# SurveyLens: A Discipline-Aware Benchmark for Automatic Survey Generation

Anonymous ACL submission


## Abstract

Automatic Survey Generation (ASG) aims to produce comprehensive literature surveys by retrieving, organizing, and synthesizing academic papers. Despite rapid progress in specialized ASG frameworks and Deep Research agents, existing evaluations largely center on Computer Science or rely on generic criteria, leaving it unclear whether current systems satisfy the survey standards of diverse disciplines. We introduce **SurveyLens**, the first discipline-aware ASG benchmark. SurveyLens comprises SurveyLens-1k, a curated dataset of 1,000 human-written surveys across 10 disciplines, and a dual-lens framework that combines discipline-aware rubric scoring with reference-based alignment to human-written surveys. Evaluating 11 state-of-the-art systems across vanilla LLMs, ASG systems, and Deep Research agents (Figure 1), we find that Deep Research agents are the only paradigm robust across all 10 disciplines, ASG systems lead on structural planning, and all paradigms remain weak on reference quality, providing practical guidance for discipline-specific tool selection and future ASG design.[1]


## 1 Introduction

The rapid growth of scientific literature has made it increasingly difficult for researchers to keep up with new developments through manual literature review. Automatic Survey Generation (ASG) addresses this challenge by combining Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to retrieve relevant papers and synthesize them into survey-style narratives. Existing ASG systems have evolved from pipeline-based approaches that streamline discrete literature review tasks (Wang et al., 2024c; Liang et al., 2025; Wang et al., 2025) to human-in-the-loop paradigms enabling iterative refinement (Wen et al., 2025; Chao et al., 2025; Zhang et al., 2025a), and finally to multi-agent systems that simulate collaborative survey workflows (Liu et al., 2025; Shi et al., 2025). In parallel, commercial Deep Research products such as Gemini Deep Research (Google, 2025) and Qwen Deep Research (Alibaba Qwen Team, 2025) have emerged as general-purpose alternatives.

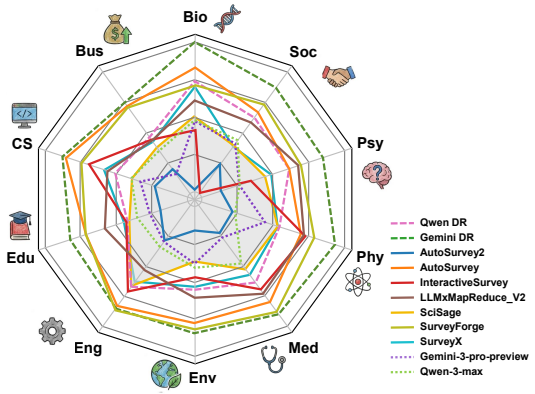


Figure 1: Performance comparison of vanilla LLMs, ASG systems, and deep research agents in generating surveys across 10 research disciplines.

Despite this progress, ASG evaluation remains inadequate along two axes: *what* surveys are evaluated on and *how* they are evaluated. On the data side, existing benchmarks (Kasanishi et al., 2023; Wang et al., 2025; Yan et al., 2025; Shi et al., 2025; Sun et al., 2025; Su et al., 2026) are predominantly restricted to Computer Science (CS) and overlook how writing conventions diverge across fields: a physics survey is anchored in equations and derivations, while a sociology review proceeds through qualitative narrative synthesis. General discipline-aware LLM benchmarks such as SuperGPQA (Team et al., 2025) do not close this gap either, as they probe atomic knowledge through Question Answering (QA) rather than the planning and coherent narrative construction required

[1] Dataset, code, results, and the annotation platform are released at https://anonymous.4open.science/r/SurveyLens-93CB.

| Assessment Methods | Disciplines | #Surveys | # Systems | Assessment Components | | | | | | Human Eval. | Discipline-Aware Assessment |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Outline | | Content | | Reference | | | |
| | | | | Ref. | Free | Ref. | Free | Ref. | Free | | |
| SciReviewGen (Kasanishi et al., 2023) | CS Only | 10K | 6 | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| LLM×MapReduce-V2 (Wang et al., 2025) | CS Only | 384 | 4 | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ |
| SurveyForge (Yan et al., 2025) | CS Only | 100 | 3 | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ |
| SciSage (Shi et al., 2025) | CS Only | 46 | 4 | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ |
| SurveyGen (Bao et al., 2025) | 4 Discip. | 4.2K | 3 | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| SurveyBench (Sun et al., 2025) | CS Only | 4.9K | 4 | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| SurveyEval (Zhao et al., 2025) | 7 Discip. | N/A | 7 | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| SurGE (Su et al., 2026) | CS Only | 205 | 3 | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ |
| **SurveyLens (Ours)** | **10 Discip.** | **1,000** | **11** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of **SurveyLens** with existing ASG assessment methods.

for survey writing. On the methodology side, semantic metrics (Kasanishi et al., 2023) such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) fail to capture the structural diversity of survey writing, while recent approaches (Wang et al., 2024c, 2025) based on LLM-as-a-judge (Zheng et al., 2023) rely on generic criteria that lack interpretability and disciplinary grounding.

To bridge these gaps, we introduce **SurveyLens**, a discipline-aware benchmark for ASG. On the data side, we construct *SurveyLens-1k*, a dataset of 1,000 human-written surveys evenly spanning 10 academic disciplines. On the methodology side, we design a dual-lens framework: (i) *Discipline-Aware Rubric Evaluation* (reference-free), which uses an LLM-as-a-judge with field-specific rubrics distilled bottom-up from SurveyLens-1k and aspect weights calibrated to expert preferences via Bradley-Terry modeling; and (ii) *Human-Reference Alignment Evaluation* (reference-based), which compares against paired human-written surveys through two metrics, Redundancy-Aware Alignment F1 (strict one-to-one Hungarian matching with a redundancy penalty, targeting paragraph bloat) and Thresholded MaxSim (margin-based max-similarity without the matching constraint, quantifying the degree of alignment).

Our analysis reveals that no single paradigm dominates: Deep Research agents are the only paradigm robust across all 10 disciplines, ASG systems and DR agents split along a structure-vs-content axis, and reference quality remains a universal bottleneck. We summarize our contributions as follows:

- **High-Quality Survey Dataset.** We construct SurveyLens-1k, a curated dataset of 1,000 human-written surveys evenly distributed across 10 academic disciplines, with each survey decomposed into outline, content, and references for component-level evaluation.
- **Discipline-Aware ASG Evaluation.** We propose a dual-lens framework that combines reference-free discipline-aware rubrics with human-reference alignment metrics to assess both disciplinary writing norms and coverage of human-written surveys, and validate it against judgments from 30 PhD-level domain experts across all 10 disciplines.
- **Cross-Paradigm Empirical Analysis.** We benchmark 11 ASG methods spanning vanilla LLMs, specialized ASG systems, and Deep Research Agents, and translate the resulting paradigm-level patterns into actionable guidance for discipline-aware system selection.

## 2 Related Work

Recent ASG systems span pipeline-based approaches (Wang et al., 2024c; Liang et al., 2025; Yan et al., 2025; Wang et al., 2025; Wu et al., 2025), human-in-the-loop paradigms (Wen et al., 2025; Chao et al., 2025; Zhang et al., 2025a), multi-agent frameworks (Shi et al., 2025; Liu et al., 2025; Nguyen et al., 2025), and commercial Deep Research agents (OpenAI, 2025; Google, 2025; Alibaba Qwen Team, 2025), yet their training corpora, retrieval pools, and self-reported benchmarks are overwhelmingly CS-centered, leaving open whether architectural conclusions transfer across disciplines.

Prior ASG assessment spans n-gram overlap (Kasanishi et al., 2023), LLM-as-a-judge content frameworks (Wang et al., 2024c; Liang et al., 2025), and structural metrics (Zhu et al., 2023; Lai et al., 2024; Bao et al., 2025; Su et al., 2026), with recent benchmarks (Wang et al., 2025; Yan et al., 2025; Su et al., 2026; Sun et al., 2025; Bao et al., 2025; Zhao et al., 2025; Guo et al., 2026) expanding scale and coverage. Two gaps persist: (i) they rely on discipline-agnostic criteria

that overlook field-specific survey-writing conventions; (ii) human validation, when reported, draws exclusively on CS experts, leaving non-CS rubrics empirically unverified. As summarized in Table 1, prior benchmarks typically expand along a single axis such as scale, system coverage, or metric type, while leaving discipline-aware assessment incomplete across survey components and evaluation settings. SurveyLens is the only benchmark covering all six (Outline/Content/Reference)×(reference-free/reference-based) cells, while also spanning 10 disciplines, 11 ASG baselines, and discipline-aware human evaluation.

## 3 SurveyLens

In this section, we introduce SurveyLens, covering (a) the construction of SurveyLens-1k, (b) Structured Survey Representation, and (c) the dual-lens evaluation framework.

### 3.1 SurveyLens-1k Construction

To ensure robust field-aware assessment of ASG methods, we constructed **SurveyLens-1k**, a large-scale dataset capturing the structural and content variances inherent to different fields, comprising 1,000 high-quality, human-written survey papers across 10 academic disciplines, with 100 surveys per field, as shown in Figure 3. Papers are sourced from Semantic Scholar via dual citation-based ranking (raw and influence-weighted) and validated through hybrid LLM-plus-human filtering; full construction details are in Appendix B.1.

Per-paper structural statistics reveal substantial cross-disciplinary variation (Figure 2). Physics is the most technically dense field, leading on equations (411.4/paper), word count (23.4k), figures, and references. Tables are the only axis where CS overtakes Physics (8.7 vs. 3.1/paper), reflecting CS's reliance on tabular comparisons. Variance reaches up to an order of magnitude across the corpus (e.g., $\sim 16\times$ on equations); the full per-discipline breakdown is given in Table 8 (Appendix B.2). These systematic differences motivate our field-aware evaluation framework as well.

### 3.2 Structured Survey Representation

To enable component-wise evaluation (§3.3), we adopt a unified *Structured Survey Representation* for both human-written references in SurveyLens-1k and ASG system outputs, formally defined as

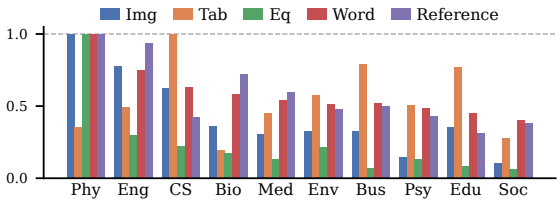


Figure 2: Cross-disciplinary variance in SurveyLens-1k: per-paper averages of figures (Img), tables (Tab), equations (Eq), word count (Word), and references, each normalized to its per-metric maximum.

the triplet

$$\mathcal{S} = (\mathcal{O}, \mathcal{C}, \mathcal{R}), \tag{1}$$

where $\mathcal{O}$ denotes the hierarchical **outline** (i.e., nested section titles), $\mathcal{C}$ represents the **content** (including textual paragraphs, figures, tables, formulas, and in-text citations), and $\mathcal{R}$ is the **reference list** (containing bibliographic metadata such as titles and publication venues). We obtain $\mathcal{S}$ from raw PDFs via a four-stage pipeline: MinerU (Wang et al., 2024a) PDF parsing, rule-based hierarchy and reference extraction, LLM-based normalization, and human verification (detailed in Appendix B.3).

### 3.3 Dual-Lens Evaluation Framework

Each component of $\mathcal{S}$ is scored by a reference-free rubric lens and a reference-based alignment lens.

#### 3.3.1 Discipline-Aware Rubric Evaluation

Existing assessment methods apply universal criteria and overlook field-specific writing conventions. We therefore distill universal evaluation aspects from existing methods (Appendix A.1, Table 7; aggregation prompt in Figure 10), instantiate them into *discipline-aware rubrics* using SurveyLens-1k, and learn *preference-aligned weights* from expert judgments to calibrate the relative importance of indicators.

**Discipline-Aware Rubric Generation.** For each component $c \in \{\mathcal{O}, \mathcal{C}, \mathcal{R}\}$, we define a set of aspects $\mathcal{A}_c$ (synthesized from existing assessments) and construct rubrics via a two-stage process:

$$a \in \mathcal{A}_c \xrightarrow[\text{per survey}]{\text{expand}} \mathcal{K}'_{d,c,a} \xrightarrow[\text{per discipline}]{\text{merge}} \mathcal{K}_{d,c,a}. \tag{2}$$

The expand stage instantiates each aspect for human-written surveys in discipline $d$, while the merge stage consolidates them into shared rubric

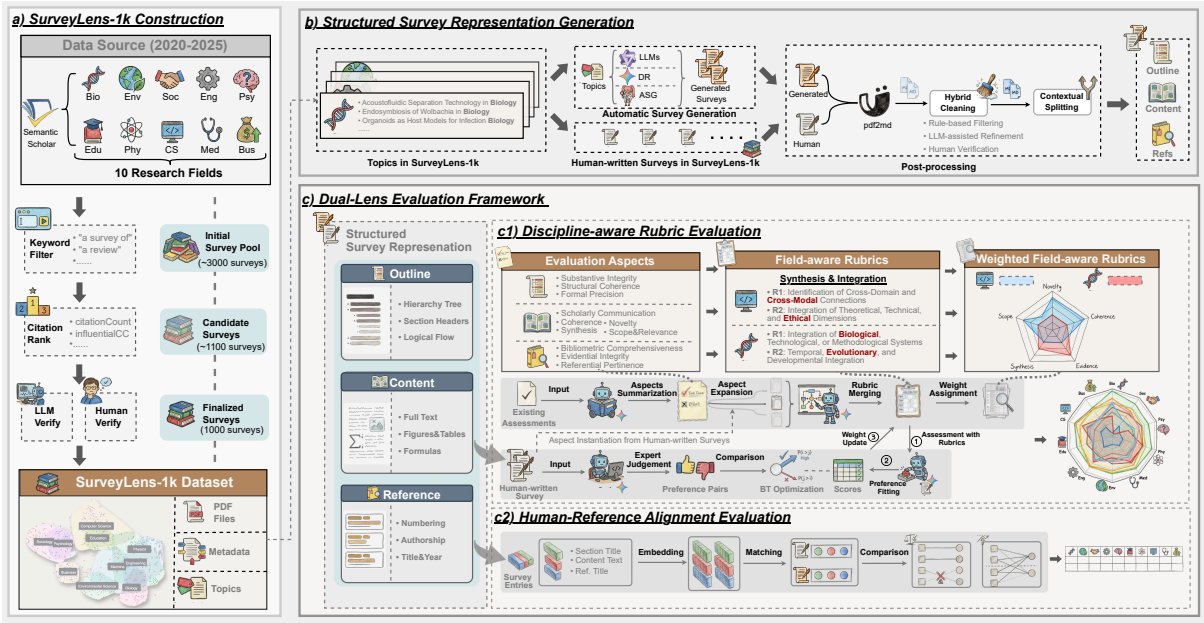


Figure 3: Overview of the proposed SurveyLens framework.

items $\mathcal{K}_{d,c,a}$. Prompts are shown in Figures 11 and 12.

