When Do Independent Scientific Alternatives Emerge? Evaluating LLM Predictions of Waiting Times

Anonymous authors
Paper under double-blind review

Abstract

The wait from a scientific contribution to its first independent, functionally equivalent alternative describes when another route to the same capability becomes available. We examine whether evaluation scores reflect how well large language models rank these reinvention times. In a post-hoc audit of an author-built dataset and saved predictions from three models, we trace waiting-time labels to their target works and decompose the ranking score by comparison type. At least 8 of 34 eligible observed held-out examples have labels inconsistent with the target work's interval under the dataset's own rules. Adding curator-specified lower bounds (claims of no alternative before a cutoff) raises the C-index ranking score from 0.42–0.45 to 0.61–0.67 on the models' respective valid subsets, without changing any observed-interval prediction ordering. The increase comes from comparisons between observed intervals and lower bounds. A diagnostic using only record type scores higher still, without ranking waiting times within either type. The score increase persists after recomputing the known mismatches from retained records, including on samples valid for all three models. Neither excluding the known mismatches nor recomputing their labels establishes above-chance observed-interval ranking. Two models exceed the specified output-failure threshold, so their affected results remain exploratory. This case shows why assessing waiting-time ranking requires checking both what each label measures and which comparisons produce the score.

1 Introduction

Why study the arrival of alternatives? Studies of scientific progress examine discovery and scholarly influence (Fortunato et al., 2018). Citation indicators track scholarly uptake (Bornmann & Daniel, 2008; Waltman, 2016), while studies of multiple discovery examine distinct researchers reaching similar advances (Merton, 1961; Bikard, 2020). The timing of independent alternatives adds a complementary question: how quickly does another route to the same capability emerge after a scientific work appears? Comparing these intervals would describe the pace of alternative discovery, beyond scholarly uptake.

From scientific prediction to waiting times. Scientific prediction benchmarks evaluate experimental outcomes (Luo et al., 2025), future paper records and research trends (Ajith et al., 2026), and scientific events' feasibility and timing (Wu et al., 2026). We consider estimating the delay from a target scientific work to its first independent, functionally equivalent alternative. We call this interval *reinvention time*, $G(a)$. Given information about the target work, the model estimates this waiting time, reported in days (Figure 1a). We evaluate historical intervals.

Two requirements for interpreting a score. Evaluation depends on both the reference interval and the comparisons used to score the predictions. A label taken from the shortest pairwise gap in a group of linked papers need not start at the work shown to the model (Figure 1b), so it can answer a different temporal question. A second issue arises when observed intervals are pooled with records claiming no alternative before a cutoff. The score then combines comparisons between observed

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.

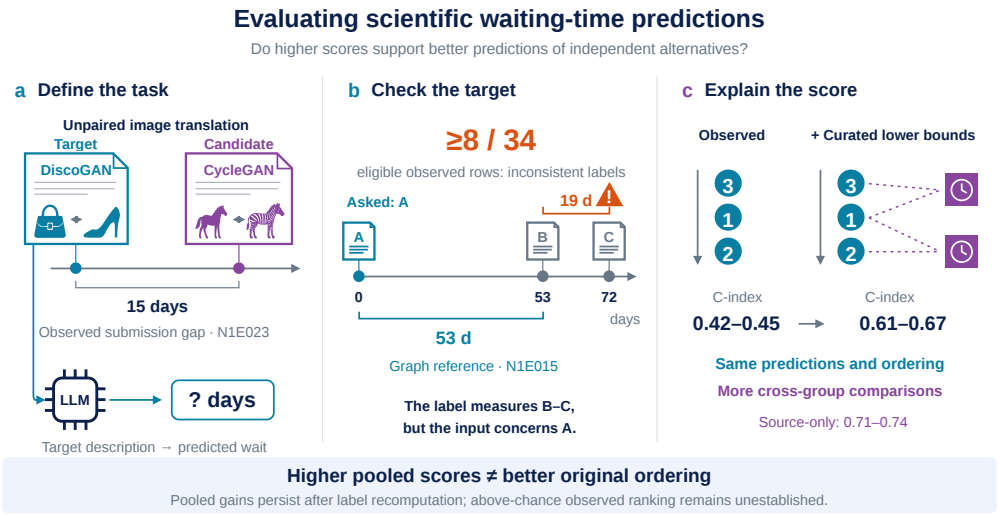


Figure 1: (a) Waiting-time prediction, illustrated by development-set candidate pair N1E023 (DiscoGAN–CycleGAN; 15-day submission gap). (b) N1E015 asks about A but labels B–C (19 days); A–B (53 days) is an internal graph reference, not validated substitute timing. (c) Adding curated lower bounds (N, squares) to mined observations (O, circles) adds ON comparisons, but no comparable NN pairs, without changing observed prediction orderings. Symbols are schematic, not sample counts.

intervals with comparisons involving lower bounds. Success on the new comparisons can raise the score while the existing prediction ordering remains unchanged (Figure 1c).

A post-hoc audit of a fixed dataset. We examine these issues in an author-built dataset and saved predictions from three models. Following measurement principles (Jacobs & Wallach, 2021; Liu et al., 2024; Bean et al., 2025), a **label consistency check** traces each stored interval to its target work under the dataset's rules and retained records. We apply established **C-index decomposition by pair type** (Alabdallah et al., 2024) to separate comparisons between observed intervals from those involving curated lower bounds. A source-only diagnostic uses just these two record types to test how much concordance is possible without within-source timing information.

Main findings. At least 8 of 34 eligible observed held-out samples have labels inconsistent with the retained records. Adding curator-specified lower bounds raises model C-indices from 0.42–0.45 to 0.61–0.67 on their respective valid subsets through new cross-source comparisons, without changing any observed–observed prediction ordering. The source-only diagnostic scores higher than each model without ordering times within either source. Excluding the known mismatches or recomputing their labels does not establish above-chance observed-interval ranking. Two models exceed the specified output-failure threshold, so their affected results remain exploratory.

Why both checks are needed. The pooled-minus-observed score difference persists after recomputing the known mismatches from retained records, including on samples valid for all three models. Label consistency and score composition therefore need separate checks. Reporting label-generating records and concordance by pair type makes both inspectable alongside the pooled score.

2 Background and Related Work

Scientific influence and independent alternatives. Quantitative studies of science use publication and citation records to examine discovery and influence (Fortunato et al., 2018). Citation-based approaches include models of a paper's future citation trajectory (Wang et al., 2013) and estimates of its causal impact on follow-up papers (Agrawal et al., 2024). These endpoints concern scholarly uptake; reinvention time concerns the arrival of an independent capability substitute. Citation

indicators depend on construction choices and varied citing motives (Bornmann & Daniel, 2008; Waltman, 2016). Matching indicators to stated research goals is a principle of responsible research assessment (Hicks et al., 2015). Historical studies of multiple discovery (Ogburn & Thomas, 1922; Merton, 1961) motivate looking for alternative routes to similar advances. Simultaneous invention and idea twins also provide research designs (Ganguli et al., 2020; Bikard, 2020). We mine candidate pairs from a frozen Papers with Code mirror and use OpenAlex metadata (Priem et al., 2022) for author/institution-overlap screening of near-simultaneous candidates. A candidate interval acquires a substitute-time interpretation only with evidence for functional equivalence, independence, and earlier-candidate coverage.

What scientific prediction benchmarks ask. BrainBench presents original and altered neuroscience abstracts, testing which experimental result is consistent with the study (Luo et al., 2025). PreScience forecasts future paper records and trends, including contributions and citations (Ajith et al., 2026). CUSP evaluates scientific events' feasibility, mechanisms, solutions, and dates of public observability (Wu et al., 2026); FIRE-Bench evaluates agents' rediscovery of established insights (Wang et al., 2026). Reinvention time is the delay from a focal work to its first independent, functionally equivalent substitute. We compare these benchmarks in terms of their prediction targets and label definitions.

