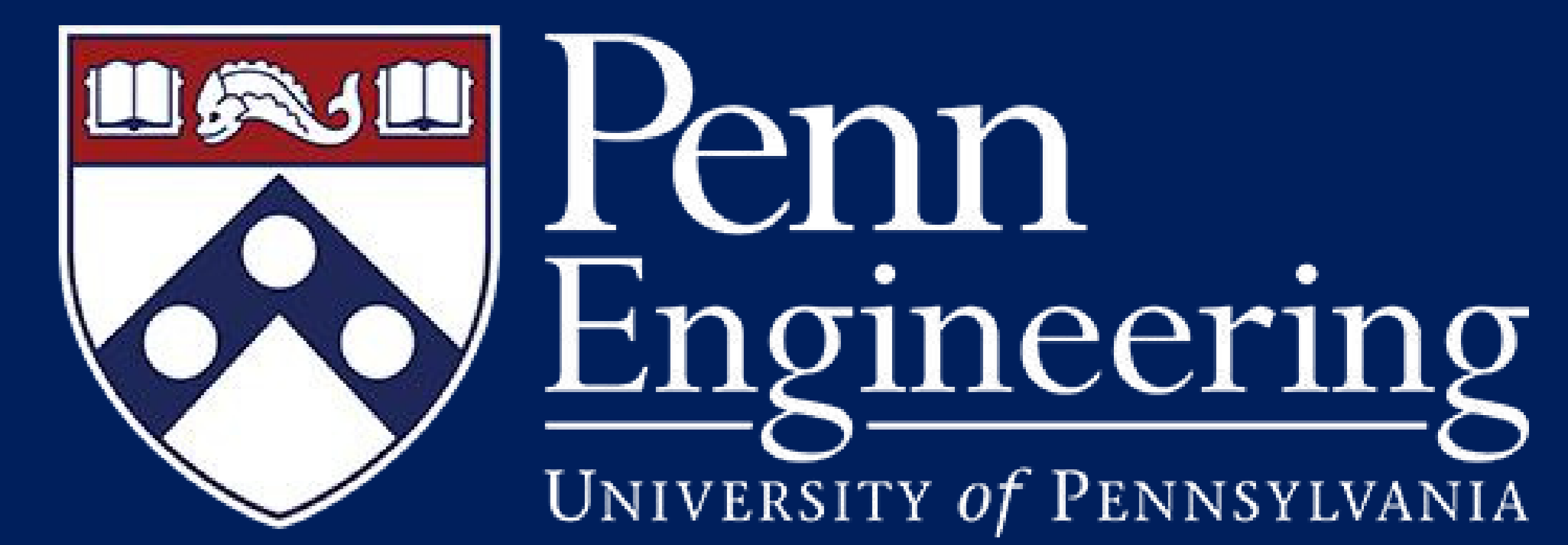


RESEARCHQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics

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



Why Evaluating Deep Research is Hard

Challenge 1: Many Valid Answers

More than one long-form answer can be correct, so a single gold answer doesn't exist.

Challenge 2: Expert Supervision Doesn't Scale

Domain experts are reliable judges, but enlisting them is costly and limits evaluations to fields like AI.

Cost of poorly scaling expert annotators:

An estimated of **~30%** of **Humanity's Last Exam** chemistry/biology answers are likely incorrect. These errors are traced to insufficient time for experts to annotate and validate answers.

Scalable Method: RESEARCHQA, which distills from Survey Articles

Solution 1: Rubric Scoring via Coverage

Score answers by **coverage** of explicit criteria instead of similarity to one gold answer.

Solution 2: Distilling from Survey Articles

Survey articles are a scalable source of expert supervision that consolidate knowledge across the field.

21K Queries **160K Rubric Items** **75 Fields**

Mined from 54K surveys across 7 Google Scholar domains, making RESEARCHQA the most diverse benchmark of its kind.