**Expert-Preference-Aligned Weighting.** We adopt the Bradley-Terry (BT) model to convert pairwise preferences into per-aspect weights, with preferences obtained from an LLM proxy (*Gemini-3-Pro*, double round-robin with position swapping; validated against genuine expert judgments in §4.3) (Zheng et al., 2023; Wang et al., 2024b). Given normalized aspect scores $\bar{x}_{i,k} = x_{i,k}/x_{\max} \in [0.2, 1]$ ($x_{\max} = 5$, judge scores satisfy $x_{i,k} \geq 1$) and pairwise preferences $\mathcal{P} = \{(s_i, s_j) : s_i \succ s_j\}$, we estimate weights $\mathbf{w}$ by maximizing:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \geq \mathbf{0}} \sum_{\mathcal{P}} \log \sigma\big(\mathbf{w}^\top (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)\big) - \alpha \|\mathbf{w}\|_2^2 \quad (3)$$

We solve Eq. (3) per $(d, c)$ pair and normalize to $\tilde{w}_{c,k} = \hat{w}_{c,k} / \sum_{k'} \hat{w}_{c,k'}$. The final compliance score is $\mathcal{S}^{\text{RUB}}_{d,c}(s_i) = 5 \sum_k \tilde{w}_{c,k} \bar{x}_{i,k} \in [1, 5]$, expressed on a 1–5 scale (Appendix A.4). The resulting discipline-aware rubric is validated against judgments from 30 PhD-level domain experts spanning all 10 disciplines (§4.3; Appendix C.1, C.2).

### 3.3.2 Human-Reference Alignment Evaluation

We use each paired human-written survey as one valid reference among many for the same topic; this lens complements rubric scoring by measuring local semantic correspondence. Conventional reference-based metrics such as ROUGE and embedding-based F1 suffer from two limitations: (1) they rely on many-to-many matching that ignores redundancy in generated content, and (2) they apply uniform thresholds that do not account for component-specific similarity distributions (Bao et al., 2025; Zhao et al., 2025). To overcome these issues, we propose two entry-level semantic alignment scores adapted for long-form survey evaluation: RA-AlignF1 (Redundancy-Aware Alignment F1) and $\tau$-MaxSim (Thresholded MaxSim). RA-AlignF1 follows the precision–recall–F1 tradition of semantic matching metrics such as BERTScore (Zhang et al., 2020), additionally enforcing one-to-one Hungarian assignment and applying a redundancy penalty to target the paragraph-bloat failure mode of ASG systems. $\tau$-MaxSim adopts a MaxSim-style operator (Khattab and Zaharia, 2020) at the entry level, introducing a component-specific threshold margin that quantifies the *degree* of alignment rather than a binary threshold decision.

Instead of taking the full component as the input, we split the components into entries for fine-grained comparison. For each component-topic pair $(c, t)$, we define the generated entry set $\mathcal{E}_{c,t}$ and ground-truth entry set $\mathcal{G}_{c,t}$ as:

$$\begin{aligned} \mathcal{E}_{c,t} &= \{e^{(1)}_{c,t}, e^{(2)}_{c,t}, \ldots, e^{(m_{c,t})}_{c,t}\}, \\ \mathcal{G}_{c,t} &= \{g^{(1)}_{c,t}, g^{(2)}_{c,t}, \ldots, g^{(n_{c,t})}_{c,t}\}, \end{aligned} \quad (4)$$

where $m_{c,t}$ and $n_{c,t}$ denote the number of en-

tries for the generated and ground-truth human-written surveys in SurveyLens-1k, respectively. Concretely, the similarity-bearing content of each entry is the section title for $\mathcal{O}$, the section body for $\mathcal{C}$, and the reference title for $\mathcal{R}$; structural metadata (hierarchical level, heading, additional bibliographic fields) is retained alongside each entry but does not enter the similarity computation.

**Redundancy-Aware Alignment F1 (RA-AlignF1).** RA-AlignF1 measures overall coverage and faithfulness while penalizing repetition. It enforces a strict one-to-one mapping $\mathcal{M}$ between $\mathcal{E}_{c,t}$ and $\mathcal{G}_{c,t}$ via the Hungarian algorithm on a thresholded similarity margin (full formulation in Appendix A.2). To target the *paragraph bloat / information dumping* failure mode where systems pad content without adding coverage, RA-AlignF1 further down-weights each generated entry by a redundancy weight $\omega(e)$ based on its maximum internal similarity to other generated entries. The Precision ($P$) and Recall ($R$) are formulated as:

$$\begin{aligned} P &= \frac{1}{|\mathcal{E}_{c,t}|} \sum_{(e,g)\in\mathcal{M}} \omega(e) \cdot \mathbb{I}[\mathrm{sim}(e,g) \geq \tau_c], \\ R &= \frac{1}{|\mathcal{G}_{c,t}|} \sum_{(e,g)\in\mathcal{M}} \mathbb{I}[\mathrm{sim}(e,g) \geq \tau_c], \end{aligned} \tag{5}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\omega(e)$ is the redundancy penalty, defined as

$$\omega(e) = \exp\big(-\lambda \cdot \max_{e' \in \mathcal{E}_{c,t}\setminus\{e\}} \mathrm{sim}(e,e')\big), \tag{6}$$

with hyperparameter $\lambda > 0$. The redundancy weight $\omega(e)$ applies only to the precision term because paragraph bloat is a generation-side failure mode; the ground-truth side $\mathcal{G}_{c,t}$ is human-curated and assumed non-redundant by construction. The final RA-AlignF1 is the harmonic mean of $P$ and $R$:

$$\text{RA-AlignF1} = \frac{2PR}{P+R}. \tag{7}$$

**Thresholded MaxSim ($\tau$-MaxSim).** With cross-topic unions $\mathcal{E}_c = \bigcup_t \mathcal{E}_{c,t}$ and $\mathcal{G}_c = \bigcup_t \mathcal{G}_{c,t}$, $\tau$-MaxSim replaces RA-AlignF1's binary indicator with a similarity margin:

$$\tau\text{-MaxSim} = \frac{1}{|\mathcal{E}_c|} \sum_{e\in\mathcal{E}_c} \max\Big\{0,\ \max_{g\in\mathcal{G}_c} \mathrm{sim}(e,g) - \tau_c\Big\}. \tag{8}$$

Unlike binary precision, the margin quantifies the *degree* of alignment, rewarding strong matches over those that barely surpass the threshold.

# 4 Experiments and Result Analysis

We apply SurveyLens to benchmark 11 representative systems across three paradigms—*ASG systems*, *Vanilla LLMs*, and *Deep Research Agents* (DR)—on their ability to generate discipline-appropriate surveys.

## 4.1 Experiment Setup

**Evaluated Systems.** We first select seven representative **ASG systems**: AutoSurvey (Wang et al., 2024c), SurveyForge (Yan et al., 2025), AutoSurvey2 (Wu et al., 2025), InteractiveSurvey (Wen et al., 2025), LLM×MapReduce-V2 (Wang et al., 2025), SurveyX (Liang et al., 2025), and SciSage (Shi et al., 2025). To provide a comprehensive comparison, we additionally evaluate two **Vanilla LLMs** (Qwen3-Max (Yang et al., 2025), Gemini-3-Pro) alongside their corresponding **Deep Research Agents** (Qwen DR (Alibaba Qwen Team, 2025), Gemini DR (Google, 2025)).

**Comparison Settings.** Each system receives 100 survey topics (10 per discipline), producing 1,062 valid surveys (96.5% yield; per-system breakdown and stability analysis in Appendix D.3). To ensure fair comparison within open-source ASG systems, we run all of them on a unified *Qwen3-30B-A3B* (Yang et al., 2025) backbone, which also serves as our primary LLM judge; *Gemini-3-Pro* is used for rubric aspect synthesis (Table 7) and as a secondary cross-judge (Appendix D.1), and *Qwen3-Embedding-8B* (Zhang et al., 2025b) for reference-based metrics. Each (system, topic) cell is averaged over three trials; per-cell standard deviations are small (mean $0.015/0.007/0.028$ for Outline/Content/Reference; max $< 0.05$). Vanilla LLMs and DR agents are evaluated using their officially deployed configurations(Appendix D.4); a leakage-controlled re-analysis is provided in Appendix D.2.

## 4.2 Results and Analysis

We present our experimental results and analysis from two complementary perspectives: *discipline-aware rubric evaluation* (§4.2.1) and *human-reference alignment analysis* (§4.2.2).

### 4.2.1 Discipline-Aware Rubric Evaluation

Based on discipline-aware rubrics, we calculate scores for each ASG method at the discipline level,

| System | Edu ↑ | Med ↑ | Bio ↑ | Soc ↑ | Phys ↑ | Psy ↑ | Eng ↑ | Env ↑ | CS ↑ | Bus ↑ | Avg. ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASG Systems** | | | | | | | | | | | |
| AutoSurvey | $3.29_{+0.37}$ | $3.68_{+0.47}$ | $\underline{3.63}_{+0.57}$ | $3.10_{+0.38}$ | $3.44_{+0.34}$ | $2.99_{+0.16}$ | $3.61_{+0.51}$ | $\mathbf{3.64}_{+0.60}$ | $\underline{3.64}_{+0.60}$ | $\underline{3.43}_{+0.60}$ | $3.45_{+0.46}$ |
| SurveyForge | $\underline{3.33}_{+0.41}$ | $\underline{3.84}_{+0.63}$ | $3.23_{+0.17}$ | $\underline{3.23}_{+0.51}$ | $\underline{3.63}_{+0.53}$ | $3.38_{+0.55}$ | $\mathbf{3.74}_{+0.64}$ | $3.52_{+0.48}$ | $3.40_{+0.36}$ | $3.41_{+0.58}$ | $\underline{3.47}_{+0.48}$ |
| AutoSurvey2 | $2.38_{-0.54}$ | $2.45_{-0.76}$ | $1.90_{-1.16}$ | $2.21_{-0.51}$ | $2.56_{-0.54}$ | $2.07_{-0.76}$ | $2.56_{-0.54}$ | $2.35_{-0.69}$ | $2.45_{-0.59}$ | $2.32_{-0.51}$ | $2.33_{-0.66}$ |
| InteractiveSurvey | $2.81_{-0.11}$ | $3.44_{+0.23}$ | $2.79_{-0.27}$ | $1.62_{-1.10}$ | $3.42_{+0.32}$ | $2.64_{-0.19}$ | $3.49_{+0.39}$ | $2.98_{-0.06}$ | $3.39_{+0.35}$ | $2.88_{+0.05}$ | $2.95_{-0.04}$ |
| LLM×MapReduce-V2 | $3.11_{+0.19}$ | $3.34_{+0.13}$ | $3.11_{+0.05}$ | $2.89_{+0.17}$ | $3.14_{+0.04}$ | $\underline{3.47}_{+0.64}$ | $3.24_{+0.14}$ | $3.19_{+0.15}$ | $3.24_{+0.20}$ | $2.73_{-0.10}$ | $3.15_{+0.16}$ |
| SurveyX | $2.85_{-0.07}$ | $3.14_{-0.07}$ | $3.30_{+0.24}$ | $2.57_{-0.15}$ | $3.04_{-0.06}$ | $2.97_{+0.14}$ | $3.41_{+0.31}$ | $3.05_{+0.01}$ | $3.05_{+0.01}$ | $2.74_{-0.09}$ | $3.01_{+0.02}$ |
| SciSage | $2.84_{-0.08}$ | $3.06_{-0.15}$ | $3.04_{-0.02}$ | $2.59_{-0.13}$ | $3.00_{-0.10}$ | $2.63_{-0.20}$ | $2.65_{-0.45}$ | $2.77_{-0.27}$ | $2.77_{-0.27}$ | $2.75_{-0.08}$ | $2.81_{-0.18}$ |
| **Avg (ASG Systems)** | $2.94_{+0.02}$ | $3.28_{+0.07}$ | $3.00_{-0.06}$ | $2.60_{-0.12}$ | $3.18_{+0.08}$ | $2.88_{+0.05}$ | $3.24_{+0.14}$ | $3.07_{+0.03}$ | $3.13_{+0.09}$ | $2.89_{+0.06}$ | $3.02_{+0.03}$ |
| **Vanilla LLMs** | | | | | | | | | | | |
| Qwen3-Max | $2.67_{-0.25}$ | $2.78_{-0.43}$ | $2.52_{-0.54}$ | $2.58_{-0.14}$ | $2.44_{-0.66}$ | $2.29_{-0.54}$ | $2.06_{-1.04}$ | $2.72_{-0.32}$ | $2.13_{-0.91}$ | $2.46_{-0.37}$ | $2.47_{-0.52}$ |
| Gemini-3-Pro | $2.28_{-0.64}$ | $2.46_{-0.75}$ | $2.75_{-0.31}$ | $2.46_{-0.26}$ | $2.41_{-0.69}$ | $2.15_{-0.68}$ | $2.46_{-0.64}$ | $2.40_{-0.64}$ | $2.60_{-0.44}$ | $2.01_{-0.82}$ | $2.40_{-0.59}$ |
| **Avg (Vanilla LLMs)** | $2.48_{-0.44}$ | $2.62_{-0.59}$ | $2.64_{-0.42}$ | $2.52_{-0.20}$ | $2.43_{-0.67}$ | $2.22_{-0.61}$ | $2.26_{-0.84}$ | $2.56_{-0.48}$ | $2.37_{-0.67}$ | $2.24_{-0.59}$ | $2.43_{-0.56}$ |
| **Deep Research Agents** | | | | | | | | | | | |
| Qwen DR | $2.85_{-0.07}$ | $3.30_{+0.09}$ | $3.46_{+0.40}$ | $3.09_{+0.37}$ | $3.14_{+0.04}$ | $3.00_{+0.17}$ | $3.42_{+0.32}$ | $3.17_{+0.13}$ | $3.08_{+0.04}$ | $2.93_{+0.10}$ | $3.14_{+0.15}$ |
| Gemini DR | $\mathbf{3.67}_{+0.75}$ | $\mathbf{3.84}_{+0.63}$ | $\mathbf{3.95}_{+0.89}$ | $\mathbf{3.53}_{+0.81}$ | $\mathbf{3.88}_{+0.78}$ | $\mathbf{3.53}_{+0.70}$ | $3.43_{+0.33}$ | $\underline{3.62}_{+0.58}$ | $\mathbf{3.71}_{+0.67}$ | $\mathbf{3.45}_{+0.62}$ | $\mathbf{3.66}_{+0.67}$ |
| **Avg (Deep Research Agents)** | $3.26_{+0.34}$ | $3.57_{+0.36}$ | $3.71_{+0.65}$ | $3.31_{+0.59}$ | $3.51_{+0.41}$ | $3.27_{+0.44}$ | $3.43_{+0.33}$ | $3.40_{+0.36}$ | $3.40_{+0.36}$ | $3.19_{+0.36}$ | $3.41_{+0.42}$ |
| **Avg (All)** | 2.92 | 3.21 | 3.06 | 2.72 | 3.10 | 2.83 | 3.10 | 3.04 | 3.04 | 2.83 | 2.99 |

Table 2: Results of discipline-aware rubric evaluation scores (1–5), best performance is in **bold**, second best is underlined. **Avg.** denotes the macro-average of the main scores across the 10 disciplines. Subscripts (+/−) indicate each system's deviation from the per-column average across all systems (last row, **Avg (All)**).

component level (outline, content, reference), and fine-grained aspect level, respectively.

