of both: survey-grounded rubric for specific contents, and parametric rubric for generic contents

ResearchQA Data Synthesis Pipeline: Step 4/4

Venue Filtering & Field Assignment → Paper Extraction → Query Generation → **Rubric Generation**

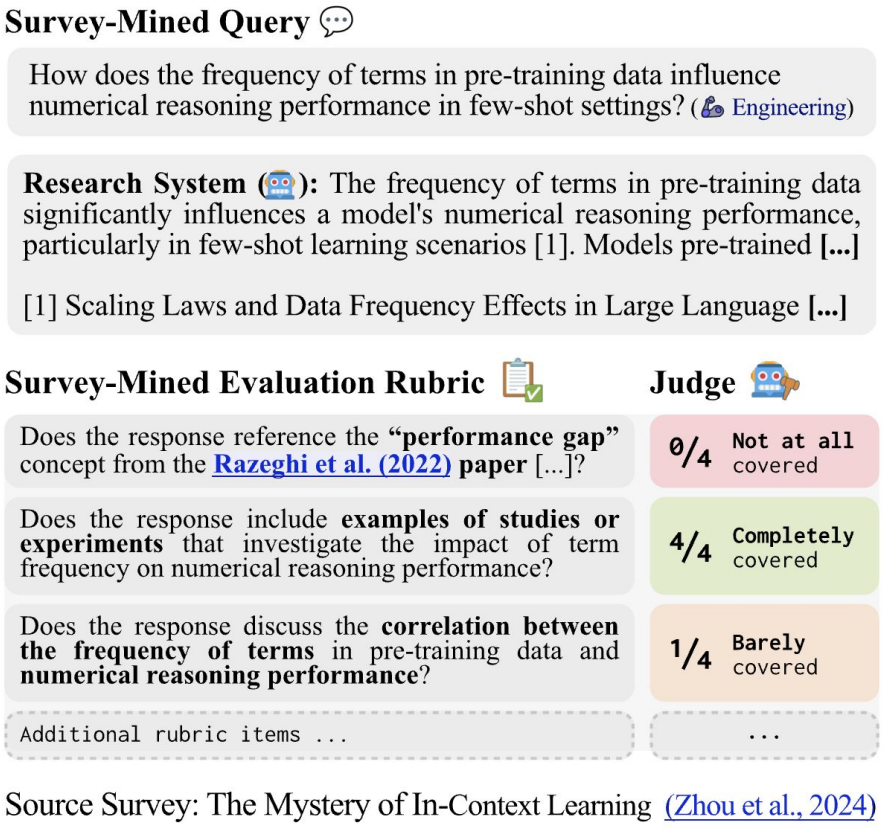


In the end, we have three types of rubric set: **hybrid rubric**, **survey rubric** and **parametric rubric**.

Recap: A Query-Rubric Sample and its Application

For each scholarly query, it provides the corresponding rubric items, grounded on the survey paper sections.

Given this query, a LLM research system (e.g. openai deep research) generates a long-form answer, and an automatic evaluator aggregates the LLM judge results for all the rubric items.



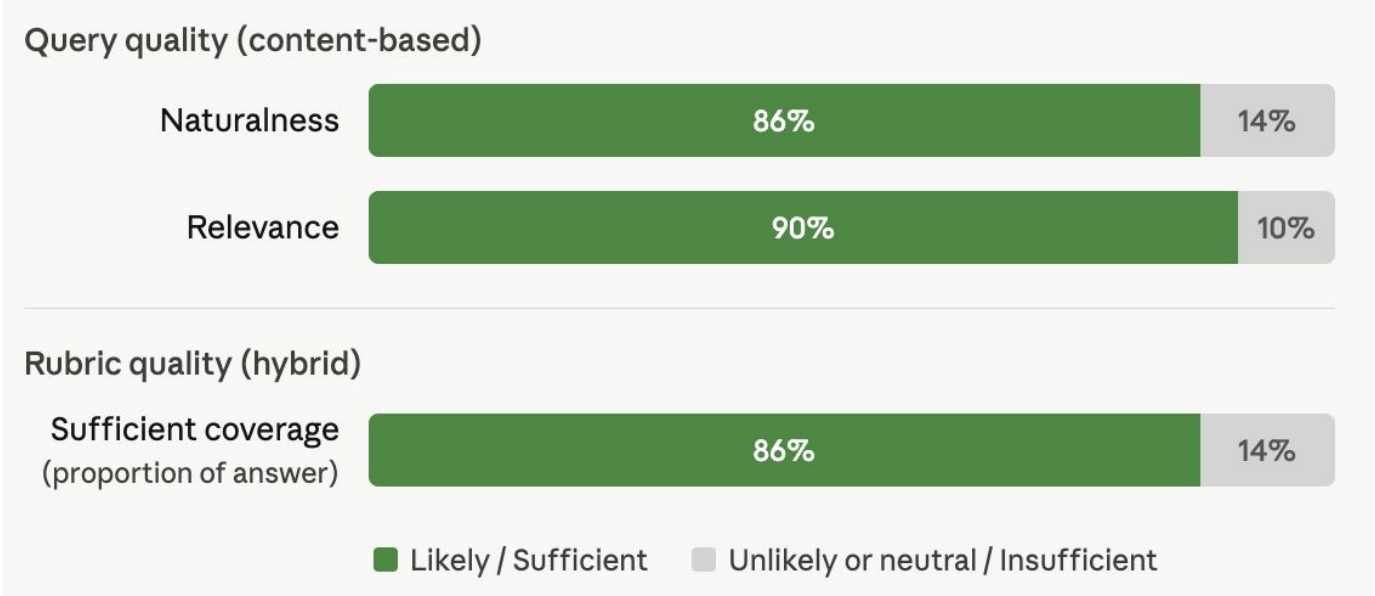
Source Survey: The Mystery of In-Context Learning (Zhou et al., 2024)