Resolve C1 & C2 via academic surveys

Survey-Mined Query

How does the frequency of terms in pre-training data influence numerical reasoning performance in few-shot settings? (Engineering)

Research System: The frequency of terms in pre-training data significantly influences a model's numerical reasoning performance, particularly in few-shot learning scenarios [1]. Models pre-trained [...]

[1] Scaling Laws and Data Frequency Effects in Large Language [...]

Survey-Mined Evaluation Rubric

Judge

Does the response reference the **"performance gap"** concept from the **Razeghi et al. (2022)** paper [...]? **0/4** Not at all covered

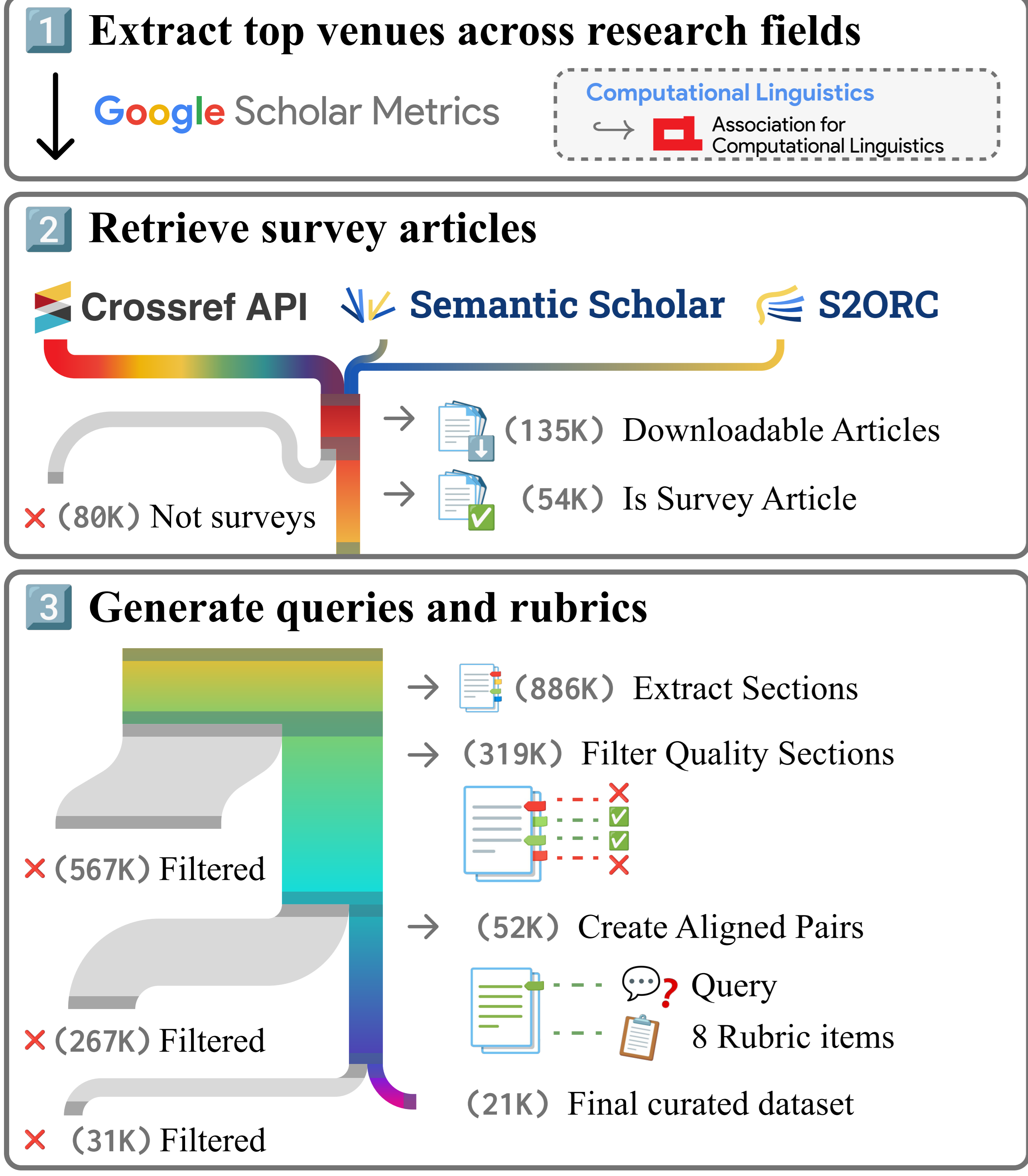
Does the response include **examples of studies or experiments** that investigate the impact of term frequency on numerical reasoning performance? **4/4** Completely covered

