SURVEYX: Academic Survey Automation via Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable comprehension capabilities and a vast knowledge base, making them promising tools for automated academic survey generation. However, existing automated survey systems face key challenges, including the lack of multimodal processing, limited context window utilization, and incomplete evaluation frameworks. Inspired by human writing workflows, we introduce SURVEYX, an efficient and organized system for automated survey generation that decomposes the survey composing process into two phases: the Preparation and Generation phases. SURVEYX is capable of automatically acquiring up-to-date literature and producing coherent, cross-disciplinary survey articles. Moreover, SURVEYX introduces a unified multimodal generation strategy that integrates textual synthesis with figure-aware content construction. Experimental results demonstrate that SURVEYX achieves state-of-the-art performance among automated survey generation systems, approaching the performance of human experts. The code of SURVEYX is available at this website.

1 Introduction

In recent years, computer science has advanced rapidly across various fields (Bick et al., 2024; Gupta, 2023; Loakman et al., 2024). Statistics indicate that arXiv.org[1] receives about a thousand new papers daily. Figure 1 highlights a remarkable trend: over the past three years (2022-2025), the number of papers published on arXiv has increased substantially, rising from 186,339 to 393,512, representing a growth of over 100%. The exponential growth of literature has made it increasingly challenging for researchers attempting to comprehend the technological evolution and developmental trajectories of specific subfields from the ground up (Zhong et al., 2024). Survey papers are instrumental in elucidating the current state of research and the historical progression within a given topic (Chang et al., 2024; Zhao et al., 2023; Zheng et al., 2025b). However, the workload involved in manually writing these surveys is continually increasing, threatening the ability to maintain comprehensive coverage and high quality of these surveys. Against the backdrop of information overload, there is an urgent need to develop efficient systems for automated survey generation. The rise of Large Language Models (LLMs) has rendered automated survey generation a viable approach (Hurst et al., 2024; Team et al., 2023; Kevian et al., 2024; Ferrag et al., 2025; Liu et al., 2025a; Kostikova et al., 2025; Chen et al., 2025). The training of large text corpora enables large language models to understand the content of academic papers and to compare, summarize, and synthesize them. (Guo et al., 2023; Minaee et al., 2024; Liu et al., 2023; Tang et al., 2024). Despite this potential, leveraging LLMs for automated survey generation poses several challenges.

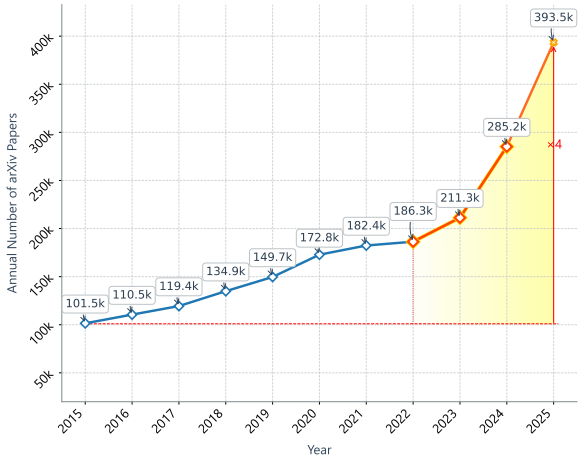


Figure 1: The number of papers received annually by the arXiv website from 2015 to 2025, with data sourced from our arXiv database. The number of submissions for 2025 is four times greater than that of 2015.

[1] https://arxiv.org/

- **Limited multimodal reasoning capability:** Current LLM-based systems are predominantly designed for text-level understanding and generation, without systematically supporting the comprehension or generation of academic visual elements such as flowcharts or structure diagrams. This is crucial for accurately capturing, synthesizing, and presenting complex scientific knowledge.

- **Inefficient utilization of context windows:** Although recent LLMs have expanded context capacities, existing methods often fail to effectively organize and exploit the available context(Li et al., 2024, 2023; Kaddour et al., 2023; Liu et al., 2025b). As a result, incorporating large-scale, heterogeneous references remains challenging, which limits the applicability of existing methods to all-domain survey generation, where reasoning over hundreds of papers is typically required.

- **Insufficient systematic evaluation frameworks:** Prior work largely focuses on evaluating textual quality while overlooking the rigorous assessment of visual components. The absence of unified and systematic evaluation protocols limits the ability to reliably measure the completeness, correctness, and communicative effectiveness of generated surveys.

To address these deficiencies, we propose SURVEYX, an efficient and well-organized system for automated survey generation. SURVEYX divides the survey composition process into two phases: the Preparation Phase and the Generation Phase. In the Preparation Phase, SURVEYX employs retrieval algorithms to search and filter highly relevant references from the internet. Inspired by human experts, it utilizes a reference pre-processing method, termed AttributeTree, to distill key information from the references, constructing a reference materials database for efficient retrieval through the Retrieval Augmented Generation (RAG) technique. In the Generation Phase, SURVEYX utilizes the AttributeTrees to sequentially generate the outline and main body of the survey, ensuring that the generated survey has a clear structure and accurate content. Additionally, SURVEYX supports the comprehension and generation of academic figures and tables, further enriching the presentation of generated surveys. Beyond the generation process, SURVEYX introduces explicit evaluation standards for visual elements in academic surveys, thereby providing a more comprehensive evaluation protocol for subsequent research.

To sum up, our main contributions are threefold:

- **Multimodal understanding and generation for survey composition.** We move beyond purely text-based survey generation by enabling the comprehension and generation of academic visual elements.

- **AttributeTree-based context optimization.** We propose an AttributeTree-based reference pre-processing strategy that mirrors the hierarchical and selective manner in which human authors organize and distill information when writing surveys.

- **Extended evaluation with visual criteria.** We expand existing evaluation frameworks by extending textual assessment dimensions and introducing explicit evaluation criteria for visual elements, enabling more comprehensive evaluation of multimodal survey quality.

2 Related Work

Retrieval Augmented Generation (RAG). RAG technique is used to help LLMs access external knowledge, thereby enhancing their text generation capabilities (Gao et al., 2023; Fan et al., 2024; Huang and Huang, 2024; Hu and Lu, 2024; Zheng et al., 2025a; Lyu et al., 2025; Singh et al., 2025). This technique has proven especially useful in tasks that require up-to-date or domain-specific information, such as QA (Question Answering) (Ke et al., 2024; Asai et al., 2024), Dialog Generation (Rajput et al., 2024; Lin et al.), Reasoning (Cheng et al., 2023; Zhang et al., 2023), etc. Beyond traditional NLG tasks, RAG is also used for long-form text generation, as this task often requires handling extensive external knowledge (Wang et al., 2024; Shao et al., 2024; Zhang et al., 2023). Compared to previous work that directly used the raw text of references as the retrieval data source, our work significantly improves retrieval efficiency and context window utilization by converting it into attribute trees.

Automated Survey Generation. Early efforts to automate academic survey writing relied on basic heuristics or multi-document summarization techniques.(Lewis et al., 2021; Mohammad et al.,

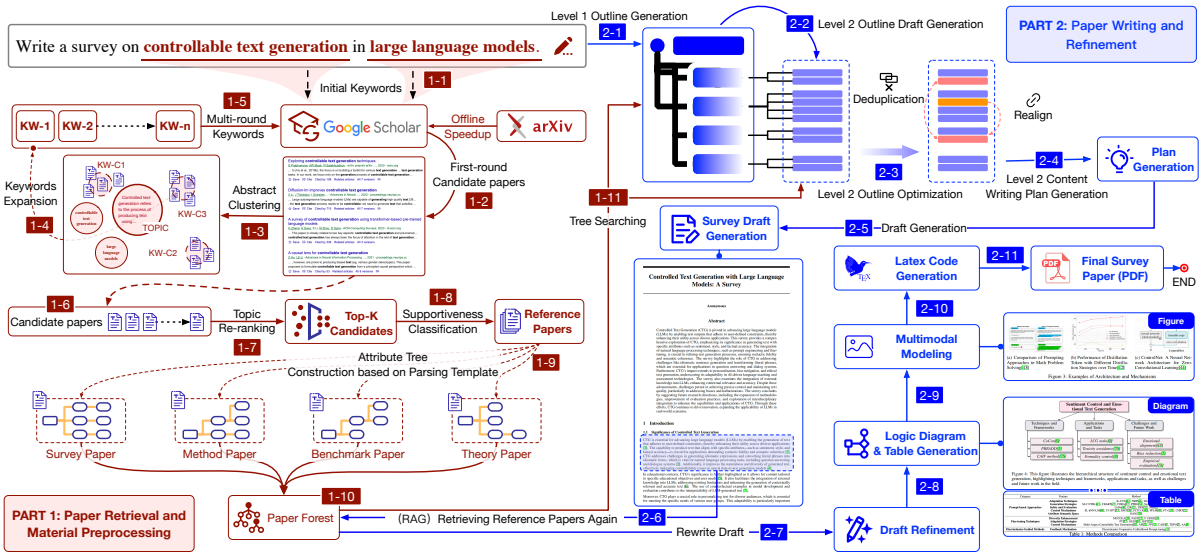


Figure 2: Pipeline of SURVEYX.

