When Better Than Chance Is Not Enough: Net-Benefit Bands for LLM Verifiers

Gi Young Lee[1,*] **Chansam Park**[1,*] **Taewan Goo**[1] **Juhee Kim**[2]
Jong Cheol Jeong[3] **Taejeong Bae**[4] **Dongwon Yun**[4] **Tae Hyung Kim**[1,†]
[1] BioNexus Inc.
Gwanggyo Business Center, 156 Gwanggyo-ro, Yeongtong-gu,
Suwon-si, Gyeonggi-do, Republic of Korea
[2] Korean Organ Transplantation Registry
[3] Department of Internal Medicine,
Seoul National University Bundang Hospital
[4] Korea University
[*]Equal first authorship. [†]Corresponding author: `thkim@bionexus.kr`

Abstract

An LLM verifier can distinguish errors yet harm a system when used to filter outputs. In a biomedical metadata pipeline, we compared a baseline verifier with seven modifications. The rebuilt verifier improved Youden's J, the difference between incorrect- and correct-call flag rates, from -10% to $+11\%$, and retained-call precision from 34% to 39%. Decision-curve analysis showed that it had the highest estimated net benefit among the compared strategies only for action thresholds $p_t \in [29.5\%, 39.2\%]$. The balanced-accuracy evaluation objective implies $p_t = 14.4\%$, outside this range; this metric-implied trade-off does not establish operator preferences. For the binary filter and reference strategies studied here, band existence is equivalent to $J > 0$. Its location adds information: the endpoints are the precisions of discarded and retained calls. A synthetic example shows that equal positive J values can lead to different deployment choices. The empirical band was empty in 10.5% of 4,000 study-level bootstrap resamples. We propose reporting the deployment band, its uncertainty, and the intended operating point alongside verifier discrimination in the Judge Deployment Disclosure Template.

1 Introduction

An LLM verifier (or judge) checks another model's output. Agreement with human ratings, stability under paraphrase, and error discrimination are useful diagnostics, but do not establish whether acting on its verdict helps. Suppressing an output, requesting human review, or triggering regeneration has consequences that depend on the base system, population, and error costs.

We discovered this gap after testing a baseline verifier, seven modifications, and eight flag-aggregation rules in a biomedical metadata pipeline. The rebuilt verifier improved discrimination, although its 95% interval included zero. Its point-estimate net benefit was highest for some error trade-offs, but not the one implied by our balanced-accuracy objective. We study binary filtering at inference time; other interventions require their own utility analysis.

Contributions. (i) A production case study of eight verifier configurations. The disease adapter shows the clearest harmful pattern, which is consistent with a legacy prompt edit; several plausible repairs do not improve the downstream decision (Section 4). (ii) A net-benefit analysis showing that, for this binary filter and the stated reference strategies, a useful band exists exactly when $J > 0$,

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: JUDGe: Can We Trust the Judge?.

but its location cannot be inferred from J alone (Section 5). (iii) A connection to decision-curve analysis, used in clinical prediction since 2006, and a proposal to add a deployment-band field to the Judge Deployment Disclosure Template (Section 6).

2 Related work

Verifier evaluation. LLM judges are commonly evaluated by agreement with human raters [21] and by probes for ordering effects and self-preference [15]. Other work asks whether model confidence tracks correctness [10, 20]. These studies characterize a verifier; they do not by themselves determine whether acting on its verdict improves a system.

Self-critique. LLM self-verification has known limits [9, 17], while tool-interactive external feedback can improve self-correction [8]. Configuration E tests a narrower idea: give the verifier the linked paper's abstract and methods while leaving the escalation gate on. Its point estimate remains harmful. F adds the full body but also disables the gate, so it is not an evidence-only test. These results do not test tool grounding; long contexts can themselves be difficult to use reliably [11].

Selective prediction and cascades. Reject-option and selective-prediction work studies classifiers that abstain to trade coverage for error [3, 4, 7, 13]. A verifier used as a filter is a related policy, but here its score is uncalibrated. Cascades and routing likewise show that a noisy second stage needs a utility-aware policy [2, 1, 14]. We did not test re-ranking, triage, or using verifier outputs as features.

Decision-analytic evaluation. Cost-sensitive learning [6], iso-performance analysis [12, 16], and cost curves [5] provide the error-cost framing behind Equation 1. Decision curve analysis [18] adds a net-benefit display and explicit comparisons with default strategies. To our knowledge it is not yet used to report LLM-verifier deployment utility. Our contribution is not a new decision-analysis method. It is a production case study showing how the deployment comparison changes the interpretation of a verifier result, together with a reporting proposal.

3 Setup

The pipeline. P converts free-text metadata from the Gene Expression Omnibus (GEO), a public genomics repository, into 18 structured fields. One study may produce several *records*, one per (case, control) contrast. The seven processing stages are detailed in Appendix A. For this evaluation, a keyword match over 14 extraction fields aggregates the record outputs into one study-level positive call.

The verifier and filter. The original sidecar verifier V has four adapters. Each reads one populated field and its evidence, then emits a verdict and categorical confidence (Appendix A). Its verdicts were logged but did not alter P. In the counterfactual filter evaluated here, a positive study is suppressed if any of its records triggers the chosen flagging rule. A missing verifier verdict passes through unchanged.

Sampling frame and analysis units. The sampling frame contained 609 GEO studies retrieved by transplantation-related keyword search. A coauthor with clinical expertise labeled each study for transplantation relevance; we excluded 40 nonbinary cases, leaving 569 studies (82 positive, 487 negative). P called 162 positive: 55 correctly and 107 incorrectly, for 34% precision, 67% recall, and 72.6% balanced accuracy (FN 27, TN 380). The original slot-level runs covered 1,479 records from 562 of the 607 studies for which P produced records and emitted 4,967 verdicts; only populated, applicable fields were judged. Configuration G instead judged all 1,599 records from those 607 studies once per record. All reported outcomes and bootstrap resamples aggregate records back to the 569 labeled studies.

