LitReview Arena: Evaluating Literature Review Agents with Battle-style Peer Review Platform

Anonymous Authors[1]

Abstract

Literature reviews are essential to reflect the landscape of research fields. Large language models, especially deep research agents, have recently shown strong capabilities in automated literature review generation. However, it remains a challenging task to rigorously evaluate the scientific value of the generated reviews, since human expert annotations are difficult to scale up and LLM-as-a-judge approaches lack of a convincing criteria. To address this gap, we introduce **LitReview Arena**, a battle-style evaluation platform with a structured protocol tailored to literature review quality. Our protocol imitates academic peer review by recruiting domain experts with research paper-writing experience, and we match each query to reviewers within their expertise. Each battle is judged with dimension-wise outcomes over five literature-review-specific criteria, enabling reproducible and diagnostic comparisons across systems. We collect a large-scale human preference dataset of expert votes (4984 votes$\times$5 dimensions) and systematically measure how far current models are from human drafts. Results show that the most advanced models win only 23.0% of decisive matches against humans on overall utility, leaving substantial room for improvement. Meanwhile, agentic LLMs, such as Sonar Deep Research, substantially outperform base language models by over 60%. We also find that existing LLM-as-a-judge evaluation methods are severely misaligned with human experts (Spearman's $\rho \approx 0.467$). Based on the collected preference data, we provide an expert-calibrated evaluator, *LitJudge*, improving alignment to $\rho \approx 0.78$, comparable to inter-expert consistency. Codes and datasets are publicly available at https://anonymous.4open.science/r/LitReview-Arena-3B82/.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Scientific literature reviews condense a fast-moving field into a usable mental model. They shape how researchers categorize methods, what evidence they treat as settled, and which gaps are worth pursuing next. As deep research systems improve, automated survey generation is no longer a speculative goal; the bottleneck is increasingly evaluation. Existing work has progressed in stages. Early evaluation systems emphasize literature retrieval and organization, but do not provide an evaluation mechanism for the *review* itself (Tang et al., 2024). More recent benchmarks and systems expand from retrieval to end-to-end review generation (e.g., DeepResearch Bench; Du et al., 2025), yet their evaluations are typically based on static reports, fixed rubrics, or proxy signals such as reference correctness, coverage heuristics, and overlap against a reference document (Tang et al., 2024; Du et al., 2025). These signals matter, but they do not resolve the central challenge in literature review assessment: many of the criteria that determine research utility are *non-verifiable* and inherently intuition-based. In a good review, structure is not a matter of formatting; it is an argument about what the field is and how its subareas relate. Likewise, strong gap and direction sections rarely come from generic templates. They require interpreting what is missing, what assumptions are fragile, and what experiments would resolve uncertainty. Without subjective, dynamic human preference as a first-class signal, evaluation struggles exactly where the review matters most to researchers.

Arena-style evaluation emerges as an effective approach to address this gap. LMArena demonstrated that arena votes can yield stable rankings for open-ended generation tasks that are difficult to score with absolute metrics (Zheng et al., 2023; Liu et al., 2023; Chiang et al., 2024). In literature-grounded scientific settings, SciArena extends this paradigm to non-verifiable tasks by using a battle platform and preference voting (Zhao et al., 2025). However, its target is generalized scientific, literature-grounded tasks and cannot provide a specialized, structured, end-to-end protocol tailored to literature review quality. We argue that literature

review evaluation requires *specialized* improvements in protocol design to make expert preference data both reliable and reproducible: (1) Annotator: we recruit domain experts and constrain participation to researchers with AI paper-writing experience, rather than open enrollment, because assessing review structure and research directions depends on field-level judgment. (2) Query selection: we derive topics from real, highly cited surveys to ensure the underlying research questions are valuable and recognizable to the community, rather than synthetic prompts with unclear academic relevance. (3) Expertise matching: we precisely match expert annotators and queries, assigning each topic to annotators within their areas of expertise, so that voters evaluate reviews in domains they can genuinely peer review. (4) Structured dimensions: we define five dimensions tailored to literature review assessment: D1-literature coverage, D2-claim support, D3-paper structure, D4-research suggestions quality, and D5-overall utility, so that both citation-facing criteria and the non-verifiable dimensions that dominate research utility are explicitly captured (Yan et al., 2025). Putting these together, our protocol is designed to simulate a peer review system that is widely recognized and used in academia, while retaining the operational simplicity of arena voting. This is the key axis on which we *specialize beyond SciArena*: the battle platform is not only a scaling device, but a mechanism to enforce scientific rigor for a single task family (literature reviews) with a complete, structured evaluation protocol.

Based on this protocol, we introduce the **LitReview Arena** platform and release LitReviewBench, an arena-grounded expert preference benchmark for literature review agents. We collect a large-scale dataset of expert votes ($\sim$ 3k $\times$ 5 dimensions) and conduct, to our knowledge, the first systematic study of how far current models are from human-level literature review performance under expert preference evaluation. We find that even the most advanced models remain far from human-level performance on overall utility (winning rate $= 23.0\%$ in decisive matches against human drafts). *Paper structure* (D3) and *research suggestions* (D4) trail human experts by roughly 200 points, indicating that coherent landscape organization and non-obvious gap identification remain difficult even for strong systems. We also observe a consistent system-level trade-off: *agentic models* substantially outperform pure *language models* across dimensions, but at a substantial test-time computing cost—on average consuming roughly $15\times$ the token budget for the same topic. To make expert-aligned evaluation practical at low ongoing cost, we release a low-cost, expert-aligned evaluator for offline evaluation. Uncalibrated LLM-as-a-judge methods are severely misaligned with expert preferences, producing leaderboards that disagree with expert rankings. Using LitReviewBench as a calibration signal, our evaluator improves expert-ranking alignment from $\rho \approx 0.467$ to $\rho \approx 0.7919$ on D0, enabling scalable offline assessment and providing a training signal that can be used to improve literature review agents with expert preference supervision.

Our contributions are as follows.

