Describe the Instrument, Declare the Plan: Preregistration for Research That Uses Generative AI

Theodore O. Cochran
AI for Altruism
theo@ai4altruism.org

AI for Altruism Position Paper A4A-PP-2026-002 Version 1, 2026-10-16
Status: Not peer reviewed.
Reviewer: Anonymous License: CC BY 4.0
https://reports.ai4altruism.org/pp-2026-002/

Abstract

Publishers now ask researchers to describe how artificial intelligence was used to conduct their research, usually in the methods section. These requirements address what was used but not when its consequential configuration was fixed, so a description written after the results are known can satisfy them even when that configuration was chosen after the results were seen. The problem grows when a model makes design choices the researcher did not specify, as can happen when an AI assistant writes analysis code. This paper draws on two preregistered evaluations by the author. In both, the registered plan changed what could be reported, and in both, assumptions introduced by an AI coding assistant were found only by later checks. For confirmatory evaluations, it recommends that researchers prospectively specify the configuration of model-based instruments and the material implementation and analysis choices, including assumptions embedded in code, before observing the resulting measurements on evaluation data. It also recommends that journals ask authors to disclose the timing of those choices and any material changes made after observation.

Keywords. preregistration, LLM-as-judge, research transparency, AI disclosure policy, AI-assisted research, evaluation methodology

1 Introduction

This paper is written for the research community, and in particular for the growing number of researchers who use generative AI in more aspects of their work. My recognition of the problem it describes came with increasing exposure to generative AI across different aspects of research. As generative AI evolved from "simple" language models into orchestrators of agents, it became apparent that these systems could do a lot more, a lot faster, than humans were accustomed to.

In May 2026, ACM updated its Policy on Authorship [1] and drew a distinction between artificial intelligence used to assist with writing and artificial intelligence used to conduct research. Writing assistance no longer requires disclosure. Research use must have its specific uses "described in detail in the methods section of the Work." The policy's enumeration of research uses is broad. Among other things, it covers the design and methodology of the research, the generation and collection of data, coding, data analysis, testing, validating results, and the creation of artifacts directly relevant to the conclusions, such as code, datasets and figures. Other publishers draw similar lines.

The ICMJE recommendations ask authors who used AI to conduct a study to describe that use in the methods section in enough detail to replicate it, including the tool, version and prompts where applicable [2]. Elsevier requires research use to be described in a reproducible manner, with the model and version [3], and Wiley asks for the technology's name and version and the date of use [4].

That distinction appears to be the right one. Policing whether prose was drafted with a model was not likely to work, and it was not the thing that mattered most. Where a model touches the evidence and shapes the methodology, however, disclosure matters a great deal.

None of the policies surveyed for this paper asks when the consequential configuration governing model-based measurement was fixed relative to observing the resulting measurements. A description written after the results are known can therefore satisfy all of these reporting requirements, even when that configuration was chosen after the results were seen. This paper argues that the timing should be disclosed and that timestamped preregistration provides a direct way to make such a statement checkable. The argument rests on two preregistered evaluations I have run [5, 6]. The registered plan changed what could be reported in both. An AI coding assistant also introduced assumptions in both that I had not specified, and those were found only by later checks.

The paper makes two contributions: evidence from practice that prospective commitments can change what is reported, and an analysis of the gap between retrospective disclosure of AI use and an auditable record of when consequential methodological choices were made.

The rest of this paper is organized as follows. Section 2 describes the gap in current disclosure requirements, and Section 3 explains why model-based instruments make that gap more consequential. Section 4 presents the two studies. Section 5 discusses related work, Section 6 sets out the scope and limitations of the argument, and Section 7 presents recommendations.

2 Disclosure Requirements

Part of the problem is that a model can make experimental design choices on its own, particularly when the human researcher has left them under-specified. Part of the problem is the timing of those choices, and in particular whether they were made and executed before or after the results were seen.

