Bloom-Eval: A Hierarchical Evaluation Benchmark for Automatic Survey Generation Based on Bloom's Taxonomy

Fei Zhang[1], Zhe Zhao[2], Haibin Wen[1], Tianshuo Wei[1], Zaixi Zhang[3], Chao Yang[4,*], and Ye Wei[1,*]
[1]City University of Hong Kong [2]Stanford University
[3]Princeton University [4]Shanghai Jiaotong University
feizhang010518@gmail.com

Abstract

The rapid advance of automatic survey generation (ASG) has created a critical evaluation challenge. Existing evaluation methods suffer from both cognitive dimensional simplification and methodological unreliability, primarily due to the over-reliance on the "LLM-as-a-Judge" approach. To bridge this gap, we establish Bloom-Eval, a six-tiered benchmark based on Bloom's Taxonomy that reliably evaluates ASG systems by prioritizing deterministic algorithms and introducing our GRADE approach for abstract abilities. Furthermore, we construct a large-scale, cross-disciplinary dataset of over 3,000 high-quality papers. Our empirical study on this benchmark reveals that while leading ASG systems are proficient format organizers, they remain unqualified knowledge integrators. This work aims to redefine ASG evaluation standards, shifting the research focus from the formal mimicry of surface structure to the cognitive deepening of intellectual content. Our method provides the ASG field with a systematic, reproducible, and theoretically grounded benchmark to guide future research.

1 Introduction

A primary objective of artificial intelligence (AI) for science is to evolve AI from a mere information processor into a genuine knowledge generator (Lu et al., 2024; Li et al., 2025; Yuan et al., 2025). Synthesizing the vast and rapidly expanding corpus of academic literature is a critical bottleneck in achieving this goal. For this challenge, automatic survey generation (ASG) (Wang et al., 2024a) serves as an indispensable cognitive bridge (Hu and Wan, 2014; Chen and Zhuge, 2019; Sun and Zhuge, 2019; Hu et al., 2024; Zhu et al., 2023). ASG is the automated process of synthesizing relevant, disparate literature into a survey given a specific research topic. This technology transforms disjointed information into structured knowledge that fuels scientific discovery. However, as ASG has made great strides in emulating human workflows and diversifying content (Yan et al., 2025; Liang et al., 2025; Wang et al., 2025), its further advancement is impeded by two profound limitations in the current evaluation paradigm.

First, a cognitive science perspective reveals that existing evaluation methods suffer from severe cognitive flatness. Metrics like reference overlap rate (Yan et al., 2025) and citation accuracy (Wang et al., 2024a) merely quantify lower-order cognitive skills like Memory and Comprehension. They fail to capture the true academic value of a survey: analytical depth, synthesis coherence, and creative originality. This lack of evaluation dimensions blurs the line between mere content organization and deep knowledge generation, misguiding the focus of ASG research. Second, an over-reliance on the "LLM-as-a-Judge" paradigm creates a critical transparency and reliability crisis. While demonstrating the potential of LLM-based evaluation, recent studies (Wang et al., 2024b; Zheng et al., 2023) also reveal that LLM judges remain prone to inherent biases when assessing subjective metrics. Given the high dependence of ASG evaluation on such subjective metrics, existing methods (Wang et al., 2024a; Liang et al., 2025) typically use simple holistic scoring. These approaches generally employ models to directly output a final score in the absence of explicit reasoning traces, a "black-box" practice that fundamentally strips the evaluation of its interpretability, transparency, and auditability. This lack of transparency veils the true performance of ASG systems, hindering the ability to guide the progress of the ASG field effectively.

To address these challenges, we introduce Bloom-Eval, a benchmark grounded in Bloom's Taxonomy (Anderson and Krathwohl, 2001). This cognitive model categorizes cognitive skills into

*Corresponding authors. E-mail: ye.wei@cityu.edu.hk, yangchao1987@sjtu.edu.cn.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28512–28544
July 2-7, 2026

a six-level hierarchy ranging from basic recall to complex creation, providing a solid theoretical basis to distinguish mere information recall from insightful knowledge construction. Encompassing Memory, Comprehension, Application, Analysis, Evaluation, and Creation, Bloom-Eval enables fine-grained analysis of model capabilities across the full spectrum of intellectual activity. We employ a dual-constraint strategy to improve reliability and transparency. Ultimately, this work provides a diagnostic benchmark to evaluate and guide the development of ASG systems capable of high-level intelligence and creativity.

The contributions of this paper are:

- We establish Bloom-Eval, the first hierarchical evaluation benchmark for ASG, which provides a profound, six-level diagnosis of a system's cognitive abilities.
- We construct a large-scale, cross-disciplinary dataset from recent (2023-2025) top-tier publications, providing a rigorous foundation for evaluating system generalizability.
- To counter the reliability crisis of "black-box" LLM judges, we introduce a dual-constraint methodology that combines deterministic algorithms with our novel and transparent GRADE approach.
- We conduct a comprehensive empirical study that quantitatively reveals a significant and pervasive gap between the higher-order cognitive abilities of ASG systems and those of human experts. We publicly release the Bloom-Eval benchmark and evaluation scripts.[1]

2 Related Work

2.1 Automatic Survey Generation

The development of ASG is rooted in the broader landscape of scientific summarization (Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2019; Wang et al., 2019; Lu et al., 2020; Jin et al., 2020; Xing et al., 2020) and long-form text generation (Fan et al., 2018; Bosselut et al., 2018; Cho et al., 2019; Mao et al., 2022; Kryscinski et al., 2022; Chang et al., 2024), where researchers have explored diverse strategies for information aggregation. Building on these foundations, the specific task of automatic survey generation (ASG) has recently attracted significant research attention (Jha et al., 2015; Sun and Zhuge, 2019; Huang, 2021; Shao et al., 2024). Researchers have proposed various approaches to address distinct challenges within ASG (Hu et al., 2024; Zhu et al., 2023). For example, AutoSurvey (Wang et al., 2024a) proposes a four-stage method for survey generation. To enhance the quality of such outputs, other systems target specific weaknesses. SurveyForge (Yan et al., 2025) learns from human patterns and enhances reference selection using a scholar navigation agent. SurveyX (Liang et al., 2025) enhances content depth and diversity through deep literature preprocessing and visual generation. Addressing the separate challenge of scale, LLMxMapReduce-V2 (Wang et al., 2025) processes long inputs using stacked convolutional scaling layers for deep synthesis.

2.2 Evaluation of ASG systems

ASG evaluation increasingly adopts the "LLM-as-a-Judge" paradigm widely used in general generation (Zheng et al., 2023; Wu et al., 2025; Chen et al., 2025). Recent works apply this paradigm across various quality dimensions. For instance, AutoSurvey (Wang et al., 2024a) employs LLMs to assess aspects like citation quality and content relevance. In terms of structure, SurveyForge (Yan et al., 2025) utilizes LLMs to evaluate outline rationality. SurveyX (Liang et al., 2025) uses semantic similarity and IoU-based reference overlap for broader evaluation. However, despite covering specific functional aspects, these metrics focus on fragmented attributes without a grounded theoretical basis. The field lacks a unified benchmark that organizes these dimensions into a cognitive hierarchy to systematically diagnose higher-order synthesis capabilities.

3 Bloom-Eval

To address the challenges of cognitive flatness and methodological unreliability, this section details the architecture of the Bloom-Eval benchmark (Figure 1). Framework reliability is validated through human correlation studies across all tiers (Appendix G). We begin in Section 3.1 by introducing our dual-constraint strategy. This strategy utilizes two foundational methodologies (distributional analysis and the GRADE assessment approach) as the algorithmic foundation for all subsequent metrics. Following this, in Section 3.2, we elaborate on the specific diagnostic metrics across the six cognitive tiers, demonstrating how the benchmark offers a comprehensive assessment.

[1] https://github.com/feizhang18/Bloom-Eval

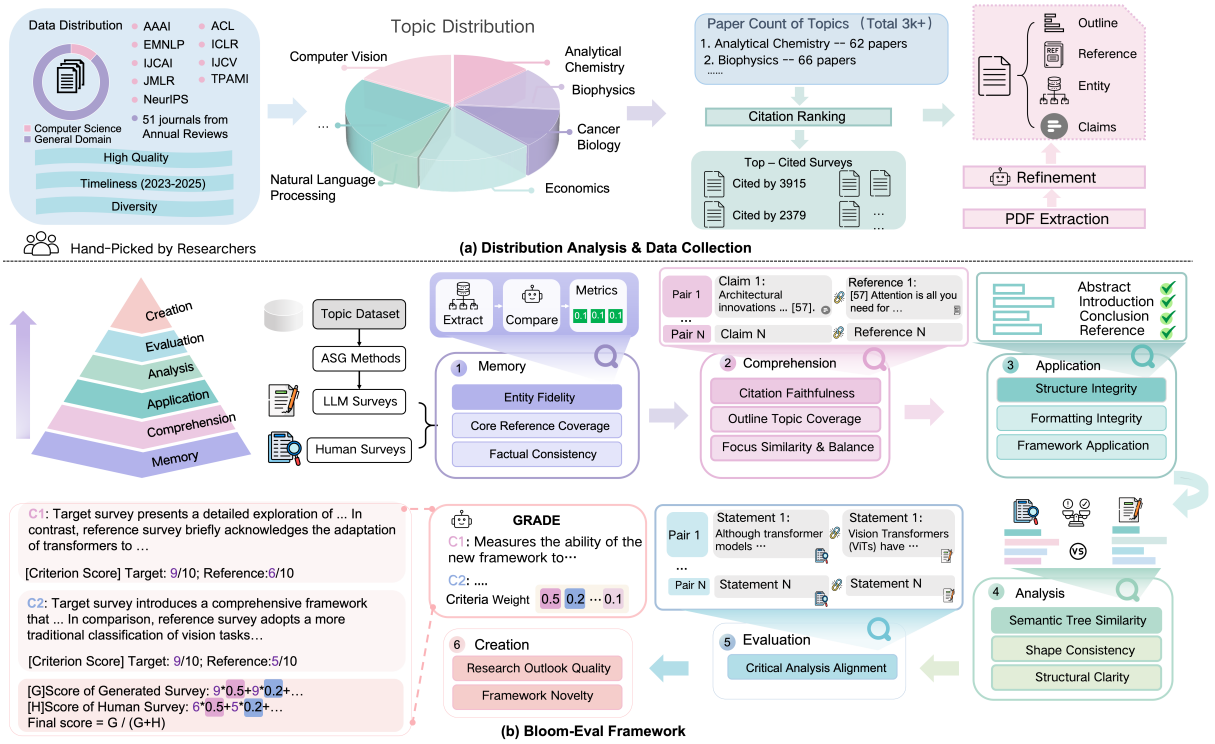

Figure 1: An overview of the Bloom-Eval framework, detailing (a) the data collection pipeline and (b) the six-tiered hierarchical evaluation across the cognitive levels of Memory, Comprehension, Application, Analysis, Evaluation, and Creation.

3.1 Evaluation Methodologies

Our framework is built upon a dual-constraint strategy designed to maximize objectivity and minimize the risk of model hallucination. As summarized in Table 1, this strategy strictly regulates the operational boundary of the LLM based on task subjectivity. For quantifiable metrics, the model is either entirely absent (relying on deterministic algorithms) or confined to specific functional tools (e.g., extraction, matching); for abstract cognitive dimensions, it operates solely as a constrained judge via the transparent GRADE approach. This principled design is the key to our framework's reliability.

To implement this strategy, we integrate standard set-comparison metrics (i.e., F1-score) with two specialized foundational methodologies tailored for distinct assessment types (implementation details in Appendix A):

- **Distributional Similarity (DS):** Standard overlap metrics (e.g., Jaccard) merely check for content presence, failing to detect focus misalignment. We propose DS, a composite metric fusing Jensen-Shannon Divergence, Hellinger Distance, and Total Variation Distance, to diagnose whether the model's focus distribution aligns with expert consensus.
- **Generative Rubric Adaptive Differential Evaluation (GRADE):** To assess abstract abilities, we introduce the GRADE approach, inspired by (Du et al., 2025). Unlike conventional "LLM-as-a-Judge" approaches that produce a single, opaque score, GRADE enforces a transparent process through a two-stage method. First, it externalizes criteria by guiding the Judge LLM to deconstruct the task into explicit, weighted rubrics. Second, it performs justified differential scoring, where the LLM compares the generated survey against a human expert reference. This step requires the model to output not just a numerical score but also a detailed textual analysis justifying the score difference based on the criteria. This ensures judgments are auditable, interpretable, and anchored to verifiable standards.

3.2 Hierarchical Evaluation Dimensions

This section details the diagnostic metrics for each cognitive tier, with mathematical formulations in Appendix B and theoretical mappings to Bloom's Taxonomy in Appendix H.