- We propose the **LitReview Arena** platform together with an evaluation protocol that emphasizes fairness and scientific rigor: battle-style blind evaluation, expert-only voting, topic–expertise matching, and structured evaluation that support diagnostic analysis.
- We collect a large-scale dataset of expert votes ($\sim$ 3k $\times$ 5 dimensions) and systematically evaluate model literature review quality against human drafts under expert preference. We find that non-human systems win only 23.0% of decisive matches against human drafts on overall utility, with the largest gaps on paper structure and research suggestions (roughly 200 points), while agentic models outperform pure language models at an average $15\times$ token cost.
- We release a low-cost, expert-aligned evaluator for offline evaluation. Using LitReviewBench as a calibration signal, it improves expert-ranking alignment from $\rho \approx 0.467$ to $\rho \approx 0.7919$ on D0, enabling scalable offline assessment and providing a training signal for improving literature review agents with expert preference supervision.

2. Related Work

2.1. Automated Literature Review Generation Systems

A growing line of work treats literature review writing as an end-to-end retrieval-and-synthesis problem: given a topic, the system collects a paper set, organizes it into an outline, and produces a narrative with citations. For example, AutoSurvey focuses on generating survey-style drafts from retrieved literature, often centering evaluation on reference generation and overall coherence (Wang et al., 2024). SurveyForge studies practical heuristics for survey writing—notably outline construction and memory-driven generation—and proposes multi-dimensional evaluations that still largely rely on operationalizable criteria (Yan et al., 2025). SurveyX frames academic survey generation as a pipeline that combines retrieval, clustering, and section-level generation (Liang et al., 2025). Other agents explore more explicit knowledge organization, such as building multiple lightweight knowledge graphs (minigraphs) and aggregating them into a review (Liu, 2024).

Despite real progress in coverage and fluency, these systems are typically positioned closer to *paper search and integration tools* than to expert-level synthesizers. In practice, they often under-specify what makes a review *usable* to researchers: a defensible global structure (what belongs

together and why) and a gap narrative that goes beyond generic future work templates. As a result, improvements in retrieval quality and citation formatting do not necessarily translate into better scholarly synthesis, especially when the target is to help readers reason about the field's organizing axes and the non-obvious directions that follow.

2.2. Static Benchmarks for Literature Review Quality

Benchmarking literature review agents has lagged behind system building, partly because expert-valued qualities are hard to reduce to deterministic checks. Existing benchmarks and datasets therefore skew toward what can be measured reliably at scale: topical relevance, summary similarity, citation matching, and claim-level factuality (Kasanishi et al., 2023; Ajith et al., 2024). Several efforts explicitly target literature-grounded generation and evaluation (e.g., SciReviewGen, SurGE, ReportBench, and related tracks), but the dominant pattern remains: evaluation prefers criteria with clear automatic signals, while higher-order synthesis is either coarse-grained or absent (Kasanishi et al., 2023; Su et al., 2025; Li et al., 2025). Recent work has also proposed benchmarks that aim to evaluate the academic value of survey-like outputs, e.g., DeepSurveyBench (Zhang et al., 2026).

A common response is to use *LLM-as-a-judge* to score richer rubrics(Hashemi et al., 2024). However, the judge model is not an oracle: its preferences can drift with prompt wording, it can overweight surface features, and it may fail precisely where domain expertise matters most. Recent meta-evaluation results on scientific, literature-grounded tasks show that even strong models only partially match expert preferences, underscoring the difficulty of using off-the-shelf judging setups as the primary development signal (Zhao et al., 2025).

LitReviewBench is designed to complement these benchmarks rather than replace them. We keep the *citation-aware* dimensions (coverage; citation–claim support) because they are necessary, but we center evaluation on two expert-facing dimensions that are routinely under-measured: *review landscape structure* and *gap/direction quality*. Importantly, our labels come from topic-matched researchers, and our offline evaluator(LitJudge) is calibrated to preserve those expert tradeoffs rather than optimizing for generic judge agreement.

2.3. Arena-Style Human Preference Evaluation

Arena-style evaluation provides an appealing alternative to static benchmarks: it supports open-ended queries, produces relative judgments that are easier to make consistently than absolute scores, and naturally yields leaderboards. Chatbot Arena (LMArena) established the now-standard recipe of blind pairwise comparison with Elo-style aggregation (Chiang et al., 2024). SciArena adapts this paradigm to *non-verifiable, literature-grounded scientific tasks*, explicitly targeting domains where expertise and retrieval quality are central (Zhao et al., 2025).

However, live arenas are not immediately usable as *benchmarks for new methods*. They are expensive to scale with experts; their query distribution evolves over time; and reproducing results for offline testing is non-trivial. LitReview Bench takes the arena advantages, topic diversity, and pairwise expert preference, but converts them into a frozen, versioned dataset with standardized per-dimension outcomes, so that progress can be measured directly offline. In addition, we introduce an expert-aligned evaluator that uses structure- and topic-matched in-context cases plus human-written gap anchors and produces quality signals, making expert-grounded evaluation and scalable training practical at low ongoing human cost.

3. The LitReviewBench Construction

LitReviewBench is constructed in three stages: (1) defining a survey draft generation task from high quality survey literature and mapping each instance into a consistent topic taxonomy, (2) collecting blind expert preferences via pairwise comparisons on LitReview Arena, which is part of the ScienceArena Community (Shao et al., 2025), and (3) freezing the resulting arena logs into a versioned offline benchmark that supports direct testing and standardized leaderboards.

3.1. Task Setup and Topic Taxonomy

LitReviewBench evaluates literature review systems on producing survey drafts that are useful to researchers. The task mirrors research practice: a good draft must not only list relevant papers but also organize a field into an interpretable landscape and surface gaps that plausibly inform next steps.

Source pool from OpenAlex. We retrieve papers from OpenAlex (Priem et al., 2022) with concept *Artificial Intelligence*, publication time in 2022 to 2025, and citation count > 50, retaining survey style papers as the seed pool (3,000+ papers). This anchors our topics in recognized research questions while staying recent enough to reflect current subfields.

One topic extraction per survey. From each selected survey, we use an LLM based extractor to produce a single topic phrase that captures the scope at the level of a survey prompt. We normalize these extracted topics into a consistent query form to reduce unintended variation, using the template: *Conduct a literature review on* $\{$*Topic*$\}$. Normalization focuses on de duplication and phrasing edits that preserve scope while improving clarity.

