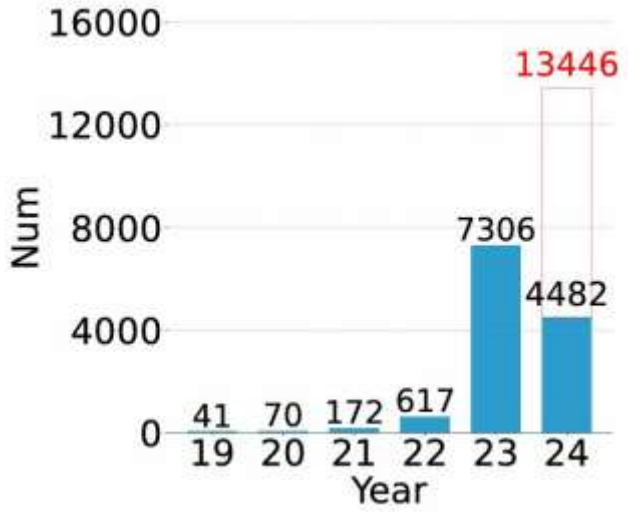
AutoSurvey: Large Language Models Can Automatically Write Surveys

Yidong Wang[1,2*], **Qi Guo**[2,3*],
Wenjin Yao[2], **Hongbo Zhang**[1], **Xin Zhang**[4], **Zhen Wu**[3],**Meishan Zhang**[4],
Xinyu Dai[3], **Min Zhang**[4], **Qingsong Wen**[5], **Wei Ye**[2†], **Shikun Zhang**[2†], **Yue Zhang**[1†]

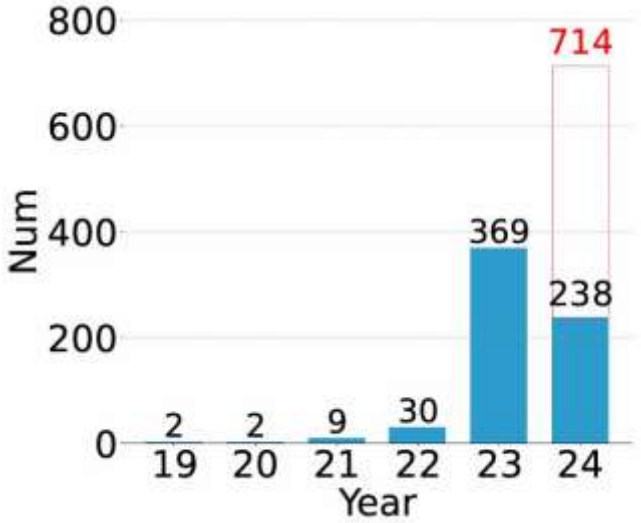
[1]Westlake University, [2]Peking University,
[3]Nanjing University, [4]Harbin Institute of Technology, Shenzhen, [5]Squirrel AI

NeurIPS 2024

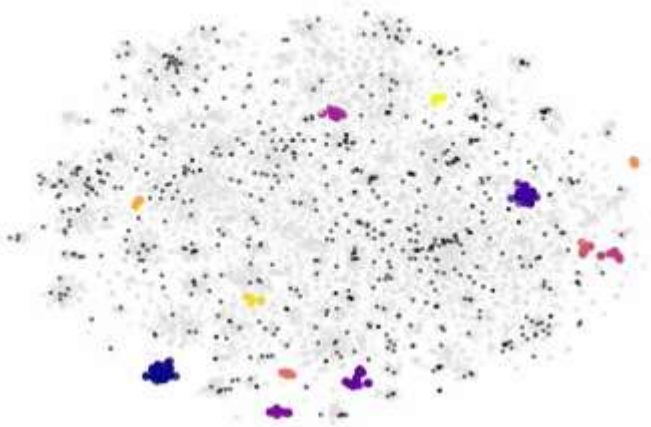
Background



(a) Paper about LLM.



(b) Survey about LLM.

