BEYOND AESTHETICS: LOGICBENCH FOR PAPER-TO-POSTER GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent multimodal models have made it increasingly feasible to generate scientific posters directly from research papers. Existing evaluation largely emphasizes visual quality or isolated fact recovery, leaving unclear whether a poster preserves the scientific logic that connects evidence to conclusions. We introduce LogicBench, a benchmark for measuring logical path coverage in generated posters with source-grounded multi-hop questions derived from validated scientific logic graphs. Questions are verified against the source paper and calibrated with reference posters before held-out poster scoring, so the benchmark measures what the poster makes recoverable. On 134 NIPS25 papers with 1,354 questions, human posters achieve 98.3% accuracy, while five poster-generation methods score 24.4–61.8%. Although these methods achieve high scores on PaperQuiz and P2PEval, which emphasize fact-level correctness, their lower LogicBench scores show that factual recovery does not ensure preserved reasoning. Our analysis shows that accuracy reaches only 66.4% even when all facts needed for a question are present on the poster, suggesting that the logical connections between these facts are not preserved. Moreover, among the evaluated methods, aesthetic scores increase over time while LogicBench accuracy decreases. These findings establish logical path coverage as a distinct dimension of poster quality and motivate methods that preserve scientific logic alongside visual quality.

1 INTRODUCTION

An academic poster presents a study in a compact visual form and serves as a medium for communicating research findings (Erren & Bourne, 2007; Rowe & Ilic, 2009). Automating the transformation from a paper to a poster is therefore important for making scientific communication more efficient and scalable (Qiang et al., 2019; Xu & Wan, 2022).

Against this background, many studies address this task, and recent systems extend earlier graphical and neural approaches with multimodal and multi-agent designs (Qiang et al., 2019; Xu & Wan, 2022; Pang et al., 2026; Choi et al., 2026; Zhang et al., 2026). Despite this progress, evaluation usually measures visual quality or isolated fact recovery (Tanaka et al., 2024; Zhong et al., 2025; Yao et al., 2025; Sun et al., 2026). It does not directly test whether the scientific logic connecting evidence to claims and conclusions is preserved and correct.

To address this gap, we introduce LogicBench, a benchmark for testing whether a generated poster preserves multi-hop relations from its source paper. Its multi-agent pipeline grounds each question in a validated scientific logic graph, while deterministic path sampling supplies connected reasoning paths. Hop ablation and reference calibration remove questions that are unsupported by the source paper or answerable through a shortcut. The source paper is used to construct and verify each question, whereas poster scoring uses the question and poster content only, so the result reflects what the poster makes recoverable. We score generated posters with the fixed question set on a held-out split, extending the evidence-grounded multimodal question-answering setting of ScienceQA (Lu et al., 2022) and SPIQA (Pramanick et al., 2024) to condensed academic posters.

Using the resulting benchmark, We test five generated poster methods on 134 NIPS25 papers with 1,354 questions. Human posters score 98.3% on LogicBench, while generated methods average 44.1%, with scores ranging from 24.4% to 61.8%. On PaperQuiz, generated methods average 63.5% on Detail and 93.4% on Understanding, compared with 54.7% and 91.4% for human posters. On

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.

P2PEval, generated methods average 91.9%, close to the human poster of 92.6%. The lower human score on PaperQuiz Detail is expected because these questions ask for minor facts and numbers that human designers often leave out to keep posters clear and focused. The human poster is therefore a reference for what can be answered from a well-made poster. Taken together, these results show that LogicBench adds a needed test to existing poster-generation benchmarks by checking whether a poster preserves the links between evidence, methods, results, and conclusions. Our analysis in Fig. 2 also shows that newer methods usually produce more attractive posters but score lower on LogicBench. This gap suggests that current evaluations place more weight on appearance than on scientific reasoning, and that future poster-generation methods should give logical fidelity equal attention.

Our contributions are summarized as follows.

1. We introduce LogicBench, a benchmark that measures logical path coverage in generated academic posters through source-grounded multi-hop questions.
2. We develop a multi-agent construction pipeline that builds a scientific logic graph, samples connected inference paths, generates multiple-choice questions, and filters them through correctness, hop-ablation, and reference-calibration checks.
3. We evaluate representative paper-to-poster methods and show that newer methods often produce more attractive posters while achieving lower logical fidelity, revealing a gap that existing benchmarks do not capture.

2 Related Work

Paper-to-Poster Generation. Early work framed paper-to-poster generation as layout planning using probabilistic graphical models and neural models (Qiang et al., 2019; Xu & Wan, 2022). Recent methods structure the generation process at different stages: Paper2Poster organizes extracted assets with a binary-tree layout, P2P divides generation into checked stages, PosterForest follows the paper hierarchy, and PosterGen separates narrative planning from layout construction (Pang et al., 2026; Sun et al., 2026; Choi et al., 2026; Zhang et al., 2026). Other approaches reuse extracted content across outputs or impose constraints for editability and auditability (Xiao et al., 2026; Yang et al., 2026; Hu et al., 2026; Vinaykumar et al., 2026). These methods improve how posters are assembled but do not test whether they preserve the scientific logic of the source paper, which LogicBench evaluates with evidence-grounded multi-hop questions.

Paper-to-Poster Evaluation. The field has developed in response to gaps in data and evaluation. To address the data gap, SciPostLayout provides paper–poster pairs with manual layout annotations and PosterSum evaluates scientific poster summarization; later datasets increase scale or variety and connect design to paper structure (Tanaka et al., 2024; Saxena et al., 2025; Zhong et al., 2025; Yao et al., 2025; Inadumi et al., 2026). To expand evaluation, existing metrics assess layout, image-text similarity, and design quality, while PaperQuiz and P2PEval assess content recovery (Tanaka et al., 2024; Yao et al., 2025; Hu et al., 2026; Yang et al., 2026; Feng et al., 2026; Pang et al., 2026; Sun et al., 2026). These resources improve coverage of poster layout and individual facts, but neither line of work measures logical path coverage in generated posters. LogicBench addresses this gap with validated multi-hop questions.

Evidence-Grounded Reasoning and Multimodal Scientific QA. We build on evidence-grounded scientific QA, where answers require multiple linked evidence pieces. ScienceQA (Lu et al., 2022) established multimodal science reasoning, and SciAssess (Cai et al., 2025) extended it to scientific papers. Later benchmarks address textual and document-level evidence through long-context reasoning, cross-document retrieval, hypothesis generation, and paper-level evaluation; other benchmarks study text–figure alignment, visual search, figure-level QA, and diagram understanding (Cui et al., 2025; Park et al., 2025; Liu et al., 2026; Huang et al., 2026; Pramanick et al., 2024; He et al., 2026; Borisova et al., 2025; Zhao et al., 2025). However, these benchmarks assume that the material defining a question is available during evaluation. LogicBench instead uses the full paper to construct and verify questions but evaluates only the compressed poster, measuring whether it preserves multi-hop links from evidence through methods and results to conclusions.