Level 1: Memory. In Bloom's Taxonomy, Memory refers to the ability to recall information, encompassing facts, concepts, definitions, and termi-

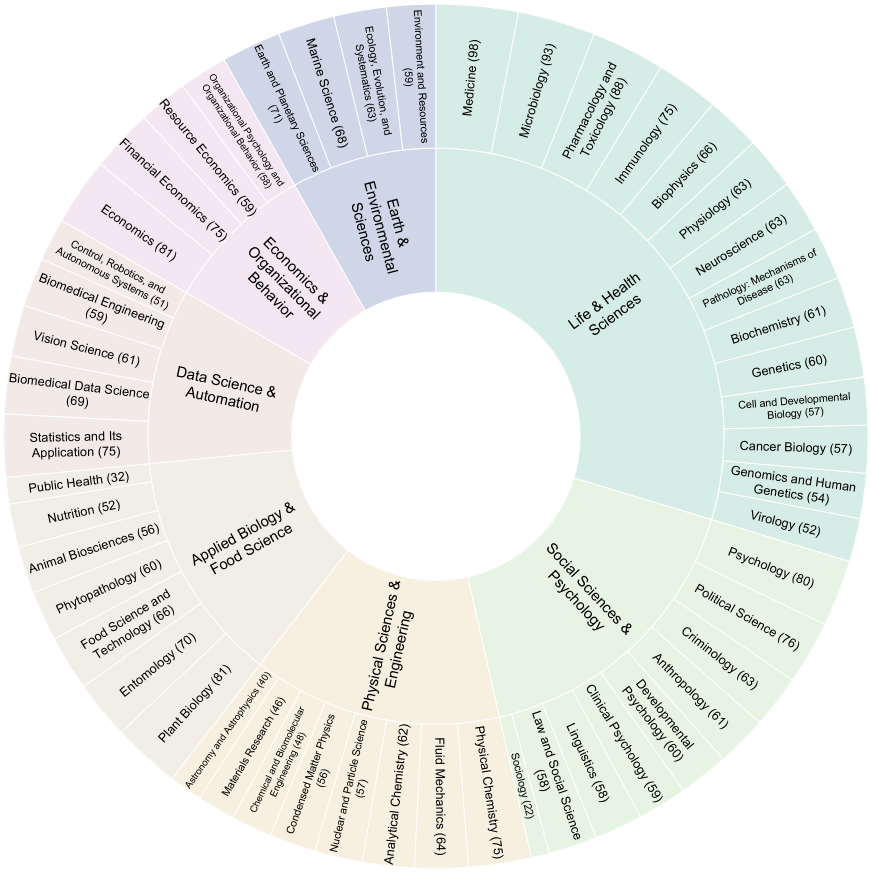

Figure 2: Distribution of Annual Reviews articles across 51 journal series (2023-2025).

Table 1: Overview of Bloom-Eval's metrics, methodologies, and the corresponding role of the LLM. **Method** abbreviations: *E.* (LLM-based Extraction), *M.* (LLM-based Matching), *Alg.* (Algorithm). **LLM Role** definitions: *Tool* for specific tasks (E./M.), *Constrained Judge* for scoring via GRADE approach.

Level	Metric	Method	LLM Role
Memory	EFid	E. + M. + Alg.	Tool
	HIRC	Alg.	None
	FCons	E. + M. + Alg.	Tool
Comprehension	OTC	E. + M. + Alg.	Tool
	CF	E. + M. + Alg.	Tool
	TFSim	Alg. + M.	Tool
	TBal	Alg.	None
Application	FMI	Alg.	None
	DSI	Alg.	None
	FAP	GRADE	Constrained Judge
Analysis	STS	Alg.	None
	SCons	Alg.	None
	SCS	M. + Alg.	Tool
Evaluation	CAA	E. + M. + Alg.	Tool
Creation	FNov	GRADE	Constrained Judge
	ROQ	GRADE	Constrained Judge

nology. Mapped to ASG, this tier mandates that the system accurately reproduce the field's foundational elements: core concepts and terms, pivotal literature, and factual statements. We diagnose these retrieval capabilities via three metrics:

- **Entity Fidelity (EFid):** To diagnose the integrity of domain entity memory, EFid combines the F1-score and DS score of shared entities to measure both the scope of entity coverage and the consistency of relative emphasis.
- **High-Impact Reference Coverage (HIRC):** To diagnose foundational literature memory, HIRC uses the F1-score to verify the inclusion of seminal works (citations > 50) representing the domain's intellectual heritage.
- **Factual Consistency (FCons):** To diagnose factual precision, FCons uses the F1-score to measure whether generated claims are logically entailed by expert ground truth via semantic matching, strictly distinguishing faithful recall from hallucination.

Level 2: Comprehension. In Bloom's Taxonomy, Comprehension refers to the ability to explain and summarize. Mapped to ASG, this requires the

model to summarize content into necessary subheadings, explain cited references, and synthesize the field's research landscape and distributional hotspots. We design four metrics to capture these capabilities, addressing potential failures in topic classification, interpretative faithfulness, and thematic synthesis:

- **Outline Topic Coverage (OTC):** To diagnose the ability to classify and identify necessary subtopics, OTC evaluates the generated survey's coverage of the field's main categories via semantic matching of outline headings.
- **Citation Faithfulness (CF):** To assess the interpretative accuracy of in-text citations, CF measures whether a statement correctly paraphrases and summarizes the cited paper. We report F1-score following (Wang et al., 2024a).
- **Thematic Focus Similarity (TFSim):** To assess the alignment of thematic understanding, TFSim uses BERTopic (Grootendorst, 2022) to categorize references and calculates the F1-score and DS score to quantify the consistency of the generated research focus relative to experts.
- **Thematic Balance (TBal):** To measure the distributional evenness of the model's research attention, TBal quantifies the equality of cited references across topics. It uses the Gini Coefficient (Gini, 1912) to detect disproportionate fixation, ensuring the model allocates attention evenly rather than exhibiting bias toward specific subfields.

Level 3: Application. In Bloom's Taxonomy, Application refers to applying learned knowledge to practice. Mapped to ASG, this entails applying standard academic survey frameworks, specifically necessitating basic formatting correctness, structural completeness, and framework application capabilities. We capture these capabilities via:

- **Formatting Integrity (FMI):** To diagnose adherence to citation conventions, FMI assesses the set consistency between in-text markers and the bibliography to verify the strict execution of academic referencing standards.
- **Document Structure Integrity (DSI):** To diagnose the application of standard academic templates, DSI evaluates the presence of essential structural components (e.g., Abstract, Introduction) to verify compliance with canonical publication norms.
- **Framework Application (FAP):** To diagnose the application of organizing paradigms, FAP uses the GRADE approach to evaluate the consistent implementation of a recognized taxonomy (e.g., chronological), assessing the transfer of abstract organizational logic to the specific survey topic.

Level 4: Analysis. In Bloom's Taxonomy, Analysis is the ability to break down information and identify structures or relationships. Mapped to ASG, Analysis implies ensuring an accurate writing structure, characterized by correct hierarchical logic, appropriate granularity, and structural clarity. Treating the survey's outline as a proxy for its knowledge structure, we propose the following metrics to capture these capabilities:

- **Semantic Tree Similarity (STS):** To diagnose the soundness of hierarchical logic, STS quantifies the structural and semantic similarity between generated and expert outlines (modeled as semantic trees) via edit distance, verifying consistency in relational logic.
- **Shape Consistency (SCons):** To diagnose the appropriateness of analytical granularity, SCons compares the structural complexity (depth and breadth) of the generated and expert outlines, ensuring the analysis avoids becoming superficially flat or overly narrow.
- **Structural Clarity Score (SCS):** To diagnose the precision of deconstruction, SCS strictly penalizes semantically redundant sections appearing in disparate branches, ensuring the clear separation of distinct concepts without confusing repetition.

Level 5: Evaluation. In Bloom's Taxonomy, Evaluation is defined as the ability to critically judge the value of information. Mapped to ASG, this entails critiquing the conclusions and limitations of existing literature. We design the following metric to capture this capability:

- **Critical Analysis Alignment (CAA):** To diagnose the quality of evaluative reasoning, CAA measures the F1-score overlap of explicit value judgments (e.g., method flaws) between generated and expert surveys, verifying if critical insights are both accurate and exhaustive.

Level 6: Creation. Representing the highest order of cognition, Creation emphasizes innovation, requiring the capacity to propose new viewpoints, models, or works. Mapped to ASG, this tier

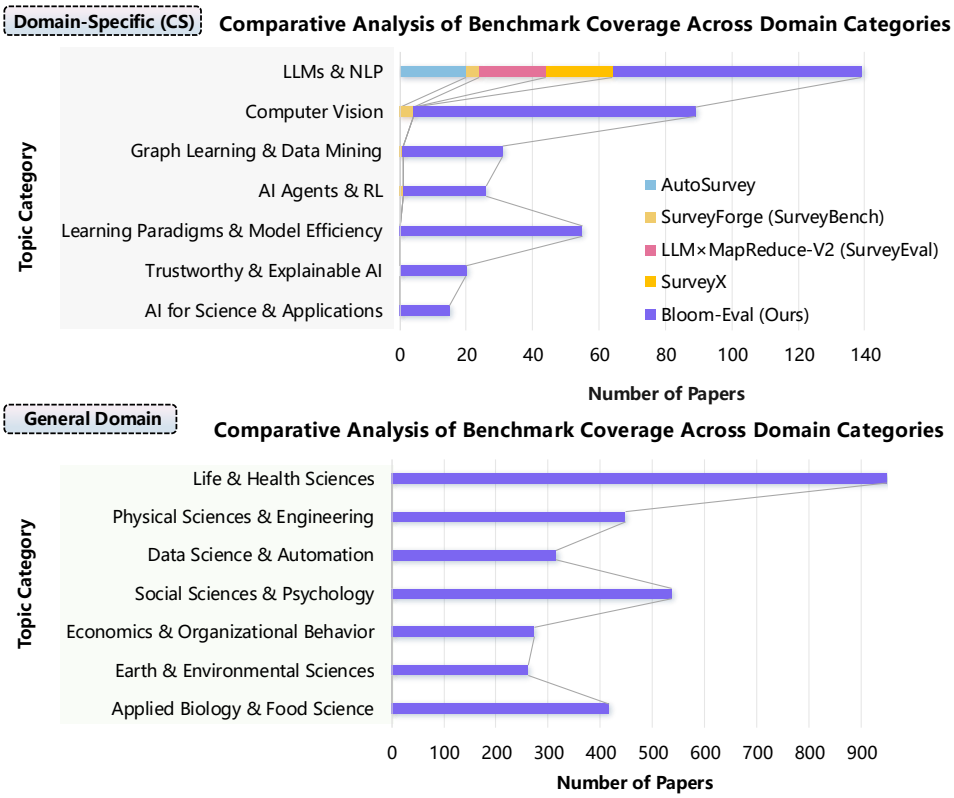

Figure 3: Comparative analysis of benchmark coverage across domain categories.

specifically mandates framework novelty (reorganizing the field with fresh perspectives) and future work discovery (identifying potential research directions). We evaluate two key capabilities:

- **Framework Novelty (FNov):** To diagnose the creation of new conceptual models, FNov uses GRADE to evaluate the originality and insightfulness of the proposed framework, assessing whether it reorganizes the research field with a fresh interpretive lens rather than merely replicating existing schemas.
- **Research Outlook Quality (ROQ):** To diagnose the discovery of future work, ROQ measures the foresight and strategic value of the research directions proposed by the system, typically found in the conclusion or future work sections.

4 Experiments

4.1 Benchmark Construction and Topic Selection

To conduct a comprehensive, equitable, and rigorous evaluation of ASG systems, we first established a benchmark characterized by its alignment with the scientific frontier, high quality, and disciplinary diversity.

Benchmark Corpus Construction. We constructed an authoritative corpus of 3,506 manually verified survey papers as the expert ground truth, curated based on three core principles: quality, scope, and timeliness. To ensure quality, all papers were selected from top-tier academic conferences (e.g., ICLR, ACL) and prestigious journal series (e.g., Annual Reviews), strictly excluding content from preprint servers like arXiv that has not undergone rigorous peer review. To achieve a broad scope for cross-domain evaluation, the corpus spans both Computer Science (CS) and a wide array of General Science disciplines. The disciplinary breakdown of the General Science portion, sourced from the Annual Reviews series, is detailed in Figure 2. Finally, all papers were published between 2023-2025 to ensure timeliness and minimize data leakage risks.

A comparative analysis presented in Figure 3 demonstrates the superiority of our benchmark over prior works (Wang et al., 2024a; Yan et al., 2025; Liang et al., 2025; Wang et al., 2025).

The analysis reveals two critical advantages of Bloom-Eval. First, regarding scope, the figure shows that while existing benchmarks are almost exclusively confined to a few computer science

Table 2: Data source composition and scale of survey generation benchmarks.