By a *model-based instrument* I mean a model whose configured behavior produces scores, labels, or other observations on which an evidential claim depends. The clearest cases are LLM-as-judge scorers, model graders applying a rubric, and automated model-based annotators. An AI assistant that writes analysis code is not an instrument in this sense, but it can embed choices in that code that shape what an instrument measures or how its results are tested.

Three categories of choice therefore need to be distinguished. *Instrument configuration* includes the model and version, prompt, rubric, and procedures for combining measurements. *Implementation choices* determine what evidence reaches the instrument, including passage selection, truncation, and preprocessing. *Analysis choices* determine how measurements support claims, including reliability gates, outcome selection, and statistical tests. In this paper, *consequential configuration* is shorthand for instrument configuration together with the implementation and analysis choices that materially determine the instrument's inputs, measurements, or interpretation. This usage does not make an AI coding assistant or a downstream statistical test part of the model-based instrument.

For evaluations of this kind, the methods section reaches its published form after the results are known. An author may have drafted it much earlier, and a preregistered analysis plan is precisely such a draft, but the version a reader sees is the one that survived to publication. Unless it explicitly

reports that chronology or links to a prospective record, it need not reveal which choices were fixed in advance and which were settled later. It can describe the instrument in complete and accurate detail and still not establish whether that instrument was specified before anyone observed the resulting measurements. In other words, disclosure without the "when" can be met even after the results were seen.

Consider an ordinary LLM-as-judge evaluation. The analyst chooses a judge model and version, a scoring prompt, and a rubric with a granularity for it, which might be holistic quality on a 1 to 5 scale or claim-by-claim citation checking. The analyst also decides whether to use one judge or two, how to aggregate across judges, what inter-rater reliability threshold to require and what happens if it is not met, and which of several criteria becomes the headline number. When an AI assistant writes the scoring and analysis code, some of these choices may be made by the assistant rather than the analyst. Others may be added that the analyst never considered, such as how much of each passage the judge is shown or which statistical test is applied.

There is already reason to treat these choices as consequential rather than clerical. LLM-as-judge studies have documented systematic evaluator effects, including position, verbosity and self-enhancement biases [7]. More recent work treats the rubric itself as a measurement specification, in which changing the rubric changes the measurement produced by an otherwise fixed judge [8]. Baumann et al. (2025) found that "with just a handful of prompt paraphrases, virtually anything can be presented as statistically significant" in research that uses LLMs to annotate text [9]. Once several defensible configurations can produce materially different measurements, the timing of their selection becomes part of the evidential record.

Several alternatives among these choices may be defensible on their own. They can be selected after seeing the scores and then described, accurately and in detail, in a methods section that satisfies the requirements above. It should be noted that this is not an accusation of bad faith. An analyst who re-runs a judge with a sharper rubric because the first rubric was not measuring the right thing is doing something that feels like craft rather than cheating, and may be right that the second rubric is better. This is the pattern Gelman and Loken (2013) named the garden of forking paths [10], and it does not require anyone to notice they are in it. The choices need only exist and be described after the fact.

3 Model-Based Instruments

This matters more when the instrument is an AI model, because we are in a period where models provide easy access to very powerful tools that users may be unfamiliar with or not fully proficient with. In the past, human analysis and the human use of laboratory instruments generally required more familiarity and proficiency. For example, I do not need to learn the details of a model instrument to use it and get results, whereas I did with more traditional instruments and methods that did not involve AI. The concern applies even when the researcher is an expert, however, because the evidential question is when a choice was fixed relative to observing its consequences.

Three properties of model-based instruments add to this concern. The first is that re-applying the instrument is inexpensive. Many physical, observational and human-subject measurements are costly or impossible to repeat once the data have been collected. An LLM judge can often be applied again to the same answers being evaluated, with a different judge model, prompt or rubric, quickly and at low marginal cost. The practical barrier to trying several defensible measurement specifications is therefore unusually low.