ResearchQA: Dataset Quality Validation

Outline

1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. LLM Eval

Expert Validation



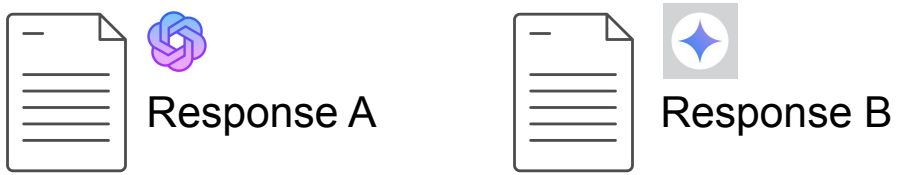
Queries are natural and realistic; Rubric items have sufficient coverage.

Outline

1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. LLM Eval

How well can our rubric-augmented eval protocol measure the llm system outputs?

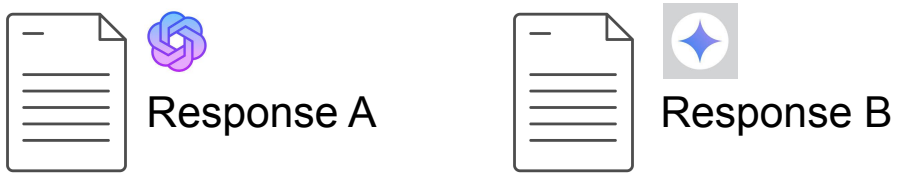
Human Study: Preference Annotation Setup



PhD Researchers
(Agreed Preference)

Given a query Q, two LLM systems generate responses A and B, and the evaluators are required to perform a blind pairwise preference assessment.

Human Study: LLM-Human Misalignment



PhD Researchers
(Agreed Preference)

LLM Direct Judge

The LLM direct judge prefers the other response,
which is **misaligned** with the agreed preference of the human experts.

Human Study: LLM-Human Alignment via Rubric

Response A

Response B

PhD Researchers
(Agreed Preference)

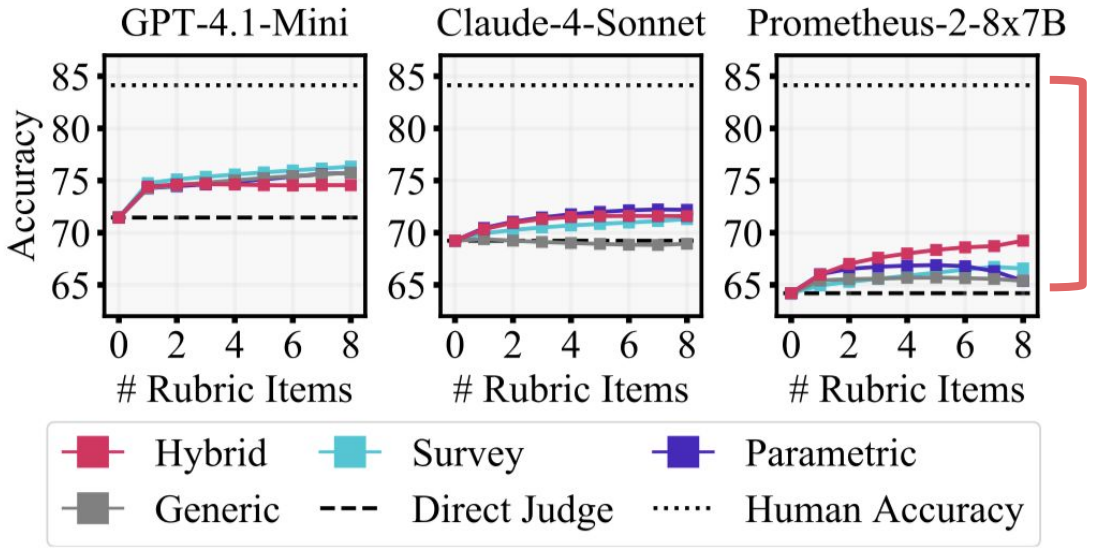
LLM Direct Judge

LLM Ensembled Judge
(with the help of rubric!)