Discrimination metric. Within P's 162 positive calls, we use Youden's J = (flag rate on false-positive calls) $-$ (flag rate on true-positive calls). Thus "flag everything" and "flag nothing" both score zero. J is a downstream-label proxy for verifier discrimination, but it omits prevalence and error costs. Its 95% intervals use the normal approximation for a difference of

two proportions. All runs use vLLM 0.19.1 at temperature 0. The extractor and primary verifier use `google/gemma-4-31B-it`; configuration C uses `Qwen/Qwen3.8-27B` with thinking disabled and a 1,024-token generation limit; the baseline uses a 512-token limit. Because both the model and output limit change, C is not a model-only ablation.

4 The discrimination view: keep fixing the verifier

Throughout this section, "correct" and "incorrect" refer to study-level relevance, not field-level factual support. Adapter-level harm therefore describes downstream filtering effects; it does not establish that a field verdict is factually wrong.

Table 1 lists the baseline and seven changes. The baseline's 95% interval lies below zero, and adding article text does not improve it. In E, both arms keep the escalation gate on, while one arm also receives the article's abstract and methods. Among prompts that neither arm escalates, the added text changes 24.6% of evidence quotes but only 0.3% of verdicts. A rerun with identical prompts changes no verdicts. J moves from -10% to -12%, with nearly overlapping intervals. F adds the full article body (median 45,278 characters) and disables the gate. Because it changes both the evidence and the gate, we cannot attribute its -12% result to added evidence alone. F also flags all 38 correct calls in the 105-study full-text subset. The evidence-only comparison is E (Appendix B). Replacing the verifier model in C is also harmful. In D, the output constraint changes one of the 3,589 covered verdicts; 1,378 escalation re-calls remain unconstrained (Appendix B). None of the eight aggregation rules improves study-level balanced accuracy (Appendix C).

The disease adapter shows the clearest harmful pattern. Only the disease adapter's 95% interval excludes zero: it has $J = -18\%$ $[-26, -9]$. Months earlier, its prompt had been changed to flag procedure names appearing where a disease was expected. The judged field carries the task's defining vocabulary in 63% of correct positive calls. Consistent with this mechanism, the adapter flags 98% of correct calls and 80% of incorrect ones (Appendix D). Removing the definition block in B′ no longer establishes harmful discrimination, but neither does it establish positive discrimination. Because that edit also changes confidence-triggered escalation, B′ is not a clean single-factor causal test.

Rebuilding improved the point estimates, but not causal certainty. Configuration G judges a whole record instead of four fields, asks whether evidence contradicts rather than supports the extraction, and removes the schema definition (Appendix B). It therefore bundles several changes. The false-flag rate among correct calls falls from 96% to 47%, J becomes positive for the first time, and retained-call precision rises from 34% to 39%. Because we compared eight configurations and did not test G's components separately, this analysis neither establishes that G will improve on new data nor identifies which change mattered.

The confidence labels do not provide a useful ranking. At baseline, 98.9% of `ok` verdicts are `high`, but confidence-dependent rules do change the filtered set; none outperforms P. In G, all 771 rejected records are `high`, while 100 of 828 accepted records are `medium`. Confidence therefore has no variation among the records G proposes to suppress, although it is not constant overall (Appendix E).

5 The deployment view: net-benefit bands

Wiring V as a filter produces a new study-level classifier. Under our primary aggregation rule, any flagged record suppresses the study's positive call. Concretely, G removes 26 of P's 55 correct positive calls and 62 of its 107 incorrect positive calls: it catches more errors, but also discards almost half the correct calls.

Let f_{TP} and f_{FP} be the fractions of P's true- and false-positive calls removed. Balanced accuracy improves iff

$$\frac{f_{TP}}{f_{FP}} < \frac{FP_P}{TP_P} \cdot \frac{n_+}{n_-} = \frac{1-\pi}{\pi} \cdot \frac{\rho}{1-\rho}, \tag{1}$$

Table 1: Eight verifier configurations. J compares flag rates among the 55 true-positive and 107 false-positive study calls; intervals are normal-approximation 95% CIs.

Configuration	What changed	J [95% CI]
Baseline	—	-10% $[-19, -2]$
A	escalation disabled	-9% $[-18, -1]$
B′	schema-definition block removed	-3% $[-18, +13]$
C	Qwen3.8-27B; 1,024-token limit	-14% $[-27, -2]$
D	output constrained (72% of verdicts)	-11% $[-20, -3]$
E	paper abstract + methods added	-12% $[-21, -4]$
F	full paper body added, gate disabled	-12% $[-20, -5]$
G	**record-level contradiction check**	$\mathbf{+11\%}$ $[-6, +27]$

where n_+ and n_- are the numbers of positive and negative studies, π is P's precision, and ρ is prevalence. In words, a filter helps only if its rate of discarding correct calls, relative to its rate of catching wrong calls, stays below a ceiling set by the base system and population. For balanced accuracy that ceiling is 0.328; G's observed ratio is 0.816, 2.49 times as large.

Equation 1 is one case of a utility criterion. If V_{FP} is the value of removing a false positive and C_{FN} the cost of removing a true positive, filtering helps iff $f_{TP}/f_{FP} < (V_{FP}/C_{FN})(FP_P/TP_P)$. Balanced accuracy sets $V_{FP}/C_{FN} = \rho/(1-\rho)$. Under a precision objective, by contrast, the ceiling is 1: $J > 0$ is then necessary and sufficient, and G passes (Appendix F). Deployability therefore cannot be assessed without specifying the objective.

Net benefit. Decision curve analysis displays this trade-off across objectives [18]. The action threshold p_t is *not* the verifier's self-reported confidence. It encodes how an operator values a false positive relative to a true positive; larger values penalize false positives more. Net benefit is

$$NB(p_t) \;=\; \frac{TP}{n} \;-\; \frac{FP}{n} \cdot \frac{p_t}{1-p_t}, \tag{2}$$

and is compared with classifying no study as positive (treat-none, $NB = 0$) and every study as positive (treat-all). The odds $p_t/(1-p_t)$ are the false-positive cost relative to the true-positive benefit; their reciprocal is the number of false positives accepted per true positive retained.

