SurveyGen: Quality-Aware Scientific Survey Generation with Large Language Models

Tong Bao[1], Mir Tafseer Nayeem[2], Davood Rafiei[2], Chengzhi Zhang[1]

1.Nanjing University of Science and Technology
2.University of Alberta

Why survey generation?

- The rapid growth of publications causes information overload, making it hard for us to keep up with through daily reading
- Survey articles help researchers summarize related work and highlight future directions in the field
- Writing a survey is a complex task (e.g., summarizing 100+ relevant papers)
- LLMs open the door for this task due to their powerful understanding and generation capabilities.

Related work

1. Retrieval → 2. Generation → 3. Evaluation

1. Semantic similarity retrieval from database (abstracts *VS.* survey topic)
2. Ranking paper based on similarity/relevance to get the final candidates
3. Generate the survey outline first, and then drafting the survey content
4. Human and LLM-as-judge for survey evaluation

[1] Autosurvey: Large language models can automatically write surveys. *NeurIPS* 2025.
[2] SurveyX: Academic survey automation via large language models. arXiv 2025.
[3] Are llms good literature review writers? evaluating the literature review writing ability of large language models. arXiv 2025.

Gaps

1. Semantic similarity-based retrieval suffers from recall issues

e.g., the Word2Vec may get a low semantic similarity score with the topic Deep learning, but it is cited by many deep-learning related papers and should be take consideration.

2. Semantic similarity ranking fails to capture the quality or impact of the retrieved papers.

e.g., the abstract of a best paper award winner may receive a similar score as a regular paper

3. Missing benchmark of human-written surveys hinders comparison with gold standards.

e.g., What are the differences between LLM-generated surveys and human-written surveys?

Our contributions

1. Introduce SurveyGen, a large-scale dataset comprising over 4,200 human-written surveys from multi domains.

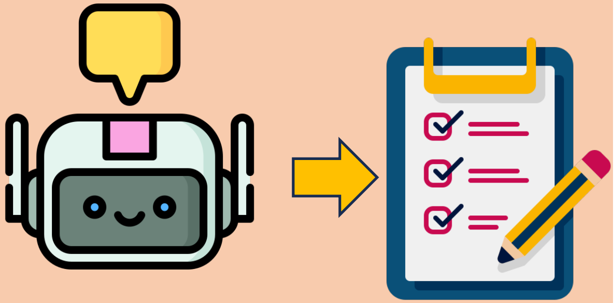
2. Propose QUAL-SG, a novel framework that extends Naive-RAG by adding academic quality evaluation into the survey generation pipeline.

3. QUAL-SG significantly improves citation quality, content relevance, and structural consistency in survey generation

SurveyGen: Task Design

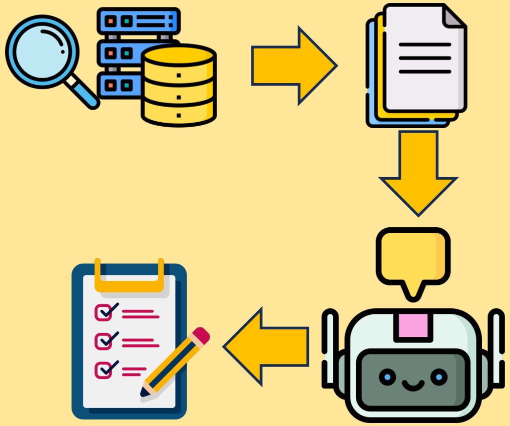
Humans may engage LLMs at different stages during survey generation, we design ***three*** tasks for comprehensive evaluation:

1.Fully LLM-based



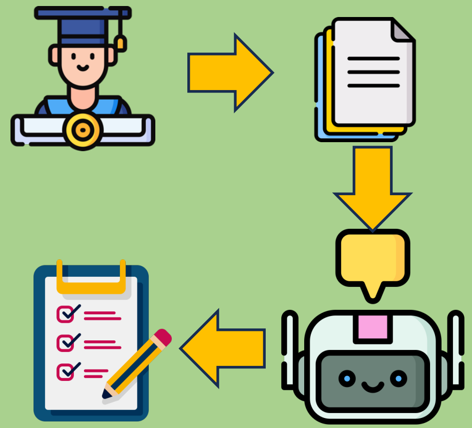
Input:
topic

2.RAG-based



Input:
topic+ RAG

3.Human-guided



Input:
topic+ outline+reference

SurveyGen: Data Collection

1. Find articles that titles contain “*a survey* ”, ”*a review* ”, ”*survey of* ”, etc., from Semantic scholar Open Research Corpus(S2ORC) from 2010-2024