2009; Jha et al., 2015; Wang et al., 2018) Such limitations persisted until the advent of large language models. AutoSurvey(Wang et al., 2024) introduced an LLM-based pipeline integrating retrieval, outline generation, section drafting, and iterative refinement for long-form survey writing. Agentic AutoSurvey(Liu et al., 2025c) extended AutoSurvey into a multi-agent framework with specialized roles for retrieval, topic organization, writing, and evaluation, improving synthesis quality and citation coverage. AutoSurvey2(Wu et al., 2025) further enhanced this paradigm by adopting a multi-stage retrieval-augmented pipeline with parallel section generation and multi-LLM evaluation to improve structural coherence and topical relevance. LLM×MapReduce-V2/V3(Wang et al., 2025; Chao et al., 2025) addressed long-form survey generation by decomposing the task into modular stages for scalable aggregation and refinement of large-scale literature. SurveyForge (Yan et al., 2025) proposed a structure-aware framework that leverages heuristic outline construction and scholar navigation agents to retrieve high-quality references and generate more logically organized surveys. Nevertheless, existing systems primarily focus on textual synthesis and do not address the generation or evaluation of visual content.

3 Methodology

In this section, we delineate the pipeline of SURVEYX, which includes two main phases: Preparation Phase (section 3.1) and Generation Phase (section 3.2). The Preparation Phase focuses on collecting and pre-processing references to build a retrievable database for RAG. The Generation Phase creates the survey draft from this database, followed by post-processing to improve quality and expression. An overview of this system is depicted in Figure 2.

3.1 Preparation Phase

The Preparation Phase consists of two stages: the references acquisition stage and the references pre-processing stage.

3.1.1 References Acquisition

Considering the current shortage of effective tools for efficiently acquiring a large volume of relevant references, we implement an LLM-driven retrieval pipeline for acquiring references.

The retrieval pipeline is divided into two steps: reference recall and reference filter (steps 1-1~1-6 and steps 1-7~1-8 in Figure 2). 1) The goal of the reference recall step is to avoid missing any references related to the topic as much as possible. 2) The paper filter step aims to filter out references unrelated to the topic as much as possible. To achieve 1), we design the *Keyword Expansion Algorithm* 1. The algorithm initializes the keyword pool with an initial keyword K_0. After searching references by newly added keywords, it performs semantic clustering on the abstracts of all retrieved references Doc and summarizes keywords for each cluster C_i. The obtained keywords K_c are then semantically compared with the existing keyword pool K_{pool} and the topic T, the most appropriate keyword k^* based on semantic distance is added to the keyword

pool. This process is repeated until the number of Doc reaches the threshold θ (set as 1000).

Algorithm 1 Keyword Expansion Algorithm

Input: initial keywords set K_0, topic T
Output: retrieved documents Doc
keyword pool: $K_{\text{pool}} \leftarrow K_0$
while LengthOf(Doc) $< \theta$ **do**
  $k \leftarrow$ newly added keywords in K_{pool}
  $Doc \leftarrow Doc + \text{Search}(k)$
  $n \leftarrow \text{LengthOf}(K_{\text{pool}})$
  $C \leftarrow \text{Cluster}(Doc, n+1)$
  $K_c \leftarrow \text{ExtractKeywords}(C)$
  $Dist \leftarrow \text{CalcDist}(K_c, K_{\text{pool}}, T, \text{Weights})$
  $k^* \leftarrow \text{Select}(Dist, K_{\text{pool}})$
  $K_{\text{pool}} \leftarrow K_{\text{pool}} + k^*$
end while
Return: Doc

Weights represent semantic distance. To emphasize topic relevance, the topic's weight is doubled. The *Select* function chooses the candidate word with the minimal average distance to existing keywords and maximal farthest distance. Specifically:

$$k^* = \arg\min_{k_c \in K_C} \left(R_1(k_c) + R_2(k_c)\right) \quad (1)$$

R_1 and R_2 represent different ranking calculation methods. Their calculation methods are as follows:

$$R_2(k_c) = \mathbb{R}_{K_c}\left(-\max_{k_e \in K_{\text{pool}}} \text{sim}(E(k_c), E(k_e))\right) \quad (2)$$

$$R_2(k_c) = \mathbb{R}_{K_c}\left(-\max_{k_e \in K_{pool}} \text{sim}(E(k_c), E(k_e))\right) \quad (3)$$

Here, $\mathbb{R}(\cdot)$ refers to the Rank Function, and $E(\cdot)$ represents the embedding model.

To achieve 2), we propose a 2-step filtration algorithm. First, an embedding model computes semantic relevance between the topic and abstracts, selecting Top-K references as coarse-grained filtering. Then, LLMs perform fine-grained filtering for higher precision. This ensures selected papers are highly relevant, enhancing survey quality.

As for the retrieval data source, it includes both offline and online data sources. The offline data source consists of 3,021,841[2] available papers downloaded from the arXiv.org, with new papers fetched daily. The online data source is a self-developed crawler system based on Google Scholar and Semantic Scholar. Offline data allows fast access to existing references, while online data ensures timely, multi-source updates.

[2]Data as of December 31, 2025.

3.1.2 References Pre-processing

Directly feeding full texts into LLMs often results in suboptimal context utilization and noisy information extraction. We propose *AttributeTree*, a structured reference modeling approach inspired by expert survey writing practices. *AttributeTree* is constructed through a two-stage process. First, LLM leverages its understanding of the target survey topic and a subset of representative references to infer a domain- and type-specific AttributeTree template. Based on this learned template, the LLM then extracts key information from individual references and instantiates each reference into a structured AttributeTree. (steps 1-9~1-10 in Figure 2). Examples of generated attribute tree templates can be found in Figure B.1 in the Appendix. This method significantly increases the information density of the reference materials and efficiently utilizes the context window of LLMs. And it supports truly all-domain surveys generation. In addition, we leverage multimodal LLMs to analyze the figures associated with each reference, integrating visual understanding with textual descriptions to accurately guide the instantiation of the AttributeTree templates.

3.2 Generation Phase

After obtaining all the reference materials needed for writing the survey, the generation phase begins. This phase is divided into three stages: **outline generation**, **main body generation**, and **post refinement**.

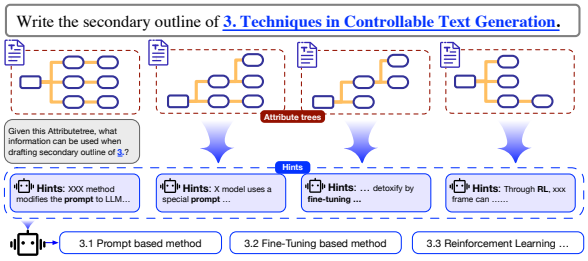


Figure 3: An example of generating secondary outlines. LLMs first generate hints based on the attribute tree to guide the generation of the secondary outline. Then, by synthesizing all hints, LLMs identify the most suitable entry points to determine the segmentation strategy and generate the secondary outline.

3.2.1 Outline Generation

Outline generation is the most critical stage of survey creation, as a clear, logical outline ensures a cohesive and informative result. Observing that hu-

mans typically categorize or summarize commonalities and differences of references to guide the writing of the survey, we propose an outline generation method called Outline Optimization (steps 2-1~2-3 in Figure 2). This method consists of two steps. In the first step, LLMs generate hints from each reference's attribute tree to guide secondary outline creation. These hints help LLMs understand how each reference contributes to the survey's systematic framework[3]. LLMs then generate the secondary outline based on these hints (see Figure 3), enabling better identification of commonalities and differences among references for efficient synthesis. In the second step, secondary outlines are stripped to headings for deduplication. LLMs then reorganize them into the primary outline. This step effectively reduces redundancy. With this outline optimization method, we generate logically rigorous and well-structured survey outlines.