From task definitions to label evidence. Measurement modeling distinguishes theoretical constructs from operational indicators (Jacobs & Wallach, 2021). Dataset-development research explains how task framing, collection, and annotation shape the resulting evaluation (Paullada et al., 2021). Datasheets make these choices explicit (Gebru et al., 2021), while Evidence-Centered Benchmark Design connects intended uses, capabilities, task content, and scoring through five design modules (Liu et al., 2024). BetterBench audits lifecycle practices (Reuel et al., 2024), and Bean et al. (2025) examine construct validity across 445 LLM benchmarks. Item Response Theory (IRT) provides response-based diagnostics: Zhou et al. (2026) estimate item properties with PSN-IRT across 11 benchmarks, and Land & Bikel (2026) use a four-parameter IRT likelihood comparison to flag likely mislabels. Our label check instead compares the focal work with the records generating its time label, independently of the evaluated models' predictions. This evidence-provenance question also appears in Huang & Yang (2026), who trace how evidence at one unit is assigned to samples in encrypted-traffic benchmarks. Our check detects internal contradictions in the retained records; label-error audits such as Northcutt et al. (2021) additionally validate candidate errors through crowdsourced human review.

Diagnosing unintended score signals. Shortcut-learning research distinguishes benchmark success from the intended ability and motivates baselines that omit supposedly necessary information (Geirhos et al., 2020). For example, a hypothesis-only classifier can exploit annotation artifacts in natural language inference without reading the premise (Gururangan et al., 2018). Our source-only diagnostic follows this logic: it asks how much concordance the comparison structure permits without any within-source timing information. It tests a property of the evaluation; identifying the LLMs' actual decision strategy is a separate task.

Interpreting concordance under censoring and pooling. Survival-analysis work shows that commonly used C-statistics can depend on a study's censoring distribution (Uno et al., 2011), while within- and between-cluster concordance serve different interpretation goals (van Klaveren et al., 2014). Most directly, Alabdallah et al. (2024) separate event–event from event–censored concordance and show how pooled scores conceal differences between them. Our OO/ON partition is that same partition: censoring status coincides with mined versus curated source membership. We apply the established decomposition to the saved predictions and connect it to the source-only diagnostic.

3 Prediction Task and Dataset Construction

3.1 The intended prediction target

The *focal work* is the scientific contribution represented by the target paper. For focal work a with initial submission date t_a, $G(a)$ is the delay to its first independent, functionally equivalent

substitute. Measuring this delay requires checking that the candidate qualifies and that no earlier qualifying substitute exists. If none has appeared by a cutoff, the elapsed time gives only a lower bound, represented as a right-censored observation. The dataset has no external validation of these claims, so $G(a)$ remains the intended prediction target rather than a verified ground truth. It also differs from the counterfactual effect of removing a.

3.2 Dataset construction

Mined candidate intervals (N1 and N2). The observed set contains 79 connected components of near-simultaneous candidate pairs (N1) and 37 selected delayed pairs (N2). Retained pairs required three valid grades, each strong or moderate, from `gpt-5.4-mini`, `qwen3.7-plus`, and `deepseek-v4-flash`. Their prompts showed full titles; only years and month names in 1,100-character abstract excerpts were regex-masked. Operational independence screens used exact shared-author strings and initial-version cross-citations; N1 additionally used limited OpenAlex author/institution-overlap checks. These gates assess agreement under the implemented rules; functional equivalence and independence require external validation.

Curated lower bounds (N3). N3 was constructed to represent landmarks with claimed long waits for an independent equivalent, complementing the mined gap windows. Fifteen separately curated N3 landmarks have curator-specified lower bounds of 2,387–4,423 days at the 2024-12-31 cutoff. They are treated as right-censored observations in scoring, based on unverified claims of no equivalent before that date. Their descriptions received no target-identity screening.

Development and held-out samples. Observed rows were outcome-sorted and interleaved into 77 development and 39 held-out samples, without enforcing disjoint paper membership. Five held-out input records are flagged by the implemented co-discoverer title-fragment scan, leaving 34 eligible observed rows. The secondary pooled set adds all 15 N3 lower bounds, giving 49 eligible rows before model-output failures. The original labels, split, input records, predictions, and primary tables remain fixed; the checks in this paper are separately versioned and post hoc. Further construction and deviation details appear in Appendix B.

3.3 Two construction challenges

The target is a work, but the mined evidence consists of pairwise relationships. N1 aggregates near-simultaneous pairs into connected components; N2 retains selected delayed pairs. Neither operation by itself guarantees that a row's label is the earliest interval associated with the work described in its prompt. Component membership also does not guarantee that every member is an independently validated substitute for the earliest member.

This mismatch precedes model prediction and can shift a label in either direction. Figure 1b and Table 1 show cases where the component minimum is smaller or larger than the target-anchored gap.

The second challenge concerns what can be compared. N3 rows encode lower bounds, not observed event times. Such a row may be ordered relative to a shorter observed gap, while two N3 rows cannot be ordered by their unknown event times. Pooling them therefore changes the question answered by the rank metric: the score now also rewards separating observed gaps from lower bounds.

4 Audit Methods and Evaluation Protocol

We check whether labels correspond to their target works, decompose the ranking score, and examine both through label and sample-selection sensitivities.

4.1 Checking label consistency

We compare each stored label with the interval obtained by applying the dataset's rules to the focal work and its retained records. A disagreement means that the stored answer and the question refer to different intervals. This checks internal consistency; candidate validity and search completeness require independent evidence.

Component labels (N1). For N1, we follow the construction's treatment of a component as one candidate discovery: the reference searches all its members, including those without a directly adjudicated edge to the target. This checks the construction's own closure assumption, not the real-world validity of every implied relationship.

For an N1 connected component $e = (V_e, E_e)$ whose earliest member a supplies the input target,

$$Y_e = \min_{(u,v)\in E_e} |t_u - t_v|, \qquad Z_e(a) = \min_{b\in V_e\setminus\{a\}} (t_b - t_a). \tag{1}$$

$Z_e(a)$ is the graph-derived reference interval within the retained component. For N1, $B_e^{\mathrm{N1}} = \mathbf{1}\{Y_e = Z_e(a)\}$. Same-day members are allowed, so $Z_e(a)$ can be zero.

Selected-pair labels (N2) and lower bounds (N3). For N2 selected pairs, we check whether an earlier candidate for the same target already exists in the retained graph. The comparison uses actual submission dates, because stored gap values include month-to-day rounding. The graph-derived reference interval $Z_e^+(a)$ uses the earliest retained candidate's gap encoding. Appendix A gives the exact date and encoding rules. For N3, the numeric predicate $B_e^{\mathrm{N3,num}} = \mathbf{1}\{Y_e = t_c - t_a\}$ passes by construction at cutoff t_c. The separate claim that no qualifying substitute occurred before t_c remains an unverified curator assertion.

4.2 C-INDEX AND DECOMPOSITION BY PAIR TYPE

The C-index measures how often predicted waiting times correctly order the pairs that can be compared. Its censoring-aware form (Harrell et al., 1982) compares two unequal observed gaps, or an observed gap strictly below another row's lower bound; prediction ties receive half credit. Two censored N3 rows are never comparable. Independent continuous random rankings have expected concordance 0.5 on the admitted pairs.

Partition the comparisons into observed–observed (OO), observed–N3 (ON), and N3–N3 (NN). Let D_g count pairs in stratum g, Q_g count concordant pairs plus half the prediction ties, and $C_g = Q_g/D_g$ when $D_g > 0$. Since $D_{NN} = 0$, the pooled score and its increase satisfy

$$C_{\mathrm{pool}} = \frac{D_{OO}C_{OO} + D_{ON}C_{ON}}{D_{OO} + D_{ON}}, \tag{2}$$

$$C_{\mathrm{pool}} - C_{OO} = \frac{D_{ON}}{D_{OO} + D_{ON}}(C_{ON} - C_{OO}). \tag{3}$$

Both expressions assume nonempty OO and ON strata. They apply the decomposition of Alabdallah et al. (2024), using pair-count weights for event–event and event–censored comparisons. Adding N3 leaves OO orderings unchanged: the increase depends exactly on the share and concordance of the new ON pairs.