The second is that some model-based instruments are non-stationary. A hosted judge model may not point to a permanently fixed thing. Providers deprecate versions, update endpoints and

ship behavioral changes, so the same model name can describe an instrument whose behavior has changed over time. For example, the comparison study described in Section 4 recorded only the alias of its scoring model, so whether a later rescoring used the same model snapshot cannot be established. This concern is smaller for pinned open-weight models run under a fixed inference stack, but it is substantial for hosted evaluators. Preregistration does not prevent drift; however, it records which instrument was intended, which separates a stated limitation from a claim no one can check.

The third is that judge measurements can appear plausible even when the measurement procedure is flawed. A model may return a well-formed score under several reasonable configurations, without an obvious failure signal to warn the analyst that a run requires scrutiny. Many paths through the available choices therefore produce something that looks publishable.

In a preregistered comparison of two retrieval architectures [6], the same 13 questions over 24 papers were put to both systems, and every answer was scored by two blinded LLM judges from different model families. The purpose of using two judges was to test whether the scoring was reproducible at all, rather than to average away disagreement. Rank agreement between the judges depended heavily on how concretely the criterion was defined, from $\rho = 0.81$ on the most concretely specified criterion to $\rho = 0.04$, near zero, on a holistic judgment of groundedness. It should be noted that thirteen questions is a small sample, and those values should not be read as estimates of judge agreement in general. What they establish is narrower: agreement varied substantially by criterion within a single study. A methods section describing both criteria in detail would have been entirely truthful and would not have revealed when either criterion, or the decision to privilege one of them, had been fixed.

4 Two Preregistered Studies

I have now run two preregistered LLM evaluations end to end [5, 6], with analysis plans registered on OSF before confirmatory scoring and study materials deposited there. Pilot work preceded the plan for the Progressive-Disclosure study, and its registration contains a draft rather than the final locked plan, as discussed below. The registration for the comparison study [6] carries a timestamp of 6 May 2026, and every judge scoring run in that study began after it. The wiki-side answer-generation run did not: it started 77 minutes before the registration and finished 13 minutes after it. The original paper reports this as a deviation and acknowledges that incidental exposure to outcome information before registration cannot be excluded [6]. The registrations constrained later decisions that would otherwise have remained open after the scores were visible. The cases also include problems, some of my own making, revealed through later checks of the implementation and the prospective record. Registration alone did not detect those problems.

The reliability gate

In the Progressive-Disclosure ablation [5], a human rater graded a stratified audit sample alongside the model judge. The registered plan required their agreement to reach a quadratic-weighted kappa of 0.60 before the judge's scores would be treated as an adequate measurement. The pooled result came in at 0.23. The numerical cutoff is not offered here as a standard. Its significance is that it was chosen when it could not yet be moved to accommodate the result.

After seeing 0.23, I might have been tempted to lower the threshold or to look for a different agreement metric. Either change would have departed from the registered confirmatory plan and would have required explicit reporting as a deviation or exploratory analysis. Had I abandoned the gate, the disclosure requirements described above would not have obliged the methods section

to mention the earlier target. Holding to it narrowed the claims the paper could make, and that narrowing proved helpful in refining my technique. The threshold was not the whole of the registered rule, however. The plan also required a rubric revision and a full re-grade when the gate failed, and I did not carry them out. The paper reports that as a deviation from the registered procedure and treats the failed gate as a material limitation [5]. The first public version of the paper reported the failed gate in its results but not in its abstract, and the revision states it in both.

The registration

The same study also shows how a registration can go wrong. Its OSF registration holds a working draft of the plan, still marked DRAFT and specifying 1,440 runs, because that copy was uploaded forty minutes after the final plan had been locked in version control. The paper reports the locked plan's 960 runs, and it states, incorrectly, that the registration holds the locked plan. The error arose from my unfamiliarity with the registration process and insufficient time to check the uploaded record against the locked plan. The paper's account of its registration has since been corrected, in its second arXiv version, to distinguish the registered draft from the final plan in version control [5]. The version-control history records when the plan was locked, although, as Section 7 notes, version control alone does not independently corroborate that timing.