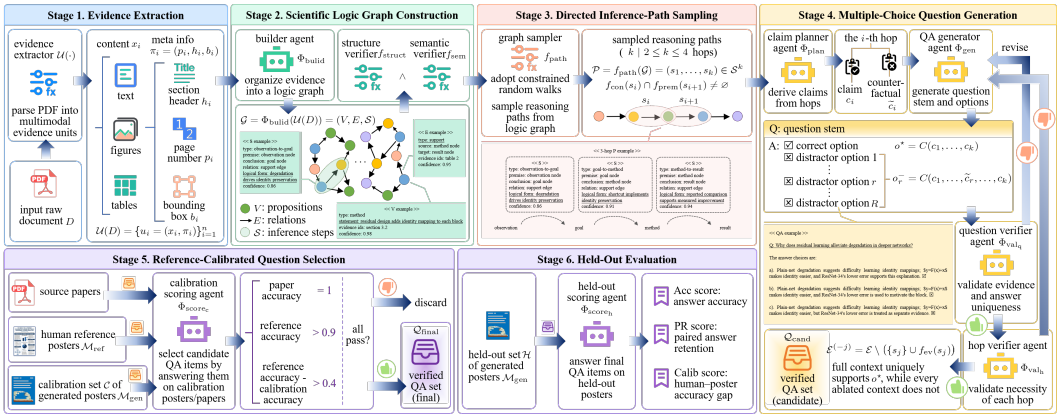


Figure 1: Overview of the LogicBench construction and evaluation pipeline. The source paper is parsed into provenance-aware evidence units and a validated scientific logic graph, from which connected multi-hop LogicQuiz questions are constructed and verified; reference calibration then selects questions for held-out evaluation of generated posters.

3 LogicBench

We use LogicBench for the benchmark and LogicQuiz for its generated question format. An academic poster is treated as a compact rendering of a paper's evidence and scientific dependencies.

3.1 Task Formulation

Traditional evaluation of academic posters focuses primarily on factual retrieval, examining whether isolated pieces of information are present on the poster. However, this approach overlooks the primary function of a poster as a scientific communication medium: preserving the causal and logical coherence of the underlying research. We therefore reframe poster evaluation as the assessment of *logical path coverage*.

Formalization of Logical Coverage. To formalize this notion, we represent a scientific paper D as a directed scientific logic graph $\mathcal{G} = (V, E, \mathcal{S})$. The node set V consists of fact-anchored propositions grounded in paper evidence, each represented as an evidence unit $u \in \mathcal{U}$ that retains provenance metadata such as page number, section header, and bounding box. The edge set $E \subseteq V \times V$ denotes inferential relations between propositions. Each explicit inference step is then represented as a triple $s = \left(v_{\text{pre}} \xrightarrow{e} v_{\text{con}}\right) \in \mathcal{S}$, consisting of its premise nodes, conclusion node, and the inferential edge connecting them. Given an academic poster H, we say that H achieves logical coverage over a reasoning path $\mathcal{P} = \{s_1, s_2, \ldots, s_k\}$ if and only if the inference steps along $\mathcal{P}$ can be traced from the content of H.

Justification for Multiple-Choice Questions. We use multiple-choice questions (MCQs) because each item has a discrete answer index, which supports objective and reproducible scoring. Each item is tied to one sampled reasoning path containing two to four inference hops, preserving the path structure used during construction. Each distractor is plausible but incorrect and changes one claim associated with one hop. This design provides an interpretable indication of where the question-answering model's reasoning may have diverged without requiring evaluation of a free-form response.

3.2 Logic QA Candidate Construction

A reliable benchmark requires a scalable and reproducible method for sourcing logically grounded questions from raw paper PDFs. We address this through a three-stage pipeline: first constructing a scientific logic graph from the paper, then sampling multi-hop reasoning paths from the graph, and finally converting each sampled path into a multiple-choice question candidate.

Scientific Logic Graph Construction. The pipeline begins by converting the raw document D into a verifiable logical graph. Evidence extractor $\mathcal{U}(\cdot)$ (Livathinos et al., 2025; Wang et al., 2024; Datalab, 2024) produces structured units

$$\mathcal{U}(D) = \{u_i = (x_i, \pi_i)\}_{i=1}^{n}, \qquad \pi_i = (p_i, h_i, b_i), \tag{1}$$

where x_i is the extracted content and π_i records the page number p_i, section header h_i, and bounding box b_i for traceability to the source. A builder agent Φ_{build} maps these units into

$$\mathcal{G} = \Phi_{\text{build}}(\mathcal{U}(D)) = (V, E, \mathcal{S}), \tag{2}$$

where V stores propositions, E records directed relations, and $\mathcal{S}$ records explicit inference steps. The generated graph is accepted only when both checks pass:

$$f_{\text{val}}(\mathcal{G}, P) = f_{\text{struct}}(\mathcal{G}) \wedge f_{\text{sem}}(\mathcal{G}, P). \tag{3}$$

Here f_{struct} checks that graph references and evidence links are valid, whereas f_{sem} checks that each proposition and inference is supported by the cited paper evidence. The validated graph is then passed to the sampling stage, where individual reasoning paths are extracted for question generation.

Directed Inference-Path Sampling. Unconstrained random walk sampling can connect statements that do not form a valid chain of reasoning. An Unconstrained walk can jump from one statement to another even when the two statements are not part of the same argument. We therefore define a hop as one inference step s and require the conclusion of one step to be used as a premise by the next. For an inference step s, $f_{\text{prem}}(s)$ returns its premise-node set and $f_{\text{con}}(s)$ returns its conclusion-node set. Valid paths are enumerated first, and a deterministic sampler then selects the paths used for question generation:

$$\mathcal{P} = f_{\text{path}}(\mathcal{G}) = \left\{(s_1, \ldots, s_k) \in \mathcal{S}^k \;\middle|\; 2 \le k \le 4,\ f_{\text{con}}(s_i) \cap f_{\text{prem}}(s_{i+1}) \neq \varnothing \text{ for } i < k\right\} \tag{4}$$

The intersection condition means that the two consecutive steps share at least one claim, so the path remains connected. Here $\mathcal{P}$ is the set of all valid paths. A seed only resolves ties and controls output order.

Multiple-Choice Question Generation. Each sampled reasoning path is converted into a multiple-choice question by three agents. The claim planner Φ_{plan} identifies one necessary claim c_i per hop and mutates it into a counterfactual $\widetilde{c}_i$ that serves as the distractor. The QA generation agent Φ_{gen} then turns these $(c_i, \widetilde{c}_i)$ pairs into the question stem and answer options. Let $o^\star$ denote the correct option and o_r^- denote the r-th distractor option, formed by replacing the claim of hop r with its counterfactual $\widetilde{c}_r$:

$$o^\star = C(c_1, \ldots, c_k), \qquad o_r^- = C(c_1, \ldots, \widetilde{c}_r, \ldots, c_k), \tag{5}$$

where C combines the hop claims in path order. Two- and three-hop paths use four options, while four-hop paths use five. Finally, the question validation agent $\Phi_{\text{val}_\text{q}}$ checks the options against the path evidence and accepts the item only when $o^\star$ is the unique supported option; failed items return to Φ_{plan} and Φ_{gen} for revision.

Hop-Ablation Validation. After question generation, we use the hop verifier agent $\Phi_{\text{val}_\text{h}}$ to test whether every hop is necessary. For a path $\mathcal{P} = (s_1, \ldots, s_k)$ with full evidence context $\mathcal{E}$, the context for hop j is

$$\mathcal{E}^{(-j)} = \mathcal{E} \setminus \big(\{s_j\} \cup f_{\text{ev}}(s_j)\big), \qquad j \in \{1, \ldots, k\}, \tag{6}$$

where $f_{\text{ev}}(s_j)$ denotes the evidence attached to step s_j. The question and answer choices remain fixed, so the verifier agent cannot use the removed material. A candidate is retained only when the full context uniquely supports the declared answer and none of the ablated contexts does. Thus, the test removes questions that can be answered through a shorter path. Finally, the per-hop outcomes are recorded, and a failed test returns the candidate to the generator agent. Candidates that pass hop-ablation validation are added to the verified candidate question set $\mathcal{Q}_{\text{cand}}$.