3.2.2 Content Generation

In the content generation stage (steps 2-4~2-5 in Figure 2), inspired by human writing habits, we continue to use the hint-based method from outline generation to enrich content depth. Here, hints are derived from the secondary outline to guide subsection-level generation. LLMs generate each subsection using these hints and the prior outline, while also referencing other subsections to maintain overall consistency.

3.2.3 Post Refinement

Inspired by human survey writing, we introduce a post-refinement stage to enhance draft quality. It has two goals: (1) improve content quality: citation accuracy, fluency, and consistency–via an RAG-based rewriting module; (2) enrich presentation by adding figures and tables using a Figure and Table Generation Module.

RAG-based Rewriting Module. Although strategies were applied during content generation to ensure consistency, they remain insufficient, and citation accuracy is unverified. We introduce a RAG-based rewriting module(steps 2-6~2-7 in Figure 2). It uses draft paragraphs as queries to retrieve relevant materials from the attribute forest, then prompts LLMs to check each paragraph. This process removes irrelevant citations, adds relevant ones, and polishes the text with contextual awareness. The module improves both citation quality and overall content consistency.

Figure and Table Generation Module. Text-only surveys often lack expressive diversity, limiting clarity. To address this, we design a figure and table generation module(steps 2-8~2-11 in Figure 2). Leveraging the AttributeTree and the current survey context, LLMs generate figure and table code, which is subsequently compiled into visual content. The resulting visual candidates are then evaluated by a multimodal LLM to assess their expressiveness and relevance, and only the most informative and well-represented figures are selected for inclusion in the final survey. Additionally, for each subsection, multimodal LLMs retrieve relevant figures from references and incorporate them when appropriate, while simultaneously analyzing reference figures to guide the overall figure generation process.

4 Experiments

4.1 Evaluation Metrics

Content Evaluation Building upon the evaluation benchmark proposed by Wang et al. (2024), we introduce two additional dimension: synthesis and critical analysis. Synthesis assesses the model's ability to integrate disparate studies, identify overarching patterns or tensions, and construct a coherent intellectual structure beyond isolated summaries. Critical analysis evaluates the depth of scholarly critique, including the identification of methodological limitations, theoretical inconsistencies, and open research gaps. The detailed prompts and evaluation protocols are provided in Appendix A.2.

Retrieved Reference Evaluation Prior work does not explicitly assess the relevance of references automatically retrieved for survey generation. We address this gap by introducing the following three evaluation metrics:

(1) IoU (Insertion over Union). This metric measures the similarity between machine-retrieved and human-retrieved references by calculating the degree of overlap between the two. The calculation method is as follows

$$IoU = \frac{Doc_{human} \cap Doc_{machine}}{Doc_{human} \cup Doc_{machine}}$$

(2) Relevance$_{semantic}$ (semantic-based reference relevance). This metric is based on an embedding model, calculating the cosine similarity between

[3]"Systematic framework" refers to the framework proposed in a survey for organizing and synthesizing all the references.

the embeddings of the retrieved references and the topic to measure their relevance:

$$R_{semantic} = \frac{1}{|Doc|} \sum_{d \in Doc} \text{sim}(E(d), E(topic))$$

(3) Relevance$_{LLM}$ (LLM-based reference relevance). This metric is to prompt LLMs to directly evaluate the relevance of the retrieved references to the topic. The calculation method is as follows:

$$R_{LLM} = \frac{1}{|Doc|} \sum_{d \in Doc} LLM(Prompt(d, topic))$$

The prompts used for $Relevance_{LLM}$ are provided in Figure A.7 in the Appendix.

Visual Elements Evaluation To evaluate the quality of generated visual elements, we adopt an LLM-as-a-Judge paradigm. Specifically, for each figure, we construct the evaluation context by extracting the continuous span of sections in which the figure is referenced, from its first appearance to its last mention in the paper. We evaluate visual quality along five complementary dimensions, each targeting a distinct aspect of figure effectiveness in academic communication:

(1) Semantic Correctness measures whether the visual elements accurately reflect the concepts described in the text, without hallucinations or conceptual errors.

(2) Expressiveness & Information Density evaluates how effectively the figure communicates core information within limited visual space while maintaining clarity.

(3) Readability & Cognitive Load assesses the ease of interpretation, considering layout clarity, visual hierarchy, and required cognitive effort.

(4) Contextual Consistency examines the alignment of the figure with its surrounding context, including terminology usage and narrative coherence.

(5) Reusability & Self-Containment evaluates whether the figure can be understood and reused independently of the full paper.

Detailed scoring criteria and prompt formulations for each evaluation dimension are provided in Appendix D.

4.2 Experiment Settings

During the retrieval and evaluation stages, we use bge-base-en-v1.5 (Xiao et al., 2023) as the embedding model. Throughout the entire process, GPT-4o serves as our LLM agent. We compare our system against several baselines: **Human**, referring to human-written surveys collected from arXiv.org; **AutoSurvey** (Wang et al., 2024), an automated survey generation system, for which we adopt the 64k-token version in our experiments; **Agentic AutoSurvey** (Liu et al., 2025c), where GPT-4o is used for all sub-agents; **SurveyForge** (Yan et al., 2025), which is evaluated using the same reference set as SURVEYX and GPT-4o as the generation model; and **SurveyGo**, evaluated under the same experimental settings as reported in (Chao et al., 2025). We select 25 representative topics to generate corresponding surveys for comparison, with topic details provided in Table 6 in the Appendix.

For visual element evaluation, we employ agent-based implementations to generate illustrative figures from the same survey text segments as SurveyX. Concretely, we use the google-adk framework with gemini-3-pro-preview, and the openai-agents framework with GPT-4.1, respectively. The generated figures are then compared against human-authored survey figures together with their original textual context, enabling a controlled assessment of visual quality and contextual alignment.

Ablation To evaluate the impact of different modules or methods on the performance of SURVEYX, we designed the following ablation setting:

(1) For the retrieval algorithms within the retrieval module of the references acquisition stage, we replace the keyword expansion–based retrieval with traditional bibliometric and citation analysis–based retrieval methods (for detailed steps, see Appendix C).

(2) For the AttributeTree method in the references pre-processing stage, we replace the originally generated attribute trees with the full text of the references.

(3) For the outline optimization method in the outline generation stage, we design prompts to let the LLM generate the entire outline in one step.

(4) For the RAG-based rewriting module in the post-refinement stage, we eliminate the entire RAG-based rewriting module.

4.3 Experiment Results and Analysis

Main results (1) The experimental results for content quality evaluation (Table 1) indicate that SURVEYX performs exceptionally well across all metrics, particularly in Coverage (4.95), Structure (4.91), and Relevance (4.94), closely approaching

Table 1: Content quality evaluation results. All LLM-Agent is GPT-4o.

Model	Coverage	Structure	Relevance	Synthesis	Critical Analysis	Avg.	Recall	Precision	F1
AutoSurvey	4.65 ± 0.104	4.30 ± 0.083	4.86 ± 0.074	3.98 ± 0.057	3.75 ± 0.170	4.31	82.25	77.41	79.76
Agentic AutoSurvey	4.62 ± 0.098	4.24 ± 0.180	4.88 ± 0.120	4.21 ± 0.166	3.98 ± 0.085	4.39	80.93	74.21	77.42
AutoSurvey2	4.69 ± 0.052	4.71 ± 0.024	4.83 ± 0.058	4.03 ± 0.098	3.89 ± 0.016	4.43	84.87	76.04	80.21
SurveyForge	4.85 ± 0.094	4.93 ± 0.004	4.90 ± 0.100	4.02 ± 0.027	4.02 ± 0.100	4.54	85.12	75.89	80.24
SurveyGo	4.96 ± 0.040	4.88 ± 0.008	4.92 ± 0.080	3.78 ± 0.112	3.88 ± 0.009	4.48	82.56	75.33	78.78
SURVEYX	$\mathbf{4.95 \pm 0.050}$	$\mathbf{4.91 \pm 0.090}$	$\mathbf{4.94 \pm 0.060}$	$\mathbf{4.10 \pm 0.116}$	$\mathbf{4.05 \pm 0.170}$	**4.59**	**85.23**	**78.12**	**81.52**
Human	5.00 ± 0.000	4.95 ± 0.050	5.00 ± 0.000	4.44 ± 0.237	4.38 ± 0.268	4.75	86.33	77.78	81.83

Table 2: SURVEYX performance on different LLMs.

