SURVEYAUDIT: EVIDENCE-GROUNDED DETECTION OF REVIEWER CONCERNS IN LITERATURE SURVEYS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of large language models has made automated survey generation increasingly practical, yet evaluation of survey quality remains unreliable: existing methods rely on opaque LLM verdicts or heuristic metrics that lack empirical grounding and evidence transparency. We propose SURVEYAUDIT, a verifiable framework focusing on the detection of reviewer concerns in literature surveys. We first introduce CONTAX, an empirically grounded taxonomy of 11 recurring quality concerns derived from 519 real peer-review records on 168 survey papers. SURVEYAUDIT then operationalizes these concerns through category-specific detection procedures and an evidence-tracing protocol: each detected concern must be supported by a verifiable source span, a bounded absence certificate, or a fully specified and reproducible computation path. Experiments on human-written, AI-generated, and perturbation-controlled surveys show that SURVEYAUDIT substantially achieves higher reviewer-aligned concern accuracy and broader coverage of observed reviewer concerns than LLM- and tool-augmented judges on reviewer-relevant concern detection. [1]

1 INTRODUCTION

The rapid growth of scientific literature has made automated survey generation (ASG) an increasingly important direction for AI-for-science systems (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025; Nguye et al., 2025; Wang et al., 2025b). As these systems become increasingly capable, reliable evaluation is needed to determine whether generated surveys provide useful and trustworthy syntheses of the literature. Existing evaluation methods, however, leave two important gaps.

First, evaluation criteria are often designed independently of the concrete concerns raised by reviewers. Metrics such as reference overlap, section-heading recall, and citation accuracy are typically derived from predefined criteria or a single reference survey (Yan et al., 2025; Su et al., 2025; Liang et al., 2025). Such criteria can measure specific properties of a survey, but do not necessarily capture the concrete deficiencies that arise in peer review. A survey may, for example, overlap substantially with a reference survey while still omitting an important development, failing to compare competing approaches, or providing insufficient synthesis across studies.

Second, recent evaluation methods increasingly use LLMs as general-purpose judges, asking them to assess survey quality or freely generate criticisms (Zheng et al., 2023). Although this provides broad coverage, unconstrained LLM judgments can be affected by systematic biases and are difficult to inspect when the resulting judgments are not explicitly tied to supporting evidence. A plausible-sounding criticism does not by itself establish which statement, source, or computation warrants the judgment. This makes it difficult to distinguish a substantive concern from an unsupported evaluation and to reproduce the reasoning underlying an output (Jiang et al., 2026; Li et al., 2025a).

We therefore study a more focused problem: **the detection of concrete reviewer concerns in literature surveys and the verification of the evidence underlying each detected concern**. We deliberately focus on concerns rather than strengths. Existing survey-evaluation methods already provide holistic or dimension-level mechanisms for assessing desirable properties of surveys (Zhang et al., 2026). Our goal is complementary, namely to identify concrete deficiencies that a reviewer may raise and make them individually inspectable.

[1] We will release our codes and datasets upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Method	Evaluation Target	# Criteria	Construct Basis	Reference Survey	Auditability
LLM Judge (Zheng et al., 2023)	Dimension-level assessment	1	Not specified	No	No
AutoSurvey (Wang et al., 2024)	Dimension-level assessment	3	Not specified	No	No
SurveyBench (Yan et al., 2025)	Dimension-level assessment	3	Not specified	Yes	No
SurGE (Su et al., 2025)	Dimension-level assessment	4	Not specified	Yes	No
ARISE (self-feedback) (Wang et al., 2025b)	Dimension-level assessment	20	Reviewer guidelines	No	No
Bloom-Eval (Zhang et al., 2026)	Dimension-level assessment	18	Bloom's Taxonomy	Yes	Partial
SURVEYAUDIT (Ours)	**Concern Detection**	11	Peer Reviews	No	Yes

Table 1: Comparison of prior frameworks.

We define a *reviewer concern* as a substantive statement in a peer review that identifies a perceived deficiency, omission, inconsistency, or inadequacy in a survey. Reviewer concerns provide an empirical basis for determining which types of problems matter in practice, but are not themselves treated as ground-truth statements of fact. Based on public peer reviews of survey papers, we construct **Con**cern **Tax**onomy (CONTAX), an empirically grounded taxonomy of reviewer concerns organized by the conditions under which each concern can be verified. CONTAX identifies 11 operational concern categories and characterizes the verification basis required to establish each category.

We then introduce SURVEYAUDIT, an evidence-grounded framework for detecting concerns instantiated by CONTAX. Its central design principle is to constrain the role of LLMs in the evaluation process. Rather than asking an LLM to judge the quality of an entire survey or freely generate weaknesses, SURVEYAUDIT uses LLMs only for bounded semantic or structural extraction and for local judgments grounded in explicitly provided evidence. Every LLM-generated judgment is required to return supporting evidence copied verbatim from its input or verification source, and the system programmatically checks this correspondence. Explicit procedures are used whenever the operation can be specified reproducibly, including structural matching, citation resolution, literature retrieval, graph-based ranking, and coverage analysis. This division of labor allows LLMs to provide flexible language understanding while keeping the verification of detected concerns explicit and inspectable.

Our contributions are: (1) **CONTAX,** an empirically grounded taxonomy of 11 reviewer-concern categories derived from public peer reviews of literature surveys and organized by their verification requirements. (2) **SURVEYAUDIT,** an evidence-grounded framework for detecting concrete reviewer concerns through category-specific detectors and constrained LLM use, with explicit verification procedures and inspectable evidence. (3) A systematic empirical evaluation of **SURVEYAUDIT** on human-written surveys and controlled perturbations, together with diagnostic analyses of AI-generated surveys and the behavior of individual design components.

2 RELATED WORK

Automated survey generation. Recent systems automate survey construction through retrieval, outline generation, and iterative refinement (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025; Nguye et al., 2025; Wang et al., 2025b). These approaches employ different organizational principles, including attribute trees, citation-graph hierarchies, and rubric-guided self-revision. However, their evaluation components primarily serve generation or refinement rather than independent auditing of completed surveys. They typically rely on LLM-based or human-preference scores whose decision processes are not exposed for verification. Our work instead focuses on the downstream auditing of finished surveys.

Survey evaluation. Survey evaluation has recently emerged as an independent research problem. SurGE evaluates computer-science survey generation along information coverage, referencing accuracy, structural organization, and content quality (Su et al., 2025), while SurveyGen provides a large-scale human-written dataset for quality-aware generation (Bao et al., 2025). More recent approaches include quiz-based evaluation in SurveyBench (Sun et al., 2025) and taxonomy-grounded diagnosis in Bloom-Eval (Zhang et al., 2026), which evaluates ASG systems across six cognitive tiers. Despite their broader coverage, these frameworks often rely on unaudited LLM judgments for open-ended evaluation, inheriting known limitations of LLM-as-a-Judge (Li et al., 2025a). In contrast, SURVEYAUDIT exposes the evidence and decision path underlying each output and reports category-specific results without collapsing them into a holistic score.

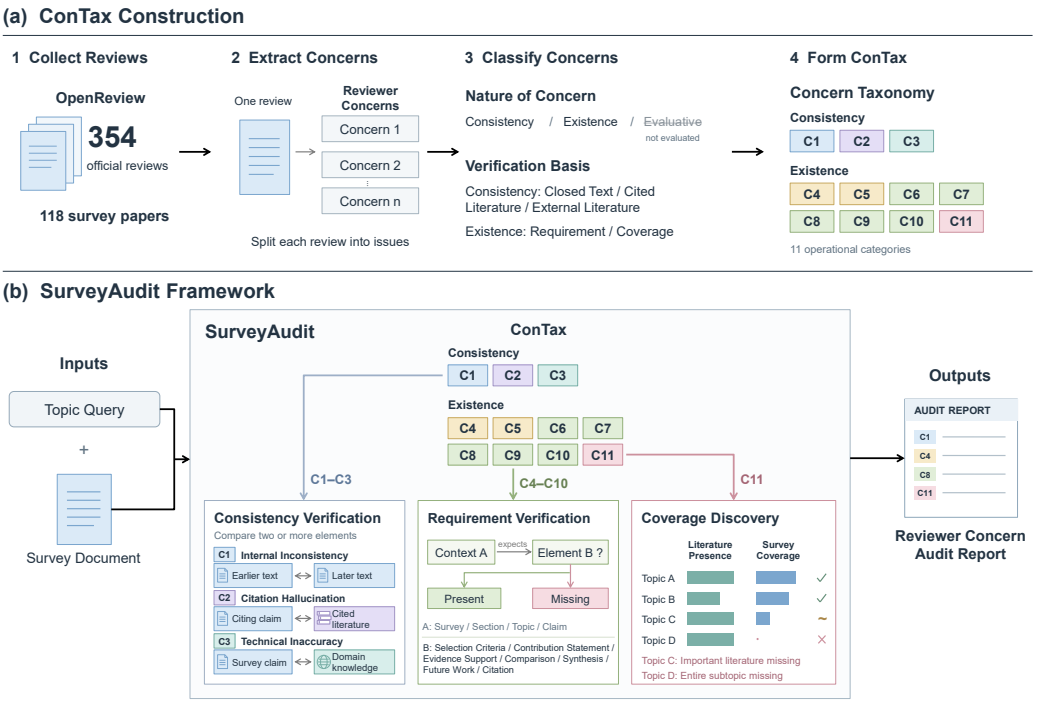


Figure 1: Overview of CONTAX construction and the SURVEYAUDIT framework.

Evidence-linked evaluation. Recent work has increasingly emphasized evidence grounding and verifiability in scientific evaluation (Ou et al., 2025). Meng et al. (2026) proposes Chain-of-Evidence, which requires claims to remain traceable to their evidence sources and instantiates this principle in ScientistOne for literature review and scientific writing. In peer review, DeepReviewer 2.0 (Weng et al., 2026) introduces process-controlled critique generation with traceability links, while ProReviewer (Fang et al., 2026) maintains structured review logs that record evidence and intermediate findings during paper investigation. Collectively, these approaches shift scientific evaluation from fluent critique toward evidence-anchored and auditable assessment. SURVEYAUDIT adopts this principle for concern detection and systematizes evidence anchoring through the CONTAX.

3 CONTAX

We illustrate the process of CONTAX construction and SURVEYAUDIT evaluation in Figure 1. CONTAX is designed to provide an explicit and operational space of reviewer-relevant concerns for concern detection. We construct the taxonomy from observed reviewer concerns and organize them by the conditions under which each concern can be verified. We restrict the taxonomy to concerns for which an objective or reproducible verification criterion can be defined.

3.1 DATA COLLECTION AND ANNOTATION

We collect 519 public official reviews associated with 168 survey papers from venues including TMLR, ICLR, NeurIPS, EMNLP, and COLM. At the paper level, we form a development set of 118 papers and 354 reviews and a held-out test set of 50 papers and 165 reviews, preserving the venue distribution and preventing any paper or review from appearing in both sets. The development split supports taxonomy construction and detector design; the held-out split is reserved for human-written survey evaluation. Collection and splitting details appear in Appendix B.

For the taxonomy, we extract reviewer-stated concerns from the 354 development reviews. Multiple annotators independently assign each instance to its evidence-structure signature, as detailed in Appendix D.2. Disagreements are reconciled through joint discussion. Because survey submissions in the corpus are heavily concentrated in TMLR, the taxonomy should be interpreted as reflecting the

review ecology of available OpenReview survey papers rather than the full population of all survey submissions.

3.2 TAXONOMY CONSTRUCTION VIA EVIDENCE-STRUCTURE ANNOTATION

We construct CONTAX by first classifying reviewer criticisms according to the *nature of the concern*. We then characterize consistency and existence concerns along different structural dimensions that reflect how each type of concern can be audited. This structure links reviewer concerns to the conditions under which they can be systematically verified.

Nature of the Concern. At the first level, we classify each concern according to the type of problem explicitly identified by the reviewer. We distinguish three fundamental forms of concern: (1) **Consistency.** The reviewer identifies an inconsistency, contradiction, or error among elements of the survey or external sources. (2) **Existence.** The reviewer identifies an element that is missing or absent from the survey. (3) **Evaluation.** The reviewer acknowledges that an element is present but considers its quality, adequacy, or degree insufficient.

Verification Basis. The second dimension characterizes what evidence or structure determines whether a concern can be verified. For consistency concerns, we distinguish three evidence sources. *Closed-text* evidence is contained within the survey itself, *cited-literature* evidence is contained in a work cited by the survey, and *external-literature* evidence is contained in an external work not cited by the survey. We also identify *external norms*, such as formatting requirements, venue conventions, and general expectations about scholarly writing, but exclude them because their verification requires open-ended normative judgment.

For existence concerns, we distinguish between *requirements* and *coverage*. A requirement concern arises when a specific element is expected in a given survey context but is absent. A coverage concern arises when relevant literature must first be identified to determine what content the survey should represent.

3.2.1 CONCERN CATEGORY

The preceding dimensions provide a general structure for organizing reviewer concerns, but the annotated concerns exhibit finer-grained distinctions within each dimension. We therefore refine these distinctions into operational categories. For consistency concerns, the three evidence bases correspond to distinct recurring types of concern because they involve different relationships between the survey and its supporting literature. For existence concerns, the annotated requirements differ according to the type of element whose absence is identified, yielding several requirement categories, while coverage concerns form a separate category because the missing content must first be identified from the surrounding literature. We retain a distinction as an operational category only when the corresponding concerns admit an explicit and reproducible verification criterion. Concern classes that require open-ended normative or evaluative judgment are therefore excluded. The resulting taxonomy is summarized in Table 2.

4 SURVEYAUDIT FRAMEWORK

As illustrated in Figure 1 (b), given a topic query T and a survey document D, SURVEYAUDIT detects concrete concerns instantiated by CONTAX and produces an evidence-grounded audit report. We use the term *reviewer concern* for a substantive statement in a peer review that identifies a perceived deficiency, omission, inconsistency, or inadequacy in a survey. A *detected concern* is an operationalized instance of such a concern that can be established under one of the CONTAX categories and its corresponding verification criterion. Reviewer concerns therefore ground the target space of the system, whereas detected concerns are the concrete outputs produced by the detectors.

The framework separates language understanding from explicit verification. LLMs are used only for bounded semantic or structural extraction and for local judgments grounded in provided evidence. Whenever a judgment depends on textual or literature evidence, the model must return the supporting evidence verbatim, which is subsequently checked against the corresponding source. Operations that admit explicit and reproducible procedures, such as structural matching, citation resolution, literature

ID	Concern Category	Nature	What it checks
C1	Internal Inconsistency	Consistency; Closed Text	Whether statements within the survey make mutually incompatible claims about the same object or condition.
C2	Citation Hallucination	Consistency; Cited Literature	Whether a cited source supports the claim attributed to it by the survey.
C3	Technical Inaccuracy	Consistency; External Literature	Whether a substantive technical claim is contradicted by relevant scholarly evidence.
C4	Missing Survey Scope or Selection	Existence; Requirement	Whether the survey explicitly defines its literature scope and, where applicable, the criteria or procedures used to identify and select the reviewed literature.
C5	Missing Contribution	Existence; Requirement	Whether the survey explicitly distinguishes its contribution or positioning from relevant prior surveys.
C6	Missing Evidence	Existence; Requirement	Whether a substantive claim that requires support is accompanied by adequate evidence within the survey.
C7	Missing Comparison	Existence; Requirement	Whether related research objects that warrant comparison are explicitly compared within the survey.
C8	Missing Cross-Study Synthesis	Existence; Requirement	Whether related studies or approaches are synthesized into higher-level relationships, trends, or abstractions.
C9	Missing Future Work	Existence; Requirement	Whether an important unresolved research direction is addressed as future work or an open problem.
C10	Missing Citation	Existence; Requirement	Whether a substantive research object or relevant prior work is attributed to an appropriate source.
C11	Missing Literature Coverage	Existence; Coverage	Whether an important and relevant literature item or research area is insufficiently represented in the survey.
Excluded	External Norms	Consistency	Concerns whose verification depends on external norms, conventions, or venue-specific expectations.
Excluded	Evaluative Concerns	Evaluative	Concerns that require open-ended judgments of quality or adequacy rather than an explicit verification criterion.

Table 2: Operational concern categories derived from reviewer concerns.

retrieval, graph-based ranking, and coverage computation, are handled algorithmically. The detectors then follow a common Extract–Check–Verify pattern, with an additional importance-based selection stage for categories that generate many potential omission obligations.

4.1 Evidence-Constrained LLM Use

SurveyAudit restricts LLM use to bounded semantic or structural extraction and local judgments grounded in explicitly provided evidence. An LLM is not asked to assess the quality of the survey as a whole or freely generate criticisms. Whenever an LLM judgment is used to establish a concern, the model must return the supporting evidence verbatim from the provided input or verification source, and the system checks that the returned span is contained in that source. Operations that admit explicit and reproducible procedures, such as structural matching, citation resolution, retrieval, graph construction, and ranking, are handled algorithmically. Thus, LLMs provide flexible language understanding within category-specific verification procedures rather than serving as unconstrained end-to-end judges.

4.2 Concern Representation

We distinguish reviewer concerns from detected concerns. A *reviewer concern* is an empirically observed statement in a peer review describing a perceived deficiency, omission, inconsistency, or inadequacy, whereas a *detected concern* is a concrete, operationalized proposition that satisfies the verification criterion of a ConTax category. We represent a detected concern as $z = (c, p, E)$, where c is the concern category, p is the detected proposition, and E is the supporting evidence or verification trace. The evidence may consist of source spans from the survey, retrieved or cited literature, or a reproducible computational procedure, depending on the category. All detectors therefore produce the same type of output, differing only in how concerns are generated and verified.

4.3 DETECTOR OPERATIONALIZATION

SURVEYAUDIT implements the 11 CONTAX categories through three recurring detection patterns: consistency verification, requirement verification, and coverage discovery. These patterns differ in how the information needed for verification is obtained. Consistency verification compares multiple elements whose relationship can be explicitly examined. Requirement verification first identifies an expected element from a given context and then checks whether it is present. Coverage discovery addresses cases where the potentially missing elements cannot be enumerated from the survey itself and must therefore be discovered from the surrounding literature. Across all three patterns, LLMs are used only for bounded semantic extraction or local evidence-grounded judgment, while explicit matching, retrieval, ranking, and structural checks are handled algorithmically. Detailed prompts and implementation configurations are provided in Appendix C.

Consistency Verification (C1–C3). Consistency Verification detect concerns by comparing two or more semantically related elements under different evidence bases. For C1, an LLM extracts structural commitments and atomic semantic statements from the survey. Statements referring to the same underlying object are then grouped and compared algorithmically, and a concern is reported only when two explicit statements are directly irreconcilable. Compatible refinements, different experimental settings, or statements that can be jointly satisfied are not treated as inconsistencies. For C2, the detector extracts citation-bearing claims, resolves their cited references, and retrieves the corresponding literature. An LLM then judges whether each claim is supported by the supplied source evidence. For C3, the same claim–evidence verification principle is applied to substantive technical claims, with candidate evidence retrieved from both cited and external scholarly literature. Background, summary, contribution, and conclusion statements are excluded from this verification. In C2 and C3, a concern is reported only when the inspected evidence establishes the relevant support, lack-of-support, or contradiction relation, rather than merely because supporting evidence could not be retrieved.