Interpreting the band. In these data P dominates treat-all throughout every candidate filter interval, so two crossings determine a filter's band. Against P, the lower crossing is $w = (TP_P - TP_V)/(FP_P - FP_V)$; converted to probability, it is the precision of the calls the filter discards. Against treat-none, the upper crossing $w = TP_V/FP_V$ becomes the precision of the calls it retains. Under this binary-filter wiring and nonzero retained and discarded counts, the band is non-empty exactly when the retained set is cleaner than the discarded set: $f_{TP}/f_{FP} < 1$, or $J > 0$. Band existence therefore repeats the sign of J. Band *location*, however, requires the separate removal rates and base counts; J alone cannot provide it.

B′ shows why a higher J value is not enough: it rises from -10% to -3% but remains negative, and its point-estimate band is empty. The sign of J and band existence therefore give the same answer. For G, the band's location provides the new information. At $p_t = 14.4\%$, the trade-off implied by balanced accuracy accepts roughly six false positives per retained true positive. G is best among the compared strategies only when that tolerance is approximately 1.55–2.38. Equivalently, G's 0.816 removal-rate ratio is 2.49 times the 0.328 ceiling; $J > 0$ would require only a ratio below 1. Even inside its band, G's gain over the better of P and treat-none is at most 0.010, about one net true positive per hundred studies.

Choosing the operating point. Balanced accuracy weights sensitivity and specificity equally despite the 82:487 class imbalance. Its implied $p_t = 82/569$ values one retained true positive as much as $487/82 \simeq 5.94$ false positives; our analysis does not establish that operators endorse this trade-off. Before deployment, an operator can specify an acceptable number k of false positives per retained true positive and use $p_t = 1/(1+k)$, or report a plausible range. Our negative deployment conclusion is conditional on the evaluation objective.

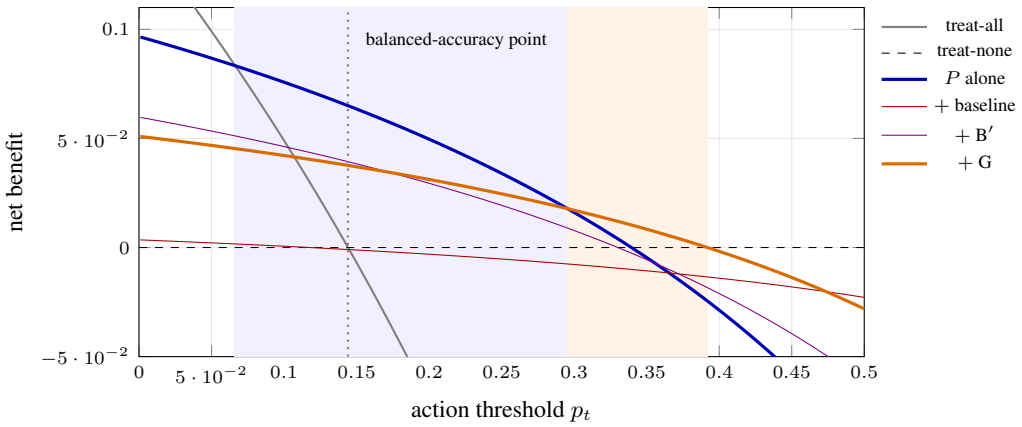


Figure 1: Decision curves for the pipeline with and without the verifier. Among the plotted strategies, P alone has the highest net benefit over $p_t \in [6.6\%, 29.5\%]$; G does so over $[29.5\%, 39.2\%]$; above 39.2%, doing nothing is best. The balanced-accuracy objective maps to $p_t = 14.4\%$, outside G's band. Shading marks the two non-default bands.

Table 2: Point-estimate deployment bands from base predictions, verifier verdicts, and labels. TP/FP are retained study calls. "Boot. empty" is the fraction of 4,000 study-level bootstrap resamples with an empty band, not a posterior probability.

Cond.	TP	FP	f_{TP}/f_{FP}	J	band	boot. empty
Baseline	2	15	1.121	-10.4%	empty	99.3%
A	2	14	1.109	-9.4%	empty	99.0%
B′	34	69	1.075	-2.7%	empty	62.8%
C	8	31	1.203	-14.4%	empty	98.9%
D	2	16	1.133	-11.3%	empty	99.6%
E	2	17	1.146	-12.3%	empty	99.8%
F	1	15	1.142	-12.2%	empty	100.0%
G	29	45	**0.816**	$\mathbf{+10.7\%}$	$[\mathbf{29.5\%}, \mathbf{39.2\%}]$	**10.5%**

Equal discrimination, different choices: a synthetic example. Consider 1,000 studies with base counts $(TP, FP, FN, TN) = (100, 200, 100, 600)$. Verifier V_1 removes $(10, 60)$ true/false positives; V_2 removes $(60, 160)$. Both have $J = +20\%$, but their bands against P, treat-all, and treat-none, evaluated separately, are $[14.3\%, 39.1\%]$ and $[27.3\%, 50.0\%]$. At $p_t = 20\%$, V_1 improves on P while V_2 worsens it; at 45%, only V_2 beats treat-none. If both are available, they cross at 33.3%: V_1 is optimal over $[14.3\%, 33.3\%]$, and V_2 over $[33.3\%, 50.0\%]$. Thus equal positive J values do not determine deployment, and a band must name its comparison strategies. These constructed counts are illustrative, not additional biomedical evidence.