Our rubric can improve the LM judges alignment with human experts!

Human Study: LLM-Human Alignment via Rubric

Rubrics Improve LLM-Human Agreement on Preference Rankings

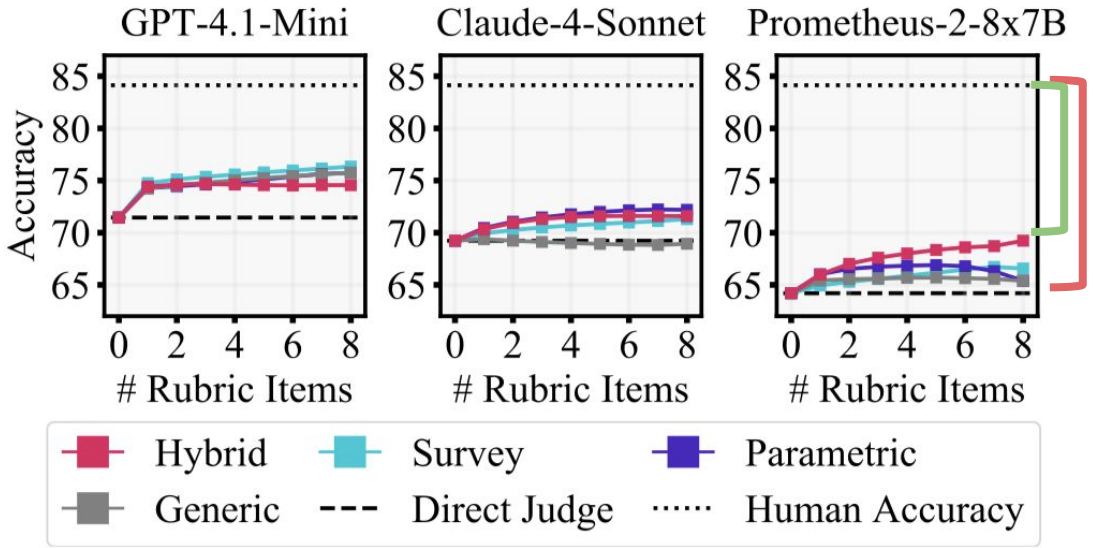


LLM-Human Preference Gap
– original gap

All rubric sets can improve the LM judges alignment with human experts!

Human Study: LLM-Human Alignment via Rubric

Rubrics Improve LLM-Human Agreement on Preference Rankings



LLM-Human Preference Gap

– original gap

+ smaller gap with the help of rubric

Our rubric can improve the LM judges alignment with human experts!
Hybrid rubric is most robust across different settings.

Outline

1. Background
   a. Deep Research Can Help Scientific Discovery
   b. Difficulty of Deep Research Evaluation
2. ResearchQA Dataset
   a. A Query-Rubric Sample (& the Benefits of Rubric)
   b. Comparison: Prior Benchmarks vs. ResearchQA
   c. Data Synthesis Pipeline
3. Validation
   a. Quality of Query and Rubric
   b. Quality of Auto-Eval Protocol
4. **LLM Eval**

ResearchQA: Evaluation and Analysis

LLM Systems Evaluation - Setup

Input a query Q and output answer A with citation support

System	Details
Parametric systems	Parametric memory
Naive retrieval systems	Grounded on top-20 retrieval from google scholar
Production retrieval systems	Advanced engineering (tool use, feedback, or refinement)
Deep research systems	Frontier deep researches (long-horizon)

Result - LLM Systems Evaluation

No performance gap between parametric LLMs and naive retrieval systems.

The production-level retrieval systems can address ResearchQA better, but to excel we need deep research systems.

System	Coverage % ↑							
	All domains							
Parametric								
llama-3.3-70b	53.42 ± 0.26	51.82	54.21	54.89	55.74	53.91	53.22	58.10
claude-4-sonnet	64.31 ± 0.31	62.92	64.96	66.71	67.08	63.33	59.85	66.17
gpt-4.1	65.43 ± 0.25	63.98	66.84	66.66	67.38	65.45	63.48	68.46
qwen-3-32b	66.64 ± 0.26	65.13	67.76	68.25	69.24	65.96	64.32	69.62
gemini-2.5-pro	**68.84 ± 0.25**	**67.42**	**70.20**	**69.82**	**71.86**	**68.28**	**65.83**	**72.06**
Retrieval (Naive)								
openscholar-8b[i]	54.71 ± 0.28	54.08	56.15	54.76	54.98	54.67	52.69	57.46
gemini-2.5-pro[ii]	59.92 ± 0.30	58.73	61.46	61.13	61.72	57.79	56.71	63.96
qwen-3-32b[iii]	60.90 ± 0.32	57.62	62.93	64.25	65.58	60.49	60.23	65.82
claude-4-sonnet[iv]	62.50 ± 0.32	61.94	63.48	62.97	64.01	61.67	58.05	65.74
gpt-4.1[v]	**64.80 ± 0.29**	**63.69**	**66.65**	**65.11**	**66.72**	**64.25**	**62.09**	**66.33**
Retrieval (Production)								
sonar	58.61 ± 0.29	56.61	60.55	59.43	61.62	59.97	57.48	62.80
openscholar-8b+feedback	58.72 ± 0.32	57.77	59.96	58.62	61.48	57.27	57.29	62.48
sonar-reasoning	64.33 ± 0.31	62.73	66.00	65.19	68.11	62.68	61.76	67.49
gpt-4o-search-preview	65.98 ± 0.27	65.52	68.21	65.01	66.60	66.07	62.62	65.63
gemini-2.5-pro+grounding	68.51 ± 0.25	67.38	70.02	**68.76**	**70.99**	68.09	**65.98**	**71.21**
claude-4-sonnet+ws	**69.18 ± 0.26**	**69.54**	**70.49**	67.59	70.28	**68.14**	64.70	67.13
Deep Research								
o4-mini-deep-research†	72.69 ± 0.54	74.02	73.58	70.57	74.04	73.25	68.99	74.54
sonar-deep-research	**75.29 ± 0.26**	**75.01**	**76.31**	**74.48**	**76.77**	**75.34**	**72.47**	**78.01**