Requirement Verification (C4–C10). C4–C10 share a common requirement-based structure. Given a context A, the detector identifies an expected element B whose presence is relevant to that context, and then verifies whether B is explicitly represented in the survey. An LLM extracts the relevant contexts and existing elements, while category-specific rules define the corresponding requirements. For C4, the context is the survey as a whole and the expected element is an explicit scope or selection boundary. For C5, the context is the survey's positioning relative to relevant prior surveys and the expected element is a distinguishing contribution or positioning statement. For C6–C10, the contexts are local claims, research objects, topics, or literature relations, and the expected elements are respectively evidential support, comparison, cross-study synthesis, future-work treatment, and citation attribution.

For requirements that are not directly determined by an explicit structural relation, the detector constructs candidate obligations from the extracted structural and topical information. Each obligation is assigned an importance score and only high-priority obligations are retained under a bounded selection policy. This prevents the detector from treating every theoretically possible omission as a concern. The selected obligations are then checked against the survey using the corresponding category criterion. A concern is reported only when the expected element is sufficiently important and is verified as absent, with the inspected survey context retained as evidence or a bounded absence certificate.

Coverage Discovery (C11). C11 addresses cases in which the expected content cannot be enumerated from the survey itself. The detector first constructs a literature pool from cited-paper neighborhoods and scholarly retrieval and represents the resulting literature as a graph. Deterministic graph procedures identify coherent literature communities and rank them according to their structural importance and representation in the survey. For high-priority candidates, an LLM examines representative literature together with the corresponding survey content to determine whether the literature is substantively covered, represents a coherent direction relevant to the survey, or is unrelated. Relevant and important literature that remains insufficiently represented is reported as a coverage concern, with representative literature and corresponding survey evidence retained for inspection.

Although C11 requires an additional discovery stage, its final verification follows the same requirement-based logic as C4–C10: a candidate literature element is treated as an expected element only after its relevance and importance have been established. The three patterns therefore

Method	DCP%↑	RCC%↑	ACpS
Conventional LLM Judge	52.03	20.76	4.62
Tool-augmented LLM Judge	56.46	16.23	4.20
SURVEYAUDIT (Ours)	**86.55**	**34.50**	**9.84**

Table 3: Evaluation on human-written surveys.

Method	P1	P2	P3	P4	P5	Overall
Conventional LLM Judge	37.0	0.0	2.0	34.0	8.0	16.2
Tool-augmented LLM Judge	46.0	57.0	29.0	44.0	30.0	41.2
SURVEYAUDIT	**73.0**	**80.0**	**31.0**	**61.0**	**89.0**	**66.8**

Table 4: Controlled perturbation evaluation. Each cell reports Perturbation Recall.

differ primarily in how their verification targets are obtained—from explicit relations, context-derived requirements, or literature-discovered candidates—while sharing the same evidence-constrained reporting principle.

5 EXPERIMENTS AND ANALYSES

We evaluate SURVEYAUDIT as a concern detection system from four perspectives: alignment with human reviewer concerns, sensitivity to controlled defects, analysis of AI-generated surveys, and the contribution of individual design components.

5.1 EXPERIMENTAL SETUP

Human-written surveys. We construct a held-out test set of 50 human-written survey papers from the corpus introduced in § 3. When partitioning the corpus into training and test sets, we aim to maintain diverse and balanced distributions of research topics and publication venues in the test set. For each test survey, publicly available peer reviews are retained as observed reviewer concerns where available. These reviews provide the basis for evaluating whether the concerns identified by an evaluator are consistent with concerns raised by human reviewers. The taxonomy and detector design are fixed before evaluation. All systems are evaluated on the same survey documents under the same annotation protocol. Human judgments are collected independently and disagreements are resolved through adjudication.

Baselines. We compare SURVEYAUDIT with two LLM-based evaluation paradigms. The first is a conventional LLM judge without external tools. The second is a tool-augmented judge implemented with Claude Code, which has access to scholarly retrieval and web-search tools. For both paradigms, we adopt a general survey-evaluation prompt requiring agent to generate a list of reviewer concerns. We use the same model Deepseek-V4-Flash-Preview with greedy sampling for all baselines and SURVEYAUDIT. The baseline systems do not have access to CONTAX or the category-specific detection procedures of SURVEYAUDIT.

5.2 REVIEWER ALIGNMENT ON HUMAN-WRITTEN SURVEYS

We first evaluate whether the concerns identified by an evaluator correspond to concerns that human reviewers consider legitimate and important. We use three complementary measures. *Detected Concern Precision* (DCP) measures the proportion of detected concerns that annotators consider correct and important enough to appear in an official review. *Reviewer Concern Coverage* (RCC) measures the proportion of concerns raised in the corresponding peer reviews that are recovered by at least one accepted concern. The matching is based on the underlying reviewer concern rather than its CONTAX category, since the same concern can be expressed at different levels of specificity. For example, a reviewer's observation that cross-paper comparisons are insufficient can be covered by a SURVEYAUDIT detected concern identifying missing comparisons for particular subtopics. Finally,

Generation System	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
SurveyG (Nguye et al., 2025)	2.1	22.7	**0.1**	**0.5**	0.1	**0.1**	**1.1**	**0.3**	**0.0**	8.1	3.0
SurveyGen-I (Chen et al., 2025)	1.7	16.8	0.3	0.8	0.1	0.2	6.4	4.7	4.9	**3.6**	**0.6**
Claude Code	**1.6**	**12.3**	**0.1**	0.7	0.1	**0.1**	3.9	2.1	3.6	6.6	5.8

Table 5: Concern profiles of AI-generated surveys. Each cell reports the number of detected concerns identified by SURVEYAUDIT for the corresponding CONTAX concern category. C1–C11 denote the automatically detectable concern categories defined in § 3. **Counts are within-category diagnostic statistics and are not intended for cross-category comparison.**

Variant	DCP%↑	RCC%↑	ACpS
SURVEYAUDIT	86.55	34.50	9.84
SURVEYAUDIT – Evidence Verification	79.51	32.31	8.72
SURVEYAUDIT – Evidence Span	70.73	27.19	6.66
SURVEYAUDIT – Algorithm	86.40	23.54	5.36
SURVEYAUDIT – CONTAX	88.98	28.36	7.40

Table 6: Ablation results on human-written surveys. The ablations isolate the effects of evidence verification, explicit evidence spans, algorithmic detector decomposition, and the CONTAX taxonomy.

Accepted Concerns per Survey (ACpS) counts the total number of accepted concerns and calculate the average on all surveys in the test set.

Table 3 summarizes the results. SURVEYAUDIT consistently outperforms the general-prompt baselines on concern detection, producing a broader set of reviewer-acceptable concerns while recovering substantially more of the concerns documented by human reviewers. This advantage is particularly notable for RCC, which is matched at the level of the underlying reviewer concern rather than the CONTAX category or surface form. Consequently, SURVEYAUDIT can recover concerns that are evaluative in their original formulation by expressing them as more concrete, verifiable concerns. For example, a reviewer may state that a survey provides insufficient comparison across methods, while SURVEYAUDIT can recover the same concern by identifying specific missing comparisons as a verifiable omission. Thus, although CONTAX restricts the concerns that SURVEYAUDIT can establish automatically to those with explicit or reproducible verification criteria, the reviewer concern being recovered need not itself be phrased in such an operational form.

5.3 CONTROLLED PERTURBATION EVALUATION

We next evaluate whether SURVEYAUDIT can recover known defects under controlled conditions. We select 20 human-written surveys from the labeled high-quality survey dataset from SurGE (Su et al., 2025) and construct five independent perturbation cases for each of five concern types, resulting in 500 perturbed surveys. Each variant contains exactly one perturbation. The perturbations instantiate representative concern types covered by the CONTAX taxonomy, while retaining unambiguous ground truth for evaluation. The five groups cover internal inconsistency (P1), citation hallucination (P2), technical inaccuracy (P3), missing content (P4), and citation omission (P5). The P4 group includes comparison, synthesis, and future-work omissions, where a coherent block of content is removed to create a substantive absence. More implementation details are in Appendix D.3. We report *Perturbation Recall*, defined as the proportion of perturbations that are recovered.

Table 4 reports the results by perturbation type. Overall, SURVEYAUDIT consistently shows stronger sensitivity to known defects across the five perturbation families, including concerns that require global reasoning or external evidence. Among all perturbations, SURVEYAUDIT performs best on P5, which most directly matches its citation-level verification procedure: removing a citation exposes an otherwise unsupported method or claim, with both the target and the evidence source explicitly specified. P2 follows closely, as citation hallucinations provide fixed candidate sentences and corresponding cited sources, making the verification target comparatively well defined. P1 requires broader global reasoning: the system must identify potential inconsistencies after reading the paper and verify them against extracted contribution statements and other evidence. P4 requires candidate generation and importance assessment; its relatively high recall is partly attributable to

the substantial nature of the perturbations, which remove complete functional elements such as future-work discussions. P3 is the most challenging setting, requiring cross-document retrieval and reasoning to verify technical inaccuracies.

5.4 Diagnostic Evaluation of AI-Generated Surveys

We further apply SurveyAudit to surveys generated by several contemporary survey-generation systems. The surveys are generated on a common set of 30 computer-science topics and then evaluated under the same SurveyAudit configuration. We report the topic details in Appendix D.4. We generate literature surveys on those topics with 3 distinct agents: two agents specially designed for literature survey generation SurveyG (Nguye et al., 2025) and SurveyGen-I (Chen et al., 2025), and a commonly-used coding agent Claude Code. Coding agents are included because they are increasingly used for research assistance and have been shown to exhibit automatic research capabilities (Rank et al., 2026).

Table 5 reports the concern profiles of the generated surveys. C2 is substantially more frequent than most other categories, but its raw count should not be interpreted as a cross-category quality measure because the category operates at the claim/citation level. Deeper analyses in Table 11 demonstrates that the three systems also exhibit different citation concern patterns: SurveyG frequently produces citation keys that are absent from its bibliography, SurveyGen-I more often exhibits mismatches between cited papers and the claims they support, while Claude Code includes cases of fabricated references. SurveyGen-I exhibits a different profile overall: its longer surveys and larger bibliographies are accompanied by more local omission errors (C6–C9), while its citation and coverage concern counts (C10, C11) are lower. These results provide distinct concern profiles for the three generation systems under the SurveyAudit detector, illustrating how the framework can be used for diagnostic analysis of generated surveys.

5.5 Ablation Study

We conduct ablation experiments on the human-written test set to examine the contribution of the major design components of SurveyAudit. We consider 4 variants: **–Evidence Verification**, which removes downstream evidence verification only from SurveyAudit; **–Evidence Span**, which removes the requirement to provide verbatim evidence spans; **–Algorithm**, which removes the category-specific detector design while retaining ConTax and evidence requirements; and **–ConTax**, which removes the ConTax taxonomy while retaining the evidence requirements. We use DCP, RCC and ACpS to evaluate the resulting detected concern sets.

The ablation results in Table 6 show that the structured detector design is important for RCC and concern yield. Removing the category-specific detector design leaves concern accuracy nearly unchanged, but substantially reduces RCC and the number of accepted concerns. Removing evidence verification or explicit evidence spans also weakens DCP, supporting the role of evidence in the concern detection process. The **–ConTax** variant exhibits a distinct trade-off: it achieves higher DCP than the full system, while producing fewer accepted concerns and recovering fewer reviewer concerns. This indicates that ConTax constrains the system toward a smaller set of more consistently acceptable concerns, at the cost of broader concern coverage.

6 Conclusion

We presented SurveyAudit, an evidence-grounded framework for detecting concrete reviewer concerns in literature surveys. We introduced ConTax, an empirically grounded taxonomy of 11 operational concern categories, and developed category-specific detectors that constrain LLM use to bounded semantic extraction and evidence-grounded judgment. Each detected concern is accompanied by supporting evidence or a reproducible verification trace. Experiments on human-written surveys and controlled perturbations demonstrate reliable detection of concrete survey deficiencies, while diagnostic analyses illustrate the framework's use on AI-generated surveys. More broadly, the results suggest that survey evaluation can complement holistic quality assessment with systematic detection and verification of individually inspectable concerns.

AI DISCLOSURE

In this work, we have not used generative AI tools for any of the required-disclosure tasks. These include generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, and interpreting results. Additionally, we used generative AI tools for editing the research paper to improve readability, clarity, and language quality. We reviewed all AI-assisted text. All research decisions, experimental procedures, implementations, data processing, and result interpretations were conducted by the authors. We take responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

REFERENCES

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. SurveyGen: Quality-aware scientific survey generation with large language models. *arXiv preprint arXiv:2508.17647*, 2025.

Tong Bao, Mir Tafseer Nayeem, Yi Zhao, Davood Rafiei, and Chengzhi Zhang. Surveyagent-hka: A multi-agent framework for scientific survey generation with llms and human knowledge augmentation, 2026. URL https://arxiv.org/abs/2609.05938.

Jennifer A Byrne. Improving the peer review of narrative literature reviews. *Research integrity and peer review*, 1:1, 2016.

Yu Chao, Siyu Lin, Xiaorong Wang, Zhu Zhang, Zihan Zhou, Haoyu Wang, Shuo Wang, Jie Zhou, Zhiyuan Liu, and Maosong Sun. Llm×mapreduce-v3: Enabling interactive in-depth survey generation through a mcp-driven hierarchically modular agent system. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 187–197. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-demos.51. URL https://aclanthology.org/2025.emnlp-demos.51/.

Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57 (3):359–377, 2006.

Jing Chen, Zhiheng Yang, Yixian Shen, Jie Liu, Adam Belloum, Paola Grosso, and Chrysa Papagianni. Surveygen-i: Consistent scientific survey generation with evolving plans and memory-guided writing. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 3687–3714, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and the Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.193. URL https://aclanthology.org/2025.ijcnlp-long.193/.

Harris M. Cooper. Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52(2):291–302, 1982. doi: 10.3102/00346543052002291.

Haishuo Fang, Yue Feng, and Iryna Gurevych. From passive generation to investigation: A proactive scientific peer review agent. *arXiv preprint arXiv:2606.13349*, 2026.

Beichen Guo, Zhiyuan Wen, Jia Gu, Haochen Shi, Jian Wang, Senzhang Wang, Haoyang Li, Ruosong Yang, and Shuaiqi Liu. Surveylens: A discipline-aware benchmark for automatic survey generation. *arXiv preprint arXiv:2602.11238*, 2026. doi: 10.48550/arXiv.2602.11238. URL https://arxiv.org/abs/2602.11238.

Fengqing Jiang, Yichen Feng, Yuetai Li, Luyao Niu, Basel Alomair, and Radha Poovendran. Badscientist: Can a research agent write convincing but unsound papers that fool llm reviewers? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24712–24727, 2026.

Thanh Gia Hieu Khuong and Benedictus Kent Rachmat. Auto-survey challenge. *arXiv preprint arXiv:2310.04480*, 2023. doi: 10.48550/arXiv.2310.04480. URL https://arxiv.org/abs/2310.04480.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, Suzhou, China, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.138.

Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. Reportbench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*, 2025b. doi: 10.48550/arXiv.2508.15804. URL https://arxiv.org/abs/2508.15804.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, and Keming Mao. SurveyX: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

Qinyue Liu, Yongxin Zhou, and Cyril Labbe. Sciciteval: A multi-domain dataset for scientific citation verification. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 1603–1611. ELRA Language Resource Association, 2026. doi: 10.63317/4m84m2k77g97. URL https://aclanthology.org/2026.lrec-1.125/.

Joanne E. McKenzie and Sue E. Brennan. Chapter 12: Synthesizing and presenting findings using other methods. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch (eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, Chichester, UK, version 6.5 edition, 2024. URL https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-12. Chapter last updated October 2019.

Rui Meng, Bhavana Dalvi Mishra, Jiefeng Chen, Chun-Liang Li, Palash Goyal, Mihir Parmar, Yiwen Song, Yale Song, Rajarishi Sinha, Parthasarathy Ranganathan, Burak Gokturk, Jinsung Yoon, and Tomas Pfister. Scientistone: Towards human-level autonomous research via chain-of-evidence. *arXiv preprint arXiv:2605.26340*, 2026.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023.

Minh-Anh Nguye, Minh-Duc Nguyen, Nguyen Thi Ha Lan, Kieu Hai Dang, Nguyen Tien Dong, and Le Duy Dung. SurveyG: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *arXiv preprint arXiv:2510.07733*, 2025.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, et al. Claimcheck: How grounded are llm critiques of scientific papers? In *EMNLP (Findings)*, pp. 21712–21735, 2025.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

Jennie Popay, Helen Roberts, Amanda Sowden, Mark Petticrew, Lisa Arai, Mark Rodgers, Nicky Britten, Katrina Roen, and Steven Duffy. Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC methods programme. Esrc methods programme report, Lancaster University, Lancaster, UK, 2006. URL https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf.

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can llm agents automate llm post-training? *arXiv preprint arXiv:2603.08640*, 2026.

Karen A. Robinson, Ian J. Saldanha, and N. A. Mckoy. Frameworks for determining research gaps during systematic reviews. Methods Future Research Needs Reports 2, Agency for Healthcare Research and Quality (US), Rockville, MD, Jun 2011. AHRQ Publication No. 11-EHC043-EF.

Hannah Snyder. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339, 2019. doi: 10.1016/j.jbusres.2019.07.039. URL https://doi.org/10.1016/j.jbusres.2019.07.039.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. Benchmarking computer science survey generation. *arXiv preprint arXiv:2508.15658*, 2025.

Zhaojun Sun et al. Surveybench: Can llm(-agents) write academic surveys that align with reader needs? *arXiv preprint*, arXiv:2510.03120, 2025. doi: 10.48550/arXiv.2510.03120. URL https://arxiv.org/abs/2510.03120.

Neset Tan, Minghao Li, and Mark Gahegan. Scitrue: Evidence-grounded claim verification in science. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 397–406. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-demo.27. URL https://aclanthology.org/2026.eacl-demo.27/.

Richard J. Torraco. Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, 4(3):356–367, 2005.

Richard J. Torraco. Writing integrative literature reviews: Using the past and present to explore the future. *Human Resource Development Review*, 15(4):404–428, 2016. doi: 10.1177/1534484316671606. URL https://doi.org/10.1177/1534484316671606.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, England, 1958. doi: 10.1017/CBO9780511840005.

Nees Jan van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010. doi: 10.1007/s11192-009-0146-3. URL https://doi.org/10.1007/s11192-009-0146-3.