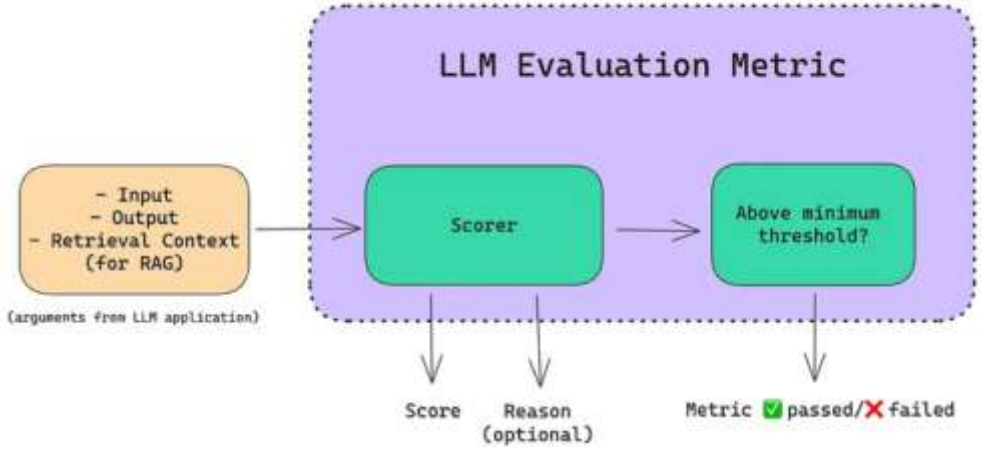


(c) A t-SNE visualization of LLM-related surveys and papers shows clusters of research papers that currently lack comprehensive survey coverage.

Background

模型[1]		deepseek-chat	deepseek-reasoner
上下文长度		64K	64K
最大思维链长度[2]		-	32K
最大输出长度[3]		8K	8K
标准时段价格（北京时间 08:30-00:30）	百万tokens输入（缓存命中）[4]	0.5元	1元
	百万tokens输入（缓存未命中）	2元	4元
	百万tokens输出[5]	8元	16元
	百万tokens输入（缓存命中）	0.25元（5折）	0.25元（2.5折）