Does the response discuss the **correlation between the frequency of terms** in pre-training data and numerical reasoning performance? **1/4** Barely covered

Additional rubric items ...

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

Data Pipeline & Expert Validation



Query Naturalness. At the date cutoff, how likely would a practitioner or Ph.D. student express their information needs in this manner?

86% of responses were **"Slightly Likely"**, **"Likely"**, or **"Very Likely"**

Query Relevance. At the date cutoff, how likely does the query reflect the information needs of a practitioner or Ph.D. student?

90% of responses were **"Slightly Likely"**, **"Likely"**, or **"Very Likely"**

Rubric Quality, measured by the amount of detail required to answer. How much emphasis should an answer place on addressing the rubric item?

(Survey) **86%** of responses were **"A sentence"** to **"Most of the answer"**

(Parametric) **79%** of responses were **"A sentence"** to **"Most of ..."**

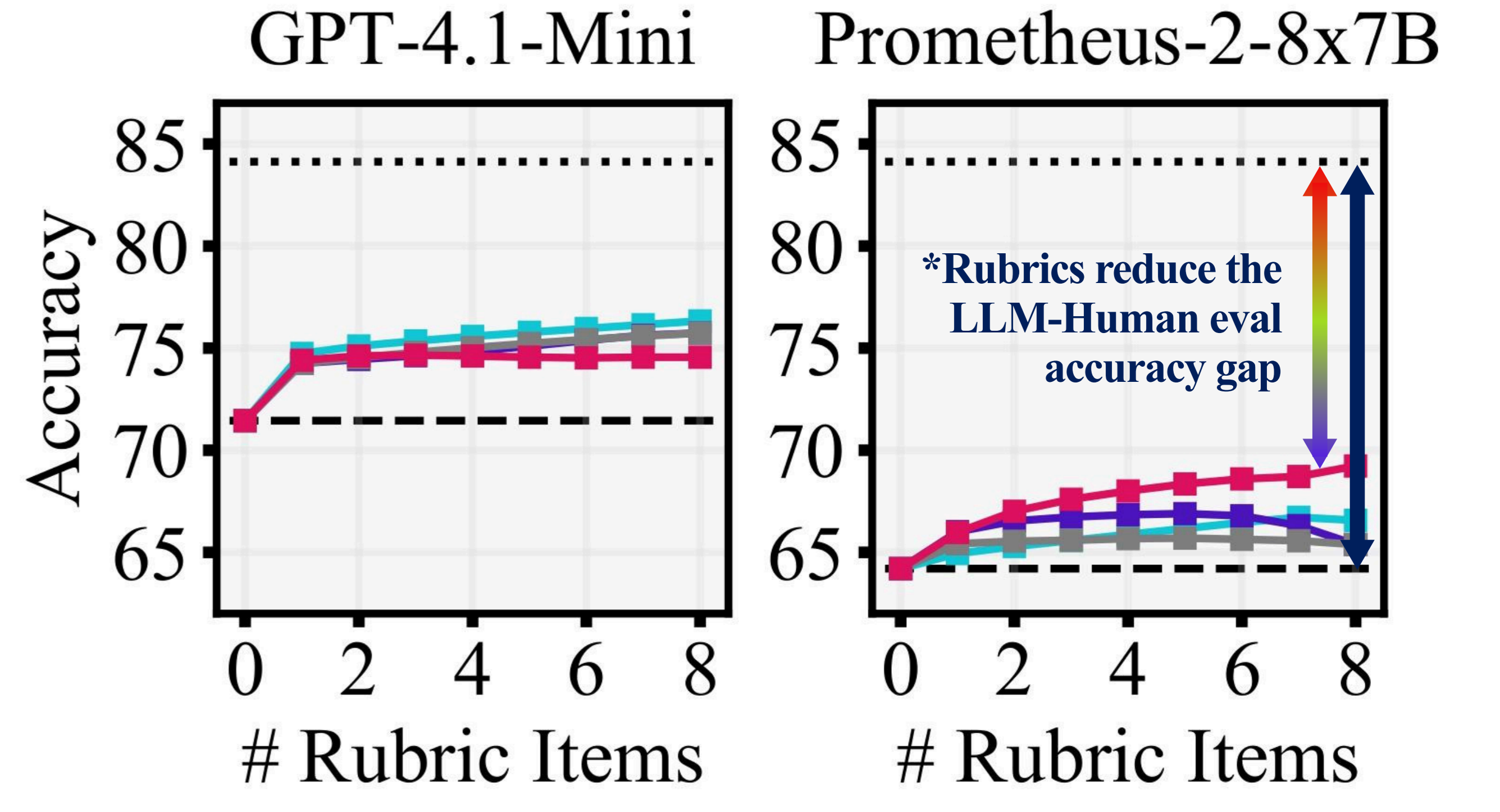
(Hybrid) **84%** of responses were **"A sentence"** to **"Most of the answer"**

