WHAT CAN A LEADERBOARD CERTIFY? COMPOSITIONAL CONTROLLABILITY FOR FAIR EVALUATION AND TRAINING OF BIOMEDICAL LITERATURE-REVIEW AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Leaderboards rank long-horizon agents by their final outputs. Yet a higher score alone does not establish whether two systems are comparable or which stage accounts for the difference. Unequal evidence, inputs, or budgets can affect scores, and statistical corrections do not remove this mismatch. We introduce *compositional controllability* to address these questions. A comparison window covers one stage, several stages, or the whole agent. Our central result bounds the gap between observed and controlled score differences using only nuisance outside the window. This yields an admissibility test applied before scores are inspected. Inadmissible comparisons are refused. For admissible pairs, an ordering is certified only when the score gap exceeds the combined sampling and nuisance radii; otherwise, it remains undecided. These decisions give each system a rank interval. We introduce *BioLitBench*, a benchmark of $2{,}042$ biomedical articles represented as structured claim graphs. We use it to compare published pipelines, commercial agents, and our own agent under a shared reference list or a shared frozen corpus. Among seven published pipelines, a conventional statistical analysis declares a winner in 14 of 21 pairwise comparisons. Yet the top-ranked system alone received the target review's bibliography, giving it access to evidence the others were not supplied. To isolate pipeline performance, our test requires matched inputs and a fixed backbone model. It refuses 11 of the 21 comparisons, including every comparison involving the top-ranked system. Seven of the 14 conventional conclusions fall within these refused pairs. The same comparison windows support stage-level training. We train *SCRIBE* (**S**tage-**C**ertified **R**eview-writing agent with **I**nterface-**B**ounded **E**vidence) on Qwen3.8-27B using rewards measured at each stage's exit. Under matched evidence, SCRIBE achieves a certified rank interval of $[1, 2]$, with certified advantages over all evaluated published pipelines and the evaluated Claude and OpenAI agents. Under same pool, SCRIBE matches the strongest published retriever and is certified above three published pipelines. The code is available at https://anonymous.4open.science/r/LitReview-SCRIBE-8E4E

1 INTRODUCTION

Leaderboards are a common way to evaluate long-horizon agents. A literature-review agent searches for papers, screens them, extracts evidence, develops claims, and writes a report. A benchmark then scores the final report. The resulting table tells us which output scored higher. But it leaves two central questions unanswered:

Are the systems comparable, and where does the difference come from?

The first question concerns the evaluation design. Systems may receive different evidence, inputs, or budgets. Their scores then reflect both these conditions and the capabilities we want to compare. More tasks, clustering, and multiplicity correction do not remove this mismatch. The second question concerns attribution. Each stage changes what the next stage receives. A better report may

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.

come from better retrieval, better writing, or both. To compare writing stages, we must account for differences in their upstream inputs. A final score alone cannot separate these effects.

We introduce *compositional controllability* to address both questions. We define a *comparison window* by its entry and exit frontiers. A window can cover one stage, several stages, or the whole agent. A common interface makes these windows comparable across different systems. For each window and readout, we define a *controlled contrast*: the score difference from a common entry under matched observation and protocol conditions. We then bound the gap between this contrast and the difference observed in an evaluation.

Our bound depends only on nuisance *outside* the window: entry mismatch, differences in exit observation, and protocol imbalance. It places no restriction on what happens inside. This distinction matters. Retrieval differences are part of the capability being tested in a whole-agent comparison. In a comparison of writing stages, those same differences affect the inputs and must be accounted for.

The bound gives an admissibility test that is applied before scores are inspected. We *refuse* a comparison if its outside nuisance cannot be bounded within a declared tolerance. For an admissible pair, we *certify* an ordering only if the score gap exceeds the sum of the sampling and nuisance radii. Otherwise, the result is *undecided*. The resulting ranking is partial. Each system receives a rank interval consistent with the certified relations. A leaderboard can therefore support an ordering only when external mismatch is controlled or bounded. Some positions may remain unresolved.

A systematic-review evaluation illustrates the problem. A production assistant reports abstract-screening sensitivity of 96.9% on 108 Cochrane reviews (Prasad, 2026). It describes this result as exceeding single-reviewer and approaching dual-reviewer human performance. The human estimates are 86.6% and 97.5%. They come from a separate trial in which crowd screeners classified $2{,}000$ abstracts from two other reviews (Gartlehner et al., 2020). These evaluations use different reviews, abstracts, and reference standards. No bound on this mismatch is stated, and no sampling radius is reported for the cross-study contrast. Two reviews also provide little basis for estimating between-review variation. Even a statistically significant gap would not resolve these differences in design. Yet the assistant's evaluation already tests individual stages against gold inputs. This is the controlled-window design we formalize. What our framework adds is an explicit admissibility test and an uncertainty radius.

We study these questions in biomedical literature review. Published human reviews provide content references against which generated reports can be scored claim by claim. Our benchmark, *BioLitBench*, contains $2{,}042$ biomedical articles, most of them reviews. It represents them as graphs with $246{,}102$ claims, $832{,}487$ typed relations between claims, $103{,}242$ hard negatives, and section hierarchies. Reviews on the same topic form set-valued references that calibrate the comparison tolerances.

We use BioLitBench to build a *certified leaderboard* of published pipelines, commercial agents, and our own agent. We evaluate two entry conditions. In the first, every system receives the target review's reference list. In the second, every system searches the same frozen corpus. The certificates concern agreement with human reference content. They do not establish how a system used its evidence. An evidence-ablation control demonstrates this limitation of the content readouts.

Certification changes which conclusions the evaluation supports. Among seven published pipelines, a cluster-robust, multiplicity-corrected analysis declares a winner in 14 of 21 pairwise comparisons. However, the top-ranked system alone received the target review's bibliography. Its score advantage therefore does not isolate pipeline quality from the benefit of supplied evidence. To compare pipelines under matched conditions, our admissibility test checks their inputs and requires a fixed backbone model. It refuses 11 of the 21 pairs, including every comparison involving the top-ranked system. Seven of the 14 statistically significant orderings fall within these refused pairs. These score differences may be statistically reliable, but they do not establish which pipeline is better under matched conditions.

The same windows also support stage-level training. A sealed entry artifact removes entry mismatch. With observation and protocol conditions matched, an exit readout can evaluate the window's contribution. This allows us to reward individual stages rather than only complete trajectories. We build *SCRIBE* (**S**tage-**C**ertified **R**eview-writing agent with **I**nterface-**B**ounded **E**vidence) around these windows. We train it with KV-Skill (Han et al., 2026) on a 27B open model. Training uses the

BioLitBench training split, which is disjoint from the 50 evaluation tasks. Under matched evidence, SCRIBE has a certified rank interval of $[1, 2]$. It has certified advantages over every evaluated published pipeline and the evaluated Claude and OpenAI agents. Its comparison with its own untrained harness remains undecided.

Our contributions are:

1. **Compositional controllability.** We define comparison windows over one stage, several stages, or the whole agent, and bound attribution error using only nuisance outside each window. This yields admissibility tests and certified partial rankings. We keep comparability separate from equivalence and substitution, and test the bound on the published pipelines themselves (Section 3; tests in Section 6.3 and Appendix F).
2. **BioLitBench.** We introduce a graph-structured benchmark of 2,042 biomedical articles with claims, typed relations, hard negatives, and section hierarchies. Same-topic reviews serve as set-valued references. A training split disjoint from the evaluation tasks lets the benchmark support both evaluation and training (Section 4).
3. **A certified leaderboard.** We compare published pipelines, commercial agents, and SCRIBE under two entry conditions: a shared reference list or a shared frozen corpus. We report certified rank intervals and explicitly identify inadmissible comparisons. These certificates establish differences in agreement with human review content (Section 6.1).
4. **SCRIBE.** We build a literature-review agent whose synthesis, planning, and writing stages each start from a sealed artifact. Each stage can then be rewarded at its own exit, rather than through the final report alone. We train SCRIBE with KV-Skill on a 27B open model and also report SCRIBE-Luna, the same harness untrained on a frontier model. Under matched evidence, SCRIBE achieves a certified rank interval of $[1, 2]$, above every evaluated published pipeline and the Claude and OpenAI agents.

2 Related work

Our machinery comes from probabilistic bisimulation (Givan et al., 2003; Ferns et al., 2004; 2011; Gelada et al., 2019) and causal abstraction (Rubenstein et al., 2017; Rischel & Weichwald, 2021; Felekis et al., 2024), which ask when one process may stand in for another. We ask the weaker question of when a contrast observed across two *separate* pipelines identifies the effect of replacing one stage, and Theorem 1 answers it with a bound in the manner of partial identification (Manski, 2003). Stage-level credit assignment (Lightman et al., 2024; Wu et al., 2026; Zhang et al., 2025) works inside one system, whereas we credit a difference between two systems to the stage in which they differ.

Survey-writing and deep-research agents (Wang et al., 2024; Yan et al., 2025; Nguye et al., 2025; Chen et al., 2025; Wang et al., 2025; Go et al., 2026; Shao et al., 2026) are benchmarked on their final reports (Su et al., 2026; Du et al., 2026). Zhu et al. (2026) argue that what such a score supports depends on the declared boundary; a comparison window is such a boundary, and the nuisance radius bounds what lies outside it. We adopt rigorous measurement practice (Miller, 2024; Bean et al., 2025; Kapoor et al., 2025; Singh et al., 2025), but no test on the scores reveals that a pair was never entitled to be compared. Confidence sets for ranks (Klein et al., 2020; Mogstad et al., 2024) give simultaneous inference on an ordering; our rank intervals are of this kind, and what we add is the admissibility test that precedes them. Other work inspects a single run, through its logs (Kirgis et al., 2026) or with an agent that grades its trajectory (Zhuge et al., 2025), rather than licensing a comparison between two systems. ADRA-Bank (Guo et al., 2025) sets stage inputs only for backbone models, whereas published pipelines expose no entry but the task.

3 Method: Compositional Controllability

Compositional controllability determines which comparisons an evaluation can support. A *comparison window* specifies the stretch of execution being compared, a *controlled contrast* defines the difference of interest, an *attribution bound* limits the gap between the observed and controlled contrasts, and a *certificate* combines this bound with sampling uncertainty. The same bound applies to one stage, several stages, or the whole agent.

An agent system $\mathcal{S}$ induces a distribution over execution histories $h \in \mathcal{H}$, since model sampling, tool calls and branching decisions all introduce randomness. An evaluation draws tasks from a distribution ρ, runs each system and scores its outputs, so all bounds below are relative to ρ and to the states reached during evaluation. We measure differences between distributions using the Wasserstein distance W_d for a bounded metric d. A Markov kernel Q is *L-Lipschitz* if $\mathsf{W}_d(\nu Q, \nu' Q) \le L\, \mathsf{W}_d(\nu, \nu')$ for the input distributions reached by the benchmark, so that L bounds how much Q can amplify a difference in its inputs. Appendix A.1 gives the formal setting.

3.1 A MOTIVATING EXAMPLE

Consider two writing modules, A and B, each receiving an evidence bundle with a fraction $e \in [0, 1]$ of the required papers. Their reports score $f_A(e) = 0.8e$ and $f_B(e) = 0.5e$, so A scores higher at any common coverage $e > 0$, but the modules belong to pipelines with different retrieval stages.

Module	Retrieved coverage	Pipeline score	Score at $e = 0.5$
A	$e_A = 0.2$	0.16	0.40
B	$e_B = 0.4$	0.20	0.25
$A - B$	—	-0.04	$+0.15$

The pipeline scores favor B by 0.04, but B receives more evidence. Given the same bundle with $e = 0.5$, A leads by 0.15, so the pipeline comparison reverses the ordering of the writing modules, and because this mismatch is systematic, more tasks do not remove it.

3.2 COMPARISON WINDOWS AND READOUTS

Windows. Different systems need not share code or internal representations, so we compare them through a *semantic interface* that maps their states into a common space; $\mathcal{Z}_M$ denotes its exit part for a window M. For a review agent, it records the evidence, the artifacts built from it, and the remaining budget.

A *comparison window* $M = [F^{\mathrm{in}}, F^{\mathrm{out}}]$ runs from an entry frontier to an exit frontier, each an antichain of histories that an execution visits at most once (Appendix A.2). For system X, all computation inside the window is represented by a *macro-kernel* $Q_X^M : F^{\mathrm{in}} \rightsquigarrow \mathcal{X}_M \times F^{\mathrm{out}}$, where $\mathcal{X}_M$ records the labeled trace, resources used and failures inside the window, and F^{out} holds the exit artifact.

A review agent works through a sequence of stages, from retrieval to writing. A *stage window* covers one of these stages and begins at its input artifact, while a *system window* covers the whole agent, from the task input to the final report. In our example, the window covers writing and begins at the evidence bundle. What lies inside the window may differ between the two systems, so a system window, which contains the backbone model, compares complete systems even across backbones.

Readouts. A benchmark evaluates a window through a readout n with score function $s_n : \mathcal{Y}_n \to [0, 1]$. The *observation channel* $C_{X,n} : \mathcal{Z}_M \rightsquigarrow \mathcal{Y}_n$ maps the window's exit state to the object being scored, including any downstream stages needed to obtain that object. When a retrieval window is scored on the final report, for instance, every later stage belongs to its channel.

Let μ_X be system X's *entry law*, the distribution of its inputs to the window. Its expected observed score is

$$\theta^{\mathrm{obs}}_{X,n} = V_n\big(\mu_X Q_X^M C_{X,n}\big), \qquad V_n(\lambda) = \int_{\mathcal{Y}_n} s_n(y)\, \lambda(dy). \tag{1}$$

3.3 CONTRASTS AND THE ATTRIBUTION BOUND

A comparison between two pipelines may change both the window and its surroundings, since the systems may enter the window with different inputs or be scored through different channels. To isolate the window, fix a common entry law μ and a common channel C_n. The *controlled contrast* and the *observed contrast* are

$$\Delta^{\star}_n(\mu, C_n) = V_n(\mu Q_A^M C_n) - V_n(\mu Q_B^M C_n), \qquad \Delta^{\mathrm{obs}}_n = \theta^{\mathrm{obs}}_{A,n} - \theta^{\mathrm{obs}}_{B,n}. \tag{2}$$

The controlled contrast changes only the window, which makes it the quantity we want to estimate, but it cannot be measured when existing pipelines cannot be run from the same entry law or through the same channel. The observed contrast uses each pipeline's actual inputs and channel, and the two can have opposite signs, as in our example, where $\Delta^\star = +0.15$ but $\Delta^{\mathrm{obs}} = -0.04$.

We bound the gap between the observed and controlled contrasts under two conditions (Assumption 1 in Appendix A.2). First, all effects of the entry on the score must pass through the window. Second, differences outside the window must have bounded effects on the score.

In the bound, the entry mismatch $\epsilon_{\mathrm{in},X}$ is the Wasserstein distance between μ_X and μ, and the channel mismatch $\epsilon_{\mathrm{cont},X,n}$ (for the continuation that forms the channel) is the largest distance between the outputs of $C_{X,n}$ and C_n from the same reachable exit. The Lipschitz constants L_{M_X} and L_{C_n} bound how much the window and the common channel amplify input differences, and the slack ξ_{M_X} allows for discontinuities such as truncation. Finally, L_{s_n} is the score's Lipschitz constant, and $\epsilon_{\mathrm{protocol},n}$ bounds any remaining protocol imbalance on the score scale.

