# 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]Universiy of Alberta

## Motivation

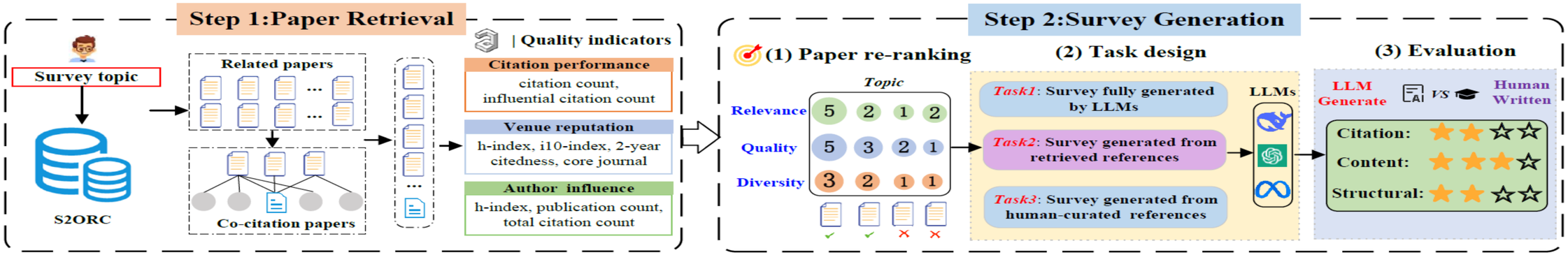
- **Automatic survey generation** has become a key task in the field of scientific document processing.
- RAG-based approaches **overlook the assessment of academic quality** during the paper retrieval stage.
- **Missing large-scale datasets for evaluating** generated surveys against human-written standards.

## Dataset

We released **SurveyGen**, which includes over 4200 human-written surveys, 115,376 sections, 242,143 references directly cited within the surveys.

| Dataset | Domains | #Docs | #Input Len | #Target Len | #Input Docs | Structural Outline | Quality Indicators | Multi-level Citation | For Survey Generation |
|---|---|---|---|---|---|---|---|---|---|
| PubMed (2018) | Bio | 133K | 3016 | 203 | 1 | ✓ | ✗ | ✗ | ✗ |
| ArXiv (2018) | Mixed | 215K | 4938 | 220 | 1 | ✓ | ✗ | ✗ | ✗ |
| SciSummNet (2019) | CL | 1K | 4417 | 151 | 61.00 | ✗ | ✗ | ✗ | ✗ |
| Multi-XScience (2020) | CS | 40.5K | 778 | 116 | 4.42 | ✗ | ✗ | ✗ | ✗ |
| BigSurvey (2022) | Mixed | 4.4K | 11893 | 1051 | 76.30 | ✗ | ✗ | ✗ | ✗ |
| SciReviewGen (2023) | CS | 10.2K | 12503 | 8082 | 68.00 | ✓ | ✗ | ✗ | ✓ |
| SurveyGen(ours) | Mixed | 4.2K | 11423 | 5115 | 57.58 | ✓ | ✓ | ✓ | ✓ |

## Our approach : QUAL-SG

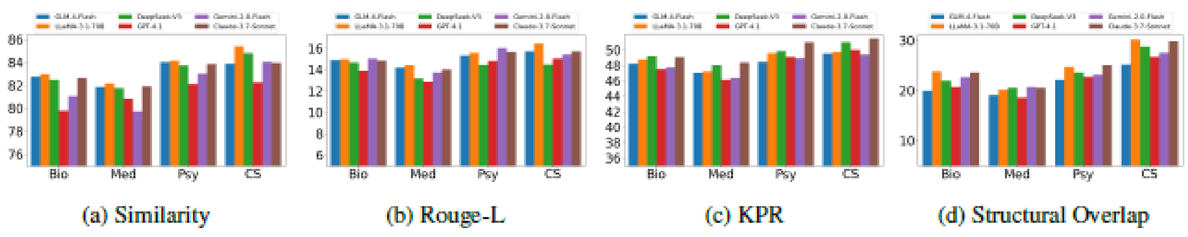



We build **QUAL-SG**, which includes two stages: **paper retrieval** and **survey generation**. The retrieval stage includes three steps: **(1)** retrieving topic-relevant papers, **(2)** expanding with frequently co-cited papers, and **(3)** enriching them with quality-related metadata. Based on the retrieved set, the generation stage first re-ranks the papers from three key aspects, then prompts LLMs to perform tasks under different input conditions. Finally, we evaluate the generated surveys against human-written ones across multiple dimensions.

## Results

| Model | Citation Quality | | | | Content Quality | | | Structural Consistency | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc. ↑ | P ↑ | R ↑ | F1 ↑ | Sim. ↑ | R-L ↑ | KPR ↑ | $Rel_{LLM}$ | 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** |

| Model | Citation Quality | | | Content Quality | | | Structural Consistency | |
|---|---|---|---|---|---|---|---|---|
| | P ↑ | R ↑ | F1 ↑ | Sim. ↑ | R-L ↑ | KPR ↑ | $Rel_{LLM}$ | 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**† |




(a) Similarity (b) Rouge-L (c) KPR (d) Structural Overlap

- Only about 36% of the LLM-generated references are authentic, the RAG-retrieved references are real but largely **different from those selected by humans**, and the structural overlap with human-written surveys is only 14.89%.
- **Our QUAL-SG significantly improves** the reference matching, and further enhances performance at both the content and structural levels.
- The performance of LLM-generated surveys **varies across disciplines**, with better results observed in computer science.

| 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% |

- **Human evaluation results** show that human–AI collaboration—where humans refine the outline and select references, and then the LLM generates the survey content—achieves higher scores than fully LLM-generated surveys. However, they still fall short in terms of critical analysis and information coverage.

## Discussions

- LLMs can assist in survey generation, they are still unable to independently craft surveys that meet academic standards at the current stage.
- Investigating human reference selection behavior is left for future work, and full-text access could further improve survey generation.

Code and data are available at:
https://github.com/tongbao96/SurveyGen