**Finding 1: Deep Research Agents are the only paradigm that performs robustly across all 10 disciplines.** As shown in Table 2, Deep Research Agents are the only paradigm that exceeds the discipline-wise average in all ten fields, whereas every ASG system and vanilla LLM falls below the paradigm average on at least one discipline. This robustness is quantitative: DR's per-discipline scores span only 0.52 points, whereas ASG systems swing by 0.68 points across disciplines. Strong ASG systems do take discipline-specific top scores on fields (e.g., SurveyForge on Engineering and Medicine, AutoSurvey on Environmental Science), but no ASG or vanilla system matches DR's breadth. Vanilla LLM underperformance is sharpest on Reference, where scores collapse into a density spike at 1.0 (Figure 9), reflecting systematic citation-quality failure. This robustness is not an artefact of paired-survey leakage: excluding the 93 leaked DR surveys shifts the rubric mean by only 0.02 with no rank flips (Appendix D.2). *Implication:* for practitioners outside CS, DR agents are the safest default in the absence of discipline-specific tuning.

**Finding 2: Data-source quality dominates architectural sophistication.** Within ASG systems, top performers (AutoSurvey, SurveyForge, LLM×MapReduce-V2) share curated academic databases with rigorous filtering (e.g., Scholar Navigation Agent in SurveyForge (Yan et al., 2025)), while underperformers (SciSage, AutoSurvey2) rely on noisier web-crawled sources or less selective retrieval pipelines. *Implication:* optimize a Scholar-style, citation-aware retrieval pipeline (over generic web search) before iterating on planner or writer architectures.

| System | Outline ↑ | Content ↑ |
|---|---|---|
| **ASG Systems** | | |
| SurveyForge | **4.48** | 3.95 |
| AutoSurvey | 4.11 | 3.67 |
| **Vanilla LLMs** | | |
| Qwen3-Max | 2.03 | <u>4.30</u> |
| Gemini-3-Pro | 2.30 | 3.73 |
| **Deep Research Agents** | | |
| Qwen DR | 2.78 | <u>4.30</u> |
| Gemini DR | <u>4.24</u> | **4.31** |

Table 3: Outline vs. Content averages (1–5) for representative systems in each paradigm. Best in **bold**, second best underlined. Per-discipline results in Appendix E.

**Finding 3: ASG systems own the skeleton, DR agents own the flesh.** As shown in Table 3 and Figure 4 (details in Appendix E), component-level analysis reveals a pronounced trade-off between structural rigor and content richness: ASG systems concentrate strength in the A1–3 (outline) band whereas Deep Research Agents dominate A4–8 (content), and all paradigms remain weak on A9–11 (reference). ASG systems excel at outline generation: SurveyForge attains the highest Outline average (4.48), ahead of every DR agent. In contrast, Deep Research Agents lead on content synthesis: Gemini DR tops Content (4.31, with the Content ranking unchanged

| System | RA-AlignF1 ↑ | | | | | | | | | | | $\tau$-MaxSim ↑ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Avg. | | | | | | | | | | | Avg. |
| **ASG Systems** | | | | | | | | | | | | | | | | | | | | | | |
| AutoSurvey | 0.05 | 0.06 | 0.07 | 0.05 | 0.05 | 0.06 | 0.09 | 0.06 | 0.09 | 0.05 | 0.06 | 0.25 | 0.19 | 0.28 | 0.22 | 0.18 | 0.20 | 0.31 | 0.20 | 0.36 | 0.24 | 0.24 |
| SurveyForge | 0.09 | 0.11 | 0.10 | 0.08 | 0.08 | 0.10 | 0.12 | 0.11 | 0.14 | 0.08 | 0.10 | 0.25 | 0.24 | 0.25 | 0.23 | 0.21 | 0.22 | 0.27 | 0.22 | 0.34 | 0.24 | 0.25 |
| AutoSurvey2 | 0.12 | 0.09 | 0.01 | 0.12 | 0.07 | 0.09 | 0.12 | 0.12 | 0.14 | 0.12 | 0.10 | 0.29 | 0.24 | 0.21 | 0.26 | 0.26 | 0.27 | 0.32 | 0.26 | 0.38 | 0.30 | 0.28 |
| InteractiveSurvey | 0.08 | 0.07 | 0.07 | 0.06 | 0.08 | 0.07 | 0.09 | 0.07 | 0.10 | 0.07 | 0.08 | 0.15 | 0.15 | 0.12 | 0.14 | 0.23 | 0.11 | 0.19 | 0.10 | 0.23 | 0.13 | 0.16 |
| LLM×MapReduce-V2 | 0.14 | 0.11 | 0.11 | 0.11 | 0.06 | 0.10 | 0.10 | 0.11 | 0.12 | 0.09 | 0.11 | 0.34 | 0.36 | 0.35 | 0.21 | 0.19 | 0.25 | 0.35 | 0.25 | 0.32 | 0.22 | 0.28 |
| SurveyX | 0.09 | 0.10 | 0.09 | 0.10 | 0.10 | 0.09 | 0.12 | 0.11 | 0.14 | 0.10 | 0.10 | 0.18 | 0.19 | 0.18 | 0.22 | 0.28 | 0.17 | 0.26 | 0.22 | 0.27 | 0.22 | 0.22 |
| SciSage | **0.18** | **0.14** | 0.16 | **0.16** | 0.10 | 0.14 | 0.17 | 0.15 | **0.18** | 0.15 | **0.15** | 0.42 | **0.40** | 0.43 | **0.34** | 0.41 | 0.36 | **0.45** | **0.39** | **0.48** | **0.43** | **0.41** |
| **Vanilla LLMs** | | | | | | | | | | | | | | | | | | | | | | |
| Qwen3-Max | 0.13 | 0.11 | 0.16 | 0.14 | 0.10 | 0.12 | 0.14 | 0.13 | 0.17 | 0.13 | 0.13 | 0.29 | 0.30 | 0.43 | 0.29 | **0.49** | 0.31 | 0.39 | 0.29 | 0.44 | 0.32 | 0.36 |
| Gemini-3-Pro | 0.12 | 0.09 | 0.18 | 0.11 | 0.12 | 0.10 | 0.12 | 0.14 | 0.14 | 0.13 | 0.13 | 0.32 | 0.28 | **0.50** | 0.31 | 0.45 | 0.31 | 0.32 | 0.33 | 0.41 | 0.39 | 0.36 |
| **Deep Research Agents** | | | | | | | | | | | | | | | | | | | | | | |
| Qwen DR | 0.18 | 0.13 | 0.15 | 0.14 | 0.08 | 0.14 | 0.15 | 0.14 | 0.15 | **0.15** | 0.14 | **0.44** | 0.36 | 0.37 | 0.31 | 0.29 | **0.37** | 0.42 | 0.35 | 0.40 | 0.37 | 0.37 |
| Gemini DR | 0.13 | 0.14 | **0.20** | 0.15 | **0.13** | **0.14** | **0.17** | **0.16** | 0.15 | 0.13 | **0.15** | 0.23 | 0.24 | 0.38 | 0.25 | 0.38 | 0.27 | 0.31 | 0.24 | 0.26 | 0.26 | 0.28 |

Table 4: Human-reference alignment evaluation results in 10 disciplines, best performance is in **bold**, second best is underlined. RA-AlignF1 = Redundancy-Aware Alignment F1, $\tau$-MaxSim = Thresholded MaxSim.

after excluding leaked outputs; Appendix D.2) and Qwen DR (4.30) is close behind, whereas their Outline averages vary widely, reflecting heterogeneous structural reliability. Two backbone-controlled sub-comparisons isolate where these advantages originate. (i) Despite running on the smaller *Qwen3-30B-A3B* backbone, SurveyForge's Outline average (4.48) far exceeds the flagship vanilla Qwen3-Max (2.03) and Gemini-3-Pro (2.30), indicating that ASG outline strength is architectural rather than backbone-driven. (ii) At fixed backbone, Gemini DR scores 4.31 on Content vs. 3.73 for Gemini-3-Pro Vanilla, attributing the gap to the DR retrieval-and-synthesis paradigm. The same comparison on the Qwen family yields no Content gap (Qwen DR 4.30 vs. Qwen3-Max Vanilla 4.30; Appendix E), indicating that the DR paradigm's content-synthesis benefit is family-dependent rather than universal. *Implication:* the non-overlapping strengths point to an ASG-planner + DR-writer hybrid; reference quality remains a shared bottleneck neither paradigm currently solves (Appendix E.2).

**Finding 4: Paradigm ranking is not constant across disciplines.** The paradigm rank is not constant across disciplines, and on social-science fields it *reverses*: Qwen3-Max Vanilla scores 2.58 on Sociology, beating several ASG systems (Table 2). Two complementary biases drive this. **ASG systems** show a strong STEM bias: AutoSurvey drops from 3.64 (CS) to 2.99 (Psychology, −0.65), reflecting arXiv-heavy corpora and rigid hierarchies ill-suited for theoretical frameworks. **Vanilla LLMs** go the other way: Qwen3-Max scores higher in Education (2.67) and Sociology than in STEM fields, where its descriptive

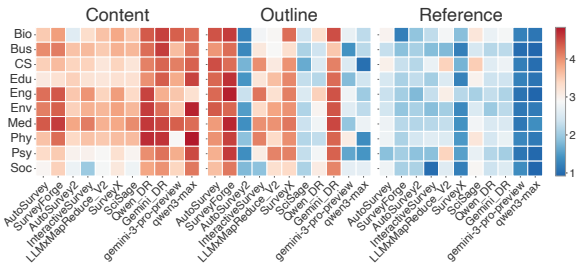


Figure 4: Aspect-level performance across the eleven sub-metrics A1–11 (grouped into Outline: A1–3; Content: A4–8; Reference: A9–11).

output matches the qualitative style of those disciplines. Only **Deep Research Agents** avoid both biases, as their agentic workflows adapt retrieval and structure per discipline. *Implication:* system selection must be done per-target-discipline; CS-derived rankings do not transfer, and discipline-aware evaluation is a prerequisite.

### 4.2.2 Human-Reference Alignment Evaluation

To complement our field-aware compliance analysis, we examine human-reference alignment through semantic similarity and structural statistics relative to paired human-written surveys.

**Finding 5: Models align with human surveys at the section level but fail at global organization.** As shown in Table 4, human-reference alignment metrics reveal a structural gap: $\tau$-MaxSim (segment-level similarity to human text) is consistently higher than RA-AlignF1 (global coverage with redundancy penalty). For example, Gemini-3-Pro averages 0.36 vs. 0.13 across the 10 disciplines, indicating local-but-not-global alignment. SciSage exemplifies this gap: high $\tau$-MaxSim (correct factual atoms) but low discipline-

| System | Img | Tab | Para | Wrd |
|---|---|---|---|---|
| **ASG Systems** | | | | |
| AutoSurvey | 0.11 | 0.02 | **7.11** | **4.00** |
| SurveyX | **0.52** | **0.32** | 4.09 | 2.16 |
| **Vanilla LLMs** | | | | |
| Qwen3-Max | 0.03 | 0.04 | 0.49 | 0.21 |
| Gemini-3-Pro | 0.01 | 0.01 | 0.24 | 0.09 |
| **Deep Research Agents** | | | | |
| Qwen DR | 0.03 | 0.26 | 0.25 | 0.22 |
| Gemini DR | 0.02 | 0.01 | 0.27 | 0.32 |

Table 5: Structural ratios vs. human surveys (1.0 = parity). Best per column in **bold**, second best underlined. Full results in Table 19.

| Setting | $\rho\uparrow$ | Conc.%$\uparrow$ |
|---|---|---|
| *(a) Method ablation on CS ($N = 7$)* | | |
| Holistic | -0.13 | 43.9 |
| Discipline-Prompted | 0.69 | 75.0 |
| Aspect (equal wt) | 0.70 | 77.3 |
| Rubric (equal wt) | **0.77** | 80.2 |
| Rubric + BT-domain | **0.77** | **81.0** |
| *(b) Cross-discipline pooled (Rubric + BT-domain)* | | |
| All non-CS ($N = 22$, 9 disciplines) | **0.90** | **83.2** |

Table 6: Human evaluation: Spearman $\rho$ and pairwise concordance (Conc.%) of LLM-judge rankings against ELO rankings. Detailed breakdown in Appendix C.2.

aware rubric score (generic template omitting the paired survey's central contributions; full case study in Appendix F.2). This also explains why DR/Vanilla's high Content rubric scores should be read as local proficiency rather than survey-level competence. *Implication:* future ASG and DR systems should treat global organization as an objective, not a byproduct of section-level synthesis.

**Finding 6: Pipeline ASG systems substitute textual bloat for substantive coverage.** Analysis of the "Richness Ratio" (Table 5) reveals a pattern of *Information Dumping*. Pipeline systems exhibit extreme textual bloat (AutoSurvey: $7.11\times$ paragraphs, $4.00\times$ words; SurveyX: $4.09\times$ / $2.16\times$ vs. human surveys) without improving quality. Multimodal synthesis remains weak even at best (SurveyX: $0.52\times$ images, $0.32\times$ tables; most other systems $\leq 0.05\times$), producing a "Wall of Text" rather than expert condensation. Appendix F.1 provides a case study where AutoSurvey expands a paired human Introduction by $27\times$ via paraphrastic redundancy across eight subsections. *Implication:* evaluation metrics that reward raw word or paragraph counts (e.g., naive ROUGE-recall) actively incentivize this failure mode; redundancy-aware metrics like RA-AlignF1 are required.

### 4.3 Human Evaluation

We validate the framework against human expert judgment through two complementary studies: (i) a *method ablation* on the CS subset (7 PhD-level CS researchers) isolating which design choices drive alignment, and (ii) a *cross-discipline generalization* test on 9 non-CS disciplines (22 PhD-level domain experts, dedicated annotation platform in Appendix C.1). We anchor presentation on outline, the most reliable signal across both human and inter-LLM-judge agreement (Appendices C.4, D.1); full results are in Appendix C.2.

**Method Ablation.** As shown in Table 6(a), holistic LLM scoring is essentially uncorrelated with human Outline judgment ($\rho = -0.13$). Discipline-Prompted ($\rho = 0.69$) helps but trails our rubric ($\rho = 0.77$, Conc. $= 80.2\%$); BT-domain weighting yields a modest gain ($81.0\%$).

**Cross-discipline Generalization.** Without recalibration, the same Rubric + BT-domain framework achieves a pooled $\rho = 0.90$ with $83.2\%$ Concordance across 9 non-CS disciplines, and *every individual discipline* attains $\rho \geq 0.70$ (Table 6(b)).

**Rubric Preference.** In a parallel criterion-recognition study, experts chose our field-aware rubric as best-fitting the work $85.3\%$ of the time (vs. $33.3\%$ chance; binomial $p < 0.001$, full per-component breakdown in Appendix C.3), indicating the rubric encodes substantive disciplinary norms rather than surface cues.

## 5 Conclusion

We introduce SurveyLens, a discipline-aware benchmark for ASG with a 1,000-survey corpus across 10 disciplines and a dual-lens framework. Benchmarking 11 state-of-the-art systems across 3 paradigms, we reveal a structure-vs-content trade-off: Deep Research agents are the most robust default for non-CS disciplines, while specialized ASG systems remain preferable when outline rigor is critical. Reference quality also remains a universal bottleneck across all surveyed paradigms. In the long term, these discipline-aware metrics can serve as optimization objectives for ASG model alignment, enabling systems to dynamically adapt to distinct disciplinary norms.