Theorem 1 (Observable attribution bound)**.** *Under Assumption 1 and the boundary factorization of Appendix A.2,* $\left|\Delta_n^{\mathrm{obs}} - \Delta_n^\star(\mu, C_n)\right| \le \epsilon_{\mathrm{outside},n}$*, where*

$$\epsilon_{\mathrm{outside},n} := L_{s_n}\Big[L_{C_n}\big(L_{M_A}\epsilon_{\mathrm{in},A} + L_{M_B}\epsilon_{\mathrm{in},B} + \xi_{M_A} + \xi_{M_B}\big) + \epsilon_{\mathrm{cont},A,n} + \epsilon_{\mathrm{cont},B,n}\Big] + \epsilon_{\mathrm{protocol},n}, \tag{3}$$

and the slack ξ_{M_X} *is charged only when* $\mu_X \neq \mu$*; it is* 0 *when* $\mu_X = \mu$*.*

Appendix A.2 gives the proof. We call $\epsilon_{\mathrm{outside},n}$ the *nuisance radius*. It does not require the two windows to behave alike: each enters only through its own Lipschitz constant and slack.

In our example, each report is scored as soon as it is written, so the window is *directly observable* (Definition 6), which gives $L_{C_n} = L_{s_n} = 1$ and no channel or protocol mismatch. The linear modules give $\xi_{M_A} = \xi_{M_B} = 0$, $L_{M_A} = 0.8$ and $L_{M_B} = 0.5$, and relative to the common bundle $e = 0.5$ the entry mismatches are $\epsilon_{\mathrm{in},A} = 0.3$ and $\epsilon_{\mathrm{in},B} = 0.1$. The bound $\epsilon_{\mathrm{outside}} = 0.8(0.3) + 0.5(0.1) = 0.29$ therefore covers the discrepancy $|-0.04 - 0.15| = 0.19$.

3.4 Certification and rank intervals

Admissibility. We first decide whether the evaluation supports the intended comparison, using the declared conditions and the nuisance radius before inspecting the score gap.

Definition 1 (Readout-relative fair comparability)**.** *Fix a tolerance* $\epsilon_{\mathrm{fair},n} > 0$ *before observing any ranking. Windows* M_A *and* M_B *are* $\epsilon_{\mathrm{fair},n}$-fairly comparable under readout n *if Assumption 1 holds, the controlled estimand in equation 2 is declared, and* $\epsilon_{\mathrm{outside},n} \le \epsilon_{\mathrm{fair},n}$*.*

We call such a pair *admissible*; otherwise the comparison is *refused* under the current design. Fair comparability concerns only the conditions surrounding a window. The stronger claims that two windows behave alike (*equivalence*) or can replace one another with a bounded change in the final outcome (*substitution*) are defined in Appendices A.5 and A.4.

Certification. For an admissible pair, let $\widehat{\Delta}_n^{\mathrm{obs}}$ be the estimated contrast and $q_n(\delta')$ its sampling radius at confidence $1 - \delta'$. We certify the ordering given by the sign of $\widehat{\Delta}_n^{\mathrm{obs}}$ only when

$$\left|\widehat{\Delta}_n^{\mathrm{obs}}\right| > q_n(\delta') + \epsilon_{\mathrm{outside},n}. \tag{4}$$

If this condition fails, the pair remains *undecided*.

In practice, we replace $\epsilon_{\mathrm{outside},n}$ in both decisions with its upper confidence bound, $\widehat{\epsilon}_{\mathrm{outside},n} + r_n(\delta)$ (Definition 7). When entries, channels and protocols are exactly matched, there is no outside mismatch or slack, so the nuisance radius is zero and certification rests on the sampling radius alone.

In our example, the gap of 0.04 lies within the nuisance radius of 0.29, so more tasks alone cannot certify the ordering. Once both modules receive the same bundle, with the channel and protocol unchanged, $\epsilon_{\mathrm{outside}} = 0$ and $\Delta^{\mathrm{obs}} = \Delta^\star = 0.15$.

Certified relations need not form a total order, so we report each system's *rank interval*, the set of positions it can take in any total order consistent with them (Appendix A.3). A rank is *pinned* when its interval contains a single position.

4 Human Reviews as Structured Ground Truth

BioLitBench provides structured human references for evaluating generated reviews claim by claim.

Construction and quality. We parse 2,063 open-access PubMed Central articles from JATS XML, mostly reviews. Qwen3.5-9B (Team, 2025) extracts sentence-level claims with their citations and section assignments. It labels embedding-selected claim pairs with eight relation types: parallel, contrast, elaboration, evidence, mechanism, context, synthesis, and condition. Topically similar pairs from different top-level sections labeled unrelated become hard negatives. We also parse section hierarchies and resolve cited papers to PubMed identifiers and abstracts.

This process yields 2,042 graphs with section, claim, and paper nodes (Table 3 in Appendix B). The mean per-article reference resolution rate is 87.4%. A PubMedBERT-based NLI model (Gu et al., 2020), fine-tuned on MNLI (Williams et al., 2018) and MedNLI, classifies 87.6% of claims as entailed by at least one cited abstract.

References and splits. Each evaluation task has 4–9 same-topic peer reviews to capture variation in acceptable content and organization. The target review supplies the upper normalization anchor; its peers define the reference band used to calibrate tolerances (Appendix A.6). We split the articles into 1,842 training and 200 held-out examples. We select 50 evaluation tasks by clustering held-out topic embeddings and sampling proportionally across clusters. Each task restricts evidence to papers published before its cutoff.

Scoring. System outputs and human peers pass through the same observation channel. We match sentences to reference claims using embedding similarity. The threshold is the larger of 0.70 and the 95th percentile of off-topic similarities. Citations are resolved in each output's format, and headings match reference sections through fuzzy titles and shared claims. A relation is reproduced when both claims match and their sentence similarity exceeds the 10th percentile for linked reference claims.

We screen 52 candidate readouts before pairwise comparisons. The screen checks discrimination, variation, saturation, channel artifacts, and whether a readout measures quality rather than compliance. The 34 admitted readouts form six axes: retrieval (6), synthesis (6), reasoning (13), planning (2), writing (2), and form (5). Quality scores are normalized between the peers' 10th percentile and the target review's score. Style scores measure distance from the peer band. The composite weights axes equally.

Entry conditions. Under *fixed input*, every system receives the task's reference list and is evaluated on five post-retrieval axes. Under *same pool*, every system searches a frozen corpus of 26.6M papers, adding the retrieval axis. We report these conditions separately.

5 Stage-Level Training with Certified Rewards

SCRIBE (**S**tage-**C**ertified **R**eview-writing agent with **I**nterface-**B**ounded **E**vidence) combines controlled stage-level rewards with KV-Skill training on a frozen open model.

Controlled stages and rewards. After retrieval, synthesis builds a claim graph, planning assigns claims to an outline, and writing drafts the sections. Each stage has a sealed, hashed input and a directly scored output. Receipts link consecutive artifacts. Two versions of a stage entered from the same sealed input and scored at its exit with the same readout and protocol are exactly matched, so $\epsilon_{\text{outside}} = 0$ and $\Delta^{\text{obs}} = \Delta^{\star}$: the exit score can serve as that stage's reward.

Rewards use BioLitBench readouts at each stage's exit. Training examples and reference graphs come only from the training split; evaluation tasks, their references, and their peers are excluded from those examples. However, reward weights are calibrated on leaderboard tasks using the initial

agent's gap to the strongest published pipeline for each readout. Planning is trained; synthesis and writing are scored but keep their initial carriers.

Training and variants. KV-Skill (Han et al., 2026) provides external operators accessed through a lightweight interface, without adding prompt tokens or changing backbone weights. SCRIBE uses Qwen3.8-27B (Team, 2025). For planning, we register a KV-Skill carrier and optimize it with GRPO on 192 training tasks. We also report *SCRIBE (untrained)*, the harness with its initial carriers.

SCRIBE-Luna runs the same harness on gpt-5.6-luna without stage-level training. This backbone also serves six of the seven published pipelines; DR-Tulu uses its own model. When retrieval is required, SCRIBE uses the shared pool service. Its search budget is capped at the largest baseline budget for that task (Appendix B).

6 Results

We evaluate on the 50 held-out tasks of BioLitBench, which fall into 23 of the 48 subfields; the subfields serve as clusters. The primary observation point truncates each report to the task's target length (*matched length*); full length is secondary. All analysis constants are recorded with input hashes.

6.1 The certified leaderboard

Protocol. The leaderboard compares complete systems; the backbone is part of each system and is reported on every row. Within each entry condition all systems share one entry: the reference list under fixed input, or the frozen corpus under same pool. We call a published pipeline's fixed-input run *reference-fed* and its same-pool run *self-retrieving*. The nuisance radius is therefore zero, so certification here is the sampling rule; the bound's role is to show that every pair may be compared, and Section 6.2 shows it refusing pairs that may not. The sampling radius is a cluster-robust t half-width over the 23 clusters, Bonferroni-corrected over all pairs of the table. A pair needs at least 10 common tasks, and a task either system failed is dropped from that pair. The two conditions have different entry laws and are never compared with each other. Figure 1 shows the certified rank intervals; Table 1 gives the fixed-input scores (same pool: Table 4).

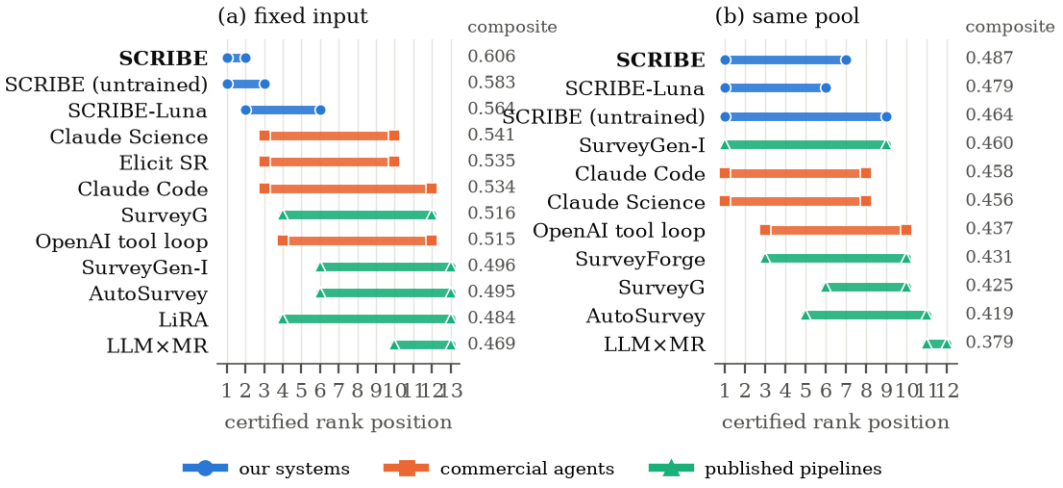


Figure 1: **Certified rank intervals.** Each bar spans all positions a system can occupy in an ordering consistent with the certified pairwise relations; a single marker denotes a pinned position. Panel (a) is fixed input and panel (b) is same pool; positions are ranked within a panel and are not comparable across panels. Systems are sorted by composite score, shown on the right. Most systems admit an interval of possible ranks rather than a unique rank.

SCRIBE. Under fixed input, SCRIBE has a certified rank interval of $[1, 2]$ and the highest composite (0.606). It is certified above every other reported system, including SCRIBE-Luna, the OpenAI tool loop, Claude Code, Claude Science and Elicit, except its own untrained harness (0.583), from which it is not separated. The gain over the untrained harness is largest on planning, the trained stage (0.319 against 0.276).

Under same pool, SCRIBE has the highest composite (0.487) and a certified interval of $[1, 7]$. It is certified above the OpenAI tool loop, AutoSurvey, SurveyG and LLM×MapReduce (LLM×MR), and is not separated from SCRIBE-Luna, SurveyGen-I or the Claude agents. Its retrieval score of 0.335 matches SurveyGen-I's 0.336.

Table 1: **Capability scores** under fixed input at the primary observation point. Rank denotes the certified rank interval shown in Figure 1. Within each column, the highest value is bold and the second highest is underlined. Our systems are shaded.

				Capability (fraction of human level)				
Backbone	System	Composite	Rank	Synthesis	Reasoning	Planning	Writing	Form
Fixed input: every system receives the review's reference list								
Qwen3.8-27B	SCRIBE	**0.606**	1–2	0.689	0.594	**0.319**	**0.613**	**0.814**
	SCRIBE (untrained)	<u>0.583</u>	1–3	0.686	0.601	<u>0.276</u>	0.565	<u>0.786</u>
gpt-5.6-luna	SCRIBE-Luna	0.564	2–6	<u>0.700</u>	**0.611**	0.177	0.568	0.767
	SurveyG	0.516	4–12	0.568	0.541	0.272	0.509	0.689
	OpenAI tool loop	0.515	4–12	**0.741**	0.552	0.136	0.487	0.660
	SurveyGen-I	0.496	6–13	0.591	0.510	0.182	0.449	0.746
	AutoSurvey	0.495	6–13	0.638	0.533	0.135	0.483	0.687
	LiRA	0.484	4–13	0.539	0.495	0.254	0.442	0.716
	LLM×MR	0.469	10–13	0.537	0.515	0.182	0.446	0.666
claude-sonnet-5	Claude Science	0.541	3–10	<u>0.700</u>	<u>0.603</u>	0.160	0.507	0.736
	Claude Code	0.534	3–12	0.650	0.547	0.185	<u>0.587</u>	0.703
proprietary	Elicit SR[§]	0.535	3–10	0.668	0.526	0.185	0.570	0.723

Published pipelines use their reference-fed configurations. [§]Elicit Systematic Review; it reads full text and cites at most 80 papers. Ranks are computed over all systems evaluated, including one commercial system that is not reported. $n = 50$ tasks in 23 subfield clusters.

Other systems. Under fixed input, Claude Science and Elicit are certified above SurveyGen-I, AutoSurvey and LLM×MR, and Claude Code above LLM×MR. The two share a backbone, so their difference comes from the product harness. The published pipelines are separated from the commercial agents, but few from one another.

What the intervals show. Most positions in both tables remain intervals, and none is pinned; a conventional leaderboard would print a distinct rank for every system. The certificate is stable: on a null with no true difference it falsely certifies near the nominal rate, and leaving out any one subfield reverses no relation (Appendix F.5). A certified relation measures agreement with the human review's content, not use of the supplied evidence (Appendix D). Under fixed input, SCRIBE stays first by point estimate at full length (Appendix C). Under same pool, SurveyGen-I retrieved the target review's own references on 33 of 50 tasks; its margins are unchanged on the other 17 (Appendix B).

6.2 Certification versus a conventional analysis

We apply a conventional analysis and our procedure to the seven published pipelines, with the backbone held fixed. Table 2 adds one requirement at a time.

The conventional analysis. R3 accounts for clustering and multiplicity, and its false-positive rate is near nominal. Yet it ranks LiRA first, and LiRA alone is given the target review's bibliography.