2. LLMs for survey-type paper classification based on the title and abstract

3. Full-text accessible, citation count >30, top-level section headers >3

4. Parse the section divisions and map each reference to the corresponding sections based on its in-text citation locations

SurveyGen: Data Supplement

1. Basic metadata: Supplement metadata (e.g., abstract, DOI, research fields) for all involved papers from S2ORC

2. Quality-related data: Retrieve citation performance, author influence, and venue reputation from OpenAlex database via DOIs

3. Metadata for second-level references: Enriched the metadata for a total of 5.06M references cited by the papers referenced in all surveys

4,200+ surveys, 115,000+ sections, 240,000+ references

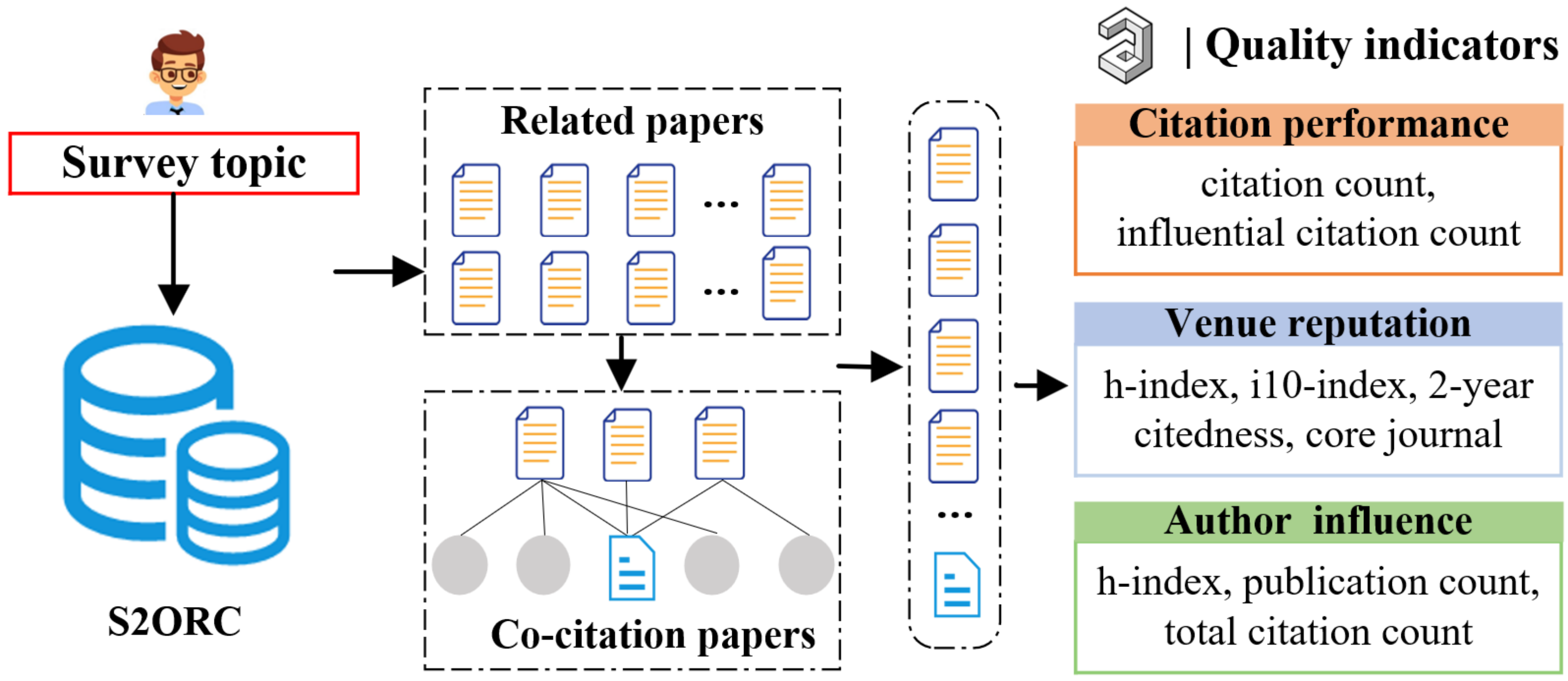
QUAL-SG: Quality-aware Survey Generation