3.3 Reference-Calibrated Selection and Held-Out Evaluation

Following hop-ablation validation, which removes questions that can be answered without every reasoning step, we calibrate the remaining questions before evaluating posters. This stage checks that each question is correct and excludes questions that a well-designed poster can answer too easily, keeping the benchmark focused on logical coverage.

Reference-Calibrated Question Selection. During calibration, the calibration scoring agent Φ_{score_c} answers the same question from the source paper and from each poster. Let $\mathcal{M}_{\text{ref}}$ and $\mathcal{M}_{\text{gen}}$ denote the human reference and generated posters. For question $q \in \mathcal{Q}_{\text{cand}}$, $a^\star(q)$ is the correct option index and $a(q)$ is the selected option index. A missing or invalid index counts as incorrect, and correctness is $\mathbf{1}[a^\star(q) = a(q)]$. For the paper check, the model receives the question and its source evidence. For poster scoring, the model receives only the question and poster content, so the answer reflects what can be recovered from the poster. The collected answers are then used to filter the candidate set. A question is retained only if it satisfies the paper, human-reference poster, and human–generated gap criteria. The retained ones form the final question set $\mathcal{Q}_{\text{final}} \subseteq \mathcal{Q}_{\text{cand}}$. We provide its exact thresholds, per-paper quotas, and fallback behavior in Sec. 4.1.

Held-Out Evaluation. For generated posters, $\mathcal{M}_{\text{gen}}$ contains between five and ten independent poster files. We split $\mathcal{M}_{\text{gen}}$ into disjoint calibration and held-out sets, $\mathcal{C}$ and $\mathcal{H}$. The calibration set contains half the runs rounded down, and the held-out set contains the remainder rounded up. Question selection uses only $\mathcal{C}$, while $\mathcal{H}$ remains sealed until $\mathcal{Q}_{\text{final}}$ is fixed. The held-out scoring agent Φ_{score_h} scores the held-out posters on $\mathcal{Q}_{\text{final}}$, and we retain the results for reporting.

4 Experiments

4.1 Experimental Setup

Benchmark and Poster Collection. We evaluate paper-to-poster systems on LogicBench using 134 paper-poster pairs randomly sampled from NIPS25. The comparison includes P2P (Sun et al., 2026), Paper2Poster (Pang et al., 2026), PosterForest (Choi et al., 2026), PosterGen (Zhang et al., 2026), and ResearchStudio (Xiao et al., 2026). Each generated poster is aligned with the paper from which its content was produced, while the human-authored poster serves as an answerability control.

Question Set and Reference Calibration. For each paper, the LogicBench split contains ten multiple-choice questions derived from validated two-, three-, or four-hop reasoning paths. A candidate enters the evaluated set only after a correctness verifier confirms that the paper supports exactly one option. We then require a correct paper answer, at least 0.90 accuracy for every human reference poster, and a mean human–poster gap of at least 0.40 whenever generated posters are available. After selection, each paper should contain at least 10 questions, yielding 1,354 questions in total.

Poster Scoring Protocol. We use the multimodal GLM-5.3 model (Z.ai Team, 2026) as the paper and poster reader. For the source paper, the reader receives the question together with the paper evidence. For each poster, it receives the question and the poster image only.

Evaluation Metrics. For a source X (poster or paper) and the question set $\mathcal{Q} \subseteq \mathcal{Q}_{\text{final}}$, let $a^\star(q)$ be the correct option and $a_X(q)$ be the option selected by the model. Accuracy is the fraction of questions answered correctly:

$$\text{Acc}(X) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[a_X(q) = a^\star(q)]\,. \tag{7}$$

For a generated poster M, paired retention measures the fraction of questions answered correctly from the paper D that are also answered correctly from M:

$$\text{PR}(M) = \frac{\sum_{q \in \mathcal{Q}} \mathbf{1}[a_D(q) = a^\star(q)]\mathbf{1}[a_M(q) = a^\star(q)]}{\sum_{q \in \mathcal{Q}} \mathbf{1}[a_D(q) = a^\star(q)]}. \tag{8}$$

During question selection, the human–poster gap is computed for each question as

$$\text{Gap}(q) = \frac{1}{|\mathcal{M}_{\text{ref}}|} \sum_{M \in \mathcal{M}_{\text{ref}}} \mathbf{1}[a_M(q) = a^\star(q)] - \frac{1}{|\mathcal{M}_{\text{gen}}|} \sum_{M \in \mathcal{M}_{\text{gen}}} \mathbf{1}[a_M(q) = a^\star(q)], \tag{9}$$

where $\mathcal{M}_{\text{ref}}$ and $\mathcal{M}_{\text{gen}}$ are as defined in Sec. 3.3.

Table 1: Comparison on LogicBench, PaperQuiz, and P2PEval.

	LogicBench (ours)		PaperQuiz Detail		PaperQuiz Underst.		P2PEval	
	Acc (%)	ΔAcc (%)	Acc (%)	ΔAcc (%)	Acc (%)	ΔAcc (%)	Acc (%)	ΔAcc (%)
Human Reference	98.3	0.0	54.7	0.0	91.4	0.0	92.6	0.0
P2P	54.7	+43.6	**75.4**	-20.7	**95.6**	-4.2	88.1	+4.5
Paper2Poster	**61.8**	+36.5	70.0	-15.3	95.0	-3.6	90.5	+2.1
PosterGen	50.3	+48.0	57.5	-2.8	91.9	-0.5	93.7	-1.1
PosterForest	29.4	+68.9	57.7	-3.0	92.2	-0.8	93.1	-0.5
ResearchStudio	24.4	+73.9	56.8	-2.1	90.8	+0.6	94.9	-2.3

Table 2: Paired retention (PR) on LogicBench. The remaining columns report PR by hop length.

	Overall PR	PR by Hop Length		
Poster Source	Overall PR (%)	2-Hop (%)	3-Hop (%)	4-Hop (%)
Human Reference	98.3	98.3	97.9	99.4
P2P	54.7	59.9	49.4	63.2
Paper2Poster	61.8	58.6	57.0	72.9
PosterGen	50.3	47.7	53.3	44.8
PosterForest	29.4	28.4	28.0	22.5
ResearchStudio	24.4	18.0	28.4	23.0

4.2 Main Results

Tab. 1 reports three sets of measurements. For LogicBench, we report accuracy, its 95% bootstrap confidence interval (CI), and the gap from the human reference (Δ Acc). For PaperQuiz (Pang et al., 2026), we report the published Detail and Understanding scores from the original Paper2Poster benchmark, where a VLM reader answers detail and understanding questions from the poster image alone and is graded by exact match. For P2PEval (Sun et al., 2026), we report the published checklist scores.

The three benchmarks rank these methods almost in reverse. On LogicBench, human-reference posters reach 98.3% accuracy, while the best generated method, Paper2Poster, reaches only 61.8%, with the remaining methods falling from 54.7% to 24.4%. On PaperQuiz, generated methods score high across the board, with detail accuracy spread between 57% and 75.4% and understanding accuracy between 92% and 95.6%, whereas human-designed posters place last at 54.7% and 91.4%. On P2PEval, all methods cluster near the top, with ResearchStudio reaching 94.9% and the human reference trailing at 92.6%. This inversion shows what each evaluation rewards. Human designers keep the core facts and logics that support the paper's claims and contributions and drop the rest. PaperQuiz and P2PEval mostly test these dropped details, such as specific data points or minor claims, so the human posters rank low on them. Methods that fill their posters with copied text, such as P2P, score highest on PaperQuiz but fall to 54.7% on LogicBench. The same pattern holds for every method with a high PaperQuiz score. This shows that strong performance on factual benchmarks does not mean the poster retains the paper's reasoning. LogicBench tests whether the chain from evidence to conclusion survives. Under that test, human posters reach the 98.3% control level while the best method reaches only 61.8%.