## Limitations

We note four limitations of SurveyLens. (1) DR agents cite or reuse the paired human-written survey on the queried topic far more often than ASG systems do; a leakage-controlled re-analysis (Appendix D.2) preserves the qualitative paradigm-level conclusions, but reference-based metrics on DR outputs should be interpreted in light of this audit. (2) Rubric scores are conditioned on a single primary judge (Qwen3-30B-A3B; Gemini-3-Pro only as a secondary cross-judge, Appendix D.1) and so carry residual judge-distribution bias; this is most pronounced on Content, where both cross-judge agreement and human alignment are markedly weaker than on Outline and Reference (Appendix C.2), reflecting low intrinsic expert consensus on long-form content quality and surface-level structural cues partially driving pairwise Content judgments (Krishna et al., 2023). (3) Coverage is bounded by API cost: each discipline is evaluated on 10 topics with three repeats (variance is small, $\sigma < 0.05$), the Discipline-Prompted ablation is conducted only on the CS subset, and SurveyLens-1k is a 2020–2025 snapshot that will require periodic refresh as both ASG systems and the literature evolve. (4) External validity: even with 10 disciplines, our findings should be read as a *representative* rather than exhaustive characterization: performance on disciplines not covered here (e.g., humanities, law) may differ systematically, and these rankings reflect the late-2025 capability frontier with rapid LLM evolution likely to shift relative paradigm standings.

## Ethics Statement

**Data sourcing and license.** All papers used to construct SurveyLens-1k are obtained through the official Semantic Scholar API and used in accordance with its Terms of Service. To avoid redistributing copyrighted full-text PDFs, our public release contains only (i) publicly available bibliographic metadata, (ii) the Structured Survey Representations we derive (outline, content segments, reference titles), and (iii) URLs / DOIs that point readers back to the original sources, following the precedent of S2ORC (Lo et al., 2020). We do not distribute the original PDFs. The released dataset and code are licensed for non-commercial academic research use only.

**Use of human annotators.** The expert annotations described in §4.3 were collected from PhD-level researchers who participated voluntarily; a subset of participants received compensation at locally-competitive hourly rates, while the remainder contributed time as a collegial courtesy within the authors' professional networks. The annotation platform records no personally identifying information; the public annotations are de-identified to (anonymous expert ID, discipline) tuples (Appendix C.1).

**Use of LLM proxy for preference labels.** Bradley-Terry weights in our rubric are learned from LLM-generated pairwise preferences rather than from large-scale human labels, as motivated in §3.3.1. We disclose this proxy explicitly and validate it post-hoc against genuine expert preferences in §4.3; readers should interpret the weights as expert-preference-aligned via an LLM proxy, not as human-aligned.

**Risks of misuse.** Automatic survey generation systems produce text that is stylistically indistinguishable from human-written surveys, and our results (§4.2.1) show that all evaluated paradigms remain weak on reference quality, including fabricated or mis-attributed citations. We intend SurveyLens as a diagnostic benchmark for system developers rather than an endorsement of using current ASG outputs as substitutes for human-authored literature reviews; downstream users should treat generated surveys, and especially their reference lists, as requiring expert verification before any scholarly use.

## References


Alibaba Qwen Team. 2025. Qwen deep research. https://chat.qwen.ai/. Accessed: 2026-01-14.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. SurveyGen: Quality-Aware Scientific Survey Generation with Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736, Suzhou, China. Association for Computational Linguistics.

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.

Google. 2025. Gemini deep research. https://gemini.google/overview/deep-research/. Accessed: 2026-01-14.

Beichen Guo, Zhiyuan Wen, Yu Yang, Peng Gao, Ruosong Yang, and Jiaxing Shen. 2026. SGSimEval: A Comprehensive Multifaceted and Similarity-Enhanced Benchmark for Automatic Survey Generation Systems. In *Advanced Data Mining and Applications*, pages 393–407, Singapore. Springer Nature Singapore.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6695–6715, Toronto, Canada. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 39–48, New York, NY, USA. Association for Computing Machinery.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2024. Instruct Large Language Models to Generate Scientific Literature Survey Step by Step. *arXiv preprint*. ArXiv:2408.07884 [cs].

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. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu li. 2025. SurveyX: Academic Survey Automation via Large Language Models. *arXiv preprint*. ArXiv:2502.14776 [cs].

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu, Yonghui Wu, Denghui Zhang, and Lichao Sun. 2025. Agentic AutoSurvey: Let LLMs Survey LLMs. *arXiv preprint*. ArXiv:2509.18661 [cs].

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Minh-Anh Nguyen, Minh-Duc Nguyen, Lan N. T. Ha, Kieu Hai Dang, Nguyen Tien Dong, and Dung D. Le. 2025. SurveyG: A Multi-Agent LLM Framework with Hierarchical Citation Graph for Automated Survey Generation. *arXiv preprint*. ArXiv:2510.07733 [cs].

OpenAI. 2025. Introducing deep research. https://openai.com/index/introducing-deep-research/. Accessed: 2026-01-14.

Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua Zhou. 2025. SciSage: A Multi-Agent Framework for High-Quality Scientific Survey Generation. *arXiv preprint*. ArXiv:2506.12689 [cs].

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. 2026. SurGE: A Benchmark and Evaluation Framework for Scientific Survey Generation. *arXiv preprint*. ArXiv:2508.15658 [cs].

Zhaojun Sun, Xuzhou Zhu, Xuanhe Zhou, Xin Tong, Shuo Wang, Jie Fu, Guoliang Li, Zhiyuan Liu, and Fan Wu. 2025. SurveyBench: Can LLM(-Agents) Write Academic Surveys that Align with Reader Needs? *arXiv preprint*. ArXiv:2510.03120 [cs].

M.-A.-P. Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 78 others. 2025. SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines. *arXiv preprint*. ArXiv:2502.14739 [cs].

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. MinerU: An Open-Source Solution for Precise Document Content Extraction. *arXiv preprint*. ArXiv:2409.18839 [cs].

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. *arXiv preprint*. ArXiv:2504.05732 [cs].

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

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. 2024c. AutoSurvey: Large Language Models Can Automatically Write Surveys. In *Advances in Neural Information Processing Systems*, volume 37, pages 115119–115145. Curran Associates, Inc.

Zhiyuan Wen, Jiannong Cao, Zian Wang, Beichen Guo, Ruosong Yang, and Shuaiqi Liu. 2025. InteractiveSurvey: An LLM-based Personalized and Interactive Survey Paper Generation System. *arXiv preprint*. ArXiv:2504.08762 [cs].

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. *arXiv preprint*. ArXiv:2510.26012 [cs].

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Hongbo Zhang, Han Cui, Yidong Wang, Yijian Tian, Qi Guo, Cunxiang Wang, Jian Wu, Chiyu Song, and Yue Zhang. 2025a. Deep Literature Survey Automation with an Iterative Workflow. *arXiv preprint*. ArXiv:2510.21900 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint*. ArXiv:1904.09675 [cs].

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint*. ArXiv:2506.05176 [cs].

Jiahao Zhao, Shuaixing Zhang, Nan Xu, and Lei Wang. 2025. SurveyEval: Towards Comprehensive Evaluation of LLM-Generated Academic Surveys. *arXiv preprint*. ArXiv:2512.02763 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical Catalogue Generation for Literature Review: A Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6790–6804, Singapore. Association for Computational Linguistics.

# A Evaluation Methodology Details

## A.1 Aspect Catalog

Table 7 lists the universal evaluation aspects $\mathcal{A}_{\mathcal{O}}, \mathcal{A}_{\mathcal{C}}, \mathcal{A}_{\mathcal{R}}$ synthesized from prior survey assessment methods, before the discipline-aware expand-and-merge process described in §3.3.1. These aspects serve as the seed set $\mathcal{A}_c$ in Eq. (2) and underlie the eleven sub-metrics (A1–A11) reported in the heatmap analysis (Figure 4).

## A.2 RA-AlignF1 Algorithm

**Algorithm 1** Calculation of RA-AlignF1

1: **Input:** Generated entries $\{e^{(i)}\}_{i=1}^{m}$, ground-truth entries $\{g^{(j)}\}_{j=1}^{n}$, component threshold $\tau_c$, penalty $\lambda$
2: **Output:** RA-AlignF1 $\in [0, 1]$
3: Compute thresholded margin matrix: $\mathbf{S}[i, j] \leftarrow \max\big(0, \operatorname{sim}(e^{(i)}, g^{(j)}) - \tau_c\big)$
4: If $m \neq n$, pad $\mathbf{S}$ with zero rows/columns to size $\max(m, n) \times \max(m, n)$
5: Find optimal matching: $\pi^* \leftarrow \arg\max_{\pi} \sum_{(i,j)\in\pi} \mathbf{S}[i, j]$ via Hungarian; retain only pairs with $i \leq m$ and $j \leq n$, so $|\pi^*| = \min(m, n)$
6: For $i = 1, \ldots, m$: $\omega(e^{(i)}) \leftarrow \exp\big(-\lambda \cdot \max_{k\neq i} \operatorname{sim}(e^{(i)}, e^{(k)})\big)$ if $m > 1$, else 1
7: $P \leftarrow \frac{1}{m} \sum_{(i,j)\in\pi^*} \omega(e^{(i)}) \cdot \mathbb{I}[\operatorname{sim}(e^{(i)}, g^{(j)}) \geq \tau_c]$
8: $R \leftarrow \frac{1}{n} \sum_{(i,j)\in\pi^*} \mathbb{I}[\operatorname{sim}(e^{(i)}, g^{(j)}) \geq \tau_c]$
9: **return** RA-AlignF1 $\leftarrow 2PR/(P + R)$ if $P + R > 0$, else 0

**Hyperparameter values.** We instantiate Algorithm 1 with cosine similarity computed over *Qwen3-Embedding-8B* embeddings (indexed in ChromaDB with `hnsw:space=cosine`). The threshold is set to $\tau_c = 0.95$ for all three components $c \in \{\mathcal{O}, \mathcal{C}, \mathcal{R}\}$, and the redundancy penalty is $\lambda = 1.0$ for all components. We choose $\tau_c = 0.95$ as a conservative paraphrase filter: near-synonymous rewrites under Qwen3-Embedding-8B typically fall in the 0.93–0.97 cosine range, so 0.95 sits at the upper end of that band and admits only substantive matches while filtering shallow lexical overlap. We use $\lambda = 1.0$ as a unit-scale default, which keeps the redundancy weight on the same dimensionless scale as the similarity itself. Both values are fixed before running the main evaluation and not tuned against the human-alignment data, so reported numbers are not a function of hyperparameter search.

## A.3 Rubric Weight Learning

The L2 regularization coefficient in Eq. (3) is set to $\alpha = 0.01$ for all (discipline, component) pairs. Like $\tau_c$ and $\lambda$ above, this value is fixed before running the main evaluation and not tuned against the human-alignment data.

## A.4 Scoring Example

Per-aspect scores $x_{i,k}$ in the rubric evaluation are obtained via LLM-based assessment. Given the discipline-specific rubric $\mathcal{K}_{d,c,a}$ and the corresponding survey component, we prompt an LLM judge to assess compliance on a 5-point scale. To ensure reliability, we adopt a structured output format requiring the judge to provide a rationale before assigning each score. An illustrative scoring example is shown below.

**Scoring Example (CS, Outline)**

**Aspect:** Critical Insight
**Learned Weight:** $\tilde{w}_{\mathcal{O},k} = 0.60$
**Rubric Criterion:** Identifies open problems & future directions
**Rubric Description:** Articulates unresolved challenges and synthesizes limitations across methods
**Normalized Score:** $\bar{x}_{i,k} = 4.2/5.0 = 0.84$

**Contribute** $\mathcal{S}_{d,\mathcal{O}}^{\text{RUB}}$**:** $5 \cdot \tilde{w}_{\mathcal{O},k} \cdot \bar{x}_{i,k} = 5 \times 0.6 \times 0.84 = \mathbf{2.52}$

# B SurveyLens-1k Dataset

## B.1 Construction Details

To ensure a robust and fair evaluation across diverse academic fields, we constructed **SurveyLens-1k**. This dataset comprises **1,000 high-quality survey papers**, evenly distributed across **10 disciplines** (100 papers per field). This section details our data collection, filtering, and verification pipeline.

We used the Semantic Scholar (S2) API[2] to obtain a comprehensive collection of survey papers published and corresponding metadata between 2020 and 2025. To guarantee domain di-

[2] https://api.semanticscholar.org

| Components | Aspects | Descriptions |
|---|---|---|
| Outline | Substantive Integrity | Evaluates depth, breadth, and scholarly merit of the content coverage. |
| | Structural Coherence | Assesses logical architecture and organizational flow of the outline. |
| | Formal Precision | Examines the technical execution of hierarchy and presentation. |
| Content | Scope and Relevance | Evaluates breadth and alignment with the central research theme. |
| | Structural Coherence | Assesses logical organization, transitions, and narrative consistency. |
| | Synthesis and Integration | Measures the ability to construct cohesive frameworks. |
| | Critical Insight and Novelty | Examines the depth of critique, original frameworks, and research gaps. |
| | Scholarly Communication | Reviews clarity, terminology precision, and citation standards. |
| Reference | Bibliometric Comprehensiveness | Evaluates coverage extent and inclusion of seminal works. |
| | Evidential Integrity | Assesses the the reliability of claims and correctness of attributions. |
| | Referential Pertinence and Compliance | Examines thematic alignment and citation formatting standards. |

Table 7: Assessment Aspects of Survey Papers Synthesized from Existing Studies (Sun et al., 2025; Zhao et al., 2025; Yan et al., 2025; Wang et al., 2024c; Liang et al., 2025; Wang et al., 2025; Liu et al., 2025; Nguyen et al., 2025; Zhao et al., 2025; Sun et al., 2025; Guo et al., 2026).

versity, we targeted ten representative disciplines: Biology, Business, Computer Science, Education, Engineering, Environmental Science, Medicine, Physics, Psychology, and Sociology.

To balance efficiency and accuracy, we restrict API queries to the "Review" document type for each discipline. Subsequently, we employ a hybrid filtering strategy to identify genuine surveys: papers with standard keywords in their titles (e.g., "survey," "review") are automatically accepted, while ambiguous entries undergo LLM-based binary classification using their titles and abstracts.

Following the initial retrieval and filtering process, we obtained an Initial Survey Pool consisting of approximately 3,000 papers. To ensure the inclusion of high-quality and influential works, we employed a citation-based selection strategy. Specifically, we ranked the papers within each discipline in descending order based on *both* (i) the raw `citationCount` and (ii) the `influentialCitationCount` metric returned by the Semantic Scholar API.[3] Combining these two signals avoids favoring merely old, highly-cited surveys while still privileging works of demonstrated impact. From this ranked list, we selected the top 110 papers per discipline. This procedure resulted in a refined set of Candidate Surveys containing roughly 1,100 papers, which serves as the foundation for our subsequent analysis.

To ensure the integrity and accessibility of the dataset, we implemented a rigorous two-step verification process on the candidate surveys. First, we employed an LLM-based verification to definitively validate the paper type, filtering out any remaining non-survey articles that passed the initial screening. Subsequently, we conducted a human verification phase focused on full-text acquisition. In this step, we manually inspected the candidates to ensure PDF accessibility, discarding entries with broken links, paywalls that could not be bypassed, or corrupted files. This process yielded a finalized collection of 1,000 high-quality survey papers, constituting the SurveyLens-1k Dataset. Figure 5 visualizes the resulting topical distribution: the 10 disciplines form well-separated clusters in topic-embedding space, indicating balanced coverage rather than concentration in any single domain.