Result - LLM Systems Evaluation

No performance gap between parametric LLMs and naive retrieval systems.

The production-level retrieval systems can address ResearchQA better, but to excel we need deep research systems.

System	Coverage % ↑							
	All domains							
Parametric								
llama-3.3-70b	53.42 ± 0.26	51.82	54.21	54.89	55.74	53.91	53.22	58.10
claude-4-sonnet	64.31 ± 0.31	62.92	64.96	66.71	67.08	63.33	59.85	66.17
gpt-4.1	65.43 ± 0.25	63.98	66.84	66.66	67.38	65.45	63.48	68.46
qwen-3-32b	66.64 ± 0.26	65.13	67.76	68.25	69.24	65.96	64.32	69.62
gemini-2.5-pro	**68.84 ± 0.25**	**67.42**	**70.20**	**69.82**	**71.86**	**68.28**	**65.83**	**72.06**
Retrieval (Naive)								
openscholar-8b[i]	54.71 ± 0.28	54.08	56.15	54.76	54.98	54.67	52.69	57.46
gemini-2.5-pro[ii]	59.92 ± 0.30	58.73	61.46	61.13	61.72	57.79	56.71	63.96
qwen-3-32b[iii]	60.90 ± 0.32	57.62	62.93	64.25	65.58	60.49	60.23	65.82
claude-4-sonnet[iv]	62.50 ± 0.32	61.94	63.48	62.97	64.01	61.67	58.05	65.74
gpt-4.1[v]	**64.80 ± 0.29**	**63.69**	**66.65**	**65.11**	**66.72**	**64.25**	**62.09**	**66.33**
Retrieval (Production)								
sonar	58.61 ± 0.29	56.61	60.55	59.43	61.62	59.97	57.48	62.80
openscholar-8b+feedback	58.72 ± 0.32	57.77	59.96	58.62	61.48	57.27	57.29	62.48
sonar-reasoning	64.33 ± 0.31	62.73	66.00	65.19	68.11	62.68	61.76	67.49
gpt-4o-search-preview	65.98 ± 0.27	65.52	68.21	65.01	66.60	66.07	62.62	65.63
gemini-2.5-pro+grounding	68.51 ± 0.25	67.38	70.02	**68.76**	**70.99**	68.09	**65.98**	**71.21**
claude-4-sonnet+ws	**69.18 ± 0.26**	**69.54**	**70.49**	67.59	70.28	**68.14**	64.70	67.13
Deep Research								
o4-mini-deep-research†	72.69 ± 0.54	74.02	73.58	70.57	74.04	73.25	68.99	74.54
sonar-deep-research	**75.29 ± 0.26**	**75.01**	**76.31**	**74.48**	**76.77**	**75.34**	**72.47**	**78.01**

Result - LLM Systems Evaluation

Among fields, deep research are less different among them, retrieval and parametric systems are more different.