Uncertainty. Band existence inherits J's sampling variation. In 4,000 ordinary study-level bootstrap resamples, G's lower endpoint had a percentile interval of $[20.3\%, 39.5\%]$ and its upper endpoint $[27.9\%, 50.6\%]$; the band was empty in 10.5% of resamples. For B′, the corresponding empty fraction was 62.8%. We use these descriptively rather than as a deployment hypothesis test, consistent with cautions about inferential use of decision curves [19]; Appendix H gives all endpoints and the fixed seed.

The required data were already available. Figure 1 requires no additional LLM calls. The base predictions, verifier verdicts, and deployment labels already existed after the baseline run, yet we tested seven more configurations before calculating net benefit. API, parsing, and schema checks had passed. Five mechanistic hypotheses were later discarded, four using existing data (Appendix G).

6 A fifth disclosure field

The Judge Deployment Disclosure Template proposed by this workshop covers training-data provenance, deployment context, known failure modes, and human validation. These fields describe the judge but do not ask whether using its verdict improves the downstream system. We propose a fifth field:

> **Deployment band.** The interval of action thresholds over which gating on this verifier has the highest net benefit among the compared strategies, using a labeled sample of the deployment population. Report its sampling uncertainty, the base confusion counts, prevalence, aggregation rule, and utility definition, and state whether the intended operating point falls inside the band.

Computing the band requires no new LLM calls once labels and verdicts exist (Appendix I); labeling remains the main cost. Here, crossings use 162 positive calls and prevalence uses all 569 labels. Comparisons must specify the base system, population, utilities, aggregation, and alternative strategies. The baseline's band was empty in the point estimate and 99.3% of resamples.

7 Limitations

This is one biomedical task, one keyword-selected population, and one primary model family. A clinician coauthor supplied the labels before verifier runs; there was no independent second reading or inter-rater estimate. The labels measure study relevance, not field-level support. G bundles three changes and was selected from eight configurations evaluated on the same labeled studies. Its J interval includes zero. The bootstrap holds the configurations fixed and does not correct selection optimism; the 10.5% empty fraction is descriptive, not a probability of population-level failure. Independent held-out or external evaluation and component ablations are needed to establish G's generalization and which change helped.

Our primary wiring is also aggressive: a flag on any record suppresses the whole study, even if another record caused the original positive call. In a post-hoc point-estimate sensitivity check using a more permissive rule in which any unflagged positive record keeps the study, G retains 33 true and 58 false positive calls, giving $J = +5.8\%$ and a band of $[31.0\%, 36.3\%]$; the balanced-accuracy point remains outside.

Human review, regeneration, reranking, and routing to a stronger model require utilities for the follow-up action, including its cost and remaining errors. The net-benefit comparison can be extended to those policies, but the present discarded/retained-precision endpoints and $J > 0$ equivalence do not transfer unchanged. We did not evaluate them or training-time rejection sampling. Our literature claim rests on a targeted, not systematic, search.

Broader impact. Utility-aware reporting can prevent a superficially accurate verifier from silently discarding useful biomedical records. It does not make the underlying labels or utility choices neutral: a narrow or biased deployment sample can yield a misleading band, and an operator can choose costs that hide harms to underrepresented cases. Bands should therefore support, not replace, domain review, subgroup checks, and monitoring after deployment.

8 Conclusion

For the binary filter and reference strategies studied here, $J > 0$ identifies band existence but not its location. G's band excludes the balanced-accuracy operating point, and the synthetic example shows different choices even at equal positive J. Report base counts, prevalence, separate removal rates, the aggregation rule, comparison strategies, band uncertainty, and the intended operating point. These quantities connect verifier discrimination to the deployment decision without additional inference once labels and verdicts exist.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25525042).

References

[1] Aggarwal, P., Madaan, A., et al. AutoMix: Automatically Mixing Language Models. *NeurIPS*, 2024.

[2] Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv:2305.05176*, 2023.

[3] Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[4] Cortes, C., DeSalvo, G., and Mohri, M. Learning with Rejection. *Algorithmic Learning Theory*, LNCS 9925, 2016.

[5] Drummond, C. and Holte, R. C. Explicitly representing expected cost: an alternative to ROC representation. *KDD*, 2000.

[6] Elkan, C. The Foundations of Cost-Sensitive Learning. *IJCAI*, 2001.

[7] Geifman, Y. and El-Yaniv, R. Selective Classification for Deep Neural Networks. *NeurIPS*, 2017.

[8] Gou, Z., Shao, Z., et al. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. *ICLR*, 2024.

[9] Huang, J., Chen, X., et al. Large Language Models Cannot Self-Correct Reasoning Yet. *ICLR*, 2024.

[10] Kadavath, S., Conerly, T., et al. Language Models (Mostly) Know What They Know. *arXiv:2207.05221*, 2022.

[11] Liu, N. F., Lin, K., Hewitt, J., et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[12] Metz, C. E. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298, 1978.

[13] Mozannar, H. and Sontag, D. Consistent Estimators for Learning to Defer to an Expert. *ICML*, 2020.

[14] Ong, I., Almahairi, A., et al. RouteLLM: Learning to Route LLMs with Preference Data. *arXiv:2406.18665*, 2024.

[15] Panickssery, A., Bowman, S. R., and Feng, S. LLM Evaluators Recognize and Favor Their Own Generations. *NeurIPS*, 2024.

[16] Provost, F. and Fawcett, T. Robust Classification for Imprecise Environments. *Machine Learning*, 42(3):203–231, 2001.

[17] Stechly, K., Valmeekam, K., and Kambhampati, S. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. *arXiv:2402.08115*, 2024.

[18] Vickers, A. J. and Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, 26(6):565–574, 2006.

[19] Vickers, A. J., Van Calster, B., Wynants, L., and Steyerberg, E. W. Decision Curve Analysis: Confidence Intervals and Hypothesis Testing for Net Benefit. *Diagnostic and Prognostic Research*, 7:11, 2023.

[20] Xiong, M., Hu, Z., et al. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation. *ICLR*, 2024.

[21] Zheng, L., Chiang, W.-L., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.