Venue	AutoSurvey	SurveyForge (SurveyBench)	LLM×MapReduce-V2 (SurveyEval)	SGSimEval	SurveyX	Bloom-Eval (Ours)
arXiv	✓	✓	✓	✓	✓	✗
AAAI	✗	✗	✗	✗	✗	✓
ACL	✗	✗	✗	✗	✗	✓
EMNLP	✗	✗	✗	✗	✗	✓
ICLR	✗	✗	✗	✗	✗	✓
IJCAI	✗	✗	✗	✗	✗	✓
IJCV	✗	✗	✗	✗	✗	✓
JMLR	✗	✗	✗	✗	✗	✓
NeurIPS	✗	✗	✗	✗	✗	✓
TPAMI	✗	✗	✗	✗	✗	✓
Annual Reviews (51)	✗	✗	✗	✗	✗	✓
Total Venues	1	1	1	1	1	**60**
Total Papers	20	100	384	80	20	**3,506**

sub-fields, our benchmark provides balanced coverage across all 14 domains, including comprehensive collections in the general sciences. Second, in terms of scale, our dataset of 3,506 papers is substantially larger than its predecessors.

Table 2 highlights the contrast in data source quality and scale. While prior works (Guo et al., 2025) rely almost exclusively on arXiv preprints, our benchmark is built upon a diverse set of 60 top-tier, peer-reviewed venues, including 9 major CS conferences and 51 journals from the Annual Reviews series. This distinction in data sources underscores the superior authoritativeness and reliability of our benchmark.

Topic Selection. As a full-scale evaluation was impractical, we selected 20 representative topics based on impact. We chose the 10 most-cited papers from both our "Computer Science" and "General Science" categories. The titles of these 20 papers became our evaluation topics, guaranteeing a test set that is authoritative, diverse, and timely. These 20 topics represent specific, well-defined academic questions rather than broad concepts, providing precise targets for assessing the performance of ASG systems.

4.2 Experimental Setup

Implementation Details. For reproducibility, we use the gpt-5-mini model with a temperature of 0.0 for internal tasks like content extraction, matching and scoring (see Appendix I for the exact prompts used). To ensure comparability with prior work, we follow the methodology of AutoSurvey (Wang et al., 2024a) for specific modules, using gpt-4o for citation metrics and nomic-ai/nomic-embed-text-v1 for text vectorization.

Evaluated ASG Methods. For a fair comparison, all ASG methods were evaluated using their official default settings from public codebases (Wang et al., 2024a; Yan et al., 2025) or publicly available online services (Liang et al., 2025; Wang et al., 2025). Our benchmark includes topics from both Computer Science (CS) and general science domains to test system robustness and generalization. We acknowledge that some systems, such as AutoSurvey (Wang et al., 2024a), are designed specifically for the CS domain. Their inclusion here is intended not for direct comparison on an unfair basis, but to systematically measure their out-of-domain performance and understand the capability boundaries of current tools. Therefore, the performance of such systems on non-CS topics should be interpreted in this context. This approach provides a valuable baseline and reference for exploring the cross-domain adaptability of ASG systems.

4.3 Main Results

Our comprehensive evaluation, with detailed results in Table 3 and a high-level summary in Figure 4, reveals a substantial capability gap between four leading ASG systems and human experts. While LLM systems show competence in structured, formal tasks, they fall significantly short in higher-order cognitive dimensions requiring deep understanding and critical thinking.

Lower-Order Skills (Memory & Comprehension). At the foundational levels, all systems exhibit significant shortcomings in factuality and faithfulness, with average Memory scores below 0.2. A severe inability to recall core, high-impact literature is evident, as no system's High-Impact Reference Coverage (HIRC) F1 score exceeded

Table 3: Comprehensive performance of ASG systems across the six cognitive tiers of Bloom-Eval. The Human column provides reference scores where a dash (-) indicates a definitional value: 1.0 for gold-standard comparisons (e.g., F1-score) and 0.5 as the reference point for human parity in the GRADE metrics. Other numerical scores are empirical measurements of the expert-written surveys. Detailed results in Appendix D.

		Domain-Specific Systems		General Domain Systems		Human
Level	**Metric**	**AutoSurvey**	**SurveyForge**	**LLMxMap Reduce-V2**	**SurveyX**	
Memory	Entity Fidelity (F1/DS)	0.10 / 0.81	0.13 / 0.73	0.12 / 0.69	0.11 / 0.78	-
	High-Impact Reference Coverage (F1)	0.03	0.07	0.02	0.04	-
	Factual Consistency (F1)	0.07	0.06	0.09	0.07	-
	Avg. Score	0.19	0.19	0.17	0.19	-
Comprehension	Outline Topic Coverage (F1)	0.40	0.61	0.53	0.56	-
	Citation Faithfulness (F1)	0.03	0.27	0.29	0.37	0.47
	Thematic Focus Similarity (F1/DS)	0.43 / 0.73	0.53 / 0.78	0.41 / 0.67	0.52 / 0.81	-
	Thematic Balance	0.72	0.70	0.68	0.68	0.70
	Avg. Score	0.43	0.56	0.51	0.57	-
Application	Formatting Integrity	1.00	1.00	0.95	1.00	0.99
	Document Structure Integrity	0.74	0.75	0.69	1.00	1.00
	Framework Application	0.38	0.42	0.52	0.41	-
	Avg. Score	0.71	0.72	0.72	0.80	-
Analysis	Semantic Tree Similarity	0.39	0.55	0.53	0.56	-
	Shape Consistency	0.54	0.74	0.72	0.75	-
	Structural Clarity Score	0.84	0.75	0.84	0.79	0.89
	Avg. Score	0.59	0.68	0.70	0.70	-
Evaluation	Critical Analysis Alignment (F1)	0.13	0.08	0.11	0.13	-
	Avg. Score	0.13	0.08	0.11	0.13	-
Creation	Framework Novelty	0.44	0.41	0.53	0.41	-
	Research Outlook Quality	0.51	0.47	0.56	0.48	-
	Avg. Score	0.48	0.44	0.55	0.45	-

0.07. Low Factual Consistency scores further expose issues with hallucination. While Entity Fidelity F1 scores were low, the high DS scores (0.69–0.81) suggest that when entities are correctly identified, their discussion frequency is human-like. Performance in Comprehension was slightly better (average scores 0.43–0.57), with models generally able to identify main topics (high OTC scores). However, universally low Citation Faithfulness (CF) scores highlight a critical failure to accurately understand and paraphrase source material.

Mid-Level Skills (Application & Analysis). In the mid-level tiers, ASG systems reveal a stark divide between superficial formatting and deep structural understanding, demonstrating a tendency for "form over substance". The Application tier is particularly illustrative of this split. While systems achieve near-perfect scores in low-level procedural tasks like Formatting Integrity (FMI) and Document Structure Integrity (DSI), their ability to apply a coherent high-level framework is notably weak. With the exception of LLMxMapReduce-V2, all systems scored well below the 0.5 human-parity baseline on Framework Application (FAP), indicating they are excellent "format organizers" but poor "architectural thinkers". In Analysis, high Structural Clarity (SCS) scores confirm that the generated outlines are internally coherent. However, this contrasts with moderate Semantic Tree Similarity (STS) scores (0.39–0.56), indicating that while structurally sound, the systems' knowledge organization still differs significantly from an expert's.

Higher-Order Skills (Evaluation & Creation). The most critical bottlenecks appear at the highest cognitive levels. In Evaluation, all systems performed exceptionally poorly, with average scores clustering around 0.1, providing strong evidence that they almost completely lack human-aligned critical judgment. The gap is even more pronounced in Creation. Contrary to being creative, most systems failed to reach the 0.5 human-expert

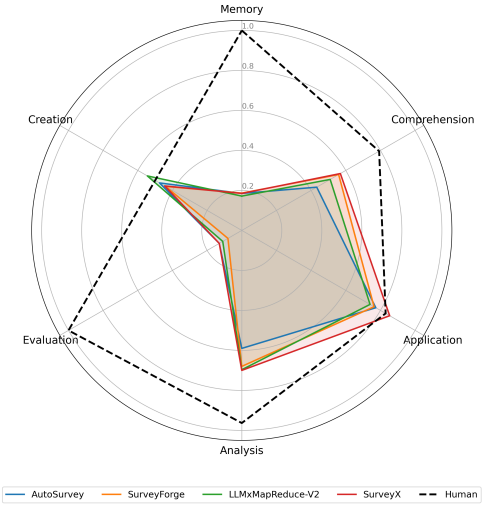

Figure 4: High-level performance of ASG systems across the six cognitive tiers of Bloom-Eval. Each axis score is the arithmetic mean of its tier's constituent metrics. For composite metrics (P/R/F1), the F1 score is used; for metrics with both F1 and DS, their average is taken.

parity baseline, with average scores as low as 0.44. This demonstrates a significant and quantifiable lack of creative ability, rather than an "illusion of creativity". Only LLMxMapReduce-V2 managed to slightly exceed this baseline, suggesting a nascent but still limited creative capacity.

Overall System Comparison. No single system dominates across all dimensions. SurveyX emerges as the strongest all-around system for processing and structuring information, achieving the top scores in Application (0.80) and Comprehension (0.57), and tying for the lead in Analysis (0.70). In contrast, LLMxMapReduce-V2 displays a more specialized strength: it is the only system to surpass the human-parity baseline in both Framework Application (0.52) and Creation (0.55), but its performance in lower-order skills like Memory is weaker. SurveyForge demonstrates solid mid-range performance, particularly in Comprehension and Analysis, but does not lead in any single category. Overall, our results profile current ASG systems as excellent "structural mimics" but deficient "knowledge integrators" and genuine "idea creators". A significant journey remains before they can be considered reliable intellectual partners.

Efficiency Analysis. We also examined operational efficiency (see Appendix Table 9). The results highlight distinct resource profiles: LLMxMapReduce-V2 is the most token-intensive (>72k tokens), whereas SurveyX exhibits the highest latency (126 min). In comparison, SurveyForge demonstrates balanced efficiency, aligning with human-level length (≈26k tokens) while accelerating the writing process by nearly $40\times$.

5 Conclusion and Outlook

We introduced Bloom-Eval, a hierarchical framework that reveals a critical gap in automatic survey generation (ASG): current systems are proficient "format organizers" but fail at higher-order cognitive tasks like critical analysis and creativity. This finding suggests that without a change in evaluation, the field risks optimizing for well-structured yet intellectually shallow content. Our work calls for a research paradigm shift from "formal mimicry" to "cognitive deepening". Bloom-Eval provides the diagnostic compass to guide this journey toward creating truly intelligent systems.

Limitations

While Bloom-Eval offers a more systematic and fine-grained framework for evaluating ASG systems, we acknowledge several limitations in our framework design and experimental methodology. First, regarding the evaluation framework itself, although our proposed GRADE approach is designed to enhance transparency, the assessment of abstract, high-order abilities such as "creativity" ultimately relies on the judgment of a large language model, which cannot fully escape its inherent subjective preferences. Second, our evaluation is constrained by the scope of the constructed benchmark. Despite our efforts to ensure its scale and interdisciplinary nature, the existing corpus of over 3,000 papers and 20 evaluation topics cannot fully represent the entire spectrum of scientific research domains.

References

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *ICLR*.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. 2025. Mlr-bench: Evaluating ai agents on open-ended machine learning research. In *NeurIPS*.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2019. Towards coherent and cohesive long-form text generation. In *Proceedings of the First Workshop on Narrative Understanding*.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*.

Josue Balandrano Coronel. 2016. University of texas rio grande valley trec liveqa 2016: Using topic modeling to answer complex questions. In *TREC*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Corrado Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill.

Beichen Guo, Zhiyuan Wen, Yu Yang, Peng Gao, Ruosong Yang, and Jiaxing Shen. 2025. Sgsimeval: A comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems. In *ADMA*.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Coling*.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *EMNLP*.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024. Hireview: Hierarchical taxonomy-driven automatic literature review generation. *arXiv preprint arXiv:2410.03761*.

Hen-Hsen Huang. 2021. Autosurvey: Automatic survey generation based on a research draft. In *IJCAI*.

Rahul Jha, Catherine Finegan-Dollak, Ben King, Reed Coke, and Dragomir Radev. 2015. Content models for survey generation: a factoid-based evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Haozhe Ji, Pei Ke, Zhipeng Hu, Rongsheng Zhang, and Minlie Huang. 2023. Tailoring language generation models under total variation distance. In *ICLR*.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *ACL*.

J. Peter Kincaid, Jr. Fishburne, Robert P., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington, TN.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *EMNLP*.

Rémi Lebret and Ronan Collobert. 2014. Word embeddings through hellinger pca. In *EACL*.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xinxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, and 1 others. 2025. Chain of ideas: Revolutionizing research in novel idea development with llm agents. In *EMNLP*.

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*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *EMNLP*.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In *ACL*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

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

Xiaoping Sun and Hai Zhuge. 2019. Automatic generation of survey paper based on template tree. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*.

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*.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the 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. 2024a. Autosurvey: Large language models can automatically write surveys. In *NeurIPS*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *ICLR*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. In *NeurIPS*.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *ACL*.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 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)*.

Jiakang Yuan, Xiangchao Yan, Botian Shi, Tao Chen, Wanli Ouyang, Bo Zhang, Lei Bai, Yu Qiao, and Bowen Zhou. 2025. Dolphin: Closed-loop open-ended auto-research through thinking, practice, and feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. *arXiv preprint arXiv:2304.03512*.

A Implementation Details of Evaluation Methodologies

This appendix presents the mathematical formulations and implementation details of the evaluation methodologies introduced in the Bloom-Eval benchmark.