Semantic retrieval

Co-cited papers expansion

Re-ranking

Generation



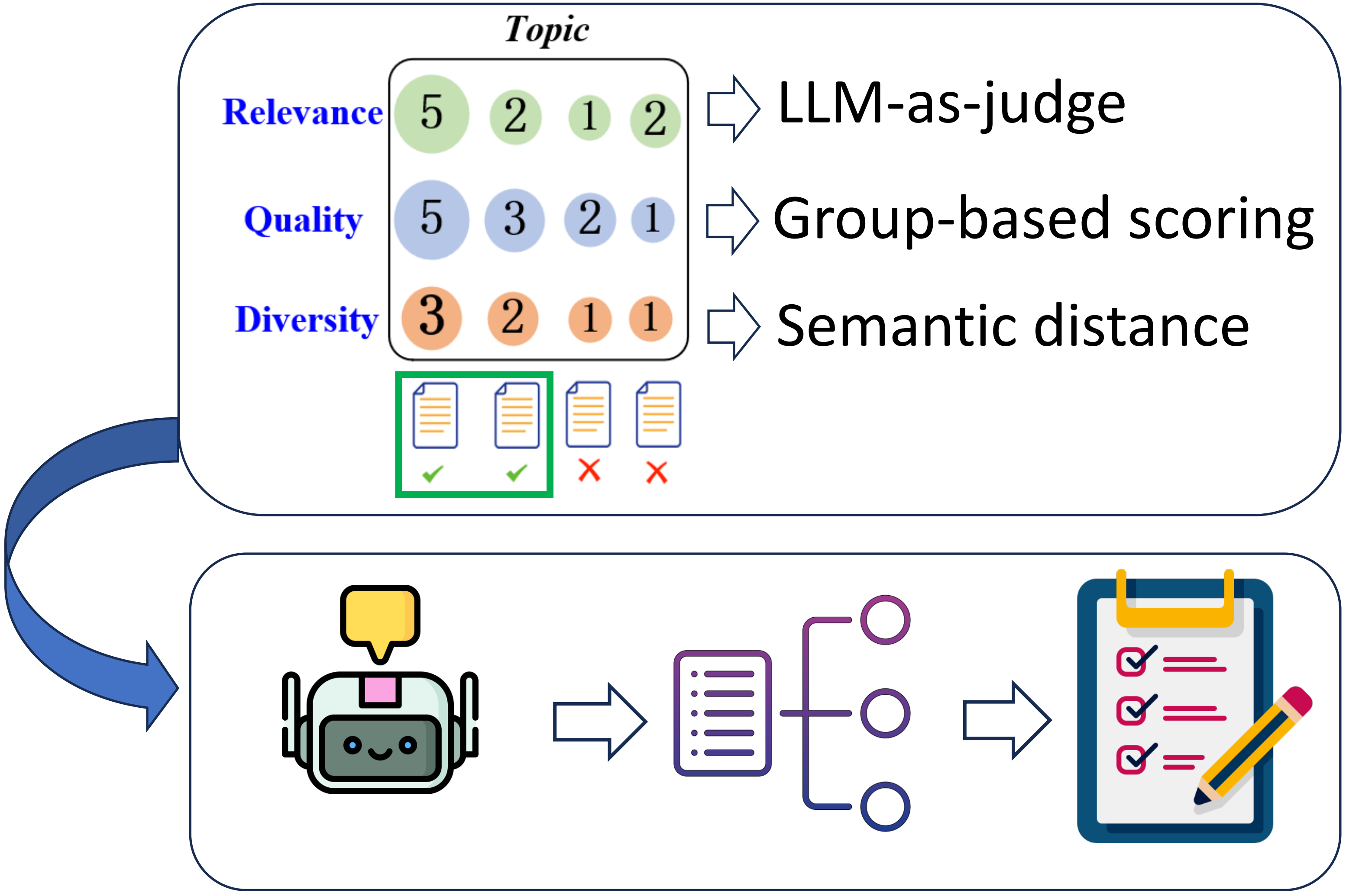
QUAL-SG: Quality-aware Survey Generation