Nees Jan van Eck and Ludo Waltman. Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4):802–823, 2014. doi: 10.1016/j.joi.2014.07.006. URL https://doi.org/10.1016/j.joi.2014.07.006.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

Haoyu Wang, Yujia Fu, Zhu Zhang, Shuo Wang, Zirui Ren, Xiaorong Wang, Zhili Li, Chaoqun He, Bo An, Zhiyuan Liu, and Maosong Sun. Llm×mapreduce-v2: Entropy-driven convolutional test-time scaling for generating long-form articles from extremely long resources. *arXiv preprint arXiv:2504.05732*, 2025a. doi: 10.48550/arXiv.2504.05732. URL https://arxiv.org/abs/2504.05732.

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. AutoSurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*, 2024.

Zi Wang, Xingqiao Wang, Sangah Lee, and Xiaowei Xu. ARISE: Agentic rubric-guided iterative survey engine for automated scholarly paper generation. *arXiv preprint arXiv:2511.17689*, 2025b.

Jane Webster and Richard T. Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii, 2002.

Yixuan Weng, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, et al. Deepreviewer 2.0: A traceable agentic system for auditable scientific peer review. *arXiv preprint arXiv:2604.09590*, 2026.

Siyi Wu, Chiaxin Liang, Ziqian Bi, Leyi Zhao, Tianyang Wang, Junhao Song, Yichao Zhang, Keyu Chen, Benji Peng, and Xinyuan Song. Autosurvey2: Empowering researchers with next level automated literature surveys. *arXiv preprint arXiv:2510.26012*, 2025. doi: 10.48550/arXiv.2510.26012. URL https://arxiv.org/abs/2510.26012.

Tianyi Xu, Zhe Zhao, Tianshuo Wei, Yiqun Kou, Liuliu Han, and Ye Wei. Feedback is the key for automated survey generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38174–38195, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1904. URL https://aclanthology.org/2026.findings-acl.1904/.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. SurveyForge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025.

Fei Zhang, Zhe Zhao, HaiBin Wen, Tianshuo Wei, Zaixi Zhang, Chao Yang, and Ye Wei. Bloomeval: A hierarchical evaluation benchmark for automatic survey generation based on bloom's taxonomy. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28512–28544, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1315.

Hongbo Zhang, Han Cui, Yidong Wang, Yijian Tian, Qi Guo, Cunxiang Wang, Jian Wu, Chiyu Song, and Yue Zhang. Deep literature survey automation with an iterative workflow. *arXiv preprint arXiv:2510.21900*, 2025. doi: 10.48550/arXiv.2510.21900. URL https://arxiv.org/abs/2510.21900.

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. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A More Related Works

A.1 Automated Survey Generation and Evaluation

Automated survey generation has developed along several complementary directions. Early end-to-end systems such as AutoSurvey (Wang et al., 2024) and SurveyX (Liang et al., 2025) decompose survey writing into retrieval, organization, and section-level generation, while SurveyForge (Yan et al., 2025) uses human-written outline structures and memory-driven scholarly retrieval to improve organization and source selection. SurveyGen-I (Chen et al., 2025) further introduces evolving plans and memory-guided writing to maintain coherence across long multi-section surveys. Other systems address the long-context bottleneck through hierarchical processing: LLM×MapReduce-V2 (Wang et al., 2025a) uses test-time scaling to integrate information from extremely long resources and underlies the SurveyGO system, while LLM×MapReduce-V3 (Chao et al., 2025) extends this paradigm to a modular multi-agent architecture. AutoSurvey2 (Wu et al., 2025) further combines retrieval-augmented synthesis, parallel section generation, iterative refinement, and automated evaluation in a unified pipeline. Together, these systems increasingly treat survey generation as a multi-stage research workflow rather than a single generation call.

A second line incorporates feedback or iterative review into the generation process. ARISE (Wang et al., 2025b) introduces a rubric-guided iterative refinement loop in which reviewer agents assess

intermediate survey drafts and their feedback is synthesized for subsequent revision. SurveyG (Nguye et al., 2025) combines multi-agent generation with an evaluation stage for checking consistency, coverage, and factual accuracy. FIKSurvey (Xu et al., 2026) explicitly organizes feedback around outline, citation, and content quality, while IterSurvey (Zhang et al., 2025) repeatedly retrieves, revises its outline, and applies a review-and-refine loop during generation. SurveyAgent-HKA (Bao et al., 2026) incorporates knowledge from published surveys and peer-review comments to guide subsequent revisions. These systems increasingly expose evaluation or review signals inside the generation loop, with the goal of improving the resulting survey.

A third line develops benchmarks and evaluation protocols for comparing generated surveys. The Auto-survey Challenge (Khuong & Rachmat, 2023) established an early shared evaluation setting in which systems were asked both to generate and critique survey papers under simulated peer review. More recent benchmarks extend evaluation to broader research-report or cross-disciplinary settings. ReportBench (Li et al., 2025b) evaluates research reports using citation quality and statement faithfulness, while SurveyLens (Guo et al., 2026) studies automatic survey generation across ten disciplines using discipline-aware rubric scoring and reference-based alignment. These benchmarks broaden the evaluation population and criteria beyond individual generation systems, while exposing persistent weaknesses in citation quality, coverage, and factual consistency.

SURVEYAUDIT addresses a different evaluation target within this literature. The above approaches primarily improve survey generation, use evaluation as feedback for generation, or compare systems using aggregate quality criteria. SURVEYAUDIT instead treats evaluation as the evidence-grounded detection of concrete reviewer concerns: given a completed survey, it identifies specific concerns that can be operationalized under CONTAX and records the evidence needed to inspect each one. This shifts the output from a holistic score or generation-time feedback signal to an auditable set of evidence-linked concerns, while leaving the survey-generation process itself independent of the evaluator.

A.2 EVIDENCE-GROUNDED SCIENTIFIC VERIFICATION

Several lines of research have studied how automatically generated scientific statements can be verified against external evidence. SciFact (Wadden et al., 2020) formulates scientific claim verification as retrieving evidence that supports or refutes a claim and identifying the corresponding rationale. SciFact-Open (Wadden et al., 2022) extends this setting to open-domain retrieval over a large scientific corpus, highlighting the difficulty of evidence retrieval when the relevant source is not known in advance. FActScore (Min et al., 2023) similarly decomposes generated text into atomic facts and evaluates each fact against an external knowledge source, demonstrating the value of fine-grained rather than purely holistic verification. More recent work directly addresses citation and attribution verification: SciCiteVal (Liu et al., 2026) constructs a benchmark for determining whether cited literature adequately supports citation contexts, while SciTrue (Tan et al., 2026) emphasizes source-level accountability and traceability for scientific claim verification.

These studies motivate the evidence requirements used by SURVEYAUDIT's consistency detectors and its evidence-tracing protocol. SURVEYAUDIT differs in that the verification target is not an isolated scientific claim or citation, but a concrete reviewer concern in a literature survey. The same evidence-oriented principle is therefore applied to heterogeneous concern types, with each reported concern required to satisfy its category-specific verification criterion and to retain the evidence or verification trace needed for inspection.

A.3 LITERATURE-REVIEW METHODOLOGY AND QUALITY CRITERIA

A separate literature-review methodology literature provides precedents for several of the substantive concerns represented in CONTAX. Snyder (Snyder, 2019) presents literature reviewing as a methodological process and identifies explicit review purpose and contribution, search strategy, inclusion and exclusion criteria, transparent data extraction, analysis, and synthesis as important aspects of review quality. Byrne (Byrne, 2016) discusses the peer review of narrative literature reviews and emphasizes rigor and synthesis as recurring quality considerations. Webster and Watson (Webster & Watson, 2002) advocate concept-centric organization and synthesis across prior work, while Torraco (Torraco, 2016) characterizes integrative reviewing as a process of combining prior findings to generate higher-level understanding and identify directions for future research.

These methodological criteria provide external grounding for several existence concerns in CONTAX, including missing selection methodology, contribution or positioning statements, evidential support, comparison, synthesis, and future-work discussion. SURVEYAUDIT does not assume that every review type must satisfy an identical checklist; instead, these sources motivate the kinds of review elements that can constitute substantively meaningful omissions, which CONTAX then converts into explicitly verifiable concern categories.

A.4 ALGORITHMIC LITERATURE MAPPING AND COVERAGE ANALYSIS

SURVEYAUDIT's coverage detector also draws on a broader literature-mapping tradition in which computational methods are used to delineate research fields and identify their intellectual structure. CiteSpace (Chen, 2006) uses citation and concept structure to detect research fronts and emerging patterns in scientific literature. VOSviewer (van Eck & Waltman, 2010) provides algorithmic construction and visualization of bibliometric maps based on relations such as co-citation and co-occurrence, while CitNetExplorer (van Eck & Waltman, 2014) explicitly uses citation networks to delineate literature on a research topic and support literature reviewing. These systems establish a precedent for using the structure of the scholarly literature to identify communities and important regions of a research field.

SURVEYAUDIT adopts similar structural signals for a different purpose. For C11, citation and literature networks are used to construct a literature pool, identify candidate literature elements ranging from individual works to coherent literature communities, and estimate whether important literature is insufficiently represented in the survey. The resulting signals support candidate generation and prioritization rather than serving as a normative rule that every structurally important work or community must appear in a survey.

B CONTAX DETAILS

We construct a multi-venue corpus of publicly available peer-review records for survey papers on OpenReview. The corpus provides a curated sample for characterizing the aspects of survey papers that reviewers tend to emphasize. Candidate submissions were first retrieved using title- and venue-based search queries, followed by regular-expression filtering for survey-related expressions such as "survey", "summary" and "literature review". We then manually review the retained candidates to verify that each submission is genuinely a survey, review, or closely related synthesis paper. We retain only submissions with sufficiently detailed publicly accessible review content and downloaded the corresponding submission and review records from OpenReview. The resulting corpus contains 168 survey papers from TMLR, ICLR, EMNLP, NeurIPS, and CoLM. TMLR contributes the majority of the corpus because its continuous journal-style OpenReview workflow provides publicly accessible reviews, while the other venues contribute fewer survey-like submissions under their respective review and track structures. We therefore treat the corpus as a curated multi-venue sample for reviewer-concern analysis rather than as a statistically representative estimate of the population of survey papers.

Table 7 reports the venue distribution of the corpus across the development and held-out test sets, while Table 8 reports the distribution of reviewer concerns annotated with CONTAX. Together, these tables characterize the venue composition of the evaluation corpus and the empirical distribution of reviewer concerns used to develop and evaluate the taxonomy.

C SURVEYAUDIT IMPLEMENTATION DETAILS

C.1 SHARED REPRESENTATION

SURVEYAUDIT takes a topic query $\mathcal{T}$ and a parsed survey document $\mathcal{D}$. The document is represented as a hierarchy of content units. Each unit is represented as $u = (s_u, p_u, n_u, d_u, X_u, T_u, O_u, A_u)$, where s_u is the unit identifier, p_u its parent identifier, n_u its section name, d_u its structural depth, X_u its text content, T_u its extracted topics, O_u its research objects, and A_u its citation anchors. A content unit can correspond to a section, subsection, subsubsection, or text directly attached to a

Venue	Development	Test	Total
TMLR	118	30	148
ICLR	0	7	7
EMNLP	0	6	6
COLM	0	4	4
NeurIPS	0	3	3
Total	118	50	168

Table 7: Distribution of the 168 human-written surveys across venues in the development and held-out test sets.

Concern Category	Development	Test	Total
C1	223	62	285
C2	36	12	48
C3	128	58	186
C4	53	14	67
C5	24	6	30
C6	37	28	65
C7	168	13	181
C8	136	17	153
C9	32	16	48
C10	108	27	135
C11	167	69	236
External norms	572	130	702
Evaluative	1000	232	1232
Total	2684	684	3368

Table 8: Distribution of annotated reviewer concern instances across CONTAX categories in the development and held-out test sets.

parent section before its first child. A terminal content unit is a unit with no child units. We use $\mathrm{Desc}(u)$ and $\mathrm{Sib}(u)$ for the descendants and siblings of u, respectively.

C.2 EVIDENCE-GROUNDED CONSISTENCY VERIFICATION

SURVEYAUDIT implements C1–C3 as evidence-grounded consistency checks. The three detectors differ in the evidence source against which a survey statement is verified: C1 uses evidence within the survey, C2 uses the literature explicitly cited by the survey, and C3 uses independently retrieved external literature. In all three cases, the detector first identifies statements that may violate the corresponding consistency criterion and then verifies them against the appropriate evidence source.

C.2.1 C1: INTERNAL INCONSISTENCY

C1 contains structural and local consistency checks. For structural checks, the system extracts contribution statements and structural commitments from the survey and associates each commitment with its intended target content units. An evidence-backed semantic judgment determines whether the target content fulfills the commitment and returns either FULFILLED or NOT_FULFILLED together with verbatim evidence. The commitment types include coverage, comparison, synthesis, future-work, and contribution statements.

For local checks, the system scans the survey paragraph by paragraph and extracts atomic notes. Each note is represented as $n_i = (\kappa_i, s_i, \pi_i, o_i, e_i)$, where κ_i is the note category, s_i the subject, π_i the predicate, o_i the object, and e_i the verbatim evidence text. The note index i provides a stable identifier for subsequent conflict judgments. The system merges notes referring to the same underlying object and compares compatible note pairs. A local inconsistency is retained only when at least two explicit

statements are directly irreconcilable. Differences in wording, compatible refinements, changes in plurality, or explicitly declared changes of setting are not treated as inconsistencies.

C.2.2 C2: CITATION HALLUCINATION

C2 operates on citation-bearing sentences identified from the parsed citation structure. Each sentence is decomposed into atomic facts using the claim decomposition procedure, while retaining the originating sentence for each fact. The verifier uses three possible judgments: SUPPORTED, REFUTED, and NEUTRAL. When the title and abstract of a cited work are insufficient to determine whether the claim is supported, the system proceeds to full-text verification. If the cited work cannot be resolved or its full text is unavailable when full-text verification is required, the result is treated as insufficient evidence rather than as an concern.

Full text is divided into chunks of at most 2,000 tokens, preferring paragraph boundaries and using sentence boundaries when necessary. The shared reranking model is applied to retrieve the five most relevant chunks. For each atomic fact, the verifier returns one of the three judgments together with verbatim evidence copied from the inspected source. A concern is retained when the cited literature reliably refutes the claim or when full-text verification establishes that the claim is unsupported. Citations within the same sentence are verified independently.

C.2.3 C3: TECHNICAL INACCURACY

C3 verifies substantive technical claims against external literature. Citation-linked claims that can be verified through the cited source are handled by C2; C3 focuses on remaining claims after excluding contribution statements, background descriptions, and conclusions. The detector first searches the literature cited by the survey using the same claim-verification procedure as C2. When the available cited literature does not determine the claim, the detector performs external scholarly retrieval using OpenAlex.

The retrieval query is generated from the original claim subject to a lexical constraint: every non-stopword in the query must occur in the original claim. OpenAlex returns up to 200 candidate papers, which are reranked using the shared reranking model. The highest-ranked papers are downloaded and processed using the same chunking procedure as C2, after which the most relevant evidence is passed to the final verifier. Only a REFUTED judgment supported by explicit external evidence produces a retained concern; insufficient or inconclusive evidence does not.

C.3 STRUCTURAL COMPLETENESS DETECTION

C4–C10 implement the structural-completeness pattern in Figure 1. The external methodological framework, argumentation principle, or scholarly practice establishes why an element is expected in a particular context. The detector then operationalizes this expectation through explicit contextual triggers, extraction rules, presence predicates, retrieval procedures, and verification criteria.

For categories that produce multiple omission instances, the detector first enumerates the missing instances and only then ranks their importance. Thus, importance affects which detected omissions are prioritized or reported, but does not determine whether an omission exists.

C.3.1 C4: MISSING SURVEY SCOPE OR SELECTION

PRISMA 2020 requires systematic reviews to report eligibility criteria, information sources, search strategies, and study-selection procedures (Page et al., 2021). Accordingly, when a survey presents itself as systematic or otherwise claims structured literature coverage, an explicit account of how the literature was identified and selected is an expected element. This expectation concerns the reporting of the literature-collection process rather than any particular search strategy.

C4 extracts sentences describing literature sources, search procedures, inclusion or exclusion criteria, selection procedures, or equivalent methodological information. If such information is identified, the requirement is considered satisfied. If no eligible statement is found, the detector reports one document-level concern; the inspected document scope provides the bounded absence certificate.

C.3.2 C5: MISSING CONTRIBUTION

Literature reviews can themselves constitute research contributions. (Torraco, 2005) characterizes the integrative literature review as a form of research that can generate new knowledge, while (Snyder, 2019) treats literature review as a research methodology and discusses its contribution and evaluation. In contexts where a survey is presented as a research contribution, an explicit statement of what the survey contributes and how that contribution relates to relevant prior surveys or recent developments is therefore an expected element. This expectation is particularly relevant in publication settings where novelty and significance are explicit review criteria, but it is not assumed to apply uniformly to every survey.

C5 retrieves potentially relevant prior surveys from the recent literature and filters them using title and abstract evidence for survey status and topical relevance. It then extracts the contribution statements of both the target survey and the retrieved prior surveys. Each target contribution is matched against the prior contributions. A concern is retained when the target contribution is fully subsumed by an identified prior contribution and the survey does not provide an explicit distinction or positioning statement. Both sides of a retained match are accompanied by verbatim evidence.

C.3.3 C6: MISSING EVIDENCE SUPPORT

Toulmin (1958)'s argument model represents a claim as supported by data or grounds through a warrant connecting the evidence to the conclusion. This framework does not imply that every sentence requires explicit evidence; rather, evidential support is relevant when an author advances a non-obvious substantive claim that requires the reader to accept a conclusion. Thus, in a survey, a claim of sufficient substantive strength creates an expectation of supporting evidence, data, or an explicit argumentative bridge.

C6 operationalizes this principle using a finite set of claim patterns commonly associated with such support requirements: numerical, empirical, comparative, causal, theoretical, and strong-generalization claims. These patterns are implementation heuristics rather than categories prescribed by Toulmin's model. The detector considers uncited claims of these forms, including claims that remain unsupported after the external-literature verification performed for C3. It excludes cited sentences, figure and table content, routing statements, normative suggestions, future-work statements, and concentrated discussion or conclusion passages.

For each candidate claim, the detector inspects the smallest complete content unit containing the claim and checks whether a reasonable direct supporting passage exists. The verifier evaluates support existence rather than truth or the overall sufficiency of the argument. A claim with an identifiable supporting passage is discarded. Otherwise, the claim is retained as an concern, with the claim itself as verbatim evidence and the inspected scope recorded as a bounded absence certificate.

C.3.4 C7: MISSING COMPARISON

Literature-review methodology treats a review as an analytical activity that relates and integrates findings across studies rather than merely listing individual papers. Cooper (1982) characterizes research reviews as a form of scientific inquiry, while Popay et al. (2006) identify the exploration of relationships within and between studies as a central component of narrative synthesis. These principles motivate checking for explicit comparison when a survey contains multiple related methods, approaches, or research objects addressing a common topic or question.