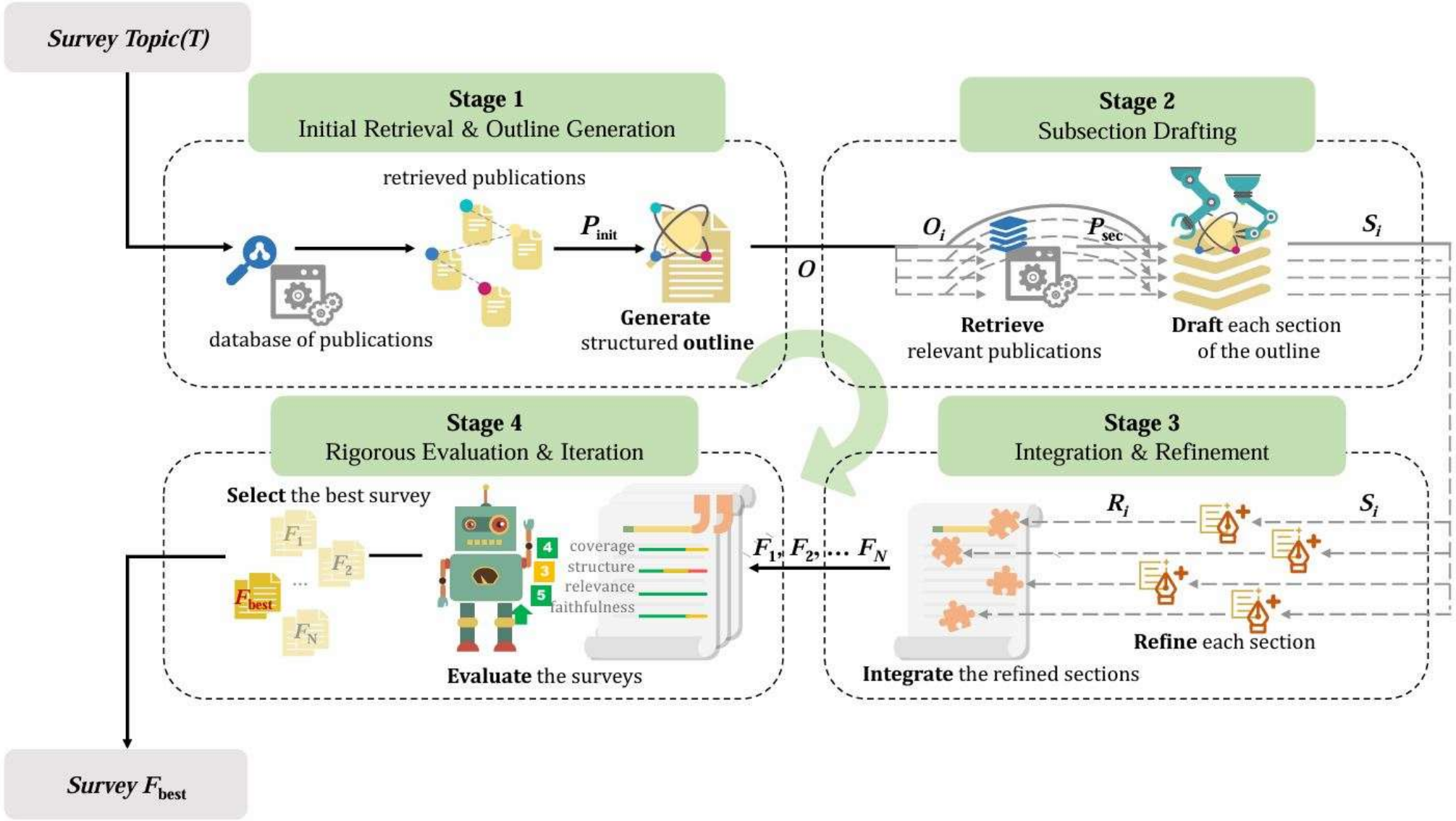
(1) Context window limitations



(3) The lack of evaluation benchmark

(2) Parametric knowledge constraints

Method


Survey Topic(T)
Stage 1
Initial Retrieval & Outline Generation
retrieved publications
database of publications
P_{init}
Generate
structured outline
O
Stage 2
Subsection Drafting
O_i
P_{sec}
S_i
Retrieve
relevant publications
Draft each section
of the outline
Stage 4
Rigorous Evaluation & Iteration
Select the best survey
F_1
F_2
...
F_{best}
F_N
coverage
structure
relevance
faithfulness
4
3
5
Evaluate the surveys
$F_1, F_2, \ldots F_N$
Stage 3
Integration & Refinement
R_i
S_i
Refine each section
Integrate the refined sections
Survey F_{best}

Experiment

Survey Creation Speed

$$Time = T_r + T_w + T_e = T_r + \frac{L}{E \times M} + \frac{1}{2} \times \frac{L}{E \times M}$$

Citation Quality

$$\text{Recall} = \frac{\sum_{i=1}^{|C|} h(c_i, \text{Ref}_i)}{|C|}$$

$$\text{Precision} = \frac{\sum_{i=1}^{|C|} \sum_{k=1}^{|\text{Ref}_i|} h(c_i, \text{Ref}_i) \cap g(c_i, r_{i_k})}{\sum_{i=1}^{|C|} |\text{Ref}_i|}$$

Where $C = \{c_1, c_2, \ldots\}$,and $\text{Ref}_i = \{r_{i_1}, r_{i_2}, \ldots\}$