Table 2: **What each analysis concludes** about the 21 pairs among the seven published pipelines at the primary observation point. R4 and R5 are counted over the 10 admissible pairs.

Rule		Pairs decided	What it adds
R0	sorted table	21 / 21	any difference counts
R1	paired t-test	17 / 21	sampling uncertainty
R2	cluster-robust t	16 / 21	task clustering
R3	+ Bonferroni	**14** / 21	multiplicity; a careful conventional analysis
R4	+ admissibility test	7 / 10	11 pairs refused before any score is read
R5	+ nuisance radius	**7** / 10	certified; the radius is zero, entries are shared

The system it ranks last, DR-Tulu, runs its own model. Significance alone does not show that a comparison isolates the component of interest.

The admissibility test. Our test refuses 11 of the 21 pairs before any score is read: the 6 pairs with DR-Tulu, whose backbone differs, and the 5 that pair LiRA's entry with a self-retrieving one. These refusals follow from Eq. 3. For the LiRA pairs, the measured entry mismatch (0.94–0.98), $L_M = 0.890$ and the score constants L_{s_n} (median 0.83; $L_C = 1$) give a radius of 0.70–0.73 per readout, against a pre-registered tolerance with median 0.21 (Appendix F.3). A backbone change lies outside a scaffold window, so the DR-Tulu pairs have no finite radius. Seven of the 14 conventional orderings fall among them.

What the admissible pairs show. With entries matched, the pipelines separate. Among the five self-retrieving pipelines, 7 of 10 pairs certify, with SurveyGen-I pinned first and LLM×MR last. Retrieval separates them most: it certifies 11 of 21 same-pool pairs in the capability analysis, and writing none (Appendix E). Among the reference-fed configurations only 1 of 10 pairs certifies. Every undecided admissible pair is limited by sampling, not nuisance.

One declared convention changes the winner. At matched length, AutoSurvey is third among the reference-fed configurations and nothing is pinned first. At full length, 6 of 10 pairs certify and AutoSurvey is pinned first (Table 9); its reports have a median of about 250,000 words, against 10,000–47,000 for the others. Both views are valid, but they answer different questions, so the convention must be declared before the comparison is read.

6.3 The theory on real systems

Table 10 in Appendix F summarizes the tests of the theory. The attribution bound is tested on the published systems in two designs and holds in both, and it decides the five LiRA refusals above. It is informative on the writing axis, where no cell is vacuous, and looser on the others; non-expansiveness depends on the declared length convention (Appendix F.2). Published pipelines cannot be re-entered at an intermediate stage, so the multi-stage composition bound is checked on synthetic chains.

7 Conclusion

Compositional controllability decides, before any score is read, whether two agents may be compared at a chosen window, and certifies an ordering only when the gap exceeds the radius. On BioLitBench, a careful conventional analysis orders 14 of the 21 published-pipeline pairs; our test refuses 11 pairs, and 7 of the 14 orderings fall among them. The same windows give stage-level rewards: SCRIBE, trained with KV-Skill on Qwen3.8-27B, reaches a certified rank interval of $[1, 2]$ under fixed input, above every published pipeline and every commercial agent we report.

The results have limits. A certified relation measures agreement with the human review's content, not use of the supplied evidence, and compares systems up to an exit, not modules in isolation. The length convention can change the winner, so it must be fixed in advance.

AI USE STATEMENT

AI assistants were used to support manuscript organization and language revision, and to implement and debug code. Language models that are part of the research itself, for building BioLitBench and as the evaluated agents, are described in Sections 4–6. The authors reviewed and modified all generated suggestions, verified the experimental results and citations, and take full responsibility for the final manuscript and released artifacts.

ETHICS STATEMENT

This work uses published biomedical literature. It involves no human subjects, patient data or private information. BioLitBench is built from open-access articles in PubMed Central, whose licences differ: 1,119 of the 2,063 source articles are CC BY or CC0 and can be redistributed, while the rest, including NIH author manuscripts, allow text mining but not redistribution. We will release the redistributable subset together with scripts that rebuild the full benchmark from PubMed Central. The benchmark measures agreement with a human review's content, not the correctness or clinical safety of a report (Appendix D). A high certified rank is therefore not evidence that a generated review is fit for clinical or policy use, and generated reviews should be checked by experts before any such use. Commercial systems were accessed through their public interfaces under their terms of service and were evaluated under the same declared conditions as every other system. We report every refusal and undecided pair rather than a total order, so that the leaderboard is not read as stronger evidence than it provides.

REPRODUCIBILITY STATEMENT

Code, the frozen leaderboard inputs and the analysis scripts are available at the anonymous repository linked in the abstract. Section 3 and Appendix A state every assumption and give complete proofs. Section 4 describes how BioLitBench is built and scored, including the readout admission screen. Section 5 and Appendix B describe SCRIBE, its backbone, carrier size, training and retrieval harness. Every published pipeline ran behind a recording proxy that logs each model call and a network guard that allows only the local corpus service; the per-task logs, integrity hashes of each pipeline's code and the cost ledgers are part of the release. The decision constants are recorded with input hashes (Appendix B), so a reader can rerun the analysis and see what changing them does. The leaderboard tables and Figure 1 are generated from the frozen board by script, and an automated check verifies every table entry against it. Appendix F gives the setting of each theory check.

REFERENCES

Babak Barazandeh, Subhabrata Majumdar, and George Michailidis. Otap:structure-aware optimal transport for evaluating planning and execution in agent trajectories. *CoRR*, abs/2607.17082, 2026. doi: 10.48550/ARXIV.2607.17082. URL https://doi.org/10.48550/arXiv.2607.17082.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle K. Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob N. Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/

2025/hash/1967e0fc3aa6cbbace562f5cb8e3954e-Abstract-Datasets_and_Benchmarks_Track.html.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov decision processes. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 10069–10076. AAAI Press, 2020. doi: 10.1609/AAAI.V34I06.6564. URL https://doi.org/10.1609/aaai.v34i06.6564.

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 Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *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, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pp. 3687–3714. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.IJCNLP-LONG.193. URL https://doi.org/10.18653/v1/2025.ijcnlp-long.193.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, pp. 42414–42448, 2026.

Yorgos Felekis, Fabio Massimo Zennaro, Nicola Branchini, and Theodoros Damoulas. Causal optimal transport of abstractions. In Francesco Locatello and Vanessa Didelez (eds.), *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, volume 236 of *Proceedings of Machine Learning Research*, pp. 462–498. PMLR, 2024. URL https://proceedings.mlr.press/v236/felekis24a.html.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In David Maxwell Chickering and Joseph Y. Halpern (eds.), *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004*, pp. 162–169. AUAI Press, 2004. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1103&proceeding_id=20.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM J. Comput.*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X. URL https://doi.org/10.1137/10080484X.

Gerald Gartlehner, Lisa Affengruber, Viktoria Titscher, Anna Noel-Storr, Gordon Dooley, Nicolas Ballarini, and Franz König. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of clinical epidemiology*, 121:20–28, 2020.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179. PMLR, 2019. URL http://proceedings.mlr.press/v97/gelada19a.html.

Robert Givan, Thomas L. Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artif. Intell.*, 147(1-2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4. URL https://doi.org/10.1016/S0004-3702(02)00376-4.

Gregory Hok Tjoan Go, Khang Ly, Anders Søgaard, Seyed Amin Tabatabaei, Maarten de Rijke, and Xinyi Chen. Lira: A multi-agent framework for reliable and readable literature review generation. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore,*

January 20-27, 2026, pp. 40456–40464. AAAI Press, 2026. doi: 10.1609/AAAI.V40I47.41489. URL https://doi.org/10.1609/aaai.v40i47.41489.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779, 2020. URL https://arxiv.org/abs/2007.15779.

Zhihan Guo, Feiyang Xu, Yifan Li, Muzhi Li, Shuai Zou, Jiele Wu, Han Shi, Haoli Bai, Ho-fung Leung, and Irwin King. Adra-bank: A modular benchmark for academic deep research agents. *arXiv preprint arXiv:2512.00986*, 2025.

Zhaowei Han, Xiang Zhang, Bing Han, Kai Liu, Danqi Hu, and Jie Liu. Kv-skill: Forging expertise in the model's native language. *arXiv preprint arXiv:2608.05475*, 2026.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Peter Kirgis, Sayash Kapoor, Stephan Rabanser, Nitya Nadgir, Cozmin Ududec, Magda Dubois, JJ Allaire, Conrad Stosz, Marius Hobbhahn, Jacob Steinhardt, and Arvind Narayanan. Log analysis is necessary for credible evaluation of AI agents. *CoRR*, abs/2605.08545, 2026. doi: 10.48550/ARXIV.2605.08545. URL https://doi.org/10.48550/arXiv.2605.08545.

Martin Klein, Tommy Wright, and Jerzy Wieczorek. A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):589–606, 2020.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *CoRR*, abs/2411.00640, 2024. doi: 10.48550/ARXIV.2411.00640. URL https://doi.org/10.48550/arXiv.2411.00640.

Magne Mogstad, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, 91(1):476–518, 2024.

Minh-Anh Nguye, Minh-Duc Nguyen, Ha Lan N. T., Hai-Dang Kieu, Nguyen Tien Dong, and Dung D. Le. Surveyg: A multi-agent LLM framework with hierarchical citation graph for automated survey generation. *CoRR*, abs/2510.07733, 2025. doi: 10.48550/ARXIV.2510.07733. URL https://doi.org/10.48550/arXiv.2510.07733.

Pradyumna Prasad. Evaluating Elicit's systematic literature review capabilities. Elicit blog, https://elicit.com/blog/evaluating-elicit-slr, May 2026. Published 6 May 2026; accessed 25 September 2026.

Eigil Fjeldgren Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1013–1023. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/rischel21a.html.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In Gal Elidan, Kristian Kersting, and Alexander Ihler (eds.), *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR tulu: Reinforcement learning with evolving rubrics for deep research. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=97NEP1pyS3.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/70a93f260a51123b3c0e33ecd1b4de97-Abstract-Datasets_and_Benchmarks_Track.html.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. Surge: A benchmark and evaluation framework for scientific survey generation. In Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (eds.), *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, pp. 3404–3412. ACM, 2026. doi: 10.1145/3805712.3808598. URL https://doi.org/10.1145/3805712.3808598.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 2019. URL http://proceedings.mlr.press/v97/titouan19a.html.

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. *CoRR*, abs/2504.05732, April 2025. URL https://doi.org/10.48550/arXiv.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. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d07a9fc7da2e2ec0574c38d5f504d105-Abstract-Conference.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018,*

New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1101. URL https://doi.org/10.18653/v1/n18-1101.

Shirley Wu, Parth Sarthi, Shiyu Zhao, Aaron Lee, Herumb Shandilya, Adrian Mladenic Grobelnik, Nurendra Choudhary, Edward Huang, Karthik Subbian, Linjun Zhang, Diyi Yang, James Zou, and Jure Leskovec. Optimas: Optimizing compound AI systems with globally aligned local rewards. In *International Conference on Learning Representations*, pp. 149685–149706, 2026.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. SURVEYFORGE : On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 12444–12465. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.609. URL https://doi.org/10.18653/v1/2025.acl-long.609.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Pengyu Zhu, Li Sun, Philip S Yu, and Sen Su. "LLM agent performance" is not a single evaluation target. *arXiv preprint arXiv:2602.03238*, 2026.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A Theory

This appendix collects the formal setting behind Section 3 and the proofs of its results, ordered by what the main text needs. Section A.1 fixes the measure-theoretic conventions and the model of an agent system. Section A.2 states Assumption 1 in full and proves Theorem 1. Section A.3 turns that bound into a finite-sample certificate. The remaining sections go beyond what the experiments exercise: Section A.4 carries error across several stages, Section A.5 builds the behavioural distance in which equivalence and substitution are stated, and Section A.6 formalises the human review graph the benchmark scores against. Appendix B validates the campaign. Appendix C gives the same-pool scores and the full-length leaderboard. Appendix D tests the readouts by evidence ablation, Appendix E reports the capability campaign, and Appendix F the theory checks, and Appendix G records where the framework stops.

A.1 Formal setting

Measurable spaces and kernels. All state, action, observation and interface spaces are standard Borel spaces, which guarantees regular conditional distributions and rules out measure-theoretic pathologies. A Markov kernel $K : X \rightsquigarrow Y$ assigns to each $x \in X$ a probability measure $K(\cdot \mid x)$ on Y, measurably in x. A probability measure μ on X is pushed through K by

$$(\mu K)(A) = \int_X K(A \mid x)\, \mu(dx), \qquad A \subseteq Y, \tag{5}$$

and for a measurable map $f : X \to Y$ we write $f_{\#}\mu$ for the ordinary pushforward.

Couplings and optimal transport. For probability measures μ on X and ν on Y, let $\Pi(\mu, \nu)$ be the set of couplings with those marginals. Given a bounded non-negative cost $d : X \times Y \to \mathbb{R}_{\geq 0}$,

$$\mathsf{W}_d(\mu, \nu) = \inf_{\Gamma \in \Pi(\mu,\nu)} \int_{X \times Y} d(x, y)\, \Gamma(dx, dy). \tag{6}$$

When $X = Y$ and d is a metric this is the 1-Wasserstein distance. We keep the same notation when X and Y differ but share a meaningful ground cost, which is what lets two systems with different internal spaces be compared at all. Several proofs below select optimal couplings measurably; Assumption 4 (Appendix A.5) is assumed throughout this appendix.

Amplification factors. A Markov kernel $Q : X \rightsquigarrow Y$ has *amplification factor* at most L on a family $\mathfrak{P}$ of input laws if $\mathsf{W}_{d_Y}(\nu Q, \nu' Q) \leq L\, \mathsf{W}_{d_X}(\nu, \nu')$ for all $\nu, \nu' \in \mathfrak{P}$. Then L is a Lipschitz constant of $\nu \mapsto \nu Q$ on $\mathfrak{P}$, and this appendix uses the two names interchangeably. The family $\mathfrak{P}$ is always the laws the benchmark actually reaches, never all probability measures. Where a kernel has discontinuities, such as a truncation or a top-k cut-off, we use the affine form $\mathsf{W}_{d_Y}(\nu Q, \nu' Q) \leq L\, \mathsf{W}_{d_X}(\nu, \nu') + \xi$ with a slack $\xi \geq 0$. Factors multiply along a chain: if Q_1 and Q_2 have factors L_1 and L_2 on the laws each receives, then $Q_1 Q_2$ has factor at most $L_1 L_2$.

Pseudometrics. A pseudometric is non-negative, symmetric and satisfies the triangle inequality, but may assign distance zero to distinct objects. Here that is a feature: two histories can be operationally different and still indistinguishable at the interface an evaluation declares.