### B.2 Per-Discipline Structural Statistics

Table 8 reports per-paper structural statistics averaged over the 100 human-written surveys in each of the 10 disciplines. These statistics quan-

[3]Per the Semantic Scholar API documentation, `influentialCitationCount` weights citations by the influence of the citing paper, so this metric is not biased toward simply older surveys with longer accumulation windows.

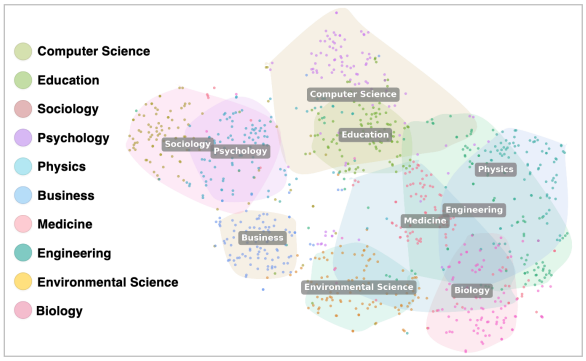


Figure 5: t-SNE visualization of survey-paper topic distribution in SurveyLens-1k, colored by discipline.

| Discipline | Img | Tab | Eq | Para | Word | Sent | Citation | Reference |
|---|---|---|---|---|---|---|---|---|
| Bio | 4.8 | 1.7 | 71.6 | 174.9 | 13.7k | 1.5k | 113.50 | 188.97 |
| Bus | 4.3 | 6.9 | 28.7 | 116.3 | 12.1k | 1.1k | 122.47 | 130.48 |
| CS | 8.3 | 8.7 | 91.7 | 146.9 | 14.7k | 1.2k | 128.15 | 110.39 |
| Edu | 4.7 | 6.7 | 35.3 | 95.8 | 10.5k | 1.0k | 98.61 | 80.71 |
| Eng | 10.3 | 4.3 | 122.1 | 203.2 | 17.5k | 1.9k | 180.36 | 245.04 |
| Env | 4.3 | 5.0 | 87.2 | 126.8 | 12.0k | 1.2k | 108.66 | 125.29 |
| Med | 4.1 | 3.9 | 54.7 | 152.1 | 12.7k | 1.3k | 88.71 | 155.34 |
| Phy | 13.3 | 3.1 | 411.4 | 256.3 | 23.4k | 2.3k | 233.03 | 261.26 |
| Psy | 1.9 | 4.4 | 54.6 | 109.6 | 11.3k | 1.1k | 117.68 | 111.84 |
| Soc | 1.4 | 2.4 | 25.8 | 92.9 | 9.4k | 0.8k | 115.57 | 98.81 |
| Avg | 5.7 | 4.7 | 98.3 | 147.5 | 13.7k | 1.3k | 130.22 | 150.41 |

Table 8: Per-paper structural statistics for SurveyLens-1k, averaged over the 100 human-written surveys in each discipline. **Para** denotes paragraphs; **Word** and **Sent** are reported in thousands.

tify the cross-disciplinary variation summarized in §3.3.1 and motivate the field-aware evaluation framework: technical density (equations, words), evidence presentation (images, tables), and citation practice all vary by an order of magnitude across disciplines.

### B.3 Structured Survey Representation Pipeline

To convert raw PDFs of both generated and human-written surveys into the Structured Survey Representation $\mathcal{S} = (\mathcal{O}, \mathcal{C}, \mathcal{R})$ used throughout our evaluation, we apply a four-stage pipeline:

1. **PDF parsing.** MinerU (Wang et al., 2024a) converts each PDF into structured Markdown, preserving font-size metadata, figure/table placement, and reference blocks.
2. **Rule-based filtering.** We recover the section hierarchy from font-size metadata, discard sections with extreme length deviations from per-discipline norms, and parse the reference list using discipline-aware heuristics (e.g., venue patterns, citation styles).
3. **LLM normalization.** An LLM pass standardizes section nesting levels across heterogeneous formats and complements the heuristic parser for reference-title extraction, particularly on entries with non-standard formatting.
4. **Human verification.** Annotators verify the resulting outlines and reference lists against the source PDFs to establish ground truth for the human-written corpus of SurveyLens-1k.

### B.4 Scalability and Maintenance

A benchmark over 10 disciplines is only useful if it can be extended and refreshed. We summarize here the operational properties of SurveyLens that make this practical.

**Automated construction pipeline.** The end-to-end pipeline is fully scripted: Semantic Scholar API retrieval, hybrid keyword + LLM survey-type filtering, dual-criterion (citation count + influential citation) ranking, and structured PDF parsing into Structured Survey Representations (Appendix B.1). Once configured for a new discipline, it requires only API access and standard compute, with no discipline-specific code changes.

**Human cost per new discipline.** Across all 10 disciplines in SurveyLens-1k, the only step that requires human attention is final verification of paper-type correctness and PDF accessibility. In our experience this consumes roughly 1–2 person-days per new discipline, which is comparable to or below the human cost of adding a new task to most NLP benchmarks.

**Discipline-aware rubric extension.** Adding a new discipline to the rubric evaluation does not require manual rubric authoring: the discipline-aware rubric is produced by the same LLM-driven *expand* $\rightarrow$ *merge* pipeline described in §3.3.1, instantiated on the new corpus. The downstream Bradley-Terry weighting step is also automated; only the optional human-alignment validation (§4.3) requires domain experts.

**Periodic refresh.** SurveyLens-1k is currently a 2020–2025 snapshot. We plan to refresh it on an annual basis by re-running the same pipeline over a rolling window (e.g., 2020–2026 in the next cycle), and to continue inviting domain experts to flag rubric items that no longer reflect current disciplinary conventions. We do not commit to a fixed update cadence in this paper, but the cost of one refresh cycle is bounded by the per-discipline figure above scaled by the number of disciplines.

# C Human Annotation and Expert Validation

## C.1 Annotation Platform and Expert Study Protocol

To support the cross-discipline expert validation described in §4.3, we built a dedicated web-based annotation platform. The platform, annotation interface screenshots, and anonymized raw annotations are all released at the project repository.[4]

**Expert Recruitment.** We recruited **30 PhD-level domain experts** in total covering all 10 disciplines: 7 CS researchers (via the method-ablation study in §4.3) and 23 non-CS experts (via this platform), distributed as Biology (1), Business (4), Education (4 recruited; 3 in Part A), Engineering (3), Environmental Science (2), Medicine (3), Physics (1), Psychology (1), and Sociology (4). All experts hold (or are completing) a PhD in their listed field. Of the 23 non-CS experts, 22 completed both annotation parts; one (Education) completed Part B (rubric preference) only and is therefore excluded from the Part A pairwise pool (so Part A has $N = 22$, Part B has $N = 23$).

**Annotation Tasks.** Each expert completed two task types on surveys from their own discipline:

- **Rubric preference (criterion recognition).** Given a survey and three rubric variants (Field-Aware, Cross-Discipline, Generic) presented in randomized order with blinded labels, the expert selects which rubric best fits the work. With 23 Part B participants and $3 + 5 + 3 = 11$ criteria per expert, the design called for 253 annotations; we collected 252 (Outline: $69 = 3 \times 23$, Content: 114 of 115 one response missing, Reference: $69 = 3 \times 23$).
- **Pairwise system comparison.** Given two anonymized system-generated surveys on the same topic, the expert selects the better one (with a no-preference option). For each expert, the 50 pairs are stratified by topic (5 pairs per topic over the 10 topics in their discipline) and drawn from the $\binom{11}{2} = 55$ system pairs; pair allocation is balanced across experts so that, pooled within a discipline, each system pair is compared in both A/B orderings (double round-robin), with partial coverage in single-expert disciplines (marked with † in Table 10). The resulting 1,100 pairwise judgments (22 Part A experts $\times$ 50 pairs each) were aggregated into per-discipline ELO rankings,[5] which we compared against SurveyLens rubric-derived rankings (Table 6, panel b).

Total expert workload was approximately 14 person-hours.

**Platform.** The annotation platform supports randomized presentation order, blinded labels for both rubrics and systems, session resume, and audit logging of all submitted judgments. Screenshots of the pairwise-comparison (Part A) and rubric-preference (Part B) interfaces are shown in Figures 6 and 7.

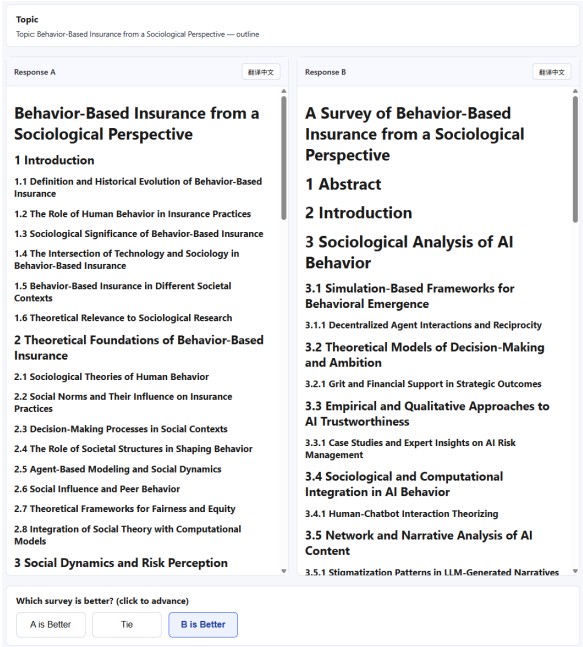


Figure 6: Part A: pairwise comparison interface. The shared survey topic is shown at the top; two anonymized system outputs are presented side-by-side under blinded labels (*Response A*, *Response B*), with model identities and any system-revealing metadata stripped. A per-pane "translate to Chinese" toggle is provided for non-English-native annotators. The annotator submits a judgment via one of three buttons at the bottom: *A is Better*, *Tie*, or *B is Better*.

[4] https://anonymous.4open.science/r/SurveyLens-93CB

[5] ELO uses standard online updates: initial rating = 1500, $K = 32$, expected score $E_A = 1/(1 + 10^{(R_B - R_A)/400})$, with win = 1 / loss = 0 / tie = 0.5. Each comparison contributes one update to each of the three per-dimension ELOs (Outline / Content / Reference), scored independently. We perform a single sweep over a randomly shuffled pair list (no iterative convergence criterion), and randomize A/B presentation order per comparison to neutralize position bias (seed = 42).

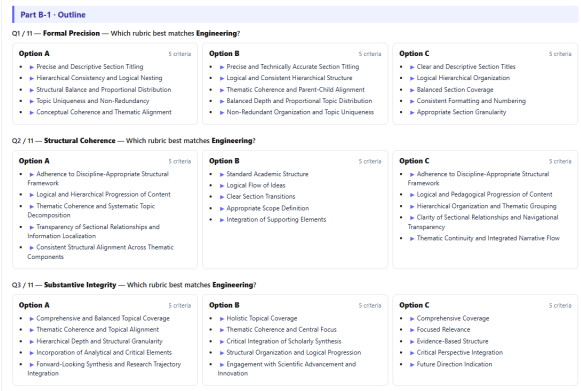


Figure 7: Part B: rubric preference interface (shown here for the Outline dimension). Each question header has the form *Q k/N* — CRITERION — *Which rubric best matches* DISCIPLINE*?*, naming the criterion under evaluation and the annotator's discipline (the example shown has 11 questions per session). The three rubric variants (Field-Aware, Cross-Discipline, Generic) are displayed as cards under blinded labels *Option A/B/C* in randomized order, each expanding to its 5 constituent criteria; the annotator selects the variant that best matches their discipline's academic standards.

**Ethics and Compensation.** Participation was voluntary; a subset of experts received compensation at locally-competitive hourly rates, while the remainder contributed time as a collegial courtesy within the authors' academic networks. No personally identifying information is released; the public annotations are de-identified to (anonymous expert ID, discipline) tuples.

### C.2 Full Human-Alignment Results (All Three Dimensions)

The main paper (§4.3) anchors its human-alignment validation on the Outline dimension, where both human inter-rater agreement and inter-LLM-judge agreement are highest. Here we report the full results across Outline, Content, and Reference.

**Method ablation (CS) across all dimensions.** Table 9 reports Spearman $\rho$ and Pairwise Concordance for all three dimensions on the CS subset ($N = 7$ experts). The rubric-based methods dominate the holistic baseline on every dimension, but the absolute alignment level is uneven: Outline and Reference reach $\rho \geq 0.47$ under Rubric + BT-domain, whereas Content alignment is substantially weaker ($\rho = 0.10$, Conc. $= 50.6\%$).

**Cross-discipline generalization (per-discipline breakdown).** Table 10 reports the per-discipline Spearman $\rho$ and Pairwise Concordance of the final framework (Rubric + BT-domain) on the Outline dimension, with per-discipline expert counts in parentheses. The pooled non-CS result ($\rho = 0.90$, Conc. $= 83.2\%$ over $N = 22$ experts), reported in Table 6(b), is supported by uniformly strong per-discipline alignment: no discipline falls below $\rho = 0.70$, and 6 of 9 disciplines reach $\rho \geq 0.90$.

| Setting | Outline | | Content | | Reference | |
|---|---|---|---|---|---|---|
| | $\rho$ | Conc.% | $\rho$ | Conc.% | $\rho$ | Conc.% |
| Holistic | -0.13 | 43.9 | -0.17 | 35.6 | 0.18 | 56.9 |
| Discipline-Prompted | 0.69 | 75.0 | -0.21 | 41.7 | 0.37 | 61.1 |
| Aspect (equal wt) | 0.70 | 77.3 | **0.20** | **52.6** | **0.52** | 64.2 |
| Rubric (equal wt) | **0.77** | 80.2 | 0.17 | 50.6 | 0.47 | 69.3 |
| Rubric + BT-domain | **0.77** | **81.0** | 0.10 | 50.6 | 0.47 | **70.2** |

Table 9: Ablation study on evaluation design choices. We report Spearman $\rho$ and Pairwise Concordance (%) between LLM-judge scores and human preference rankings (ELO) on the CS subset. **Discipline-Prompted** injects discipline-specific instructions into the judge prompt without using our rubric.

| Discipline | $\rho\uparrow$ | **Conc.%** $\uparrow$ |
|---|---|---|
| Biology (1)$^\dagger$ | 0.80 | 80.0 |
| Business (4) | 0.70 | 78.7 |
| Education (3) | 0.70 | 81.2 |
| Engineering (3) | 1.00 | 84.0 |
| Environmental Science (2) | 0.90 | 75.3 |
| Medicine (3) | 1.00 | 89.3 |
| Physics (1)$^\dagger$ | 0.90 | 84.0 |
| Psychology (1)$^\dagger$ | 1.00 | 90.0 |
| Sociology (4) | 0.90 | 86.6 |
| **All non-CS pooled** ($N = 22$) | **0.90** | **83.2** |

Table 10: Per-discipline human-alignment on Outline (Rubric + BT-domain). Numbers in parentheses are expert counts; $\rho$ is computed over the 11 systems. Cells marked $^\dagger$ are single-annotator disciplines; the reported $\rho$ reflects rank agreement between one expert's ELO and the LLM judge, and should be read as indicative rather than statistically aggregated.