Paired Retention Variants. Tab. 2 reports paired retention (PR), defined in Eq. 8, together with its breakdown by hop length. Because the paper baseline answers almost all questions correctly, PR equals accuracy, and the three aggregation schemes coincide; the table therefore shows their common value once as Overall PR. The hop-length columns show where chains break. The human reference stays flat across all lengths (98–99%). Paper2Poster improves with depth, rising from 58.6% at two hops to 72.9% at four, whereas PosterForest and ResearchStudio decline. Chain length therefore separates methods that connect reasoning steps from those that preserve isolated facts.

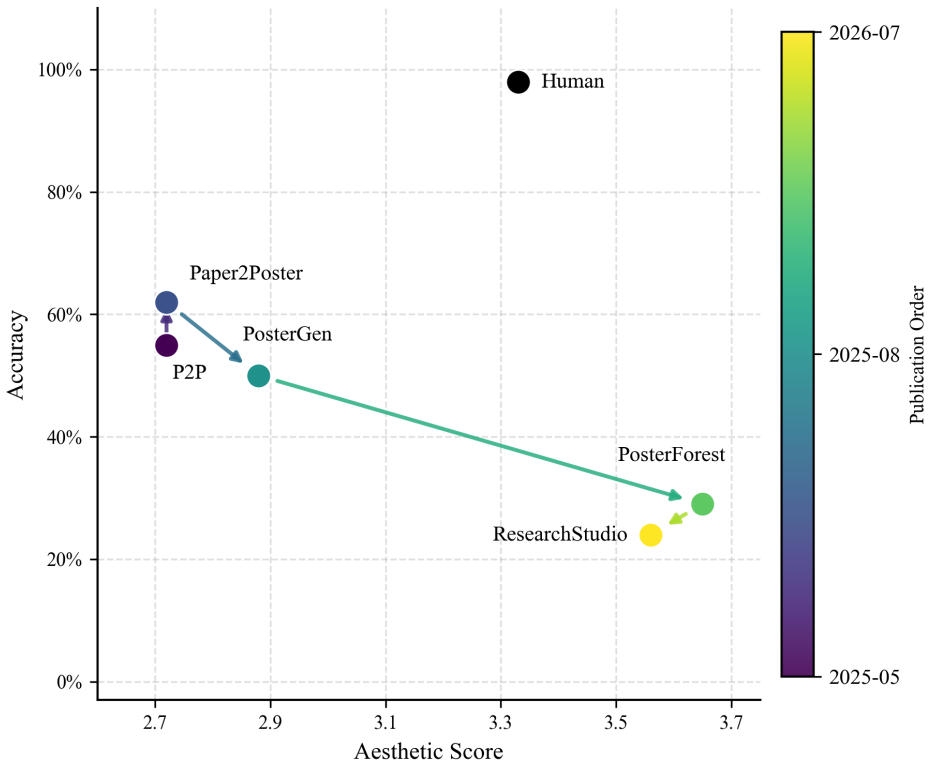


Figure 2: Aesthetic score vs. LogicBench accuracy for the six poster sources. Newer methods achieve higher aesthetics but lower logical fidelity, yielding a weak negative correlation.

4.3 Analysis

Our analysis is organized around the following perspectives: **(A) Benchmark validation** asks whether our experiments support our claims (Aesthetics-Logic Disconnect, Distractor Diagnosis, Calibration-Test Generalization Gap, Fact Availability vs. Error Rate). **(B) Method comparison** asks how existing poster-generation methods trade off aesthetics, layout, and logical fidelity against each other (Aesthetics-Logic Disconnect); detailed layout and inference-edge results are reported in Appendix A.

Aesthetics-Logic Disconnect. To answer whether high aesthetic quality implies logical fidelity, we compare each method's aesthetic score with its LogicBench accuracy. We use the aesthetic scores on their 1–5 scale judged by VLM models, following the aesthetic metric in the Paper2Poster benchmark. As shown in Fig. 2, the six scores range from 2.71 to 3.65 while accuracy ranges from 24.4% to 98.3%, yielding a negative Pearson correlation of $r = -0.56$ ($n = 6$). The scores also follow the release timeline: P2P and Paper2Poster from early 2025 score 2.72 and retain 54.7% and 61.8%; PosterGen and PosterForest from later 2025 score 2.88 and 3.65 with retention of 50.3% and 29.4%; and ResearchStudio from 2026 scores 3.56 with the lowest retention at 24.4%. Newer methods generally receive higher aesthetic scores while retaining less of the paper's multi-hop reasoning. This is where our benchmark matters: it calls for a more holistic evaluation of poster quality that includes logical fidelity.

Distractor Diagnosis. To identify which step of a reasoning path is most likely to be lost when a paper becomes a poster, we exploit the design of our distractors. Each wrong option is identical to the correct one except for a single step whose claim is replaced by its counterfactual. When a reader selects a wrong option, the modified step therefore marks the step the poster failed to preserve. For questions of each length, we count how often each step's distractor is selected among all wrong answers. As shown in Fig. 3, the first step dominates the errors, and its share decreases monotonically along the path. This pattern holds across all question lengths. Among four-hop wrong