Stochastic agent systems. The main text treats an agent system as a distribution over executions. Formally:

Definition 2 (Stochastic agent system)**.** *A finite-horizon stochastic agent system is a tuple* $\mathcal{S} = (\mathcal{C}, \mathcal{H}, \mathcal{A}, \mathcal{O}, \iota, \pi, K, H)$*, where*

- $\mathcal{C}$ *is a space of task contexts;*
- $\mathcal{H} = \bigsqcup_{t=0}^{H} \mathcal{H}_t$ *is a disjoint union of history spaces,* $\mathcal{H}_t$ *holding histories of length* t*;*
- $\mathcal{A}$ *and* $\mathcal{O}$ *are action and observation spaces;*
- $\iota : \mathcal{C} \rightsquigarrow \mathcal{H}_0$ *is an initial-history kernel;*
- $\pi : \mathcal{H} \rightsquigarrow \mathcal{A}$ *is a policy kernel supported on the admissible actions at each history;*
- $K : \mathcal{H} \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{H}$ *is an environment-and-tool kernel whose next-history component lies in* $\mathcal{H}_{t+1}$ *when the current history lies in* $\mathcal{H}_t$*;*
- $H < \infty$ *is the maximum horizon.*

Histories that terminate before H *are made absorbing.*

The context holds every externally controlled variable relevant to a comparison. For a literature-review task it may be $c = (q, \mathcal{D}, \mathcal{R}, t_{\max}, B, \mathcal{U})$: the research question, the accessible corpus, scope and writing requirements, a publication cutoff, a resource budget and the available tools. The formulation is deliberately more general than a Markov decision process. If a sufficient state summary exists it may replace the history; if none does, nothing here assumes one.

Branch kernels and computation trees. Two systems can carry out the same step through very different internal actions, so their executions are compared through a shared vocabulary of labels rather than through those actions. Let $(\mathcal{L}, d_{\mathcal{L}})$ be a common label space, whose labels record externally meaningful facts about an executed branch — stage, action class, tool class, resource use and observable effect — and let each system carry a measurable labelling map $\ell_{\mathcal{S}} : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{H} \to \mathcal{L}$. The induced *branch kernel* $P_{\mathcal{S}} : \mathcal{H} \rightsquigarrow \mathcal{L} \times \mathcal{H}$ is

$$P_{\mathcal{S}}(D \mid h) = \int_{\mathcal{A}} \int_{\mathcal{O} \times \mathcal{H}} \mathbf{1}\big\{(\ell_{\mathcal{S}}(h, a, o, h'), h') \in D\big\}\, K(do, dh' \mid h, a)\, \pi(da \mid h) \tag{7}$$

for measurable $D \subseteq \mathcal{L} \times \mathcal{H}$. Sampling from $P_{\mathcal{S}}$ recursively unfolds a rooted stochastic computation tree whose branches are all possible executions; a rollout is one sampled root-to-leaf path. Parallel or dependency-structured executions can be written as DAGs and unfolded into histories without changing the probability distribution over executions.

Semantic interfaces. Systems need not share states or actions, but they must expose comparable artifacts.

Definition 3 (Semantic interface)**.** *Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a common metric space. A* semantic interface *for $\mathcal{S}$ is a measurable map $\phi_{\mathcal{S}} : \mathcal{H} \to \mathcal{Z}$. It is* comparison-sufficient *for a family $\mathcal{F}$ of evaluations if every $f \in \mathcal{F}$ factors through it: $f(h) = \bar{f}(\phi_{\mathcal{S}}(h))$ for some $\bar{f} : \mathcal{Z} \to \mathbb{R}$.*

For a review agent an interface state is $z = (P, E, G, O, b, r)$: a paper set, an evidence ledger, a synthesis graph, an outline, the remaining budget and a vector of stage-level verification results. Private chain-of-thought need not be observed.

A.2 Windows, readouts and the attribution bound

Section 3 introduces windows and readouts in enough detail to carry its argument. This section gives their precise forms, states Assumption 1 in full, and proves Theorem 1.

Frontiers and windows. A stage is delimited by when an execution stops, not by where one source file ends.

Definition 4 (Execution frontier)**.** *An* execution frontier *$F \subseteq \mathcal{H}$ is a measurable antichain of histories: no member is a strict prefix of another. An execution reaches F at the stopping time $\tau_F = \inf\{t : h_t \in F\}$.*

Definition 5 (Comparison window)**.** *A* comparison window *$M = [F^{\mathrm{in}}, F^{\mathrm{out}}]$ is an entry frontier and a later exit frontier such that every evaluated execution entering F^{in} reaches F^{out}, or an explicitly labelled failure state, within a bounded number of steps. It induces the* macro-kernel

$$Q_{\mathcal{S}}^{M} : F^{\mathrm{in}} \rightsquigarrow \mathcal{X}_M \times F^{\mathrm{out}}, \tag{8}$$

where $\mathcal{X}_M$ records the labelled local trace, cumulative resources and failure indicators inside the window.

The internal trace may be a sequence, a dependency DAG, or its tree unfolding. Let d_M be a bounded, structure-aware discrepancy on $\mathcal{X}_M \times F^{\mathrm{out}}$ — for instance the recursive distance of Section A.5, a tree edit distance, or an optimal-transport distance on attributed execution DAGs Barazandeh et al. (2026). Entry histories are mapped into a common boundary space $(\mathcal{B}_M, d_{\mathcal{B}_M})$ by maps ψ_X^{in}, and exits into a common space $(\mathcal{Z}_M, d_M^{\mathrm{out}})$. Throughout, $\mathcal{R}_M(\rho) \subseteq \mathcal{Z}_M$ is the set of exit states reachable under the benchmark distribution ρ.

Assumption 1 applies Q_X^M to laws on $\mathcal{B}_M$, meaning the exit-interface marginal of equation 8 after the two interface maps. Writing it as a kernel on $\mathcal{B}_M$ rather than on entry histories presupposes a *boundary factorization*: the exit law depends on the entry history h only through $\psi_X^{\mathrm{in}}(h)$, so that $Q_X^M(\cdot \mid h) = Q_X^M(\cdot \mid \psi_X^{\mathrm{in}}(h))$ for almost every entry history reached under ρ. This is stronger than boundary sufficiency, condition (i) below, which only places the confounders in the boundary: it also rules out any other route from the entry history to the exit. It is assumed throughout this subsection.

The identification assumption. Assumption 1 has two halves, which do different work. Conditions (i)–(iv) make the window the only thing the evaluation credits; the constants in (v) bound, on the score scale, everything else that differs between the two pipelines.

Assumption 1 (Identification and nuisance bounds)**.** On the declared benchmark support:

(i) *Boundary sufficiency.* Every upstream variable causally influencing both the selected module and its evaluated outcome lies in the entry boundary or is fixed by protocol.

(ii) *No hidden bypass.* Every evaluated causal path from the entry frontier to the exit artifact passes through the window or appears as an explicit branch of its macro-kernel.

(iii) *Resource accounting.* Tokens, tool calls, latency, monetary cost and accessible tools that can affect behaviour are matched, or carried in the branch labels and the boundary metric.

(iv) *Common exit semantics.* Both systems expose exits in $(\mathcal{Z}_M, d_M^{\mathrm{out}})$, and d_M dominates exit discrepancy: $d_M((x,z),(x',z')) \geq d_M^{\mathrm{out}}(z,z')$.

(v) *Nuisance bounds.* For $X \in \{A,B\}$ there are constants with

$$\begin{aligned} \mathsf{W}_{d_{\mathcal{B}_M}}(\mu_X, \mu) &\leq \epsilon_{\mathrm{in},X}, \\ \mathsf{W}_{d_{\mathcal{Y}_n}}\big(C_{X,n}(\cdot \mid z), C_n(\cdot \mid z)\big) &\leq \epsilon_{\mathrm{cont},X,n} \quad \text{for all } z \in \mathcal{R}_M(\rho), \\ \mathsf{W}_{d_M^{\mathrm{out}}}(\nu Q_X^M, \nu' Q_X^M) &\leq L_{M_X} \mathsf{W}_{d_{\mathcal{B}_M}}(\nu,\nu') + \xi_{M_X}, \\ \mathsf{W}_{d_{\mathcal{Y}_n}}(\nu C_n, \nu' C_n) &\leq L_{C_n} \mathsf{W}_{d_M^{\mathrm{out}}}(\nu,\nu'); \end{aligned}$$

s_n is L_{s_n}-Lipschitz, and $\epsilon_{\mathrm{protocol},n}$ bounds on the score scale all remaining resource, alignment, annotation and protocol imbalance. The last two displays need hold only for the laws ν, ν' that occur on the declared benchmark support, including the declared common entry law μ and its image μQ_X^M, not for arbitrary pairs of measures. The slack $\xi_{M_X} \geq 0$ absorbs discontinuities such as truncation or a top-k cut-off, as ξ_k does in Assumption 3 (Appendix A.4); $\xi_{M_X} = 0$ is the pure Lipschitz case.

Each of (i)–(iv) is necessary, and none asks the two modules to behave alike. If boundary sufficiency fails, a local difference can be an upstream confound. If a path bypasses the window, credit for it is misattributed to the window. If resources go unaccounted, an apparent improvement may be additional computation. If exit semantics differ, there is nothing well-defined to compare.

Condition (v) is of a different kind. Conditions (i)–(iv) are qualitative, and if one fails the bound identifies nothing. Condition (v) is quantitative: it asks that the constants exist but places no limit on their size. Theorem 1 holds whatever they are, and their size decides only whether the resulting radius is small enough to pass the tolerance of Definition 1. Its constants L_{M_X} bound each module's sensitivity to a perturbation of its entry; they do *not* bound the behavioural difference between M_A and M_B.

Readouts. For one observed $y \sim \mu_X Q_X^M C_{X,n}$, $s_n(y)$ is a number. Because the window is stochastic, the observed value $\theta_{X,n}^{\mathrm{obs}}$ of equation 1 is a population expectation, and must in general be estimated (Section A.3). A failure or missing readout is recorded as a declared symbol $\perp_n$ with a prespecified score, never by discarding the failed run. A readout index n may encode a stage, a document section and a criterion, for instance $n = (\text{retrieval}, \text{introduction}, \text{citation coverage})$.

Definition 6 (Direct observability)**.** *Fix a canonical measurable readout map $r_n : \mathcal{Z}_M \to \mathcal{Y}_n$. A window M_X is* directly observable under n *if the readout is available at its exit frontier and $C_{X,n}(\cdot \mid z) = \delta_{r_n(z)}$ on the reachable exit support. Write $\mathfrak{D}_n$ for the class of such windows. Equivalently, the score factors through the common exit interface with no downstream stage.*

Whether a window is directly observable depends on the window, the interface and the readout map. It does not depend on which runs happened to produce a readout: a window does not become directly observable because its failed runs are set aside.

Proof of the attribution bound.

Proof of Theorem 1. Write the discrepancy as a difference of one term per system:

$$\Delta_n^{\mathrm{obs}} - \Delta_n^{\star} = \big[V_n(\mu_A Q_A^M C_{A,n}) - V_n(\mu Q_A^M C_n)\big] - \big[V_n(\mu_B Q_B^M C_{B,n}) - V_n(\mu Q_B^M C_n)\big].$$

By the triangle inequality it is enough to bound each bracket and add.

Fix X. Its natural pipeline differs from the controlled one in two places, the entry law and the channel. Insert the law that changes only the channel, $\mu_X Q_X^M C_n$:

$$\begin{aligned} V_n(\mu_X Q_X^M C_{X,n}) - V_n(\mu Q_X^M C_n) = &\underbrace{V_n(\mu_X Q_X^M C_{X,n}) - V_n(\mu_X Q_X^M C_n)}_{\text{channel}} \\ &+ \underbrace{V_n(\mu_X Q_X^M C_n) - V_n(\mu Q_X^M C_n)}_{\text{entry}}. \end{aligned}$$

Step 1: from scores to distributions. Because s_n is L_{s_n}-Lipschitz, Kantorovich–Rubinstein duality gives $|V_n(\lambda) - V_n(\lambda')| \le L_{s_n} \mathsf{W}_{d_{\mathcal{Y}_n}}(\lambda, \lambda')$ for any two laws on $\mathcal{Y}_n$. Each term is therefore at most L_{s_n} times a transport distance.

Step 2: the channel term. Both laws push the same exit law $\mu_X Q_X^M$, which is supported on $\mathcal{R}_M(\rho)$, through different channels. Coupling the two channels optimally at each exit z and integrating over z gives $\mathsf{W}(\mu_X Q_X^M C_{X,n}, \mu_X Q_X^M C_n) \le \epsilon_{\mathrm{cont},X,n}$.

Step 3: the entry term. Both laws share the window and the channel and differ only in entry. If $\mu_X = \mu$ the two laws are identical and the term is zero, with no appeal to L_{M_X} or ξ_{M_X}: the slack is charged only for a system whose entry differs from μ. Otherwise, apply the Lipschitz bound of C_n and then that of Q_X^M with $\nu = \mu_X$ and $\nu' = \mu$:

$$\begin{aligned}\mathsf{W}(\mu_X Q_X^M C_n, \mu Q_X^M C_n) &\le L_{C_n} \mathsf{W}(\mu_X Q_X^M, \mu Q_X^M) \le L_{C_n}\big(L_{M_X} \mathsf{W}(\mu_X, \mu) + \xi_{M_X}\big) \\ &\le L_{C_n}\big(L_{M_X} \epsilon_{\mathrm{in},X} + \xi_{M_X}\big).\end{aligned}$$

Step 4: assemble. Each bracket is at most $L_{s_n}\big(L_{C_n}(L_{M_X}\epsilon_{\mathrm{in},X} + \xi_{M_X}) + \epsilon_{\mathrm{cont},X,n}\big)$. Summing over $X \in \{A, B\}$ and adding $\epsilon_{\mathrm{protocol},n}$, which covers every imbalance the named constants do not, gives equation 3. □

Corollary 1 (Non-expansive windows)**.** *If $L_{M_A}, L_{M_B} \le 1$ the windows do not amplify entry mismatch, and the bracket in equation 3 is at most $L_{C_n}(\epsilon_{\mathrm{in},A} + \epsilon_{\mathrm{in},B} + \xi_{M_A} + \xi_{M_B}) + \epsilon_{\mathrm{cont},A,n} + \epsilon_{\mathrm{cont},B,n}$. Its multi-stage analogue is Corollary 2. Estimating L_{M_X} requires entering the same window from two different entry laws. For a system window, which begins at the task input, this is possible: the benchmark sets the task and the searchable corpus, and Appendix F.1 estimates L_{M_X} for the published pipelines by entering each twice on the same task. For a stage window of a published pipeline it is not, because a black-box pipeline lets the stage's input artifact be observed but not set (Appendix G).*

When terms vanish. If $M_A, M_B \in \mathfrak{D}_n$ and share a canonical readout, both channel terms are zero and the bound reduces to its entry, slack and protocol terms. If one or both windows are read through a downstream stage, they remain certifiably comparable provided the channels end in the same observation space with common semantics and $\epsilon_{\mathrm{outside},n} \le \epsilon_{\mathrm{fair},n}$. Two identical channels contribute no channel mismatch, however long, though a large L_{C_n} still amplifies whatever entry mismatch remains. If in addition both entry laws equal μ, Step 3 charges neither entry nor slack, and the radius reduces to $\epsilon_{\mathrm{protocol},n}$. With the protocol matched too, $\epsilon_{\mathrm{outside},n} = 0$: this is the exact regime of Section 6.1.

A.3 Estimation and certification

Theorem 1 bounds the gap between two population quantities. An evaluation observes neither: it sees finitely many rollouts on finitely many tasks. This section says what can be certified from them.