**Why Content alignment is intrinsically harder.** The weaker Content alignment is not specific to our rubric: holistic scoring fares worse ($\rho = -0.17$, Conc. $= 35.6\%$), and Discipline-Prompted does too ($\rho = -0.21$, Conc. $= 41.7\%$). Two independent lines of evidence indicate that this gap reflects an intrinsic property of long-form content judgment rather than a deficiency of the rubric design: (i) *Human inter-dimension concordance.* Expert Content judgments concord with Outline judgments at $69.4\%$ but with Reference judgments at only $55.6\%$ (Appendix C.4), pointing to structural-cue interference in how humans rate content on

| Component | Field-Aware | Cross-Discipline | Generic | N | Binomial $p$ |
|---|---|---|---|---|---|
| Outline (3 criteria) | **94.2%** | 5.8% | 0.0% | 69 | $< 0.001$ |
| Content (5 criteria) | **88.6%** | 9.6% | 1.8% | 114 | $< 0.001$ |
| Reference (3 criteria) | **71.0%** | 24.6% | 4.3% | 69 | $< 0.001$ |
| **Overall** | **85.3%** | 12.7% | 2.0% | 252 | $< 0.001$ |

Table 11: Rubric preference study across 9 non-CS disciplines ($N = 23$ Part B experts). Experts judge which of three rubric variants best fits a given survey: **Field-Aware** (ours), **Cross-Discipline** (rubric from a different field), or **Generic** (field-agnostic). Binomial test against the 33.3% chance baseline.

| Dimension Pair | Concordance | Spearman $\rho$ |
|---|---|---|
| Outline × Content | 69.4% | 0.450 |
| Outline × Reference | 58.3% | 0.267 |
| Content × Reference | 55.6% | 0.067 |

Table 12: Inter-dimension concordance among human expert pairwise judgments. For each pair of dimensions, we report the rate at which expert preferences agree on the winning system (Concordance) and the Spearman $\rho$ between the per-system ELO rankings derived from each dimension.

long surveys. (ii) *Inter-LLM-judge agreement.* Replicating with a second judge (Gemini-3-Pro) yields Outline and Reference $\rho \approx 0.88$ but Content $\rho = 0.34$ (Appendix D.1). The same Content gap is reproduced across independent LLM judges, ruling out a rubric-side cause. We discuss mitigation strategies (content-only redaction protocols, instrument-level changes) in Limitations.

We separately note that Discipline-Prompted fares worse than Holistic on Content ($\rho = -0.21$ vs. $-0.17$), despite improving Outline ($\rho = 0.69$ vs. $-0.13$) and Reference ($\rho = 0.37$ vs. $0.18$). We tentatively attribute this to the discipline prompt steering the judge toward surface-level disciplinary cues (canonical structure, characteristic terminology), which aids judgment on Outline and Reference but actively crowds out substantive content assessment when applied to long-form content. The rubric, in contrast, operationalizes per-aspect criteria that constrain the judge's attention. We treat this as a hypothesis rather than a verified mechanism.

### C.3 Rubric Preference Study: Detailed Results

Table 11 reports the per-component breakdown of the 252 rubric-preference annotations summarized in §4.3. Across all three components, the Field-Aware rubric was chosen substantially more often than chance (33.3%): Outline 94.2%, Content 88.6%, Reference 71.0%, pooled 85.3% (binomial $p < 0.001$). The relatively lower preference on Reference is consistent with citation-quality criteria being more universal across disciplines than outline conventions or content depth.

### C.4 Inter-dimension Concordance Analysis

To diagnose why the Content dimension shows weaker alignment with expert judgment than Outline or Reference (§4.3), we examine the consistency of human pairwise preferences *across* the three dimensions. For each pair of dimensions, we compute (i) the rate at which an expert who prefers system $A$ over system $B$ on one dimension also prefers $A$ over $B$ on the other (Concordance), and (ii) the Spearman $\rho$ between the per-system ELO rankings derived independently from each dimension.

As shown in Table 12, Content judgments concord with Outline judgments at 69.4% ($\rho = 0.450$), substantially more than with Reference judgments (55.6%, $\rho = 0.067$). If Content were judged in isolation, we would expect symmetric, near-chance concordance with both other dimensions. The observed asymmetry suggests that human raters, when making pairwise Content judgments on long-form surveys, are partially influenced by surface-level structural cues such as section titles and headings rather than purely by content quality, an effect documented in prior long-form generation studies (Krishna et al., 2023). This structural-cue interference, rather than a deficiency of the rubric itself, is the most plausible explanation for the residual Content alignment gap in Table 9; mitigating it likely requires content-only redaction protocols or instrument-level changes to the annotation procedure, which we leave to future work.

## D Evaluation Robustness and Audits

### D.1 Multi-Judge Robustness Check

To assess whether our primary results are sensitive to the choice of LLM judge, we replicate the rubric evaluation with a second independent judge, *Gemini-3-Pro*, on the 4 directly comparable systems (Qwen3-Max, Qwen DR, Gemini-3-Pro, and Gemini DR) across all 10 disciplines. These four systems are chosen because they form two backbone-matched Vanilla↔DR pairs, making them the most informative slice for diagnos-

| Dimension | Inter-judge $\rho$ | Concordance |
|---|---|---|
| Outline | 0.88 | 91.7% |
| Content | 0.34 | 63.3% |
| Reference | 0.88 | 90.0% |
| **Average** | **0.84** | **88.3%** |

Table 13: Inter-judge agreement between the primary judge (*Qwen3-30B-A3B*) and a second independent judge (*Gemini-3-Pro*) on 4 directly comparable systems (Qwen3-Max, Qwen DR, Gemini-3-Pro, Gemini DR) across 10 disciplines. Outline and Reference exhibit near-identical rankings; the lower Content $\rho$ mirrors the human-judge structural-cue interference reported in Appendix C.4.

ing both inter-judge agreement and potential self-preference.

**Inter-judge agreement.** Table 13 reports the per-dimension Spearman $\rho$ and Pairwise Concordance between the primary judge (*Qwen3-30B-A3B*) and *Gemini-3-Pro*. Average agreement is high ($\rho = 0.84$, Concordance $= 88.3\%$). Outline and Reference rankings are near-identical across judges ($\rho = 0.88$ on both), whereas Content shows lower agreement ($\rho = 0.34$). This Content gap mirrors the structural-cue interference we observe in human expert judgments (Appendix C.4): long-form content judgments are intrinsically noisier across raters (LLM or human) than judgments anchored on outline structure or citation evidence. We note that with $N = 4$ systems, Spearman $\rho$ is sensitive to single rank flips; this check should therefore be read as evidence of *rank-stability under judge swap on backbone-matched pairs* rather than as a population-level inter-judge agreement statistic.

**Absence of systematic self-preference.** A natural concern is that each judge might favor outputs from systems sharing its model family (Qwen judge favoring Qwen3-Max/Qwen DR, Gemini judge favoring Gemini-3-Pro/Gemini DR). Examining the two pairs in question, rank shifts between the two judges for Qwen3-Max and Gemini-3-Pro are bounded by $\pm 1$ position. Crucially, the direction of any small shift is not consistent with self-preference (e.g., the Qwen judge does not uniformly rank Qwen systems higher than the Gemini judge does). We therefore find no evidence of systematic self-preference at the magnitude relevant to our paradigm-level conclusions.

### D.2 Retrieval Leakage Audit Protocol

This appendix details a two-stage retrieval-leakage audit motivated by a concern about human-reference alignment metrics: for every generated survey, we determine whether the system reproduced or directly retrieved the paired human-written survey on the queried topic, which would inflate human-reference alignment metrics in a way that does not reflect genuine synthesis ability.

**Stage 1: automatic candidate flagging.** For each of the $1{,}062$ generated outputs, we (i) extract the full reference list and (ii) scan the generated body text for occurrences of the paired survey title (matched both verbatim and under a normalized form that strips punctuation, case, and version suffixes). Any output that cites the paired survey title in either location is flagged as a candidate leak.

**Stage 2: manual verification.** Every flagged output is then verified by hand to remove false positives such as incidental citation of a same-topic but non-paired reference, or citation of a related survey on a strictly different sub-topic. Only outputs that demonstrably ground a substantial portion of their content on the paired survey are counted as "confirmed leakage" in Table 14 (panel a).

**Coverage and per-paradigm rates.** We apply the protocol uniformly across all $1{,}062$ outputs. Confirmed leakage rates are $2/200$ ($1.00\%$) for Vanilla LLMs, $46/665$ ($6.92\%$) for ASG systems, and $93/197$ ($47.21\%$) for Deep Research Agents. The DR category is internally split: Qwen DR $8.2\%$ vs. Gemini DR $85.9\%$. We report this asymmetry transparently rather than aggregate it away.

**Leakage-excluded re-analysis.** Table 14 (panel b) repeats our human-reference alignment analysis with all confirmed-leaked DR outputs removed. On RA-AlignF1, the paradigm ordering DR (0.14) $>$ Vanilla (0.13) $>$ ASG (0.10) is preserved after exclusion, indicating that the DR advantage over ASG is not an artifact of citing the paired survey; RA-AlignF1 is structurally robust to this form of leakage because its Hungarian one-to-one matching prevents repeated alignment with a single source paper. On $\tau$-MaxSim, we note transparently that Vanilla LLMs (0.36) edge out leakage-excluded DR (0.35), with both clearly above ASG (0.26). We attribute the Vanilla $\tau$-MaxSim edge to output-length effects: Vanilla

*(a) Confirmed leakage rate by paradigm*

| Group | N | Leakage Rate↓ |
|---|---|---|
| Vanilla LLMs | 200 | 1.00% |
| ASG Systems | 665 | 6.92% |
| Deep Research Agents | 197 | 47.21% |
| **Overall** | **1,062** | **13.29%** |

*(b) Human-reference alignment after leakage exclusion*

| Group | N | RA-AlignF1↑ | $\tau$-MaxSim↑ |
|---|---|---|---|
| DR (all) | 197 | 0.15 | 0.32 |
| DR (w/o leakage) | 104 | 0.14 | 0.35 |
| Vanilla LLM | 200 | 0.13 | **0.36** |
| ASG | 665 | 0.10 | 0.26 |

*(c) Rubric scores after leakage exclusion (DR)*

| Group | N | Out. | Cont. | Ref. | Mean |
|---|---|---|---|---|---|
| DR (all) | 197 | 3.53 | 4.32 | 2.43 | 3.43 |
| DR (clean) | 104 | 3.50 | 4.29 | 2.42 | 3.40 |
| $\Delta$ | | $-0.03$ | $-0.03$ | $-0.01$ | $-0.02$ |
| Qwen DR (all) | 98 | 2.76 | 4.28 | 2.40 | 3.15 |
| Qwen DR (clean) | 90 | 2.78 | 4.26 | 2.40 | 3.15 |
| Gemini DR (all) | 99 | 4.30 | 4.36 | 2.47 | 3.71 |
| Gemini DR (clean)$^{\dagger}$ | 14 | 4.23 | 4.32 | 2.44 | 3.66 |

Table 14: Retrieval-leakage audit. (a) Confirmed leakage rates by paradigm. (b) Human-reference alignment (RA-AlignF1/ $\tau$-MaxSim) after removing leaked DR outputs. (c) Rubric scores (1–5) before and after leakage exclusion. The DR aggregate is stable ($\Delta = -0.02$ on the mean, $|\Delta| \leq 0.03$ on all three dimensions) and *no per-discipline rank flips* are observed. Per-system splits are reported for completeness; $^{\dagger}$Gemini DR clean ($n = 14$, covering 6 of 10 disciplines) has high variance and should be read alongside the more robust DR aggregate. Gemini DR remains first on Content after exclusion (4.32), preserving the Finding 3 (Content) ordering.

outputs are short and contain little textual bloat, so their per-segment maximum similarity to the paired survey is mechanically inflated. We therefore read RA-AlignF1, which jointly penalizes redundancy and rewards coverage, as the more reliable human-reference alignment signal, on which the three-way ordering remains clean.

**Limitations of the audit.** Our protocol detects leakage that surfaces as citation of, or near-verbatim reuse from, the paired survey title. It does not detect (i) paraphrased reuse without title citation, (ii) reuse of other same-topic surveys not designated as the paired reference, or (iii) reuse of underlying primary sources cited by the paired survey. A leakage-controlled evaluation protocol using post-cutoff topics is the natural next step, which we leave to future work.

| System | Missing | Cause |
|---|---|---|
| SurveyX | 7 | Three retries still failed; server-side issue confirmed by authors. |
| AutoSurvey2 | 13 | Biology outputs written to a wrong directory; generation itself succeeded. |
| SciSage | 10 | Late re-runs not in initial counts; included in current results. |
| Others | 8 | Intermittent LLM/retrieval API failures; no call-level logs retained. |
| **Total** | **38** | Yield rate $1{,}062/1{,}100 = 96.5\%$. |

Table 15: Per-system generation failure breakdown. Failures are concentrated in system-level issues and show no systematic correlation with discipline or topic.

### D.3 Per-System Generation Failure Analysis

Of the 1,100 ($11 \times 100$) system-topic generation attempts, 1,062 produced valid outputs (yield 96.5%). Table 15 breaks down the 38 missing outputs by system and cause. Failures fall into four buckets: persistent server-side failures (SurveyX, 7), output-handling mistakes on our side that did not affect generation itself (AutoSurvey2 Biology directory misplacement, 13), late re-runs that are now included in the reported results (SciSage, 10), and intermittent third-party API failures without retained call-level logs (others, 8). None of these failure modes are correlated with specific disciplines or topics in a way that would bias the cross-discipline conclusions: missing outputs are distributed across disciplines roughly proportional to the per-system attempt count. The 8 intermittent failures are sparsely scattered across affected systems ($\leq 2$ outputs each) and are not large enough to flip any per-discipline ranking at the resolution reported in §4.2.1.

### D.4 Comparison Scope

Our benchmark compares three paradigms that differ not only in algorithmic design but also in deployment form, and we structure the comparison accordingly. For the seven open-source ASG systems, we control backbone (a unified *Qwen3-30B-A3B*) so that any performance gap among them isolates the architectural contribution of each pipeline. Vanilla LLMs and Deep Research agents, by contrast, are evaluated using their officially deployed configurations: their backbones are integral to the deployed product and

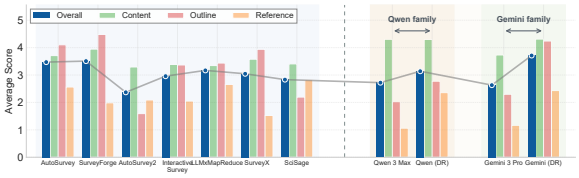


Figure 8: Dimension-level evaluation results of the 11 systems across Overall, Outline, Content, and Reference, highlighting the trade-offs between structural planning and content synthesis.