A source-only diagnostic. To test how much the score rewards source separation alone, define $S_i = \mathbf{1}\{i \in \mathrm{N3}\}$. It ties every OO prediction and correctly orders every comparable ON pair, giving

$$C_{\mathrm{source}} = \frac{0.5D_{OO} + D_{ON}}{D_{OO} + D_{ON}} = \frac{1}{2} + \frac{1}{2}\frac{D_{ON}}{D_{OO} + D_{ON}}. \tag{4}$$

The diagnostic assigns one value to all observed rows and another to all N3 rows. It uses known dataset membership, so it is a check on the score rather than a deployable forecast. It contains no within-source timing information.

4.3 MODEL INPUTS AND PREDICTIONS

Inputs are anchored at the earliest member's initial submission date t_a. The masked condition includes a title-withheld, heuristically masked abstract excerpt (up to 1,400 characters), three date-filtered field-activity features, and up to six earlier-work titles. Retrieval uses the unmasked target. The static mirror and model parameters can retain later information, leaving temporal leakage unresolved. Appendix B details masking and the auxiliary raw-abstract condition.

Gemini-3.1-pro-preview, kimi-k3, and glm-5.2 output month-valued point estimates, which we convert to days ($\times 30.4$) for reporting. The code neither fits model parameters nor supplies development

labels as examples. Development inspection preceded the internal plan, without probability calibration. Masked N1/N2 rows use calls at temperatures 0.0, 0.4, and 0.8; the middle ordered valid point estimate is scored (the larger value when only two survive). N3 and raw-abstract rows use one temperature-0 call. Evaluation and adjudication used distinct requested API model identifiers; backend identities were not independently verified.

4.4 Analyses and uncertainty

The primary ranking analysis reports observed held-out C_{OO} from saved predictions on model-specific subsets with valid predictions. A secondary pooled analysis includes N3. Citation and pre-t_a field-activity baselines are evaluated on matched observed rows; the citation count is a retrospective collection-time indicator and is excluded from evidence of leakage-free forecasting. A later Cox gradient-boosting baseline fitted to the 77 development samples is exploratory. Primary intervals use 2,000 sample-level bootstrap attempts and nearest-rank percentiles. Resamples with no comparable pair are discarded; at least 90% of attempts must remain. A post-hoc sensitivity resamples groups of events linked by shared papers, retaining all evaluated members of each drawn group. Groups are built before projection onto each model's observed subset. This addresses recorded within-group dependence in the resampling, while preserving the original split and point estimates. All intervals remain descriptive and unadjusted for multiplicity.

Deterministic diagnostics use only saved predictions. We compute C_{ON} and actual D_{OO}/D_{ON} from the pooled records, then evaluate Equation 4 on each model-specific valid subset. We assess the observed score's sensitivity to known mismatches in two ways: excluding the eight eligible rows already flagged by the prediction-free audit, and replacing only their stored gaps with $Z_e(a)$ or $Z_e^+(a)$. The replacement probes label-definition sensitivity using retained-graph references; it supplies no external truth. A separate, previously computed sensitivity restricts analysis to samples in singleton components of the graph linking samples that share a paper, excluding every component with multiple samples. Appendix D crosses label definitions with sample-selection rules in a post-hoc point analysis. We also compare upper-median, ordinary-median, and mean aggregation while holding valid calls and sample membership fixed. Appendix C reports all aggregation and group-bootstrap sensitivities separately from the unchanged primary analysis.

Missing predictions and failed calls. We also compute conservative bounds on the observed-sample C-index under completion of fully missing predictions and hypothetical completion of failed observed-sample calls (Appendix Table 5). They are deterministic sensitivity envelopes for these two completion regimes.

The 49-row eligible masked set schedules 117 calls ($34\times3+15\times1$). The implemented output-quality rule flags a model when either call-level failure or sample-level missingness exceeds 5%. In the masked held-out-plus-N3 run, Kimi and GLM exceed the call-level threshold (GLM also exceeds the sample-level threshold), so all Kimi/GLM results using these masked calls are exploratory; Gemini passes both thresholds. A sample is nevertheless scored whenever at least one call is valid, so surviving one- and two-call records enter model-specific subsets with valid predictions. Further input record, missing-output, split, and planning deviations are listed in Appendix B.

5 Results

5.1 Do the labels answer the target-specific question?

All 59 two-member N1 components satisfy $Y_e = Z_e(a)$ by construction, so the nontrivial audit covers the 20 larger components. Eleven of 20 fail, five held out. The N2 check finds a strictly earlier retained candidate for 9/37 rows, three held out. These held-out sets are disjoint, and all eight rows pass the implemented title-fragment scan. At least 8/34 eligible observed held-out rows have labels inconsistent with the retained graph. The other 26 rows have not been externally validated. Across these eight rows, absolute original-to-reference differences are 1–34 days (median 6) for the five N1 cases and 43–359 days (median 53) for the three N2 cases.

Examples of label mismatches. Table 1 shows how labels can refer to a different pair or miss an earlier retained candidate. In N1E007, a component member appears two days after the target,

but the component label comes from an eight-day retained edge. In N1E015, A is Parallel Instance Query Network for Named Entity Recognition (arXiv:2203.10545); the label instead comes from the later B–C edge. In N2G010, an earlier retained candidate exists for the same target. The N2 calendar dates are 65 days apart, but the original pipeline re-encoded its stored 2.1-month gap as 63 days; the original primary analyses preserve that stored encoding. Public identifiers and initial submission dates support the identity/date statements.

Table 1: Three documented label inconsistencies. A/B/C denote temporal order. Y_e is unchanged; Z is a graph-derived reference interval from the retained graph.

Row	Recorded timeline	Consistency check
N1E007	A: 2105.14217 (May 29); earliest component member B: 2105.14734 (May 31); retained edge A–C: 2106.03180 (Jun 6)	$Y_e = 8$d (A–C); $Z_e(A) = 2$d
N1E015	A: 2203.10545 (Mar 20) → B: 2205.05904 (May 12) → C: 2205.15485 (May 31)	$Y_e = 19$d (B–C); $Z_e(A) = 53$d
N2G010	A: 2211.09394 (Nov 17) → earlier retained B: 2212.03506 (Dec 7); selected C: 2301.08855 (Jan 21)	$Y_e = 63$d; $Z_e^+(A) = 20$d

The complete eight-row held-out case list appears in Appendix A. These cases establish that the stored label can answer a different temporal question even before a model is evaluated.

5.2 What creates the pooled score increase?

Table 2 reports the C-index decomposition by pair type, Figure 2 visualizes it, and Table 3 reports sensitivity to labels and shared papers. Observed-sample C_{OO} is 0.424–0.452, and every descriptive interval contains 0.5. Adding N3 raises the pooled statistic to 0.609–0.674. All observed–observed prediction orderings remain unchanged. The largest eligible observed label is 1,231 days and the smallest N3 lower bound is 2,387 days, so every available observed–N3 pair is comparable: $D_{ON} = n_O n_3$. Depending on model missingness, this adds 352–510 ON pairs, accounting for 41.6–47.7% of all comparable pairs. Their concordance is 0.869–0.935, whereas C_{OO} remains unchanged. There are no comparable NN pairs. The higher pooled score therefore reflects separation of curated lower bounds from mined observations. It does not test the ordering of N3 event times: none is observed, so $D_{NN} = 0$.