A.1 Distributional Similarity (DS)

To quantify the distributional alignment between sets (e.g., entities or topics), we propose a composite metric named Distributional Similarity (DS). We model item frequencies as two probability distributions, P and Q. Frequency counts from the expert (C_{E}) and generated (C_{G}) documents are normalized into the probability mass functions P and Q as follows:

$$p_i = \frac{c_{\mathrm{E},i}}{\sum_{j=1}^{n} c_{\mathrm{E},j}} \quad \text{and} \quad q_i = \frac{c_{\mathrm{G},i}}{\sum_{j=1}^{n} c_{\mathrm{G},j}}. \quad (1)$$

The DS score is then calculated as a multi-metric fusion of three distances empirically validated in NLP contexts (Coronel, 2016; Lebret and Collobert, 2014; Ji et al., 2023): the Jensen-Shannon Divergence d_{JS}, the Hellinger Distance d_{H}, and the Total Variation Distance d_{TV}. Each distance is converted to a similarity score between 0 and 1, and then averaged. The Hellinger and Total Variation distances are used directly as their values naturally fall within the $[0, 1]$ range. The Jensen-Shannon Divergence is first normalized by its theoretical maximum:

$$\hat{d}_{\mathrm{JS}} = \frac{d_{\mathrm{JS}}(P, Q)}{\ln 2}. \quad (2)$$

The final DS score is then computed as

$$\mathrm{DS}(P, Q) = \frac{1}{3}\left((1 - \hat{d}_{\mathrm{JS}}) + (1 - d_{\mathrm{H}}) + (1 - d_{\mathrm{TV}})\right).$$

A higher DS score indicates greater similarity, with 1 signifying identical distributions.

A.2 Generative Rubric Adaptive Differential Evaluation (GRADE)

1. **Rubric Generation:** A "Judge LLM" first deconstructs the evaluation task by generating a specific rubric. This yields a set of k criteria $C = \{c_1, \ldots, c_k\}$ and corresponding weights $W = \{w_1, \ldots, w_k\}$, where $\sum w_i = 1$. This step externalizes the LLM's reasoning, making it explicit and auditable.
2. **Differential Scoring:** The LLM then uses this public rubric to perform a comparative evaluation, assigning a score $s_i \in [0, 10]$ to each criterion for both the generated survey and a human expert's reference.

An intermediate score for any document is calculated as the weighted sum of its criterion scores:

$$\text{Score}_{\text{doc}} = \sum_{i=1}^{k} s_i \cdot w_i. \tag{3}$$

The final GRADE score is then computed as a relative measure between the generated intermediate score $\text{Score}_{\text{Generated}}$ and the human's score $\text{Score}_{\text{Human}}$:

$$\text{GRADE Score} = \frac{\text{Score}_{\text{Generated}}}{\text{Score}_{\text{Generated}} + \text{Score}_{\text{Human}}}. \tag{4}$$

A score of 0.5 indicates parity with the human expert, while a score greater than 0.5 suggests superior performance by the ASG system on that specific abstract ability. This process transforms the LLM from an unaccountable "black-box" judge into a tool for structured, auditable assessment.

B Implementation Details of Hierarchical Metrics

This section provides the specific mathematical formulations and implementation parameters for the diagnostic metrics across the six cognitive tiers.

B.1 Level 1: Memory Metrics

Entity Fidelity (EFid). To assess domain knowledge retention, we employ a Judge LLM to extract fine-grained entities from both the generated and expert surveys, specifically targeting four categories: methods, models, datasets, and evaluation metrics. Based on these extractions, we calculate the standard F1-score to quantify the set overlap between the generated and expert entities, as well as the Distributional Similarity (DS) score (defined in Appendix A.1) to evaluate the alignment of frequency vectors for shared entities.

High-Impact Reference Coverage (HIRC). This metric measures the F1-score of seminal-work identification. Citation counts for references are retrieved via Google Scholar. We define the expert high-impact set R_{high} as references in the expert bibliography with citation counts $c > 50$, and let R_{gen} denote all references cited in the generated survey. We compute precision and recall and the final HIRC score is the F1-score. This metric assesses whether the generated content preserves the cornerstone research that constitutes the domain's intellectual heritage.

Factual Consistency (FCons). To distinguish faithful recall from hallucination, we implement a two-stage LLM-based verification pipeline that begins by extracting atomic factual claims from the generated text. A Judge LLM then evaluates each claim through semantic matching against the expert survey to determine entailment, verifying support from the ground truth.

B.2 Level 2: Comprehension Metrics

Outline Topic Coverage (OTC). We employ an LLM to extract all headings from the outlines of both the generated and expert surveys. Subsequently, the LLM performs semantic matching between the two sets of headings to identify overlapping topics. Based on the matched set, we compute the F1-score.

Citation Faithfulness (CF). We adopt the Citation Recall and Precision metrics from (Wang et al., 2024a) to distinguish faithful interpretation from hallucination. Specifically, for each citation, we retrieve the title and abstract of the referenced paper. A Judge LLM then determines whether the specific statement in the survey is textually supported by the retrieved reference context. Based on these judgments, we compute the final F1-score.

Thematic Focus Similarity (TFSim). TFSim measures whether the model captures the same research clusters as experts. We utilize the nomic-embed-text-v1 model (Nussbaum et al., 2025) to generate semantic embeddings for the concatenated titles and abstracts of all cited references in both bibliographies. These embeddings are then clustered to identify shared research themes. Based on the clustering results, we calculate the F1-score for theme overlap and the Distributional Similarity (DS) score for the frequency distribution of these themes.

Thematic Balance (TBal). Using the topic counts derived from the bibliography clustering analysis, TBal quantifies the balance of topic coverage within a single survey. We apply the standard Gini Coefficient (G) (Gini, 1912) to capture the distributional inequality. The final metric is defined as $1 - G$:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i}, \tag{5}$$

where x_i is the number of references in theme i. A higher TBal score thus reflects a more comprehensive research focus.

B.3 Level 3: Application Metrics

Formatting Integrity (FMI). This metric assesses the consistency between in-text citations and the bibliography to detect errors such as undefined citation markers and extraneous bibliographic entries. It is calculated as the Jaccard similarity between the set of citation identifiers found in the text C_{cited} and those defined in the bibliography R_{defined}:

$$\text{FMI} = \frac{|C_{\text{cited}} \cap R_{\text{defined}}|}{|C_{\text{cited}} \cup R_{\text{defined}}|}. \tag{6}$$

Document Structure Integrity (DSI). This checklist-based metric measures the presence of pre-defined, essential sections S_{req} (e.g., Abstract, Introduction, Conclusion, References) within the set of sections extracted from the generated document, denoted as S_{found}. It verifies whether the model correctly constructs the necessary skeleton of a formal survey paper. The score is calculated as:

$$\text{DSI} = \frac{\sum_{s \in S_{\text{req}}} \mathbb{I}(s \in S_{\text{found}})}{|S_{\text{req}}|}, \tag{7}$$

where $\mathbb{I}(\cdot)$ is an indicator function.

Framework Application (FAP). To evaluate the application of organizing paradigms, FAP adopts the GRADE approach. The Judge LLM is first instructed to synthesize bespoke, topic-specific scoring rubrics regarding the selection and consistent implementation of recognized taxonomies (e.g., chronological, methodological). Subsequently, utilizing these generated rubrics, the LLM evaluates both the ASG-generated survey and the human expert survey to assess the transfer of abstract organizational logic, ultimately deriving the relative GRADE score.

B.4 Level 4: Analysis Metrics

Semantic Tree Similarity (STS). This metric quantifies the hierarchical and semantic similarity between the generated outline (Tree A) and the expert outline (Tree B). We parse both outlines into tree structures and compute their similarity using the Tree Edit Distance (TED) algorithm. The cost for node insertion and deletion is 1, while the update cost for a node (i.e., a section heading) is its semantic distance, defined as $1 - \cos(\theta)$ based on vector embedding similarity. The final normalized score is calculated as:

$$\text{Similarity} = 1 - \frac{\text{TED}(A, B)}{|A| + |B|}, \tag{8}$$

where $|A|$ and $|B|$ are the number of nodes in each tree. A score approaching 1 indicates high structural and semantic congruence. By modeling outlines as semantic trees and computing their edit distance, this metric verifies whether the model organizes sub-topics with relational logic (e.g., parent-child relationships) consistent with human experts.

Shape Consistency (SCons). This diagnostic metric compares the structural style of two outlines by quantifying the similarity of their depth and breadth. It is calculated as the geometric mean of two ratios: **Depth Consistency (DC)** and **Breadth Consistency (BC)**. Let the depths of the two outlines be d_1 and d_2, and their total node counts be n_1 and n_2. The consistency ratios are defined as the smaller value divided by the larger value for each dimension:

$$\text{DC} = \frac{\min(d_1, d_2)}{\max(d_1, d_2)} \quad \text{and} \quad \text{BC} = \frac{\min(n_1, n_2)}{\max(n_1, n_2)}. \tag{9}$$

The final Shape Consistency score is then computed as:

$$\text{SCons} = \sqrt{\text{DC} \times \text{BC}}. \tag{10}$$

Structural Clarity Score (SCS). This metric assesses an outline's logical coherence by quantifying improper content duplication. We use an LLM to identify semantically equivalent heading pairs that do not share an immediate parent, a rule designed to prevent misidentifying normal sibling sections as redundant. These pairs are counted as redundant $N_{\text{redundant}}$. The clarity score is then calculated as:

$$\text{SCS} = 1 - \frac{N_{\text{redundant}}}{N_{\text{total}}}, \tag{11}$$

where N_{total} is the total number of sections. A score near 1 suggests a clear structure with minimal

content overlap.

B.5 Level 5: Evaluation Metrics

Critical Analysis Alignment (CAA). This metric quantifies the extent to which the model's critical stance aligns with expert consensus. First, a Judge LLM is employed to extract explicit critical statements from both the generated and expert surveys. The prompt specifically instructs the model to identify sentences that describe limitations, methodological flaws, or comparative disadvantages of existing works. Second, to measure the overlap, a Judge LLM performs semantic matching between the two sets of extracted statements. For each critical statement in the generated set, the Judge determines if it is semantically equivalent to (or entailed by) any statement in the expert set.

Based on the number of LLM-verified matches, we report the F1-score to verify if the model's critical insights are both accurate (Precision) and exhaustive (Recall) relative to human experts.

B.6 Level 6: Creation Metrics

Framework Novelty (FNov). To evaluate the framework novelty of the survey, FNov adopts the GRADE approach. The Judge LLM is first instructed to synthesize bespoke, topic-specific scoring rubrics regarding framework innovation. Subsequently, utilizing these generated rubrics, the LLM evaluates both the ASG-generated survey and the human expert survey, ultimately deriving the relative GRADE score.

Research Outlook Quality (ROQ). Similarly, ROQ utilizes the GRADE approach to assess the quality of future research directions. The Judge LLM is prompted to synthesize bespoke, topic-specific scoring rubrics concerning the foresight and value of the proposed research agenda. Based on these dynamic criteria, the LLM scores both the generated and expert surveys to calculate the final relative GRADE score.

C Experimental Topic Details

To ground our evaluation in a rigorous and representative context, we selected 20 topics for our experiments. These topics correspond to the titles of the 20 most-cited survey papers published between 2023 and 2025 within our benchmark corpus, ensuring relevance to the current scientific frontier. The selection is evenly split, with 10 topics drawn from the Computer Science domain and 10 from a diverse range of General Science disciplines. This balanced approach allows for a robust assessment of the generalization capabilities of the evaluated ASG systems. The complete list of these 20 topics, along with their publication details and citation counts, is provided in Table 4.

C.1 Exact 10-Gram Overlap Analysis

To investigate whether the benchmark performance is attributed to memorization or direct plagiarism, we computed the exact 10-gram overlap between the system-generated and expert-authored surveys across all 20 experimental topics. As detailed in Table 5, the average overlap rates across all systems are exceptionally low. This minimal verbatim overlap indicates that the benchmark results are not simply artifacts of copying contiguous text spans from the reference data. Although exact n-gram matching cannot definitively preclude all forms of data contamination, these findings provide empirical evidence that the evaluated systems generate original content rather than merely reproducing the ground-truth texts.

D Detailed Experimental Results

This section presents the complete and disaggregated results of our evaluation. For each of the six cognitive tiers in the Bloom-Eval framework, we provide a detailed table showing the Precision (P), Recall (R), F1-Score (F1), and Distributional Similarity (DS) for every metric where applicable. Additionally, we report on several supplementary metrics, including generation speed and economic cost, to provide a more holistic view of system performance.

D.1 Core Cognitive Performance Metrics

For each of the evaluated cognitive tiers, Table 6, Table 7, and Table 8 show the detailed Precision (P), Recall (R), F1-Score (F1), and Distributional Similarity (DS) for every metric where applicable.

D.2 Supplementary Metrics: Efficiency and Readability

To provide a more holistic view of system performance, we report on several supplementary metrics, with detailed results presented in Table 9. These include the generation speed, total output token count, and a composite score for Readability.