We explored three knowledge sources for generating rubrics:

- Survey:** Rubric items are extracted from sections of survey articles. These items are typically higher in specificity (e.g., a detailed claim).
- Parametric:** Rubric items are generated solely from querying an LLM. These items are more general (e.g., mentioning a concept).
- Hybrid:** Combines the specificity of survey rubric items with the generality of parametric rubric items. We additionally apply reranking and deduplication procedures to retain the most meaningful rubric items.

Rubrics for Pairwise Judgements

Rubrics Improve LLM-Human Agreement on Pairwise Preference Rankings



Legend: Hybrid (red), Survey (blue), Parametric (green), Generic (grey), Direct Judge (dashed), Human Accuracy (dotted)

Benchmarking LLM Systems

LLM System	Coverage % ↑							
	All domains	AI	Chem	Eng	Math	Med	Phy	Psych
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	54.71 ± 0.28	54.08	56.15	54.76	54.98	54.67	52.69	57.46
gemini-2.5-pro	59.92 ± 0.30	58.73	61.46	61.13	61.72	57.79	56.71	63.96
qwen-3-32b	60.90 ± 0.32	57.62	62.93	64.25	65.58	60.49	60.23	65.82
claude-4-sonnet	62.50 ± 0.32	61.94	63.48	62.97	64.01	61.67	58.05	65.74
gpt-4.1	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	71.21	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

Conclusions

Four Findings of Benchmarking

All systems struggle. No parametric or retrieval system we evaluate exceeds 70% coverage. The highest-performing system we evaluate is Perplexity's deep research, with 75.29% coverage, indicating room for improvement.

Models selectively benefit from retrieval. When only relying on parametric memory, Claude Sonnet-4 ranks lower than Gemini-2.5-Pro (30% win rate), but when both models use retrieval Claude Sonnet-4 ranks higher (75% win rate).

Retrieval optimization is critical. We observe that a naive retrieval implementation is often insufficient to improve coverage. On average, the same models differ by 7.6% coverage between naive and production retrieval.

Deep research is far ahead. Perplexity's deep research obtains 82% win rate over the next best system, revealing substantial gaps between the helpfulness of research-oriented tools.

Takeaways

Survey articles are a scalable source for distilling expert supervision. **Rubrics** can help to:

- Evaluate more than one correct long-form answer
- Align LLM-as-a-Judge with human preferences

We need **expert-level long-form QA** such as RESEARCHQA to evaluate and reveal the needs of deep research systems.



Website