A Pipeline and verifier architecture

Stages. P runs seven stages. (1) Group inference clusters samples within a study by normalizing sample titles and characteristics. (2) Phrase compression extracts common substrings across each group's sample titles. (3) Paper retrieval resolves the study's linked PubMed identifiers against a local corpus of full-text XML. (4) Pair assembly identifies one control group and emits one record per (case, control) pair; the pipeline output unit therefore changes from a study to a contrast record. (5) Rule-based prefilling populates slots reachable by static keyword lookup and regular expressions. (6) LLM extraction emits 18 slots as strict JSON. (7) Post-validation merges stages 4 and 6 and attaches exception flags. Evaluation aggregates the records back to the study level.

Adapters. V comprises four adapters keyed to slots: disease, tissue, cell type, and organism. Each receives the slot value plus an evidence block assembled from the study's title, summary, design, and sample characteristics, and returns $\{\text{verdict}, \text{evidence quote}, \text{confidence}\}$. Verdict vocabularies differ by adapter: disease emits `ok/partly_hallucinated/hallucinated`; tissue and cell type emit `ok/mismatch/unsupported`; and organism emits `ok/inconsistent`. A gate re-invokes an adapter with expanded evidence when confidence is `low` or the verdict is high-stakes; configuration A disables this gate.

Evidence coverage. The evidence block is assembled from repository metadata and is available for every study (median 1,626 characters of series text plus 1,070 of sample characteristics). Linked full text is available for only part of the labeled set, which is why conditions E and F are also reported on the 105 studies that P called positive and that received full text.

Model serving. The extractor, baseline verifier, and G use `google/gemma-4-31B-it`; C uses `Qwen/Qwen3.8-27B`. Both were served with vLLM 0.19.1 in bfloat16 at temperature 0. The default request limit was 512 generated tokens. For C, thinking was disabled and the limit was 1,024 tokens. The Gemma server context was 65,536 tokens and the Qwen context was 32,768 tokens.

Compute resources. All inference ran on an internal server with NVIDIA H100 80-GB HBM3 accelerators, an Intel Xeon Gold 6448Y host (124 logical CPUs), and about 1.77 TiB of RAM. The frozen seven-stage pipeline took 15:30 on two GPUs (0.52 H100-hours). The measured verifier wall times were 26:38 for baseline (2 GPUs; 0.89 H100-hours), 7:20 for A (2; 0.24), 12:42 for B′ (2; 0.42), 18:17 for C (1; 0.30), 27:15 for D (2; 0.91), and 23:09 for G (2; 0.77). Server activity gave approximate times of 14:00 for E (2; 0.47) and 60:00 for F (2; 2.00). Gemma used either two one-GPU replicas or one two-GPU tensor-parallel server; C used one Qwen GPU. The reported runs therefore used approximately 6.5 H100-hours. E and F are estimates from server activity because their end-to-end start times were not retained. Recorded preliminary, debugging, and one invalid full-body run add at least 2.0 H100-hours; not every exploratory run was metered.

Existing assets and terms. We used public records from the NCBI Gene Expression Omnibus (GEO) under the NCBI GEO Disclaimer (`https://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html`), which places no NCBI restriction on use or distribution while noting that submitter-held rights may still apply. Linked PMC full text was retrieved programmatically through an NCBI API, cached locally, and parsed as JATS XML (`https://pmc.ncbi.nlm.nih.gov/tools/textmining/`). PMC rights are article-specific; the 280-paper corpus included CC BY, CC BY-NC, CC BY-NC-ND, CC BY-NC-SA, CC BY-ND, and PMC text-and-data-mining or custom terms. We used article text only for computational analysis and do not redistribute it. Google DeepMind's `google/gemma-4-31B-it`, the Qwen Team's `Qwen/Qwen3.8-27B`, and the vLLM Project's vLLM 0.19.1 are distributed under the Apache License 2.0. We redistribute neither model weights nor third-party software. Official sources and terms are the Gemma model card (`https://ai.google.dev/gemma/docs/core/model_card_4`), Qwen checkpoint page (`https://huggingface.co/Qwen/Qwen3.8-27B`), and vLLM license (`https://github.com/vllm-project/vllm/blob/main/LICENSE`).

Ethical scope and annotation. The study used public, unrestricted-access GEO series and sample metadata, together with linked publications accessed under article-specific terms. It did not use controlled-access records or raw participant-level molecular measurements, and the research team

did not attempt to identify individuals. No participants were recruited, contacted, or intervened upon, and no prospective data were collected. A clinician coauthor, acting as a member of the research team rather than as a recruited annotator or crowd worker, labeled study-level relevance. Accordingly, the participant-facing and IRB items in the checklist were not applicable. All authors reviewed the NeurIPS Code of Ethics and confirm that the work conforms to it.

Labeling rule and exclusions. For each study, the annotator judged whether it concerned transplantation in humans. The review used the GEO title, organism, assay type, sample count, matched keyword and its location, and, when available, the linked publication. Forty of the 609 studies received no binary label and are excluded from every number in the paper: 21 whose source publication could not be located or offered only an abstract, 9 xenograft studies in which human cells were engrafted into an animal host, 6 *ex vivo* machine-perfusion studies of organs before transplantation, and 4 remaining borderline cases (deceased-donor tissue without a transplant contrast, fecal microbiota transfer, autologous fat grafting, and normal islet profiling). The perfusion and xenograft rules are the consequential ones: both describe procedures adjacent to transplantation, and admitting either would change the denominator and the prevalence on which the deployment analysis rests.

B The eight configurations

Table 3: What each configuration changed and what evidence the verifier saw.