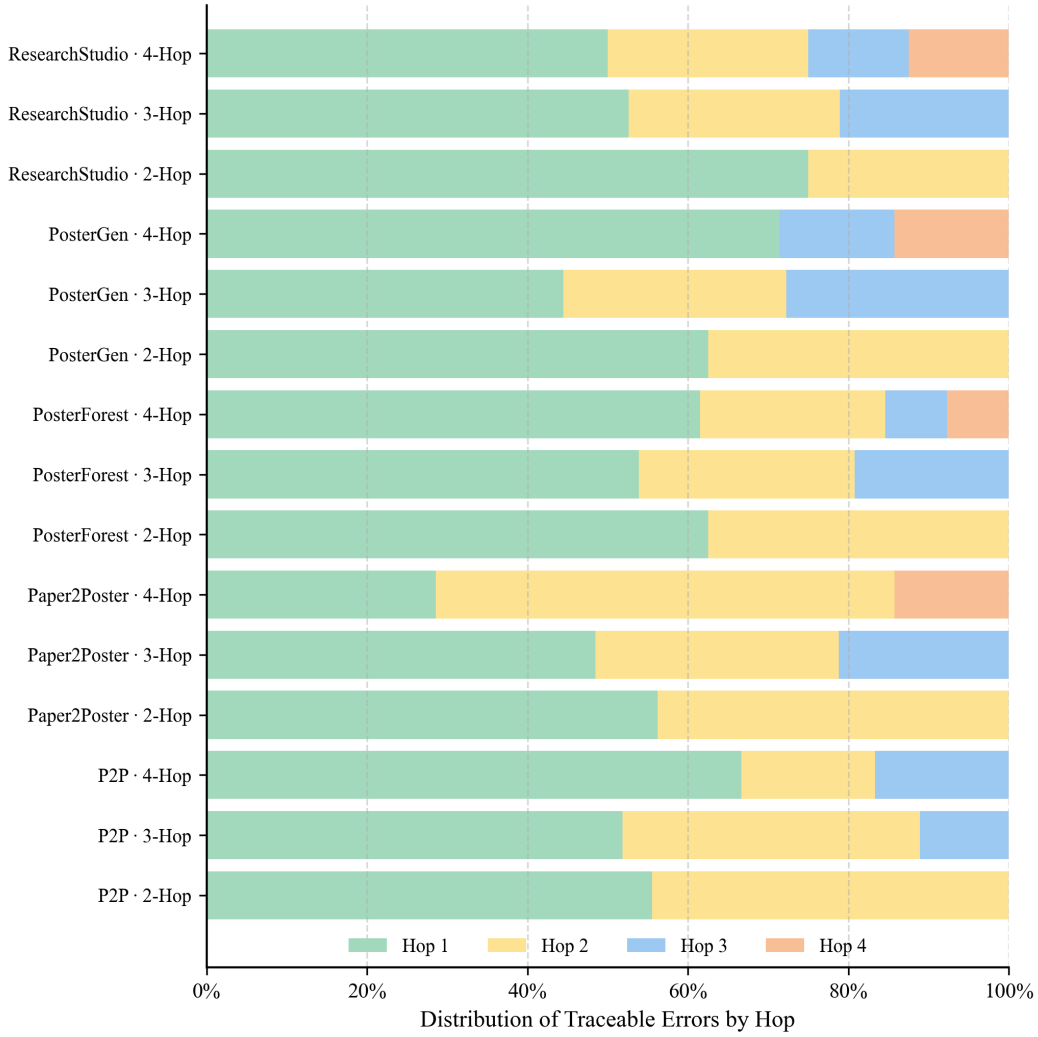


Figure 3: Share of traceable wrong answers attributed to each reasoning step, grouped by question length. Errors are dominated by early steps across all lengths, indicating that generated posters preserve conclusions while dropping the logical motivations that precede them.

answers that can be traced to a step, 56.1% change step one, 24.4% step two, and 9.8% each for steps three and four. Two- and three-hop questions show the same ordering. This pattern indicates that generated posters tend to keep conclusions and results while dropping the logical motivations.

Calibration-Test Generalization Gap. As mentioned in Sec. 3.3, our selection stage scores every candidate question on the calibration split $\mathcal{C}$ and retains the filtered questions for final evaluation on the held-out split $\mathcal{H}$. Because selection never accesses the held-out split, the held-out results serve as a sealed test of how well the selected questions transfer. Tab. 3 compares the two splits. All generated methods score substantially lower on the calibration split than on the held-out split, with held-out scores exceeding calibration scores by 11.5% to 32.6% (e.g., P2P: 22.1% vs. 54.7%). The human reference also drops but remains high (89.4% vs. 98.3%). This gap arises because the filter ran on the calibration posters and retained questions that the methods failed there, making the calibration split harder. Consequently, scores on the split that shaped selection are not clean test performance, which is why we seal the held-out split.

Fact Availability vs. Error Rate. We analyze whether posters fail because facts are missing or because the connections between them are lost. For each question and method, we compute the

Table 3: Accuracy on the calibration split and the held-out split. Calibration scores are lower across all methods because that split retains the questions they failed during selection, making the held-out split an unbiased estimate of logical retention.

Poster Source	Calibration (%)	Held-Out Test (%)	Gap (%)
Human Reference	89.4	98.3	8.9
P2P	22.1	54.7	32.6
Paper2Poster	38.5	61.8	23.3
PosterGen	27.6	50.3	22.7
PosterForest	15.7	29.4	13.7
ResearchStudio	12.9	24.4	11.5

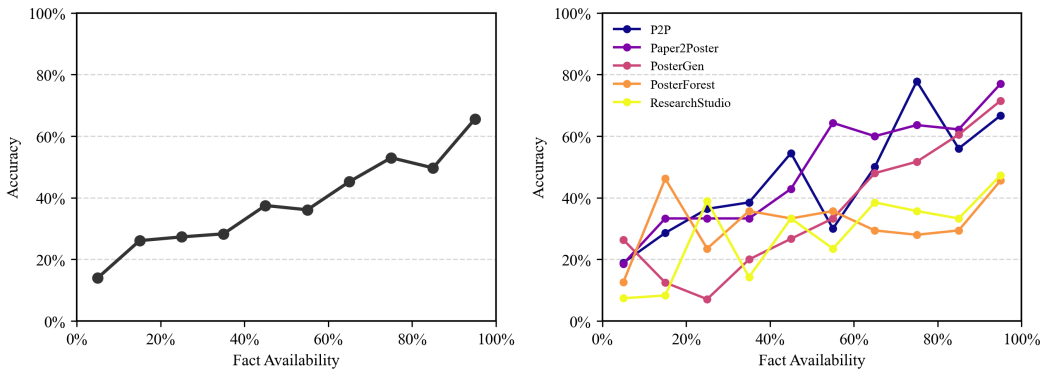


Figure 4: Question accuracy vs. the proportion of evidence units present on the generated poster. Even with all facts present, accuracy remains limited, indicating that failures stem from lost logical connections rather than missing facts.

fraction of the question's evidence units that appear on the poster, and Fig. 4 plots accuracy against this fact availability, with all methods pooled on the left and per-method curves on the right. Even when all facts are confirmed present, accuracy reaches only 66.4%, meaning one in three questions still fails. This demonstrates that our questions test the logical connections between facts, not the facts themselves. A benchmark that only checked fact presence would report near-perfect scores here, since the facts are largely on the poster. The residual failures we measure are therefore pure connection loss, which is exactly the reasoning capacity LogicBench is designed to probe.

Appendix B reports ablations on the gap threshold and the number of questions per paper.

5 Conclusion

In this paper, we introduce LogicBench, a benchmark for evaluating whether generated academic posters preserve the reasoning that connects evidence to conclusions in the source paper. It measures logical path coverage with source-grounded multi-hop questions derived from a validated scientific logic graph. Although the generated methods achieve high scores on PaperQuiz and P2PEval, which emphasize fact-level correctness, they reach only 24.4–61.8% on LogicBench, compared with 98.3% for human posters, across 134 NIPS25 papers and 1,354 questions. Our analyses show that generated posters often omit early steps in a reasoning path and that visual appeal has little relation to logical fidelity. These findings show that poster evaluation should measure logical path coverage alongside visual quality and factual recovery. Future work can test LogicBench across more research areas, poster formats, and reader models.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, code testing, experiment running, debugging and LaTeX formatting, including adaptation to the ICLR 2027 template. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript, including its text, claims, citations, results, and artifacts.

REPRODUCIBILITY STATEMENT

The main text specifies the LogicBench and its contents. Code and artifacts are in the Supp Materials during double-blind review.

REFERENCES

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. Scivqa 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 182–210, 2025.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, et al. Sciassess: Benchmarking llm proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2335–2357, 2025.

Jiho Choi, Seojeong Park, Seongjong Song, and Hyunjung Shim. Posterforest: Hierarchical multi-agent collaboration for scientific poster generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 379–401, 2026.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. In *International Conference on Learning Representations*, pp. 9859–9906, 2025.

Datalab. Marker: Convert pdf to markdown + json quickly with high accuracy. `https://github.com/datalab-to/marker`, 2024.

Thomas C. Erren and Philip E. Bourne. Ten simple rules for a good poster presentation. *PLoS Computational Biology*, 3(5):e102, 2007.

Yuheng Feng, Wen Zhang, Haodong Duan, and Xingxing Zou. Posteriq: A design perspective benchmark for poster understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29295–29304, 2026.