Field and subfield assignment. Each topic is assigned a

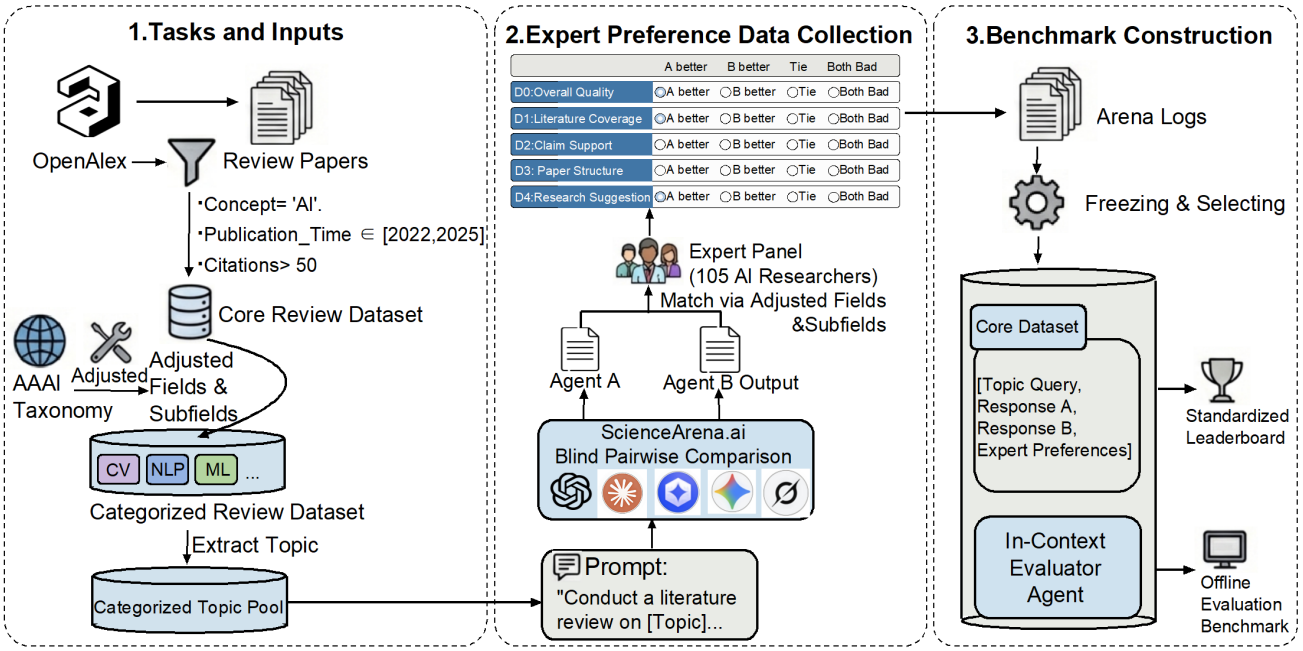

Figure 1. **LitReviewBench construction overview.** Left: topic extraction from high quality AI survey papers curated from OpenAlex and assignment into an AAAI based field and subfield taxonomy. Middle: expert preference collection on LitReview Arena using blind pairwise comparisons and five dimension wise four way outcomes (D1 to D5). Right: benchmark construction by freezing arena logs into a versioned dataset for offline evaluation and standardized leaderboards.

field and subfield using an AAAI based taxonomy refined for modern AI. These tags enable expertise matching during annotation and support stratified analysis to test generalization beyond dominant areas.

3.2. LitReview Arena Pairwise Evaluation Protocol

LitReviewBench collects judgments through LitReview Arena, leveraging the stability of relative comparisons for open ended generation (Chiang et al., 2024; Zhao et al., 2025). For each topic, two candidate drafts generated by different systems are displayed side by side with identities hidden and order randomized.

The atomic unit is a *battle record* containing the topic query, paired drafts, and expert outcomes. This format supports standard pairwise aggregation methods used in arenas, including Bradley Terry models and Elo style ratings (Bradley & Terry, 1952; Elo, 1978). LitReviewBench preserves this battle format in the released dataset to ensure downstream evaluation follows a consistent interface.

3.3. Expert Preference Data Collection

We recruit 105 annotators with AI paper writing experience, matched to topics within their areas of familiarity using the field and subfield tags from Section 3.1. This matching is critical for dimensions relying on disciplinary norms, particularly review structure and gap quality.

Each comparison uses a four way outcome set: **A**, **B**, **Tie**, and **Both Bad**. Every battle receives five separate votes:

- **D1 Literature Coverage:** Which draft cites a more complete set of relevant papers with fewer omissions?
- **D2 Claim Support:** Which draft more reliably supports key claims with relevant citations?
- **D3 Paper Structure:** Which draft better organizes prior work into categories or comparisons that clarify relationships among approaches?
- **D4 Research Suggestions Quality:** Which draft more clearly identifies important, non obvious gaps or future directions?
- **D5 Overall Utility:** From the perspective of a researcher, which draft is preferred as a starting point?

This design yields an overall preference signal (D5) alongside four diagnostic signals (D1 to D4) that isolate aspects of scholarly synthesis where automated setups often drift. Records store minimal metadata for auditing, including pseudonymized annotator identifiers and field tags, allowing for granular analysis without embedding personal information.

3.4. From Arena Logs to a Reproducible Offline Benchmark

We freeze LitReview Arena logs by selecting a stable set of topics and associated battles, packaging expert outcomes as fixed labels. The resulting dataset is released in versioned form to ensure reproducibility over time.

Each benchmark instance contains the topic query, paired drafts, four way outcomes for D1 to D5, and aggregation metadata. Preserving Tie and Both Bad as explicit outcomes enables downstream evaluators to handle neutrality and quality failure separately. Finally, we support standardized leaderboards by applying the same BT and Elo style aggregation used in arena settings to the frozen outcomes (Bradley & Terry, 1952; Elo, 1978), producing comparable system ratings per dimension independent of specific judge models.

4. Results on SOTA Models and Agents

4.1. Overall Performance and Leaderboards

LitReviewBench comprises 3k expert votes across five dimensions. We aggregate outcomes into dimension-wise leaderboards using a standard Bradley–Terry model with Elo-style ratings (Bradley & Terry, 1952; Elo, 1978). Table 1 reports the ratings for the four diagnostic dimensions (D1–D4) and Overall Utility (D5).