Cond.	Changed	Evidence	J
Baseline	—	repository metadata, gate on	-10%
A	gate disabled	repository metadata	-9%
B′	schema definition removed	repository metadata	-3%
C	verifier → `Qwen3.8-27B`; 1,024-token limit	repository metadata	-14%
D	output constrained for 72% of verdicts	repository metadata	-11%
E	paper abstract + methods appended	$+\,9,221$ chars (median)	-12%
F	full paper body appended, gate disabled	$+\,45,278$ chars (median)	-12%
G	record-level, contradiction-seeking, no definition	repository metadata	$+11\%$

Conditions E and F on the subset with full text. Among the 105 studies that P called positive and that received full text, the baseline has $J = -9.3\%$ and F has $J = -11.9\%$; F flags all 38 correct calls. Thus the overall null result is not explained by studies that lacked a linked paper. The direction remains harmful in the subset, although it is not more negative than F's result on all 162 positive calls.

F mixes two changes and cannot be decomposed. F appends the full paper body *and* disables the escalation gate. We cannot separate them. A gate-on full-body run exists but is invalid: before the truncation limit was consolidated in one place, each adapter capped its evidence block at 4,000 characters. That run supplied about as much text as E rather than the full body, and we therefore do not use it to estimate F's effect. In the valid evidence-only comparison E, among prompts that neither arm escalated, verdicts differ from baseline on 0.3% and evidence quotes on 24.6%. The nominally deterministic control A, which leaves the evidence unchanged, still differs on one of 3,593 verdicts and 33 evidence quotes; temperature 0 therefore did not eliminate run-to-run variation. E's observed differences are larger than this control, but its J interval still overlaps the baseline's. We treat F as a single mixed condition and place the evidence-only comparison in E, where the gate is on in both arms.

The format constraint reaches 72% of verdicts. The guided decoder selects a per-adapter response schema by prompt hash. Escalation re-calls carry augmented evidence, so their prompts hash differently and are never registered; the implementation then sends them without a schema rather than forcing the schema of the wrong adapter. In D this affects 1,378 of 4,967 verdicts (27.7%), and the run log records 397 unconstrained calls against 2,200 constrained. The single verdict that changed lies in the constrained 3,589, and no escalation-derived verdict changed.

Configuration G in detail. G replaces the four slot adapters with one record-level verifier. Its prompt shows the record's populated slots and the repository evidence, and asks whether the ev-

idence *contradicts* the extraction, with three explicit instructions: do not reject merely because a term is absent verbatim; accept when the evidence neither confirms nor denies; reject only when the evidence asserts the opposite. No task vocabulary is injected through the fixed instructions or schema; task terms may still occur in the dynamic record and evidence. An earlier revision of the disease adapter had included transplantation terms as instruction examples; we do not repeat that task-specific leakage.

C Threshold sweep

Table 4: Eight ways to combine the baseline verifier's flags. The evaluation unit is a study. No rule improves balanced accuracy; the only tie, "all four," suppresses nothing. Differences are percentage points.

Rule	Balanced accuracy	vs. P alone
(no filter)	72.6%	—
any adapter flags	49.7%	-22.9
disease adapter only	49.7%	-22.9
≥ 2 adapters flag	62.7%	-9.8
≥ 3 adapters flag	71.9%	-0.6
all 4 adapters flag	72.6%	$+0.0$
strong verdicts only	51.4%	-21.2
strong + confidence high	67.0%	-5.6
low-confidence flags only	51.9%	-20.7

D Per-adapter analysis

Table 5: Adapter-level discrimination at baseline among studies for which that adapter returned an applicable verdict. Only the disease adapter's interval excludes zero. "Task vocabulary" is the share of correct calls whose judged slot contains transplantation-defining terms.

Adapter	flags correct	flags incorrect	J [95% CI]	task vocab. in slot
disease	52/53 (98.1%)	86/107 (80.4%)	-17.7% $[-26.1, -9.4]$	**63%**
tissue	13/50 (26.0%)	27/91 (29.7%)	$+3.7\%$ $[-11.7, +19.0]$	0%
cell type	10/36 (27.8%)	13/90 (14.4%)	-13.3% $[-29.7, +3.0]$	0%
organism	1/53 (1.9%)	2/107 (1.9%)	0.0% $[-4.5, +4.5]$	—

The disease adapter has little room to discriminate: it flags 98% of correct calls and 80% of incorrect calls. Its most frequent objections on correct calls are the terms that identify the task, including graft-versus-host disease, antibody-mediated rejection, and the procedure name itself. The other adapters' intervals include zero and should not be read as stable design lessons.

We also tested whether "unsupported" merely meant "not copied verbatim." Of components the disease adapter called unsupported, 50.8% were absent verbatim from the repository metadata; the corresponding rate across all extracted disease names was 49.8%. The nearly equal rates do not support verbatim absence as the mechanism.

E Confidence

The `low` signal changes sharply when the schema-definition block is removed: B′ reduces $\Pr(\text{low} \mid \text{suspicious})$ from 76.7% to 0.7%. In G, 1,499 of 1,599 records are `high` and 100 are `medium`; all 771 rejected records are `high`. Confidence is therefore not constant overall, but it supplies no ranking signal within the set G proposes to suppress. This experiment does not establish that every confidence-elicitation scheme behaves this way.

Table 6: Self-reported confidence against verdict. In G, `ok` means accepted and `suspicious` means rejected.

Condition	Pr(high | ok)	Pr(low | suspicious)
Baseline	98.9%	76.7%
B′ (definition removed)	98.9%	**0.7%**
C (verifier swapped)	99.5%	6.2%
F (full paper)	99.8%	52.8%
G (rebuilt)	87.9%	0.0%

F Objective dependence of the threshold

Writing V_{FP} for the value of removing a false positive and C_{FN} for the cost of removing a true positive, a filter improves expected utility iff $f_{TP}/f_{FP} < (V_{FP}/C_{FN}) \cdot (FP_P/TP_P)$. Balanced accuracy is the case $V_{FP}/C_{FN} = \rho/(1-\rho)$.

Table 7: The ceiling depends on the objective. "Passes" means improvement over P alone, not superiority to every default strategy. Configuration G has $f_{TP}/f_{FP} = 0.816$. The F_1 row is a local comparison at the observed base confusion counts, not a universal F_1 threshold.