Model	Coverage	Structure	Relevance	Synthesis	Critical Analysis	Avg.	Recall	Precison	F1
ChatGPT-5.2	4.95 ± 0.017	4.91 ± 0.090	4.94 ± 0.036	4.18 ± 0.247	4.09 ± 0.331	4.61	85.23	78.24	81.58
ClaudeSonnet4	4.96 ± 0.040	4.93 ± 0.070	4.94 ± 0.060	4.01 ± 0.456	4.12 ± 0.117	4.59	85.55	76.23	80.62
Gemini2.5pro	4.92 ± 0.080	4.89 ± 0.110	4.94 ± 0.060	4.13 ± 0.139	4.08 ± 0.357	4.59	84.32	77.88	80.97
Deepseek-V3.2	4.86 ± 0.140	4.55 ± 0.298	4.85 ± 0.150	3.83 ± 0.389	4.01 ± 0.309	4.42	84.87	76.04	80.21
Llama3.3-Instruct-70B	4.87 ± 0.130	4.59 ± 0.387	4.90 ± 0.100	4.02 ± 0.350	3.76 ± 0.138	4.43	84.87	76.32	80.37

Table 3: Evaluation results of reference relevance.

Model	IoU	$\text{Relevance}_{semantic}$	Relevance_{LLM}
Human	1	0.4455	0.9485
SURVEYX	0.55	0.4226	0.7689

the performance of human experts. Compared to other methods, SURVEYX shows a clear advantage across all metrics, with significant improvements in Structure (4.91) and Critical Analysis (4.05). Additionally, SURVEYX leads existing automated survey generation systems with an average score of 4.590, proving its outstanding performance in survey generation.

(2) The results of the citation quality experiments (Table 1) show that SURVEYX outperforms existing automated survey generation systems and closely approaches the performance of human experts in Citation Recall (85.23), Citation Precision (78.12), and F1 Score (81.52). Notably, in the Precision metric (78.12), SURVEYX even slightly surpasses human experts. This indicates that SURVEYX can significantly enhance the comprehensiveness and accuracy of citations in its generated surveys, ensuring that all statements are well-supported by literature while effectively reducing irrelevant citations, thereby increasing the reliability and credibility of the generated surveys.

(3) Table 2 demonstrates that our method exhibits strong robustness across different base models. This consistency indicates that the effectiveness of our approach does not rely on a specific base model, but rather stems from the model-agnostic design of the framework, enabling reliable survey generation across a wide range of large language models.

(4) The results of the reference relevance experiments (Table 3) indicate that SURVEYX approaches the performance of human experts in the $\text{Relevance}_{semantic}$ metric. However, there is a certain gap compared to human experts in the IoU and Relevance_{LLM} metrics.

(5) Table 4 reports the visual quality evaluation results across different base models. SurveyX consistently outperforms agent-based baselines built on GPT-4.1 and Gemini-3-Pro-Preview across all five dimensions, demonstrating the effectiveness of its structured figure generation pipeline. In particular, SurveyX achieves notable improvements in semantic correctness and contextual consistency, indicating stronger alignment between generated figures and their surrounding survey text.

Ablation results The results of the ablation experiments are shown in Table 5. Key findings are as follows:

(1) After ablating the retrieval algorithm, the Coverage (4.95→4.74) and Relevance (4.94→4.79) metrics showed the most significant decreases. This indicates that the improvement in reference quality brought by the retrieval algorithm is crucial for ensuring comprehensive content coverage and high topic relevance in the generated surveys.

(2) After ablating the AttributeTree method, the Structure (4.91→4.08), Synthesis (4.10→3.88), and Critical Analysis (4.05→3.93) metrics showed significant declines. Besides, the Recall (85.23→60.09), Precision (78.12→56.49), and F1 score (81.52→58.23) metrics also decreased notice-

Table 4: Evaluation results of visual elements in academic surveys.

Model	Semantic Correctness	Expressiveness & Information Density	Readability & Cognitive Load	Contextual Consistency	Reusability & Self-Containment	Avg.
gpt-4.1	3.83 ± 0.151	3.12 ± 0.069	4.54 ± 0.083	4.12 ± 0.128	2.89 ± 0.008	3.70
gemini-3-pro-preview	3.94 ± 0.102	2.88 ± 0.036	4.01 ± 0.011	3.95 ± 0.139	2.78 ± 0.044	3.51
SuveyX	4.33 ± 0.105	3.67 ± 0.096	4.52 ± 0.049	4.43 ± 0.075	2.94 ± 0.055	3.98
Human	4.67 ± 0.159	4.33 ± 0.133	4.64 ± 0.033	4.53 ± 0.194	4.67 ± 0.075	4.57

Table 5: Ablation study results for SurveyX with different components removed.

Ablation Object	Coverage	Structure	Relevance	Synthesis	Critical Analysis	Avg.	Recall	Precision	F1
Retrieval	4.74 ± 0.260	4.88 ± 0.120	4.79 ± 0.210	3.98 ± 0.237	4.02 ± 0.116	4.48	78.88	73.34	76.01
AttributeTree	4.84 ± 0.160	4.08 ± 0.046	4.89 ± 0.110	3.88 ± 0.455	3.93 ± 0.224	4.32	60.09	56.49	58.23
Outl. Optim.	4.90 ± 0.100	3.80 ± 0.470	4.91 ± 0.090	3.98 ± 0.038	4.02 ± 0.096	4.32	85.1	77.13	80.92
Rewriting	4.92 ± 0.080	4.89 ± 0.110	4.93 ± 0.070	4.00 ± 0.104	4.00 ± 0.359	4.55	55.37	54.95	55.16
No Ablation	$\mathbf{4.95 \pm 0.050}$	$\mathbf{4.91 \pm 0.090}$	$\mathbf{4.94 \pm 0.060}$	$\mathbf{4.10 \pm 0.374}$	$\mathbf{4.05 \pm 0.169}$	**4.59**	**85.23**	**78.12**	**81.52**

ably. This indicates that converting the content of the reference into an attribute tree enables the LLM to understand the core information better, thereby improving its ability to organize outlines, integrate information from different references, conduct in-depth analysis, and add references more accurately.

(3) After ablating the outline optimization method, the Structure metric (4.91→3.80) showed the most significant decline. This indicates that this method effectively enhances the logical coherence and structural clarity of the generated survey outline.

(4) After ablating the RAG-based rewriting module, the Recall (85.23→55.37), Precision (78.12→54.95), and F1 score (81.52→55.16) metrics declined significantly. This indicates that the RAG-based rewriting module can considerably enhance the citation quality of the generated survey by more accurately adding relevant citations and removing irrelevant ones.

5 Conclusion

In this paper, we present SURVEYX, a multimodal and context-efficient system for automated survey generation. SURVEYX effectively addresses issues in LLM-based automated survey generation, such as the lack of multimodal reasoning, inefficient utilization of context windows, and insufficient evaluation frameworks. Moreover, SURVEYX also overcomes shortcomings of existing automatic generation systems, including constrained retrieval strategies, limited support for all-domain survey generation, and the absence of dedicated evaluation frameworks for visual elements. Extensive experiments demonstrate that SURVEYX consistently outperforms existing automated survey generation systems across multiple evaluation dimensions, with performance approaching that of human experts. These results suggest that SURVEYX represents a meaningful step toward practical, scalable, and high-quality automated survey generation, offering valuable support for researchers navigating rapidly growing scientific literature.

Limitations

Despite its effectiveness, SURVEYX has several limitations. First, the construction of AttributeTree schemas depends on the LLM's understanding of the target survey topic, which may be less stable for emerging or rapidly evolving research areas. Second, the system relies on powerful LLMs and multimodal models, whose stochastic behavior and version updates may affect reproducibility. Third, while multimodal analysis enables figure-aware generation, the current visual understanding is limited to high-level semantics and may struggle with complex or low-quality figures. Finally, parts of the evaluation rely on LLM-based judgments, which may introduce subjective bias despite structured prompts.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Alexander Bick, Adam Blandin, and David J Deming. 2024. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi,

Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Yu Chao, Siyu Lin, Xiaorong Wang, Zhu Zhang, Zihan Zhou, Haoyu Wang, Shuo Wang, Jie Zhou, Zhiyuan Liu, and Maosong Sun. 2025. LLM×MapReduce-v3: Enabling interactive in-depth survey generation through a MCP-driven hierarchically modular agent system. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 688–695, Suzhou, China. Association for Computational Linguistics.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, and 1 others. 2025. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, and 1 others. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Vikram Gupta. 2023. Recent advancements in computer science: A comprehensive review of emerging technologies and innovations. In *International Journal for Research Publication and Seminar*, volume 14, pages 329–334.