Hang He, Chuhuai Yue, Chengqi Dong, Chengcheng Wan, Ting Su, Haiying Sun, Jiajun Chai, Xiaohan Wang, and Guojun Yin. Vistahop: Benchmarking multi-hop visual reasoning for visual deepsearch. *arXiv preprint arXiv:2606.03273*, 2026.

Haojie Hu, Chenhao Dang, Yaojia Liu, Hengrui Kang, Conghui He, and Weijia Li. Postermeld: Multi-agent paper-to-poster generation for controllable design diversity with editable print-ready outputs. *arXiv preprint arXiv:2608.02218*, 2026.

Tiancheng Huang, Ruisheng Cao, Yuxin Zhang, Zhangyi Kang, Zijian Wang, Chenrun Wang, Yijie Luo, Hang Zheng, Lirong Qian, Lu Chen, et al. Airqa: A comprehensive qa dataset for ai research with instance-level evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 9927–9955, 2026.

Shun Inadumi, Shohei Tanaka, Tosho Hirasawa, Atsushi Hashimoto, Koichiro Yoshino, and Yoshitaka Ushiku. Scipostgen: Bridging the gap between scientific papers and poster layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2131–2141, 2026.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 2026.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*, 2025.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. Paper2poster: Towards multimodal poster automation from scientific papers. *Advances in Neural Information Processing Systems*, 38, 2026.

Jiwon Park, Seohyun Pyeon, Jinwoo Kim, Rina Carines Cabal, Zhenyuan He, Yihao Ding, and Soyeon Caren Han. Dochop-qa: Towards multi-hop reasoning over multimodal document collections. *arXiv preprint arXiv:2508.15851*, 2025.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024.

Yu-Ting Qiang, Yan-Wei Fu, Xiao Yu, Yan-Wen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers by probabilistic graphical models. *Journal of Computer Science and Technology*, 34(1):155–169, 2019.

Nicholas Rowe and Dragan Ilic. What impact do posters have on academic knowledge transfer? a pilot survey on author attitudes and experiences. *BMC Medical Education*, 9(1):71, 2009.

Rohit Saxena, Pasquale Minervini, and Frank Keller. Postersum: A multimodal benchmark for scientific poster summarization. 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. 1828–1844, 2025.

Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Ge Zhang, Wenhao Huang, Jian Yang, et al. P2p: Automated paper-to-poster generation and fine-grained benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 18756–18798, 2026.

Shohei Tanaka, Hao Wang, and Yoshitaka Ushiku. Scipostlayout: A dataset for layout analysis and layout generation of scientific posters. In *35th British Machine Vision Conference*, 2024.

Amogh Vinaykumar, Aiden Li, Suozhi Huang, and Shilong Liu. Any2poster: Any-source poster generation across modalities and domains. *arXiv preprint arXiv:2606.02915*, 2026.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Lingao Xiao, Yalun Dai, Yangyu Huang, Qihao Zhao, Wenshan Wu, Hugo He, Ruishuo Chen, Jin Jiang, Qianli Ma, Jiahuan Zhang, et al. Researchstudio-reel: Automate the last mile of research from paper to poster, video, and blog. *arXiv preprint arXiv:2607.04438*, 2026.

Sheng Xu and Xiaojun Wan. Posterbot: A system for generating posters of scientific papers with neural models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 13233–13235, 2022.

Tianyi Yang, Dawei Fu, Youpeng Wu, Zixun Kou, Linrui Chen, Ruobing Jiang, Zijian Wang, and Qiang Li. Posterharness: Turning scientific poster generation into an auditable instruction-following benchmark. *arXiv preprint arXiv:2607.03006*, 2026.

Kaichun Yao, Lan Zeng, Chuan Qin, Hengshu Zhu, Yanjun Wu, and Libo Zhang. Scipg: A new benchmark and approach for layout-aware scientific poster generation. 2025.

Z.ai Team. GLM-5.3: Frontier coding with emergent cyber capabilities. `https://z.ai/blog/glm-5.3`, 2026.

Zhilin Zhang, Xiang Zhang, Jiaqi Wei, Yiwei Xu, and Chenyu You. Postergen: Aesthetic-aware multi-modal paper-to-poster generation via multi-agent llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9813–9823, 2026.

Yilun Zhao, Chengye Wang, Chuhan Li, and Arman Cohan. Can multimodal foundation models understand schematic diagrams? an empirical study on information-seeking qa over scientific papers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18598–18631, 2025.

Xinyi Zhong, Zusheng Tan, Jing Li, Shen Gao, Jing Ma, Shanshan Feng, and Billy Chiu. Scientific poster generation: A new dataset and approach. *Pattern Recognition*, 164:111507, 2025.

Table 4: Within-method Pearson correlation between a poster's white-space ratio and its LogicBench accuracy. More white space generally correlates with lower accuracy, indicating that it leaves less room for the evidence that multi-hop questions rely on.

	P2P	Paper2Poster	PosterGen	PosterForest	ResearchStudio
Within-Method r	+0.05	-0.34	+0.13	-0.40	-0.42

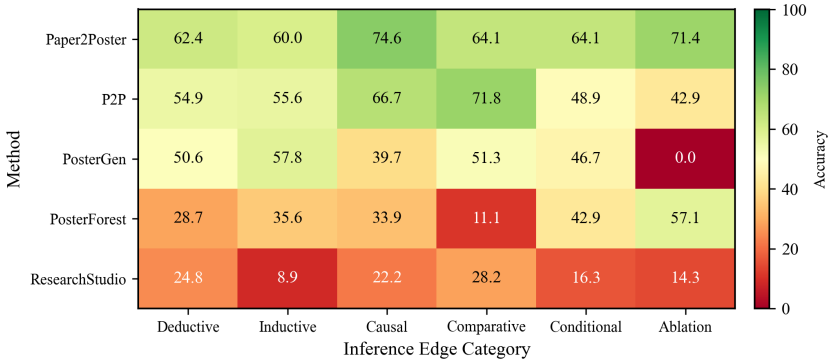


Figure 5: Accuracy by method and inference-edge category. Each question counts under every category in its chain. The methods preserve different parts of the argument, and none preserves all of them.

A Additional Analysis

Layout-Logic Correlation. We examine whether poster layout correlates with logical accuracy by computing the white-space ratio (proportion of near-white pixels, grayscale > 240) for each generated poster and correlating it with its LogicBench accuracy.

As shown in Tab. 4, the within-method Pearson correlation r is negative for Paper2Poster, PosterForest, and ResearchStudio: posters with more white space from these methods tend to score lower. P2P and PosterGen show weak positive trends, but neither is significant. This pattern is consistent with the fact that more white space leaves less room for content: when a method compresses aggressively and leaves the canvas sparse, it also drops the evidence units that multi-hop questions rely on.

Inference-Edge Category Analysis. We examine which categories of inference survive poster compression. Each inference step is assigned one of six categories: deductive (derived from assumptions and definitions), inductive (generalized from experimental observation), causal (states what causes what), comparative (ranks methods or settings), conditional (bounds where a conclusion applies), and ablation (attributes effects to components).

We group questions by the categories present in their reasoning chains and recompute each method's accuracy, as shown in Fig. 5. Deductive questions dominate the benchmark, and all methods struggle with them, ranging from 24.8% (ResearchStudio) to 62.4% (Paper2Poster). Beyond this, the methods diverge. P2P performs best on comparative questions (71.8%), a category that rewards copying tables and rankings verbatim, whereas PosterForest collapses on the same category (11.1%), as its tree layout drops comparative content. ResearchStudio scores near the floor on inductive questions (8.9%), which require reading experimental trends from the poster. Paper2Poster is the only approach above 50% on every category, consistent with its overall lead. This spread shows that each generator preserves a different part of the paper's argument, and none preserves all of it.