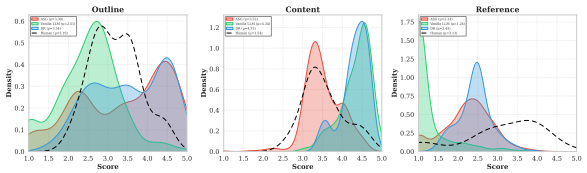


Figure 9: Paradigm-level score distributions on Outline, Content, and Reference for the three machine paradigms (ASG, Vanilla LLM, DR) and human-written surveys. The Reference panel exhibits a persistent human advantage absent from Outline and Content; Vanilla LLM exhibits a density spike at score 1.0 on Reference; ASG Outline is bimodal at the paradigm level.

not user-substitutable, so swapping them would no longer reflect how practitioners actually consume these systems. We therefore position the cross-paradigm results as a practitioner-oriented guide at the paradigm level. Where backbone-controlled architectural claims are made, they are supported by targeted sub-comparisons in the main text: Finding 3 (§4.2.1) isolates the ASG outline advantage by holding the backbone fixed at Qwen3-30B-A3B (e.g., SurveyForge 4.48 vs. flagship Qwen3-Max 2.03), and the Vanilla↔DR pairs at fixed backbone family (Qwen3-Max vs. Qwen DR; Gemini-3-Pro vs. Gemini DR) isolate the DR retrieval-and-synthesis contribution. Together, these within-family controls let us attribute the headline paradigm-level findings to architecture rather than backbone capacity, while keeping the deployment-form comparison faithful to real usage.

# E Detailed Results

In this section, we report the detailed performance of the evaluated systems. We first present a fine-grained visualization of aspect-wise metrics and then provide a breakdown of scores across different academic disciplines.

## E.1 Dimension-Level Grouped Comparison

Figure 8 complements the aspect-level heatmap in the main text (Figure 4) by aggregating the eleven sub-metrics back into the three primary dimensions (Outline, Content, Reference) plus the Overall score, giving a side-by-side view of the 11 systems at the dimension granularity.

For reference, the aspect grouping behind both figures is as follows. The **Outline** dimension covers A1–A3, evaluating **Substantive Integrity**, **Structural Coherence**, and **Formal Precision**. The **Content** dimension spans A4–A8, assessing **Scope and Relevance**, **Structural Coherence**, **Synthesis and Integration**, **Critical Insight and Novelty**, and **Scholarly Communication**. Finally, the **Reference** dimension includes A9–A11, examining **Bibliometric Comprehensiveness**, **Evidential Integrity**, and **Referential Pertinence and Compliance**.

## E.2 Paradigm-Level Score Distributions

To complement the system-level bar chart (Figure 8) and the aspect-level heatmap (Figure 4), Figure 9 reports score distributions of the three machine paradigms (ASG, Vanilla LLM, DR) alongside human-written surveys on the three primary components. Three visualisation-level observations support findings in the main text: (i) on Reference, the human distribution ($\mu = 3.13$) sits clearly above all machine paradigms (DR $\mu = 2.43$, ASG $\mu = 2.24$, Vanilla LLM $\mu = 1.24$), reinforcing Finding 3's identification of Reference as the shared bottleneck; (ii) Vanilla LLM exhibits a density spike at score 1.0 on Reference, indicating a catastrophic citation-quality failure mode for systems without external retrieval, consistent with Finding 1's observation that vanilla LLMs consistently underperform; (iii) the ASG Outline distribution is bimodal at the paradigm level (with modes near 2.5 and 4.5), directly visualising the within-paradigm reliability split that Finding 2 attributes to data-source quality.

## E.3 Discipline-Specific Performance

Tables 16, 17, and 18 list the Reference-Free scores (1–5) for the three key dimensions across ten distinct disciplines.

- **Outline Quality (Table 16):** This table compares the structural quality of the generated surveys. It includes scores for ASG Systems, Vanilla LLMs, and Deep Research Agents,

highlighting the best and second-best performing models in each field.

- **Content Quality (Table 17):** This table evaluates the coherence, relevance, and depth of the generated content. It allows for a comparison of how well different approaches handle domain-specific knowledge generation.
- **Reference Quality (Table 18):** This table assesses the accuracy and relevance of the citations provided by the systems. It particularly highlights the significant performance gap between systems with retrieval capabilities (ASG Systems and Deep Research Agents) and Vanilla LLMs in terms of hallucination and citation correctness.

| System(Outline) | Edu | Med | Bio | Soc | Phys | Psy | Eng | Env | CS | Bus | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASG Systems** | | | | | | | | | | | |
| AutoSurvey | 4.28 | 3.92 | 4.43 | 3.62 | 3.99 | 3.75 | 4.27 | 4.38 | **4.46** | 3.95 | 4.11 |
| SurveyForge | **4.65** | **4.73** | **4.58** | **4.16** | **4.54** | **4.60** | **4.78** | 4.29 | 4.24 | **4.24** | **4.48** |
| AutoSurvey2 | 1.43 | 1.37 | 1.23 | 1.80 | 1.88 | 1.35 | 2.33 | 1.51 | 1.73 | 1.30 | 1.59 |
| InteractiveSurvey | 3.18 | 4.00 | 2.81 | 1.88 | 4.11 | 3.06 | 4.20 | 3.35 | 4.01 | 3.14 | 3.37 |
| LLM×MapReduce-V2 | 3.62 | 3.99 | 2.92 | 3.28 | 3.35 | 4.09 | 3.26 | 3.85 | 3.11 | 2.88 | 3.44 |
| SurveyX | 4.06 | 4.19 | 4.24 | 3.36 | 4.02 | 4.10 | 4.12 | 3.98 | 3.99 | 3.28 | 3.93 |
| SciSage | 2.28 | 2.35 | 2.38 | 2.10 | 2.17 | 2.28 | 2.51 | 2.13 | 1.64 | 2.11 | 2.20 |
| **Vanilla LLMs** | | | | | | | | | | | |
| Qwen3-Max | 2.40 | 2.73 | 2.24 | 2.64 | 1.33 | 1.43 | 2.00 | 2.24 | 1.03 | 2.21 | 2.03 |
| Gemini-3-Pro | 1.82 | 1.92 | 2.69 | 2.64 | 3.06 | 1.60 | 2.76 | 2.58 | 2.51 | 1.40 | 2.30 |
| **Deep Research Agents** | | | | | | | | | | | |
| Qwen DR | 2.19 | 2.79 | 3.28 | 3.08 | 2.28 | 2.82 | 3.45 | 2.95 | 2.54 | 2.38 | 2.78 |
| Gemini DR | 4.26 | 4.50 | 4.46 | 4.08 | 4.31 | 4.32 | 4.42 | **4.41** | 4.11 | 3.54 | 4.24 |

Table 16: Results of discipline-aware rubric evaluation scores (1–5) in terms of Outline. Best performance is in **bold**, second best is underlined.

| System(Content) | Edu | Med | Bio | Soc | Phys | Psy | Eng | Env | CS | Bus | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASG Systems** | | | | | | | | | | | |
| AutoSurvey | 3.11 | 4.39 | 3.52 | 2.93 | 3.59 | 3.34 | 4.23 | 4.12 | 3.47 | 4.04 | 3.67 |
| SurveyForge | 3.25 | 4.58 | 3.90 | 3.45 | 4.24 | 3.50 | 4.40 | 4.34 | 3.65 | 4.14 | 3.95 |
| AutoSurvey2 | 3.44 | 3.83 | 2.57 | 2.73 | 3.37 | 2.97 | 3.54 | 3.44 | 3.42 | 3.59 | 3.29 |
| InteractiveSurvey | 3.22 | 4.13 | 3.31 | 1.98 | 3.59 | 2.99 | 3.97 | 3.74 | 3.43 | 3.67 | 3.40 |
| LLM×MapReduce-V2 | 3.18 | 3.74 | 3.59 | 2.84 | 3.52 | 2.90 | 3.42 | 3.67 | 3.17 | 3.51 | 3.35 |
| SurveyX | 3.25 | 3.82 | 3.81 | 3.14 | 3.64 | 3.26 | 3.91 | 3.86 | 3.42 | 3.75 | 3.59 |
| SciSage | 3.25 | 4.04 | 3.54 | 2.75 | 3.65 | 2.94 | 3.36 | 3.67 | 3.18 | 3.68 | 3.41 |
| **Vanilla LLMs** | | | | | | | | | | | |
| Qwen3-Max | **4.53** | 4.25 | 4.22 | **4.16** | **4.83** | **4.48** | 3.27 | **4.76** | **4.32** | 4.22 | 4.30 |
| Gemini-3-Pro | 3.79 | 4.33 | 4.34 | 3.41 | 2.89 | 3.76 | 3.56 | 3.49 | 4.11 | 3.63 | 3.73 |
| **Deep Research Agents** | | | | | | | | | | | |
| Qwen DR | 3.99 | **4.61** | 4.38 | 3.84 | 4.69 | 4.01 | **4.53** | 4.57 | 4.11 | 4.25 | 4.30 |
| Gemini DR | 4.35 | 4.52 | **4.55** | 3.98 | 4.63 | 4.22 | 3.75 | 4.22 | 4.30 | **4.58** | **4.31** |

Table 17: Results of discipline-aware rubric evaluation scores (1–5) in terms of Content. Best performance is in **bold**, second best is underlined.

| System(Reference) | Edu | Med | Bio | Soc | Phys | Psy | Eng | Env | CS | Bus | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASG Systems** | | | | | | | | | | | |
| AutoSurvey | 2.47 | 2.74 | 2.96 | 2.76 | 2.76 | 1.87 | 2.34 | 2.41 | 3.00 | 2.30 | 2.56 |
| SurveyForge | 2.09 | 2.20 | 1.20 | 2.08 | 2.10 | 2.04 | 2.04 | 1.93 | 2.32 | 1.86 | 1.99 |
| AutoSurvey2 | 2.28 | 2.16 | 1.90 | 2.11 | 2.44 | 1.88 | 1.80 | 2.10 | 2.18 | 2.08 | 2.09 |
| InteractiveSurvey | 2.04 | 2.20 | 2.25 | 1.00 | 2.58 | 1.86 | 2.30 | 1.85 | 2.72 | 1.84 | 2.06 |
| LLM×MapReduce-V2 | 2.54 | 2.28 | 2.82 | 2.56 | 2.54 | **3.42** | 3.04 | 2.05 | 3.44 | 1.79 | 2.65 |
| SurveyX | 1.24 | 1.40 | 1.84 | 1.20 | 1.47 | 1.53 | 2.20 | 1.30 | 1.74 | 1.20 | 1.51 |
| SciSage | **2.98** | 2.80 | **3.18** | 2.92 | **3.18** | 2.68 | 2.08 | 2.51 | **3.48** | 2.48 | **2.83** |
| **Vanilla LLMs** | | | | | | | | | | | |
| Qwen3-Max | 1.08 | 1.36 | 1.10 | 0.94 | 1.14 | 0.96 | 0.92 | 1.16 | 1.02 | 0.94 | 1.06 |
| Gemini-3-Pro | 1.22 | 1.12 | 1.22 | 1.34 | 1.28 | 1.10 | 1.06 | 1.12 | 1.18 | 1.01 | 1.17 |
| **Deep Research Agents** | | | | | | | | | | | |
| Qwen DR | 2.36 | 2.50 | 2.72 | 2.36 | 2.46 | 2.16 | 2.28 | 1.98 | 2.58 | 2.17 | 2.36 |
| Gemini DR | 2.38 | 2.50 | 2.84 | 2.54 | 2.70 | 2.04 | 2.12 | 2.23 | 2.74 | 2.23 | 2.43 |

Table 18: Results of discipline-aware rubric evaluation scores (1–5) in terms of Reference. Best performance is in **bold**, second best is underlined.

| System | Img | Tab | Eq | Para | Wrd | Sent | Cite | Ref | Char |
|---|---|---|---|---|---|---|---|---|---|
| **ASG Systems** | | | | | | | | | |
| AutoSurvey | 0.11 | 0.02 | 0.04 | 7.11 | 4.00 | 2.18 | **1.03** | **0.99** | 4.64 |
| SurveyForge | 0.04 | 0.01 | 0.01 | 1.51 | 0.73 | 0.35 | 0.27 | 0.24 | 0.87 |
| AutoSurvey2 | 0.04 | 0.00 | 0.00 | 0.18 | 0.14 | 0.07 | 0.11 | 0.10 | 0.16 |
| InteractiveSurvey | 0.05 | 0.00 | 0.01 | 0.56 | 0.39 | 0.19 | 0.09 | 0.08 | 0.47 |
| LLM×MapReduce-V2 | 0.02 | 0.02 | 0.01 | 0.46 | 0.37 | 0.18 | 0.04 | 0.03 | 0.44 |
| SurveyX | **0.52** | **0.32** | **0.14** | **4.09** | **2.16** | **0.85** | 0.25 | 0.22 | **2.42** |
| SciSage | 0.01 | 0.00 | 0.00 | 0.25 | 0.13 | 0.08 | 0.07 | 0.07 | 0.16 |
| **Vanilla LLMs** | | | | | | | | | |
| Qwen3-Max | 0.03 | 0.04 | 0.04 | 0.49 | 0.21 | 0.20 | 0.33 | 0.18 | 0.24 |
| Gemini-3-Pro | 0.01 | 0.01 | 0.03 | 0.24 | 0.09 | 0.09 | 0.11 | 0.06 | 0.11 |
| **Deep Research Agents** | | | | | | | | | |
| Qwen DR | 0.03 | 0.26 | 0.01 | 0.25 | 0.22 | 0.14 | 0.10 | 0.10 | 0.25 |
| Gemini DR | 0.02 | 0.01 | 0.04 | 0.27 | 0.32 | 0.29 | 0.07 | 0.33 | 0.29 |

Table 19: Full structural statistics of generated surveys relative to human-written references (1.0 = human parity). Compact four-column version used in the main text in Table 5. Highest value per column in **bold**, second highest underlined.

# F Qualitative Case Studies on ASG Failure Modes

To make the abstract failure modes identified in Findings 5 and 6 concrete, we conduct two qualitative case studies on a single shared topic, *"Abusive Language Training Data in Natural Language Processing"*, for which a high-quality human-written survey (Vidgen and Derczynski, 2020) is available as the paired reference. We compare it against generated outputs from two representative ASG systems (**AutoSurvey** (Wang et al., 2024c) and **SciSage** (Shi et al., 2025)), which exemplify two opposite failure patterns at the structural level.

## F.1 Case A: Information Dumping in Pipeline ASG (Finding 6)

**Setup.** Finding 6 observes that pipeline ASG systems like AutoSurvey exhibit extreme textual bloat ($7.11\times$ paragraphs, $4.00\times$ words relative to human surveys) without improving substantive coverage. We zoom into the *Introduction* chapter of both surveys, which targets the same rhetorical goal: motivating the importance of the topic.

**Quantitative comparison.** Table 20 reports per-section structural statistics. AutoSurvey expands the Introduction into 8 third-level subsections, each averaging $\approx 880$ words. The whole AutoSurvey Introduction occupies roughly **7,001 words / 47,567 characters**, whereas the paired human Introduction completes the same rhetorical function in **259 words / 1,851 characters**, a $27\times$ **length expansion** with no corresponding gain in coverage breadth (both introduce the same set of motivations).

| Section | Words | Chars | Sents |
|---|---|---|---|
| Human *Introduction* (full) | 259 | 1,851 | 13 |
| AutoSurvey §1.1 *Importance* | 974 | 6,733 | 35 |
| AutoSurvey §1.2 *Role of Data* | 912 | 6,219 | 35 |
| AutoSurvey §1.3 *Societal Impact* | 888 | 6,234 | 36 |
| AutoSurvey §1.4 *Curation Challenges* | 853 | 5,993 | 35 |
| AutoSurvey §1.5–1.8 (4 subsec.) | 3,374 | 22,388 | 137 |
| AutoSurvey *Intro.* (full, 8 subsec.) | **7,001** | **47,567** | 278 |

Table 20: Per-section statistics for the *Introduction* chapter of the human-written survey (Vidgen and Derczynski, 2020) versus AutoSurvey's generated output on the same topic.