Semantic retrieval

Co-cited papers expansion

Re-ranking

Generation

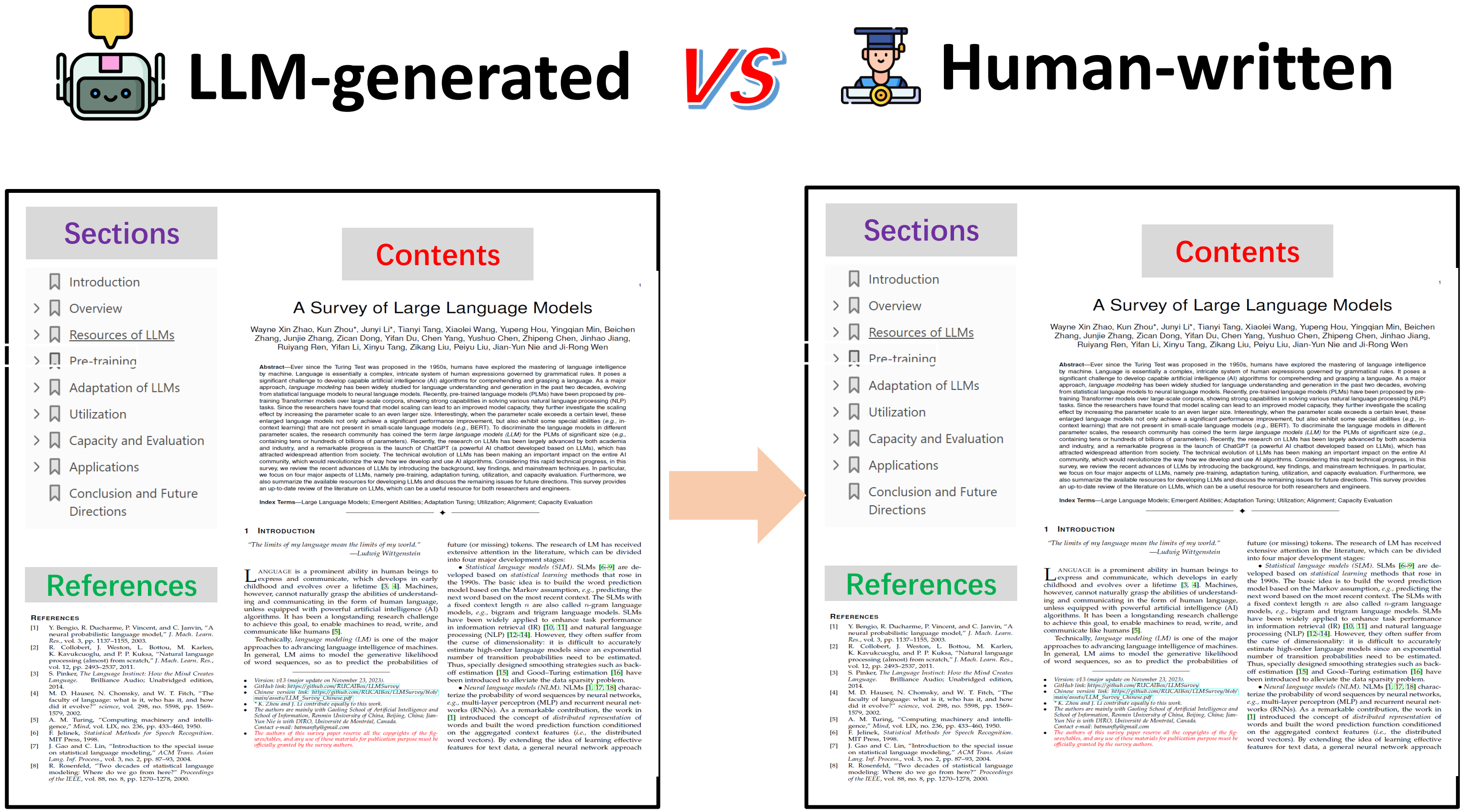


Evaluation: Automatic Evaluation

1.Citation quality
- ✓ Acc.(hallucinated)
- ✓ P, R, F1 (human)