When competing directly against human drafts (the upper-bound reference), state-of-the-art models and agents win only 208 out of 904 decisive matches on D5, a win rate of 23.0%.[1] This gap confirms qualitative expert feedback: functional research starting points require more than well-formed summaries.

The human baseline ranks first across all five dimensions. Among non-human systems, GPT-5.2 leads on D1–D3, while Sonar Deep Research leads on D4 (Research Suggestions), though still trailing humans. These standings indicate that current systems struggle with field organization and actionable direction-setting.

High performance correlates with computational cost. As detailed in Table 1, agentic systems consume 122.3K tokens per query on average—15× the budget of standalone models (8.1K). While allocating computation to search and synthesis yields comprehensive gains over raw models, it has not yet closed the utility gap to expert expectations.

4.2. Fine-grained Analysis of Model Capability Gaps

Inspection of battle records identifies four persistent failure modes: incomplete or poor literature coverage (D1), weak claim support (D2), shallow organization (D3), and generic or impractical research suggestions, such as research gaps and future directions (D4). Data from the performance gap between the top overall model, GPT-5.2, and human experts reveal that, across diagnostic dimensions D1–D4, systems consistently lag behind expert drafts, with the most pronounced deficits in Paper Structure (D3) and Research Suggestions (D4). These two dimensions are the strongest predictors of expert preference: Spearman correlations with Overall Utility (D5) are 0.99 (D3) and 0.96 (D4), exceeding that of Literature Coverage (0.90). Expert utility thus depends less on citation volume than on conceptual coherence and non-trivial future directions. Qualitatively, drafts often resemble efficient bibliographies rather than synthetic reviews.

Increasing draft length or interaction turns does not automatically yield higher D3/D4 scores. Experts reward drafts that make latent relationships legible, not those that simply expand coverage. LitReviewBench therefore exposes a reasoning gap invisible to checklist evaluations. Since D3 and D4 are also where off-the-shelf LLM judges diverge most from experts (Section 5), reliance on uncalibrated automated metrics risks misguiding development.

5. Meta-evaluation of LLM-based Judges

Collecting expert preferences is resource-intensive, raising the question of whether LLM-based judges serve as reliable proxies. Prior work suggests scientific literature tasks pose greater challenges than general chat evaluation, with evaluators often failing to match expert consensus (Zhao et al., 2025). We conduct a meta-evaluation on LitReviewBench to quantify this alignment gap.

Setup. We sample 500 battle instances from LitReviewBench, preserving the task format and 4-way expert votes. We employ **Qwen/Qwen3-235B-A22B-Instruct-2507** as the automated judge using a minimal prompt that forces a decision for each of the five dimensions. We run the judge once per instance and dimension.

Scoring and aggregation. We evaluate alignment via instance-level accuracy, assigning 0.5 credit for neutral expert outcomes, and leaderboard-level Spearman correlation. Judge-induced system rankings are derived using the same Bradley–Terry aggregation as Section 4.1 to enable direct comparison with expert standings (Table 2). Table 3 summarizes agreement and reliability.

5.1. Blindspots of LLM Judges: Systematic Mismatches with Human Experts

We observe that the judge's reliability varies fundamentally depending on the nature of the evaluation sub-task. On

[1] Win rate calculated on decisive outcomes only.

Category	Methods	Token Cost/Query	Literature Coverage	Claim Support	Paper Structure	Research Suggestions	Overall Utility
Human	**human**	**N/A**	**1787.4**	**1565.6**	**1502.5**	**1521.5**	**1668.8**
Agentic Models	GPT-5.2	38.096K	1632.8	1536.5	1322.4	1272.7	1449.1
	Sonar Deep Research	322.080K	1175.9	1106.3	1262.1	1322.7	1285.9
	Qwen Deep Research	70.265K	863.3	925.4	1125.0	1178.4	1117.3
	OpenAI Deep Research	58.745K	882.2	943.3	857.5	867.3	874.1
	Average	**122.297K**	**1138.6**	**1127.9**	**1141.8**	**1160.3**	**1181.6**
Language Models	Claude Opus 4.5	5.490K	1129.0	1032.8	1177.6	1099.9	1135.5
	Qwen3 235B	5.276K	759.3	826.3	886.0	877.4	836.2
	Grok 4	16.399K	886.8	908.5	780.2	778.2	799.2
	GLM 4.6	4.741K	436.7	553.6	564.0	608.5	434.3
	Gemini 2.5 Pro	8.586K	446.1	601.8	526.0	469.8	400.2
	Average	**8.098K**	**731.6**	**784.6**	**786.8**	**766.8**	**721.1**

Table 1. Expert-preference leaderboards on LitReviewBench. *Token Cost/Query* (third column): lower is better, reported in thousands of tokens based on exact measurements. *Utility metrics* (columns 4 to 8): higher is better, ordered as Literature Coverage, Claim Support, Paper Structure, Research Suggestions, and Overall Utility. Systems are grouped as Human, Agentic Models, and Language Models. Human achieves the highest utility scores across all metrics; GPT-5.2 leads non-human systems in Overall Utility. .

Category	Methods	Literature Coverage	Claim Support	Paper Structure	Research Suggestions	Overall Utility
Human	**human**	**378**	**321**	**317**	**407**	**310**
Agentic Models	GPT-5.2	2439	2470	2485	2367	2490
	Sonar Deep Research	2012	2071	2094	1864	2096
	Qwen Deep Research	1320	1320	1371	1392	1370
	OpenAI Deep Research	802	767	763	775	761
	Average	**1643.3**	**1657.0**	**1678.3**	**1599.5**	**1679.3**
Language Models	Claude Opus 4.5	1163	1033	1035	915	1068
	Qwen3 235B	848	926	945	1069	923
	Grok 4	575	559	524	625	521
	GLM 4.6	293	305	255	350	251
	Gemini 2.5 Pro	129	174	165	203	165
	Average	**601.6**	**599.4**	**584.8**	**632.4**	**585.6**

Table 2. Judge-induced leaderboards on LitReviewBench using Qwen/Qwen3-235B-A22B-Instruct-2507 as the evaluator. Scores are Elo ratings derived from Bradley–Terry aggregation, and higher is better.

Literature Coverage (D1), which primarily tests retrieval recall, Qwen correlates moderately well with experts ($\rho = 0.552$). The model effectively identifies whether a draft includes the expected set of canonical papers.