The Readability (R) score is designed to measure stylistic similarity. It is not an absolute measure of simplicity but rather a score from 0 to 1 indicating

Table 4: The 20 experimental topics used for evaluation, selected based on the most-cited survey papers from our benchmark corpus (2023-2025).

Title	Venue	Year	Citations
Computer Science Domain			
A Survey on Vision Transformer	TPAMI	2023	3915
A Survey on In-context Learning	EMNLP	2023	2379
Diffusion Models in Vision: A Survey	TPAMI	2023	2037
Domain Generalization: A Survey	TPAMI	2023	1720
Transformers in Time Series: A Survey	IJCAI	2023	1458
Generalized Out-of-Distribution Detection: A Survey	IJCV	2024	1377
A Comprehensive Survey of Continual Learning: Theory, Method and Application	TPAMI	2024	1313
Class-Incremental Learning: Survey and Performance Evaluation on Image Classification	TPAMI	2023	1038
Transfer Learning in Deep Reinforcement Learning: A Survey	TPAMI	2023	1031
Multimodal Learning With Transformers: A Survey	TPAMI	2023	990
General Science Domain			
Structural Mechanisms of NLRP3 Inflammasome Assembly and Activation	Immunology	2023	781
The New Economics of Industrial Policy	Economics	2023	567
Random Quantum Circuits	Condensed Matter Physics	2023	494
Biological Impacts of Marine Heatwaves	Marine Science	2023	468
How to Run Surveys: A Guide to Creating Your Own, Identifying Variation and Revealing the Invisible	Economics	2023	443
A Review of Generalizability and Transportability	Statistics and Its Application	2023	387
Non-Hermitian Topological Phenomena: A Review	Condensed Matter Physics	2023	358
Hepcidin and Iron in Health and Disease	Medicine	2023	340
Generalized Symmetries in Condensed Matter	Condensed Matter Physics	2023	335
Everything You Wanted to Know about Deep Eutectic Solvents but Were Afraid to Be Told	Chemical and Biomolecular Engineering	2023	258

System	Average 10-Gram Overlap
AutoSurvey	0.02%
LLMxMapReduce-V2	0.01%
SurveyForge	0.04%
SurveyX	0.09%

Table 5: Average exact 10-gram overlap rates between system-generated and expert-authored surveys across the 20 experimental topics.

how closely an ASG system's writing style mimics that of the human-written gold standard. By definition, the human-authored text has a perfect score of 1. The final score is a composite, calculated as the arithmetic mean of similarity scores from three established readability formulas: the Flesch-Kincaid Grade Level (Kincaid et al., 1975), the Gunning Fog index (Gunning, 1952), and the Coleman-Liau Index (Coleman and Liau, 1975).

For each formula, a similarity ratio (sim) is first calculated by comparing the system's metric value ($v_{\text{generated}}$) to the human's (v_{human}):

$$\text{sim}(v_{\text{human}}, v_{\text{generated}}) = \frac{\min(v_{\text{human}}, v_{\text{generated}})}{\max(v_{\text{human}}, v_{\text{generated}})}.$$

This ensures that a larger divergence from the human baseline in either direction results in a lower score. The final Readability score (R) is the unweighted average of the three individual similarity scores ($\text{sim}_1, \text{sim}_2, \text{sim}_3$):

$$R = \frac{1}{3}\sum_{i=1}^{3} \text{sim}_i. \tag{12}$$

This approach leverages these well-established formulas to quantify the stylistic similarity in terms of overall text complexity. Each formula computes a score, typically representing a U.S. grade level, by combining sentence-level features (e.g., average sentence length) with word-level features (e.g., syllable or character counts). By averaging the similarity scores derived from these distinct formulas, we mitigate the biases inherent in any single metric and obtain a more robust and comprehensive assessment of how closely the machine-generated text's complexity profile matches the human standard.

The evaluation of speed distinguishes between automated systems and human experts. For ASG systems, speed is measured as the time from the initial submission of a research topic to the final gen-

Table 6: Detailed results for Level 1: Memory. EFid scores are in P/R/F1/DS format. HIRC and FCons are in P/R/F1 format.

Metric	Values	AutoSurvey	SurveyForge	LLMxMapReduce-V2	SurveyX
EFid	Precision	0.11	0.17	0.11	0.13
	Recall	0.10	0.11	0.17	0.12
	F1-Score	0.10	0.13	0.12	0.11
	DS-Score	0.81	0.73	0.69	0.78
HIRC	Precision	0.03	0.11	0.04	0.06
	Recall	0.05	0.06	0.01	0.04
	F1-Score	0.03	0.07	0.02	0.04
FCons	Precision	0.07	0.06	0.19	0.10
	Recall	0.10	0.09	0.07	0.06
	F1-Score	0.07	0.06	0.09	0.07

Table 7: Detailed results for Level 2: Comprehension. TFSim is in P/R/F1/DS format. OTC and CF are in P/R/F1 format.

Metric	Values	AutoSurvey	SurveyForge	LLMxMapReduce-V2	SurveyX
OTC	Precision	0.29	0.52	0.41	0.43
	Recall	0.76	0.78	0.82	0.81
	F1-Score	0.40	0.61	0.53	0.56
CF	Precision	0.03	0.27	0.39	0.44
	Recall	0.03	0.26	0.23	0.32
	F1-Score	0.03	0.27	0.29	0.37
TFSim	Precision	0.35	0.68	0.73	0.62
	Recall	0.76	0.56	0.32	0.55
	F1-Score	0.43	0.53	0.41	0.52
	DS-Score	0.73	0.78	0.67	0.81
TBal	1-Gini	0.72	0.70	0.68	0.68

Table 8: Detailed results for Level 5: Evaluation. Critical Analysis Alignment (CAA) scores are in P/R/F1 format.

Metric	Values	AutoSurvey	SurveyForge	LLMxMapReduce-V2	SurveyX
CAA	Precision	0.09	0.04	0.07	0.08
	Recall	0.35	0.38	0.45	0.43
	F1-Score	0.13	0.08	0.11	0.13

eration of the complete survey. For human experts, we follow the established estimation methodology (Wang et al., 2024a).

E Comparative Performance Across Scientific Domains

While many automatic survey generation (ASG) systems are capable of retrieving and processing information across multiple scientific disciplines, some systems, such as AutoSurvey (Wang et al., 2024a) and SurveyForge (Yan et al., 2025), are architecturally designed or configured to operate primarily within the Computer Science (CS) domain. This presents a critical question regarding their cross-domain generalization capabilities. To investigate this limitation, this section provides a detailed performance analysis of these two systems, contrasting their efficacy on their native CS topics with a broader set of General Science topics. The specific topics for the CS domain correspond to

Table 9: Supplementary metrics assessing efficiency and cost. Readability (R) is on a 0-1 scale. Speed (S) is in minutes. Tokens (T) is the total output count.

Metric	AutoSurvey	SurveyForge	LLMxMapReduce-V2	SurveyX	Human
Readability (R)	0.73	0.71	0.75	0.68	1.00
Speed (S)	12	11	60	126	434
Output Tokens (T)	65,907	25,911	72,217	23,630	29,274

the first ten entries in our experimental corpus (see Table 4), while the General Science topics are represented by the subsequent ten entries.

The following radar charts visualize the performance drop-off across key cognitive tiers of the Bloom-Eval framework. Figures 5 to 8 illustrate the domain-specific performance for AutoSurvey, while Figures 9 to 12 show the results for SurveyForge. This analysis quantifies the performance penalty of domain-specific architectures, highlighting the critical value of cross-disciplinary benchmarks for testing true generalization.

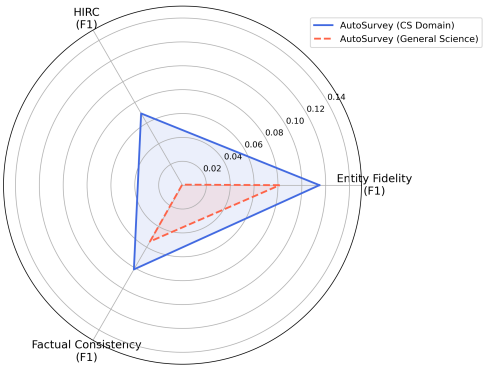

Figure 5: The performance of AutoSurvey on Level 1 (Memory) metrics, comparing Computer Science and General Science domains.

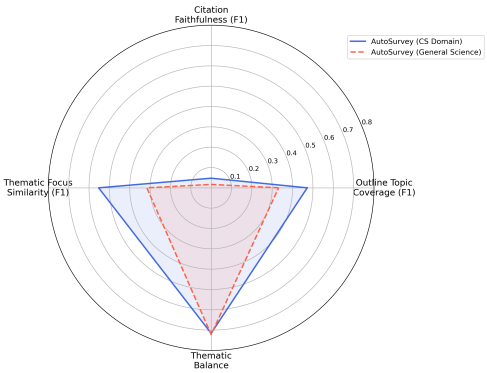

Figure 6: The performance of AutoSurvey on Level 2 (Comprehension) metrics, comparing CS and General Science domains.

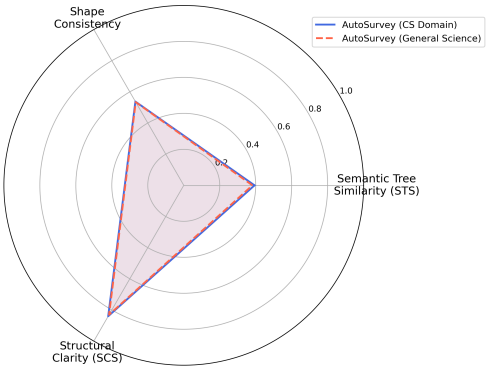

Figure 7: The performance of AutoSurvey on Level 4 (Analysis) metrics, comparing CS and General Science domains.

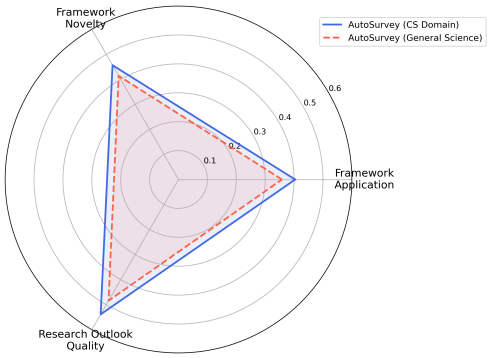

Figure 8: The performance of AutoSurvey on the GRADE-based metrics, comparing CS and General Science domains.

The performance of AutoSurvey demonstrates a strong dependency on its native Computer Science domain, with a notable degradation observed when applied to General Science topics. This performance gap is most pronounced in knowledge-intensive, memory-based tasks (Level 1). Specifically, the system's ability to identify high-impact references (HIRC) is almost entirely nullified in the General Science domain, accompanied by significant reductions in entity fidelity and factual consistency. Similarly, comprehension-level metrics such as Citation Faithfulness (Level 2) also exhibit a marked decline.

In stark contrast, the system's analytical capabil-

ities for structuring content (Level 4) appear to be largely domain-agnostic. Metrics such as Semantic Tree Similarity, Shape Consistency, and Structural Clarity Score remain remarkably stable across both domains. For higher-order GRADE-based tasks, while AutoSurvey's ability to generate novel frameworks and high-quality research outlooks is strong within CS, this capacity is diminished in the broader scientific context. This suggests that the primary limitation stems not from the model's inherent reasoning or structural generation abilities, but rather from its retrieval mechanism, which is heavily restricted to the Computer Science literature.

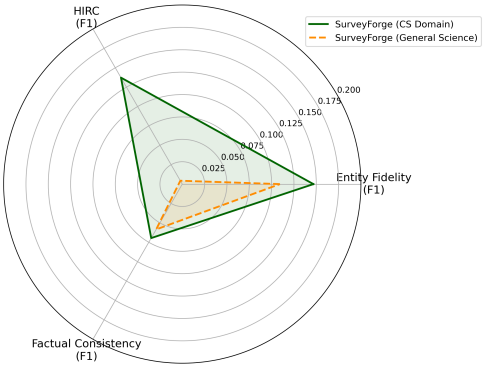

Figure 9: The performance of SurveyForge on Level 1 (Memory) metrics, comparing Computer Science and General Science domains.

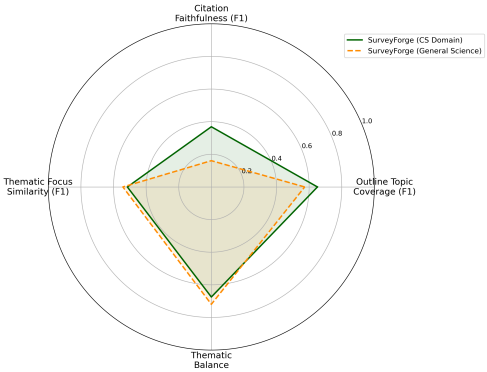

Figure 10: The performance of SurveyForge on Level 2 (Comprehension) metrics, comparing CS and General Science domains.

SurveyForge also exhibits a significant performance dependency on the Computer Science (CS) domain. The system's most critical failure outside of CS lies in its knowledge grounding and citation abilities. In the General Science domain, the capacity to identify high-impact references is severely diminished, and Citation Faithfulness drops by more than half, indicating a fundamental breakdown in connecting claims to correct, relevant sources.