2.Content quality
- ✓ Semantic similarity.
- ✓ Rouge-L
- ✓ Key Point Recall

3.Structural consistency
- ✓ Section overlap (%)
- ✓ Overall relevance



Evaluation: Human Evaluation

Criteria

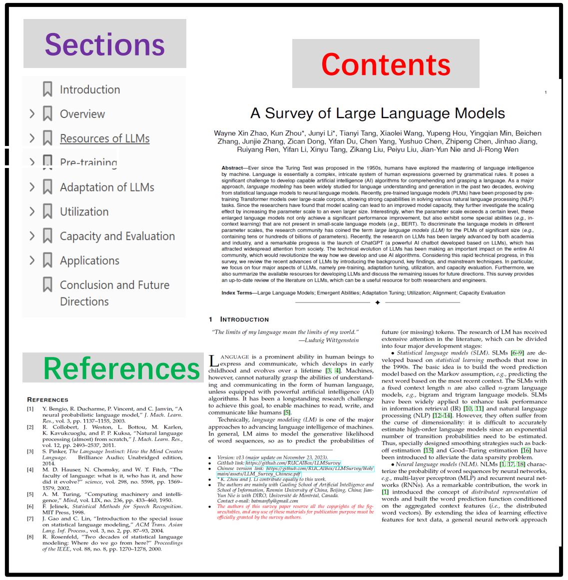
Topic Relevance: whether the survey maintains a clear focus on the assigned topic?

Information Courage: whether the survey includes key papers, major developments, and diverse research approaches relevant to the topic?

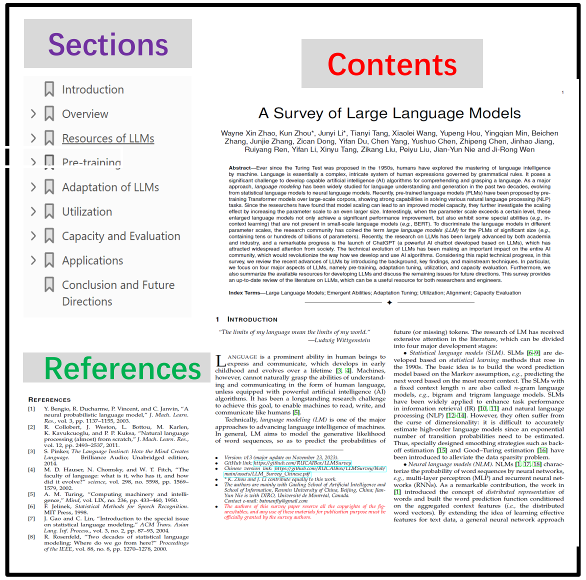
Critical Analysis: whether the survey compares methods or findings, identifies limitations or open challenges, and offers insight rather than descriptive summaries?

Overall Rating: whether the survey is well-written, logically structured, and academically appropriate, and would be considered the better survey in comparison?

Which one is better, comparable, or worse?



VS



Results: Task1→Fully LLM-based

- LLMs is not reliable at reference generation (Acc.35.84%)
- Good similarity but lower KPR
- Closed-source LLMs show better structural consistency
- Open-soured deliver comparable results in content generation

Model	Citation Quality				Content Quality			Structural Consistency	
	Acc. ↑	P ↑	R ↑	F1 ↑	Sim. ↑	R-L ↑	KPR ↑	$\mathrm{Rel}_{\mathrm{LLM}}$	Overlap (%)
Open-source LLMs									
GLM-4-Flash	9.27	9.03	3.26	4.79	81.27	15.04	41.71	2.44	10.62
LLaMA-3.1-70B	15.43	11.48	2.74	4.42	**82.43**	**15.36**	44.36	2.62	13.48
DeepSeek-V3	33.63	10.85	4.09	5.94	82.05	14.18	43.53	2.57	11.03
Closed-source LLMs									
GPT-4.1	21.07	**12.31**	3.72	5.71	79.51	13.48	39.21	2.39	10.95
Gemini-2.0-Flash	22.20	8.97	3.59	5.13	80.20	14.65	42.67	2.50	12.39
Claude-3.7-Sonnet	**35.84**	11.79	**5.78**	**7.76**	81.32	13.77	**46.59**	**2.65**	**14.89**

Table 2: Performance comparison of different LLMs on Task 1. "Acc" indicates whether the generated references are factually accurate and correspond to real papers. "Sim", "R-L", and "KPR" represent "Semantic similarity", "Rouge-L", and "Key Point Recall", respectively. "$\mathrm{Rel}_{\mathrm{LLM}}$" represents structural consistency in LLM evaluations. The best results are marked **bold** and the second-best are underlined.