C7 first enumerates all potential comparison opportunities and identifies which of them lack an explicit comparison. The detector uses two structural patterns. For a sibling group $U_p = \{u : p_u = p\}$, a local comparison opportunity is created when $|U_p| \geq 2$ and the corresponding joint scope contains no extracted comparison relation. For an individual content unit u, a topic-group comparison opportunity is created when $|T_u| \geq 2$, $|O_u| \geq 1$, and the corresponding subtree contains no extracted comparison relation. Existing semantic comparison records therefore determine whether each opportunity is missing; no importance score is used at this stage.

For each comparison opportunity o, let $\Sigma(o)$ denote the content units in which the comparison is expected, and let $T(o)$ and $O(o)$ denote the associated topics and research objects. We represent the

Algorithm 1 Non-redundant selection of missing structural obligations for C7 and C8

Require: Set of missing structural obligations $\mathcal{M}$
$\mathcal{S} \leftarrow \emptyset$
sort $\mathcal{M}$ by descending $I(o)$
for all $o \in \mathcal{M}$ **do**
  if o is not subsumed by any obligation in $\mathcal{S}$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{o\}$
  end if
end for
return $\mathcal{S}$

required content by

$$C(o) = \Sigma(o) \cup \{\texttt{topic:}t : t \in T(o)\} \cup \{\texttt{object:}x : x \in O(o)\}.$$

For a cited paper anchor a, $C(a)$ denotes the corresponding normalized signature over content units, topics, and research objects associated with that anchor. Given the set A of citation anchors associated with the obligation scope, the structural gap is

$$G(o \mid A) = \frac{\left|C(o) \setminus \bigcup_{a \in A} C(a)\right|}{\max(1, |C(o)|)}.$$

The quantity measures how much of the expected comparison structure is not represented by the current anchors. Let

$$D(o) = \max_{u \in \Sigma(o)} \frac{d_u}{d_{\max}},$$

where d_u is the structural depth of content unit u in the survey hierarchy and $d_{\max}$ is the maximum depth of any content unit. We rank all detected missing comparison opportunities using

$$I(o) = \frac{D(o) + G(o \mid A)}{2}.$$

The ranking is performed only after missing comparison opportunities have been enumerated. Redundant opportunities are then removed in descending order of $I(o)$.

C.3.5 C8: Missing Cross-Study Synthesis

The same literature-review methodology provides the grounding for synthesis. Popay et al. (2006) describe synthesis as an analysis of relationships and patterns across studies, while McKenzie & Brennan (2024) Handbook treats synthesis as a formal stage of review analysis rather than a simple collection of individual study descriptions. Accordingly, when a survey contains multiple studies or approaches whose findings can be related, an explicit synthesis that integrates them into higher-level conclusions is an expected element.

C8 first enumerates all potential synthesis opportunities and identifies those lacking an explicit synthesis relation. For each content unit u, the normalized semantic-key set is defined as $K(u) = \{\text{lower}(x) : x \in T_u \cup O_u\}$. For sibling units u_i and u_j, a synthesis opportunity is created when $K(u_i) \cap K(u_j) \neq \emptyset$, and no synthesis relation is extracted from their joint scope. For an individual content unit u, a topic-group synthesis opportunity is created when $|T_u| \geq 2$, $|O_u| \geq 1$, and no synthesis relation is extracted from the corresponding subtree. The set of opportunities lacking such a relation forms the missing-synthesis set. These instances are ranked using the same $I(o)$ definition and the non-redundant selection procedure in Algorithm 1.

C.3.6 C9: Missing Future Work and Research Gaps

Following the research-gap framework of (Robinson et al., 2011), we treat a research gap as a context in which the available evidence is insufficient to support a satisfactory conclusion and further investigation is therefore warranted. For survey evaluation, the following evidence patterns define subtopics that may incur a future-work obligation: conflicting conclusions across otherwise comparable studies; systematically under-supported evidence dimensions; materially non-comparable

methods or experimental designs; concentration of evidence in an earlier research focus despite later developments; and a distinct recent research paradigm supported by multiple studies but not yet adequately integrated into the survey.

C9 first enumerates future-work and research-gap signals across the survey rather than assigning importance before detecting them. At the document level, the detector extracts explicit future-work, research-gap, limitation, and open-problem discussion. If none is found, the detector reports one document-level missing-future-work concern.

Otherwise, each survey subtopic is screened against at most $M = 10$ representative papers using title, abstract, and publication year. One LLM call jointly evaluates all five evidence patterns for the subtopic and returns potential gap propositions together with the supporting papers and evidence dimensions involved. For a gap proposition g, let P_g be its supporting paper set, Φ_g its distinct evidence-dimension set, and $E_g \in \{0, 1\}$ indicate whether at least two independent papers explicitly identify the concern as unresolved, limited, or otherwise insufficient. Potential gaps are ranked lexicographically by $P_{\text{screen}}(g) = \big(|P_g|, |\Phi_g|, E_g\big)$. The top $K_{\text{verify}} = 10$ ranked gap propositions are retained for targeted full-text verification, with at most three supporting papers inspected for each. The shared C2 retrieval procedure is used for full-text evidence collection. The verifier then checks whether the target survey substantively covers the proposed gap and returns COVERED or NOT_COVERED with verbatim evidence. Only NOT_COVERED propositions remain as detected missing future-work or research-gap instances. Their verified priority is recomputed as

$$P_{\text{final}}(g) = \big(|P_g^\star|, |\Phi_g^\star|, E_g^\star\big),$$

after which semantically equivalent instances are deduplicated and the top $K_{\text{out}} = 5$ are retained.

The detector does not infer a research gap from a low publication count alone. In particular, a structural evidence gap requires positive evidence that the missing dimension is relevant and systematic evidence that it remains under-supported within the inspected literature.

C.3.7 C10: Missing Citation

C10 follows the scholarly attribution principle that relevant prior work should be identified when it establishes the context, positioning, or attribution of a survey's claims. Literature-review methodology emphasizes positioning a review within the existing knowledge base (Snyder, 2019), and guidance on reviewing narrative literature reviews likewise treats engagement with relevant prior literature as part of review quality (Byrne, 2016). Development-set analysis further indicated two recurring citation-omission patterns, which are implemented as separate detection procedures within the same C10 category.

The first procedure checks whether relevant prior surveys are cited or explicitly positioned. The detector retrieves recent surveys that match the target topic and checks whether each relevant prior survey is present in the target bibliography or explicitly discussed in the survey's contribution positioning. A relevant prior survey with neither form of attribution produces a retained concern.

The second procedure checks attribution of externally established scientific entities. The detector reuses named methods, datasets, benchmarks, models, frameworks, and other substantive research objects from the shared content map rather than issuing a new document-level semantic extraction call. Repeated mentions, abbreviation/full-name variants, and equivalent naming forms are merged before retrieval. For each remaining entity, the entity name is used as a retrieval query against the existing literature pool and OpenAlex. Up to three candidate papers are retained, and an evidence-backed semantic judgment verifies whether the source actually introduces, defines, or establishes the entity using title and abstract evidence. A verified source produces a retained concern; failure to identify a reliable source does not.

C.4 Literature Coverage

C11 uses two complementary coverage detectors and a budget of $K = 5$ retained missing-coverage instances. The detector first enumerates potentially missing literature instances and then ranks their importance. The budget controls the number of retained instances or supporting papers reported in the audit report rather than creating a separate concern for every supporting paper.

C.4.1 KEY-LITERATURE COVERAGE

For each eligible content unit u, the detector forms the relevance query $q_u = \mathcal{T} \oplus T_u$ and the local citation graph $V_u = A_u \cup N(A_u)$, where $N(A_u)$ contains first-order citation neighbors of the unit's cited papers. The local graph is capped at 100 most-related papers. All citation anchors are retained, and the remaining nodes are selected by citation connectivity.

Key-literature detection first constructs the set of uncited papers that are structurally connected to the survey's cited literature and whose best local PPR rank is at most 10. These papers form the initial missing-literature pool. Importance ranking is then performed over this pool.

The detector uses a directed local citation graph and Personalized PageRank (PPR) (Page et al., 1999). The personalization vector is

$$z_u(v) = \begin{cases} 1/|A_u|, & v \in A_u, \\ 0, & \text{otherwise.} \end{cases}$$

The resulting local importance is denoted by $\pi_u(v)$. Let $n_u = |A_u|$, let $\mathcal{N}$ be the collection of citation counts across eligible content units, and let $a_u = Y_{\text{eval}} - \max_{p \in A_u} Y(p)$, where Y_{eval} is the evaluation year. The citation-sparsity and citation-staleness signals are

$$F_{\text{few}}(u) = 1 - \frac{|\{n \in \mathcal{N} : n \le n_u\}|}{|\mathcal{N}|},$$

and

$$F_{\text{old}}(u) = \frac{|\{a \in \mathcal{A} : a \le a_u\}|}{|\mathcal{A}|},$$

where $\mathcal{A}$ is the collection of corresponding citation ages. Let h_u be the number of the five highest-ranked local PPR papers already cited by the survey. Then

$$F_{\text{core}}(u) = 1 - \frac{h_u}{5}, \qquad N(u) = \frac{F_{\text{few}}(u) + F_{\text{old}}(u) + F_{\text{core}}(u)}{3}.$$

For an uncited paper p, its local support weight is

$$S(p, u) = \frac{1}{\log_2(\text{rank}_u(p) + 1)}.$$

The initial importance score over the full missing-literature pool is

$$B(p) = \sum_{u : p \in V_u} N(u) S(p, u).$$

To diversify the final set across subtopics, the detector maintains cumulative coverage C_u and uses the marginal score

$$\Delta(p \mid A) = \sum_{u : p \in V_u} \frac{N(u) S(p, u)}{1 + C_u}.$$

After each selection, the detector updates $C_u \leftarrow C_u + S(p, u)$. Selection terminates after $K = 5$ supporting papers or when no candidate has positive marginal gain.

The selected supporting papers are then checked by an evidence-backed semantic judgment using the corresponding survey subtopic, local survey context, and candidate title and abstract. A selected paper is retained only when the supplied evidence establishes its relevance to the target subtopic. All retained supporting papers are attached to the corresponding missing-coverage concern rather than counted as separate concerns.

C.4.2 KEY-SUBFIELD COVERAGE

Let $V_{\mathcal{D}}^{\text{cited}} = \bigcup_u A_u$ be the set of papers cited by the survey. The key-subfield coverage detector constructs the undirected induced graph $G_U = G[V \setminus V_{\mathcal{D}}^{\text{cited}}]$, and applies Leiden community detection at resolutions $\Gamma = \{0.5, 1, 2, 4, 8\}$. Communities smaller than $\max\left(3, \lceil 0.01 |\mathcal{L}(\mathcal{T})| \rceil\right)$ are removed, and identical node sets obtained at different resolutions are merged.

Algorithm 2 Key-literature support selection

Require: Missing-literature pool $\mathcal{M}$, budget K
$A \leftarrow \emptyset$ and $C_u \leftarrow 0$ for all eligible u
compute $B(p)$ for all $p \in \mathcal{M}$
while $|A| < K$ and eligible papers remain **do**
  compute $\Delta(p \mid A)$ for each eligible p
  if $\max_p \Delta(p \mid A) \leq 0$ **then**
    break
  end if
  $p^\star \leftarrow \arg\max_p \Delta(p \mid A)$
  $A \leftarrow A \cup \{p^\star\}$
  for all $u : p^\star \in V_u$ **do**
    $C_u \leftarrow C_u + S(p^\star, u)$
  end for
end while
return A

Algorithm 3 Key-subfield gap selection

Require: Literature graph G, survey citation set $V_{\mathcal{D}}^{\text{cited}}$, budget K
construct $G_U = G[V \setminus V_{\mathcal{D}}^{\text{cited}}]$ and run Leiden at all $\gamma \in \Gamma$
remove small communities and merge identical node sets
semantically classify every remaining community using representative papers
verify substantive survey coverage for communities initially labeled COVERED
$\mathcal{M} \leftarrow$ uncovered coherent directions not explicitly excluded by the survey
compute $R_{\text{topic}}(c)$ for every $c \in \mathcal{M}$
sort $\mathcal{M}$ by descending R_{topic}, using M_{global} as a tie breaker
remove semantically duplicate communities
$\mathcal{S} \leftarrow$ top K remaining communities
return $\mathcal{S}$

The detector first semantically classifies all remaining literature communities and checks whether communities initially judged COVERED receive substantive discussion in the survey rather than only a name-only mention. For each community c, the five globally highest-ranked papers serve as semantic representatives. An evidence-backed semantic judgment assigns one of COVERED, COHERENT_NEW_DIRECTION, MIXED, or UNRELATED and supplies verbatim evidence from the representative papers. A community is included in the missing-subfield pool only when it represents a coherent direction that is not substantively covered and is not explicitly excluded by the survey boundary.

Only after the missing-subfield pool has been constructed do we compute importance. For each retained community, we denote $M_{\text{global}}(c) = \sum_{v \in c} PR(v)$ and $M_{\text{cited}}(c) = \sum_{v \in c} PPR_{\mathcal{D}}(v)$. The community importance score is

$$R_{\text{topic}}(c) = \log(1 + |c|) \left(1 - \frac{M_{\text{cited}}(c)}{M_{\text{global}}(c)}\right).$$

Missing subfields are sorted by $R_{\text{topic}}(c)$, with $M_{\text{global}}(c)$ as the tie breaker. Semantically duplicate communities are then removed; two communities are treated as duplicates if their names are identical or $\frac{|c_i \cap c_j|}{\min(|c_i|, |c_j|)} \geq 0.8$. The top $K_{subfield} = 5$ missing subfield instances are retained.

D Main Experiment Details

D.1 Evaluation Label Distribution

Table 9 summarizes the final accept/reject labels of the 50 human-written surveys used in our held-out evaluation set. The overall set is balanced between accepted and rejected surveys. The TMLR and non-TMLR subsets have different label distributions, which we preserve throughout the corresponding subgroup analysis in Appendix E.

Subset	Accept	Reject	Total
TMLR	11	19	30
Non-TMLR	14	6	20
Total	25	25	50

Table 9: Accept/reject label distribution of the 50 human-written surveys in the held-out evaluation set.

D.2 Human Annotation Details

Detected Concern Precision. We recruited three Ph.D. students with backgrounds in computer science or artificial intelligence. Two annotators independently labeled each detected concern, while the third annotator reviewed the annotations and resolved disagreements. Given the full text of a survey and a set of detected concerns, annotators independently answered whether they would include each concern in their official review if they were reviewing the survey. The two independent annotators agreed on 76.3% of the concern-level judgments.

Reviewer Concern Coverage. We used the same three-annotator procedure for RCC. Two Ph.D. students independently annotated the reviewer concerns, while the third reviewed the annotations and resolved disagreements. For each survey, annotators were given all official reviews and the final audit report generated by an evaluator. Following the concern decomposition procedure used to construct ConTax, each official review was first decomposed into individual reviewer concerns. For each concern, annotators independently determined whether it was covered by at least one concern in the audit report. We define a reviewer concern as *covered* when the report contains an concern whose underlying failure addresses the same substantive concern, even if the concern is more specific or is formulated under a different ConTax category. In particular, addressing the detected concern should constitute a concrete way of addressing the substantive concern expressed by the reviewer. For example, a reviewer concern that a survey insufficiently compares different methods is considered covered if the audit report identifies a specific missing comparison and provides corresponding evidence. The two independent annotators agreed on 80.2% of the concern-level judgments.

DCP and RCC are reviewer-grounded measures rather than objective correctness measures, and their annotation agreements (76.3% and 80.2%) indicate residual subjectivity in issue acceptance and concern matching.

D.3 Controlled Perturbation Details

We construct controlled perturbations on 20 human-written surveys, with 5 perturbation cases for each of 5 failure types, yielding 500 perturbed variants in total. Each variant contains exactly one injected defect. The perturbations are designed to create a direct correspondence between the injected defect and a ConTax category while preserving the remainder of the survey.

Internal inconsistency (P1). P1 targets C1. For each survey, the first two perturbations modify statements about the paper's contributions, while the remaining three modify symbols, terminology definitions, or section numbering. These perturbations introduce inconsistencies between statements that should remain mutually consistent while leaving the surrounding content unchanged.

Citation hallucination (P2). P2 targets C2. We select sentences that contain citations and whose factual claims can be verified against the cited source. We then inject a citation hallucination into the selected sentence while preserving its citation structure. Background, conclusion, and contribution statements are excluded from the candidate pool, as these sentences are less directly amenable to source-based verification.

Technical inaccuracy (P3). P3 targets C3. We select sentences that contain no citations but make factual or technical claims whose correctness can be determined from domain knowledge. We inject a technical error into the selected sentence. Background, conclusion, and contribution statements are excluded from the candidate pool.

#	Topic	Domain
1	Aspect-Based Sentiment Analysis	NLP
2	Commonsense Reasoning	NLP
3	Coreference Resolution	NLP
4	Dialogue Systems	NLP
5	Information Extraction	NLP
6	Large Language Models	NLP
7	Multilingual NLP	NLP
8	3D Point Cloud Processing	CV
9	Action Recognition	CV
10	Autonomous Driving Perception	CV
11	Depth Estimation	CV
12	Few-Shot Image Recognition	CV
13	Image-To-Image Translation	CV
14	Automatic Speech Recognition	Speech
15	Active Learning	Core ML
16	Adversarial Robustness	Core ML
17	Community Detection	Core ML
18	Continual Learning	Core ML
19	Cross-modal Retrieval	Core ML
20	Diffusion Models	Core ML
21	Graph Classification	Core ML
22	Efficient Transformers	Core ML
23	Multi-agent Reinforcement Learning	Core ML
24	Graph Neural Networks	Core ML
25	Imbalanced Learning	Core ML
26	Bayesian Deep Learning	Extended ML
27	AI for Code Generation	Application
28	AI for Drug Discovery	Application
29	AI Fairess and Bias	Application
30	Embodied AI	Application

Table 10: The 30 survey topics used for AI survey generation.

Missing content (P4). P4 targets C4, C5, C7, C8, and C9 through deletion-based perturbations. Each perturbation removes the sentences and sections that collectively serve a particular function. For C7–C9, which generate candidate failures, the deletion is restricted to the content associated with a single subtopic: all sentences and sections serving the corresponding function for that subtopic are removed. Consequently, P4 removes substantially more content than the other perturbation types, reflecting the functional nature of these failure categories rather than deletion of an isolated sentence.

Citation omission (P5). P5 targets C10. We remove a citation that has a unique association with a particular entity, such that the deleted citation can be unambiguously linked to that entity. This construction provides a well-defined ground-truth citation omission while avoiding cases where multiple citations could independently support the same entity.

C6 and C11 are excluded from the controlled perturbation benchmark because their ground-truth conditions cannot be isolated reliably. For C6, constructing an isolated perturbation would require a reliable ground-truth set of claims that objectively warrant additional evidence, which is not directly observable from the original survey. For C11, a coverage perturbation would require a ground-truth ranking over topics that are absent from the original survey. The absence of a topic alone does not establish a coverage failure, and artificially deleting a topic would therefore require an external judgment about whether it should have been covered and how it should rank relative to other omitted topics.