**Qualitative excerpt.** The contrast becomes sharper when reading opening sentences of each. The human introduction grounds claims in specific policies, venues, and counts:

> *"… policy and regulatory developments, such as the EU's Code of Conduct on Hate, the UK Government's 'Online Harms' white paper, Germany's NetzDG laws, the Public Pledge on Self-Discipline for the Chinese Internet Industry, and France's anti-hate regulation. In 2020 alone, at least three computer science venues will host workshops on online hate (TRAC and STOC at LREC, and WOAH at EMNLP), and a shared task at 2019's SemEval … reported that 800 teams downloaded the training data and 115 submitted detection systems." (Vidgen and Derczynski, 2020)*

AutoSurvey's §1.1, addressing the same motivational goal, instead chains generic intensifiers without verifiable evidence:

> *"Abusive language detection in Natural Language Processing (NLP) has emerged as a critical area of research … The significance of this task cannot be overstated, as it plays a vital role in creating safer online environments and promoting the ethical development of AI systems. As online interactions become more frequent and diverse, the need to identify and mitigate abusive language has become imperative to ensure that digital spaces remain inclusive and respectful for all users … "*

The next three subsections (§1.2 *Role of Training Data*, §1.3 *Societal Impact*, §1.4 *Curation Challenges*) re-invoke near-identical hedges (*"plays a critical role"*, *"cannot be overstated"*, *"it is essential to ensure"*) and largely overlap in claim space, producing horizontal redundancy rather than incremental coverage.

**Takeaway.** The case directly instantiates Finding 6: AutoSurvey produces a $27\times$-longer Introduction *by paraphrasing the same motivational claims across eight subsections*, rather than by expert condensation. This is also precisely the failure mode the redundancy penalty $\omega(e)$ in our RA-AlignF1 metric (§4.2.2) is designed to surface.

## F.2 Case B: Local Alignment without Global Structure in SciSage (Finding 5)

**Setup.** Finding 5 documents a systematic gap between high $\tau$-MaxSim (local segment-level similarity to human text) and low RA-AlignF1/ rubric scores (global coverage and narrative organization). SciSage on this topic provides a particularly clean illustration: its individual paragraphs reference correct facts, but its structural footprint is dominated by a generic "introduction $\rightarrow$ literature review $\rightarrow$ conclusion" template that omits the paired survey's central contributions.

**Structural footprint.** Table 21 contrasts the second-level outline coverage. The paired human-written survey (Vidgen and Derczynski, 2020) organizes its body around the *full life-cycle* of an abusive-language training dataset: task formation, content properties, annotation process, annotator identity, annotation guidelines, dataset sharing, an open-science infrastructure proposal (`hatespeechdata.com`), and an evidence-based best-practice checklist. SciSage's outline collapses this seven-stage life cycle into a generic *Literature Review* section with two thematic subsections.

| Topical block in human survey | Human | SciSage |
|---|---|---|
| Task formation & detection typology | ✓ | — |
| Dataset content properties | ✓ | — |
| Annotation process & methodology | ✓ | partial |
| Annotator identity & demographics | ✓ | — |
| Annotation guidelines | ✓ | — |
| Dataset sharing & open science | ✓ | — |
| Best-practice recommendations | ✓ | — |
| Total L1 chapters | 8 | 4 |
| Total L2 subsections | 17 | 4 |

Table 21: Outline-level coverage comparison on the same topic. SciSage retains a generic survey template, fully omitting six of the paired survey's seven central contributions and only partially covering the seventh (annotation process).

**Local fidelity is preserved.** Despite this structural collapse, individual SciSage paragraphs are factually grounded and cite the same evidence that appears in the paired survey. For example, the *Recent Advancements* subsection reads:

> *"A systematic review of 63 publicly available datasets highlights the difficulties in producing large, diverse, and theoretically informed collections while minimizing biases … specialized datasets like L-HSAB … emphasize rigorous data collection procedures and annotation guidelines … inter-annotator agreement metrics like Cohen's Kappa and Krippendorff's alpha to validate annotation quality."*

All three referenced facts (63-dataset systematic review, L-HSAB, Cohen's $\kappa$ / Krippendorff's $\alpha$) align with the paired human-written paper, which explains why such outputs accumulate non-trivial token-level overlap ($\tau$-MaxSim) with the human reference.

**Global narrative is missing.** However, this local fidelity does not survive aggregation. SciSage's two thematic subsections (*Recent Advancements*, *Current State of Research*) collapse the paired paper's distinct chapters on (i) *who annotated* the data, (ii) *how the annotation guidelines treat irony, calumny, and intent*, (iii) *how datasets degrade under platform Terms-of-Service constraints*, and (iv) *the proposed data-trust infrastructure for sharing* into a single undifferentiated "advances and challenges" block. The paired survey's central proposals (the `hatespeechdata.com` repository, the four-stage best-practice checklist, and the open-science / data-trust framework) have no corresponding section in SciSage's output.

**Takeaway.** The case operationalises the *local-alignment / global-structure gap* flagged in Finding 5: SciSage retrieves correct factual atoms (high $\tau$-MaxSim), but the discipline-specific narrative scaffold that turns those atoms into a usable survey is replaced by a one-size-fits-all template (low RA-AlignF1, low rubric Outline score). Together, Cases A and B show that ASG failure modes are not symmetric: **AutoSurvey overproduces text without adding coverage**, whereas **SciSage retrieves correct facts without producing the right structure**, two opposite manifestations of the same underlying gap between local segment-level alignment and global disciplinary organization.

## G Prompts

To ensure reproducibility and provide a clear understanding of our framework, we detail the specific prompts used in our experiments. These prompts were designed to guide the LLM effectively across different stages of the pipeline. They are categorized into three main functional phases:

- **Criteria Construction Phase:** This phase focuses on establishing comprehensive evaluation standards. Figures 10, 11, and 12 illustrate the prompts used to aggregate high-level aspects, expand them into detailed criteria based on survey content, and merge criteria from various sources, respectively. This process ensures that the resulting criteria are both diverse and specific to the domain.
- **Evaluation Phase:** We adopt a modular strategy to evaluate the quality of the surveys. We present the prompts designed for the per-aspect evaluation of the survey's **outline** (Figure 13), **content** (Figure 14), and **references** (Figure 15). Evaluating these components independently helps minimize interference and ensures a more accurate assessment.
- **Generation Phase:** Figure 16 presents the prompt used for the automated generation of systematic literature reviews. This prompt guides the model to synthesize the collected references into a complete paper based on the topic provided by the user.

## Aggregate Aspects Prompt

**System prompt:** You are an academic evaluation expert. I need you to aggregate the following evaluation criteria for {category_name} into {target_count} universal, highly aggregated aspects.

**Current aspects list:**
{formatted_aspects}

**Requirements:**

- Aggregate these {len(aspects_list)} aspects into {target_count} universal aspects
- Each aggregated aspect should:
  - Be a highly abstract, universal concept
  - Cover multiple related original aspects
  - Have a concise and generalizable name
  - Have clear descriptions that reflect its universal nature
- Maintain academic rigor, ensuring the aggregated aspects comprehensively cover the core content of the original aspects
- The number of returned aspects must be exactly {target_count}

**Please return the result strictly in the following JSON array format, without any additional text:**

```
[
{"Aspect Name 1": "Detailed description 1"},
{"Aspect Name 2": "Detailed description 2"},
...
]
```

**Note:** Return only the JSON array, without any markdown code block markers or other text.

Figure 10: Prompt used to aggregate multiple evaluation aspects into a smaller set of universal, highly aggregated aspects. Used in the `call_llm_for_aggregation` function.

## Expand Aspects Prompt

**System prompt:** You are an academic survey expert. Now you are refining the writing guideline of survey writing.

**Given:**

1. Component Type: {component_name} (outline/content/reference)
2. Component Content: The actual {component_name} from a survey paper
3. Aspect: {aspect_name} - {aspect_description}
4. Target Number: {n}

**Task:**
Expand the given aspect "{aspect_name}" into {n} specific, detailed criteria. Each criterion should:

- Provide a clear, actionable explanation of what "{aspect_name}" means in the context of {component_name}
- Include a concrete example from the provided component content that demonstrates this criterion
- The example should be a summary/synthesis of relevant parts from the component content, NOT a direct verbatim quote
- Be specific enough to guide evaluation of survey {component_name}

**Component Content:**
{component_content}

**Aspect to Expand:**
Name: {aspect_name}
Description: {aspect_description}

**Please return the result in the following JSON format:**

```
{
"aspect_name": "{aspect_name}",
"expanded_criteria": [
{
"criterion_name": "Specific criterion name 1",
"description": "Detailed explanation of what this criterion means",
"example": "Summarized example from the component content that demonstrates
this criterion"
},
{
"criterion_name": "Specific criterion name 2",
"description": "Detailed explanation of what this criterion means",
"example": "Summarized example from the component content that demonstrates
this criterion"
},
...
]
}
```

**Note:** Return only the JSON object, without any markdown code block markers or additional text.

Figure 11: Prompt used to expand a single aspect into detailed criteria using survey content. Each criterion includes description and example. Defined as `EXPAND_ASPECT_PROMPT`.

## Merge Aspect Criteria Prompt

**System prompt:** You are an academic survey expert. Now you are merging multiple expanded criteria from the same aspect across multiple survey papers into a consolidated set.

**Given:**

1. Aspect Name: {aspect_name}
2. Component Type: {component_name} (outline/content/reference)
3. All Expanded Criteria: A list of criteria from the same aspect "{aspect_name}" collected from multiple survey papers
4. Target Number: {n}

**Task:**
Merge all the given expanded criteria into {n} comprehensive, consolidated criteria. Each merged criterion should:

- Synthesize related criteria from different survey papers into a unified, coherent criterion
- Preserve the essential meaning and evaluation focus of the original criteria
- Be general enough to apply across different survey papers while remaining specific enough to guide evaluation
- Avoid redundancy while maintaining comprehensiveness
- Each merged criterion should have a clear, distinct focus
- Extract common patterns and best practices from the examples

**All Expanded Criteria:**
{all_criteria_json}

**Please return the result in the following JSON format:**

```
{
"aspect_name": "{aspect_name}",
"component_name": "{component_name}",
"merged_expanded_criteria": [
{
"criterion_name": "Consolidated criterion name 1",
"description": "Comprehensive description that synthesizes related criteria",
"example": "Highly synthesized and aggregated example that demonstrates universally applicable criteria"
},
{
"criterion_name": "Consolidated criterion name 2",
"description": "Comprehensive description that synthesizes related criteria",
"example": "Highly synthesized and aggregated example that demonstrates universally applicable criteria"
},
...
]
}
```

**Note:** Return only the JSON object, without any markdown code block markers or additional text.

Figure 12: Prompt used to merge criteria from multiple survey papers into consolidated criteria. Synthesizes common patterns across different examples. Defined as `MERGE_ASPECT_CRITERIA_PROMPT`.

## Per-Aspect Outline Evaluation Prompt

**System prompt:** You are a domain reviewer for category: {category}.

**Domain-specific outline aspects and their criteria (JSON):**
{json.dumps(criteria_payload, ensure_ascii=False)}

**Survey outline (cleaned):**
{json.dumps(outline, ensure_ascii=False)}

**Instructions:**
For EACH aspect below, give:
- an integer score from 1 to 5
- a short justification in one sentence

**Return JSON exactly in this format:**

```
{"aspects": [
{"aspect_name": "<string>", "score": <int>, "notes": "<string>"}
]}
```

Figure 13: Prompt for per-aspect evaluation of survey outline quality. Returns individual scores for each aspect, which are then averaged. Used in `_build_prompt_per_aspect` method for outline.

## Per-Aspect Content Evaluation Prompt

**System prompt:** You are a domain reviewer for category: {category}.

**Domain-specific content aspects and their criteria (JSON):**
{json.dumps(criteria_payload, ensure_ascii=False)}

**Survey content sections (trimmed for length):**
{json.dumps(self._trim_sections(content), ensure_ascii=False)}

**Quick stats:**
{json.dumps(stats, ensure_ascii=False)}

**Instructions:**
For EACH aspect below, give:
- an integer score from 1 to 5
- a short justification in one sentence

**Return JSON exactly in this format:**

```
{"aspects": [
{"aspect_name": "<string>", "score": <int>, "notes": "<string>"}
]}
```

Figure 14: Prompt for per-aspect evaluation of survey content quality. Returns individual scores for each aspect, which are then averaged. Used in `_build_prompt_per_aspect` method for content.

## Per-Aspect Reference Evaluation Prompt

**System prompt:** You are a domain reviewer for category: {category}.

**Domain-specific reference aspects and their criteria (JSON):**
{json.dumps(criteria_payload, ensure_ascii=False)}

**Survey reference titles (cleaned):**
{json.dumps([r.get('title') for r in references], ensure_ascii=False)}

**Instructions:**
For EACH aspect below, give:

- an integer score from 1 to 5
- a short justification in one sentence

**Return JSON exactly in this format:**

```
{"aspects": [
{"aspect_name": "<string>", "score": <int>, "notes": "<string>"}
]}
```

Figure 15: Prompt for per-aspect evaluation of survey reference quality. Returns individual scores for each aspect, which are then averaged. Used in `_build_prompt_per_aspect` method for reference.

## Deep Research Agent Survey Generation Prompt

**Given:**

1. Topic: {topic} (user-provided literature review topic)
2. Output Style: Formal academic style suitable for a top-tier journal
3. Output Format: A full systematic literature review with section headings and references

**Task (Step 1: Initial Request):**
Please write a comprehensive systematic literature review paper on: "{topic}"
The review should cover:

1. Introduction and background
2. Key concepts and definitions
3. Historical development and milestones
4. Current state-of-the-art methods and techniques
5. Applications and case studies
6. Challenges and open problems
7. Future research directions
8. Conclusion

Please write in a formal academic style suitable for a top-tier journal. Include section headings and provide references for all sources used.

**Task (Step 2: Full Paper Generation):**
Great. Now please proceed to write the complete literature review as outlined. The review should be comprehensive, well-structured, and rigorous.
Please include:

1. A clear title and abstract
2. Introduction with research motivation and objectives
3. Comprehensive review of existing literature organized by themes or chronology
4. Critical analysis of the field's strengths and weaknesses
5. Identification of research gaps
6. Conclusion summarizing key findings
7. References section with all cited sources

The review should be detailed enough for publication in a reputable academic journal, typically between 3000–8000 words. Please begin writing now.

**Notes:**

- The code sends a multi-turn message sequence:
  1. User: Step 1 initial request (above)
  2. Assistant: Model "clarification" (captured from Step 1 output; if missing, replaced by a short default acknowledgement)
  3. User: Step 2 full-paper instruction (above)

Figure 16: Prompt for batch generation of systematic literature reviews. The script first issues an initial scope request, then prompts the model to produce a full paper for each user-provided topic; references returned by the deep-research model are collected and saved.