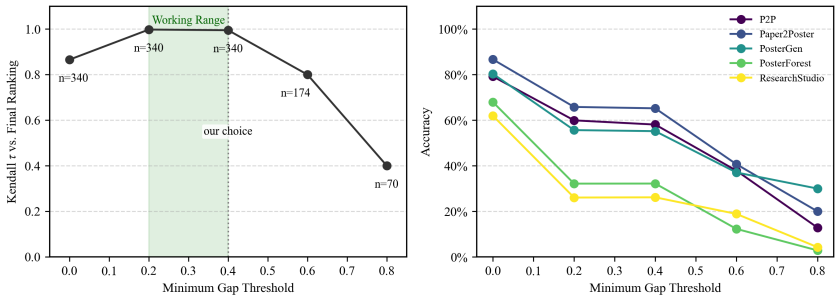


Figure 6: Ranking robustness under random question selection at each gap threshold. Method rankings remain stable across the working range and shift only when filtering is too weak or too strong.

Table 5: Bootstrap over question subsets. Five questions per paper suffice to stabilize the ranking.

k	Top-1 Agreement (%)	Mean Kendall τ	CI Half-Width (%)
2	84.6	0.87	8.0
5	99.2	0.96	4.3
10	100.0	1.00	0.0

B Additional Ablations

Gap-Threshold Ablation. To test whether the method ranking is robust to the choice of gap threshold, we redraw the question set at each threshold. For every gate, we randomly draw ten questions per paper, repeat the draw 200 times, and average the results. Fig. 6 reports the Kendall τ against the final ranking. The curve is flat across the working range: at 0.2 and at our choice of 0.4, every random draw recovers the final ranking ($\tau \geq 0.995$). The ranking moves only at the edges. With no filter, easy questions flood the set and PosterGen overtakes P2P ($\tau = 0.87$). With too strong a filter, too few questions remain, with 0.6 keeping 174 ($\tau = 0.8$) and 0.8 keeping 70 ($\tau = 0.4$). Between the edges, the threshold does not affect the ranking.

Questions-per-Paper Saturation. To determine how many questions per paper the benchmark requires, we draw k questions from the frozen ten, for $k = 2$, 5, and 10, with 500 draws each, and recompute all final scores. Tab. 5 reports the results. With two questions per paper, the 95% interval spans about ± 8 points and the top method is recovered in 84.6% of draws. With five questions, the interval tightens to about ± 4 points, the mean Kendall τ against the full ranking reaches 0.96, and the top method is recovered in 99.2% of draws. Five questions per paper therefore suffice to recover the ranking, showing that the benchmark is robust to the number of questions drawn per paper.

C Example Questions

This section presents three representative questions from LogicBench, one for each major reasoning motif. Each question tests a multi-hop inference chain drawn from the paper's scientific logic graph. The correct option preserves the full chain; each distractor changes exactly one hop to a counterfactual.

C.1 Mechanism: How Theoretical Assumptions Support a Framework (3 Hops)

Paper: Provably Efficient Multi-Task Meta Bandit Learning via Shared Representations

Question: Which of the following best describes how the paper's theoretical assumptions and regret analyses jointly support its proposed multi-task meta bandit framework?

Reasoning chain: Assumption (Gaussian features) $\rightarrow$ spectral initialization guarantee $\rightarrow$ EtC source-task regret bound $\rightarrow$ OFUL target-task regret bound $\rightarrow$ framework justification.

A. Assumption 3.1 assumes source-task feature vectors are standard Gaussian and source noise is i.i.d. Gaussian; this Gaussian assumption is used for the spectral-initialization guarantee and applies to source tasks, not to the target transfer phase. Column-wise incoherence of the true reward matrix, implied by Assumption 3.2, is essential for the EtC regret analysis and turns the source-task cumulative regret sublinear, $\tilde{O}(\sqrt{rNT})$. The OFUL target-task regret is $\tilde{O}(\sqrt{drN_2})$, which improves on the direct-learning regret $\tilde{O}(d\sqrt{N_2})$ because $r \ll d$, validating the benefit of transferring the learned representation. **(Correct)**

B. Assumption 3.1 assumes source-task features are only sub-Gaussian, so the spectral-initialization guarantee holds even without the standard Gaussian model. (Distractor: changes the Assumption hop)

C. Without column-wise incoherence, the EtC source-task regret would still be sublinear because the analysis does not depend on incoherence. (Distractor: changes the Incoherence hop)

D. The OFUL target-task regret is $\tilde{O}(d\sqrt{N_2})$, the same order as direct learning, so transferring provides no improvement. (Distractor: changes the Regret Comparison hop)

C.2 Scope: Justification for Restricting Computation (3 Hops)

Paper: The Flood Complex: Large-Scale Persistent Homology on Millions of Points

Question: Which statement accurately describes the paper's justification for restricting masking and distance computations to maximal simplices?

Reasoning chain: Delaunay simplex $\rightarrow$ barycentric grid $P_\sigma \subset \mathrm{conv}(\sigma) \rightarrow$ face-filtration containment condition $\rightarrow$ maximal-simplex sufficiency.

A. P_σ is chosen within $\mathrm{conv}(\sigma)$ as a barycentric grid (Lemma 6). A simplex σ is included only if $P_\tau \subseteq X_r$ for every face $\tau \subseteq \sigma$. Because $P_\tau \subseteq P_\sigma$ for all faces, face filtration values are extracted from maximal-simplex rows, so masking and distance computation is performed only on maximal simplices. **(Correct)**

B. A simplex σ is included as soon as $P_\sigma \subseteq X_r$ without requiring the condition for any face. (Distractor: changes the Containment hop)

C. P_σ is chosen over the entire ambient Delaunay triangulation rather than within $\mathrm{conv}(\sigma)$. (Distractor: changes the Grid Construction hop)

D. Each face τ must be processed with its own P_τ, so computation must be repeated for every simplex. (Distractor: changes the Maximal Simplex hop)

C.3 Comparison: Method-Design Trade-off (3 Hops)

Paper: Problem-Parameter-Free Decentralized Bilevel Optimization

Question: According to the paper's method-design reasoning, which chain of claims best explains why AdaSDBO is described as having only modest additional communication overhead?

Reasoning chain: Scalar accumulator design $\rightarrow$ consensus synchronization $\rightarrow$ overhead comparison with gradient-tracking methods.

A. Adaptive stepsizes are computed from scalar accumulators; trackers synchronize the accumulators; AdaSDBO *also* communicates tracker states, making its overhead comparable to gradient tracking. (Distractor: changes the Overhead Comparison hop)

B. Stepsizes are computed from full gradient vectors rather than scalar squared norms. (Distractor: changes the Accumulator Design hop)

C. The tracking mechanism averages final stepsizes locally rather than synchronizing the underlying accumulators. (Distractor: changes the Synchronization hop)

D. AdaSDBO's adaptive stepsizes are computed from scalar accumulators of squared gradient norms. The tracking mechanism synchronizes these accumulators via consensus before

computing stepsizes. As a result, AdaSDBO communicates only scalars, whereas gradient-tracking methods communicate full tracker states, so the additional overhead is modest. **(Correct)**

C.4 Discussion

These examples illustrate the key properties of LogicBench questions. First, each option is a complete multi-hop argument; a reader must trace the entire chain to distinguish the correct option from distractors that differ at exactly one hop. Second, the chains span different reasoning motifs: Example A tests how theoretical assumptions (mechanism) justify a framework, Example B tests how a scope restriction follows from a geometric containment, and Example C tests how a comparative claim (communication overhead) follows from design choices. Third, answering correctly from the poster requires the poster to preserve the inferential connections between facts, not just the facts themselves — a poster that lists the assumptions and the regret bounds without connecting them will not support a correct answer.