Results: Task2→RAG-based

- QUAL-SG outperforms baselines
- Directly and significantly improves citation quality
- Also achieves notable gains in content quality and structural consistency

Model	Citation Quality			Content Quality			Structural Consistency	
	P ↑	R ↑	F1 ↑	Sim. ↑	R-L ↑	KPR ↑	$\mathrm{Rel}_{\mathrm{LLM}}$	Overlap (%)
Fully-LLMGen	11.79	5.78	7.76	81.32	13.77	46.59	2.65	14.89
Naive-RAG	5.18	6.94	5.93	82.37	12.90	42.17	2.43	12.22
QUAL-SG (Ours)	**15.87**†	**17.71**†	**16.73**†	**83.10**†	**15.17**†	**50.25**†	**2.81**†	**24.76**†

Table 3: Performance of different models on Task 2. For Fully-LLMGen (Tang et al., 2025), we directly report the results from Task 1. In the Naive-RAG setting (Wu et al., 2025), retrieval is based on the semantic similarity between the survey topic and candidate abstracts. Claude-3.7-Sonnet is used as the backbone for all methods. The best results are marked **bold**. † denotes significant differences to baselines (p-value < 0.001).

Results: Task3→Human-guided

- Compared to Task 1 and Task 2, human-guided method achieve best content quality
- Closed-source LLMs are a cost-effective option at content generation
- Even with perfect references and an outline, there remains a gap compared to humans

Model	Sim. ↑	R-L ↑	KPR ↑
Open-source LLMs			
GLM-4-Flash	82.04	<u>16.29</u>	46.88
LLaMA-3.1-70B	**84.39**	**17.16**	<u>52.13</u>
DeepSeek-V3	<u>83.97</u>	15.25	49.50
Closed-source LLMs			
GPT-4.1	82.59	13.82	50.02
Gemini-2.0-Flash	83.74	15.62	51.76
Claude-3.7-Sonnet	84.22	15.43	**54.67**

Table 4: Content quality evaluation results of different LLMs on Task 3. The best results are marked **bold** and the second-best are <u>underlined</u>.

Results: Reference Selections Analysis

- QUAL-SG show the best alignment with human-written survey
- Fully-LLMGen show a pronounced long-tail distribution
- Poor performance of Naive-RAG highlights the limitation of purely semantic retrieval