Yucheng Hu and Yuxing Lu. 2024. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: A system for generating coherent survey articles for scientific topics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10438–10451, Bangkok, Thailand. Association for Computational Linguistics.

Darioush Kevian, Usman Syed, Xingang Guo, Aaron Havens, Geir Dullerud, Peter Seiler, Lianhui Qin, and Bin Hu. 2024. Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra. *arXiv preprint arXiv:2404.03647*.

Aida Kostikova, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger. 2025. Lllms: A data-driven survey of evolving research on limitations of large language models. *arXiv preprint arXiv:2505.19240*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source LLMs truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Chengwei Liu, Chong Wang, Jiayue Cao, Jingquan Ge, Kun Wang, Lyuye Zhang, Ming-Ming Cheng, Penghai Zhao, Tianlin Li, Xiaojun Jia, and 1 others. 2025a. A vision for auto research with llm agents. *arXiv preprint arXiv:2504.18765*.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, and 1 others. 2025b. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc.

Yixin Liu, Yonghui Wu, Denghui Zhang, and Lichao Sun. 2025c. Agentic autosurvey: Let llms survey llms. *Preprint*, arXiv:2509.18661.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2024. Train & constrain: Phonologically informed tongue-twister generation from topics and paraphrases. *arXiv preprint arXiv:2403.13901*.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2025. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2):1–32.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592, Boulder, Colorado. Association for Computational Linguistics.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2024. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Chen Tang, Tyler Loakman, and Chenghua Lin. 2024. A cross-attention augmented model for event-triggered context-aware story generation. *Computer Speech & Language*, 88:101662.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Haoyu Wang, Yujia Fu, Zhu Zhang, Shuo Wang, Zirui Ren, Xiaorong Wang, Zhili Li, Chaoqun He, Bo An, Zhiyuan Liu, and Maosong Sun. 2025. Llm×mapreduce-v2: Entropy-driven convolutional test-time scaling for generating long-form articles from extremely long resources. *Preprint*, arXiv:2504.05732.

Jie Wang, Chengzhi Zhang, Mengying Zhang, and Sanhong Deng. 2018. Citationas: A tool of automatic survey generation based on citation content. *J. Data Inf. Sci.*, 3(2):20–37.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: Large language models can automatically write surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siyi Wu, Chiaxin Liang, Ziqian Bi, Leyi Zhao, Tianyang Wang, Junhao Song, Yichao Zhang, Keyu Chen, Benji Peng, and Xinyuan Song. 2025. Autosurvey2: Empowering researchers with next level automated literature surveys. *Preprint*, arXiv:2510.26012.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. 2025. SURVEYFORGE : On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12444–12465, Vienna, Austria. Association for Computational Linguistics.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023. IAG: Induction-augmented generation

framework for answering reasoning questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025a. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025b. Attention heads of large language models. *Patterns*, page 101176.

Shupeng Zhong, Xinger Li, Shushan Jin, and Yang Yang. 2024. The solution for the pst-kdd-2024 oag-challenge. *arXiv preprint arXiv:2407.12827*.

A Evaluation details

A.1 Survey Topics

Table 6 presents the 20 survey topics used in the evaluation experiment, along with the titles of the corresponding survey written by human experts.

A.2 Evaluation Prompt

Prompts used for content evaluation are presented in figure A.1, A.2, A.3, A.4, A.5. Prompts used for citation quality are presented in figure A.6. Prompts used for reference papers are presented in figure A.7.

B Attribute Tree Templates

Attribute tree templates are shown in figure B.1, B.2, B.3, B.4.

C Bibliometrics Method

For each survey S, we manually extracted keywords (including synonyms, hypernyms, and related terms) to create a controlled vocabulary list V, and then computed:

$$B = \Big\{ p \;\Big|\; p \in R \cup Q, \max\big(\mathrm{CoCiteCos}(p, R), \mathrm{Coupling}(p, R)\big) \geq \tau \Big\}$$

where:

- R is the initial set of papers retrieved via Boolean search based on controlled vocabulary V.
- Q is the set of papers obtained through forward and backward citation chaining from R.
- $\mathrm{CoCiteCos}(p, R)$ is the average co-citation similarity between paper p and the papers in R.
- $\mathrm{Coupling}(p, R)$ is the average bibliographic coupling between paper p and the papers in R.
- τ is a threshold (set to 0.2) for selecting papers with high similarity.

The process for constructing B follows the traditional bibliometric method:

1. Perform a Boolean search based on V to retrieve an initial set R from open datasets OpenAlex and PubMed.
2. Conduct forward and backward citation chaining on R to expand the set to Q, with a depth of one hop and a time window of the last 5 years.
3. Select documents $q \in Q$ to add to R, based on the criteria:

   $$\max(\mathrm{CoCiteCos}(q, R), \mathrm{Coupling}(q, R)) \geq \tau,$$

   where $\tau = 0.2$
4. Filter R to retain only articles and reviews, with a **domain-normalized percentile score** of at least 50%, then we get B.

D Visual elements evaluation prompts.

The prompts used for visual elements evaluation is shown in figure D.2, D.3, D.4, D.5, D.6.

E App. Development and Case Study

To validate the effectiveness of the system, we developed the initial version of the product and launched it online . Within three months of its release, the platform attracted over 4,000 natural users for trial use, resulting in the generation of 12,012 survey articles. In this chapter, we will introduce the industrial-grade modules and tools used to deploy SURVEYX, and perform a statistical analysis of generated articles, combined with user behavior data, to better guide the direction of future system iterations and optimizations.

E.1 Application Development of SURVEYX

SURVEYX is an industrial-grade system for automated academic survey generation, built on a modular and scalable architecture. Beyond the core algorithms previously introduced, the deployment of SURVEYX in real-world scenarios involves a range of engineering techniques to ensure reliability, performance, and extensibility. As illustrated in Figure E.7, the system forms a complete pipeline supported by five key layers: application algorithms, system processing, data storage, data acquisition, and LLMs. To enhance user experience, we have developed a user-friendly and easy-to-navigate interface. It includes features such as a login system, email notifications, and user-customizable features, allowing users to fully customize the content of the generated surveys. A screenshot of part of the interface is shown in Figure B.1.

In the preprocessing stage, we developed a custom crawler system that can flexibly access the

Table 6: Survey Topics

Topic	Survey Title
NoSQL Database	A Survey of NoSQL Database Systems for Flexible and Scalable Data Management
Graph Databases	Graph Databases: A Survey on Models, Data Modeling, and Applications
Contemporary Data Management	A Survey of Temporal Databases Real-Time Databases and Data Management Systems
Vector Database	Vector Databases and Their Role in Modern Data Management and Retrieval: A Survey
Large Language Model Meets Database	A Survey on Large Language Model Integration with Databases
In-context Learning	A survey for in-context learning
LLMs for Recommendation	A Survey on Large Language Models for Recommendation
LLM-Generated Texts Detection	A Survey of Detecting LLM-Generated Texts
Explainability for LLMs	Explainability for Large Language Models
Evaluation of LLMs	A Survey on Evaluation of Large Language Models
LLMs-based Agents	A Survey on Large Language Model based Autonomous Agents
LLMs in Medicine	A Survey of Large Language Models in Medicine
Domain Specialization of LLMs	Domain Specialization as the Key to Make Large Language Models Disruptive
Challenges of LLMs in Education	Practical and Ethical Challenges of Large Language Models in Education
Alignment of LLMs	Aligning Large Language Models with Human
ChatGPT	A Survey on ChatGPT and Beyond
Instruction Tuning for LLMs	Instruction Tuning for Large Language Models
LLMs for Information Retrieval	Large Language Models for Information Retrieval
Safety in LLMs	Towards Safer Generative Language Models: Safety Risks, Evaluations, and Improvements
Chain of Thought	A Survey of Chain of Thought Reasoning
Hallucination in LLMs	A Survey on Hallucination in Large Language Models
Bias and Fairness in LLMs	Bias and Fairness in Large Language Models
Large Multi-Modal Language Models	Large-scale Multi-Modal Pre-trained Models
Acceleration for LLMs	A Survey on Model Compression and Acceleration for Pretrained Language Models
LLMs for Software Engineering	Large Language Models for Software Engineering

APIs of Google Scholar and Semantic Scholar to automatically retrieve references, and has also scraped all papers from the arXiv website. To handle raw PDFs, we deploy a PDF parsing module based on `minerU`, which converts the documents into structured markdown representations, enabling accurate downstream information extraction. Retrieved documents are stored in a structured format using MongoDB for metadata management. This preprocessing pipeline ensures a high-quality, well-organized reference corpus for subsequent content generation.