The second judge

In the comparison study [6], adding a second judge to a supplementary analysis materially changed the interpretation of a result that had already been written up. On holistic groundedness, GPT-5.4 scored a decomposition-retrieval variant 1.15 points above the compiled wiki, while Gemini 2.5 Pro scored the same contrast 0.85 points in the other direction. With the registered adjustment for inter-rater disagreement applied to both judges' scores, the difference fell from +1.15 to +0.15, a small observed difference. The supplementary comparison was exploratory and was not itself preregistered. The adjustment rule, however, had been declared for the original confirmatory study before either judge's measurements were observed. Applying it prevented me from choosing an aggregation method according to which judge's result was more favorable. The shift illustrates substantial differences between these two judges in their measurements of the same contrast. It does not establish a general difference between model families.

The human check and the code

Reviewers of the comparison study asked for human validation of its claim-level analysis, in which a model judge decided whether each cited claim was supported by the passage it cited. A human check was specified before any human annotation label existed, and two annotators labeled a sample of 100 cited claims. They agreed with each other on 94 of the 100, a kappa of 0.81, but agreed with the model judge on only 54 and 50. The rule set in advance required both 70% agreement and a kappa of at least 0.40 [6].

The main cause was a cap that cut every cited passage to its first 1,500 characters before the judge saw it. Retrieved passages ran to about 2,500 characters and wiki excerpts to about 430, so the cap cut 49 of the 50 sampled retrieval items and none of the wiki items, while the annotators read the full passages. Rescored without the cap, the judge's agreement with the annotators rose to between 65 and 69%, which still failed the rule against either annotator, and the judge's claim-level contrast between the two systems reversed. That reversal is in the automated scorer's measurements, not in a human-validated result. The design also called for adjudicating the annotators' disagreements, which was not done, so agreement with an adjudicated consensus remains unassessed. A second

problem was found in a later precision review. The p-values reported for the claim-level contrasts came from a test with a finite-population correction that the pre-specification had named only in a power calculation. On one annotator's labels that test gave $p = 0.049$, where a conventional two-proportion test gives 0.11. Neither figure is fully adequate. The finite-population correction supports inference about the sampled claims rather than about the systems, and both tests treat claims as independent although claims cluster within answers, so both overstate precision [6].

The code for both studies was written by an AI coding assistant under my direction and review. Two assumptions it introduced without instruction, the passage cap and the non-traditional method of computing key statistics, were not caught at the time and were found by the checks described above.

The lesson is that assumptions need testing before they are fixed in code, and more thoroughly than these were. Beyond the time and effort, the cost was to credibility: once an implementation assumption reverses a reported result, subsequent claims reasonably receive greater scrutiny. Correction is necessary, but it does not make the original evidential record equivalent to one in which the assumption had been tested prospectively.

The test was fixed in an analysis script committed seventeen minutes after the pre-specification, nine days before any label existed, and the script did not change afterward. Including the relevant code in a prospective record can make such choices inspectable and establish when they were implemented. Here, however, the test sat in committed code from the start and still went unnoticed until the precision review. The commit history recorded when the test was written, but not whether the statistical method was appropriate. Code review and methodological validation remain separate requirements.

5 Related Work

The case for preregistration in machine learning is not new. The NeurIPS pre-registration workshops of 2020 and 2021 piloted a review model in which a study's protocol was accepted before any results existed [11], and van Miltenburg et al. (2021) presented preregistration questions for different kinds of NLP studies [12]. Søgaard et al. (2023) weighed the pros and cons, including the risk that preregistration biases NLP toward confirmatory research [13], and Leech et al. (2024) catalogued questionable practices in machine learning research, among them choosing evaluation details after seeing test results [14].

Closer to the present argument, Baumann et al. (2025) recommend preregistration of LLM configuration choices for text annotation, together with reporting of all tested combinations [9]. Vaccaro (2026) proposes a preregistration template for experiments with AI agents in which researchers commit to the evaluation prompt and the evaluator model version, and asks that submission guidelines require a statement of whether a study was preregistered [15]. Neither addresses disclosure policies of the kind described in Section 2, and neither reports a case in which a registration changed what was published.