A high score without within-source ranking. The source-only diagnostic (Equation 4) scores 0.739, 0.734, and 0.708 on the Gemini, Kimi, and GLM valid subsets, respectively. Each exceeds the corresponding pooled model score despite carrying no timing signal within either source.

Comparison with simple baselines. On matched observed subsets ($n = 33/32/31$ for Gemini/Kimi/GLM), model-minus-citation C-index differences are $-.104$, $-.072$, and $-.099$; all three descriptive intervals cross zero. The exploratory structured baseline scores 0.572 [.442,.693] on 34 rows, and matched model-minus-structured intervals also cross zero. These comparisons establish no clear advantage for the LLMs. The citation indicator is retrospective; the structured baseline was specified after held-out access and fitted on development samples. Appendix C gives the full differences.

Table 2: Observed and pooled scores. Brackets are descriptive, unadjusted 95% sample-level bootstrap intervals; C_{ON} is a point decomposition with no interval estimated. Parentheses are valid observed/N3 rows. †: the masked +N3 run exceeded the 5% output-quality rule through call-level failures; GLM also exceeded it at sample level.

Model (n_O/n_3)	C_{OO} [interval]	D_{OO}/D_{ON}	C_{ON}	C_{pool} [interval]
Gemini (34/15)	.436 [.305,.570]	559/510	.935	.674 [.563,.775]
Kimi† (33/14)	.452 [.315,.611]	526/462	.878	.651 [.519,.777]
GLM† (32/11)	.424 [.312,.541]	494/352	.869	.609 [.496,.711]

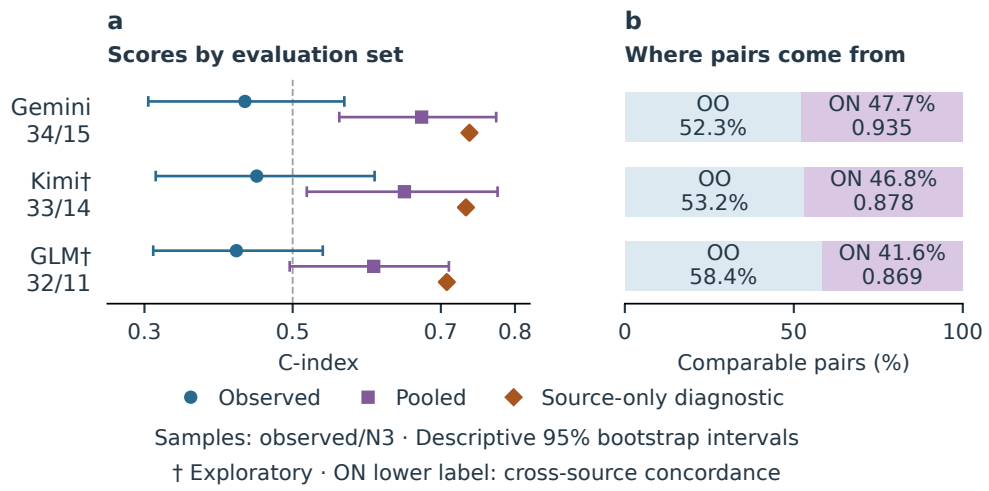


Figure 2: C-index decomposition. (a) Observed and pooled model scores with descriptive 95% intervals, plus the interval-free source-only diagnostic. (b) Pooled comparable-pair shares for observed–observed (OO) and observed–N3 (ON); the lower number in each ON segment is its concordance. Counts below model names give valid observed/N3 rows. Missing outputs make the three valid subsets model-specific; † marks exploratory Kimi/GLM results under the output-quality rule.

Table 3: Post-hoc observed-sample sensitivities, reported as C-index [descriptive 95% interval] (n). Label recomputation substitutes graph-derived focal-work references for the eight flagged rows; these are internal audit quantities. †: same masked-call output-quality boundary as Table 2.

Model	Exclude 8 mismatches	Label recomputation	Singleton components
Gemini	.362 [.229,.503] (26)	.450 [.319,.582] (34)	.476 [.314,.630] (27)
Kimi†	.362 [.222,.532] (25)	.459 [.315,.598] (33)	.448 [.293,.617] (26)
GLM†	.370 [.252,.490] (24)	.437 [.327,.542] (32)	.367 [.257,.482] (26)

5.3 Do label and overlap sensitivities recover ranking performance?

Removing all eight eligible known-mismatch rows lowers the three point estimates to 0.362–0.370. Replacing their original labels with the retained-graph references in Equations 1 and 5 moves the estimates only to 0.437–0.459. At the label level, recomputation reverses the order of 41 of the 559 originally comparable pairs among all 34 eligible rows and turns two more into ties, leaving 557 comparable pairs.

Sensitivity to shared papers. Restricting to singleton components of the graph linking samples that share a paper yields 0.367–0.476. None of these descriptive intervals lies wholly above 0.5. The GLM exclusion and singleton-component intervals lie below 0.5; both remain exploratory under the output-quality rule. Thus the known mismatches change pair ordering, but their exclusion or label recomputation does not establish above-chance ranking in these saved outputs. With original labels and upper-median predictions, resampling shared-paper groups gives intervals of [.322,.574], [.320,.613], and [.288,.555] for Gemini/Kimi/GLM; all contain 0.5. The 34/33/32 observed samples occupy 31/30/30 groups. This changes uncertainty estimation, not the split: 7/7/6 evaluated events remain linked to development samples.

5.4 Does the pooled score increase survive label recomputation?

After label recomputation, pooled scores are 0.682, 0.655, and 0.617 for Gemini, Kimi, and GLM, respectively: still 0.232, 0.196, and 0.180 above their observed-sample scores. Every ON pair retains its comparability and concordance; recomputation changes only the OO comparisons. On the common valid subset (31 observed rows and 10 N3 rows), the reference-label pooled-minus-

observed differences remain 0.167–0.174. Appendix D reports these post-hoc point estimates and complete-call and N1/N2 sensitivities.

Complete calls and aggregation. Requiring every scheduled call to succeed across all three models leaves only 8 observed rows and 10 N3 rows. Their observed scores are .537/.574/.389 (Gemini/Kimi/GLM), and the 80 ON pairs account for 74.8% of comparisons. These point estimates describe a small, execution-selected subset; the original Kimi/GLM quality flags remain. With original labels and fixed model-specific valid samples, switching among upper median, ordinary median, and mean gives observed scores of .411–.452 and pooled scores of .609–.674. The pooled increase persists under all three rules (Appendix C).

Prediction-scale, missing-output, and raw-abstract diagnostics appear in Appendix C; the raw contrast changes content, call count, and instruction–input alignment together.

6 Discussion and Limitations

What a higher score means. The pooled score rewards separating short observed gaps from long lower bounds as well as ordering observed gaps. The source-only diagnostic gives the score attainable from source separation with all within-source predictions tied. Because source and censoring status coincide, it diagnoses the evaluation structure; identifying whether the LLMs actually use a shortcut requires separate evidence.

What inconsistent labels reveal. Label consistency and score composition require separate checks. Recomputing the known mismatches leaves the pooled increase largely intact, including on common valid samples. A label mismatch calls for revisiting the label rule; a pooled gain calls for examining which comparisons support its interpretation.

What to report alongside the score. Dataset documentation should preserve each row's target, label-generating records and rule, and source (Gebru et al., 2021; Paullada et al., 2021). Report inconsistencies with supporting records, and each pair type's D_g and C_g beside the pooled score. Keep exclusion and definition sensitivities separate from the original primary result.

Label validity and sampling. The evidence concerns internal consistency in an arXiv-era ML/NLP construction. Agreement among three model adjudicators can reflect correlated errors; functional equivalence, independence, and search completeness lack external domain validation. N1 components also imply relationships without direct adjudication. Bounded mining windows, ranked candidate selection, and curated N3 assertions provide no population sampling frame or justified mixture weights. Thus the 8/34 finding is a retained-graph lower bound, not a population error estimate; transfer to other publication practices remains untested.