D.4 Topic Pool and Sampling for AI-generated Surveys

We directly select 30 ML/AI survey topics for generation. Table 10 lists the selected topics, domain strata, and reference-survey counts.

C2 Issue Type	SurveyG	SurveyGen-I	Claude Code
Citation key absent from bibliography	20.7	2.5	0.1
Citation–claim mismatch	0.2	4.2	10.3
Fabricated reference	1.8	9.8	1.9
Total	**22.7**	**16.5**	**12.3**

Table 11: Distribution of C2 (Citation Hallucination) issues across the three survey generators. Values denote the average number of issues per generated survey.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	EN	Eval.
Conventional	4/8	0/5	3/5	6/6	0/0	10/11	18/20	6/7	1/2	0/0	18/27	24/137	141/216
Tool-augmented	4/6	2/6	1/2	3/3	1/1	4/4	15/18	3/3	1/1	0/1	23/28	37/160	115/217
SURVEYAUDIT	54/71	10/12	18/21	16/16	9/9	40/52	89/103	60/68	24/29	102/105	70/82	–	–

Table 12: Per-category DCP. Each cell reports accepted concerns over all evaluated concerns. We use "Conventional" and "Tool-augmented" for "Conventional LLM Judge" and "Tool-augmented LLM Judge" respectively. EN stands for consistency concerns with external norms and Eval. stands for evaluative concerns.

E SUPPLEMENTARY EXPERIMENTS

This section provides additional analyses of main results. These experiments examine category-level behavior, subgroup differences, the granularity of detected concerns, case analysis and error analysis.

E.1 HUMAN-WRITTEN SURVEYS

We further analyze the human-written survey evaluation from 3 perspectives: per-category reviewer alignment, differences between TMLR and non-TMLR surveys, and the granularity of accepted concerns relative to underlying reviewer concerns.

E.1.1 PER-CATEGORY DCP

Table 12 shows that SURVEYAUDIT achieves higher DCP across most CONTAX categories, but its performance varies substantially across failure types. Within C1–C11, DCP varies substantially across failure categories. C1, C7–C9, and C11 account for a larger share of unacceptable issues, whereas C2–C6 and C10 are comparatively more reliable. This difference is closely related to the operational requirements of the corresponding detectors. C7–C9 require the LLM to extract textual elements that serve specific functional roles in a survey, which is more demanding than identifying explicit lexical or structural patterns. C1 also relies more heavily on semantic judgment to determine whether two statements are genuinely inconsistent, increasing the risk of false positives. C11 is particularly challenging because many OpenReview surveys combine two or more technical concepts, making it difficult to determine whether a given paper or research community falls within the intended scope of the survey. In contrast, the detectors for C2–C6 and C10 are deliberately designed to favor precision over recall, using more constrained verification procedures to compensate for the unreliability of LLM-based judgments.

By comparison, the general-prompt baselines more frequently produce External Norms and Evaluative concerns, which are substantially less likely to be accepted by reviewers. Thus, unconstrained failure generation can increase the range of reported concerns without yielding a comparable increase in reviewer-acceptable concerns. SURVEYAUDIT also produces substantially more accepted concerns overall, suggesting that its higher acceptance is not achieved simply by being conservative in the number of concerns it reports.

E.1.2 PER-CATEGORY RCC

Table 13 shows a different pattern across categories. C1, C3, and C11 account for a relatively large share of the observed reviewer concerns, and SURVEYAUDIT recovers substantially more of the C1 and C11 concerns than the baselines. C3 remains more difficult for all systems. The current

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	EN	Eval.
Conventional	9/62	3/12	2/58	3/14	0/6	4/28	7/13	9/17	5/16	2/27	10/69	28/130	60/232
Tool-augmented	3/62	3/12	4/58	3/14	1/6	3/28	5/13	6/17	2/16	5/27	8/69	23/130	45/232
SURVEYAUDIT	30/62	10/12	16/58	13/14	3/6	12/28	10/13	13/17	11/16	21/27	43/69	6/130	48/232

Table 13: Per-category RCC. Each cell reports covered observed reviewer concerns over all golden concerns in the category.

	TMLR			Non-TMLR		
Method	DCP%	RCC%	ACpS	DCP%	RCC%	ACpS
Conventional	57.41 (151/263)	16.77 (83/495)	5.03	44.20 (80/181)	31.22 (59/189)	4.00
Tool-augmented	48.11 (127/264)	13.33 (66/495)	4.23	43.78 (81/185)	23.81 (45/189)	4.05
SURVEYAUDIT	89.06 (350/393)	32.12 (159/495)	11.67	81.14 (142/175)	40.74 (77/189)	7.10

Table 14: Evaluation results on TMLR and non-TMLR surveys.

C3 detector focuses on single-document factual verification for claims without an explicit literature source, whereas many technical inaccuracies require evidence from multiple documents and reasoning across those sources, in addition to appropriate domain knowledge. This limitation is particularly relevant to RCC because a reviewer may recognize a technical error even when establishing it requires substantially more evidence than is available from the claim and its immediate context.

Interestingly, SURVEYAUDIT achieves reviewer-concern coverage on Evaluative concerns comparable to the baselines, despite not directly targeting Evaluative failures. This follows from the way RCC is defined: it matches the substantive concern expressed by the reviewer rather than requiring the detected concern to have the same evaluative form. A reviewer concern such as insufficient comparison can therefore be covered by a concrete, verifiable omission that addresses the same underlying problem.

E.1.3 TMLR VS. NON-TMLR

Table 14 shows that the gains of SURVEYAUDIT over the baseline evaluators on reviewer-grounded concern detection extend from TMLR surveys to non-TMLR surveys. The improvement in reviewer-concern coverage is larger on TMLR, where surveys contain substantially more reviewer concerns on average (16.5 vs. 9.45 per survey). This higher concern density provides more opportunities for concern detection and is consistent with the larger improvement observed on TMLR. Overall, these results suggest that the observed advantage of SURVEYAUDIT is not specific to the TMLR review setting and the same qualitative pattern is observed on the non-TMLR subset.

E.1.4 ISSUE-TO-CONCERN RATIO

ACpS can increase either because an evaluator identifies more underlying reviewer concerns or because it decomposes a single reviewer concern into multiple concrete concerns. We therefore examine the relationship between accepted concerns and the observed reviewer concerns they cover. We define the Issue-to-Concern Ratio as the number of accepted concerns that cover at least one golden reviewer concern divided by the total number of observed reviewer concerns:

$$\mathrm{ICR} = \frac{\#\{\text{accepted concerns covering at least one golden reviewer concern}\}}{\#\{\text{observed reviewer concerns}\}}.$$

A concern is counted at most once, even if it covers multiple reviewer concerns.

The Issue-to-Concern Ratio changes relatively little across the three systems. This suggests that the substantially higher ACpS of SURVEYAUDIT is not primarily explained by decomposing a small number of reviewer concerns into many separate concerns. Instead, the increase is accompanied by broader identification of underlying reviewer concerns, consistent with the higher Reviewer Concern Coverage.

Method	Issue-to-Concern Ratio
Conventional LLM Judge + General Prompt	1.29
Tool-augmented LLM Judge + General Prompt	1.40
SURVEYAUDIT	1.67

Table 15: Issue-to-Concern Ratio.

E.2 ERROR ANALYSIS

We analyze errors produced by SURVEYAUDIT by tracing them to the underlying detection process rather than grouping them by CONTAX category. We identify four recurring failure modes: *functional extraction*, *coverage-necessity judgment*, *evidence sufficiency/necessity judgment*, and *semantic comparability and consistency judgment*. Their estimated distribution is summarized in Table 16.

Functional Extraction Errors. The most common failure mode is the inability to reliably extract textual elements serving a specific survey function. This issue is particularly relevant to C7–C10, whose detectors first identify sentences or passages corresponding to functions such as comparison, synthesis, or discussion of a particular aspect of the literature. Unlike surface-level pattern matching, this step requires the LLM to infer the function served by a sentence from its local context. The difficulty is further amplified by jointly extracting multiple functional elements within a paragraph to reduce inference cost. In contrast, the extraction performed for C1 is more repetitive but targets a narrower and more homogeneous semantic pattern, making it less susceptible to extraction failures. Table 17 shows a representative example.

Semantic Comparability and Consistency Judgment Errors. A second failure mode occurs when the LLM identifies a superficial difference between statements without establishing that the compared quantities or propositions are semantically comparable. For example, the system may interpret two numerical statements as contradictory because their magnitudes differ, even though the statements describe different quantities or operate at different levels of scope. Such errors arise when surface-level discrepancies are used as a proxy for semantic inconsistency, without first resolving the objects, units, or scope being compared. Table 20 provides a representative example.

Evidence Sufficiency/Necessity Judgment Errors. The third failure mode arises when the detector incorrectly determines whether an observed element is sufficiently supported by the available evidence, or whether additional evidence is required. These errors are most relevant to detectors whose decisions depend on whether a claim, element, or omission satisfies an evidence-related obligation in its specific context. The underlying difficulty is that the applicability of such an obligation depends on contextual features that are not always captured by fixed decision rules. In the absence of reliable training annotations for these judgments, the current detectors rely on hand-designed criteria and therefore remain vulnerable to generalization failures across heterogeneous surveys. A representative case is presented in Table 19.

Coverage-Necessity Judgment Errors. The final failure mode concerns determining whether a paper or research community is sufficiently related to a survey query to warrant coverage. This decision becomes difficult for surveys whose scope combines multiple technical concepts, where relevance cannot be reliably determined from an isolated keyword or community label. The detector may therefore overestimate or underestimate whether a candidate falls within the intended coverage scope, even when the candidate itself has been correctly retrieved. This failure is particularly relevant to C11, where coverage requires reasoning about the relationship between a candidate and the survey's overall topic rather than verifying a localized textual property. A representative case is shown in Table 18.

Taken together, these cases indicate that the remaining errors are not concentrated in a single CONTAX category. Instead, they arise at distinct stages of the detection pipeline: extracting functionally defined elements, determining whether coverage or evidence is warranted, and establishing semantic comparability before making consistency judgments. This distinction helps explain the category-level performance differences reported in Table 16 and identifies limitations that are not visible from category-level recall or acceptance rates alone.

Failure Mode	Ratio	Most Affected
Functional Extraction	34%	C7,C8,C9,C10
Semantic Comparability/Consistency	20%	C1
Evidence Sufficiency/Necessity	16%	C6
Coverage Necessity	16%	C11
Four Types Combined	86%	

Table 16: Estimated distribution of major error types in SURVEYAUDIT.

Context	Incorrect Output	Diagnosis
To limit the scope of the survey, we focus on supervised learning. Interested readers may find references for collaborative unsupervised and self-supervised learning in (	The survey explicitly limits scope to supervised learning only, excluding unsupervised, self-supervised, and self-training approaches which represent active research directions in both federated learning and graph learning	The extraction error is that the later selection-relevant sentence was not extracted as a SELECTION item. The paragraph-level extraction retained the scope statement but missed the sentence that should have represented the survey's literature-selection boundary.

Table 17: Representative case of a FUNCTIONAL EXTRACTION error.

E.3 CASE STUDY

Quantitative experiments evaluate whether SURVEYAUDIT can reliably detect survey concerns at scale. We complement these results with a qualitative case study to illustrate how a reviewer concern is transformed into a concrete, verifiable concern. We select three concerns from the held-out human-written surveys for which the original reviewer comment expresses a relatively high-level or evaluative judgment that can be operationalized by an in-scope ConTax category. We additionally report the output of a generic LLM judge on the same survey context, providing a qualitative comparison of unconstrained and evidence-grounded concern detection.

Table 21 presents the three cases. In each case, the human reviewer provides the original concern, while the generic LLM judge produces a free-form description of a related problem. SURVEYAUDIT, in contrast, maps the concern to a specific ConTax category and instantiates it as a concrete proposition that can be checked under the corresponding verification criterion. The resulting concern is accompanied by the evidence or verification trace used to establish it, allowing the reader to inspect both the target proposition and the basis for reporting it.

The cases illustrate three complementary aspects of the framework. First, high-level reviewer judgments can often be decomposed into more concrete concerns without requiring the system to reproduce the reviewer's wording. Second, the ConTax category determines what constitutes sufficient evidence for reporting the concern, rather than leaving the judgment to an unconstrained LLM assessment. Third, the resulting output makes the basis of an individual concern explicit: rather than providing only a natural-language criticism, SURVEYAUDIT identifies what is being checked and retains the evidence or verification trace needed to inspect the result. The case study is intended as a qualitative illustration of this evidence-grounded formulation rather than as an additional quantitative evaluation.

F PROMPT DETAILS

We report the baseline and ablation prompts and all LLM prompts used internally by SURVEYAUDIT. Prompt text is reproduced verbatim after LaTeX escaping. ARISE prompts and the report aggregation and scoring prompts are omitted.

Context	Incorrect Output	Diagnosis
To limit the scope of the survey, we focus on supervised learning. Interested readers may find references for collaborative unsupervised and self-supervised learning in (	Survey explicitly limits scope to supervised learning without providing alternative references for unsupervised/self-supervised graph collaborative learning	The query is collaborative learning on graph-structured data. The retrieved community is unsupervised and self-supervised graph collaborative learning, including graph contrastive learning methods such as Deep Graph Infomax (DGI) [17041] and GraphCL [16848]. The error is judging this community as outside the query even though the survey itself mentions those methods and the query is not restricted to supervised learning.

Table 18: Representative case of a COVERAGE-NECESSITY JUDGMENT error.

Context	Incorrect Output	Diagnosis
This metric does not guarantee equalized odds since the TPR and FPR might differ between groups, even though groups have equal accuracy (Angwin et al., 2016).	The paper conflates Rawlsian min-max fairness with demographic group fairness but acknowledges the gap in the Future Directions section.	The cited passage distinguishes worst-case group fairness from sensitive-attribute-aware group fairness, but the issue treats this related discussion as evidence that the broader conflation claim is established. The error is an evidence-sufficiency judgment: the evidence is present in the search scope, but its entailment is overstated.

Table 19: Representative case of a EVIDENCE SUFFICIENCY/NECESSITY JUDGMENT error.

F.1 BASELINE AND ABLATION PROMPTS

F.2 SURVEYAUDIT INTERNAL PROMPTS

F.2.1 PREPROCESSING PROMPTS

F.2.2 C1: INTERNAL INCONSISTENCY

F.2.3 C2–C3: CITATION HALLUCINATION AND TECHNICAL INACCURACY

F.2.4 C4–C9: REQUIREMENTS VERIFICATION

The structural-obligation planner, retriever, and verifier are deterministic and introduce no additional LLM prompt. They consume the preprocessing extractions reported in Table 28.

F.2.5 C6: MISSING EVIDENCE SUPPORT

F.2.6 C10: MISSING CITATION

F.2.7 C11: MISSING LITERATURE COVERAGE

Context	Incorrect Output	Diagnosis
There are 10^{26} possible choices for each time-step in StarCraft II, and the number of possible games in AlphaGo is 10^{761}.	The survey contains inconsistent numerical claims: it gives10^{26} possible choices for StarCraft II and 10^{761} possible games for AlphaGo.	The input contains many notes at once. The model matches the two numbers on the surface and treats them as competing values, without checking the conditioning terms: one is the number of choices at one time-step in StarCraft II, while the other is the number of complete games in AlphaGo.

Table 20: Representative case of a SEMANTIC COMPARABILITY AND CONSISTENCY JUDGMENT error.

Human Reviewer	Generic LLM Judge	SURVEYAUDIT
Case 1: Missing literature-selection methodology (C4)		
The survey claims to provide a 'systematic review' but does not report its retrieval strategy, inclusion/exclusion criteria, or search scope.	The paper does not describe a systematic literature search methodology (e.g., which databases were searched, inclusion/exclusion criteria, search terms used).	**C4: Missing Survey Scope or Selection.** The original evidence is: 'We provide a systematic review of fairness notions that do not (fully) rely on sensitive attributes and recent algorithms to achieve them.' (Abstract, line 870); 'Under each category in the taxonomy, the paper provides a systematic review of existing techniques, their pros and cons, and the fairness metric they can handle.' (Introduction, Section 1, line 1871). The SELECTION extraction searched the whole paper and returned no SELECTION-class sentence for this survey. The subsequent C4 check therefore had no extracted selection-methodology item to match against the survey-level systematic-review claim.
Case 2: Paper-by-paper enumeration without synthesis (C8)		
Replace paper-by-paper enumeration with top-down categorization and synthesis of core methodologies.	No corresponding issue.	**C8: Missing synthesis.** The discussion of Methods Based on Unconditional Generation does not include cross-paper synthesis. **Evidence:** Paragraph 37: Methods Based on Unconditional Generation. Topics: input-dependent vs universal adversarial perturbations; Cross Domain Adversarial Perturbation. Objects: Generative Adversarial Pertruabtions (GAP); universal perturbations; Cross Domain Adversarial Perturbation; relativistic loss; relativistic cross entropy; Distribution and Neighbourhood Similarity Matching; Attentive-Diversity Attack (ADA); Conditional Adversarial Distribution (CAD); maximum model discrepancy (M3D). Role: trigger evidence.
Case 3: Omitted literature coverage (C11)		
The survey overlooks substantial spectral-GNN and related message-passing literature.	The survey focuses exclusively on supervised learning as stated, but mentions unsupervised/self-supervised learning only in passing with a reference, yet the graph-specific sections extensively discuss techniques like contrastive learning that are essentially self-supervised.	**C11: Missing domain coverage.** The original triage issue is: 'The survey overlooks substantial spectral-GNN and related message-passing literature.' Its evidence is the survey's graph-structured collaborative-learning scope together with the retrieved spectral-GNN and related message-passing literature; the triage reasoning treats this as an omitted subfield relevant to the stated scope.

Table 21: Qualitative comparison of reviewer concerns identified by a human reviewer, a generic LLM judge, and SURVEYAUDIT. The cases illustrate how high-level reviewer concerns can be operationalized into category-specific, evidence-grounded concerns.

Role: System instruction for the conventional LLM judge.

```
You are an expert reviewer with broad knowledge of machine learning and artificial intelligence research. Identify concrete problems in the supplied survey paper. Be exhaustive, but do not

summarize or praise the paper, score it, recommend acceptance or rejection, or assess its overall quality.
```

Table 22: System prompt for the conventional baseline evaluator.

Role: System instruction for the tool-augmented LLM judge.

```
You are an expert reviewer with broad knowledge of machine learning and artificial intelligence research. The survey paper to inspect is in {input_dir}, as a LaTeX source directory/file or
parsed JSON file. Identify as many distinct, concrete problems in the survey as the evidence supports. Do not summarize or praise the paper, score it, recommend acceptance or rejection, or
assess its overall quality.
{requirements}

Write your output as a JSON file in {output_file}.
```

Table 23: System prompt for the tool-augmented baseline evaluator.