Content Quality

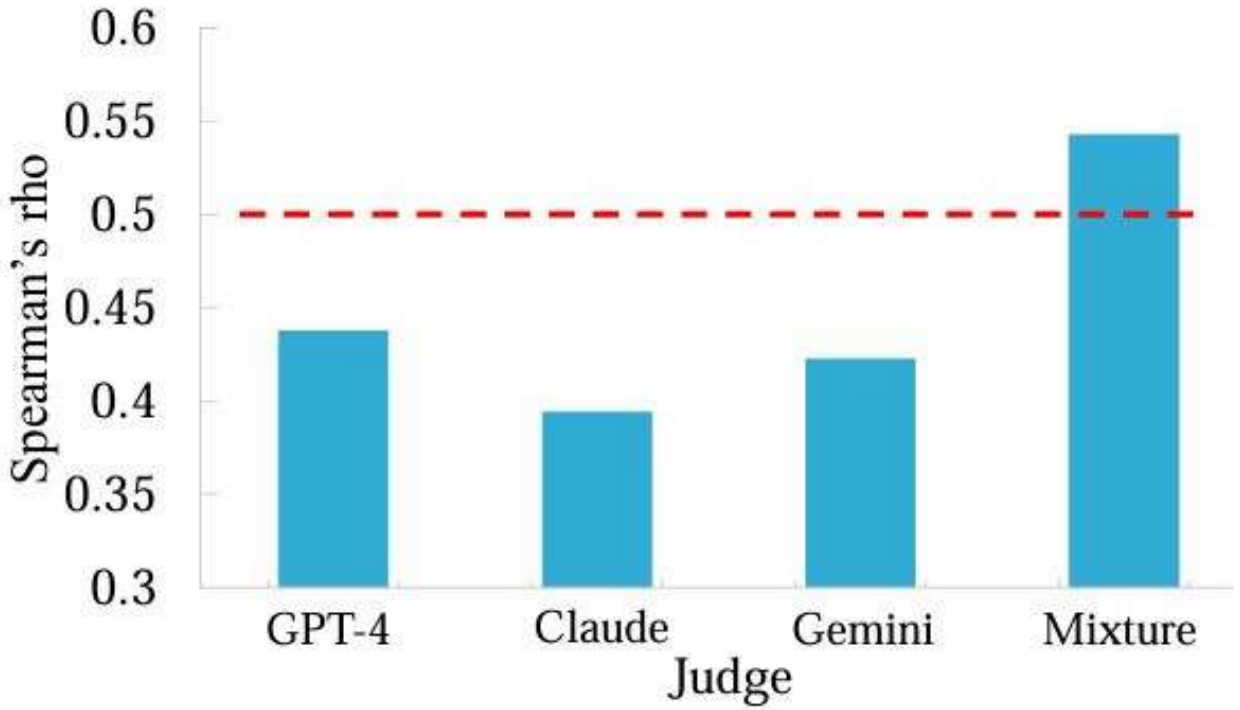
Coverage: Assesses the extent to which the survey encapsulates all aspects of the topic.
Structure: Evaluates the logical organization and coherence of each section.
Relevance: Measures how well the content aligns with the research topic.

Experiment

Table 2: Results of naive RAG-based LLM generation, Human writing and AutoSurvey. Note that AutoSurvey and naive RAG-based LLM generation both use Claude-haiku as the writer. **Note that human writing surveys used for evaluation are excluded during the retrieval process.**

Survey Length (#tokens)	Methods	Speed	Citation quality		Content Quality			
			Recall	Precision	Coverage	Structure	Relevance	Avg.
8k	Human writing	0.16	80.00	87.50	4.50	4.16	5.00	4.52
	Naive RAG-based LLM generation	79.67	$78.14_{\pm 5.23}$	$71.92_{\pm 6.83}$	$4.40_{\pm 0.48}$	$3.86_{\pm 0.71}$	$4.86_{\pm 0.33}$	4.33
	AutoSurvey	107.00	$82.48_{\pm 2.77}$	$77.42_{\pm 3.28}$	$4.60_{\pm 0.48}$	$4.46_{\pm 0.49}$	$4.8_{\pm 0.39}$	4.61
16k	Human writing	0.14	88.52	79.63	4.66	4.38	5.00	4.66
	Naive RAG-based LLM generation	43.41	$71.48_{\pm 12.50}$	$65.31_{\pm 15.36}$	$4.46_{\pm 0.49}$	$3.66_{\pm 0.69}$	$4.73_{\pm 0.44}$	4.23
	AutoSurvey	95.51	$81.34_{\pm 3.65}$	$76.94_{\pm 1.93}$	$4.66_{\pm 0.47}$	$4.33_{\pm 0.59}$	$4.86_{\pm 0.33}$	4.60
32k	Human writing	0.10	88.57	77.14	4.66	4.50	5.00	4.71
	Naive RAG-based LLM generation	22.64	$79.88_{\pm 4.35}$	$65.03_{\pm 8.39}$	$4.41_{\pm 0.64}$	$3.75_{\pm 0.72}$	$4.66_{\pm 0.47}$	4.23
	AutoSurvey	91.46	$83.14_{\pm 2.44}$	$78.04_{\pm 3.14}$	$4.73_{\pm 0.44}$	$4.26_{\pm 0.69}$	$4.8_{\pm 0.54}$	4.58
64k	Human writing	0.07	86.33	77.78	5.00	4.66	5.00	4.88
	Naive RAG-based LLM generation	12.56	$68.79_{\pm 11.00}$	$61.97_{\pm 13.45}$	$4.4_{\pm 0.61}$	$3.66_{\pm 0.47}$	$4.66_{\pm 0.47}$	4.19
	AutoSurvey	73.59	$82.25_{\pm 3.64}$	$77.41_{\pm 3.84}$	$4.73_{\pm 0.44}$	$4.33_{\pm 0.47}$	$4.86_{\pm 0.33}$	4.62