A finite-sample certificate. Let $c_1, \dots, c_N \overset{\text{iid}}{\sim} \rho$ be task contexts. A context may carry several paired interventions and repeated rollouts, but the context is the statistical cluster. Let $\ell(c) \in [0,1]$ be an ideal context-level loss with population mean $\bar{\ell} = \mathbb{E}_{c\sim\rho}[\ell(c)]$, let $\ell_i \in [0,1]$ be its rollout-based estimate on c_i, and set $\widehat{\bar{\ell}}_N = N^{-1}\sum_i \ell_i$. Let η_{roll}, η_{align} and η_{ann} be deterministic bounds on finite-rollout approximation, learned alignment error and reference annotation error, whose sum bounds $\big|N^{-1}\sum_i \ell_i - N^{-1}\sum_i \ell(c_i)\big|$. These bounds are *assumed* rather than derived. The rollout error in particular is random, so a deterministic bound on it is a modelling assumption that can hold only with high probability.

Proposition 1 (Finite-sample upper certificate)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\bar{\ell} \le \widehat{\bar{\ell}}_N + \sqrt{\frac{\log(1/\delta)}{2N}} + \eta_{\mathrm{roll}} + \eta_{\mathrm{align}} + \eta_{\mathrm{ann}}. \tag{9}$$

Proof. The ideal losses $\ell(c_i)$ are independent and lie in $[0,1]$, so Hoeffding's inequality gives, with probability at least $1-\delta$, $\bar{\ell} \le N^{-1}\sum_i \ell(c_i) + \sqrt{\log(1/\delta)/(2N)}$. Replacing the ideal empirical

mean by $\widehat{\widehat{\ell}}_N$ costs at most $\eta_{\text{roll}} + \eta_{\text{align}} + \eta_{\text{ann}}$ by assumption, and the triangle inequality adds the two. □

Repeated rollouts of one context do not increase N: they tighten η_{roll}, not the sampling term. In practice η_{roll} can be estimated from repeated runs and a hierarchical bootstrap reported beside equation 9. The certificate bounds whatever loss ℓ measures. When ℓ is a controllability loss it certifies equivalence or a substitution component — but not fairness, because such a loss includes the target difference inside the window.

Certifying a comparison. Fairness is certified from an estimate of the nuisance radius itself.

Definition 7 (Empirically certified fair comparability)**.** *Let $\widehat{\epsilon}_{\text{outside},n}$ estimate equation 3, with a one-sided radius $r_n(\delta)$ such that $\mathbb{P}\big[\epsilon_{\text{outside},n} \le \widehat{\epsilon}_{\text{outside},n} + r_n(\delta)\big] \ge 1-\delta$. The pair is* empirically certified $\epsilon_{\text{fair},n}$-fair under n *at confidence* $1-\delta$ *if* $\widehat{\epsilon}_{\text{outside},n} + r_n(\delta) \le \epsilon_{\text{fair},n}$.

The precision of the score contrast is a separate matter. If $\widehat{\Delta}_n^{\text{obs}}$ has two-sided sampling radius $q_n(\delta')$, then Theorem 1 gives, with probability at least $1-\delta'$,

$$\Delta_n^\star \in \big[\widehat{\Delta}_n^{\text{obs}} - q_n(\delta') - \epsilon_{\text{outside},n},\ \widehat{\Delta}_n^{\text{obs}} + q_n(\delta') + \epsilon_{\text{outside},n}\big], \tag{10}$$

and a direction is certified when this interval excludes zero, which is equation 4. In practice $\epsilon_{\text{outside},n}$ is unknown and is replaced by its upper bound $\widehat{\epsilon}_{\text{outside},n} + r_n(\delta)$, so a direction is certified when $\big|\widehat{\Delta}_n^{\text{obs}}\big| > q_n(\delta') + \widehat{\epsilon}_{\text{outside},n} + r_n(\delta)$. The interval then contains $\Delta_n^\star$ whenever both the sampling event and the event of Definition 7 hold, which by a union bound has probability at least $1-\delta-\delta'$. A certified direction is therefore wrong with probability at most $\delta+\delta'$. In exact cells the radius is zero by construction rather than estimated, so $r_n = 0$ and only δ' is spent.

A pair that fails this test is *undecided*, for one of two reasons. It is *attribution-limited* when the population gap $|\Delta_n^{\text{obs}}|$ lies within the nuisance radius $\epsilon_{\text{outside},n}$, so that more samples alone cannot certify the ordering, and *sampling-limited* when the gap exceeds the nuisance radius but sampling uncertainty prevents certification.

Rank intervals. Certified relations define a partially resolved relation rather than a total order. Because pair-specific radii differ, the certified relation need not be transitive, and in principle it could contain a cycle. When it is acyclic, its transitive closure is a strict partial order, and a system's *rank interval* is the range of positions it takes across all linear extensions of that order. Provided the sampling radii are valid, the family-wise correction makes every certified relation hold simultaneously with the stated probability, and the relations the closure adds hold on the same event, since true utilities are ordered transitively. Appendix F.5 checks the realised rate against a sign-flip null.

A.4 Composition across stages

Theorem 1 concerns one window. A pipeline has several stages, and a difference introduced at one is carried — and possibly amplified — by every stage after it. This section bounds that propagation. It is stated in general; testing it needs every stage to be run from a controlled input, which published pipelines do not allow: their intermediate artifacts can be recovered from logs but not set (Appendix G).

Setting. Let $(\mathcal{Z}_k, d_k)$, $k = 0, \dots, K$, be the interfaces between K stages. System $X \in \{A, B\}$ has stage kernels $Q_{X,k} : \mathcal{Z}_{k-1} \rightsquigarrow \mathcal{Z}_k$ and initial law $\nu_{X,0}$, and its law at interface k is $\nu_{X,k} = \nu_{X,k-1} Q_{X,k}$. Let $\mathcal{R}_k(\rho) \subseteq \mathcal{Z}_k$ be the interface states reachable under the benchmark and its declared interventions, and $\mathfrak{P}_k(\rho)$ an admissible family of laws supported there. Every law below, including those in the proofs, is assumed admissible; this avoids claiming stability on histories no benchmark reaches.

Assumption 2 (Local replacement error)**.** For each k there is $\epsilon_k \ge 0$ with $\sup_{z \in \mathcal{R}_{k-1}(\rho)} \mathsf{W}_{d_k}\big(Q_{A,k}(\cdot \mid z), Q_{B,k}(\cdot \mid z)\big) \le \epsilon_k$.

Assumption 3 (Approximate downstream stability)**.** For each k there are $L_k, \xi_k \ge 0$ with

$$\mathsf{W}_{d_k}(\nu Q_{B,k}, \nu' Q_{B,k}) \le L_k \mathsf{W}_{d_{k-1}}(\nu, \nu') + \xi_k \qquad \text{for all } \nu, \nu' \in \mathfrak{P}_{k-1}(\rho). \tag{11}$$

The slack ξ_k absorbs discontinuities a benchmark actually meets — routing thresholds, top-k changes, parsing failures, context truncation. System B's kernels serve only as a reference path; a symmetric result may use either system's, or a common upper bound. The constants belong to the declared interfaces, metrics and support, not to the implementation alone, and no condition $L_k < 1$ is needed for a finite-stage bound.

Theorem 2 (Compositional substitution bound)**.** *Under Assumptions 2 and 3, if* $\mathsf{W}_{d_0}(\nu_{A,0}, \nu_{B,0}) \le \epsilon_0$ *then for every* $m \in \{1, \dots, K\}$

$$\mathsf{W}_{d_m}(\nu_{A,m}, \nu_{B,m}) \le \epsilon_0 \prod_{j=1}^{m} L_j + \sum_{k=1}^{m} (\epsilon_k + \xi_k) \prod_{j=k+1}^{m} L_j, \tag{12}$$

an empty product being 1.

Proof. Write $\Delta_k = \mathsf{W}_{d_k}(\nu_{A,k}, \nu_{B,k})$. The idea is to change one thing at a time: first the stage, then the input to it. Insert $\nu_{A,k-1}Q_{B,k}$, which runs system B's stage on system A's input:

$$\Delta_k \le \mathsf{W}_{d_k}(\nu_{A,k-1}Q_{A,k}, \nu_{A,k-1}Q_{B,k}) + \mathsf{W}_{d_k}(\nu_{A,k-1}Q_{B,k}, \nu_{B,k-1}Q_{B,k}).$$

The first term compares the two stages on a common input. Coupling them optimally at each $z \in \mathcal{R}_{k-1}(\rho)$ and integrating over $\nu_{A,k-1}$ bounds it by ϵ_k. The second term runs one stage on two inputs, and equation 11 bounds it by $L_k\Delta_{k-1} + \xi_k$. Hence

$$\Delta_k \le \epsilon_k + \xi_k + L_k\Delta_{k-1}, \qquad \Delta_0 \le \epsilon_0. \tag{13}$$

Unrolling from $k = m$ down to $k = 1$, each error $\epsilon_k + \xi_k$ is multiplied by the constants of the stages after it, and ϵ_0 by all of them, which is equation 12. □

Corollary 2 (Non-expansive stages)**.** *If* $L_k \le 1$ *for every* k*, then* $\mathsf{W}_{d_K}(\nu_{A,K}, \nu_{B,K}) \le \epsilon_0 + \sum_{k=1}^{K}(\epsilon_k + \xi_k)$*: errors add, and no stage amplifies them.*

When two systems differ only on a block of stages, the same recursion gives a bound for the block propagated to the end of the pipeline.

Corollary 3 (Propagated bound for a window)**.** *Suppose* A *and* B *differ on stages* $i{:}j$*, with* $\mathsf{W}_{d_{i-1}}(\nu_{A,i-1}, \nu_{B,i-1}) \le \epsilon_{i-1}$*, and the common downstream stages after stage* j *have constant* $L_{j+1:K} = \prod_{k=j+1}^{K} L_k$*. Then the terminal discrepancy is at most*

$$\epsilon^{\text{global}}_{\text{subst}} = L_{j+1:K}\Big[\epsilon_{i-1}\prod_{k=i}^{j} L_k + \sum_{k=i}^{j}(\epsilon_k + \xi_k)\prod_{r=k+1}^{j} L_r\Big] + \sum_{k=j+1}^{K} \xi_k \prod_{r=k+1}^{K} L_r. \tag{14}$$

Any independently bounded difference between the two systems' downstream stages must also be propagated to the end and added.

Because equation 14 contains the in-window errors $\epsilon_i, \dots, \epsilon_j$, it is a substitution bound, not a fairness criterion. The two windows are *ϵ-substitutable* in the pipeline if equation 14 is at most ϵ: replacing one with the other then moves the pipeline's terminal law by at most ϵ in d_K. The same bound controls any downstream quantity that is Lipschitz in the final interface.

Theorem 3 (Utility stability)**.** *Let* $U : \mathcal{Z}_K \to \mathbb{R}$ *be* L_U*-Lipschitz with respect to* d_K*. Under the conditions of Theorem 2,*

$$\big|\mathbb{E}_{\nu_{A,K}}[U] - \mathbb{E}_{\nu_{B,K}}[U]\big| \le L_U\Big[\epsilon_0 \prod_{j=1}^{K} L_j + \sum_{k=1}^{K}(\epsilon_k + \xi_k)\prod_{j=k+1}^{K} L_j\Big]. \tag{15}$$

Proof. Kantorovich–Rubinstein duality bounds the difference of expectations by $L_U\mathsf{W}_{d_K}(\nu_{A,K}, \nu_{B,K})$, and Theorem 2 with $m = K$ bounds the distance. □

A.5 Behavioural geometry: equivalence and substitution

Fairness constrains only what surrounds a window. The two stronger claims the main text defers — that two windows behave alike, and that one can replace the other — constrain what happens inside it. Stating them needs a distance between the behaviours of systems that share neither states nor actions. This section constructs one and shows it is well defined.

The controllability operator. Consider systems $\mathcal{S}_A, \mathcal{S}_B$ with history spaces $\mathcal{H}_A, \mathcal{H}_B$, branch kernels P_A, P_B and interfaces ϕ_A, ϕ_B into shared spaces $(\mathcal{L}, d_{\mathcal{L}})$ and $(\mathcal{Z}, d_{\mathcal{Z}})$, with $d_{\mathcal{L}}, d_{\mathcal{Z}} \in [0,1]$.

Assumption 4 (Optimal-transport regularity)**.** Every ground cost below is bounded and measurable, every displayed optimal-transport infimum is attained, and each optimal value is measurable in its conditioning histories. These hold automatically for finite discretisations, and on Polish spaces under standard lower-semicontinuity and tightness conditions.

Fix a continuation weight $\gamma \in [0,1)$ and a label weight $\beta \in (0,1]$. For a bounded $d : \mathcal{H}_A \times \mathcal{H}_B \to [0,1]$, the *lifted branch cost* compares two next branches by their labels and their successors:

$$\bar{d}\big((\ell_A, h'_A), (\ell_B, h'_B)\big) = \beta\, d_{\mathcal{L}}(\ell_A, \ell_B) + (1-\beta)\, d(h'_A, h'_B). \tag{16}$$

Definition 8 (Controllability operator)**.** *The operator $\mathcal{T}_{AB}$ acts on bounded cross-system distances by*

$$(\mathcal{T}_{AB} d)(h_A, h_B) = (1-\gamma)\, d_{\mathcal{Z}}\big(\phi_A(h_A), \phi_B(h_B)\big) + \gamma\, \mathsf{W}_{\bar{d}}\big(P_A(\cdot \mid h_A), P_B(\cdot \mid h_B)\big). \tag{17}$$

The first term measures how far apart the two systems are now; the second finds the cheapest probabilistic alignment of what they do next. Two systems can therefore use different tools or decompose the work differently and still be close, provided their labels, effects and future semantic states can be coupled cheaply.

Theorem 4 (Existence and uniqueness)**.** *Under Assumption 4, $\mathcal{T}_{AB}$ maps bounded measurable functions $\mathcal{H}_A \times \mathcal{H}_B \to [0,1]$ into themselves and is a contraction in the supremum norm with modulus at most $\gamma(1-\beta) < 1$. It therefore has a unique fixed point $d^{\dagger}_{AB} = \mathcal{T}_{AB} d^{\dagger}_{AB}$.*

Proof. Self-map. Since $d_{\mathcal{L}}$, $d_{\mathcal{Z}}$ and d take values in $[0,1]$, so does the lifted cost equation 16, hence so does its transport value, and $\mathcal{T}_{AB} d$ is a convex combination of two $[0,1]$-valued terms.

Contraction. Take candidate distances d_1, d_2. Their lifted costs differ pointwise by at most $(1-\beta)\|d_1 - d_2\|_\infty$, since only the successor term changes. For fixed marginals, perturbing a transport cost by at most r everywhere moves its optimum by at most r — any coupling's cost moves by at most r, and so does the infimum. The interface term does not depend on d at all. Hence $\|\mathcal{T}_{AB} d_1 - \mathcal{T}_{AB} d_2\|_\infty \le \gamma(1-\beta)\|d_1 - d_2\|_\infty$.