Role: User prompt for the holistic baseline.

```
{SURVEY_FULL_TEXT}
<input_processing_note>
Ignore formatting artifacts, missing rendered tables/figures, reference formatting issues, and placeholder metadata like "Author names". Do not report these as reviewer concerns.
The survey's written date is {publication_date}. Use this date as a hard temporal boundary: do not recommend any work published after this date as a "missing" reference, and do not flag as
insufficient coverage any topic community whose publications postdate the survey.
</input_processing_note>
Inspect the entire survey and report as many distinct, concrete problems as you can support. Focus only on problems: do not provide a summary, strengths, neutral comments, a score, an
overall-quality judgment, or an accept/reject recommendation. Put each independently actionable problem in a separate item. Return exactly one top-level field in the following JSON format,
with no text outside the JSON object.
{{
"reviewer_concerns": [
{{
"issue": "<concise description of the problem>",
"location": "<section/paragraph/sentence where this occurs, quoted or paraphrased>"
}}
]

}}
```

Table 24: User prompt for holistic baseline evaluation.

Role: User prompt for the "–Algorithm" ablation setting.

```
Given a survey paper, identify as many distinct, concrete reviewer concerns as the evidence supports. Use the predefined CONTAX categories to organize the detected concerns, but do not rely
on category-specific detector procedures, structural matching algorithms, candidate-ranking rules, or other implementation-specific heuristics. Your task is to identify concerns directly from
the survey and available evidence under the category definitions below.
{SURVEY_FULL_TEXT}
<input_processing_note>
Ignore formatting artifacts, missing rendered tables/figures, reference formatting issues, and placeholder metadata such as "Author names". Do not report these as reviewer concerns. The
survey's written date is {publication_date}. Use this date as a hard temporal boundary: do not recommend work published after this date as a missing reference, and do not flag as insufficient
coverage any literature community whose relevant publications postdate the survey. </input_processing_note>
<categories>
The concern taxonomy distinguishes three natures. Report only concerns that fall within the 11 operational CONTAX categories below. Do not report open-ended evaluative concerns.
## N1: Consistency Concerns
These concerns arise when the survey contains mutually inconsistent statements or when a claim is contradicted or unsupported by relevant literature evidence.
- **C1: Internal Inconsistency.** Statements within the survey make mutually incompatible claims about the same object or condition, such as contradictory definitions, terminology, symbols,
quantities, structural claims, or other explicitly incompatible statements.
- **C2: Citation Hallucination.** A cited source does not support the claim attributed to it, or the cited source contradicts or is unrelated to the claim. Non-existent or unresolved citations
alone are insufficient unless the available evidence establishes the problem.
- **C3: Technical Inaccuracy.** A substantive technical claim in the survey is contradicted by relevant scholarly evidence, including claims about methods, algorithms, theory, empirical findings,
or other technical content.
## N2: Existence Concerns
These concerns arise when an expected element is absent from the survey.
- **C4: Missing Survey Scope or Selection.** The survey does not explicitly define the literature, topics, or criteria that determine its scope. When the survey presents itself as systematic
or otherwise claims structured literature coverage, this also includes missing information about how the literature was identified or selected, such as search sources or inclusion/exclusion
criteria. Do not require a particular search methodology unless the survey makes such a methodological claim.
- **C5: Missing Contribution.** The survey does not explicitly state its contribution or distinguish its positioning from relevant prior surveys when such positioning is part of the survey's
claimed contribution.
- **C6: Missing Evidence.** A substantive claim that requires evidential support is not accompanied by adequate support within the inspected survey context. Relevant claims include numerical,
empirical, comparative, causal, theoretical, and strong-generalization claims. Do not report a concern merely because a claim lacks a citation if the available survey content provides an
adequate supporting argument or evidence.
- **C7: Missing Comparison.** Related research objects that warrant comparison are presented without an explicit comparison. A concern may apply to the survey as a whole or to a specific
subtopic or local context when the absence is established within that scope.
- **C8: Missing Cross-Study Synthesis.** Related studies or approaches are presented without an integrative synthesis that relates them through higher-level relationships, trends, trade-offs,
or abstractions. A concern may apply to the survey as a whole or to a specific subtopic or local context when the absence is established within that scope.
- **C9: Missing Future Work.** The survey lacks an explicit discussion of future research, research gaps, unresolved problems, or future directions. A document-level concern may be reported when
no such discussion is found in the survey; a more localized concern may be reported when an important unresolved direction is established for a specific subtopic but is not addressed there.
- **C10: Missing Citation.** A specific relevant prior work or substantive research object is present or discussed in the survey but is not appropriately attributed to a source. This includes
relevant prior surveys that are neither cited nor explicitly positioned, and named methods, datasets, benchmarks, models, frameworks, or other research objects for which a reliable source can
be identified.
- **C11: Missing Literature Coverage.** Important and relevant literature is insufficiently represented in the survey. The missing literature element or research area must be supported by
retrieved evidence and should not be inferred from absence alone.
## N3: Evaluative Concerns
Do not report concerns that require open-ended judgments of quality or adequacy, such as writing clarity, insight quality, visualization quality, elegance, significance, or venue fit, unless
the concern can be instantiated as one of the operational C1–C11 categories above.
</categories>
<evidence_tracing>
Every reported concern must include an evidence anchor.
- **For C1:** provide the verbatim survey spans establishing the incompatible statements.
- **For C2–C3:** provide the relevant claim and verbatim evidence from the cited or retrieved literature that establishes the support, contradiction, or lack of support.
- **For C4–C10:** provide the relevant survey text and, when the concern is based on absence, a bounded absence certificate specifying the inspected scope, what was sought, and that no matching
element was found. For localized concerns, specify the corresponding section, topic, claim, or research-object context.
- **For C11:** provide the survey scope evidence together with the retrieved literature or literature-community evidence establishing that the missing material is relevant and important.
Do not treat retrieval failure or unavailable evidence as evidence of a concern. When the available evidence is insufficient to establish the corresponding category criterion, do not report
the concern.
</evidence_tracing>
<output_format>
Report all supported concerns. For each concern, return:
1. **Category ID** (C1–C11)
2. **Nature** (N1 or N2)
3. **Proposition** — a concrete, falsifiable statement describing what is inconsistent or missing
4. **Evidence** — the evidence anchor required by the corresponding category
5. **Reasoning** — a concise explanation of how the evidence establishes the concern
</output_format>
<rules>
- Report only concerns supported by the survey, cited literature, or explicitly retrieved literature.
- Evidence must be copied verbatim from the inspected source whenever a source span is available. Do not use a paraphrase as evidence.
- For existence concerns, an absence claim must specify the inspected scope rather than asserting that an element is missing from an unspecified portion of the survey.
- Report multiple distinct concerns when they concern genuinely different problems.
- For C7–C9, do not infer that every possible comparison, synthesis, or future direction is required; report only omissions supported by the corresponding survey context or, for future-work
and coverage concerns, by relevant literature evidence.
- For C11, discovering a relevant paper or community is not by itself sufficient; the available evidence must establish that the material is relevant to the survey and insufficiently represented.
- Do not report N3 evaluative concerns unless they can be concretely instantiated as C1–C11.
- Do not provide a summary, praise, neutral comments, scores, holistic judgments, or accept/reject recommendations.
</rules>
Return exactly one top-level field in the following JSON format and no text outside the JSON object.
{{
"reviewer_concerns": [
{{
"category": "C1|C2|C3|C4|C5|C6|C7|C8|C9|C10|C11",
"nature": "N1|N2",
"description": "<concise problem description>",
"evidence": [
{{
"verbatim": "<exact evidence passage>",
"location": "<section/paragraph/sentence or retrieval source>",
"role": "<what this passage establishes>"
}}
],
"reasoning": "<concise reasoning from the evidence to the concern>"
}}
]

}}
```

Table 25: User prompt for the algorithm-guided ablation setting.

Role: User prompt for the "–ConTax" ablation setting.

```
{SURVEY_FULL_TEXT}
<input_processing_note>
Ignore formatting artifacts, missing rendered tables/figures, reference formatting issues, and placeholder metadata like "Author names". Do not report these as reviewer concerns.
The survey's written date is {publication_date}. Use this date as a hard temporal boundary: do not recommend any work published after this date as a "missing" reference, and do not flag as
insufficient coverage any topic community whose publications postdate the survey.
</input_processing_note>
Requirements:
- Inspect the entire survey and report as many distinct, concrete problems as the evidence supports. Do not cap the number of concerns.
- Focus only on problems. Do not provide a summary, strengths, neutral comments, a score, an overall-quality judgment, or an accept/reject recommendation.
- Every reviewer concern must include supporting evidence.
- Evidence must be copied verbatim from the source. Do NOT paraphrase.
- The evidence field must contain all evidence you used to justify that concern, not just a summary.
- Evidence may come from the survey text, the survey's cited sources, or external literature that you explicitly retrieved for this judgment.
- If a concern depends on external literature or domain knowledge, state the exact external claim you relied on and name its source in the evidence field.
- Do not add any category labels, taxonomy labels, verifiability labels, or other classification traces.
- Do not report an unsupported suspicion as a concern.
Return exactly one top-level field in the following JSON format. Do not include any text outside the JSON object.
{{
"reviewer_concerns": [
{{
"issue": "<concise description of the problem>",
"location": "<section/paragraph/sentence where this occurs, quoted or paraphrased>",
"evidence": [
{{
"verbatim": "<exact evidence passage>",
"location": "<where this evidence came from>",
"role": "<what this passage establishes>"
}}
]
}}
]

}}
```

Table 26: User prompt for the evidence-tracing ablation setting.

Role: Parse section topics and reviewed objects.

```
### Task
You will read one section of a survey paper. Extract:
(1) research objects - named or explicitly identified methods, models, datasets, benchmarks, frameworks, systems, or tasks from the literature that this section discusses
(2) the topic(s) this section is organized around, each anchored to verbatim textual evidence
(3) which topic each object belongs to, anchored to verbatim textual evidence
(4) whether each object is a concrete named artifact that a later module should search for its original paper
### Topics
Use a list of topic objects. Each topic object must contain:
- label: copied verbatim from the text, not invented.
- anchor_type: one of section_title, text_span, or inferred.
- evidence_span: exact substring from section_text that anchors the topic, or null when the section title alone is the anchor.
Default: the section's own title is one topic, with anchor_type = section_title and evidence_span = null.
Only add additional topics if the text contains an explicit sentence that itself names sub-categories AND the rest of the section actually follows that split. For each such topic, set anchor_type
= text_span and evidence_span to the exact clause that states the defining criterion.
### Object types and search eligibility
For every object, set object_type to exactly one of: named_method, named_model, named_dataset, named_benchmark, named_framework, named_system, named_task, method_family, metric, concept,
application, other.
Set entity_candidate=true only for a concrete, individually identifiable artifact that a later module could search for its original or defining paper. This normally includes a proper name,
acronym, named model, named dataset, named benchmark, named framework, or explicitly named method/system. You may keep an unfamiliar proper name as entity_candidate=true; you do not need to
know what it is.
Set entity_candidate=false for broad research paradigms, method families, task categories, metrics, properties, goals, principles, applications, or ordinary noun phrases. Examples that must
be false include reinforcement learning, representation learning, embodied control, success rate, FID, object search, task planning, CLIP embeddings, uncertainty, and evaluation.
Do not convert a generic phrase into a named artifact. If a sentence lists several areas or capabilities without naming a specific artifact, either omit them or mark them with the appropriate
non-entity object_type and entity_candidate=false.
### Assigning objects to topics
For each object, only assign it to a specific sub-topic if you can quote a verbatim span from section_text that directly ties that object to that sub-topic. If no such direct textual link
exists, assign the object to the default section topic.
### Output format (JSON only, no other text)
{{
"topics": [
{{"label": "...", "anchor_type": "section_title|text_span|inferred", "evidence_span": "..." or null}}
],
"objects": [
{{
"citation_keys": ["..."],
"name": "..." or null,
"topics": ["..."],
"evidence_span": "..." or null,
"object_type": "named_method|named_model|named_dataset|named_benchmark|named_framework|named_system|named_task|method_family|metric|concept|application|other",
"entity_candidate": true
}}
]
}}
### Input
Paper title: "{paper_title}"
Section title: "{section_title}"
Section text:
"{section_text}"

Output:
```

Table 27: Prompt for section-level topic and research-object parsing.

Role: Extract survey self-claims, literature-selection statements, comparisons, syntheses, and future-work statements; the latter four functions supply C4–C9.

```
### Task
You are an academic survey-structure analyst. Read the complete paragraph below and independently extract every expressed instance of the five requested functions. A paragraph may contain several functions, and one function may span several items; preserve all relevant item_ids in one match. Do not extract SUMMARY or BACKGROUND unless the passage also performs SYNTHESIS. Do not infer an instance that is not expressed.
### Evidence and identifiers
Every match must contain item_id as an array of one or more exact input identifiers. Every match must contain verbatim_evidence copied exactly from the supplied items; do not paraphrase, normalize, or return only a trigger phrase. For a multi-item match, join the copied evidence with a newline. The evidence must be sufficient for an independent reader to inspect the decision.
### Functions and fields
1. SURVEY_SELF_CLAIM: a statement about what the current survey presents or how its content is organized. This unifies document-level CONTRIBUTION and component-level TEXTUAL. Set claim_kind to "coverage" for statements such as "This survey covers X", to "organization" for statements such as "The survey is structured around three axes" or "This review is organized from foundations to applications", and to "component_function" for statements such as "Section 3 discusses X". Use scope_type="document" for the first two and scope_type="component" for the last. For document claims return promised_scope; for component claims return component, component_target, and function. Do not extract a statement about a prior work. Do not treat a table/figure/equation as a checkable component; those remain out of scope downstream.
2. SELECTION: a statement about how this survey selects, retrieves, filters, includes, excludes, or organizes the literature or reviewed work set. Return polarity, operation, source, boundary, and criterion. It must describe the survey's own literature-selection process or boundary. Do not label ordinary topic coverage, a research-object property, or a statement that a section discusses a topic as SELECTION.
3. COMPARISON: a substantive comparison that states how two or more research objects differ, relate, or perform along an explicit dimension. Return comparison_targets, comparison_dimensions, and comparison_relation. Mere juxtaposition, historical change, the word contrast, or a sentence mentioning a comparative experiment without stating its relation is not COMPARISON.
4. SYNTHESIS: an integrative or interpretive statement that combines multiple prior works, methods, topics, results, or background facts into a higher-level relation or insight. Return synthesis_type, integrated_objects, and insight. integrated_objects must contain only the objects that participate in the integration, not every noun phrase. Valid synthesis_type values include trend, tradeoff, taxonomy, generalization, evaluation, mechanism, field_framing, and other. A single-work summary, a list of facts, or ordinary background is not SYNTHESIS.
5. FUTURE_WORK: an explicit prospective direction or open problem for future research. Extract it only when the sentence proposes what researchers should develop, study, evaluate, or resolve next, or states that a field, subfield, task, or topic-level question remains unresolved as a research agenda. Do not infer FUTURE_WORK merely because one or more existing studies have limitations. A criticism of one cited work used to motivate another work is not FUTURE_WORK. A factual limitation of one method, a historical observation, a present-tense challenge, or a deployment recommendation is not FUTURE_WORK unless explicitly framed as future research. Return future_work_scope, future_work_type, and target.
### Positive examples
- SURVEY_SELF_CLAIM
Input: "This review is organized from foundational concepts through applications."
Output: {"label":"SURVEY_SELF_CLAIM","item_id":["ex-o1"],"claim_kind":"organization","scope_type":"document","promised_scope":"organizes the review from foundational concepts through applications"}
- SURVEY_SELF_CLAIM
Input: "This survey covers multilingual benchmarks for evaluating factual consistency in generated summaries."
Output: {"label":"SURVEY_SELF_CLAIM","item_id":["ex-c1"],"verbatim_evidence":"This survey covers multilingual benchmarks for evaluating factual consistency in generated summaries.","claim_kind":"coverage","scope_type":"document","promised_scope":"covers multilingual benchmarks for evaluating factual consistency in generated summaries"}
- SURVEY_SELF_CLAIM
Input: "Section 3 reviews graph-level tasks."
Output: {"label":"SURVEY_SELF_CLAIM","item_id":["ex-t1"],"verbatim_evidence":"Section 3 reviews graph-level tasks.","claim_kind":"component_function","scope_type":"component","component":"section","component_target":"Section 3","function":"reviews graph-level tasks"}
- SELECTION
Input: "We include studies published between 2018 and 2024 retrieved from ACL Anthology."
Output: {"label":"SELECTION","item_id":["ex-s1"],"verbatim_evidence":"We include studies published between 2018 and 2024 retrieved from ACL Anthology.","polarity":"inclusion","operation":"retrieve/include","source":"ACL Anthology","boundary":"studies published between 2018 and 2024","criterion":""}
- SYNTHESIS
Input: "Across retrieval-based and generative systems, greater response diversity improves engagement but makes factual verification more difficult."
Output: {"label":"SYNTHESIS","item_id":["ex-y1"],"verbatim_evidence":"Across retrieval-based and generative systems, greater response diversity improves engagement but makes factual verification more difficult.","synthesis_type":"tradeoff","integrated_objects":["retrieval-based systems","generative systems","response diversity","user engagement","factual verification"],"insight":"Greater response diversity can improve engagement while making factual verification more difficult."}
- COMPARISON
Input: "Retrieval-based systems are more controllable than generative systems, but they provide less diverse responses."
Output: {"label":"COMPARISON","item_id":["ex-m1"],"verbatim_evidence":"Retrieval-based systems are more controllable than generative systems, but they provide less diverse responses.","comparison_targets":["retrieval-based systems","generative systems"],"comparison_dimensions":["controllability","response diversity"],"comparison_relation":"Retrieval-based systems are more controllable, whereas generative systems provide more diverse responses."}
- FUTURE_WORK
Input: "Future work should develop evaluation protocols for long-term user adaptation in conversational agents."
Output: {"label":"FUTURE_WORK","item_id":["ex-f1"],"verbatim_evidence":"Future work should develop evaluation protocols for long-term user adaptation in conversational agents.","future_work_scope":"field","future_work_type":"future_direction","target":"evaluation protocols for long-term user adaptation"}
### Negative examples
- Do not label "Method A cannot handle long sequences, so Method B introduces recurrence" as FUTURE_WORK. This is a prior-work critique and method motivation, not a future research direction.
- Do not label "Task-oriented systems operate in bounded domains" as SELECTION. Bounded domains describe the research object, not the survey's literature selection.
- Do not label "Model X fails to preserve long-range coherence" as FUTURE_WORK unless the passage explicitly proposes future research to address it.
- Do not label "Table 4 presents a taxonomy" as SURVEY_SELF_CLAIM unless the passage states that the current survey presents or organizes that taxonomy.
- Do not label "Our taxonomy is more elegant than prior taxonomies" as SURVEY_SELF_CLAIM. This is an uncheckable quality claim, not a coverage or organization declaration.
- Do not label "Early systems used pattern matching without semantic understanding" as FUTURE_WORK. This is a description of a prior work.
### Input
{paragraph}
### Output
Return JSON only:
{
"matches": [
{
"label": "SURVEY_SELF_CLAIM|SELECTION|COMPARISON|SYNTHESIS|FUTURE_WORK",
"item_id": ["p0-s0"],
"verbatim_evidence": "...",
"...label-specific fields...": "..."
}
]
}
Return an empty matches array when no requested function is expressed. Do not include source_text because the input paragraph is already stored by its parent document.
```

Table 28: Prompt for paragraph-level survey-function extraction.