Objective	V_{FP}/C_{FN}	Ceiling	G
Balanced accuracy	0.168	0.328	fails
F_1 (local)	0.291	0.566	fails
Precision	—	1.000	**passes**
Equal per-item value/cost	1.000	1.945	passes

Under a precision objective the condition reduces to $f_{TP} < f_{FP}$, exactly $J > 0$; G raises precision from 34.0% to 39.2%. Thus the claim that $J > 0$ is insufficient applies to objectives whose ceiling is below one, including balanced accuracy here, but not to precision. A fixed "required J" is likewise misleading because the required difference depends on the observed removal rate.

Equal per-item value/cost corresponds to $p_t = 0.5$. G improves on P at this threshold, but its net benefit is $(29 - 45)/569 = -0.028$, below treat-none's zero. The pairwise improvement in Table 7 therefore does not make G the preferred deployment strategy. Precision alone also omits the cost of lost coverage.

G Hypotheses we discarded

We include these hypotheses because four were testable with data already available and their sequence is central to the postmortem.

Table 8: Discarded hypotheses about the verifier's failure.

Hypothesis	Why discarded
The verifier lacks evidence	E and F do not repair it; F flags every correct subset call
Unsupported means absent verbatim	absence is 50.8% in flagged vs. 49.8% overall
Complex values attract flags	correct calls average 1.04 components, incorrect 1.23
Entanglement predicts harm	entanglement does not correlate in rank with J
Definition removal repairs it	B′ has an empty point-estimate band

We initially proposed the fourth hypothesis as a contribution, so we report its test in more detail. We hypothesized that the verifier was biased by the task's defining vocabulary and that this bias could be detected without labels by conditioning the flag rate on whether the judged value contains that vocabulary. The test detects an association: the disease adapter flags 100% of vocabulary-bearing values and 69% of other values ($\chi^2 = 159.7$, $p < 0.001$).

The test does not predict harm. Ranked by this index, the cell-type adapter is the most entangled and only moderately harmful, while the tissue adapter is highly entangled with positive point-estimate J. Although the index detects vocabulary-associated flagging, it does not rank adapters by downstream harm. Statistical significance did not translate into predictive validity, so we withdraw the proposal.

H Sampling uncertainty

Table 9: Crossing points and ordinary-bootstrap variability. The first crossing is the precision of discarded calls and the second is the precision of retained calls. A point-estimate band is non-empty only when the first is below the second. Intervals are 2.5th–97.5th bootstrap percentiles.

Cond.	discarded crossing [95% interval]	retained crossing [95% interval]	empty resamples
Baseline	36.6% [29.0,44.3]	11.8% [0.0,30.0]	99.3%
A	36.3% [28.8,44.0]	12.5% [0.0,31.2]	99.0%
B′	35.6% [24.0,48.1]	33.0% [24.3,42.0]	62.8%
C	38.2% [30.0,46.6]	20.5% [8.6,34.1]	98.9%
D	36.8% [29.1,44.5]	11.1% [0.0,27.8]	99.6%
E	37.1% [29.3,44.8]	10.5% [0.0,26.7]	99.8%
F	37.0% [29.3,44.6]	6.2% [0.0,20.0]	100.0%
G	29.5% [20.3,39.5]	39.2% [27.9,50.6]	10.5%

We resampled the 569 labeled studies with replacement 4,000 times and recomputed every strategy's confusion counts and crossings. We used the Python standard library's `random.Random` with seed 20260826 and shared one study resample across all strategies within each replicate. For each endpoint, we sorted the 4,000 values and selected the value at zero-based index $round(q(B-1))$, where $B = 4,000$ and $q \in \{0.025, 0.975\}$; this is a noninterpolating nearest-index percentile. No replicate had an undefined crossing denominator. We counted a replicate as empty when its discarded-set crossing was not below its retained-set crossing. These intervals describe sampling variation and are not confidence intervals for the endpoints of a conditionally selected, necessarily non-empty band [19].

I Reproduction

Frozen inputs for the resampling table. Table 10 represents the 162 positive calls as 25 distinct patterns of clinician label and eight verifier flags; n is the multiplicity. Complete the 569-study sampling population by adding 27 positive-labeled and 380 negative-labeled studies that P did not call positive. They contribute zero to every strategy's retained and discarded positive-call counts, but must remain in the ordinary bootstrap sample.

Table 10: Frozen study patterns for reproducing the deployment analysis. Flags are one when a verifier suppresses a call; Base denotes the baseline verifier, not the pipeline prediction.

Key	Label	Base	A	B′	C	D	E	F	G	n
A	0	1	1	0	1	1	1	1	1	22
B	0	1	1	1	1	1	1	1	1	21
C	1	1	1	0	1	1	1	1	0	15
D	0	1	1	0	1	1	1	1	0	14
E	1	1	1	1	1	1	1	1	1	13
F	0	1	1	1	1	1	1	1	0	10
G	1	1	1	0	1	1	1	1	1	10
H	0	1	1	0	0	1	1	1	1	9
I	1	1	1	1	1	1	1	1	0	8
J	0	1	1	0	0	1	1	1	0	7
K	0	1	1	1	0	1	1	1	0	5
L	0	0	0	0	0	0	0	0	1	4
M	0	0	0	0	1	0	0	0	1	4
N	1	1	1	0	0	1	1	1	0	4
O	0	0	0	0	1	0	0	0	0	3
P	1	1	1	0	0	1	1	1	1	3
Q	0	0	0	0	0	0	0	0	0	2
R	0	0	0	0	1	0	0	1	0	1
S	0	0	1	1	1	0	0	1	1	1
T	0	1	1	0	0	1	0	1	0	1
U	0	1	1	0	0	1	1	0	0	1
V	0	1	1	0	0	1	1	0	1	1
W	0	1	1	1	0	0	0	1	0	1
X	1	0	0	0	0	0	0	1	0	1
Y	1	0	0	0	1	0	0	0	0	1