During Generation stage, SURVEYX employs Elasticsearch (ES) for efficient full-text search. To support large-scale intermediate data processing, Object Storage Service (OSS) is used for storing and accessing a vast collection of academic documents. Upon completing generation, all outputs are compiled into PDF documents via a LaTeX compilation module. Additionally, the system supports automatic watermark embedding to avoid risks of ethical violations. This entire pipeline is powered by a robust LLM layer, which integrates multiple state-of-the-art models such as GPT, Qwen, and DeepSeek.

Together, these stages and components enable SURVEYX to operate as a complete, automated, and academically aligned survey generation system.

E.2 Statistical Analysis of User Cases

To provide a good user experience, we designed a basic UI in SURVEYX. It includes fundamental features such as account login, task submission, task management, and email notifications.

User Characteristics Analysis To better understand our users, we collected usage data from six months and conducted a user survey, resulting in the user profile shown in Figure E.8. The word cloud highlights frequently mentioned keywords, such as "language," "data," "model," "system," "technology," "learning," and "intelligence," indicating a user focus on topics like NLP, data analysis, and machine learning. We also observed that most users hold doctoral (46.4%) or master's degrees (39.3%), with primary interests in science, engineering, and medicine. These findings underscore a strong demand for automated literature survey systems among researchers seeking efficient access to research trends and developments.

E.3 Statistical Analysis of Generation Cases

Production Case Analysis We conducted a statistical analysis of 12,012 genuine production samples provided by users. Table 7 presents a comprehensive overview of the specifics of these auto-

Evaluation prompt: content - coverage

Here is an academic survey about the topic "{topic}":

{content}

<instruction>
Please evaluate this survey about the topic {topic} based on the criterion above provided below, and give a score from 1 to 5 according to the score description:

Criterion Description: Coverage: Coverage assesses the extent to which the survey encapsulates all relevant aspects of the topic, ensuring comprehensive discussion on both central and peripheral topics.

Score 1 Description: The survey has very limited coverage, only touching on a small portion of the topic and lacking discussion on key areas.
Score 2 Description: The survey covers some parts of the topic but has noticeable omissions, with significant areas either underrepresented or missing.
Score 3 Description: The survey is generally comprehensive in coverage but still misses a few key points that are not fully discussed.
Score 4 Description: The survey covers most key areas of the topic comprehensively, with only very minor topics left out.
Score 5 Description: The survey comprehensively covers all key and peripheral topics, providing detailed discussions and extensive information.

Return the score without any other information:

Figure A.1: Content coverage prompt for evaluation.

Evaluation prompt: content - structure

Here is an academic survey about the topic "{topic}":

{content}

<instruction>
Please evaluate this survey about the topic {topic} based on the criterion above provided below, and give a score from 1 to 5 according to the score description:

Criterion Description: Structure: Structure evaluates the logical organization and coherence of sections and subsections, ensuring that they are logically connected.

Score 1 Description: The survey lacks logic, with no clear connections between sections, making it difficult to understand the overall framework.
Score 2 Description: The survey has weak logical flow with some content arranged in a disordered or unreasonable manner.
Score 3 Description: The survey has a generally reasonable logical structure, with most content arranged orderly, though some links and transitions could be improved such as repeated subsections.
Score 4 Description: The survey has good logical consistency, with content well arranged and natural transitions, only slightly rigid in a few parts.
Score 5 Description: The survey is tightly structured and logically clear, with all sections and content arranged most reasonably, and transitions between adajecent sections smooth without redundancy.

Return the score without any other information:

Figure A.2: Content structure prompt for evaluation.

Table 7: Production Statistics. Data comes from 810 survey papers generated by real users.

Metrics	Mean Value
PDF Pages per Paper	21.41
Paper Length (tokens)	9976.68
Figures Generated per Paper	3.52
Tables Generated per Paper	2.78
References Retrieved per Paper	174.94

matically generated survey papers. It is noteworthy that the average length of a review article's main text is 9.9k tokens, a figure comparable to the length of review articles manually written by domain experts, thereby confirming the potential of automatically generated reviews in terms of content comprehensiveness and depth. Furthermore, the average number of genuine references cited per review article is 174.94, a statistic that not only highlights the efficiency of the system's literature Retrieval Algorithm but also significantly alleviates the workload of researchers in the process of gathering and organizing literature. These findings suggest that the automatic survey paper generation system is capable not only of emulating the writing style of expert-level authors but also of

Evaluation prompt: content - relevance

Here is an academic survey about the topic "{topic}":

{content}

<instruction>
Please evaluate this survey about the topic {topic} based on the criterion above provided below, and give a score from 1 to 5 according to the score description:

Criterion Description: Relevance: Relevance measures how well the content of the survey aligns with the research topic and maintain a clear focus.

Score 1 Description: The content is outdated or unrelated to the field it purports to review, offering no alignment with the topic
Score 2 Description: The survey is somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
Score 3 Description: The survey is generally on topic, despite a few unrelated details.
Score 4 Description: The survey is mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
Score 5 Description: The survey is exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.

Return the score without any other information:

Figure A.3: Content relevance prompt for evaluation.

Evaluation prompt: content - synthesis

Here is an academic survey about the topic "{topic}":

{content}

<instruction>
Please evaluate this survey about the topic {topic} based on the criterion above provided below, and give a score from 1 to 5 according to the score description:

Criterion Description: Synthesis: Synthesis evaluates the ability to interconnect disparate studies, identify overarching patterns or contradictions, and construct a cohesive intellectual framework beyond individual summaries.

Score 1 Description: The survey is purely a collection of isolated study summaries with no attempt to connect ideas or identify broader trends.
Score 2 Description: The survey occasionally links studies but fails to synthesize them into meaningful patterns; connections are superficial.
Score 3 Description: The survey identifies some thematic relationships between studies but lacks a unified framework to explain their significance.
Score 4 Description: The survey synthesizes most studies into coherent themes or debates, though some connections remain underdeveloped.
Score 5 Description: The survey masterfully integrates studies into a novel framework, revealing latent trends, resolving contradictions, and proposing paradigm-shifting perspectives.

Return the score without any other information:

Figure A.4: Content synthesis prompt for evaluation.

Table 8: Computational Cost Analysis of SURVEYX (Mean Values, N=810)

Metrics	Mean Value
Processing Time (min)	44.3
Prompt Tokens	244,389.46
Completion Tokens	24,226.08

providing high-quality literature review services to users through its advanced retrieval algorithms, thus holding significant promise and research value in the development of research assistance tools.

Computational Cost Analysis We evaluated the system's computational cost (Table 8) and found that the average processing time per task is 44.3 minutes, substantially faster than the months typically needed by human experts. Compared to the AutoSurvey model, our system uses only 8% of its prompt tokens, thanks to the efficient AttributeTree method, which minimizes preprocessing overhead and reduces hallucination by focusing resources on content generation. Moreover, SURVEYX generates twice the number of completion tokens, indicating deeper content refinement and resulting in higher-quality survey articles. Overall, SURVEYX outperforms existing models in speed, resource efficiency, and output quality, making it highly valuable for research support.

Evaluation prompt: content - critical analysis

Here is an academic survey about the topic "{topic}":

{content}

<instruction>
Please evaluate this survey about the topic {topic} based on the criterion above provided below, and give a score from 1 to 5 according to the score description:

Criterion Description: Critical Analysis: Critical Analysis examines the depth of critique applied to existing studies, including identification of methodological limitations, theoretical inconsistencies, and research gaps.