System	Coverage % ↑							
	All domains							
Parametric								
llama-3.3-70b	53.42 ± 0.26	51.82	54.21	54.89	55.74	53.91	53.22	58.10
claude-4-sonnet	64.31 ± 0.31	62.92	64.96	66.71	67.08	63.33	59.85	66.17
gpt-4.1	65.43 ± 0.25	63.98	66.84	66.66	67.38	65.45	63.48	68.46
qwen-3-32b	66.64 ± 0.26	65.13	67.76	68.25	69.24	65.96	64.32	69.62
gemini-2.5-pro	**68.84 ± 0.25**	**67.42**	**70.20**	**69.82**	**71.86**	**68.28**	**65.83**	**72.06**
Retrieval (Naive)								
openscholar-8b[i]	54.71 ± 0.28	54.08	56.15	54.76	54.98	54.67	52.69	57.46
gemini-2.5-pro[ii]	59.92 ± 0.30	58.73	61.46	61.13	61.72	57.79	56.71	63.96
qwen-3-32b[iii]	60.90 ± 0.32	57.62	62.93	64.25	65.58	60.49	60.23	65.82
claude-4-sonnet[iv]	62.50 ± 0.32	61.94	63.48	62.97	64.01	61.67	58.05	65.74
gpt-4.1[v]	**64.80 ± 0.29**	**63.69**	**66.65**	**65.11**	**66.72**	**64.25**	**62.09**	**66.33**
Retrieval (Production)								
sonar	58.61 ± 0.29	56.61	60.55	59.43	61.62	59.97	57.48	62.80
openscholar-8b+feedback	58.72 ± 0.32	57.77	59.96	58.62	61.48	57.27	57.29	62.48
sonar-reasoning	64.33 ± 0.31	62.73	66.00	65.19	68.11	62.68	61.76	67.49
gpt-4o-search-preview	65.98 ± 0.27	65.52	68.21	65.01	66.60	66.07	62.62	65.63
gemini-2.5-pro+grounding	68.51 ± 0.25	67.38	70.02	**68.76**	**70.99**	68.09	**65.98**	**71.21**
claude-4-sonnet+ws	**69.18 ± 0.26**	**69.54**	**70.49**	67.59	70.28	**68.14**	64.70	67.13
Deep Research								
o4-mini-deep-research†	72.69 ± 0.54	74.02	73.58	70.57	74.04	73.25	68.99	74.54
sonar-deep-research	**75.29 ± 0.26**	**75.01**	**76.31**	**74.48**	**76.77**	**75.34**	**72.47**	**78.01**

Result - Rubric Type

Item Type	Description	Example	Frequency %	Error %
Citation	X is cited	Does the response cite Kvaskoff et al. (2015) (title: [...]) that links endometriosis with elevated cardiovascular risk?	8.3	89.3
Limitation	Limitations of X are mentioned	Does the response address limitations of CTC in detecting small polyps and flat adenomas?	2.7	52.4
Comparison	X and Y are compared	Does the response compare reaction rates before and after catalyst saturation occurs?	14.2	51.9
Example	Examples of X are mentioned	Does the response include forage species (e.g., legumes, chicory) affecting lamb meat's fatty acid profile?	11.2	46.8
Impact	Cause or impact of X is mentioned	Does the response mention the preservation of the anterior cruciate ligament (ACL) as a benefit of UKA?	15.5	46.3
Other	None of the above	Does the response discuss METEOR's fragmentation penalty and its role in evaluating word order?	48.1	43.6

Table 5: Error rates for different rubric types, measured as the percentage of items not rated as *Completely* covered by the best-performing system (`sonar-deep-research`). Each rubric type provides a description (with X representing a concept, method, or paper being evaluated) along with an illustrated example.

Takeaways

- To address the limited, unscalable access to expert supervision, survey papers can serve as a scalable source for distilling expert supervision

Takeaways

- To address the limited, unscalable access to expert supervision, survey papers can serve as a scalable source for distilling expert supervision
- Rubric can help to
  - Evaluate more than one correct long-form answers
  - Align the LLM judge with human preference

Takeaways

- To address the limited, unscalable access to expert supervision, survey papers can serve as a scalable source for distilling expert supervision
- Rubric can help to
  - Evaluate more than one correct long-form answers
  - Align the LLM judge with human preference
- We need expert-level long-form QA such as ResearchQA
  - To evaluate deep research
  - To reflect the needs of deep research

Impacts as Benchmark (ResearchQA has been used by)

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research

Rulin Shao[1♡†], **Akari Asai**[2,3♡†], **Shannon Zejiang Shen**[4♡†], **Hamish Ivison**[1,2♡†], **Varsha Kishore**[1,2†], **Jingming Zhuo**[1†], **Xinran Zhao**[3], **Molly Park**[1], **Samuel G. Finlayson**[1,5], **David Sontag**[4], **Tyler Murray**[2], **Sewon Min**[2,6], **Pradeep Dasigi**[2], **Luca Soldaini**[2], **Faeze Brahman**[2], **Wen-tau Yih**[1], **Tongshuang Wu**[3], **Luke Zettlemoyer**[1], **Yoon Kim**[4], **Hannaneh Hajishirzi**[1,2], **Pang Wei Koh**[1,2]

[1]University of Washington, [2]Allen Institute for AI, [3]Carnegie Mellon University
[4]Massachusetts Institute of Technology, [5]Seattle Children's Hospital, [6]University of California, Berkeley

[♡]Joint first authors, [†]Core contributors. See full author contributions here.

salesforce research

Enterprise Deep Research: Steerable Multi-Agent Deep Research for Enterprise Analytics

Akshara Prabhakar, Roshan Ram, Zixiang Chen, Silvio Savarese, Frank Wang*, Caiming Xiong, Huan Wang, Weiran Yao