However, performance degrades as the task shifts from retrieval to synthesis. On Claim Support (D2), Paper Structure (D3), and Research Suggestions (D4), the uncalibrated judge struggles to differentiate high-quality analysis from plausible but incorrect statements.

More critically, Table 2 reveals a severe **AI to AI bias**. While experts rank human drafts first (Table 1), the automated judge penalizes human writing aggressively. On Overall Utility (D5), it assigns humans an Elo score of 310 while boosting GPT-5.2 to 2490. This inversion suggests the judge conflates model-like fluency with quality, making it unreliable for comparing human and machine outputs without calibration.

5.2. Consistency and Reliability

A common hypothesis attributes disagreement on high-level dimensions to inherent task subjectivity. We test this by comparing agreement accuracy for expert pairs versus judge pairs. For judge pairs, we measure cross-model agreement between **Qwen/Qwen3-235B-A22B-Instruct-2507** and **DeepSeek-V3.2** to assess whether failure patterns remain consistent across different LLM judges.

Expert agreement is consistently higher than judge agreement on D3 and D4, reaching 0.861 on Overall Utility (D5) (Table 3). This high consensus indicates that the quality signal is not dominated by noise. Notably, **judge-to-judge agreement remains high** (> 0.7) **across all dimensions**, even where alignment with experts is poor. This indicates that judge failures are systematic rather than random, reflecting shared biases toward surface polish over deep structure.

Dimension	Judge–Expert Accuracy	Spearman's ρ	Expert–Expert Agreement Accuracy	JudgeLLM–JudgeLLM Agreement Accuracy
D1 (Literature Coverage)	0.586	**0.552**	0.833	**0.783**
D2 (Claim Support)	0.554	0.442	0.556	0.747
D3 (Paper Structure)	0.598	0.467	0.639	0.747
D4 (Research Suggestions)	**0.620**	0.430	0.556	0.739
D5 (Overall Utility)	0.606	0.467	**0.861**	0.747

Table 3. Agreement with expert preference and reliability. Judge–expert accuracy merges Tie and Both Bad into a neutral outcome and assigns 0.5 credit regardless of the judge decision. Spearman's ρ is the rank correlation between judge-induced and expert-induced BT and Elo leaderboards. Expert–expert agreement accuracy and JudgeLLM–JudgeLLM agreement accuracy are reported as mean pairwise accuracy. JudgeLLM–JudgeLLM agreement measures cross-model consistency between two judge models, Qwen/Qwen3-235B-A22B-Instruct-2507 and DeepSeek-V3.2, indicating whether failure patterns remain stable across different LLM judges.

5.3. Implications for Evaluator Calibration

Naive LLM judging is an insufficient substitute for expert preference. The capabilities required to assess claim grounding, structure, and research gaps differ significantly from those required for simple coverage checking.

Leaderboards derived from uncalibrated judges are deceptively stable yet inaccurate. As shown in Table 2, judge-based rankings systematically undervalue human baselines and over-weight surface-level mechanics. These limitations motivate calibrating the evaluator with task-specific in-context examples, leading to our calibrated evaluator in Section 6.1.

6. Closing the Loop: Expert-aligned Evaluator

Progress on literature review agents is constrained less by generation capabilities than by evaluation reliability. Standard LLM-based evaluators perform adequately on Literature Coverage (D1) but fail to capture expert utility on Claim Support (D2), Paper Structure (D3), and Research Suggestions (D4). We address this gap by using LitReviewBench as a calibration signal and instantiating an expert-aligned evaluator, **LitJudge**. As shown in Figure 2, LitJudge conditions on case context that matches the test instance in structure and topic, and it grounds D4 decisions with gap anchors derived from expert-written drafts.

6.1. Expert-in-the-loop Calibration

We employ **Qwen/Qwen3-235B-A22B-Instruct-2507**, consistent with Section 5, and use a 500 instance subset from LitReviewBench for calibration and evaluation.

Calibrated evaluator context (Figure 2a). For each battle, we construct an in-context packet with up to three demonstrations per group:

- **Group S (Structure-similar; for D3).** We extract a skeleton text (headers and lead sentences) to derive a paragraph relationship network capturing discourse transitions. We retrieve battles with the most similar networks based on normalized graph similarity.
- **Group C (Content-similar; for D1/D2).** We retrieve battles with topics closest to the query using LLM-based embedding matching. These cases provide local standards for sufficient coverage and plausible citation–claim support.
- **Group G (Gap anchors; for D4).** We extract gap anchors exclusively from expert-written drafts (`lit-review_human`). These bullet points serve as grounding exemplars for high-quality, non-hallucinated directions.

6.2. Efficacy of the Calibrated Evaluator

We evaluate judge alignment by agreement accuracy against expert votes, using the same scoring protocol as Section 5. Figure 2b summarizes the gains from calibration.

LitJudge improves alignment most on synthesis dimensions, while keeping coverage behavior comparable. On Literature Coverage (D1), the gain is modest, from 0.552 to 0.5758. On Claim Support (D2) and Paper Structure (D3), alignment increases substantially, from 0.442 to 0.6727 and from 0.467 to 0.6485. The largest jump is on Research Suggestions (D4), from 0.43 to 0.8424, consistent with the role of expert-derived gap anchors. Overall Utility (D5) also rises sharply, from 0.467 to 0.7919, indicating that calibration helps recover the expert holistic preference signal.

This enables repeated offline evaluation that tracks expert tradeoffs with low marginal cost.