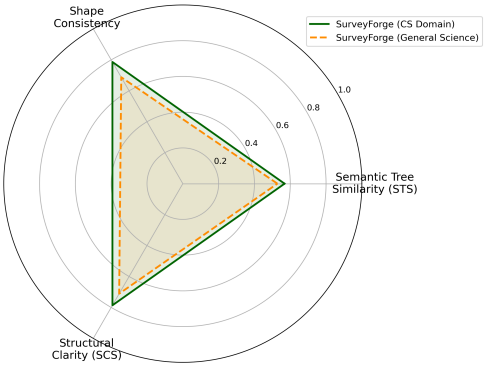

Figure 11: The performance of SurveyForge on Level 4 (Analysis) metrics, comparing CS and General Science domains.

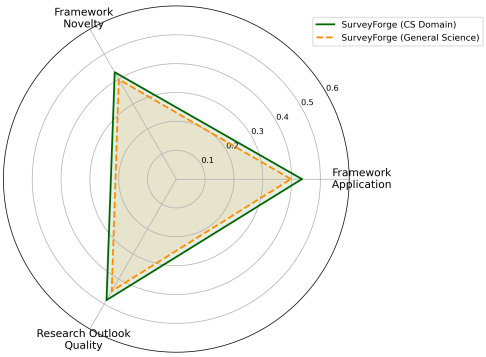

Figure 12: The performance of SurveyForge on the GRADE-based metrics, comparing CS and General Science domains.

F Diagnostic Analysis of Performance Bottlenecks

To better understand the capability gap identified in our experiments, specifically why ASG systems fail at higher-order cognitive tiers, we analyze the systemic bottlenecks of current architectures.

- **Retrieval Misalignment.** Existing systems universally failed the High-Impact Reference Coverage (HIRC) metric, even in their native CS domain. Our analysis reveals that current retrieval algorithms (e.g., in AutoSurvey (Wang et al., 2024a)) rely heavily on vector embedding similarity while ignoring bibliometric impact. This leads to the retrieval of "semantically overlapping but mediocre" papers, missing foundational works with high citation counts. Furthermore, the reliance on static, closed-source CS databases leads to a severe degradation in cross-disciplinary tasks. Future work must integrate bibliometric in-

dicators (e.g., citation velocity) into the retrieval ranker. Our contribution of a cross-disciplinary benchmark is specifically designed to guide the field away from closed CS datasets toward robust, open-domain mechanisms.

- **Information Processing via "Lossy Compression".** Systems scored poorly in Creation and Evaluation tiers because they lack deep insight. To fit limited context windows, baselines employ aggressive compression strategies. For instance, SurveyX (Liang et al., 2025) compresses papers into "AttributeTrees", and LLMxMapReduce-V2 (Wang et al., 2025) relies on "Digests". This processing strips away subtle derivations, negative results, and complex arguments, effectively depriving the LLM of the raw materials needed for deep synthesis and innovation. The field must move from lossy summarization to long-context architectures that can process high-fidelity full text, preserving the nuance required for higher-order cognition.

- **Underutilization of Meta-Knowledge.** Existing methods like SurveyForge (Yan et al., 2025) utilize related survey papers primarily to mimic their "outline structure", rather than understanding the intellectual lineage or core debates within the field. This strategy results in models that produce a "flat" listing of literature, failing to synthesize knowledge depth.

G Reliability and Validation of Evaluation Metrics

In this section, we provide a comprehensive validation of the Bloom-Eval framework through three complementary analyses: a verification of the intermediate LLM modules (Section G.1), a global correlation study covering all cognitive tiers (Section G.2), and a specific sensitivity analysis for the abstract GRADE metrics (Section G.3).

G.1 Verification of LLM-based Extraction and Matching

Since several Bloom-Eval metrics employ LLM-based extraction and matching as intermediate steps, we additionally validated these modules to assess whether errors in these stages propagate and contaminate the final scores. Importantly, these modules do not directly generate final metric scores. Instead, they produce structured intermediate artifacts, such as extracted entity lists, factual statements, topic headings, matched pairs, and entailment decisions. Final scores are subsequently computed by deterministic algorithms operating on these artifacts. Consequently, each LLM decision can be inspected and audited at the item level.

We randomly sampled 100 extraction and matching instances from our evaluation pipeline and tasked two human experts with independently validating the LLM outputs. Disagreements were resolved through discussion. Table 10 summarizes the validation results. The extraction modules achieved 83% precision and 85% recall compared to human annotations, while the matching modules reached a 78% agreement rate with human judgments on semantic equivalence. These results indicate that although the LLM-based middleware is not flawless, its errors are observable through explicit logs rather than obscured within an end-to-end black-box judgment.

Table 10: Validation results of intermediate LLM-based extraction and matching modules. Scores are calculated based on 100 randomly sampled instances verified by two human experts.

Task Type	Metric	Score
Extraction	Precision	83%
	Recall	85%
Matching	Agreement Rate	78%

G.2 Alignment with Human Judgments

To validate the overall effectiveness of Bloom-Eval, we assessed whether our automated metrics, ranging from deterministic algorithms to LLM-based evaluations, align with human consensus across the six cognitive tiers.

We recruited five independent evaluators to provide human judgments, including three Ph.D. candidates and two Master's students. All evaluators had formal training in computer science, and several had research experience in interdisciplinary areas connecting computer science with life sciences and quantum computing. Collectively, their expertise spans the Computer Science and General Science topics covered in this validation study. All evaluators had prior experience reading and assessing academic literature, and none was involved in the development of the evaluated ASG systems. To ensure domain diversity, we selected the three most-cited topics from both the Computer Science and General Science domains (six topics in total). For each topic, we collected survey papers generated by

four different ASG systems plus the expert-written references, resulting in a total of 30 papers. To ensure the human evaluation was both rigorous and tractable, we selected one representative metric for each cognitive tier that captures the core definition of that level for human assessment.

We calculated the Pearson and Spearman correlations between the scores of these selected representative metrics (computed using GPT-5-mini for GRADE) and the human ratings. As shown in Table 11, Bloom-Eval demonstrates strong alignment (≥ 0.7) with human experts across all dimensions. This result confirms two key findings:

- **Validation of Engineered Metrics:** Our metrics (e.g., Factual Consistency, Citation Faithfulness) successfully capture the nuances of human assessment.

- **Evaluator Reliability:** The choice of GPT-5-mini as the underlying model for GRADE provides a rigorous standard consistent with expert reviewers, justifying its use as the primary judge in our main experiments.

G.3 Sensitivity Analysis of the GRADE Approach

Notwithstanding the general consistency demonstrated above, the inherent subjectivity of abstract reasoning metrics (specifically in Levels 3 and 6) warrants deeper scrutiny regarding their sensitivity to the choice of the Judge LLM.

To investigate this, we conducted a sensitivity analysis on the Framework Application (FAP) metric, a key indicator of a system's ability to apply knowledge.

In this experiment, we selected three distinct large language models, GPT-4.1-mini, GPT-5-mini, and GPT-4o-mini, to serve as independent judges. Each model evaluated the same 20 generated surveys against their human-expert counterparts and produced a GRADE score for the FAP metric. The results reveal a crucial finding: the evaluation outcome is sensitive to the choice of the Judge LLM, exposing significant "Rater Bias" inherent in the "LLM-as-a-Judge" paradigm. This finding does not weaken our framework. On the contrary, it highlights the critical necessity of our transparent, rubric-based design.

G.3.1 Divergence in Scoring Bias and Qualitative Conclusions

The most immediate finding is the profound difference in the scoring tendencies of the three models, as illustrated in Figure 13.

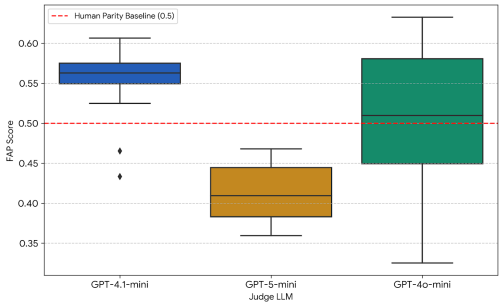

Figure 13: Distribution of Framework Application (FAP) scores across 20 tasks, as judged by three different LLMs. The red dashed line indicates the human parity baseline. The plot reveals starkly different scoring biases: optimistic (GPT-4.1-mini), pessimistic (GPT-5-mini), and unstable (GPT-4o-mini).

The analysis reveals three distinct rater profiles:

- **The Optimistic Rater (GPT-4.1-mini):** This model exhibits a strong positive bias, with a mean score of 0.554. In 18 out of 20 cases (90%), it judged the generated survey's framework application as superior to the human expert's. Its scoring is also relatively consistent, with a small standard deviation (0.042).

- **The Pessimistic Rater (GPT-5-mini):** In stark contrast, this model shows a consistent negative bias. Its mean score is only 0.407, and it judged the system's performance to be inferior to the human baseline in all 20 cases (100%). It is the most consistent rater, with the smallest standard deviation (0.035).

- **The Unstable Rater (GPT-4o-mini):** This model lacks a clear, consistent standard. Its mean score of 0.506 hovers around the human parity line. Critically, its standard deviation (0.089) is more than double that of the other two models, indicating highly variable and unpredictable judgments.

This divergence is alarming: a researcher's conclusion about a system's capability on this metric is predetermined by their choice of Judge LLM.

Directional Agreement. We calculated the rate at which any two models agreed on whether a score was ≥ 0.5 or < 0.5. The results in Table 12 are presented below.

Table 11: Correlations between Bloom-Eval metrics and human expert ratings on a subset of 30 papers.

Cognitive Tier	Automated Metric	Pearson	Spearman
Level 1: Memory	Factual Consistency	0.76	0.70
Level 2: Comprehension	Citation Faithfulness	0.92	0.92
Level 3: Application	Framework Application (GRADE)	0.83	0.84
Level 4: Analysis	Structural Clarity Score	0.75	0.76
Level 5: Evaluation	Critical Analysis Alignment	0.79	0.84
Level 6: Creation	Research Outlook Quality (GRADE)	0.70	0.75

Table 12: Directional Agreement Rate between Judge LLMs.

Model Pair	**Agreement Rate**
GPT-4.1-mini vs. GPT-5-mini	10%
GPT-4.1-mini vs. GPT-4o-mini	60%
GPT-5-mini vs. GPT-4o-mini	50%

The minimal 10% agreement rate between the optimistic and pessimistic raters signifies a fundamental breakdown in consensus, highlighting a systematic opposition in their evaluative judgments. Furthermore, the moderate agreement levels observed in the other pairings remain below the threshold required for reliable inter-rater consistency.

Analysis of Correlational Agreement. To assess a deeper level of inter-rater reliability, we performed a correlational analysis. We investigated two forms of correlation: linear association using Pearson's correlation and ordinal consistency using Spearman's rank-order correlation.

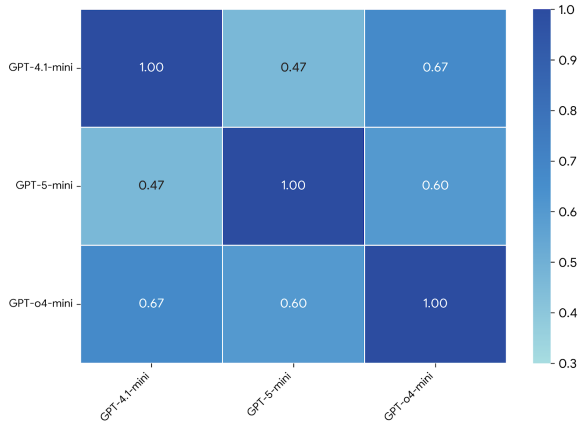

Figure 14: Pearson correlation heatmap of FAP scores.

First, we examined the linear relationship between the models' scores. As shown in Figure 14, the scores exhibit a weak to moderate positive correlation, with coefficients ranging from 0.47 to 0.67. While this suggests a general tendency for scores to move in the same direction, it does not sufficiently establish a robust, shared evaluative standard.

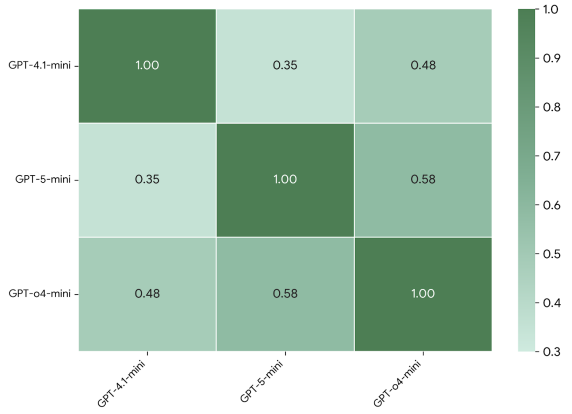

Figure 15: Spearman's rank-order correlation heatmap.