Salesforce AI Research

Other Works - Starling

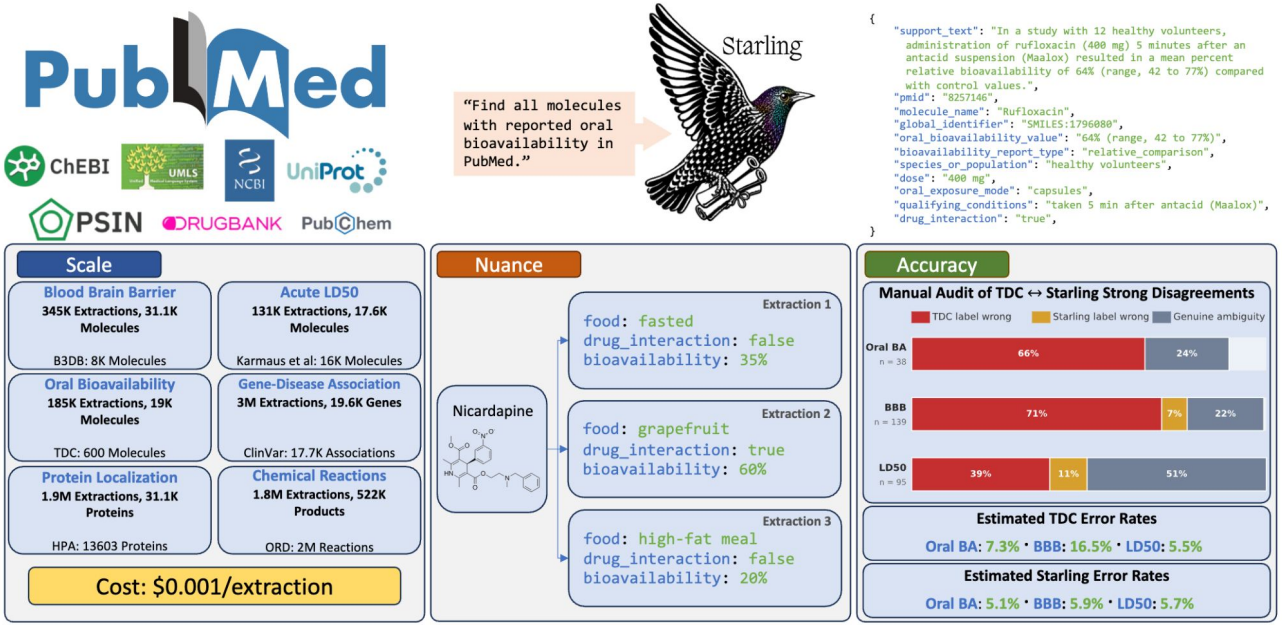


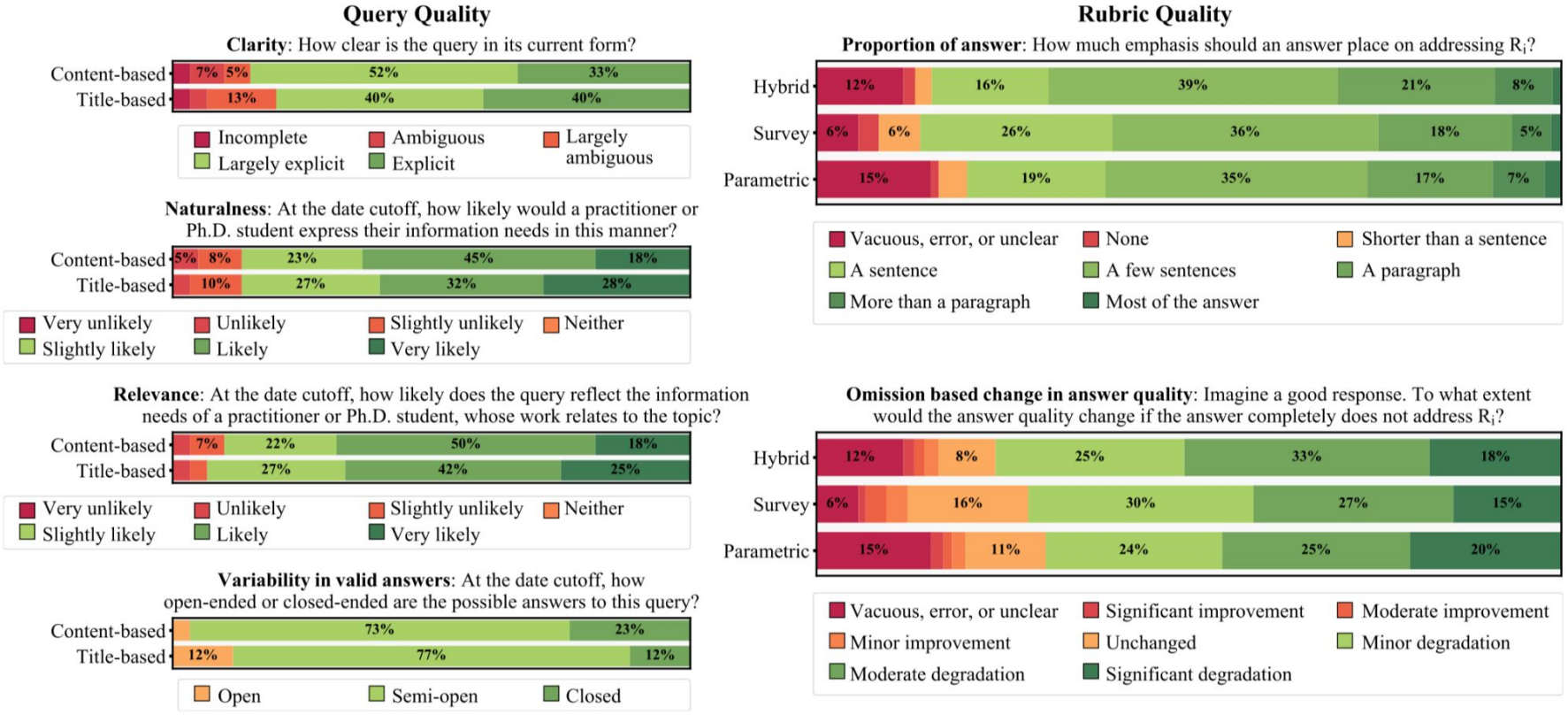
Figure 1: `Starling` is a promptable agentic system that automatically creates datasets from biomedical literature. `Starling` is fully self-driving, requiring no human intervention, and produces datasets at comparable or larger scale than manual resources, capturing nuanced context often missing in existing datasets, with similar or better accuracy. `Starling` can keep datasets better synchronized to quickly expanding information in literature, at under 1 cent per datum it extracts.