Inputs, failed outputs, and uncertainty. Heuristic masking and retrospective model access leave temporal leakage unresolved. Failed calls affect aggregation and sample selection; complete-call filtering does not remove the Kimi/GLM quality flags. Shared-paper resampling retains the original development/held-out overlap. The internal plan followed development and N3 inspection, lacks external registration, and omitted multiplicity adjustment; the structured baseline followed held-out access. All uncertainty remains descriptive. Estimating a causal effect of label mismatch would require an exogenous design.

7 Conclusion

This audit finds labels inconsistent with retained target-specific records in at least 8/34 eligible observed held-out rows. Adding curator-specified lower bounds raises the C-index from 0.42–0.45 to 0.61–0.67 through cross-source comparisons, with observed-interval prediction orderings unchanged. The increase persists after label recomputation, including on samples valid for all three models. Label and sample-selection sensitivities do not establish above-chance observed-interval ranking. Interpreting pooled scores therefore requires both target-specific label records and concordance by pair type.

AI USE STATEMENT

Generative AI tools assisted literature discovery and synthesis, conceptual and mathematical formulation, hypothesis and methodology refinement, method implementation and analysis code, research artifacts, statistical interpretation, figure preparation, and manuscript drafting and editing. The research use of models as adjudicators and evaluated systems is described separately in Sections 3 and 4. We checked literature claims against the cited primary records, quantitative claims against internal frozen records, and AI-assisted code against fixed expected outputs. We reviewed all AI-assisted work and retain responsibility for the final content.

ETHICS STATEMENT

The audit reports paper identifiers to make construction inconsistencies inspectable; its evidence concerns label consistency for individual works. Because labels such as "necessary" can affect reputations, the proxy and its outputs must not be used for person- or institution-level assessment. Quantitative indicators should inform expert judgment, never replace it (Hicks et al., 2015).

REPRODUCIBILITY STATEMENT

An anonymous artifact provides the saved predictions, retained numerical and categorical construction records, analysis code, and figure-generation scripts needed to reproduce the saved-prediction analyses and quantitative result plots. It also supplies Figure 1 as an editable diagram and PDF. It does not replay the original model calls or reconstruct early data collection. Full text inputs are outside the artifact, and raw relay requests/responses were not retained. The internal plan predates held-out runs but follows development and N3 inspection and lacks external registration. Deviations and post-hoc audits are documented.

REFERENCES

Ishan Agrawal, Zhijing Jin, Ehsan Mokhtarian, Siyuan Guo, Yuen Chen, Mrinmaya Sachan, and Bernhard Schölkopf. CausalCite: A causal formulation of paper citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8395–8410, 2024. doi: 10.18653/v1/2024.findings-acl.497.

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. PreScience: A dataset and benchmark for scientific forecasting. *arXiv preprint arXiv:2602.20459*, 2026. URL `https://arxiv.org/abs/2602.20459v2`.

Abdallah Alabdallah, Mattias Ohlsson, Sepideh Pashami, and Thorsteinn Rögnvaldsson. The concordance index decomposition: A measure for a deeper understanding of survival prediction models. *Artificial Intelligence in Medicine*, 148:102781, 2024. doi: 10.1016/j.artmed.2024.102781.

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 Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott Hale, Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19868–19949, 2025. doi: 10.52202/085713-0590. Datasets and Benchmarks Track.

Michaël Bikard. Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8):1528–1543, 2020. doi: 10.1002/smj.3162.

Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008. doi: 10.1108/00220410810844150.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359: eaao0185, 2018. doi: 10.1126/science.aao0185.

Ina Ganguli, Jeffrey Lin, and Nicholas Reynolds. The paper trail of knowledge spillovers: Evidence from patent interferences. *American Economic Journal: Applied Economics*, 12(2):278–302, 2020. doi: 10.1257/app.20180017.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017.

Frank E. Harrell, Jr., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982. doi: 10.1001/jama.1982.03320430047030.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. Bibliometrics: The Leiden manifesto for research metrics. *Nature*, 520:429–431, 2015. doi: 10.1038/520429a.

Sizhe Huang and Shujie Yang. SoK: Where do flow labels come from? Auditing label provenance in encrypted traffic benchmarks. *arXiv preprint arXiv:2609.02140*, 2026. URL https://arxiv.org/abs/2609.02140v1.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Sander Land and Daniel M. Bikel. Auditing LLM benchmarks with item response theory. *arXiv preprint arXiv:2605.30504*, 2026. doi: 10.48550/arXiv.2605.30504. URL https://arxiv.org/abs/2605.30504v2. Version 2.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, 2024. doi: 10.18653/v1/2024.acl-long.861.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K. Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O. Cohen, Valentina Borghesani, Anton Pashkov, Daniele Marinazzo, Jonathan Nicholas, Alessandro Salatiello, Ilia Sucholutsky, Pasquale Minervini, Sepehr Razavi, Roberta Rocca, Elkhan Yusifov, Tereza Okalova, Nianlong Gu, Martin Ferianc, Mikail Khona, Kaustubh R. Patil, Pui-Shee Lee, Rui Mata, Nicholas E. Myers, Jennifer K. Bizley, Sebastian Musslick, Isil Poyraz Bilgin, Guiomar Niso, Justin M. Ales, Michael Gaebler, N. Apurva Ratan Murty, Leyla Loued-Khenissi, Anna Behler, Chloe M. Hall, Jessica Dafflon, Sherry Dongqi Bao, and Bradley C. Love. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9:305–315, 2025. doi: 10.1038/s41562-024-02046-9.

Robert K. Merton. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470–486, 1961. URL https://www.jstor.org/stable/985546.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

William F. Ogburn and Dorothy Thomas. Are inventions inevitable? A note on social evolution. *Political Science Quarterly*, 37(1):83–98, 1922. doi: 10.2307/2142320.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. doi: 10.1016/j.patter.2021.100336.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In *Proceedings of the 26th International Conference on Science and Technology Indicators (STI 2022)*, 2022. doi: 10.5281/zenodo.6936227. URL https://zenodo.org/records/6936227.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21763–21813, 2024. doi: 10.52202/079017-0685.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011. doi: 10.1002/sim.4154.

David van Klaveren, Ewout W. Steyerberg, Pablo Perel, and Yvonne Vergouwe. Assessing discriminative ability of risk models in clustered data. *BMC Medical Research Methodology*, 14(1):5, 2014. doi: 10.1186/1471-2288-14-5.

Ludo Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10 (2):365–391, 2016. doi: 10.1016/j.joi.2016.02.007.

Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013. doi: 10.1126/science.1237825.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02905v2.

Sean Wu, Pan Lu, Yupeng Chen, Jonathan Bragg, Yutaro Yamada, Peter Clark, David Clifton, Philip Torr, James Zou, and Junchi Yu. Scientific reasoning does not reliably translate into scientific forecasting in frontier AI. *arXiv preprint arXiv:2605.22681*, 2026. URL https://arxiv.org/abs/2605.22681v2.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? Rethinking large language model benchmarking with item response theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35085–35093, 2026. doi: 10.1609/aaai.v40i41.40814. URL https://ojs.aaai.org/index.php/AAAI/article/view/40814.

A Complete Label Inconsistency Cases

Table 4 lists all eight rows identified by the two retained-graph checks among the 34 eligible observed held-out rows. The calculation does not use model predictions. The reference interval is computed for the focal work from the retained graph under the rules in Section 4; external reinvention-time validation remains a separate task.

Table 4: Complete eligible held-out case list. Original and reference values are in days. For N1, the original label basis is the minimum retained edge and the reference candidate is the earliest later component member. For N2, they are the selected and earlier retained candidates. A larger N1 audit reference is still a mismatch because the component minimum need not equal the target-anchored interval.