Experiment



Spearman's rho shows how well LLM rankings align with human experts, with values above 0.3 indicating positive correlation and above 0.5 indicating strong positive correlation.

Experiment

Table 3: Ablation study results for AutoSurvey with different components removed.

Methods	Citation Quality		Content Quality			
	Recall	Precision	Coverage	Structure	Relevance	Avg.
AutoSurvey	$83.48_{\pm 5.05}$	$77.15_{\pm 6.05}$	$4.7_{\pm 0.45}$	$4.16_{\pm 0.73}$	$4.93_{\pm 0.30}$	4.57
AutoSurvey w/o retrieve	$60.11_{\pm 6.42}$	$51.65_{\pm 6.33}$	$4.51_{\pm 0.49}$	$4.01_{\pm 0.74}$	$4.88_{\pm 0.32}$	4.44
AutoSurvey w/o reflection	$83.23_{\pm 3.82}$	$76.36_{\pm 4.08}$	$4.76_{\pm 0.42}$	$4.13_{\pm 0.76}$	$4.88_{\pm 0.32}$	4.56

Table 4: Performance of AutoSurvey with different base LLM writers.

Base LLM writer	Citation Quality		Content Quality			
	Recall	Precision	Coverage	Structure	Relevance	Avg.
GPT-4	$80.25_{\pm 4.19}$	$78.83_{\pm 7.00}$	$4.8_{\pm 0.54}$	$4.46_{\pm 0.49}$	$4.86_{\pm 0.33}$	4.70
Claude-haiku	$82.45_{\pm 2.77}$	$76.31_{\pm 2.18}$	$4.66_{\pm 0.47}$	$4.26_{\pm 0.67}$	$4.86_{\pm 0.33}$	4.58
Gemini-1.5-pro	$78.13_{\pm 2.39}$	$71.24_{\pm 3.28}$	$4.86_{\pm 0.33}$	$4.33_{\pm 0.78}$	$4.93_{\pm 0.25}$	4.69
Human	85.86	80.51	4.71	4.43	5	4.70