Score 1 Description: The survey merely lists existing studies without any analytical commentary or critique.
Score 2 Description: The survey occasionally mentions limitations of studies but lacks systematic analysis or synthesis of gaps.
Score 3 Description: The survey provides sporadic critical evaluations of some studies, but the critique is shallow or inconsistent.
Score 4 Description: The survey systematically critiques most key studies and identifies research gaps, though some areas lack depth.
Score 5 Description: The survey demonstrates rigorous critical analysis of methodologies and theories, clearly maps research frontiers, and proposes novel directions based on synthesized gaps.

Return the score without any other information:

Figure A.5: Content critical analysis prompt for evaluation.

Evaluation prompt: citation

Claim:
{claim}

Source:
{source}

Claim:
{claim}

Is the Claim faithful to the Source?
A Claim is faithful to the Source if the core part in the Claim can be supported by the Source.\n
Only reply with 'Yes' or 'No':

Figure A.6: Citation prompt for evaluation.

Evaluation prompt: reference paper relevance

I will give you a paper abstract and a topic, you decide if this paper is relevant to this topic.
Only return "yes" or "no".
Here is the paper abstract:
{abstract}
Here is the topic:
{topic}
So, is this paper relevant to this topic?
Only output "yes" or "no".

Figure A.7: Judge relevance for reference paper.

Method Paper

1. ***Background***: *problem background, including previous methods, the need of new breakthrough.*
2. ***Problem***
   a. *definition: specific description of problem.*
   b. *key obstacle: main difficulty, main challenge.*
3. ***Idea***
   a. *intuition: idea was inspired by what.*
   b. *opinion: what's the idea*
   c. *innovation: what's the main difference compared to previous method, or where is the primary improvement.*
4. ***Method***
   a. *method definition: given the problem, what's the definiton of method.*
   b. *method description: in one sentence, describe the method.*
   c. *method steps: procedures of method.*
   d. *principle: why this method is effective.*
5. ***Experiments***
   a. *experiments setting: including dataset, baseline and so on.*
   b. *experiments progress: specific evaluation steps*
6. ***Conclusion***: *what's the conclusion of experiments/paper.*
7. ***Discussison***
   a. *advantage: what's the advantages of this paper.*
   b. *limitation: what's the disadvantages of this paper.*
   c. *future work: based on the adv and disadv, what and where can be improved in the future.*
8. ***Other Info***: *Is there any other info not mentioned above?*

Figure B.1: Method paper attribute tree.

Benchmark Paper

1. ***Background***: *problem background, including previous methods, the need of new breakthrough.*
2. ***Problem***
   a. *definition: specific description of problem.*
   b. *key obstacle: main difficulty, main challenge.*
3. ***Idea***
   a. *intuition: idea was inspired by what.*
   b. *opinion: what's the idea*
   c. *innovation: what's the main difference compared to previous method, or where is the primary improvement.*
4. ***Dataset***
   a. *source: how this dataset generated.*
   b. *desc: a description of dataset, eg. scale, accessibility, diversity, quality and so on.*
   c. *content: the specific content of dataset.*
5. ***Metrics***
   a. *aspect: what aspects of the model performance are being measured.*
   b. *principle: the underlying principles of the metrics used.*
   c. *procedure: how to evaluate performance of a model?*
6. ***Experiments***
   a. *model: which models used in the experiments.*
   b. *procedure: how the experiment conducted, also as the experiments settings.*
   c. *result: what's the result of experiments. How models perform on this benchmark, if is statistically significant.*
   d. *variability: how was the variability in the results accounted for?*
7. ***Conclusion***: *what's the conclusion of experiments/paper.*
8. ***Discussion***
   a. *advantage: what's the advantages of this paper.*
   b. *limitation: what's the disadvantages of this paper.*
   c. *future work: based on the adv and disadv, what and where can be improved in the future.*
9. ***Other Info***: *Is there any other info not mentioned above?*

Figure B.2: Benchmark paper attribute tree.

Theory Paper
1. ***Background***: *problem background, including previous methods, the need of new breakthrough.* 2. ***Problem*** a. *definition: specific description of problem.* b. *key obstacle: main difficulty, main challenge.* 3. ***Idea*** a. *intuition: idea was inspired by what.* b. *opinion: what's the idea* c. *innovation: what's the main difference compared to previous method, or where is the primary improvement.* 4. ***Theory*** a. *perspective: the perspective of theory and the architecture of the perspective.* b. *opinion: the view or assumption about a problem.* c. *proof: the proof or derivation of a theory.* 5. ***Experiments*** a. *experiments setting: including dataset, baseline and so on.* b. *experiments progress: specific evaluation steps* 6. ***Conclusion***: *what's the conclusion of experiments/paper.* 7. ***Discussison*** a. *advantage: what's the advantages of this paper.* b. *limitation: what's the disadvantages of this paper.* c. *future work: based on the adv and disadv, what and where can be improved in the future.* 8. ***Other Info***: *Is there any other info not mentioned above?*

Figure B.3: Theory paper attribute tree.

Survey Paper
1. ***Background:*** a. *purpose: the purpose of this survey.* b. *scope: topics covered and what not covered by this survey.* 2. ***Problem*** a. *definition: specific description of problem.* b. *key obstacle: main difficulty, main challenge.* 3. ***Architecture:*** a. *perspective: the new view this survey proposed, which could summarize each method into a fields/stages.* b. *fields/stages: what fields or stages the survey summarize the current methods to?* 4. ***Conclusion:*** *what's the conclusion of experiments/paper.* a. *comparisons: comparisons between each research or methods.* b. *results: the main conclusion or discovery.* 5. ***Discussion*** a. *advantage: what's the advantages of existing research.* b. *limitation: what's the disadvantages of existing research.* c. *gaps: gaps existed in the current researches.* d. *future work/trends: based on the adv and disadv, what and where can be improved in the future.* 6. ***Other Info:*** *is there any other info not mentioned above? List them in json format.*

Figure B.4: Survey paper attribute tree.

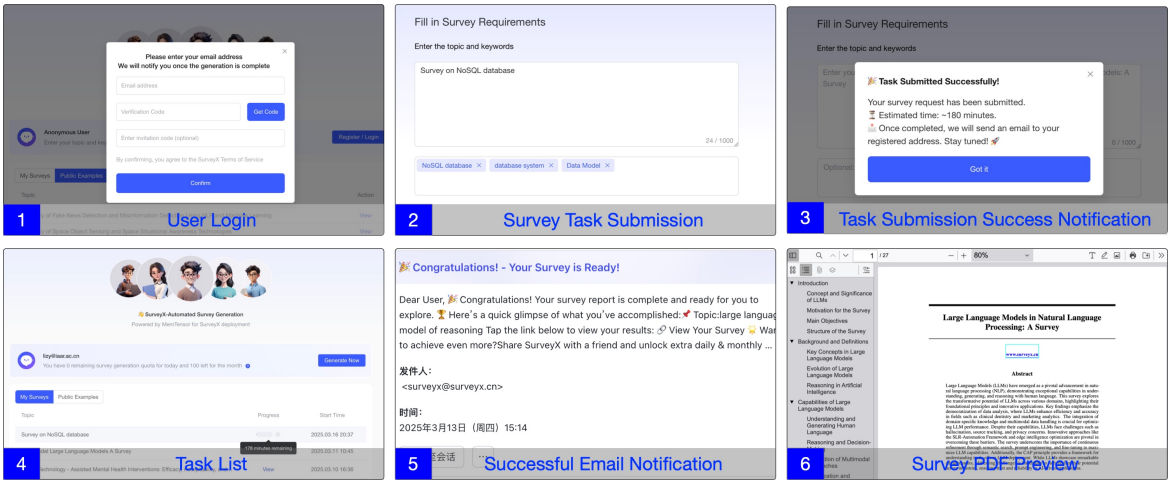


Figure B.1: Screenshot of User Interface.