Row	Target	Original label basis	Reference candidate	Y/Z
N1E007	2105.14217	2105.14217–2106.03180	2105.14734, next member	8/2
N1E009	2203.03825	2203.10743–2204.00811	2203.10743, next member	12/13
N1E011	2302.06397	2302.07739–2302.13610	2302.07739, next member	13/1
N1E015	2203.10545	2205.05904–2205.15485	2205.05904, next member	19/53
N1E017	2202.09741	2203.03937–2203.11987	2203.03937, next member	14/16
N2E010	2302.13610	2404.17178, selected	2305.01914 in N2G012	422/63
N2G010	2211.09394	2301.08855, selected	2212.03506 in N1E008	63/20
N2G011	2308.15846	2312.10439, selected	2310.16667 in N1E016	109/56

The frozen construction also retains arXiv:2302.13610, whose public record is marked withdrawn. Among these cases, its initial submission date enters N1E011 and N2E010; these remain checks of the retained records.

For N2G010, public initial dates are 65 days apart, but the original pipeline encoded $\lfloor 2.1 \times 30.4 \rfloor = 63$. Across N2, this rounding changes 24/37 encodings by at most two days (13 exact; stored-minus-calendar range -2 to $+1$); it is excluded from the 9/37 earlier-candidate finding.

Exact N2 consistency and recomputation rule. For an N2 row containing target a and selected candidate record (b_e, t_{b_e}, d_e), $Y_e = d_e$ is the original day encoding of the selected-pair gap. The pipeline computed it as $\lfloor 30.4 m_e \rfloor$ from a previously rounded month value m_e, so it need not equal the exact calendar difference $t_{b_e} - t_a$. Let $\mathcal{R}(a)$ collect retained candidate records (b, t_b, d) for target a: an N1 component member contributes its target-to-member calendar interval d, whereas an N2 row contributes its original day encoding. Define

$$T_e^+(a) = \min_{(b,t_b,d)\in\mathcal{R}(a):t_b>t_a} t_b, \qquad Z_e^+(a) = \min_{(b,t_b,d)\in\mathcal{R}(a):t_b=T_e^+(a)} d. \tag{5}$$

For N2, $B_e^{\mathrm{N2}} = \mathbf{1}\{Y_e = d_e \ \wedge \ T_e^+(a) = t_{b_e}\}$; it fails here when $T_e^+(a) < t_{b_e}$. $Z_e^+(a)$ is the corresponding graph-derived reference interval used in the sensitivity analysis. Comparing dates rather than day encodings prevents the rounding step from creating an earlier-candidate flag. The check tests whether the selected pair is earliest under the construction's retained-graph closure.

B Dataset Construction and Protocol Details

Candidate funnel and adjudication. Author-side execution records report 576,261 Papers with Code rows, 430,769 enumerated same-task pairs within 60 days, a ranked top-800 stage, and 725 rows after rule-based exclusions. The transient top-800 table was not retained. The adjudication covered 723 mined candidates and six canonical checks, yielding 729 rows, of which 727 have three valid grades. Initial-version citation screening excluded two candidates before first-batch adjudication; in the second batch, it followed initial adjudication and gated component construction. Five of the six canonical checks pass the final gate. Four historical batches record Fleiss $\kappa = 0.44$–0.58 on the four-category scale (Fleiss, 1971). Previous and final pools agree on 87.1% of 634 rows with complete binary decisions (Cohen's $\kappa = 0.679$). These numbers describe agreement, while label validity requires external evidence. N2 combines a 6–42-month stream with a 2–6-month gap-fill stream, although the older parameter table mentions only the former.

Independence and components. The N1 pipeline screens exact author strings, initial-version cross-citations, and OpenAlex author/institution overlap, retaining unresolved cases; N2 has no parallel institution screen. Latest-version reference lists are unsuitable for submission-time awareness because later revisions can add citations. N1 closes retained links transitively into components, including within-component pairs that received no direct judgment. A global overlap audit maps 116 observed rows to 95 components linking samples that share a paper; four components cross the development and held-out partitions, affecting seven eligible held-out rows.

Scope of the N3 rationales. The retained curator reasons use different time scopes from the encoded lower bounds. For example, DQN is described as having "no near-term equivalent," and Mixed Precision Training as having "no contemporaneous equivalent." The builder nevertheless assigns each row the full interval from its initial submission to 2024-12-31 (4,030 and 2,639 days, respectively), without a rationale-dependent follow-up check. These reasons alone do not document the absence of a substitute throughout those intervals. This provenance gap does not establish that a qualifying substitute actually appeared; it explains why the encoded lower bounds remain curator assertions rather than validated censoring observations.

Model inputs and masking. The implemented question reads: "starting from that moment, how long would it take until an INDEPENDENT team (not building on this work) produces a functionally equivalent method?" The input builder uses the earliest member as the target. We formalize this waiting-time question as a first qualifying substitute endpoint; the prompt does not define functional equivalence further. The prompt also requests an unused distribution over five gap bins. The input record includes 18- and 36-month same-task paper counts, a recent activity ratio, a masked target description, and six of up to eight stored earlier works. The implemented regex masks four-digit years and acronym/CamelCase tokens, but ordinary-capitalization names, venues, and month names can remain. A co-discoverer title-fragment scan excludes five of 39 held-out rows. N3 is assigned `leak_flag=false` by construction; for example, masked N3E001 still contains "Transformer." The stored input record includes the title, but the runner omits it from both masked and raw-abstract prompts. The raw-abstract condition uses the same excerpt cap without masking: at most 1,400 characters. Both conditions reuse a fixed instruction claiming that all dates, names, and venues are masked.

Execution and planning deviations. Transport exceptions were retried; an unparseable point estimate produced a parse failure without retry. The "seed" field indexes calls; the API received no seed parameter. Raw-abstract and N3 rows schedule one temperature-0 call, whereas masked observed rows schedule three. The internal plan was finalized after development and N3 inspection, before held-out runs, and has no external timestamp. It left Holm adjustment, publication-year and CausalCite baselines, a placebo cohort, and a censoring grid undone; the structured baseline postdates held-out access. The 5% output-quality rule checks call failures and sample-level missingness. Any sample with a valid call remains analyzable. Aggregation uses the sole estimate for one valid call and the larger estimate for two, rather than their arithmetic mean. The raw-abstract run has 1/49, 9/49, and 8/49 failed single-call rows for Gemini/Kimi/GLM; the latter two also exceed the rule. The post-output scan checked only the parsed reason field for six model-family terms (GPT-4, GPT-5, Llama, ChatGPT, Claude, and Gemini) and stored a Boolean. It did not invalidate or rerun a call. All 1,362 stored flags are false, indicating that these six substrings are absent from the parsed reason field.

C Additional Results

Prediction scale and missing outputs. Across eligible masked development, held-out, and N3 records, all 354 valid model–row aggregates lie between 30.4 and 912 days, with 306 in the 243.2–912-day band, while labels and lower bounds span 0–4,423 days (Figure 3). This scale compression produces overprediction for many short observed gaps and places all 40 valid N3 aggregates below their respective lower bounds. Kimi and GLM have 23/117 and 28/117 failed calls and 47/49 and 43/49 valid samples in the masked held-out-plus-N3 run, compared with 0/117 and 49/49 for Gemini; all results using those Kimi/GLM calls are exploratory. The daggers in Tables 2–3 mark this boundary.

Holding surviving aggregates fixed, conservative bounds on the C-index for samples with no valid predictions are [.425,.484] for Kimi and [.375,.491] for GLM. Allowing arbitrary completion of all failed observed-sample calls widens them to [.254,.662] and [.200,.691]. These wider envelopes cross 0.5; the original output-quality flags remain (Table 5).