Thank You!
Q&A

Project Website
https://researchqa.cylumn.com/

Additional Materials

Expert Validation: Query and Rubric Items



LLM Systems Evaluation - Results (Leaderboard score version)

System	Coverage % ↑								Leaderboard Score ↑	Avg Length (Words)
	All domains									
Parametric										
llama-3.3-70b	53.42 ± 0.26	51.82	54.21	54.89	55.74	53.91	53.22	58.10	617 ± 13	167.4
claude-4-sonnet	64.31 ± 0.31	62.92	64.96	66.71	67.08	63.33	59.85	66.17	1099 ± 09	226.5
gpt-4.1	65.43 ± 0.25	63.98	66.84	66.66	67.38	65.45	63.48	68.46	1080 ± 09	241.5
qwen-3-32b	66.64 ± 0.26	65.13	67.76	68.25	69.24	65.96	64.32	69.62	1038 ± 09	219.3
gemini-2.5-pro	**68.84 ± 0.25**	**67.42**	**70.20**	**69.82**	**71.86**	**68.28**	**65.83**	**72.06**	**1244 ± 10**	267.1
Retrieval (Naive)										
openscholar-8b[i]	54.71 ± 0.28	54.08	56.15	54.76	54.98	54.67	52.69	57.46	478 ± 17	499.9
gemini-2.5-pro[ii]	59.92 ± 0.30	58.73	61.46	61.13	61.72	57.79	56.71	63.96	945 ± 10	270.4
qwen-3-32b[iii]	60.90 ± 0.32	57.62	62.93	64.25	65.58	60.49	60.23	65.82	1011 ± 09	265.5
claude-4-sonnet[iv]	62.50 ± 0.32	61.94	63.48	62.97	64.01	61.67	58.05	65.74	972 ± 09	238.4
gpt-4.1[v]	**64.80 ± 0.29**	**63.69**	**66.65**	**65.11**	**66.72**	**64.25**	**62.09**	**66.33**	**1020 ± 09**	263.6
Retrieval (Production)										
sonar	58.61 ± 0.29	56.61	60.55	59.43	61.62	59.97	57.48	62.80	862 ± 10	242.2
openscholar-8b+feedback	58.72 ± 0.32	57.77	59.96	58.62	61.48	57.27	57.29	62.48	769 ± 12	788.8
sonar-reasoning	64.33 ± 0.31	62.73	66.00	65.19	68.11	62.68	61.76	67.49	1115 ± 10	280.5
gpt-4o-search-preview	65.98 ± 0.27	65.52	68.21	65.01	66.60	66.07	62.62	65.63	992 ± 09	255.0
gemini-2.5-pro+grounding	68.51 ± 0.25	67.38	70.02	**68.76**	**70.99**	68.09	**65.98**	**71.21**	960 ± 09	278.5
claude-4-sonnet+ws	**69.18 ± 0.26**	**69.54**	**70.49**	67.59	70.28	**68.14**	64.70	67.13	**1149 ± 10**	327.8
Deep Research										
o4-mini-deep-research†	72.69 ± 0.54	74.02	73.58	70.57	74.04	73.25	68.99	74.54	1145 ± 10	271.6
sonar-deep-research	**75.29 ± 0.26**	**75.01**	**76.31**	**74.48**	**76.77**	**75.34**	**72.47**	**78.01**	**1505 ± 17**	267.3