C.5 Evidence: How Theory Motivates Method Choice (3 Hops)

Paper: Provably Efficient Multi-Task Meta Bandit Learning via Shared Representations

Question: Which statement best explains how the paper's Theorem 4.1 connects the spectral-initialization guarantee, the per-task estimator, and the sample-accuracy trade-off?

Reasoning chain: Spectral initialization $\rightarrow$ subspace-distance guarantee $\rightarrow$ least-squares estimator $\rightarrow$ sample-size requirement $N_1 T \geq C\mu^2\kappa^4 dr^2/\delta_0^2$.

A. The subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ relies on Assumption 3.1's standard Gaussian features/noise together with Assumption 3.2 and the NSR condition. After fixing $\hat{B}$, each task parameter is estimated by *ridge regression* instead of least squares, and the same Theorem 4.1 error bound still holds. The proof requires $N_1 T \geq C\mu^2\kappa^4 dr^2/\delta_0^2$, so smaller δ_0 means a larger required source-sample size. **(Distractor: changes the Estimator hop)**

B. The subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ relies on Assumption 3.1's standard Gaussian features/noise together with Assumption 3.2 and the NSR condition. After fixing $\hat{B}$, each task parameter is estimated by least squares, $\hat{w}_t = (\Phi_{N_1,t}\hat{B})^\dagger Y_{N_1,t}$, and this least-squares estimator enters the Theorem 4.1 error bound. The proof requires $N_1 T \geq C\mu^2\kappa^4 dr^2\delta_0^2$, so *smaller* δ_0 allows a *smaller* required source-sample size. **(Distractor: changes the Sample-Size hop)**

C. The subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ relies on Assumption 3.1's standard Gaussian features/noise together with Assumption 3.2 and the NSR condition. After fixing $\hat{B}$, each task parameter is estimated by least squares, $\hat{w}_t = (\Phi_{N_1,t}\hat{B})^\dagger Y_{N_1,t}$, and this least-squares estimator enters the Theorem 4.1 error bound. The proof of Theorem 4.1 requires $N_1 T \geq C\mu^2\kappa^4 dr^2/\delta_0^2$, so smaller δ_0 means a larger required source-sample size. **(Correct)**

D. The subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ holds under Assumption 3.2 and the NSR condition *even if* Assumption 3.1's Gaussian feature/noise model is dropped. After fixing $\hat{B}$, each task parameter is estimated by least squares, $\hat{w}_t = (\Phi_{N_1,t}\hat{B})^\dagger Y_{N_1,t}$, and this least-squares estimator enters the Theorem 4.1 error bound. The proof requires $N_1 T \geq C\mu^2\kappa^4 dr^2/\delta_0^2$, so smaller δ_0 means a larger required source-sample size. **(Distractor: changes the Assumption hop)**

C.6 Contribution: How Auxiliary Results Prove the Main Theorem (4 Hops)

Paper: Regularized Least Squares Learning with Heavy-Tailed Noise is Minimax Optimal

Question: Which statement accurately describes how the paper's refined margin equalities, auxiliary functions, and sub-task bounds combine to prove the main generalization bound?

Reasoning chain: Refined margin equalities (three-term decomposition) → auxiliary functions ϕ, ρ (Lipschitz constants) → balancing $k = c\gamma_{i+1}^{-2} \ln(e/\ell_{j+1})$ → union bound over sub-tasks (Claims 1–2) → main generalization bound.

A. The refined bounds give a three-term upper bound for $\sup_{w \in H(\Gamma_i, L_j)}(L_D(w) - L_S^\gamma(w))$: the (20) transformed margin-loss term plus the (21) ϕ and (22) ρ population-vs-empirical expectation gaps. Lemma 8 requires $k \geq c'\gamma_{i+1}^{-2}$; its Rademacher bound for (21)/(22) uses the Lipschitz constants of ϕ and ρ, which are at most $c_L \exp(-\gamma_{i+1}^2 k/c_L)(\sqrt{k} + \gamma_{i+1}^{-1})$. Balancing sets $k = c\gamma_{i+1}^{-2} \ln(e/\ell_{j+1})$, ensuring $\exp(-\gamma_{i+1}^2 k/c) \leq \ell_{j+1}/e$ and $k \geq c'\gamma_{i+1}^{-2}$, so Lemma 7 and Lemma 8 combine by a union bound with $\delta' = \delta/2$. Claim 1 comes from a union bound over all margin/norm sub-tasks with sub-task-specific confidences $\delta_{i,j} \approx \delta \exp(-\gamma_{i+1}^2 \ln(e/\ell_{j+1}))$ and $\delta_i \approx \delta \exp(-\gamma_{i+1}^{-2} \ln(e\gamma_{i+1}^2 n))$; Claim 2 then lifts this to every $\gamma \in (n^{-1/2}, c_\gamma]$ via Lemma 4. **(Correct)**

B. The refined bounds make the supremum of the generalization gap *exactly equal* to the product of the three sub-task terms (20), (21), and (22). **(Distractor: changes the Decomposition hop)**

C. Balancing sets $k = c\gamma_{i+1}^{-2}$ *independent of* ℓ_{j+1}, so the exponential term is replaced by 1 and the Lemma 3 bound no longer depends on ℓ_{j+1}. **(Distractor: changes the Balancing hop)**

D. The refined bounds give a three-term upper bound, but Lemma 8 *applies for all* $k > 0$, since the Lipschitz constants of ϕ and ρ are uniformly bounded by an absolute constant independent of k and γ_{i+1}. **(Distractor: changes the Lipschitz hop)**

C.7 Rationale: Why a Confidence Ellipsoid Scales with $\sqrt{r}$ (3 Hops)

Paper: Provably Efficient Multi-Task Meta Bandit Learning via Shared Representations

Question: Which chain of reasoning best explains why the OFUL meta-bandit learner can construct a confidence ellipsoid whose size scales with $\sqrt{r}$ rather than $\sqrt{d}$?

Reasoning chain: MoM inapplicability ($T \leq O(d)$ restriction) → spectral initialization guarantee → rank-r representation → confidence ellipsoid scaling with $\sqrt{r}$.

A. The MoM estimator is not directly applicable because its near-optimal subspace guarantee holds only when $T \leq O(d)$. The spectral subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ requires Assumption 3.1: standard Gaussian source features and i.i.d. Gaussian source noise. The Theorem 5.1 confidence ellipsoid scales with $\sqrt{r}$ rather than $\sqrt{d}$ because it uses the shared rank-r representation. **(Correct)**

B. The MoM estimator is directly applicable even when T is independent of d, with no $T \leq O(d)$ restriction. **(Distractor: changes the Applicability hop)**

C. The Theorem 5.1 confidence ellipsoid scales with $\sqrt{d}$, exactly like single-task OFUL, because the target parameter θ_{T+1}^* is still d-dimensional. **(Distractor: changes the Dimensionality hop)**

D. The spectral subspace-distance guarantee $SD(\hat{B}, B^*) \leq \delta_0$ remains valid for arbitrary heavy-tailed source features and noise even if Assumption 3.1 is dropped. **(Distractor: changes the Assumption hop)**