Auxiliary comparisons. Against the retrospective collection-time citation indicator on matched observed subsets ($n = 33/32/31$), model-minus-citation differences are $-.104$ $[-.328, .113]$, $-.072$ $[-.307, .174]$, and $-.099$ $[-.294, .093]$ for Gemini/Kimi/GLM; all three intervals cross zero. The exploratory development-set-fitted structured baseline scores 0.572 [.442,.693] on all 34 rows; model-minus-structured intervals on matched subsets also cross zero.

The raw-abstract minus masked pooled C-index differences are $-.015$, $+.043$, and $-.008$ on common-row subsets. Raw used one call per row versus three scheduled for masked observed rows, retained the mismatched masking instruction, and omitted the title; the Kimi/GLM raw runs also exceed the output-quality rule. The contrast jointly changes content, call count, and instruction–input alignment; the latter two estimates are exploratory. Full intervals and auxiliary baselines appear below.

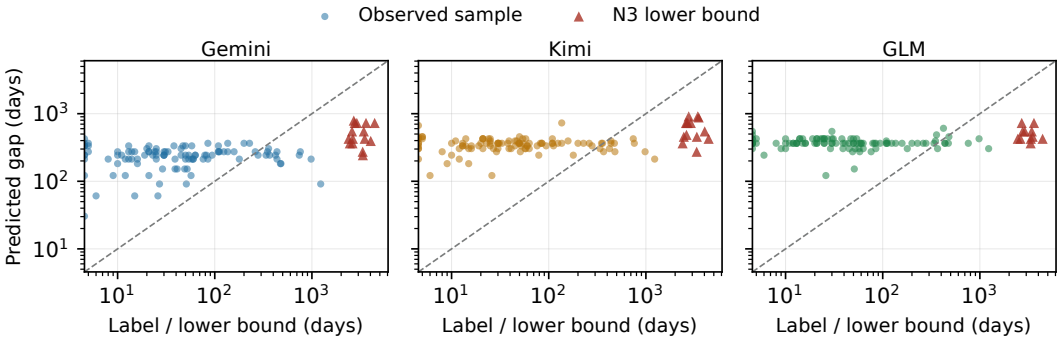


Figure 3: Saved masked-condition predictions against original labels or N3 lower bounds for valid development, held-out, and N3 samples. Predictions use the upper median (the larger value for two valid calls); values below 4.56 days are plotted at 4.56 days. Triangles mark lower bounds; all 40 N3 predictions fall below their lower bounds; the dashed line is equality.

Table 5: Missing-prediction sensitivity over 559 comparable pairs among 34 observed samples. Missing-sample bounds vary predictions for samples with no valid output, holding existing aggregates fixed; call-completion bounds also vary incomplete aggregates and ambiguous-pair credits. Both are deterministic envelopes.

Model	Samples (0/1/2/3 calls)	Missing-sample bounds	Call-completion bounds
Gemini	0/0/0/34	[.436,.436]	[.436,.436]
Kimi	1/2/15/16	[.425,.484]	[.254,.662]
GLM	2/4/10/18	[.375,.491]	[.200,.691]

Across 108 citation-complete observed rows, collection-time OpenAlex citations have Spearman $\rho = -0.05$ with original labels and the pre-t_a 18-month shared-task count has $\rho = 0.22$. On matched held-out subsets, their C-indices are 0.52–0.53 and 0.55–0.59, respectively. The development-set-fitted structured model has C-index 0.572 [.442,.693] on 34 observed rows and 0.393 [.268,.513] after adding N3. Refit on singleton-component development samples, it scores 0.551 [.416,.673] on 27 singleton held-out rows. On matched observed subsets ($n = 34/33/32$ for Gemini/Kimi/GLM), model-minus-structured differences are $-.137$ $[-.352, .055]$, $-.140$ $[-.362, .077]$, and $-.123$ $[-.305, .064]$; all three intervals cross zero.

On common pooled subsets of $n = 48/38/36$ for Gemini/Kimi/GLM, raw-minus-masked differences are $-.015$ $[-.063, .034]$, $+.043$ $[-.063, .159]$, and $-.008$ $[-.125, .127]$. Content and call count differ, so these intervals quantify a combined contrast; equivalence and identity effects remain unidentified.

Aggregation-rule sensitivity. We compare the original upper median with the ordinary median and arithmetic mean of the same valid calls. With two valid calls, the ordinary median uses their average; with three, it equals the original rule. The mean can also change three-call predictions. Samples with no valid call remain excluded, and the one-call N3 predictions are unchanged. Table 6 uses each model's original valid subset; Table 7 fixes the same 31 observed and 10 N3 rows for all models. Each table reports all three rules, with original labels and fixed membership. These post-hoc point comparisons have no new intervals or significance tests.

Ordinary-median aggregation changes 12 Kimi and four GLM predictions on the original valid subsets, all from two-call records. Mean aggregation additionally changes 28 Gemini, 13 Kimi, and 12 GLM three-call records. Observed scores remain below 0.5 in all 18 comparisons, and each pooled score exceeds its corresponding observed score. The changes do not remove the original output-quality flags or identify the outcomes of failed calls.

Table 6: Aggregation sensitivity on original model-specific valid subsets. U/M/A denotes upper median, ordinary median, and arithmetic mean. Changed (2/3) counts observed predictions changed from U among two-/three-call records; no one-call prediction changes. All scores are point estimates. †: exploratory under the original masked-call output-quality rule.

Model (n_O/n_3)	Rule	Changed (2/3)	C_{OO}	C_{ON}	C_{pool}	D_{OO}/D_{ON}
Gemini (34/15)	U	0/0	.436	.935	.674	559/510
	M	0/0	.436	.935	.674	559/510
	A	0/28	.415	.936	.664	559/510
Kimi† (33/14)	U	0/0	.452	.878	.651	526/462
	M	12/0	.447	.892	.655	526/462
	A	12/13	.450	.892	.656	526/462
GLM† (32/11)	U	0/0	.424	.869	.609	494/352
	M	4/0	.413	.891	.612	494/352
	A	4/12	.411	.901	.615	494/352

Table 7: Aggregation sensitivity on the common valid subset (31 observed and 10 N3 rows). Each row has 463 OO and 310 ON pairs. Notation and exploratory flags follow Table 6.

Model	Rule	Changed (2/3)	C_{OO}	C_{ON}	C_{pool}
Gemini	U	0/0	.461	.905	.639
	M	0/0	.461	.905	.639
	A	0/26	.454	.906	.635
Kimi†	U	0/0	.454	.885	.627
	M	11/0	.445	.892	.624
	A	11/12	.448	.890	.625
GLM†	U	0/0	.435	.861	.606
	M	3/0	.417	.879	.602
	A	3/12	.417	.890	.607

Shared-paper group bootstrap. We form 110 connected groups from all 131 development, held-out, and N3 events by linking events that share a paper, then project these groups onto each model's original valid observed subset. Gemini has 29 singleton groups, one two-event group, and one three-event group; Kimi has 28 singletons, one pair, and one triple; GLM has 29 singletons and one triple. Thus the corresponding group counts are 31, 30, and 30. Groups can be connected through development events even though only eligible observed held-out members are resampled.

For each of 2,000 attempts, we sample K of the K nonempty groups with replacement and retain every evaluated member of each draw, including repeated copies. We recompute the C-index over this expanded sample. Equal observed times, including copies of the same event, remain incomparable. Intervals use the nearest-rank 2.5th and 97.5th percentiles; at least 90% of attempts must have a comparable pair. All 2,000 attempts were valid for every model. Resampled event counts range from 31–45, 30–42, and 30–42 for Gemini/Kimi/GLM. Seeds are 20260920, 20260922, and

20260921, respectively. The original event-level intervals are reproduced using their saved seeds and event-ID ordering.

Table 8: Observed-sample C-index with descriptive, unadjusted 95% intervals from event-level and shared-paper group bootstrap. Predictions, labels, and point estimates are identical across the two resampling schemes. Dev-linked counts evaluated events whose global group also contains a development event. †: original output-quality flag retained.