Analysis - Length Bias

Systems	Avg Length			Coverage %		
	L=250	$\neg L$	Δ%	L=250	$\neg L$	Δ
Parametric						
`qwen-3-32b`	219.5	176.5	−19%	66.8	64.0	−2.8
`gemini-2.5-pro`	268.5	241.4	−10%	69.9	67.5	−2.5
Retrieval (Naive)						
`claude-4-sonnet`	238.9	299.5	+25%	63.6	62.8	−0.8
`gpt-4.1`	264.0	246.1	−6%	65.0	63.7	−1.3
Deep Research						
`o4-mini-deep-research`	268.0	595.0	+122%	72.2	78.9	+6.7
`sonar-deep-research`	267.2	1431.0	+435%	76.2	85.3	+9.1

Table 4: Removing the L=250 words length guidance prompt can affect average answer length (words). Longer answers tend to score higher coverage, because coverage is a recall-based measure. Analysis is performed on a 225 query subset of $\mathcal{D}_{\text{test}}$ (3 per field).

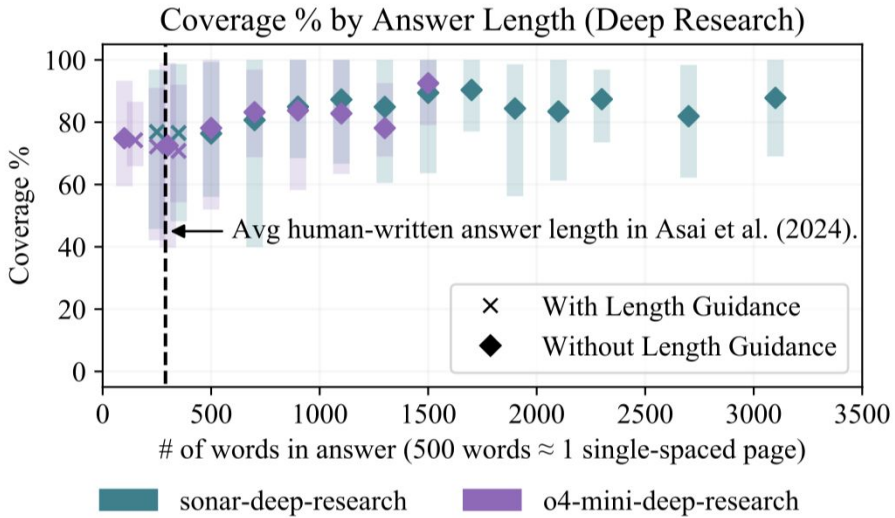


Figure 6: Coverage % increases with answer length, up until ∼2K words (about 4 pages of text).

Analysis - Temporal Bias

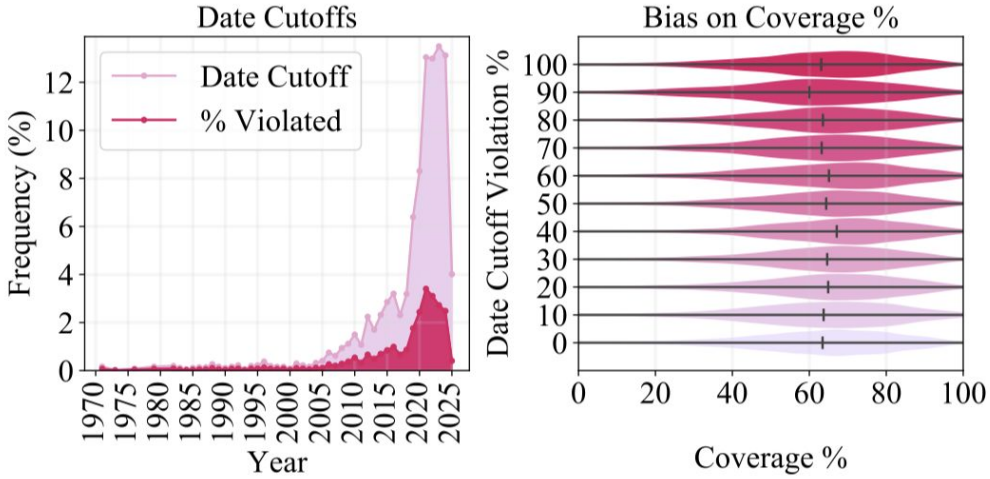


Figure 5: Rubrics are recent, mostly originating from surveys in the past decade. When cited sources violate date cutoff years, there is little bias on Coverage %.

Analysis - Survey Leakage

System	L%	Coverage %		
		¬ Leaked	Leaked	Δ
claude-4-sonnet+ws	30	**71.27**	68.30	−3.0
gemini-2.5-pro+grounding	2	66.91	**68.54**	+1.6
o4-mini-deep-research	28	**74.82**	71.85	−3.0
sonar-deep-research	21	74.46	**75.51**	+1.1

Table 6: Production systems often cite the related survey as a source, at 20-30% leakage (L%). However, Coverage % roughly stays the same with leakage (Leaked) and without leakage (¬ Leaked).