7. Discussion

LitReview Arena targets a practical bottleneck for literature review agents: progress is now limited more by how we measure scholarly value than by how fluently systems can write. Our results show a clear split across dimensions. When the judgment is close to retrieval verification, such as Literature Coverage (D1), uncalibrated LLM judges track

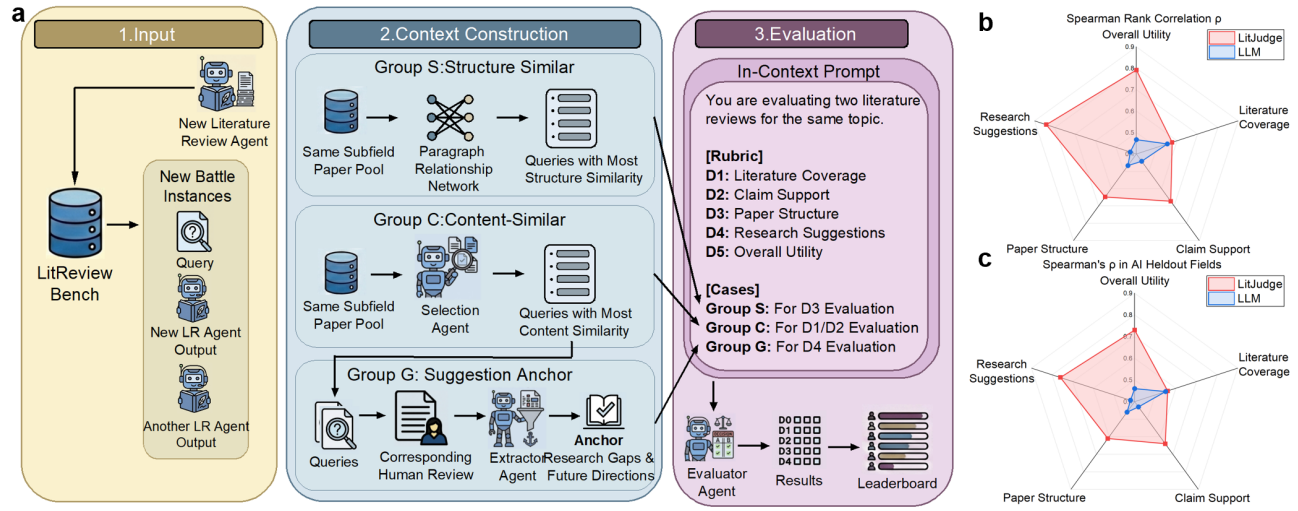

Figure 2. **Expert-aligned evaluator workflow and results.** (a) Context construction for the calibrated evaluator. (b) Calibration improves alignment between judge and expert votes. (c) In-domain generalization to held-out AI subfields (20%) shows robust transfer.

experts moderately well. When the judgment depends on synthesis, especially Paper Structure (D3) and Research Suggestions (D4), the same judges become unreliable, including large ranking inversions between human drafts and model outputs. This is consistent with expert feedback that utility is driven by organizing axes and actionable gaps, not by surface completeness alone.

Methodologically, LitReview Arena provides a battle-style peer review interface that makes expert evaluation scalable and repeatable. Relative votes are easier to apply consistently than absolute scores, and they support robust aggregation with Bradley–Terry and Elo-style estimators (Bradley & Terry, 1952; Elo, 1978). The explicit outcomes Tie and Both Bad help separate indifference from quality failure, which is important when systems are close in quality or both are weak. Freezing arena logs into LitReviewBench then turns this live preference signal into a stable offline benchmark and a comparable leaderboard interface.

The calibrated evaluator closes part of the loop. The meta-evaluation suggests that judge errors are not merely noise, since different judge models can agree strongly while still disagreeing with experts. LitJudge addresses this by conditioning the judge on task-specific context, including structure-matched cases and expert-derived gap anchors. This design improves alignment on the dimensions that matter most for utility, and it makes offline iteration feasible without repeatedly convening large expert panels. A natural workflow is to use LitJudge for inner-loop development and reserve expert votes on LitReview Arena for periodic audits and refreshes.

Several limitations remain. First, our current scope is anchored in AI topics, so generalization to domains with different evidential norms, such as clinical medicine, is not yet verified. Second, the notion of good structure and good gaps is shaped by the preferences of our annotator pool, and different communities may weight virtues differently. Third, calibration can entrench biases if structure matching favors familiar rhetorical forms or if gap anchors over-represent fashionable directions, so transparency about exemplar selection and periodic refreshes are necessary. Finally, pairwise preferences capture relative judgments but do not fully disentangle writing style from substance, and they do not cover long-horizon settings such as maintaining living reviews. Complementary evaluations, including targeted factual audits and longitudinal update tasks, are promising directions.

8. Conclusion

We introduced LitReview Arena, a battle-style peer review platform for literature review agents, and LitReviewBench, an offline benchmark distilled from arena logs. Analyzing 3k+ expert votes, we show that current SOTA systems still lag behind human drafts, particularly in organizing coherent landscapes and generating non-obvious research insights. We also find that uncalibrated LLM judges are unreliable for synthesis-heavy tasks and introduce systematic biases when comparing human and model outputs. To address this, we developed LitJudge, a calibrated evaluator using structure-matched cases and expert-derived anchors, which significantly improves agreement with experts and enables efficient offline iteration. Overall, LitReview Arena provides a robust evaluation interface for complex scientific writing, while LitReviewBench and LitJudge bridge the gap between expert feedback and reproducible testing, aligning development signals with actual researcher needs.

References

Ajith, A., Xia, M., Chevalier, A., Goyal, T., Chen, D., and Gao, T. Litsearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, 2024. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Du, M. et al. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

Elo, A. E. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, 1978. ISBN 9780668047210.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., and Kedzie, C. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*, 2024. URL https://arxiv.org/abs/2501.00274.

Kasanishi, T., Isonuma, M., Mori, J., and Sakata, I. Scireviewgen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6695–6715, 2023. doi: 10.18653/v1/2023.findings-acl.418. URL https://aclanthology.org/2023.findings-acl.418/.

Li, M., Zeng, Y., Cheng, Z., Ma, C., and Jia, K. Reportbench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*, 2025. URL https://arxiv.org/abs/2508.15804.

Liang, X., Yang, J., Wang, Y., Tang, C., Zheng, Z., Song, S., Lin, Z., Yang, Y., Niu, S., Wang, H., Tang, B., Xiong, F., Mao, K., and Li, Z. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025. URL https://arxiv.org/abs/2502.14776.

Liu, e. Mixture of knowledge minigraph agents for literature review generation. *arXiv preprint arXiv:2410.17166*, 2024.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.153/.

Priem, J., Piwowar, H., and Orr, R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Shao, C., Huang, D., Li, Y., Zhao, K., Lin, W., Zhang, Y., Zeng, Q., Chen, Z., Li, T., Huang, Y., Wu, T., Liu, X., Zhao, R., Zhao, M., Li, J., Zhang, X., Wang, Y., Zhen, Y., Xu, F., Li, Y., and Liu, T.-Y. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists, 2025. URL https://arxiv.org/abs/2511.16931.