Fixed point. Bounded measurable $[0,1]$-valued functions form a complete metric space under the supremum norm, and Banach's fixed-point theorem gives a unique fixed point. □

Definition 9 (Recursive controllability distance)**.** *The fixed point $d^{\dagger}_{AB}$ of Theorem 4 is the* recursive controllability distance *between histories of A and B.*

A distance between two systems' histories is more useful if it is a single pseudometric over all systems at once, so that the triangle inequality holds across a whole panel.

Proposition 2 (Pseudometric realisation)**.** *Let $\mathcal{H}_{\sqcup} = \bigsqcup_{X \in \mathcal{I}} \mathcal{H}_X$ be the disjoint union of the history spaces of a collection of systems, with $P(\cdot \mid h) = P_X(\cdot \mid h)$ and $\phi(h) = \phi_X(h)$ for $h \in \mathcal{H}_X$. If $d_{\mathcal{L}}$ and $d_{\mathcal{Z}}$ are pseudometrics, the fixed point of equation 17 on $\mathcal{H}_{\sqcup} \times \mathcal{H}_{\sqcup}$ is a pseudometric, and its restriction to $\mathcal{H}_A \times \mathcal{H}_B$ is $d^{\dagger}_{AB}$.*

Proof. Iterate from $d_0 \equiv 0$, which is a pseudometric, and show the property survives each step. Suppose d_n is a pseudometric. Then the lifted cost $\bar{d}_n$ is a pseudometric on $\mathcal{L} \times \mathcal{H}_{\sqcup}$, as a non-negative combination of two. The Wasserstein lift of a pseudometric is a pseudometric on probability measures, and its pullback through $h \mapsto P(\cdot \mid h)$ is again one. The interface term is the pullback of $d_{\mathcal{Z}}$ through ϕ. Their non-negative weighted sum $d_{n+1} = \mathcal{T} d_n$ is therefore a pseudometric. By Theorem 4 the iterates converge uniformly, and non-negativity, symmetry and the triangle inequality are all preserved under uniform limits. Uniqueness of the fixed point gives the restriction claim. □

From histories to systems. Comparing two whole systems means comparing their initial histories task by task.

Definition 10 (Whole-system controllability)**.** *For a task distribution ρ on $\mathcal{C}$,*

$$\mathfrak{C}(\mathcal{S}_A, \mathcal{S}_B; \rho) = \mathbb{E}_{c\sim\rho}\Big[\mathsf{W}_{d^\dagger_{AB}}\big(\iota_A(\cdot \mid c), \iota_B(\cdot \mid c)\big)\Big]. \tag{18}$$

Coupling within each context prevents an easy task for one system being matched with a different task for the other. The uniform version $\mathfrak{C}^\infty$ replaces the expectation over c by the ρ-essential supremum.

Distance zero means what it should, provided the underlying costs vanish only on genuinely equivalent objects.

Proposition 3 (Exact behavioural equivalence)**.** *Let $\gamma \in (0,1)$ and $\beta \in (0,1)$. Suppose $d_{\mathcal{Z}}(z,z') = 0$ only when z, z' are evaluation-equivalent, and $d_{\mathcal{L}}(\ell,\ell') = 0$ only when ℓ, ℓ' are functionally equivalent. If $d^\dagger_{AB}(h_A, h_B) = 0$, then the current interface states are evaluation-equivalent, and there is a coupling of the next branches supported only on functionally equivalent labels and on successor pairs at recursive distance zero.*

Proof. Both terms on the right of equation 17 are non-negative and carry positive weight, since $1-\gamma > 0$ and $\gamma > 0$, so both vanish. The interface term vanishing gives evaluation-equivalence. For the transport term, attainment (Assumption 4) gives an optimal coupling of cost zero; since $\beta > 0$ and $1-\beta > 0$, that coupling must be supported where both the label cost and the successor cost are zero. □

Remark 1 (Relation to bisimulation). When the systems share an MDP's states and actions, labels identify actions, and $d_{\mathcal{Z}}$ measures immediate reward difference, equation 17 reduces to a bisimulation-style recursive metric Ferns et al. (2004; 2011). Because the branch laws are those of the systems' own policies rather than a maximum over actions, the closest analogue is the on-policy (π-)bisimulation metric Castro (2020). The cross-system form replaces the maximum over shared actions with an optimal coupling over heterogeneous labelled branches.

Equivalence of partial systems. The same construction applies to a window. Let η^M_A, η^M_B be the entry-history laws induced by ρ and the upstream parts of A and B.

Definition 11 (Controllability radius and ϵ-controlled equivalence)**.** *For $\lambda_{\mathrm{in}} \in (0,1)$,*

$$\mathfrak{C}_M(A,B;\rho) = \inf_{\Lambda\in\Pi(\eta^M_A,\eta^M_B)} \int \Big[\lambda_{\mathrm{in}}\, d_{\mathcal{B}_M}\big(\psi^{\mathrm{in}}_A(h_A), \psi^{\mathrm{in}}_B(h_B)\big) + (1-\lambda_{\mathrm{in}})\, \mathsf{W}_{d_M}\big(Q^M_A(\cdot \mid h_A), Q^M_B(\cdot \mid h_B)\big)\Big] \Lambda(dh_A, dh_B). \tag{19}$$

*The windows are ϵ-*controllably equivalent *under ρ if $\mathfrak{C}_M(A,B;\rho) \le \epsilon$, and* uniformly *so if the same holds with the essential supremum over the optimal entry coupling in place of the integral.*

The infimum in equation 19 ranges over all couplings of the entry laws, so it may pair an entry reached on one task with an entry reached on a different task. When the comparison is meant to be task-matched, as in Definition 10, we use the *contextwise* version $\mathfrak{C}^\rho_M$, in which Λ is restricted to couplings that are the identity on the task context $c \sim \rho$ and arbitrary within it. Restricting the feasible set can only raise the value, so $\mathfrak{C}_M \le \mathfrak{C}^\rho_M$.

$\mathfrak{C}_M$ is the smallest residual radius the declared interface, branch cost and task distribution support; it does not claim the implementations are identical. Because it includes differences *inside* the window, it supports equivalence and substitution but is not a fairness criterion.

Remark 2 (Fairness, equivalence and substitution). Fairness requires only $\epsilon_{\mathrm{outside},n} \le \epsilon_{\mathrm{fair},n}$ (Definition 1). Approximate equivalence additionally requires a small $\mathfrak{C}_M$, and ϵ-substitutability (Section A.4) a small propagated bound equation 14. A large controlled module effect is therefore compatible with a perfectly fair comparison.

A.6 Reference specifications from human review graphs

Definition 12 (Executable reference specification)**.** *For each reference boundary state $b \in \mathcal{B}^\star$, let $\mathcal{K}^\star(b)$ be a nonempty set of acceptable continuation kernels over the next standardized artifact.*

Together with the reference graph $G^\star$, these define an executable reference specification $\mathcal{S}^\star = (\mathcal{B}^\star, \mathcal{K}^\star, G^\star)$.

The benchmark scores every system against a human review. A review is not one correct execution, and treating it as one would penalise valid alternative syntheses. This section makes the reference precise.

The human review graph. The reference artifact for scientific synthesis is neither a report string nor a trajectory, but a structured account of what the review claims and on what evidence.

Definition 13 (Human review graph)**.** *A* human review graph *is a typed attributed directed graph $G^\star = (V^\star, E^\star, \tau_V, \tau_E, a_V, w_V)$. The map τ_V assigns node types in $\{$topic, claim, evidence, paper, condition$\}$; τ_E assigns relation types such as* `supports`, `contradicts`, `qualifies`, `evidence-from`, `belongs-to-topic` *and* `valid-under-condition`*; a_V holds semantic and provenance attributes; and $w_V \geq 0$ holds importance weights.*

The graph is generally a DAG rather than a tree: one paper may support several claims, one claim may synthesise several pieces of evidence, and one condition may qualify several claims. A topic hierarchy or outline can remain a tree inside it.

Graph discrepancy. Let $G = (V, E)$ and $G' = (V', E')$ carry node masses $m \in \Delta^{|V|}$ and $m' \in \Delta^{|V'|}$, where Δ^p is the probability simplex. Let $C_{ii'}$ compare node type, semantics, evidence identity and provenance, and let R^G_{ij}, $R^{G'}_{i'j'}$ encode typed relational dissimilarity. A fused discrepancy that matches nodes and relations together is

$$d^2_{\mathrm{FGW}}(G, G') = \inf_{\Gamma \in \Pi(m, m')} \Big\{ (1 - \lambda_G) \sum_{i,i'} C_{ii'} \Gamma_{ii'} + \lambda_G \sum_{i,j,i',j'} \big| R^G_{ij} - R^{G'}_{i'j'} \big|^2 \Gamma_{ii'} \Gamma_{jj'} \Big\}, \quad (20)$$

with $\lambda_G \in [0, 1]$. This is the fused Gromov–Wasserstein objective of Vayer et al. (2019), with a linear feature term and a squared structure term. The square on the left is notational: the right-hand side is not itself the square of a metric, and d_{FGW} is a metric only under the conditions stated there, satisfying in general a relaxed triangle inequality. Paper identifiers and evidence spans can act as hard anchors, while claim and topic nodes are aligned semantically; null nodes or unbalanced transport represent omitted or novel material. The choice of metric must be validated against expert judgement. Equation equation 20 is one explicit instantiation, not a claim that it captures all graph semantics.

Set-valued reference behaviour. A human review specifies acceptable synthesis only partially, so the reference is a *set* of acceptable continuations — the executable reference specification $\mathcal{S}^\star = (\mathcal{B}^\star, \mathcal{K}^\star, G^\star)$ of Definition 12. The set $\mathcal{K}^\star(b)$ can admit alternative searches, paper sets, taxonomies and organisations while still enforcing evidence support, scope, contradiction handling and provenance.

To measure a system against a set, compare it with the closest acceptable member. Let $\nu^\star$ be a reference boundary-state distribution with $\mathrm{supp}(\nu^\star) = \mathcal{B}^\star$, a non-empty compact subset of a common boundary space $(\mathcal{B}, d_{\mathcal{B}})$, and fix a measurable nearest-reference selector

$$a^\star(b) \in \arg\min_{b^\star \in \mathcal{B}^\star} d_{\mathcal{B}}(b, b^\star), \quad (21)$$

whose existence is a modelling assumption that holds under standard compactness and measurability conditions. Kernels at b and at $a^\star(b)$ produce artifacts in a common output space $(\mathcal{Z}_{\mathrm{out}}, d_{\mathrm{out}})$. For a system with continuation kernel $Q(\cdot \mid b)$, its *reference error* at $b \in \mathcal{B}$ is

$$e(b) = \inf_{Q^\star \in \mathcal{K}^\star(a^\star(b))} \mathsf{W}_{d_{\mathrm{out}}}\big(Q(\cdot \mid b), Q^\star\big). \quad (22)$$

Likewise, $\mathfrak{C}(\mathcal{S}, \mathcal{S}^\star; \rho)$ for a set-valued reference means the infimum of equation 18 over all measurable executable selections $Q^\star(\cdot \mid b) \in \mathcal{K}^\star(b)$. Both compare a system with the nearest acceptable behaviour rather than with one arbitrarily chosen human trajectory.

Remark 3 (What the reference does not specify). The review graph does not normally identify private reasoning or one exact tool-call sequence. Claims about internal process equality are therefore valid only at declared observable interfaces. Evaluating unrestricted chain-of-thought would need further assumptions, and the theory does not require it.

B CAMPAIGN VALIDATION

Table 3 summarizes BioLitBench.

Table 3: **BioLitBench at a glance.** Statistics for the 2,042 usable graphs.

Articles	2,042	Fields / subfields / topics	6 / 48 / 230
Claims	246,102	Median claims per article	108
Typed relations	832,487	Hard negatives	103,242
Unique cited papers	224,140	Review publication years	2016–2025
Training / held-out articles	1,842 / 200	Evaluation tasks	50

SCRIBE. SCRIBE's backbone is a frozen 27B open model (Qwen3.8-27B, bf16, greedy decoding at seed 0, 65,536-token context for generation), and its trained object is a KV-Skill carrier of 14–16M parameters per trained stage, served through a patched inference server. It exposes three windows — synthesis, planning and writing — each emitting its exit artifact. Its entry is sealed before the first model call, and a check at the start of synthesis fails the run if any retrieval capability is still reachable, so the absence of search downstream holds by construction. Each window appends a receipt whose exit hash is the next window's entry hash, so a claim about a run can be checked against the run itself. The trained SCRIBE runs the current pipeline. With the same pipeline and its initial carriers, the composite is 0.595 under fixed input and 0.494 under same pool, so the planning carrier's paired gain is $+0.011$ and -0.010, neither certified; SCRIBE (untrained) in the tables is the earlier pipeline.

SCRIBE's retrieval harness. Under same pool SCRIBE retrieves in two stages. The model first decomposes the question into subtopics, proposes queries under a duplicate guard and ranks the candidates. A selection rule then fills the bundle from the top 1,000 ranked papers in two-year windows anchored at the cutoff, ordering each window by in-pool citation count with citers strictly before the cutoff and the evaluated review removed. That rule reads the corpus citation graph directly, so it is part of the system rather than a tool call, and about 40 % of the final bundle is chosen by it rather than by the model's own ranking. The per-task search, row and document budgets are each the largest that any baseline spent on that task. The harness was selected on screening and held-out tasks disjoint from the 50 evaluation tasks. The earlier harness, without the selection rule, is the SCRIBE row of Table 8; the leaderboard (Section 6.1) uses the final one.

Entry disclosures. Three facts bear on the same-pool table (Table 4), and the fixed-input table is untouched by all of them. (i) The pool service the published pipelines ran exposes a citation route on which the evaluated review's own reference list can be requested; SurveyGen-I requested it on 33 of 50 tasks, and its paired margins are the same on the 17 tasks where it did not (e.g. $+0.080$ and $+0.081$ over LLM×MR), so the route does not produce its position. (ii) SCRIBE is restricted by a route guard to plain search and title, abstract and year, and the commercial agents' tools exclude the evaluated review. (iii) A SCRIBE acquisition that stops at its document budget is scored as a normal exit when its bundle is sealed, its hashes match, its audits are clean and all three downstream windows succeed; the published pipelines' own budget stops are likewise recorded as normal.

Report length. Report length is a genuine confounder: the primary view holds every report to the task's target length, and on entry-conditioned synthesis the full-length view reverses the conclusion, certifying a pipeline whose reports have a median of about 250,000 words over SCRIBE's reports of under 5,000. All analysis constants are recorded with input hashes, so a reader can rerun the analysis and see what changing them does.

Assertions. Twenty-nine assertions run with every rebuild and all pass; four carry most of the weight. The table reproduces the previous analysis cell by cell (288 cells, $\max|\Delta| = 1.11 \times 10^{-16}$). A replicate null — three seeds of one model against each other, where nothing should separate — certifies 0 of 114 cells, the control showing the threshold has not been loosened. Entry byte-equality is re-derived rather than assumed and holds on 50 of 50 tasks, without which the fixed-input certifications would not stand. And rescoring the reference-fed reports on an independent implementation differs on 4 of 81,575 cells, all on readouts the admission rule already excludes.

C FURTHER LEADERBOARD RESULTS

This appendix gives the same-pool capability scores behind Figure 1(b) (Table 4) and the leaderboard at full length.