We also assessed the consistency of ordinal judgments using Spearman's rank correlation. The results, visualized in Figure 15, reveal a significant lack of consensus on relative quality. The correlation coefficients are uniformly low, with the value between the optimistic rater (GPT-4.1-mini) and the pessimistic rater (GPT-5-mini) being a mere 0.35. This weak rank-order correlation is a key finding: it indicates that the models do not share a stable, underlying logic for what constitutes a "better" or "worse" application of a framework. Their disagreement is therefore not just a matter of scale, but a fundamental divergence in evaluative criteria.

The Value of Transparency in a Biased World. The severe sensitivity to rater selection presents a major challenge to the "LLM-as-a-Judge" paradigm. However, rather than invalidating our framework, this finding powerfully underscores the value of the GRADE approach's core design principle: transparency.

1. **Rubrics as an Auditable Artifact:** Since any Judge LLM is inherently biased, we cannot trust a single, opaque score. The first step of GRADE forces the LLM to generate an

explicit, weighted rubric. This rubric becomes a public, auditable artifact. It shifts the debate from an inscrutable final score to a transparent set of criteria that can be debated, refined, and analyzed, thus mitigating the black-box problem.

2. **A Methodological Guideline:** Our results serve as a critical methodological warning for the field. Any study employing an LLM-as-a-Judge must report the specific model version used and acknowledge that results may not be directly comparable to studies using different judges. Moreover, they should report quantitative evidence of the judge's alignment with human experts (e.g., correlation coefficients).

3. **Ensuring Internal Consistency:** Given the strong consistency between GPT-5-mini and human expert judgments, we ultimately selected GPT-5-mini as the judge.

This sensitivity analysis confirms that no LLM is an objective, "god-like" evaluator. Recognizing this reality, the GRADE approach provides the necessary methodological scaffolding through the transparency of explicit rubrics, making LLM-based evaluation more rigorous, interpretable, and scientifically sound.

G.3.2 Case Study: Deconstructing Rater Bias Through Transparent Rubrics

To understand the origin of rater bias, we conducted a case study on a task where the Judge LLMs profoundly disagreed. Our analysis of their explicit, model-generated rubrics reveals that the scoring divergence stems from three distinct rater archetypes with different "academic values": GPT-4o-mini, the "Pragmatic Balancer", seeks a compromise between structure and content (Table 13a); GPT-5-mini, the "Rigorous Methodologist", demands scientific rigor like a peer reviewer (Table 13b); and GPT-4.1-mini, the "Structuralist", prioritizes formal organization like an editor (Table 13c). This vividly demonstrates that rater bias is not random error but a systematic manifestation of different latent definitions of "quality". Crucially, this analysis is only possible due to our transparent, rubric-based methodology, which transforms the LLM's "black box" reasoning into a "glass box" and makes the evaluation process itself an object of scientific inquiry.

G.3.3 Case Study: Auditability Comparison

To demonstrate how the GRADE approach transforms a "black-box" evaluation into an auditable process, we provide a visual contrast between a standard baseline judge and our multi-step framework in Table 14. As illustrated in Table 14, while a baseline judge only provides an unexplainable score, the GRADE approach forces the model to externalize its criteria and justify its judgment with specific evidence from the text. This multi-step trace allows researchers to diagnose exactly why a model received a particular score, meeting the requirements for academic auditability.

H Mapping Metrics to Bloom's Taxonomy

Table 15 provides a systematic mapping of Bloom-Eval's 16 metrics to the six tiers of Bloom's Taxonomy. It details the theoretical alignment between standard cognitive definitions and our specific diagnostic purposes within the context of ASG.

I Prompt Templates

This section provides the detailed prompt templates used for the various LLM-based tasks within the Bloom-Eval framework (see Figures 16 through 24).

Table 13: Evaluation rubrics generated by the three different Judge LLMs.

(a) Evaluation rubric generated by GPT-4o-mini.

Criterion	Weight
Relevance and Appropriateness of Framework	30%
Clarity and Definition of Framework Components	20%
Consistent Application of Framework	20%
Analytical Depth and Insight	20%
Domain-Specific Adaptability	10%

(b) Evaluation rubric generated by GPT-5-mini.

Criterion	Weight
Clarity and Relevance of the Chosen Framework	22%
Taxonomy Coverage and Granularity	18%
Consistent Application of the Framework	17%
Analytical Depth and Trade-off Analysis Enabled by the Framework	16%
Mapping to Tasks, Datasets, and Evaluation Protocols	10%
Evidence-based Classification and Traceability	9%
Actionability: Gaps, Best Practices, and Research Agenda	8%

(c) Evaluation rubric generated by GPT-4.1-mini.

Criterion	Weight
Clarity and Relevance of Framework	40%
Consistent Application of Framework	35%
Analytical Depth Enabled by Framework	25%

Table 14: Comparison of evaluation transparency between a Baseline Judge and our GRADE Approach.

Evaluation Stage	Baseline Judge ((Wang et al., 2024a))	Ours: GRADE Approach (Auditable Trace)
1. Rubric	Uses pre-defined generic 1–5 definitions (e.g., "The survey has good logical consistency") without domain context.	**[Self-Generated Rubric]** Automatically creates 7 domain-specific criteria such as: *Coverage of Key ViT Axes*, *Taxonomy Granularity*, etc.
2. Rubric Explanation	**[Missing]** Provides no explanation for what it specifically measures.	**[Detailed Criterion Definition]** Explicitly defines each item, e.g., for ViT Axes: "Measures whether the framework covers architectural variants, attention modifications, and training regimes."
3. Rubric Weighting	**[Missing]**	**[Explicit Weights]** Assigns precise importance (e.g., 0.20 for Clarity, 0.14 for Granularity).
4. Comparative Analysis	**[Missing]** Output is often just a scalar score (e.g., *Score: 3*) with no comparative grounding between articles.	**[Evidence-based Analysis]** Provides cross-article comparisons, e.g., "Article 2 explicitly states its framework early... making it easier to apply than Article 1's implicit organization."
Conclusion	**Black Box**: Human auditors cannot verify if the final score is grounded in the source text or the rubric.	**Auditable**: Human experts can inspect the rubric definitions, weights, and evidence-based analysis to validate the results.

Table 15: Systematic mapping of Bloom-Eval's 16 metrics to Bloom's Taxonomy. The table illustrates how each metric aligns with the theoretical definition of a cognitive tier to diagnose specific ASG capabilities.

Cognitive Level	**Bloom's Definition**	**Bloom-Eval Metrics**	**Theoretical Alignment**
L1: Memory	The ability to recall information, encompassing facts, concepts, definitions, and terminology.	• Entity Fidelity (EFid) • High-Impact Reference Coverage (HIRC) • Factual Consistency (FCons)	**Foundational Reproduction:** Diagnoses the system's proficiency in accurately retrieving and reproducing the field's foundational elements, specifically domain entities (EFid), pivotal literature (HIRC), and factual descriptions (FCons).
L2: Comprehension	The ability to explain and summarize information.	• Outline Topic Coverage (OTC) • Citation Faithfulness (CF) • Thematic Focus Similarity (TFSim) • Thematic Balance (TBal)	**Interpretative Synthesis:** Diagnoses the ability to summarize content into necessary sub-headings (OTC), accurately explain cited references without hallucination (CF), and synthesize the research landscape's distributional hotspots (TFSim, TBal).
L3: Application	The ability to apply learned knowledge to practical problems.	• Formatting Integrity (FMI) • Document Structure Integrity (DSI) • Framework Application (FAP)	**Standard Execution:** Diagnoses the application of standard academic survey frameworks, necessitating strict adherence to citation conventions (FMI), structural skeleton completeness (DSI), and the consistent implementation of organizing paradigms (FAP).
L4: Analysis	The ability to break down information and identify structures or relationships.	• Semantic Tree Similarity (STS) • Shape Consistency (SCons) • Structural Clarity Score (SCS)	**Structural Deconstruction:** Diagnoses the accurate construction of writing structure, characterized by sound hierarchical logic (STS), appropriate analytical granularity (SCons), and clear separation of distinct concepts without redundancy (SCS).
L5: Evaluation	The ability to critically judge the value of ideas.	• Critical Analysis Alignment (CAA)	**Critical Reasoning:** Diagnoses the quality of evaluative reasoning by measuring the alignment of explicit value judgments (e.g., identifying method flaws) between the generated survey and expert consensus.
L6: Creation	The capacity to propose new viewpoints, models, or works, representing the highest order of cognition.	• Framework Novelty (FNov) • Research Outlook Quality (ROQ)	**Innovative Synthesis:** Diagnoses the capacity for innovation, specifically mandating the reorganization of the field with fresh conceptual models (FNov) and the foresight to identify potential future research directions (ROQ).

Prompt for Entity Extraction

```
You are a meticulous text scanner. Your task is to extract EVERY occurrence of technical entities from the academic survey text below. You will act like a machine, scanning the text from beginning to end and listing entities as you find them.

    **CRITICAL INSTRUCTIONS:**
    1.  **DO NOT DEDUPLICATE**: If an entity like "3DGS" appears 10 times, you MUST list it 10 times. If "3D Gaussian Splatting" appears 5 times, you MUST list it 5 times. List every single instance you find.
    2.  **EXTRACT EXACT PHRASES**: Extract the exact wording as it appears in the text. Do not normalize or change the entities.
    3.  **CATEGORIZE EACH INSTANCE**: For each entity instance you find, classify it into one of the three categories below.

    **Categories to extract:**
    1.  `methods_models`: Any named technique, algorithm, framework, architecture, or specific system (e.g., "3D Gaussian Splatting", "NeRF", "SfM").
    2.  `datasets`: Standardized data collections or benchmarks (e.g., "Neuman", "Stereo Blur").
    3.  `evaluation_metrics`: Quantitative metrics used to measure performance (e.g., "PSNR", "Chamfer Distance").

    **OUTPUT FORMAT:**
    You MUST return the output as a single, valid JSON object. Do not add any explanatory text before or after the JSON. The lists in the JSON should contain every single occurrence of the entities.

    **JSON format example:**
    If the text says "We use PSNR. Our method improves PSNR.", the output should be:
    {{
      "methods_models": [],
      "datasets": [],
      "evaluation_metrics": ["PSNR", "PSNR"]
    }}

    Now, analyze the text below and extract all entity occurrences:
    --- TEXT START ---
    {text}
    --- TEXT END ---
```

Figure 16: Prompt template for entity extraction.

Prompt for Domain-Adaptive Entity Matching

```
# Role
You are an expert in Entity Resolution for scientific literature. Your first step is to analyze the content of the `expert_data` and `llm_data` JSON objects to identify the specific academic domain they belong to. Then, you will act as a specialist in THAT domain to perform the matching task. You are highly familiar with the names, abbreviations, and common variants of models, datasets, and evaluation metrics found in your specialized domain. Your task is to accurately identify the common entities between the two provided JSON objects (`expert_data` and `llm_data`). For each common entity you identify, you must provide its corresponding main name from both of the original JSON files.
# Core Requirements
1. **Definition**: A "common entity" refers to the same underlying concept existing in both JSON files, **even if their main names or aliases have different spellings or variations**.
2. **Method**: You must use the entity's main name, its list of aliases, and your specialized domain knowledge to determine if they refer to the same concept.
3. **Confidence**: Match entities only when you are highly confident they are the same.
4. **Categorization**: Matching must occur *within* the same top-level category (e.g., a `methods_models` from expert_data can only match a `methods_models` from llm_data).
# Examples of Correct Matching
- An entity named "Convolutional Neural Network" (with an alias "CNN") in `expert_data` should be matched with an entity named "CNN" (with an alias "convolutional neural networks") in `llm_data`.
- An entity named "IPT" (with an alias "Image Processing Transformer (IPT)") in `expert_data` should be matched with "Image Processing Transformer" (with an alias "IPT") in `llm_data`.
# Input Data Format
I am providing two JSON objects: `expert_data` and `llm_data`. Each contains three categories: `methods_models`, `datasets`, and `evaluation_metrics`.
# Output Requirements
1. **Format**: The output must be a **single, valid JSON object**.
2. **No Extra Text**: **Do not** add any explanations, comments, or additional text outside of the final JSON code block.
3. **Structure**: The output JSON must contain the same three top-level keys: `methods_models`, `datasets`, and `evaluation_metrics`.
4. **Content**: Under each key, provide a list of matched pairs. Each element in the list must be an object with exactly two keys:
    * `expert_main_name`: The main name of the entity from `expert_data`.
    * `llm_main_name`: The main name of the entity from `llm_data`.
5. **Accuracy**: If an entity exists in only one file, it must not be included. If a category has no matches, the list for that category should be empty (`[]`).
---
### INPUT DATA
#### expert_data
{expert_data_json}
#### llm_data
{llm_data_json}
---
Now, perform the entity resolution and provide the final JSON output.
```

Figure 17: Prompt template for domain-adaptive entity matching.