Su, W., Xie, A., Ai, Q., Long, J., Mao, J., Ye, Z., and Liu, Y. Benchmarking computer science survey generation. *arXiv preprint arXiv:2508.15658*, 2025. URL https://arxiv.org/abs/2508.15658.

Tang, X. et al. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. *arXiv preprint arXiv:2412.13612*, 2024. URL https://arxiv.org/abs/2412.13612.

Wang, Y., Guo, Q., Yao, W., Zhang, H., Zhang, X., Wu, Z., Zhang, M., Dai, X., Zhang, M., Wen, Q., Ye, W., Zhang, S., and Zhang, Y. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*, 2024. URL https://arxiv.org/abs/2406.10252.

Yan, X., Feng, S., Yuan, J., Xia, R., Wang, B., Zhang, B., and Bai, L. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025. URL https://arxiv.org/abs/2503.04629.

Zhang, G.-B., Liu, D.-Y., Wu, D.-Y., Lan, T., Huang, H., Wu, Z., and Mao, X.-L. Deepsurvey-bench: Evaluating academic value of automatically generated scientific survey, 2026. URL https://arxiv.org/abs/2601.15307.

Zhao, Y., Zhang, K., Hu, T., Wu, S., Le Bras, R., McGrady, C., Anderson, T., Bragg, J., Chang, J. C., Dodge, J., Latzke, M., Liu, Y., Tang, X., Wang, Z., Zhao, C., Hajishirzi, H., Downey, D., and Cohan, A. Sciarena: An open evaluation platform for non-verifiable

scientific literature-grounded tasks. *arXiv preprint arXiv:2507.01001*, 2025. URL https://arxiv.org/abs/2507.01001.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z. L., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Zhang, H. Judging llm-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

A. APPENDIX

A.1. A.1 Expert Evaluation Protocol

Outcome set. Each dimension is labeled with one of {A, B, Tie, BothBad}.

- **A**: Draft A is better on this dimension.
- **B**: Draft B is better on this dimension.
- **Tie**: both drafts are comparably good.
- **BothBad**: neither draft is acceptable.

Dimension questions (verbatim).

- **D0 Overall Utility:** From a researcher's perspective, which response would you prefer to use as a starting point for a literature review on this topic?
- **D1 Literature Coverage:** Which response cites a more complete and appropriate set of relevant papers for this topic, with fewer obvious omissions of important work?
- **D2 Claim Support:** Which response more reliably grounds its key claims in the cited literature, with citations that are relevant and appropriately support the statements made?
- **D3 Paper Structure:** Which response better structures the existing literature by organizing prior work into clear categories or comparisons that help a researcher understand relationships among approaches, rather than listing papers independently?
- **D4 Research Suggestions Quality:** Which response more clearly identifies important and non-obvious research gaps or future directions that go beyond generic summaries and would meaningfully inform a researcher's next steps?

Interface rules. Two drafts are presented side-by-side with anonymized system identity and randomized left/right ordering. Annotators are allowed to consult external resources (e.g., follow citations / search for papers) when forming judgments.

A.2. A.2 Data Schema (Battle Record)

Fields. Each battle record contains:

- `query`: normalized topic prompt, formatted as `Conduct a literature review on {Topic}.`
- `response_a`, `response_b`: draft texts.
- `label_d0`...`label_d4`: {A,B,Tie,BothBad}.
- `field`, `subfield`: taxonomy tags.
- `annotator_id`: pseudonymized identifier.
- `metadata`: optional audit fields (e.g., hashed system ids).

JSON example.

```
{
  "query": "Conduct a literature review on ...",
  "response_a": "...",
  "response_b": "...",
  "label_d0": "A",
  "label_d1": "Tie",
```

```
  "label_d2": "B",
  "label_d3": "A",
  "label_d4": "BothBad",
  "field": "...",
  "subfield": "...",
  "annotator_id": "anon_XXXX",
  "metadata": {"...": "..."}
}
```

A.3. A.3 Aggregation and Scoring Conventions

Bradley–Terry (BT). We fit a BT model per dimension with system-level parameters. Ties are treated as half wins for each side:

$$\texttt{Tie} \Rightarrow 0.5 \text{ win for A } + 0.5 \text{ win for B}.$$

Elo. We compute Elo per dimension with `init`=1500 and $K = 32$. Tie is treated as a 0.5 score for each side (consistent with BT).

Judge–expert alignment (neutral credit). For alignment accuracy, `Tie` and `BothBad` are treated as neutral expert outcomes. If the expert label is `Tie` or `BothBad`, the judge receives 0.5 credit regardless of output `A` or `B`.

A.4. A.4 Meta-evaluation Judge

Judge model. Qwen/Qwen3-235B-A22B-Instruct-2507.

Prompt.

Evaluator Prompt.

```
{"role": "system",
 "content": "You are an expert evaluator for literature reviews. You must provide judgments a

{"role": "user", "content": """
You are an expert evaluator for literature reviews. Your task is to compare two draft literat
(Draft A and Draft B) and make judgments across five dimensions.

Evaluation Dimensions:
D0: Overall Utility | From a researcher's perspective, which response would you prefer to use
for a literature review on this topic?
D1: Literature Coverage | Which response cites a more complete and appropriate set of relevan
with fewer obvious omissions of important work?
D2: Claim Support | Which response more reliably grounds its key claims in the cited literatu
relevant and appropriately support the statements made?
D3: Paper Structure | Which response better structures the existing literature by organizing
categories or comparisons that help a researcher understand relationships among approaches, r
independently?
D4: Research Suggestions Quality | Which response more clearly identifies important and non-o
future directions that go beyond generic summaries and would meaningfully inform a researcher

For each dimension, you must choose one of: A, B, Tie, or BothBad.
- A: Draft A is better
- B: Draft B is better
- Tie: Both drafts are equally good
- BothBad: Neither draft is acceptable
```

A.5. A.5 Calibrated Expert-aligned Evaluator

Group sizes. $k_S = 3$ (structure-similar examples), $k_C = 3$ (content-similar examples), $k_G = 3$ (gap anchors).