Table 4: **Capability scores under same pool** at the primary observation point. Conventions as in Table 1.

				Capability (fraction of human level)					
Backbone	System	Composite	Rank	Retrieval	Synthesis	Reasoning	Planning	Writing	Form
Same pool: every system searches one frozen corpus									
Qwen3.8-27B	SCRIBE	**0.487**	1–7	0.335	0.523	0.545	0.250	0.446	0.826
	SCRIBE (untrained)	0.464	1–9	0.335	0.512	0.538	0.232	0.404	0.766
gpt-5.6-luna	SCRIBE-Luna	0.479	1–6	0.335	0.577	**0.550**	0.185	0.386	0.842
	SurveyGen-I	0.460	1–9	**0.336**	0.535	0.485	0.181	0.371	**0.850**
	OpenAI tool loop	0.437	3–10	0.172	**0.665**	0.544	0.126	0.425	0.689
	SurveyForge	0.431	3–10	0.220	0.529	0.439	0.255	0.382	0.760
	SurveyG	0.425	6–10	0.155	0.487	0.486	**0.284**	0.451	0.686
	AutoSurvey	0.419	5–11	0.238	0.540	0.444	0.117	0.411	0.766
	LLM×MR	0.379	11–12	0.077	0.504	0.445	0.175	0.403	0.671
claude-sonnet-5	Claude Code	0.458	1–8	0.227	0.573	0.520	0.188	**0.467**	0.775
	Claude Science	0.456	1–8	0.287	0.598	0.546	0.164	0.375	0.767

The leaderboard at full length. Table 5 reads the leaderboard's systems at full report length, the secondary and descriptive view; conventions as in Table 1. The top of the fixed-input table is unchanged, while AutoSurvey — a median of about 250,000 words per task — rises to third, because the full-length composite rewards material that longer reports contain more of. Under same pool the certified intervals widen and nothing is pinned.

Table 5: **The leaderboard at full length** (secondary, descriptive). Conventions as in Tables 1 and 4; the footnotes of Table 1 apply to panel (a).

Backbone	System	Composite	Rank	Capability (fraction of human level) Synthesis	Reasoning	Planning	Writing	Form
(a) Fixed input: every system receives the review's reference list								
Qwen3.8-27B	SCRIBE	**0.609**	1–4	0.695	0.596	**0.319**	**0.619**	**0.816**
	SCRIBE (untrained)	0.585	1–6	0.690	0.603	0.276	0.569	0.787
gpt-5.6-luna	SCRIBE-Luna	0.573	2–8	0.710	0.622	0.177	0.592	0.763
	AutoSurvey	0.567	1–8	**0.958**	**0.769**	0.135	0.370	0.605
	SurveyG	0.530	6–12	0.705	0.640	0.272	0.353	0.678
	SurveyGen-I	0.520	6–13	0.695	0.580	0.182	0.415	0.729
	OpenAI tool loop	0.515	7–13	0.741	0.552	0.136	0.487	0.660
	LiRA	0.499	5–13	0.755	0.634	0.254	0.281	0.594
	LLM×MR	0.491	10–13	0.749	0.663	0.182	0.280	0.578
claude-sonnet-5	Claude Science	0.546	3–12	0.702	0.604	0.160	0.526	0.736
	Claude Code	0.535	5–12	0.650	0.547	0.185	0.588	0.703
proprietary	Elicit SR[§]	0.555	2–11	0.709	0.542	0.185	0.610	0.731

Backbone	System	Composite	Rank	Capability (fraction of human level) Retrieval	Synthesis	Reasoning	Planning	Writing	Form
(b) Same pool: every system searches one frozen corpus									
Qwen3.8-27B	SCRIBE	**0.506**	1–8	0.335	0.570	0.574	0.250	**0.468**	**0.841**
	SCRIBE (untrained)	0.477	1–10	0.335	0.579	0.578	0.232	0.353	0.787
gpt-5.6-luna	AutoSurvey	0.490	1–8	0.238	**0.906**	**0.746**	0.117	0.257	0.677
	SCRIBE-Luna	0.489	1–8	0.335	0.599	0.561	0.185	0.423	0.832
	SurveyForge	0.476	1–8	0.220	0.744	0.618	0.255	0.274	0.746
	SurveyGen-I	0.473	1–9	**0.336**	0.655	0.547	0.181	0.283	0.835
	OpenAI tool loop	0.438	6–11	0.172	0.668	0.545	0.126	0.427	0.689
	SurveyG	0.434	5–11	0.155	0.621	0.597	**0.284**	0.278	0.670
	LLM×MR	0.404	9–12	0.077	0.708	0.635	0.175	0.249	0.579
claude-sonnet-5	Claude Science	0.455	1–12	0.281	0.601	0.550	0.164	0.416	0.770
	Claude Code	0.449	1–12	0.222	0.573	0.520	0.188	0.467	0.775

Commercial systems. Claude Science is Anthropic's research workbench and Claude Code is Anthropic's coding agent run as a tool loop at low effort; both use claude-sonnet-5 and reach the corpus only through the tools we provide. The OpenAI tool loop is gpt-5.6-luna called through the OpenAI Responses API with the same tools. Under fixed input these tools return only the task's reference list. Elicit Systematic Review uses its own model, receives the same reference list, reads full text where the other systems read abstracts, and cites at most 80 papers; all of its citation marks resolve. Those are properties of the system and sit inside the system window, so they are declared rather than refused.

D READOUT VALIDITY UNDER EVIDENCE ABLATION

The admission screen asks whether a readout carries topical signal: scoring a system against an unrelated review must do worse than against its own. It does not ask whether a readout responds to the *evidence* a system was given. We test that on the controlled writing window, where six published pipelines were each handed the same sealed entry (papers, outline and leaf allocation) on 20 tasks,

one per subfield cluster. The same driver, AutoSurvey's writer, was then run with no papers, with another section's papers, and with half of each section's papers.

Given the same driver, a writer with no papers scores *higher* than one given them (Table 6) on form (-0.114 for with-evidence minus no-evidence, 95% CI $[-0.170, -0.059]$), reasoning (-0.061, $[-0.110, -0.013]$) and synthesis (-0.044, $[-0.087, -0.002]$), and lower on writing ($+0.071$, $[+0.040, +0.102]$). Averaged over the six pipelines, the only readout that separates full evidence from both the empty and the shuffled controls is evidence agreement ($+0.145$ and $+0.078$). The content readouts therefore measure agreement with the human review's claims, relations and organisation, which a strong backbone partly supplies from its own knowledge. That is the reading the leaderboard adopts: its relations certify content agreement with the human review, the ground truth of BioLitBench, and not the use of the evidence provided. An evidence-ablation control belongs in any admission screen that intends the stronger reading.

Table 6: **Evidence ablation at the controlled writing window** (20 tasks, matched length, fraction-of-human-level scale). Each entry is the paired difference between the writer given the full evidence and a control; the 95% interval is below it. *Same driver*: AutoSurvey's own writer against its controls. *Pipeline mean*: the six pipelines' mean against the same controls. A negative value means the control scores higher.

	same driver			pipeline mean	
Axis	no evidence	shuffled	half	no evidence	shuffled
synthesis	-0.044 $[-.087, -.002]$	-0.036 $[-.075, +.003]$	-0.019 $[-.056, +.018]$	-0.053 $[-.088, -.018]$	-0.045 $[-.082, -.008]$
reasoning	-0.061 $[-.110, -.013]$	-0.035 $[-.100, +.029]$	-0.027 $[-.080, +.027]$	-0.081 $[-.119, -.042]$	-0.055 $[-.109, -.000]$
writing	$+0.071$ $[+.040, +.102]$	$+0.011$ $[-.027, +.048]$	-0.001 $[-.023, +.021]$	$+0.094$ $[+.067, +.122]$	$+0.034$ $[+.005, +.063]$
form	-0.114 $[-.170, -.059]$	$+0.022$ $[-.024, +.068]$	$+0.019$ $[+.001, +.037]$	-0.115 $[-.170, -.059]$	$+0.022$ $[-.019, +.062]$

Scores do not depend significantly on the target review's publication year (pooled slope -0.007 per year, $p = 0.15$–0.17), and the content axes do not rise with a review's age more than planning, writing or retrieval ($p \geq 0.34$). Publication year is only a proxy for a review's exposure during pretraining.

E THE CAPABILITY CAMPAIGN AND THE PUBLISHED PIPELINES

The leaderboard ranks systems by one composite. This appendix opens the same kind of comparison capability by capability, on a smaller roster decided as its own family: the six fixed-input systems of Table 7, and seven same-pool systems in which SCRIBE runs its earlier retrieval harness (Table 8); our system in this appendix is SCRIBE (untrained). Its rank intervals are therefore not those of Tables 1 and 4.

Given a shared entry, the pipelines separate — on some capabilities and not others. Under fixed input the ten admissible pairs across five axes give 50 cells, of which 18 certify: five on synthesis, five on planning, three on writing, three on form and two on reasoning. There is no retrieval axis here, because every arm was handed its evidence. The confirmatory family decides 5 of the 10 pairs. Four involve SCRIBE, which is certified above each of its four comparable opponents ($\widehat{\Delta}$ between 0.076 and 0.110 against q between 0.037 and 0.050, $n = 50$); the fifth is SurveyG above LLM×MR ($+0.035$ against $q = 0.028$). Under same pool the seven-system roster gives 21 pairs over six axes, 126 cells, of which 32 certify — eleven on retrieval, eight on planning, eight on form and five on synthesis. Reasoning and writing certify *nothing at all*: once reports are held to a common length, no two of these systems can be told apart on either, at this sample size.

Table 7: **Capability scores as a fraction of human level** under *fixed input*, where every arm receives the task's reference bundle, with certified rank intervals in brackets — the positions consistent with every certified relation. LiRA receives the review's bibliography in its own input format rather than through the shared allowlist, so this campaign does not pair it with the other arms, and its whole row is unpinned. The same-pool scores are in Table 8. Scores come from this campaign's own per-axis scorer; SCRIBE (untrained)'s form score therefore reads 0.783 here and 0.786 on the merged board of Table 1.

fixed input (reference bundle; no retrieval axis)

System	synthesis	reasoning	planning	writing	form
SCRIBE (untrained)	**0.686** [1,3]	**0.601** [1,4]	**0.276** [1,4]	**0.565** [1,4]	**0.783** [1,3]
AutoSurvey	0.638 [1,5]	0.533 [1,6]	0.135 [3,6]	0.483 [1,6]	0.687 [2,6]
SurveyGen-I	0.591 [2,5]	0.510 [2,6]	0.182 [3,6]	0.449 [2,6]	0.746 [1,6]
SurveyG	0.568 [2,6]	0.541 [1,6]	0.272 [1,3]	0.509 [1,5]	0.689 [2,6]
LiRA	0.539 [1,6]	0.495 [1,6]	0.254 [1,6]	0.442 [1,6]	0.716 [1,6]
LLM×MR	0.537 [4,6]	0.515 [2,6]	0.182 [2,6]	0.446 [3,6]	0.666 [2,6]

Table 8: **Capability scores under *same pool***, where every system retrieves for itself from one frozen corpus and the retrieval axis becomes observable. Conventions as in Table 7. Writing, which certifies no pair, is not shown. The SCRIBE row uses its earlier retrieval harness; the final harness, used in the leaderboard, reaches retrieval 0.335 (Table 4).

System	synthesis	reasoning	planning	retrieval	form
SurveyGen-I	0.535 [1,6]	0.485 [1,7]	0.181 [3,6]	**0.336** [1,1]	**0.850** [1,4]
AutoSurvey	0.540 [1,6]	0.444 [1,7]	0.117 [6,7]	0.238 [2,3]	0.766 [1,5]
SurveyForge	0.529 [1,6]	0.439 [1,7]	0.255 [1,5]	0.220 [2,4]	0.760 [1,5]
SCRIBE (untrained, earlier harness)	0.496 [1,6]	0.480 [1,7]	0.257 [1,6]	0.156 [3,6]	0.786 [1,6]
SurveyG	0.487 [1,6]	0.486 [1,7]	**0.284** [1,4]	0.155 [4,5]	0.686 [4,7]
DR-Tulu	0.369 [6,7]	0.419 [1,7]	0.251 [1,6]	0.084 [5,7]	0.663 [3,7]
LLM×MR	0.504 [1,7]	0.445 [1,7]	0.175 [2,7]	0.077 [6,7]	0.671 [4,7]

The two entry conditions diagnose a system rather than rank it. Reading Tables 7 and 8 together is what the framework is for. Given the same papers, SCRIBE is first or joint-first on every axis it is comparable on and no baseline is certified above it anywhere; retrieving for itself with its earlier harness, its one certified deficit was retrieval. The mechanism is in the artifacts: that harness recovered 50 papers per task against 109 under fixed input, its claim count fell from 118 to 52 and its reports from 3,947 to 2,644 words. Everything downstream was working on a smaller bundle. A single leaderboard number would have called the system mid-field and said nothing about where to look; the per-window diagnosis located the deficit, and the final retrieval harness closed it (retrieval 0.335 against SurveyGen-I's 0.336, Table 4).

Refinement trades content for form. Scoring the draft and the final exit separately makes the refinement pass its own comparison, with entries identical by construction. Five refinement effects certify and three are *negative*, all on synthesis — pipelines deleting human claims their own drafts had covered (-0.072, -0.113, -0.076 against q of 0.035, 0.100, 0.051) — against two positive effects, on writing ($+0.068$) and form ($+0.115$). The same pass improves the prose and loses the content, exactly the conflation that reading only a final report cannot detect. Across the panel, on the four axes with a draft exit (all but planning), 10 of the 13 pair-by-axis relations certified at the final exit under fixed input are already present at the draft exit, and 21 of 24 under same pool: most of what separates these systems is settled before refinement begins.

A capability visible only at its own exit. Allocation — which papers are handed to the writer for each section — exists only at its own exit. A human places a paper in 0.14 of the reference's top-level sections; pipelines that broadcast their whole evidence set to nearly every section sit between 0.94 and 1.00. Broadcasting makes raw co-location recall 1.000 by construction, carrying the broadcaster on one task to an F_1 of 0.800 against 0.286 for a system genuinely allocating. Chance correction removes that expectation and returns both to near zero, which is why the adjusted Rand index is

the only certifying readout here. Measured that way every system is far below human, SCRIBE at 0.152–0.174 against 0.099 for the best baseline.

The published pipelines alone. Table 9 gives the certified tiers of the published pipelines with the backbone held fixed, as discussed in Section 6.2. Under a sign-flip null the per-pair false-certification rate is 0.0035–0.0047 against a nominal 0.005, and leaving out one subfield gives a mean Jaccard overlap of 0.994 with no winner flips.

Table 9: **The published pipelines alone** (scaffold family, 10 pairs per tier; compare Tables 1 and 4, whose larger families give wider intervals). *Display rank* orders point estimates and is not a claim; *certified rank* is the interval of positions consistent with the certified relations over every linear extension of their transitive closure, so a single value is a position the data pin down. In the reference-fed tier, each pipeline runs its reference-fed entry. Composite is the equal-weight mean over the tier's capability axes on the fraction-of-human-level scale, at matched length. The last column gives the certified rank at full length.