Model	Events/groups	C_{OO}	Event interval	Group interval	Dev-linked
Gemini	34/31	.436	[.305,.570]	[.322,.574]	7
Kimi†	33/30	.452	[.315,.611]	[.320,.613]	7
GLM†	32/30	.424	[.312,.541]	[.288,.555]	6

All group-bootstrap intervals contain 0.5 (Table 8). Grouping changes the resampling unit, while leaving the development/held-out overlap and nonrandom dataset construction intact. It accounts for the recorded paper-sharing groups in this descriptive sensitivity, not unrecorded dependence or population sampling uncertainty.

D Joint Label and Sample-Selection Sensitivities

Question and design. The label audit and the pair-type decomposition concern different stages of evaluation. To examine them jointly, we cross two label definitions with four sample-selection rules, separately for all observed rows, N1, and N2. This post-hoc analysis uses saved masked-condition predictions and the eight previously flagged label inconsistencies. Predictions and membership are fixed within each original/reference comparison. Reference labels replace only flagged observed intervals present in that subset; N3 lower bounds and all other labels remain unchanged. Tables 9–11 report all 72 combinations of model, subset, observed group, and label definition. They contain point estimates only, with no new confidence intervals or hypothesis tests.

Sample-selection rules. *Available* uses each model's rows with at least one valid point estimate, preserving the original upper-median aggregation rule. *Common valid* intersects those rows across all three models. *Complete* retains only rows with every scheduled call valid for that model: three of three for observed rows and one of one for N3. *Common complete* intersects the three complete-call subsets. Completeness is therefore a stricter requirement for observed rows than for N3. It selects on execution success rather than recovering a random sample. For N1 or N2 analyses, the observed part is restricted to that source and the pooled score retains all N3 rows admitted by the subset rule. The subgroup analyses consequently share N3 rows and are not independent replications.

Joint results and common membership. On the available subsets, recomputation increases C_{OO} by 0.014, 0.007, and 0.013 for Gemini, Kimi, and GLM, respectively. The pooled-minus-observed differences change from 0.238, 0.199, and 0.185 to 0.232, 0.196, and 0.180. All ON comparisons remain unchanged: the revised observed intervals still lie below every retained N3 lower bound. The common-valid subset contains 31 observed rows (21 N1 and 10 N2), all eight flagged cases, and 10 N3 rows. Here, recomputation leaves 461 OO pairs and 310 ON pairs, with pooled-minus-observed differences of 0.174, 0.172, and 0.167.

Table 9: Joint label and sample-selection sensitivities, with N1 and N2 combined. O/R denotes original/reference labels; n_O/n_3 denotes observed/N3 sample counts. D_{ON} is unchanged by recomputation. All scores are post-hoc point estimates, with no new intervals or tests. †: exploratory under the original masked-call output-quality rule, including after complete-call filtering.

Subset	Model	n_O/n_3	D_{OO} (O/R)	D_{ON}	C_{OO} (O/R)	C_{pool} (O/R)
Available	Gemini	34/15	559/557	510	.436/.450	.674/.682
	Kimi†	33/14	526/524	462	.452/.459	.651/.655
	GLM†	32/11	494/492	352	.424/.437	.609/.617
Common valid	Gemini	31/10	463/461	310	.461/.472	.639/.646
	Kimi†	31/10	463/461	310	.454/.458	.627/.630
	GLM†	31/10	463/461	310	.435/.447	.606/.613
Complete	Gemini	34/15	559/557	510	.436/.450	.674/.682
	Kimi†	16/14	119/119	224	.483/.517	.751/.762
	GLM†	18/11	152/152	198	.375/.414	.667/.684
Common complete	Gemini	8/10	27/27	80	.537/.537	.813/.813
	Kimi†	8/10	27/27	80	.574/.574	.813/.813
	GLM†	8/10	27/27	80	.389/.389	.738/.738

Complete-call selection. Complete-call filtering leaves 34, 16, and 18 observed rows for Gemini, Kimi, and GLM. On their common intersection, only eight observed rows remain (six N1 and two N2), alongside 10 N3 rows. The original and reference scores coincide there: two label values change, but the 27 comparable OO pair orders and the remaining label tie do not. Observed scores are 0.537, 0.574, and 0.389; the 80 ON comparisons constitute 74.8% of admitted pairs. These higher Gemini/Kimi point estimates describe a small selected subset. Filtering does not remove the original Kimi/GLM output-quality failures or establish above-chance performance.

N1/N2 heterogeneity. Tables 10 and 11 show the source-specific results. On common-valid rows with reference labels, C_{OO} is 0.498/0.515/0.510 for N1 and 0.400/0.456/0.422 for N2 (Gemini/Kimi/GLM). Label recomputation can move subgroup scores in either direction: GLM's common-valid N1 score rises from 0.474 to 0.510, while its N2 score falls from 0.556 to 0.422. Thus the combined observed result does not imply uniformly below-chance point estimates in every subgroup. N1 and N2 also have different construction and gap-selection rules (Appendix B), so these contrasts are descriptive.

Table 10: N1 sensitivities, using the same notation and subset definitions as Table 9. Each pooled score combines the retained N1 rows with all N3 rows admitted by that subset rule; it excludes N2.

Subset	Model	n_O/n_3	D_{OO} (O/R)	D_{ON}	C_{OO} (O/R)	C_{pool} (O/R)
Available	Gemini	23/15	251/249	345	.458/.474	.732/.740
	Kimi†	22/14	229/227	308	.483/.482	.700/.701
	GLM†	22/11	229/227	242	.461/.498	.668/.687
Common valid	Gemini	21/10	208/206	210	.493/.498	.699/.702
	Kimi†	21/10	208/206	210	.524/.515	.697/.694
	GLM†	21/10	208/206	210	.474/.510	.666/.685
Complete	Gemini	23/15	251/249	345	.458/.474	.732/.740
	Kimi†	10/14	44/44	140	.625/.625	.823/.823
	GLM†	12/11	65/65	132	.400/.492	.708/.739
Common complete	Gemini	6/10	14/14	60	.643/.643	.851/.851
	Kimi†	6/10	14/14	60	.571/.571	.831/.831
	GLM†	6/10	14/14	60	.464/.464	.757/.757

Table 11: N2 sensitivities, using the same notation and subset definitions as Table 9. Each pooled score combines the retained N2 rows with all admitted N3 rows; it excludes N1. The common-complete N2 subset has just one OO pair.

Subset	Model	n_O/n_3	D_{OO} (O/R)	D_{ON}	C_{OO} (O/R)	C_{pool} (O/R)
Available	Gemini	11/15	55/55	165	.391/.391	.805/.805
	Kimi†	11/14	55/55	154	.500/.536	.801/.811
	GLM†	10/11	45/45	110	.556/.422	.787/.748
Common valid	Gemini	10/10	45/45	100	.400/.400	.752/.752
	Kimi†	10/10	45/45	100	.411/.456	.762/.776
	GLM†	10/10	45/45	100	.556/.422	.772/.731
Complete	Gemini	11/15	55/55	165	.391/.391	.805/.805
	Kimi†	6/14	15/15	84	.367/.500	.823/.843
	GLM†	6/11	15/15	66	.667/.667	.901/.901
Common complete	Gemini	2/10	1/1	20	1.000/1.000	.929/.929
	Kimi†	2/10	1/1	20	.500/.500	.881/.881
	GLM†	2/10	1/1	20	1.000/1.000	.952/.952

The common-complete N2 subset has only two observed rows and one comparable OO pair. Gemini and GLM each score 1.000 on OO, but 0.929 and 0.952 after pooling. Equation 3 gives the exact condition: pooling raises the score when $C_{ON} > C_{OO}$ and lowers it when the inequality reverses. The pooled increases in the main results are empirical properties of those comparisons, not a universal effect of adding censored rows.