Role: Classify section functions for downstream scope control.

```
### Task
Classify the rhetorical function of one section of an academic survey. Use the section title, its position in the document, its parent section type, and its opening text. A subsection normally inherits the rhetorical context of its parent, but may receive a different label when its own purpose clearly differs.
### Labels
- SCOPE: the survey's literature search, inclusion/exclusion, filtering, selection, or coverage boundary.
- BACKGROUND: prerequisites, definitions, motivation, or field context that prepares the reader; an Introduction subsection remains BACKGROUND unless it actually reviews a research area.
- CONTENT: organized review of specific methods, systems, datasets, tasks, or research subareas.
- FUTURE_WORK: open problems, limitations of the field, or future research directions.
- CONCLUSION: a closing synthesis of the survey without substantial new review content.
Do not label a section CONTENT merely because its title is technical. In an Introduction or Background branch, use BACKGROUND unless the opening text shows that the section systematically reviews prior work. Do not infer SCOPE from a topic boundary such as "bounded domains"; SCOPE requires a boundary of this survey's literature selection.
### Input
Document title: "{DOCUMENT_TITLE}"
Section path: "{SECTION_PATH}"
Parent type: "{PARENT_TYPE}"
Section title: "{SECTION_TITLE}"
Opening text: "{PREAMBLE}"
### Output

Return JSON only: {{"functional_type": "SCOPE|BACKGROUND|CONTENT|FUTURE_WORK|CONCLUSION", "confidence": 0.0}}
```

Table 29: Prompt for hierarchical section-function classification.

Role: Filter candidate papers for the evaluation literature pool.

```
### Task
You are filtering candidate papers for a survey evaluation literature pool. Given the survey topic and a batch of candidate titles and abstracts, decide whether each candidate is substantively related to the survey topic. Mark relevant=true only if the paper studies the same research topic, a direct subtopic, a method family, a task, a dataset/benchmark, or an application branch that a survey on this topic could reasonably cover. Reject candidates that only share generic words, belong to a different sense of an ambiguous phrase, or are merely broad background. For each relevant candidate, set verbatim_evidence to the candidate's complete JSON title field value, copied exactly as written. Do not use the abstract for evidence. Do not paraphrase, translate, normalize capitalization, alter punctuation, alter hyphenation, or add ellipses. For each irrelevant candidate, set verbatim_evidence to an empty string.
### Output format (JSON only, no other text)
{{
 "papers": [
{{"paper_id": "...", "relevant": true, "verbatim_evidence": "..."}}
]
}}
### Input
Survey topic: "{query}"
Candidate papers:
{papers}

Output:
```

Table 30: Prompt for literature-pool relevance filtering.

Role: C1: extract reusable anchors for symbols, terminology, definitions, assumptions, and quantities.

```
### Task
Extract only small, reusable internal anchors that could be contradicted elsewhere in this survey. The allowed categories are: symbol, terminology, definition, assumption, and quantity.
Create a note only when the paragraph explicitly introduces or fixes an anchor:
- symbol: notation and what it denotes, such as "G=(V,E)" or "L" for a loss;
- terminology: an explicit naming or abbreviation relation;
- definition: what a named concept means in this paper;
- assumption: an explicit condition required by a method or argument;
- quantity: a numeric value tied to a named entity, metric, dataset, count, or complexity expression.
Do not extract ordinary scientific facts, method descriptions, results, motivations, conclusions, topic coverage, section summaries, literature summaries, or future work. Those belong to other checks. Do not make one note for every sentence. A paragraph normally yields zero, one, or a few notes. Extract only anchors that would matter if the same symbol, term, definition, assumption, or quantity were later reused inconsistently. Copy the complete supporting span verbatim; do not paraphrase or invent text. Return JSON only:
{{"notes":[{{"note_id":"N1","category":"symbol|definition|terminology|assumption|quantity","subject":"...","predicate":"...","object":"...","verbatim_evidence":"..."}}]}}.
Paragraph:
{paragraph}
```

Table 31: Prompt for extracting internal-consistency notes.

Role: C1: identify conflicts within the complete note set.

```
### Task
Inspect the supplied reading notes from one survey paper for small, directly checkable internal inconsistencies. Report only conflicts supported by at least two notes and involving one of the five allowed objects: symbol, terminology, definition, assumption, or quantity.
Rules:
- symbol: the same notation denotes incompatible objects or roles.
- terminology: the same term or abbreviation is explicitly used for incompatible concepts. Normal singular/plural, capitalization, hyphenation, or other ordinary grammatical variants are consistent unless the paper explicitly assigns them different meanings.
- definition: the paper gives incompatible definitions for the same named concept, or uses a concept as if defined when its own definition conflicts with that use.
- assumption: the same argument or result is presented with incompatible prerequisite conditions.
- quantity: the same named quantity, count, unit, or numerical value is inconsistent in comparable contexts.
- Do not report ordinary scientific facts, method descriptions, results, section scope, contribution claims, prose quality, redundancy, figure/table presentation, or taxonomy preferences.
- Do not infer a conflict from different topics, a refinement, a qualification, a broader/narrower wording, or a mere stylistic variant.
- Do not treat a mathematically incorrect formula or derivation as an internal conflict; that requires external technical verification unless the issue is only an internally inconsistent symbol or definition.
Copy each evidence span verbatim from the supplied notes. Return JSON only: {{"conflicts":[{{"note_ids":["N1","N2"],"conflict_type":"symbol|definition|terminology|assumption|quantity","explanation":"...","evidence":["...","..."]}}]}}.
Reading notes:
{notes}
```

Table 32: Prompt for unified internal-conflict verification.

Role: C1: identify conflicts between current and earlier notes in serial processing.

```
### Task
Inspect the CURRENT reading notes together with PRIOR notes for small, literal or shallow internal inconsistencies. Report only conflicts supported by at least two notes and involving symbol, terminology, definition, assumption, or quantity.
Normal singular/plural, capitalization, hyphenation, and grammatical variants are consistent unless the paper explicitly gives them different meanings. Do not report ordinary descriptions, method/result differences, section scope, contribution claims, refinements, qualifications, different topics, presentation preferences, or mathematical errors that require external technical verification. A formula may be reported only when its symbols or definitions conflict internally. The current notes are marked CURRENT and earlier notes are marked PRIOR. Copy every evidence span verbatim from the supplied notes. Return JSON only:
{{"conflicts":[{{"note_ids":["N1","N2"],"conflict_type":"symbol|definition|terminology|assumption|quantity","explanation":"...","evidence":["...","..."]}}]}}.
Earlier notes:
{prior_notes}
Current notes:
{current_notes}
```

Table 33: Prompt for incremental internal-conflict verification.

Role: C1: retain only checkable survey coverage and organization declarations.

```
### Task
Decide whether one extracted SURVEY_SELF_CLAIM item is a checkable declaration about the current survey's coverage or organization.
Return exactly one decision:
- KEEP_SECTION: a concrete statement that the survey covers, discusses, examines, or devotes a named topic to a paper section. Check it with the complete heading outline.
- KEEP_ORGANIZATION: a concrete statement about how the survey is organized, such as a taxonomy, progression, or set of organizing axes. Check it with the complete heading outline; do not judge taxonomy quality.
- KEEP_SENTENCE: an explicit current-survey statement about comparison, synthesis, survey methodology, or future work, with a concrete target. Check it only with extracted COMPARISON, SYNTHESIS, SCOPE, or FUTURE_WORK items.
- REJECT: motivation, background, a research gap, an open problem, a conclusion, a quality claim, a summary of literature, or a statement about a cited work.
- NOT_CHECKED: cited claims, figures/tables/equations, or an unresolved/ambiguous paper component.
Be conservative. Do not split one item into subclaims. A coordinated list remains one item. Reject examples: "our analysis reveals several insights", "existing methods neglect safety", "this work proposes a framework", "the paper identifies a gap", and "world models address grounding limitations". Accept examples: "This survey covers video transformers", "Section 3 reviews graph-level tasks", "We compare the surveyed approaches", and "The survey is organized around three taxonomic axes". The words "this paper", "this work", or "the paper" do not identify the current survey; reject them unless the sentence explicitly says "this survey" or "this review".
Input:
claim_label={claim_label}
source_sentence={source_sentence}
paragraph_context={paragraph_context}
component={component}
component_target={component_target}
function={function}
Return JSON only:

{{"decision":"KEEP_SECTION|KEEP_ORGANIZATION|KEEP_SENTENCE|REJECT|NOT_CHECKED","reason":"brief reason"}}
```

Table 34: Prompt for filtering survey self-claims.

Role: C1: verify grouped survey self-claims against their complete inspection scope.

```
### Task
Judge each numbered coverage declaration against the supplied current-survey material. Each declaration is one indivisible item; do not split it and do not infer obligations that are not stated.
KEEP_SECTION and KEEP_ORGANIZATION declarations are checked only against the complete heading outline. KEEP_SENTENCE declarations are checked only against the supplied typed function items. A
declaration is FULFILLED when the supplied material clearly establishes the declared coverage, organization, or component/function. Otherwise it is NOT_FULFILLED. Use NOT_CHECKED only when the
supplied material is structurally unusable. Do not judge motivation, research gaps, future suggestions, literature conclusions, taxonomy quality, or the quality of the survey.
For every NOT_FULFILLED item, the supplied candidate list is the complete inspection scope. Return all candidate item IDs that materially support a positive decision; for a negative decision
return an empty list. Reasoning must explain the match or the absence without relying on material outside this request.
Claims:
{claims}
Complete heading outline:
{outline}
Typed function items:
{function_items}
Return JSON only:

{{"results":[{{"claim_id":"...","verdict":"FULFILLED|NOT_FULFILLED|NOT_CHECKED","supporting_item_ids":["..."],"reasoning":"..."}}]}}
```

Table 35: Prompt for survey self-claim consistency verification.

Role: C2: recover a cited-work title when structured bibliography parsing fails.

```
# Task
Extract the title of the cited work from the following reference string. Return only the title, enclosed in quotation marks. Return the following JSON object only:
{{
"title": "..."
}}
# Example
Input: Tianhe Lin, Jian Xie, Siyu Yuan, et al. (2025). Implicit Reasoning in Transformers is Reasoning through Shortcuts. Annual Meeting of the Association for Computational Linguistics.
Output:
{{
"title": "Implicit Reasoning in Transformers is Reasoning through Shortcuts"
}}
# Input
Input: {info}

Output:
```

Table 36: Prompt for cited-title extraction.

Role: C2: match academic web-search results to a cited title.

```
You are filtering academic web search results for a cited paper lookup.
Target paper title:
{title}
Candidates:
{candidates}
Return a JSON object only:
"'json
{{
"matched_indices": [1, 3],
"reason": "brief explanation"
}}
"'
Rules:
- keep only candidates that are very likely to refer to the same paper,
- prefer publisher, DOI, arXiv, OpenReview, ACL Anthology, Semantic Scholar, DBLP, or author pages,
- it is acceptable to return an empty list if the evidence is weak.
```

Table 37: Prompt for cited-source search-result matching.

Role: C2–C3: decompose cited and uncited sentences into atomic claims.

```
You are a precise claim extractor for citation verification. Extract minimal, independently verifiable claims from the sentences marked <E id="...">...</E> below. Unmarked sentences are context
only - never extract claims from them, use them only to resolve references inside <E> sentences.
### INPUT
"""{range}"""
### RULES
1. Atomicity: split "and"-joined coordinate predicates/objects sharing one citation into separate claims, one per item - UNLESS the joint statement only makes sense together (e.g., "X and Y
jointly demonstrated Z"). Do NOT split "or"-joined items; keep them as one claim (splitting would turn "at least one holds" into two independently-required claims).
2. Citation attachment:
- Multiple markers supporting the same statement stay in one claim.
- A pronoun/implicit reference inherits its antecedent's citation key(s); substitute the resolved entity name for the pronoun.
- No citation, no valid antecedent to inherit from -citation_keys = [].
3. Source tracking: every claim must list the id(s) of every <E> sentence it draws content from - the sentence containing the predicate, plus (if a pronoun/implicit reference was resolved) the
sentence containing the antecedent. A claim built entirely from one sentence has one id.
4. Ambiguity flag - mark "verifiable": false if ANY apply:
- The reference (pronoun/implicit mention) cannot be resolved to a specific antecedent within the given text.
- It is a categorical/definitional statement about a class of things, backed by multiple citations, not specific to any single cited work.
- It is a catch-all enumeration tail with no specific referent (e.g., "...and other applications").
- It only restates/continues a prior statement without adding a new checkable detail.
- Its subject is a generic technique/design pattern/paradigm rather than one identifiable named entity (paper, method, model, dataset, system).
Otherwise mark "verifiable": true.
5. Skip pure transitions/hedges with no checkable statement at all (e.g., "this remains an active research area") - do not output these as claims.
6. Verbatim constraint: every word must come from the text, except resolved entity names substituted for pronouns.
### OUTPUT (JSON only)
{{
"claims": [
{{"claim": "...", "verifiable": true|false, "citation_keys": ["..."], "source_ids": ["..."]}}
]
}}
### EXAMPLE 1 - and/or splitting, cross-sentence coreference, generic-subject ambiguity
"""<E id="1">[12] introduced GraphSAGE, an inductive framework for learning node embeddings on large graphs.</E> <E id="2">It achieves strong results on node classification and link prediction.</E>
<E id="3">Sinusoidal or learned time embeddings can be used for time conditioning [8].</E>"""
Output:
{{
"claims": [
{{"claim": "[12] introduced GraphSAGE, an inductive framework for learning node embeddings on large graphs.", "verifiable": true, "citation_keys": ["12"], "source_ids": ["1"]}},
{{"claim": "GraphSAGE achieves strong results on node classification.", "verifiable": true, "citation_keys": ["12"], "source_ids": ["1", "2"]}},
{{"claim": "GraphSAGE achieves strong results on link prediction.", "verifiable": true, "citation_keys": ["12"], "source_ids": ["1", "2"]}},
{{"claim": "Sinusoidal or learned time embeddings can be used for time conditioning.", "verifiable": false, "citation_keys": ["8"], "source_ids": ["3"]}}
]
}}
### EXAMPLE 2 - unresolved reference, categorical statement, enumeration tail
"""He proposed a two-stage training approach for it, achieving strong performance on the benchmark [44]. <E id="4">Large Language Models (LLMs) [19, 91, 255] are advanced language models with
massive parameter sizes.</E> <E id="5">This has enabled progress on visual question answering, and other multimodal applications.</E>"""
Output:
{{
"claims": [
{{"claim": "Large Language Models are advanced language models with massive parameter sizes.", "verifiable": false, "citation_keys": ["19", "91", "255"], "source_ids": ["4"]}},
{{"claim": "Large Language Models have enabled progress on visual question answering.", "verifiable": true, "citation_keys": ["19", "91", "255"], "source_ids": ["5"]}},
{{"claim": "Large Language Models have enabled progress on other multimodal applications.", "verifiable": false, "citation_keys": ["19", "91", "255"], "source_ids": ["5"]}}
]
}}
CHECK before output: every claim has a non-empty source pointing to <E> sentence id(s) actually used, "or" was not split while "and" was, every unresolved reference is flagged verifiable:false
rather than guessed, and no pure transition/hedge sentence was extracted.
```

Table 38: Prompt for atomic-claim extraction.

Role: C3: rewrite an uncited atomic claim as an academic-search query.

```
You are a query rewriting assistant for an academic literature search system. You will be given an Atomic Fact (AF) - a single claim extracted from a survey paper that currently has no citation.
Rewrite it into a clean keyword query for an academic database (OpenAlex), so we can find candidate papers that might support or refute this claim.
## Task
Step 1 - Assess searchability:
Mark `searchable: false` if the claim:
- contains no specific method/dataset/task entity (i.e., a generic statement that could apply to many works)
- is a meta-statement about the survey itself, not about prior work
- is a value judgement with no factual anchor (e.g., "this remains an important direction")
Step 2 - If searchable, construct the query:
- 3-8 words, keyword-style (NOT a natural language question or full sentence)
- The query MUST be composed only of words/phrases taken verbatim from the claim. Do NOT introduce synonyms, paraphrases, or new terms not present in the claim text.
- No punctuation except hyphens within compound terms (e.g., "few-shot")
- No stopwords, no filler verbs ("shows that", "demonstrates")
- Prioritize the most specific/rare entity first (a specific model or dataset name outranks a generic task name)
- If the claim compares two things, include both entities
- Do NOT include the survey's own framing language (e.g., "recent work")
- `key_entities` MUST be a subset of the words/phrases actually used in the query. Do not list any entity that does not appear in the query string.
## Output format (JSON only, no other text):
{{
"searchable": true/false,
"reason_if_not_searchable": "<brief reason, empty string if searchable>",
"query": "<the search query, empty string if not searchable>",
"key_entities": ["<entity1>", "<entity2>", ...]
}}
## Examples
### Example 1 (searchable, single entity)
AF: "Speculative decoding techniques reduce inference latency by over 40% compared to standard transformer decoding."
Output:
{{
"searchable": true,
"reason_if_not_searchable": "",
"query": "speculative decoding transformer inference latency",
"key_entities": ["speculative decoding", "transformer inference latency"]
}}
### Example 2 (searchable, comparison)
AF: "Symbolic approaches tend to generalize better to unseen compositional structures compared to neural approaches."
Output:
{{
"searchable": true,
"reason_if_not_searchable": "",
"query": "symbolic approaches compositional structures neural approaches",
"key_entities": ["symbolic approaches", "compositional structures", "neural approaches"]
}}
### Example 3 (not searchable)
AF: "Robustness to distribution shift in low-resource settings remains a challenging open problem."
Output:
{{
"searchable": false,
"reason_if_not_searchable": "General characterization of a research gap, not a specific factual claim attributable to a particular finding",
"query": "",
"key_entities": []
}}
Now process the following AF:
AF: {af_text}
```

Table 39: Prompt for technical-claim query construction.