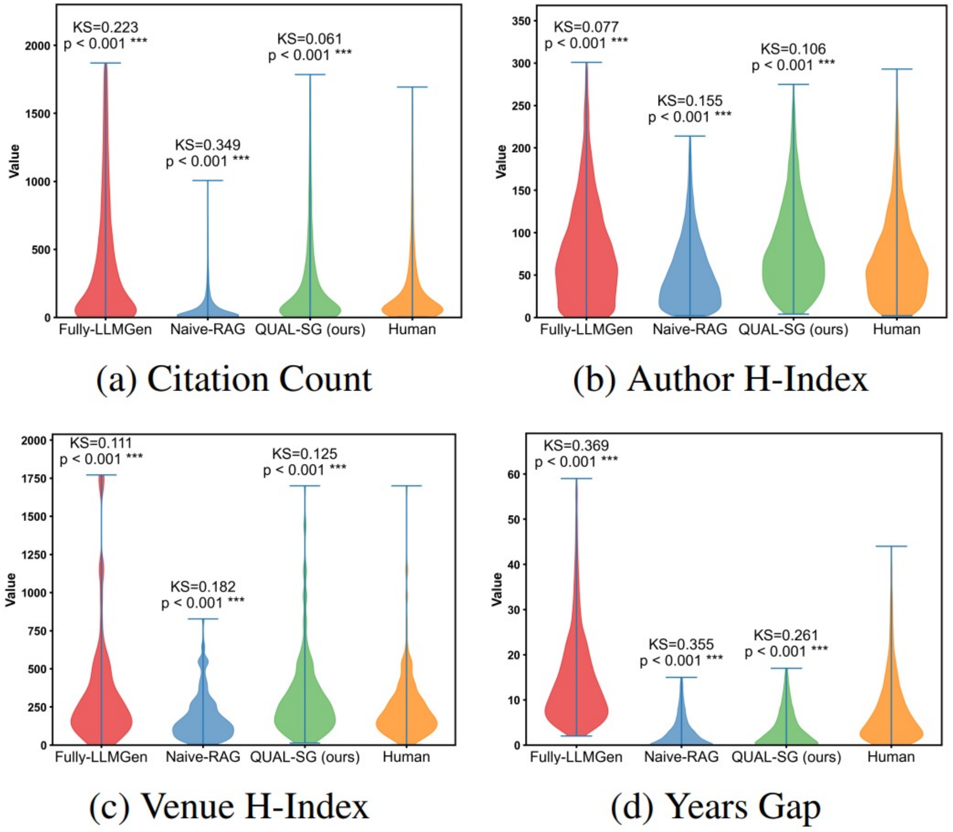


Figure 3: Comparison of reference selection distributions across models: "KS" denotes the Kolmogorov–Smirnov statistic against the human baseline (lower values indicate closer alignment), "*p*" is the associated p-value, and "Years Gap" denotes the difference in publication years between reference and the survey. For Fully-LLMGen, the survey year is set to 2025. Claude-3.7-Sonnet is used as the backbone LLM for all methods.

Results: Compare with Other Ranking Models

- QUAL-SG outperforms UPR and RankGPT
- RankGPT (Prompt-based ranking) performs poorly in distinguishing paper quality

Model	P%↑	R%↑	F1%↑
UPR (Sachan et al., 2022)	10.28	10.63	10.45
RankGPT (Sun et al., 2023)	14.55	15.09	14.81
QUAL-SG (ours)	**15.87**	**17.71**	**16.73**

Table 5: Citation quality comparison of different ranking models. For RankGPT, we instruct it via prompt to rank based on the same three criteria (§2.4) used in our QUAL-SG. The best results are marked **bold** and the second-best are underlined.

Results: Human Evaluation

- Survey generated from Human-guided setting rated more acceptable by human evaluators
- In general, the generated surveys currently fail to provide sufficient information coverage and critical analysis

Task	Comparison	Topic Relevance	Information Coverage	Critical Analysis	Overall Rating
Task 1	Comparable	33.3%	33.3%	26.7%	20.0%
	LLM-Generated > Human-written	20.0%	26.7%	26.7%	13.3%
Task 2	Comparable	33.3%	46.7%	40.0%	26.7%
	LLM-Generated > Human-written	33.3%	20.0%	20.0%	13.3%
Task 3	Comparable	40.0%	53.3%	46.7%	26.7%
	LLM-Generated > Human-written	26.7%	20.0%	20.0%	20.0%

Table 6: Human evaluation results across tasks. Each task includes five surveys from the Computer Science domain, all generated using Claude-3.7-Sonnet. For Task 2, the surveys were generated from the QUAL-SG pipeline.

Results: Ablation Study

- The performance of QUAL-SG declines across all ablation settings.

- Academic ranking is the most Important components, then co-cited expansion, relevance, and content diversity

Ablation Setting	P ↑	R ↑	F1 ↑
QUAL-SG	**15.87**	**17.71**	**16.73**
w/o co-cited expansion	10.07 (↓5.80)	11.52 (↓6.19)	10.75 (↓5.98)
w/o topical relevance	11.54 (↓4.33)	13.15 (↓4.56)	12.29 (↓4.44)
w/o academic impact	8.76 (↓7.11)	9.28 (↓8.43)	9.01 (↓7.72)
w/o content diversity	13.16 (↓2.71)	14.34 (↓3.37)	13.72 (↓3.01)

Table 7: Ablation study of QUAL-SG in the literature retrieval stage.

Future works

- Analyzing human citation behavior—such as citation intent, frequency, and location in the textual context for better paper selection
- Using full-body text of a paper may provide more useful information
- Improving survey quality via human-in-the-loop structure control, factual verification, and advanced long-document modeling to improve the quality

Thank you!

Paper