Exact study order. The sequence below encodes the original bootstrap input in numeric GEO accession order. Uppercase letters refer to Table 10; `p` and `n` denote the 27 positive and 380 negative

studies that P did not call positive. Concatenate the lines without whitespace. Sampling 569 indices per replicate with `random.Random(20260826).randrange(569)` and applying the percentile rule above reproduces every displayed entry in Table 9. The order matters for reproducing a fixed random seed, although it does not change the underlying resampling distribution.

```
nnnnnIBnnnnnnnnnpnnnnBnnAnnDFBpnnnnnHnnnKnnnnnnCnnnMpBnnnpXIEDnA
CnnTnnnnnnnpBnnKGBnnnnnnnnCnnnnCCnpAFnLnnnInnnnnnnnnBUnnnnAApnnn
npnAEECnnQnnNnpnnnnnnnnnnnCnnAppnnQGnEYnnnnnnnnOAnnnEnGInEnnnnnnnB
DnInnnCGnAnnnnInVnnEnDnnFnnnnEnnnnnKSnnnppnnnHnnnGnnnnHnnnnnBDBn
DnnJnnnInnMnnnnEnAnpnFPEnFFnnnnnnnnpnnnnnnnnnnBpnnDnApHAnnnnnPnnnF
GnpnpnnnAnnnNnnBnpnnnpMnnJnnnnnFCnnAnnnnnnnnnJnDWnMGnnCnFNGnnEInn
nnBnnnnnLBnnAnnnGCnnEnnCBHEHPnnnnnKnGpnpBnnnnnnnAAnAnnnnnnnnnnnn
nJHnnnCnJnnDpnCnOnNnnnnnnnAnnnOnnBBnnnnnnAnnnnDDnnnnnnnnnnnnnAnp
pnLJnnnnnnnnDnnnnnBnnnnLKnnRnJnnnDnAHHnnnnnnCBnnBDnnFnnnp
```

The deployment calculations use frozen verifier verdicts, deployment labels, and base predictions; they do not require new model inference. On $n = 569$ studies, the strategy counts (TP, FP) are treat-all $(82, 487)$, P alone $(55, 107)$, baseline $(2, 15)$, B$'$ $(34, 69)$, and G $(29, 45)$. The primary rule suppresses a positive study if any of its records is flagged. A missing verifier verdict passes through. Under the sensitivity rule that any unflagged positive record keeps the study, G instead yields $(33, 58)$.

Band endpoints follow from two crossings per configuration: against P at $w = (TP_P - TP_V)/(FP_P - FP_V)$ and against treat-none at $w = TP_V/FP_V$, converted by $p_t = w/(1 + w)$. Given study identifiers, record-level base calls, verifier flags, and labels, Figure 1 and Tables 2, 7, and 9 can be reproduced directly from these counts and formulas.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract, Introduction, and Limitations distinguish point estimates from uncertainty, scope the result to a binary inference-time filter, and avoid claiming that decision-curve mathematics is novel.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the single task and annotator, configuration selection, bundled changes, label mismatch, aggregation choice, uncertainty, and external-validity limits.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper applies established decision-curve analysis and gives the algebra needed to interpret its crossings, but it does not claim a new theorem or other theoretical result that requires a formal proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper's main empirical conclusion concerns deployment utility. Appendix I provides the frozen pattern table and exact encoded study order, base counts, aggregation rules, crossing formulas, and exact resampling procedure needed to recompute every J value, deployment band, and bootstrap summary supporting that conclusion without new model inference. Secondary mechanistic diagnostics are reported as analyses of frozen observations, not as independently rerunnable LLM experiments. The implementation source code is not released.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Appendix I provides paper-embedded sufficient statistics for auditing the main deployment calculations, but the implementation source code and full record-level verification package are not publicly released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: This is an inference-only evaluation with no training, optimizer, or train/test split. Section 3 and Appendix A specify the sampling and exclusion criteria, labels, analysis units, aggregation and missing-verdict rules, exact model identifiers, serving and decoding settings, and all eight configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 1 reports normal-approximation 95% intervals for differences of proportions. Section 5 and Appendix H report a 4,000-resample ordinary study-level bootstrap, the seed, the noninterpolating nearest-index percentile rule, and the descriptive interpretation of the resulting ranges.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A reports the accelerator and host memory, serving topology, measured or explicitly identified approximate wall time and accelerator-hours for every reported run, their total, and a lower bound for preliminary, debugging, and invalid runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: All authors reviewed the NeurIPS Code of Ethics and confirm that the research conforms to it. The study uses public, unrestricted-access GEO records together with linked publications accessed under article-specific terms, reports aggregate results, uses no controlled-access records, and involved no participant recruitment or intervention.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader impact paragraph in Section 7 discusses the benefit of preventing silent data loss and the risks of biased labels, utility choices, and underrepresented cases, together with monitoring and review mitigations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases neither a new model nor a scraped dataset with a high risk of misuse; it analyzes an inference-time verifier used on public biomedical metadata.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix A credits GEO, PMC, both model checkpoints, and vLLM; it states their official sources, versions, licenses or article-level terms, programmatic NCBI API access, non-redistribution, and how the study respected those conditions.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission releases no standalone dataset, model, or software asset. Appendix I contains paper-embedded sufficient statistics for auditing the reported deployment calculations rather than a separate reusable asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or participant study was conducted. Relevance labels were produced by a clinician who is a coauthor and a member of the research team, not by recruited participants or crowd workers; participant instructions, screenshots, and compensation are therefore not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study analyzed public GEO series and sample metadata and linked publications. It used no controlled-access records and involved no participant recruitment, contact, intervention, or prospective data collection; the research team did not attempt re-identification. The IRB item was therefore not applicable to this secondary analysis.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are core components of both the extractor and verifier. Sections 3 and Appendix A describe their roles, exact model identifiers, serving and decoding settings, evidence inputs, and structured outputs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.