Prompt for Extracting Factual Statements

```
# ROLE
You are a meticulous, detail-oriented research assistant tasked with extracting all **Factual Statements** from a given academic article.
# TASK DEFINITION
A "Factual Statement" is a complete sentence that is **verifiable, objective, contains no subjective evaluation, and its factual content can be directly understood from the current text without consulting external citations.**
It typically describes:
- **Definitions**: e.g., "A radiance field is a function..."
- **Methods**: e.g., "The model is trained using the Adam optimizer..."
- **Specific Data or History**: e.g., "The Transformer architecture was proposed in 2017."

Please DO NOT extract the following:
- **Statements reliant on external citations**: Claims whose validity can only be verified by reading a cited reference. This is the most important rule.
- **Subjective evaluations or opinions**: e.g., "A key limitation of this approach is...", "This is a promising direction..."
- **Future outlooks or suggestions**: e.g., "Future work should focus on..."
- **Structural descriptions of the article**: e.g., "Section 3 describes our architecture..."
---
### POSITIVE EXAMPLES (These are the types of sentences you SHOULD extract)
1. "A radiance field is a function that maps a 5D coordinate (spatial location and viewing direction) to a color and density."
2. "The model is trained using the Adam optimizer with a learning rate of 1e-4."
3. "The Transformer architecture was first proposed in the 2017 paper 'Attention Is All You Need'."
---
### NEGATIVE EXAMPLES (Do NOT extract these sentences)
1. "However, a key limitation of this approach is its significant memory footprint, making it impractical for consumer-grade hardware." (This is an evaluation/limitation)
2. "Handling dynamic and deforming objects remains a largely unexplored challenge." (This identifies a research gap/analysis)
3. "Future work should focus on integrating a simultaneous localization and mapping (SLAM) component." (This is a future direction/suggestion)
4. "As demonstrated by Smith et al. [25], this phenomenon is caused by quantum entanglement." (This claim's validity depends on external reference [25] and should not be extracted.)
---
# TASK
Follow these steps carefully:
1.  First, read through the entire article below and identify a preliminary list of all potential "Factual Statements".
2.  Next, perform a **mandatory final review** of your preliminary list. For each sentence in your list, you must ask yourself: **"Does this sentence contain a citation marker (e.g., [9], [60]) or mention a specific author's name (e.g., 'Smith et al.')?"**
3.  If the answer is YES, you **MUST** delete that sentence from your list. This final filtering step is the most critical part of your task.
4.  Finally, present the fully cleaned and filtered list according to the output rules.
# OUTPUT RULES
- Your output **MUST** be a single, valid JSON object.
- The JSON object must contain only one key: `"factual_statements"`.
- The value for this key must be a **list of strings**, where each string is a complete, standalone factual statement you extracted after the final review.
- If no factual statements remain after filtering, return an empty list: `[]`.
- **DO NOT** add any explanations or extra text outside the JSON code block.
---
### ARTICLE TO ANALYZE
{text}
---
```

Figure 18: Prompt template for factual statement extraction.

Prompt for Matching Factual Statements

```
# ROLE
You are an expert in Natural Language Understanding and semantic analysis. Your task is to act as a meticulous fact-checker.
# TASK
Your goal is to identify all pairs of semantically equivalent factual statements from two provided lists: `expert_factual_claims` and `llm_factual_claims`.
# DEFINITION OF "SEMANTICALLY EQUIVALENT"
Two statements are semantically equivalent if and only if they assert the exact same core fact.
- **Good Match (Equivalent):** "ViT was developed by Google researchers in 2020." vs. "In 2020, researchers at Google created the Vision Transformer (ViT)."
- **Bad Match (Not Equivalent):** "BERT uses an encoder." vs. "GPT uses a decoder." (Different facts).
# RULES
1. **Strictness is Key:** Only match statements if you are highly confident they convey the identical meaning. If there is any factual discrepancy (e.g., different numbers, dates, or subtle meanings), do not match them.
2. **One-to-One Matching:** Each statement from one list should be matched to at most one statement from the other list. Find the best possible pairings.
3. **Ignore Wording:** Do not be distracted by different phrasing or word order if the core fact is the same.
---
### INPUT LISTS
#### expert_factual_claims
{expert_statements_json}
#### llm_factual_claims
{llm_statements_json}
---
# OUTPUT RULES
- Your output **MUST** be a single, valid JSON object and nothing else.
- The JSON object must contain a single key: `"matched_pairs"`.
- The value of `"matched_pairs"` must be a **LIST** of JSON objects.
- Each object in the list must have exactly two keys: `"expert_factual_claims"` and `"llm_factual_claims"`.
- The values for these keys must be the full, original strings of the statements that you have matched.
- If no matches are found, return an empty list: `{{"matched_pairs": []}}`.
- **DO NOT** add any explanations, comments, or text outside the final JSON code block.
```

Figure 19: Prompt template for factual statement matching.

Prompt for Semantic Alignment of Survey Paper Outlines

ROLE
You are an expert academic researcher specializing in scientific literature analysis. Your task is to meticulously compare two outlines for a survey paper and identify all pairs of section headings that refer to the same core topic.
TASK DEFINITION
Analyze the two lists of section headings provided below: `EXPERT_HEADINGS` and `LLM_HEADINGS`. Identify all pairs of headings that are semantically equivalent. A match occurs if a heading from the LLM list is a direct synonym, a clear paraphrase, or covers the same conceptual ground as a heading from the Expert list.
EXAMPLES
- If `EXPERT_HEADINGS` has "Historical Development" and `LLM_HEADINGS` has "The Rise of Transformers", they are a match.
- If `EXPERT_HEADINGS` has "Conclusion" and `LLM_HEADINGS` has "Summary and Future Work", they are a match.
- If `EXPERT_HEADINGS` has "Core Mechanisms" and `LLM_HEADINGS` has "Applications", they are NOT a match as they cover different concepts.
INPUT DATA
EXPERT_HEADINGS
{expert_topics_str}
LLM_HEADINGS
{llm_topics_str}
OUTPUT RULES
Your response MUST be a single, valid JSON object.
This object must contain only one key: `"matched_pairs"`.
The value for this key must be a list of objects. Each object in the list represents one matched pair and must have exactly two keys:
1. `"expert_heading"`: The heading from the `EXPERT_HEADINGS` list.
2. `"llm_heading"`: The corresponding heading from the `LLM_HEADINGS` list.
If no matches are found, return an empty list: `[]`.
DO NOT include any explanations or extra text outside the JSON code block.

Figure 20: Prompt template for semantic alignment of outline topics.

Prompt for Matching Critical Statements

```
# ROLE
You are an expert in Natural Language Understanding and semantic analysis, specializing in academic and research-oriented texts.
# TASK
Your goal is to meticulously identify all pairs of **semantically equivalent "critical statements"** from two provided lists: `expert_critical_statements` and `llm_critical_statements`.
# DEFINITION OF "SEMANTICALLY EQUIVALENT CRITICAL STATEMENT"
Two critical statements are semantically equivalent if they express the **same core evaluation, analysis, limitation, research gap, or future direction**, even if the wording is different.
- **Good Match (Equivalent):**
  - Statement 1: "A key drawback of this model is its significant computational overhead."
  - Statement 2: "The model is computationally expensive, which limits its practical use."
- **Bad Match (Not Equivalent):**
  - Statement 1: "The model struggles with small objects."
  - Statement 2: "The model requires extensive training data."
# RULES
1. **Strictness is Paramount:** Only match statements if you are highly confident they convey the identical critique or suggestion.
2. **One-to-One Matching:** Each statement from one list can be matched to at most one statement from the other list.
3. **Focus on Meaning, Not Phrasing:** Ignore differences in wording if the core critical point is the same.
---
### INPUT LISTS
#### expert_critical_statements
{expert_statements_str}
#### llm_critical_statements
{llm_statements_str}
---
# OUTPUT RULES
- Your output **MUST** be a single, valid JSON object.
- The JSON object must contain a single key: `"matched_critical_pairs"`.
- The value must be a **LIST** of objects, each with two keys: `"expert_critical_statement"` and `"llm_critical_statement"`.
- The values must be the full, original strings of the statements you have matched.
- If no matches are found, return an empty list: `{{"matched_critical_pairs": []}}`.
```

Figure 21: Prompt template for matching critical statements.

Prompt Templates for Evaluating "Research Outlook Quality" - Dynamic Criteria Generation

<system_role>
You are an experienced research mentor and academic editor, specializing in identifying high-impact future research directions. You excel at deconstructing the abstract concept of "good research questions" into concrete, weighted criteria tailored to a specific research field.
</system_role>

<user_prompt>
Background: We are evaluating the quality of the "future research directions" section of a survey paper. The survey was written to address the following research task:

<task>
{task_prompt}
</task>

Core Evaluation Dimension: Research Outlook Quality
This metric evaluates the quality of the proposed directions for future research. The assessment focuses on the novelty, insightfulness, and feasibility of the new questions or hypotheses, reflecting the ability to identify promising avenues for future inquiry.

<Instruction>
Your Goal: For the **Research Outlook Quality** dimension, develop a detailed, specific, and logically sound set of evaluation criteria tailored to the research `<task>`. You must:

1. **Analyze the Task & Dimension**: Based on the `<task>`, identify what constitutes insightful and high-impact future research questions for this specific topic.
2. **Formulate Task-Specific Criteria**: Propose criteria to evaluate the quality of future research directions.
3. **Provide Rationale**: For each criterion, provide a concise explanation (`explanation`).
4. **Assign Weights**: Assign a weight (`weight`) to each criterion, ensuring the sum of all weights is exactly **1.0**.

Core Requirements:
1. **Task-Centric:** Your criteria must be directly linked to the nuances of the `<task>`.
2. **Sufficient Justification:** The `<analysis>` must explain your reasoning for the criteria and weights.
3. **Standard Output Format:** Strictly follow the example format.
</Instruction>

<example>
<task>
"A comprehensive survey on Vision Transformers."
</task>
<output>
<analysis>
For a Vision Transformer survey, strong future research directions must move beyond generic suggestions. Key criteria should assess whether the suggestions are grounded in the identified limitations (e.g., data-hungriness, computational cost) and whether they propose specific, feasible research avenues. Therefore, "Specificity and Feasibility" and "Grounding in Literature Gaps" are weighted most heavily.
</analysis>

Figure 22: GRADE prompt for Research Outlook Quality (ROQ): Step 1(a) - rubric generation.

<json_output>
[
  {{
    "criterion": "Specificity and Feasibility",
    "explanation": "Assesses if the proposed research questions are concrete and actionable, rather than vague, high-level suggestions. A good direction is one that a researcher could immediately start designing an experiment for.",
    "weight": 0.3
  }},
  {{
    "criterion": "Grounding in Literature Gaps",
    "explanation": "Evaluates whether the proposed directions logically follow from the limitations, challenges, and open problems identified in the main body of the survey.",
    "weight": 0.3
  }},
  {{
    "criterion": "Potential Impact and Insightfulness",
    "explanation": "Measures the potential of the proposed research to significantly advance the field, open new sub-fields, or resolve major theoretical or practical bottlenecks.",
    "weight": 0.25
  }},
  {{
    "criterion": "Novelty of Questions",
    "explanation": "Assesses whether the research directions are fresh and forward-looking, rather than restating well-known, incremental next steps.",
    "weight": 0.15
  }}
]
</json_output>
</output>
</example>

Now, please perform this task for the following research prompt:
<task>
{task_prompt}
</task>
</user_prompt>

Figure 23: GRADE prompt for Research Outlook Quality (ROQ): Step 1(b) - rubric generation.

Prompt Templates for Evaluating "Research Outlook Quality" - Criteria-Based Scoring

<system_role>
You are a rigorous, meticulous, and objective academic reviewer. You are skilled at deeply comparing the "**research outlook quality**" sections of two research reviews based on specific evaluation criteria, providing precise scores with clear justifications.
</system_role>

<user_prompt>
Task Background
You need to evaluate the quality of the proposed future research directions in two survey papers. The surveys were written to address the following research task:
<task>
{task_prompt}
</task>

Articles to be Evaluated
<article_1>
{article_1_text}
</article_1>

<article_2>
{article_2_text}
</article_2>

Evaluation Criteria: Research Outlook Quality
Now, you must evaluate and compare the future research directions proposed in these two articles on a criterion-by-criterion basis according to the following **list of criteria**.

<criteria_list>
{criteria_json_string}
</criteria_list>

<Instruction>
Your Task
Strictly following **each criterion** in the `<criteria_list>`, compare and evaluate the performance of `<article_1>` and `<article_2>` on that criterion. Your analysis should focus ONLY on the sections discussing conclusions, future work, challenges, and future prospects.

Scoring Rubric
For each criterion, score both articles on a continuous scale from 0 to 10.

Output Format Requirement
Please **strictly** follow the `<output_format>` below. Ensure the final output is a single, valid JSON object that can be parsed directly.
</Instruction>

<output_format>
```
{{
   "research_heuristics": [
      {{
         "criterion": "[Text of the first evaluation criterion]",
         "analysis": "[Comparative analysis]",
         "article_1_score": [0-10 continuous score],
         "article_2_score": [0-10 continuous score]
      }},
      ...
   ]
}}
```
</output_format>

Now, based on the evaluation criteria, please assess the two articles and provide a detailed comparative analysis and scores as required.
</user_prompt>

Figure 24: GRADE prompt for Research Outlook Quality (ROQ): Step 2 - criteria-based scoring.