Role: C2–C3: determine whether retrieved source material supports, refutes, or cannot verify a claim.

```
You are a factual correctness verifier for academic surveys. Given:
- A claim extracted from a survey, and
- The paper that it cites (including {content_type})
Determine whether the claim is supported by the cited paper. Your judgment should be one of the following:
- SUPPORTED: the claim is clearly supported by the evidence.
- REFUTED: the claim is clearly contradicted by the evidence.
- NEUTRAL: the claim is not mentioned in the evidence, or there's no sufficient information to verify if the claim is supported or refuted.
**Important:** If your judgment is "SUPPORTED" or "REFUTED", you MUST provide verbatim evidence from the content of the cited paper to support that.
Use a strict contradiction standard:
- Mark REFUTED only when the evidence explicitly contradicts the core factual content of the claim.
- Do not mark REFUTED merely because the evidence uses different terminology, a broader/narrower formulation, or an implementation-level description that is compatible with the claim.
- For method descriptions, treat high-level descriptions and implementation descriptions as compatible unless they cannot both be true.
- If the evidence is related but does not clearly support or clearly contradict the claim, choose NEUTRAL.
- If you mark REFUTED, include a concise contradiction_reason explaining why the claim and evidence cannot both be true.
Your output should be a single JSON object only:
"'json
{{
"judgment": "SUPPORTED" | "REFUTED" | "NEUTRAL",
"evidence": "verbatim evidence from the cited paper, if judgment == SUPPORTED or REFUTED" | "" (if judgment == NEUTRAL),
"contradiction_reason": "why the claim and evidence cannot both be true, if judgment == REFUTED" | ""
}}
"'
### Claim
{claim}
### Evidence
{text}
```

Table 40: Prompt for source-grounded factual verification.

Role: C6: determine whether a high-risk uncited claim lacks any internal support trace.

```
### Task
Extract whether this uncited, falsifiable empirical, theoretical, comparative, causal, or quantified claim has any direct internal support in the supplied survey span. This is a zero-support
check, not an assessment of whether the explanation is complete, rigorous, persuasive, or appropriate for a particular reader. Any directly connected observation, reported result, example,
analysis, cited observation explained by the span, or explicit explanatory link counts as support. A neighboring citation counts when the span explains what it supports. Do not treat
recommendations, future work, task definitions, section roadmaps, or strong wording alone as failures. Do not verify external factual truth; that belongs to fact checking.
Survey topic: {topic}
Candidate claim: {claim}
Search span ({search_scope}): {support_span}
### Decision
Return ARGUMENT_SUPPORT_FAILURE only when the claim is genuinely falsifiable and the inspected span contains no connected internal support at all. If even one plausible support span exists, or
the claim type or support status is ambiguous, return NO_COMMENT. Do not use external literature or unstated domain knowledge. Do not decide that a claim deserves more detail merely because
more detail could be useful.
### Evidence rule
For ARGUMENT_SUPPORT_FAILURE, claim_quote must copy the claim verbatim and support_quote must be an empty string. For NO_COMMENT, support_quote must copy the supporting bridge verbatim from
the search span. Do not paraphrase either field.
### Output JSON only
{{
"problem_type": "ARGUMENT_SUPPORT_FAILURE" | "NO_COMMENT",
"comment": "one concise reviewer-facing finding, or empty string for NO_COMMENT",
"claim_quote": "verbatim claim quote for a failure, otherwise empty string",
"support_quote": "verbatim reasoning bridge for NO_COMMENT, otherwise empty string",
"reasoning": "why the span does or does not supply the required bridge",
"confidence": 0.0

}}
```

Table 41: Prompt for zero-support claim verification.

Role: C10: select prior surveys suitable for reference anchoring.

```
You are a professional academic researcher selecting reference surveys to evaluate a target survey titled "{query}".
### Input
You are given a list of candidate papers with their titles and abstracts.
### Definition: Reference Survey
A **reference survey** is a field-level survey that:
- Treats the query topic as its PRIMARY organizing focus, not as a tool or method applied within a different domain.
- Organizes the literature into coherent conceptual or methodological dimensions (e.g., taxonomies, categorizations, design spaces).
- Covers multiple distinct sub-topics within the query field.
- Would be consulted by an expert to judge whether another survey on this topic has missed important topics or references.
### Inclusion Criteria for Reference Surveys (ALL must be satisfied)
1. The query topic is the main research object AND the primary organizing principle of the paper.
2. At least THREE distinct sub-topics within the query field are covered.
3. The paper synthesizes existing literature rather than reporting original experimental results.
### Exclusion Criteria (ANY triggers exclusion from BOTH tiers)
- Not a survey: excludes benchmarks, position papers, tutorials, or original research papers.
- Primary subject is a downstream application domain, with the query topic appearing only as the method used (e.g., "Transformers for Medical Imaging" is excluded when evaluating a survey on
Transformers).
- Covers only ONE task or sub-area within the query field.
- Mentions the query topic only as background or one method among many.
### Classification Instructions
- First apply exclusion criteria. If any exclusion criterion is met, discard the candidate entirely.
- For remaining candidates, apply inclusion criteria. If ALL inclusion criteria are satisfied → reference survey. Otherwise → discard.
- Select at most 3 surveys in total.
- Be conservative: fewer is better than including a marginal candidate.
- If no candidate meets reference survey criteria, return an empty list.
### Candidate Surveys
{candidates}
### Output Format
Return JSON only, no extra text.
{{
"reference_surveys": [
{{
"title": "...",
"covered_subtopics": ["subtopic1", "subtopic2", "subtopic3"],
"reason": "one sentence explaining why this qualifies as a reference survey"
}}
]
}}
```

Table 42: Prompt for selecting reference surveys.

Role: C10: determine whether a retrieved paper is the original source of an uncited named artifact.

```
### Task
You are a strict academic verifier. Determine whether each candidate paper is the original source of a named research artifact (model, dataset, method, benchmark, metric, task, concept, or
framework).
Candidate artifact: {entity_name}
Candidate abstracts:
{candidate_papers}
### Decision rules
Judge yes only when the abstract explicitly says that the paper proposes, introduces, presents, names, defines, releases, or creates the artifact. Ignore titles completely: title occurrence is
not evidence of authorship. Judge no when the artifact is only used, compared, evaluated, or mentioned as prior work. Return uncertain when the abstract lacks enough information. For every yes
decision, copy the proposing sentence verbatim from the abstract; do not paraphrase.
### Field guide
* `entity`: copy the named artifact exactly as given in the input.
* `results`: return one entry per candidate paper, in the same order as the candidate abstracts. Each entry must include `paper_index`, `decision`, `matched_rule`, `evidence`, and `confidence`.
* `decision`: use `yes` only for the original source, `no` for clear non-sources, and `uncertain` when the abstract does not say enough.
* `matched_rule`: use `abstract` for explicit source claims, `exclusion` for explicit non-source claims, and `none` when the abstract is too weak to decide.
* `evidence`: for `yes`, copy the exact proposing sentence from the abstract; for `no` or `uncertain`, leave it empty.
* `confidence`: choose `high`, `medium`, or `low` according to how explicit the abstract is.
* `most_likely_source`: choose the index of the strongest `yes` decision, or null if there is no `yes`.
### Output JSON only
{{
"results": [
{{
"paper_index": 1,
"decision": "yes" | "no" | "uncertain",
"matched_rule": "abstract" | "exclusion" | "none",
"evidence": "verbatim abstract quote or empty string",
"confidence": "high" | "medium" | "low"
}}, ...
],
"most_likely_source": <integer> or null
}}

Use null whenever no paper is judged yes. Otherwise return the index of the strongest yes decision.
```

Table 43: Prompt for uncited-entity source verification.

Role: C11: determine the concrete citation role of a graph-retrieved candidate.

```
### Task
Judge whether each candidate paper is substantively relevant to the supplied survey subtopic and therefore merits inclusion in a short missing-reference shortlist. Use the subtopic, the survey evidence card, the graph evidence card, and each candidate abstract. Do not infer relevance from a title alone. A relevant candidate must directly study the subtopic, one of its named research objects, or a method/task/dataset central to it, and it must add semantic gain: filling a sparsely cited subtopic, updating an old subtopic, supplying an omitted core method, adding an important benchmark/dataset/evaluation object, or adding a necessary theoretical or tool foundation.
### Decision rules
1. Return ‘relevant=true‘ only when both sides are supported: the survey evidence card shows a concrete subtopic/gap, and the candidate abstract shows that the paper’s own contribution fits that subtopic/gap.
2. Treat citation-neighborhood evidence as retrieval provenance. It can explain why the candidate was inspected, but it cannot by itself justify relevance.
3. Prefer method, dataset, benchmark, evaluation, tool/system, and theoretical-foundation papers. Do not mark a survey/review/tutorial paper relevant unless the survey topic itself is about surveys or the supplied subtopic explicitly asks for survey papers.
4. Reject papers that are only broadly in the survey topic but do not match the supplied subtopic, papers whose contribution is only a downstream application not discussed by the subtopic, and papers whose abstract only mentions the subtopic as background.
5. A paper may fill ‘few_reference_gap‘ or ‘old_reference_gap‘ only when the survey evidence card states a small or old cited-paper set and the candidate abstract shows a concrete contribution that would improve that section.
6. A paper may fill ‘missing_core_method‘ only when the candidate abstract presents a method/object that is central to the supplied subtopic rather than a minor variant or generic application.
### Field guide
* ‘paper_id‘: copy the candidate paper id exactly as given.
* ‘relevant‘: set to true only when the abstract clearly supports the supplied subtopic and the paper adds a real semantic gain; otherwise set false.
* ‘matched_subtopic‘: copy the exact supplied subtopic name when relevant; leave it empty when irrelevant.
* ‘candidate_role‘: choose the paper’s role in the subtopic from ‘core_method‘, ‘supporting_method‘, ‘benchmark_dataset‘, ‘application_case‘, ‘survey_or_review‘, ‘tool_or_system‘, ‘theoretical_foundation‘, or ‘no_role‘. Use ‘no_role‘ for irrelevant papers.
* ‘gap_type‘: choose the gap this paper helps fill from ‘missing_core_method‘, ‘few_reference_gap‘, ‘old_reference_gap‘, ‘coverage_gap‘, ‘benchmark_gap‘, or ‘no_gap‘. Use ‘no_gap‘ for irrelevant papers.
* ‘survey_evidence‘: for a relevant paper, copy one short verbatim span from the survey evidence card that shows the matched subtopic or the gap being filled, such as a sentence about few cited papers, old cited papers, uncovered high-ranked papers, or the subtopic’s own discussion. Leave it empty when irrelevant.
* ‘verbatim_evidence‘: for a relevant paper, copy one short verbatim span from the candidate abstract that shows the candidate’s actual contribution, role, method, dataset, benchmark, tool, theory, or evaluation object. Leave it empty when irrelevant.
* ‘reason‘: give one brief human-readable reason explaining how the two evidence spans support the semantic gain. Do not restate the quotes and do not cite graph rank as the main reason.
### Evidence rules
For every relevant candidate, ‘survey_evidence‘ must be copied verbatim from the supplied survey evidence card, and ‘verbatim_evidence‘ must be copied verbatim from the supplied candidate abstract. Do not paraphrase, compress, translate, normalize punctuation, or mix text from multiple places. For every irrelevant candidate, leave both evidence fields empty, set ‘candidate_role‘ to ‘no_role‘, and set ‘gap_type‘ to ‘no_gap‘.
Survey topic: {query}
Subtopic: {topic}
Survey evidence card:
{survey_evidence}
Graph evidence card:
{graph_evidence}
Candidate papers:
{candidate_papers}
Graph evidence describes retrieval provenance only. It is not, by itself, evidence that a citation obligation exists. Do not copy graph evidence into either quoted evidence field.
### Output JSON only
{{
"papers": [
{{"paper_id": "...", "relevant": true, "matched_subtopic": "...", "candidate_role": "...", "gap_type": "...", "survey_evidence": "...", "verbatim_evidence": "...", "reason": "..."}}
]

}}
```

Table 44: Prompt for reference-anchor relevance verification.

Role: C11: classify a discovered literature community relative to existing survey topics.

```
### Task
Analyze the semantic relationship between one Leiden-discovered candidate community and the survey query and existing survey topics.
Survey query: {query}
Existing survey topics:
{existing_topics}
Representative community papers:
{papers}
### Semantic relations
* COVERED: The community is represented by one or more existing survey topics at a granularity that is reasonable given the survey query. A broader existing topic may cover a narrower community when the query itself is broad and the broader organization is a reasonable survey structure. However, do not treat a broad topic as automatically covering every distinct specialization within it. List every matching topic exactly in covered_topics.
* COHERENT_NEW_DIRECTION: The community is relevant to the survey query, forms a coherent and sufficiently distinct research direction, and is not represented by any existing survey topic at a granularity that is reasonable for the query. Give a concise community_name and explain the semantic evidence.
* MIXED: The representative papers clearly span two or more distinct research directions that would reasonably be treated as separate topics at a finer granularity. List these directions in mixed_topics and explain the semantic distinction.
* UNRELATED: The community does not fall within the intended scope of the survey query. Give a concrete reason.
### Decision rules
1. Judge the community from the supplied paper titles and abstracts only.
2. Treat the existing survey topics as semantic scope descriptions, not as lists of papers that must exactly match the community.
3. In this first pass, judge whether an existing topic is a plausible semantic match. A later verification pass will inspect the original subsection text for every COVERED decision.
4. Judge the semantic structure of the community at its current granularity. Do not infer or propose a different clustering.
5. Use MIXED when the representative papers clearly span multiple distinct directions. Do not use MIXED merely because papers differ in application, task, dataset, model, or implementation, when they share the same underlying research direction.
6. Use COHERENT_NEW_DIRECTION only when the community is both query-relevant and semantically distinct from all existing survey topics.
7. Do not infer information that is not supported by the supplied titles, abstracts, query, or topic labels.
### Field guide
* ‘semantic_relation‘: choose exactly one label that matches the community-level relation.
* ‘covered_topics‘: this is a closed-set field. Every item must be copied character-for-character from the supplied topic list. Do not paraphrase, abbreviate, translate, singularize, pluralize, reorder, or replace a word with a synonym. In particular, do not turn a supplied topic such as ‘LLMs and VLMs as Planners and Reasoners‘ into a similar name such as ‘LLMs and VLMs as Planners and Executors‘. If no supplied topic is an exact match, return an empty list and use ‘COHERENT_NEW_DIRECTION‘, ‘MIXED‘, or ‘UNRELATED‘ as appropriate.
* ‘community_name‘: write a concise label only for ‘COHERENT_NEW_DIRECTION‘; otherwise leave it empty.
* ‘mixed_topics‘: list the distinct directions that make the community mixed; keep it empty unless the label is ‘MIXED‘.
* ‘reason‘: one short explanation of the final relation.
* ‘evidence‘: for each supporting paper, copy one verbatim quote from that paper’s title or abstract that justifies the decision.
### Evidence rules
* Every evidence quote must be copied verbatim from the matching paper title or abstract.
* COVERED requires at least one exact supplied topic name in covered_topics.
* Before returning JSON, compare every ‘covered_topics‘ string character-by-character with the supplied topic list. A semantically similar but differently worded topic is invalid.
* COHERENT_NEW_DIRECTION requires a non-empty community_name and reason, with empty covered_topics and mixed_topics.
* MIXED requires at least two distinct entries in mixed_topics and a non-empty reason, with empty covered_topics.
* UNRELATED requires a non-empty reason, with empty covered_topics and mixed_topics.
* The community_name is a descriptive label for the community’s research direction; it does not imply that the survey should cover it.
### Output JSON only
{{
"semantic_relation": "COVERED" | "COHERENT_NEW_DIRECTION" | "MIXED" | "UNRELATED",
"covered_topics": ["exact existing topic name"],
"community_name": "concise semantic label required for COHERENT_NEW_DIRECTION, otherwise empty string",
"mixed_topics": ["distinct research direction represented in the community"],
"reason": "required for COHERENT_NEW_DIRECTION, MIXED, and UNRELATED; otherwise an empty string",
"evidence": [
{{"paper_index": 1, "quote": "..."}}
]

}}
```

Table 45: Prompt for missing-topic community assessment.

Role: C11: verify whether an apparently covered community is substantively discussed at comparable granularity.

```
### Task
Verify whether a Leiden-discovered paper community is substantively discussed by the existing survey topics selected by a previous semantic pass.
Survey query: {query}
Community representative papers:
{papers}
Existing topics proposed by the previous pass:
{topics}
Original survey subsections corresponding to those topics:
{subsections}
For each existing topic, answer this question: "Does this existing topic represent the community as a distinct topic at an appropriate granularity for this survey?" The subsection must substantively discuss the community's defining research direction, not merely a broader parent task that the community specializes in. A subsection title or a passing mention is not sufficient. The subsection must discuss the distinctive problem setting, research object, method family, data modality, dataset type, or evaluation regime visible in the community papers. Do not count generic discussion of a parent task, generic sensor fusion, generic benchmarks, or a shared application domain as coverage of a distinct specialization unless the subsection explicitly discusses that specialization.
If at least one topic represents the community at a comparable granularity, return COVERED and list only those topics. If the supplied subsections discuss only broader parent topics while omitting the community's defining specialization, return COHERENT_NEW_DIRECTION when the community is query-relevant and coherent. In that case, provide a concise community name. Do not change the community membership or infer a different clustering.
### Field guide
* `final_relation`: choose `COVERED` only when at least one supplied topic truly discusses the community at a comparable granularity; otherwise choose `COHERENT_NEW_DIRECTION`.
* `covered_topics`: copy only exact supplied topic names that genuinely cover the community.
* `community_name`: provide a concise label only when the final relation is `COHERENT_NEW_DIRECTION`.
* `topic_checks`: include one entry per requested topic, with `substantively_discussed` set by whether the subsection actually discusses the community, `evidence` copied verbatim from the subsection sentence when true, and `reason` explaining the judgment.
* 'reason': brief explanation of the final relation.
### Evidence rules
For every topic judged substantively_discussed=true, copy one sentence verbatim from that topic's supplied subsection that explicitly supports the community's defining direction, not merely its parent task. Do not paraphrase. For a false judgment, return an empty evidence string.
### Output JSON only
{{
"final_relation": "COVERED" | "COHERENT_NEW_DIRECTION",
"covered_topics": ["exact topic name from the supplied list"],
"community_name": "concise name when final_relation is COHERENT_NEW_DIRECTION, otherwise empty string",
"topic_checks": [
{{
"topic": "exact topic name",
"substantively_discussed": true,
"evidence": "verbatim subsection sentence or empty string",
"reason": "brief reason"
}}
],
"reason": "brief explanation of the final relation"

}}
```

Table 46: Prompt for topic-coverage verification.

Role: C11: determine whether the survey explicitly excludes a candidate topic.

````
Determine whether the paper **explicitly states** that a given topic is excluded, and why.
### Instructions
1. Search for explicit scope limitation statements.
2. Only accept **clear declarative claims** (e.g., We do not cover X / We exclude X)
3. Do not infer justification.
### Constraints
* If justification is implicit or vague, return false.
### Field guide
* `has_claim`: set to true only when the paper explicitly states the topic is excluded.
* `evidence`: copy one verbatim scope sentence when `has_claim` is true; otherwise leave it empty.
### Output Format
```json
{
"has_claim": true | false,
"evidence": "verbatim text if present, otherwise empty"
}

```
````

Table 47: Prompt for explicit topic-exclusion detection.