Evaluation prompt: visual - semantic correctness

You are an expert reviewer evaluating the SEMANTIC CORRECTNESS of an academic figure.
Semantic correctness measures whether the visual elements (modules, symbols, arrows, flows, and relationships) in the figure accurately and faithfully represent the concepts described in the provided text, without hallucination or conceptual errors.
Assign a score from 1 to 5 according to the following criteria:
Score 5 (Excellent):
- All components, concepts, and relationships in the figure precisely match the textual description.
- Causal, hierarchical, or sequential relations are correct and unambiguous.
- No hallucinated, incorrect, or undefined elements are present.
- The figure can be considered semantically authoritative.
Score 4 (Good):
- Core concepts and relationships are correct.
- Minor abstraction differences or non-critical simplifications exist.
- These issues do not affect the overall semantic validity.
Score 3 (Fair):
- The overall semantic intent is correct.
- Some relationships are vague, oversimplified, or partially misrepresented.
- Readers must rely on the text to avoid misunderstanding.
Score 2 (Poor):
- Multiple important relationships or components are incorrect or misleading.
- Conceptual confusion is evident.
- Semantic reliability is low.
Score 1 (Unacceptable):
- The figure fundamentally contradicts the text.
- Severe hallucinations or false concepts are present.
- The figure is unusable for academic communication.
Here is the context of the academic figure:
{context}
Return the score and a brief justification.

Figure D.2: Visual elements evaluation prompt for semantic correctness.

Evaluation prompt: visual - expressiveness & information density

You are evaluating the INFORMATION DENSITY AND EXPRESSIVENESS of an academic figure.
This dimension assesses how effectively the figure conveys the core information from the text within limited visual space, balancing informativeness and simplicity.
Assign a score from 1 to 5 using the following rubric:
Score 5 (Excellent):
- The figure covers most key points from the text.
- Information is well-structured and efficiently encoded.
- No significant redundancy or unnecessary elements exist.
- The figure substantially reduces the need for textual explanation.
Score 4 (Good):
- The main ideas are clearly conveyed.
- Minor redundancy or slight information imbalance is present.
- Overall communication remains efficient.
Score 3 (Fair):
- Some important information is missing or over-compressed.
- The figure requires frequent reference to the text.
- Information efficiency is moderate.
Score 2 (Poor):
- The figure is either information-poor or overly cluttered.
- It fails to effectively summarize the textual content.
- Visual space is inefficiently used.
Score 1 (Unacceptable):
- The figure conveys little to no meaningful information.
- Structure is confusing or incoherent.
- It provides no advantage over text.
Here is the context of the academic figure:
{context}
Return the score and a brief justification.

Figure D.3: Visual elements evaluation prompt for expressiveness & information density.

Evaluation prompt: visual - readability & cognitive load

You are evaluating the READABILITY AND COGNITIVE LOAD of an academic figure.
This dimension assesses how easily a reader can understand the figure, considering visual hierarchy, layout clarity, and mental effort required
Assign a score from 1 to 5 according to the following rubric:
Score 5 (Excellent):
- Visual hierarchy is clear and well-organized.
- Reading flow is intuitive (e.g., left-to-right or top-to-bottom).
- The main idea can be grasped within a few seconds.
- Cognitive load is minimal.
Score 4 (Good):
- Overall structure is clear.
- Some local regions require extra attention.
- Cognitive effort is low to moderate.
Score 3 (Fair):
- The figure is understandable but not immediately intuitive.
- Readers must invest noticeable effort to parse the structure.
- Cognitive load is moderate.
Score 2 (Poor):
- Visual hierarchy is weak or confusing.
- Layout causes frequent backtracking or confusion.
- Cognitive load is high.
Score 1 (Unacceptable):
- The figure is extremely difficult or impossible to understand.
- Visual clutter or disorganization dominates.
- Cognitive load is excessive.
- It provides no advantage over text.
Here is the context of the academic figure:
{context}
Return the score and a brief justification.

Figure D.4: Visual elements evaluation prompt for readability & cognitive load.

Evaluation prompt: visual - contextual consistency

You are evaluating the CONTEXTUAL CONSISTENCY of an academic figure.
This dimension measures how well the figure aligns with its surrounding paper context, including terminology, narrative role, and referenced discussion.
Assign a score from 1 to 5 using the following criteria:
Score 5 (Excellent):
- The figure perfectly matches the terminology and focus of the text.
- It clearly supports the specific argument or section where it appears.
- Figure and text strongly complement each other.
Score 4 (Good):
- The figure aligns well with the paper context.
- Minor mismatches in emphasis or naming exist.
- Overall coherence is strong.
Score 3 (Fair):
- The figure is generally relevant but feels somewhat detached.
- Its role in the narrative is not fully clear.
- Contextual alignment is moderate.
Score 2 (Poor):
- The connection between the figure and the text is weak.
- It is unclear why the figure appears in this context.
- Terminology or focus diverges noticeably.
Score 1 (Unacceptable):
- The figure is inconsistent with the paper context.
- It does not support the surrounding discussion.
- Contextual mismatch is severe.
Here is the context of the academic figure:
{context}
Return the score and a brief justification.

Figure D.5: Visual elements evaluation prompt for contextual consistency.

Evaluation prompt: visual - reusability & self-containment

You are evaluating the REUSABILITY AND SELF-CONTAINMENT of an academic figure.
This dimension assesses whether the figure can be understood and reused independently of the full paper, such as in surveys, presentations, or citations.
Assign a score from 1 to 5 according to the following rubric:
Score 5 (Excellent):
- The figure is largely self-explanatory without relying on the full text.
- Labels and structure provide sufficient context.
- It has high potential for reuse in other academic works.
Score 4 (Good):
- The figure is understandable with minimal additional explanation.
- Reusability is high with slight context assumptions.
Score 3 (Fair):
- The figure requires substantial reference to the main text.
- Independent reuse is limited.
Score 2 (Poor):
- The figure heavily depends on the surrounding paper.
- It is difficult to interpret in isolation.
- Reusability is low.
Score 1 (Unacceptable):
- The figure cannot be understood without extensive external context.
- It has no practical reuse value.
Here is the context of the academic figure:
{context}
Return the score and a brief justification.

Figure D.6: Visual elements evaluation prompt for reusability & self-containment.

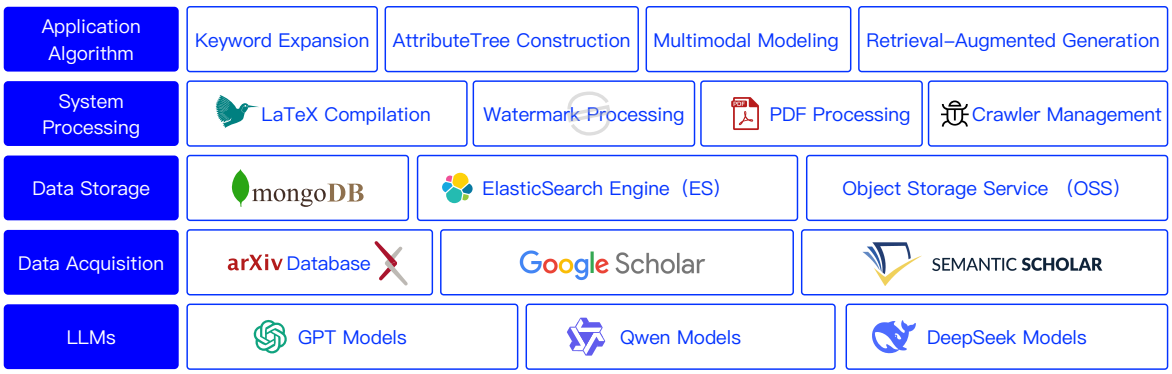
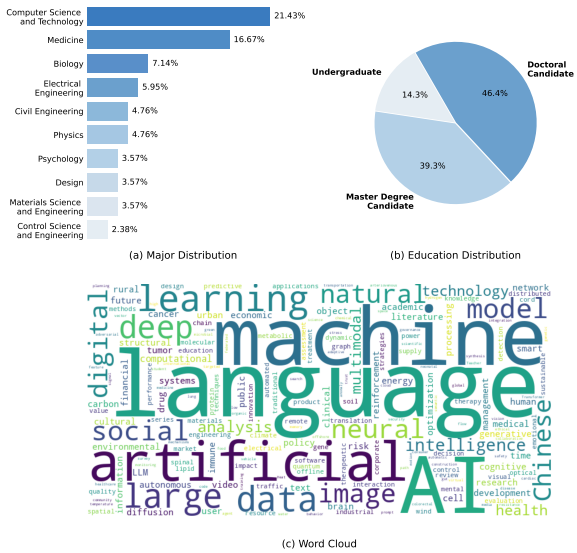


Figure E.7: System framework of SURVEYX.



(a) Major Distribution (b) Education Distribution

(c) Word Cloud

Figure E.8: User portrait.