Evaluator Prompt.

```
{"role": "system",
 "content": "You are an expert evaluator for literature reviews. You must provide judgments a

{"role": "user", "content": """
You are an expert evaluator for literature reviews. Your task is to compare two draft literat
(Draft A and Draft B) and make judgments across five dimensions.

Evaluation Dimensions:
D0: Overall Utility | From a researcher's perspective, which response would you prefer to use
for a literature review on this topic?
D1: Literature Coverage | Which response cites a more complete and appropriate set of relevan
with fewer obvious omissions of important work?
D2: Claim Support | Which response more reliably grounds its key claims in the cited literatu
relevant and appropriately support the statements made?
D3: Paper Structure | Which response better structures the existing literature by organizing
categories or comparisons that help a researcher understand relationships among approaches, r
independently?
D4: Research Suggestions Quality | Which response more clearly identifies important and non-o
future directions that go beyond generic summaries and would meaningfully inform a researcher

For each dimension, you must choose one of: A, B, Tie, or BothBad.
- A: Draft A is better
- B: Draft B is better
- Tie: Both drafts are equally good
- BothBad: Neither draft is acceptable

=== Structure-Similar Examples (for D3) ===
Example 1:
Query: [STRUCT_EX_QUERY_1]
Draft A:
[STRUCT_EX_A_1]
Draft B:
[STRUCT_EX_B_1]
Expert Outcomes:
  D0: [STRUCT_EX_D0_1]
  D1: [STRUCT_EX_D1_1]
  D2: [STRUCT_EX_D2_1]
  D3: [STRUCT_EX_D3_1]
  D4: [STRUCT_EX_D4_1]
[... up to k_S=3 ...]

=== Content-Similar Examples (for D1/D2) ===
Example 1:
Query: [CONT_EX_QUERY_1]
Draft A:
[CONT_EX_A_1]
Draft B:
[CONT_EX_B_1]
Expert Outcomes:
```

```
  D0: [CONT_EX_D0_1]
  D1: [CONT_EX_D1_1]
  D2: [CONT_EX_D2_1]
  D3: [CONT_EX_D3_1]
  D4: [CONT_EX_D4_1]
[... up to k_C=3 ...]

=== Gap Anchors (for D4) ===
Examples of meaningful research gaps and future directions from expert-written reviews:
1. [GAP_ANCHOR_1]
2. [GAP_ANCHOR_2]
3. [GAP_ANCHOR_3]

=== Current Battle to Evaluate ===
Query: [CUR_QUERY]
Draft A:
[CUR_DRAFT_A]
Draft B:
[CUR_DRAFT_B]

Return a JSON object:
{
  "D0": "A|B|Tie|BothBad",
  "D1": "A|B|Tie|BothBad",
  "D2": "A|B|Tie|BothBad",
  "D3": "A|B|Tie|BothBad",
  "D4": "A|B|Tie|BothBad"
}
Do not output any other keys or values.
"""}
```

Output schema. A single JSON object with keys `D0`–`D4` and values in {A,B,Tie,BothBad}.

A.6. A.6 Annotator Survey (Background) and Summary Statistics

Questionnaire items. Tencent questionnaire items: Q01 Name; Q02 pre-survey completion; Q03 phone number; Q04 NLP; Q05 CV; Q06 ML; Q07 Reasoning & Symbolic AI; Q08 Data Mining & Big Data; Q09 Robotics & Embodied AI; Q10 Multi-Agent & Game Theory; Q11 Interdisciplinary Applications; Q12 other directions; Q13 prior publications (DOIs); Q14 most familiar papers (at least 5 DOIs).

Survey statistics (from responses). Number of respondents: $N = 107$. Completion time (seconds): median $= 510$, IQR $= [293, 990]$, min $= 31$, max $= 107443$.

Top-level area coverage. Count of respondents selecting at least one subtopic: NLP (91), CV (71), ML (74), Reasoning & Symbolic AI (60), Data Mining & Big Data (64), Robotics & Embodied AI (50), Multi-Agent & Game Theory (66), Interdisciplinary Applications (76).

A.7. A.7 Field/Subfield Taxonomy (Full List)

The field/subfield taxonomy is directly instantiated from the annotator background questionnaire (Tencent Survey, Q04–Q11). Each field corresponds to one multi-select question, and each subfield corresponds to a selectable option.

Q04: Natural Language Processing (NLP).

- Large Language Models (LLMs)

- Prompt Engineering and In-Context Learning
- Text Generation and Summarization
- Conversational AI and Dialogue Systems
- Machine Translation and Multilingual NLP
- Information Extraction and Knowledge-related NLP
- NLP Evaluation, Analysis, and Interpretability
- NLP Safety, Bias, and Fact-checking

Q05: Computer Vision (CV).

- Generative Vision Models
- Large Vision Models and Foundation Models
- Vision–Language and Multimodal Learning
- Image and Video Understanding
- 3D Computer Vision
- Medical and Biological Imaging
- Low-level Vision and Computational Photography

Q06: Machine Learning (Methodologies).

- Deep Learning Architectures
- Reinforcement Learning
- Graph Machine Learning
- Trustworthy and Robust Machine Learning
- Optimization and Learning Theory
- Self-supervised and Unsupervised Learning
- Federated and Distributed Learning
- Neuro-Symbolic AI

Q07: Reasoning, Planning, and Symbolic AI.

- Knowledge Representation and Reasoning
- Knowledge Graphs
- Automated Planning and Scheduling
- Search, Optimization, and Constraint Satisfaction
- Causality
- Reasoning under Uncertainty

Q08: Data Mining and Big Data.

- Recommender Systems
- Time-series Analysis
- Anomaly and Outlier Detection
- Web Mining and Social Computing
- Spatio-temporal Data Mining
- Databases and Data Management for AI

Q09: Robotics and Embodied AI.

- Embodied AI
- Robot Learning
- SLAM and Navigation
- Multi-robot Systems
- Human–Robot Interaction (HRI)

Q10: Multi-Agent Systems and Game Theory.

- Multi-agent Coordination and Collaboration
- Game Theory and Economic Paradigms
- Agent-based Modeling and Simulation
- Social Choice and Voting

Q11: Interdisciplinary Applications and Society.

- AI for Science
- Cognitive Modeling and Cognitive Systems
- Human–AI Collaboration and HCI
- AI Ethics, Law, and Governance
- Smart Cities and Transportation
- Financial Technology (FinTech)