			Certified rank	
Tier	System	Composite	matched	full length
self-retrieving	SurveyGen-I	0.460	**1**	1–3
(certified: 7 / 7 of 10)	SurveyForge	0.431	2–4	1–3
	SurveyG	0.425	2–4	**4**
	AutoSurvey	0.419	2–4	1–3
	LLM×MR	0.379	**5**	**5**
reference-fed	SurveyG	0.516	1–4	2–4
(certified: 1 / 6 of 10)	SurveyGen-I	0.496	1–5	2–4
	AutoSurvey	0.495	1–5	**1**
	LiRA	0.484	1–5	2–5
	LLM×MR	0.469	2–5	4–5

F THEORY CHECKS IN DETAIL

Table 10 summarizes the tests of the theory. The subsections below give each test in detail.

Table 10: **Verification summary.** Each part of the theory, what it is tested on, and the key result. The subsection with the full test is given in parentheses.

Component	Tested on	Key result	Verdict
On the published systems			
Attribution bound, Thm. 1 (F.1)	two entries per pipeline, 118 runs	L_M transfers to held-out pipelines (coverage 0.79–1.00)	✓ holds
Bound at a stage window (F.2)	one sealed writing input, 20 tasks; fit on 5 pipelines, check on 15 pairs of 6	$420/420$ cells hold at matched length; 44% vacuous	✓ holds; informative on writing
Admissibility, Def. 1 (F.3)	five LiRA pairs, measured entry mismatch	radius 0.70–0.73 against tolerance 0.21; refused on $15/15$ readouts	✓ refuses
Non-expansive windows, Cor. 1 (F.1, F.2)	both designs above	0 runs expand at matched length; $62/100$ at full length	✓ at matched length
Certification rule, Eq. 4 (F.5)	sign-flip null; leave one subfield out	false certification 0.035 / 0.062; no flips	✓ near nominal
Readout validity (D)	evidence ablation, writing stage	no-evidence writer scores higher on synthesis, reasoning, form	$\sim$ content agreement
Comparability vs. equivalence, Def. 11 (F.6)	$\mathfrak{C}_M$ on reference-fed pairs	a fair pair: 0.07 at retrieval, 0.51 at writing	✓ distinct
Synthetic chains or implementation			
Composition, Thm. 2 (F.4)	synthetic four-stage chains	holds on 95–100% of chains	✓ synthetic only
Fixed point, Thm. 4 (F.7)	empirical kernels	contraction $0.449 \leq 0.45$; unique	○ implementation

✓ holds as stated; $\sim$ holds with the scope given; ○ implementation check only. False certification: same pool / fixed input.

F.1 The attribution bound on two entries

Each of four published pipelines ran twice on every task it completed under both entries: searching the frozen corpus, and with the corpus restricted to the review's reference bundle. For each task, x is the entry mismatch of the self-retrieving run to the reference bundle, and y is the distance between the two runs' writing exits. This gives 118 paired runs with $x \in [0.74, 1.00]$. The selected estimator fits the module-sensitivity constant as the 95th percentile of y/x, giving $L_M = 0.890$. Table 12 reports the tests that could fail. The leave-one-pipeline-out test checks transfer, because the panel applies one constant to every pipeline. Non-expansiveness is falsifiable: the same estimator returns 1.74 on the writing stage of the *development instrument*, an in-house four-stage agent used only as a measuring instrument, whose stages can be entered from chosen inputs and so probed at smaller x. None of the 118 runs expands. Table 11 gives the estimator designs that were compared; the fitted constant is a 95th-percentile ratio, so the radius it gives is a high-probability envelope rather than a worst-case bound.

Table 11: **The module-sensitivity constant L_M, measured on the panel.** Fitted on the paired gap over 118 paired runs from four pipelines, $x \in [0.74, 1.00]$.

L_M design	L_M	ξ	note
ratio q_{95}	**0.890**	**0.000**	**selected design**
ratio sup	0.967	0.000	supremum form
affine	0.803	0.085	envelope, slope + slack
per system (ratio q_{95}): AS 0.911 (n=19), SGI 0.910 (19), LMR 0.880 (31), SG 0.863 (49)			

Table 12: **Falsification tests for Assumption 1.** (a) A constant fitted on the other three pipelines, evaluated on the held-out one — the transferability the panel-level radius actually assumes. (b) The non-expansive claim is falsifiable and is not falsified; the instrument's writing stage is a positive control.

(a) leave-one-system-out held out	n test	L_M (train)	L_M (own)	held-out coverage
AutoSurvey	19	0.881	0.911	0.789
SurveyGen-I	19	0.882	0.910	0.842
LLM×MR	31	0.895	0.880	0.968
SurveyG	49	0.908	0.863	1.000
(b) falsifiability	locally expansive runs ($y > x$): $0/118$	max ratio 0.967	instrument writing stage: 1.74	

F.2 The bound at a controlled writing window

Six published pipelines were each handed the same sealed writing-window entry on 20 tasks, one per subfield cluster: the papers, the outline, and the assignment of papers to sections, built from the review's own sections and citations. The drivers call each pipeline's own stage code. This exit is the controlled contrast $\Delta^\star$; each pipeline's own writing exit on the same task gives Δ^{obs}. Table 13 reports the fit and the score-level check of Theorem 1 over the 30 admitted readouts scored at the writing window and 15 pairs. The 5 pairs with LiRA, which does not retrieve, carry no retrieval readouts, which leaves 420 cells. The shared entry's paper set is essentially the reference bundle, so the self-retrieving pipelines still sit at $x \in [0.74, 1.00]$; LiRA, which is reference-fed natively, sits at $x = 0$ with a nonzero exit displacement. A pure ratio cannot describe that case, so the fit uses the five self-retrieving pipelines; the slack ξ of the theorem is the term that covers it. The bound is loose: the median bound is 0.955 against a median gap of 0.060. It is informative on the writing axis, where no cell is vacuous, and vacuous on every retrieval cell, because retrieval readouts read the given entry rather than the writing exit. Non-expansiveness holds at matched length but fails at full length, because report length dominates the exit distance, so the corollary's hypothesis depends on the declared length convention.

Table 13: **The attribution bound at a controlled writing window** (5 self-retrieving pipelines × 20 tasks for the fit; 420 pair-by-readout cells for the score-level check).

	matched length	full length
fitted L_M (95th pct. / sup)	0.906 / 0.940	1.042 / 1.188
runs with $y > x$	$0/100$	$62/100$
leave-one-pipeline-out coverage (mean / min)	0.93 / 0.80	0.94 / 0.90
bound holds (all cells)	$420/420$	$419/420$
vacuous cells ($\epsilon_{\text{outside}} \geq 1$)	43.6%	46.0%
median $\epsilon_{\text{outside}}$ / median $\lvert\Delta^{\text{obs}} - \Delta^\star\rvert$	0.955 / 0.060	0.978 / 0.066

F.3 Refusals computed from the bound

Table 14 applies Eq. 3 to the five pairs of LiRA with a self-retrieving pipeline. The common entry μ is the reference bundle, which is LiRA's own entry, so $\epsilon_{\text{in,LiRA}} = 0.002$ and $\epsilon_{\text{in},P}$ is the mean entry mismatch of pipeline P's self-retrieved papers to the bundle. With $L_C = 1$ and no channel, protocol or slack term (the selected ratio fit has $\xi = 0$), Eq. 3 reduces to $\epsilon_{\text{outside},n} = L_{s_n} L_M(\epsilon_{\text{in},P} + \epsilon_{\text{in,LiRA}})$, with $L_M = 0.890$ and L_{s_n} the score constant of readout n at the writing window (median 0.833). For AutoSurvey, $0.833 \times 0.890 \times (0.975 + 0.002) = 0.725$. The radius is compared with the pre-registered tolerance of each readout, the median gap between human peer reviews. Of the 47 readouts in the pre-registered tolerance table, 15 have an admissible tolerance ($0 < \epsilon_{\text{fair}} < 0.5$), with median 0.208. Every pair is refused on all 15; the smallest margin is $+0.077$, and a readout would be admitted only if $L_M \leq 0.554$, below every observed ratio. The verdicts are unchanged under the leave-one-pipeline-out constants (0.881–0.908).

Table 14: **Refusals computed from Eq. 3** for LiRA against each self-retrieving pipeline.

Pair	$\epsilon_{\text{in},P}$	median $\epsilon_{\text{outside}}$	median ϵ_{fair}	readouts refused
AutoSurvey – LiRA	0.975	0.725	0.208	15/15
SurveyForge – LiRA	0.968	0.720	0.208	15/15
SurveyG – LiRA	0.972	0.723	0.208	15/15
SurveyGen-I – LiRA	0.942	0.700	0.208	15/15
LLM×MR – LiRA	0.979	0.728	0.208	15/15

F.4 Composition across stages

Published pipelines cannot be re-entered at an intermediate artifact, so the multi-stage bound is checked on synthetic four-stage chains built from the human graphs, with known knobs for retrieval recall, claim retention, and section order (Table 15). The bound holds everywhere and is informative through the organisation stage. At the writing stage it exceeds one on $[0, 1]$-valued distances.

Table 15: **Theorem 2 on synthetic four-stage chains built from the human graphs.** Observed chain discrepancy against the composed bound (affine at q_{95}, and ratio form); fraction of chains on which the bound holds; tightness = observed/bound.

Stage m	observed W	bound (affine)	bound (ratio)	holds	non-vacuous	tightness
1 retrieval	0.605	0.744	0.744	95 %	yes	0.79
2 synthesis	0.347	0.844	0.891	100 %	yes	0.38
3 organization	0.261	1.312	0.981	100 %	yes	0.26
4 writing	0.685	1.465	3.378	100 %	**no**	0.21

F.5 Calibration of the certification rule

We apply the leaderboard's decision rule to a null in which no system differs, on the ten-system version of each table that preceded the addition of the trained SCRIBE and Elicit. Each pair's paired differences are centred, and their signs are flipped per subfield cluster, with one sign vector per draw shared by every pair so that the family is defined. Table 16 reports the family-wise false-certification rate over 2,000 draws, the radius scale $c^\star$ at which that rate equals the nominal 0.05, and the certified relations that $c^\star$ would remove. It also reports leave-one-subfield-out stability over the 23 clusters. The rebuilt tables reproduce the published leaderboard exactly before any perturbation.

Table 16: **Calibration and stability of the leaderboard's decision rule.** Family-wise rates are over 2,000 sign-flip draws (nominal 0.05). $c^\star$: radius scale giving the nominal rate. Jaccard: overlap of the certified set when one of the 23 subfields is left out (mean / minimum).

Table	View	pairs	certified	family-wise	$c^\star$	Jaccard	flips
fixed input	matched length	45	21	0.062	1.027 (drops 1)	0.952 / 0.870	0
same pool	matched length	45	20	0.035	0.966 (drops 0)	0.918 / 0.727	0
fixed input	full length	45	20	0.051	1.003 (drops 0)	0.965 / 0.905	0
same pool	full length	45	16	0.041	0.980 (drops 0)	0.877 / 0.750	0

In that version, the one relation that calibration would remove is Claude Code over AutoSurvey in the fixed-input table of the leaderboard, whose gap exceeds its radius by a factor of only 1.02. No relation involving our system is affected.

F.6 Comparability versus equivalence

Fair comparability constrains only what surrounds a window; equivalence constrains what happens inside it. $\mathfrak{C}_M$ (Definition 11) measures the second. For task-matched pairs we use the contextwise radius, whose coupling is the task diagonal, so $\mathfrak{C}_M(A, B) = \text{mean}_t[\lambda_{\text{in}}\, d(\text{entry}_A, \text{entry}_B) + (1 -$

$\lambda_{\text{in}})\, d(\text{exit}_A, \text{exit}_B)]$, at $\lambda_{\text{in}} = 0.5$. Where a direct entry distance is not stored, the entry term is bracketed by $|e_A - e_B|$ and $e_A + e_B$, using each system's distance to the reference, and both ends are reported.

Table 17 shows the case that Remark 2 separates. The reference-fed AutoSurvey and LiRA are a fair pair. At the retrieval window they share the entry exactly and $\mathfrak{C}_M = 0.070$, so their retrieval windows are nearly equivalent. At the writing window the same pair has $\mathfrak{C}_M \approx 0.51$: comparable, but not equivalent, since substituting one writer for the other moves the exit by half the scale of the readout.

Table 17: Ctrl_M **for the reference-fed pairs** at $\lambda_{\text{in}} = 0.5$ (Eq. 19); mean over the tasks both ran, with the essential-supremum form. Pairs with the reference-fed SurveyGen-I are omitted, because they were not recomputed after its re-run. Entry lo/hi bracket the entry term where only each system's distance to the reference is stored. At retrieval the entry is the shared task specification, so the entry term is 0 and Ctrl_M measures behaviour inside the window alone.

		retrieval		writing	
Pair	n	Ctrl_M	sup	Ctrl_M (lo–hi)	sup
AutoSurvey – LiRA	50	0.070	0.161	0.504–0.506	0.606
LiRA – SurveyG	49	0.092	0.293	0.509–0.511	0.749
AutoSurvey – SurveyG	49	0.127	0.322	0.392–0.505	0.901
AutoSurvey – LLM×MR	50	0.424	0.472	0.800–0.940	1.116
LLM×MR – SurveyG	49	0.429	0.485	0.757–0.941	1.232
LiRA – LLM×MR	50	0.430	0.478	0.852–0.854	0.971

F.7 The controllability distance

The operator of Theorem 4 is iterated with exact optimal transport on empirical branch kernels of the logged systems (Table 18). The check confirms the implementation: the contraction modulus, the unique fixed point, and the pseudometric properties.

Table 18: Theorem 4 and its corollaries on the empirical kernels of the twelve logged configurations.

Property	Requirement	Measured	
contraction	ratio $\leq \gamma(1-\beta) = 0.45$	0.449, 0.425, 0.428 (random inputs); 0.450 at convergence	✓
existence / uniqueness	same fixed point from $d_0 \equiv 0$ and $d_0 \equiv 1$	18 / 18 iterations, sup gap 5.7×10^{-7}	✓
pseudometric	symmetry, triangle inequality	symmetry gap 0; 0 / 20,000 triangle violations	✓
zero self-distance	$d^\dagger(S, S) = 0$	0 for all 12 systems	✓

G Scope

A certificate compares systems up to an exit. A certified relation between system windows compares complete systems as built, backbone included (Section 3). A scaffold comparison asks which scaffold (the prompts, tools and control flow around the model) is better for a fixed model. It holds the backbone fixed, so what it certifies is an ordering of scaffolds around one generator, and a pair on different backbones is comparable only if that difference is accounted for. Comparability is indexed by the declared interface, observation channel and readout family.

The tasks come from one distribution. The 50 tasks are biomedical review questions drawn from one corpus. The subfield clustering makes the radius account for dependence within that distribution; transfer beyond it is a separate question.

Stage windows of published pipelines cannot be re-entered. Theorem 2 needs each stage to be run from a controlled input. The published pipelines log every model call, so their intermediate artifacts can be recovered and scored, but they cannot be re-entered at an intermediate artifact of our choosing; only their system windows can be perturbed (Corollary 1). Pipelines that accept an intermediate artifact as input, and record its hash, would make every stage testable.