Experiment

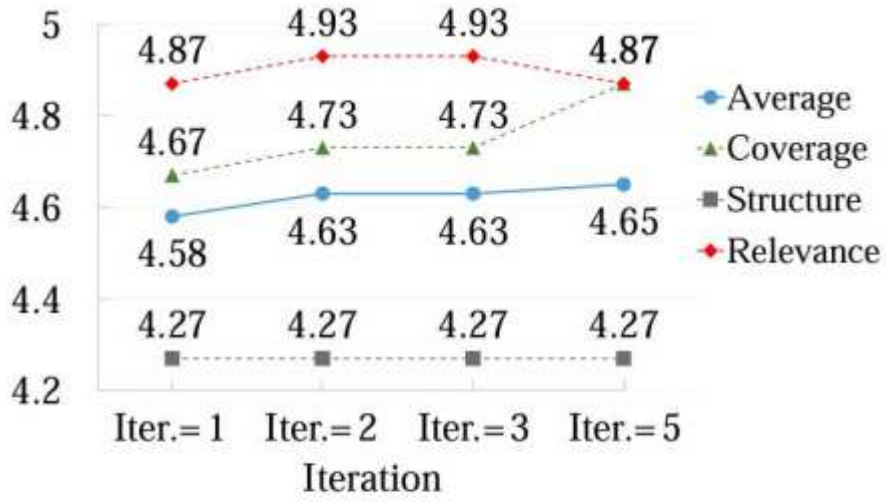
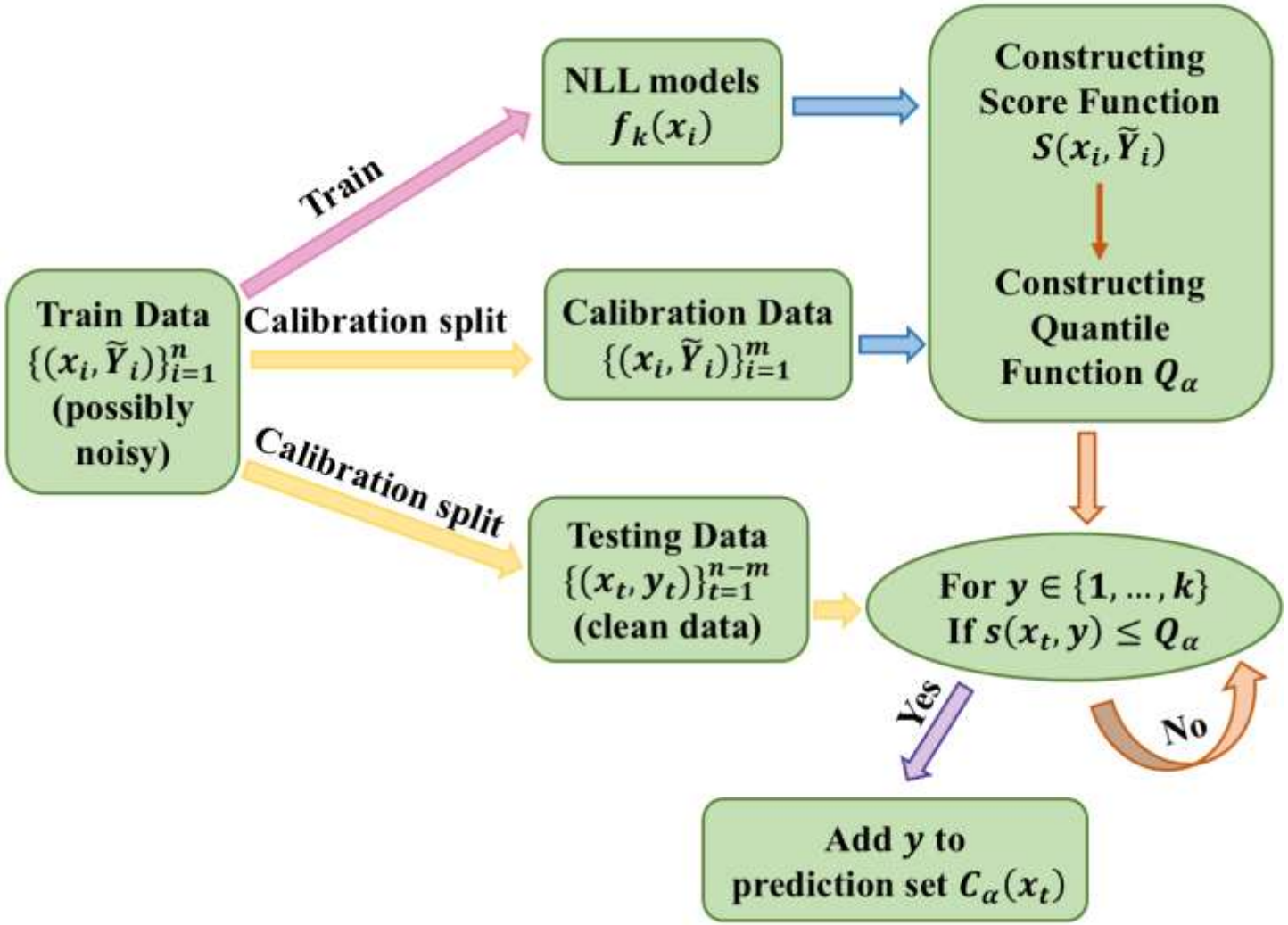


Figure 4: Impact of Iteration on Auto-Survey Performance.

Table 5: Performances given different references.

Methods	Accuracy
Direct	$58.40_{\pm 4.96}$
Naive RAG-based LLMs	$65.20_{\pm 8.06}$
Upper-bound	$73.60_{\pm 3.44}$
AutoSurvey	$67.60_{\pm 4.96}$

Progress Report



Thanks