I have not found the observation about publisher policies set out in Section 1 in the work reviewed for this paper, though Vaccaro's (2026) call for a preregistration statement in submission guidelines comes close [15].

6 Scope and Limitations

A registration does not prove that a result is correct, but it does show which methodological commitments were made before their consequences were known. Two studies do not show how common

the problem is. They illustrate what a timestamp changes when methodological choices remain open; they do not estimate how often such choices are made after the fact. The cases also suggest a need to inspect choices embedded in AI-generated code, whether or not the researcher is familiar with the underlying method. Inspection and registration may add the most where researchers now use AI for practices they know less well, or that they previously handed to other people, because those are the places where a choice they did not make is least likely to be noticed. These cases do not establish that such errors are more common with AI assistance than with other forms of software development.

Much machine learning research is also exploratory, and a universal preregistration requirement would fit exploratory work poorly. The discipline worth adopting is to distinguish confirmatory claims from exploratory analyses, and to preregister the former when the distinction matters [16]. Both studies described here contain both kinds, and in both the exploratory analyses produced the more interesting findings. A registration covering three predictions still allows a fourth analysis, provided the fourth is reported as exploratory.

A related objection comes from work on evaluation practice. Shankar et al. (2024) observed that people grading model outputs often need to see outputs before they can settle on their criteria, a pattern they call *criteria drift* [17]. Fixing a rubric in advance does not preclude a limited amount of rehearsal with real examples in order to refine it. Researchers should allocate examples for that express purpose and keep them apart from the examples used in the registered evaluation. For example, the plan for the Progressive-Disclosure study was locked only after a pilot run had been used to fix the number of replicates [5]. The relevant boundary is therefore the first confirmatory scoring run on evaluation data, not every preliminary use of the instrument. If evaluation measurements inform a later configuration change, that change should be identified, and its analysis distinguished from the original confirmatory plan.

7 Recommendations

These recommendations are addressed to researchers and to journals in general.

For confirmatory evaluations, researchers should register the configuration of the model-based instrument and the material implementation and analysis choices, together with the decision rule for adequate measurement, before the first confirmatory scoring run on evaluation data. The prospective record should identify the model and version, prompt, rubric, input-processing rules, aggregation procedures, reliability gates, and statistical analysis relevant to the claims. It should include or link to the version of the code in which those choices are implemented, so that assumptions introduced by an AI coding assistant are inspectable alongside the plan. Those assumptions should be tested before the confirmatory configuration is locked. A further recommendation comes from my own experience: researchers should become proficient with the registration process, allow sufficient time for it, and verify that the registered materials match the intended final plan.

For journals and other venues, a single addition to existing disclosure requirements would close the gap described in Section 2:

> For AI systems used to produce measurements or evaluations directly relevant to the conclusions, authors should state whether the instrument configuration, and the material implementation or analysis choices affecting its measurements or their interpretation, including assumptions in the code, were specified before or after the resulting measurements on evaluation data were observed. Authors should identify material changes made after observation.

This is a disclosure requirement rather than a preregistration mandate. It compels no particular research design and costs an author one or two sentences. On its own, the statement remains self-report. When supported by a timestamped registration, an archived code version with a verifiable timestamp, or an equivalent immutable record, the chronology becomes independently auditable. Version control records the sequence of changes, but independent evidence of when a version existed is needed to corroborate its timing.

Recording registration identifiers in publication metadata, alongside the DOI and any artifact links, is also recommended. A reader should be able to find the prospective specification without searching the acknowledgments or reconstructing the chronology from prose.

8 Conclusion

As generative AI becomes more deeply embedded in the research process, it lowers the cost of methodological iteration. Researchers can now re-prompt evaluators, alter rubrics, adjust preprocessing caps, or swap judge models in a matter of seconds. Yet this unprecedented agility introduces a subtle vulnerability into scientific disclosure. Current journal policies appropriately focus on "what" AI systems were used, but they remain silent on "when" those systems' consequential configurations were fixed. A perfectly detailed, retrospective methods section can fully satisfy today's reporting standards while quietly masking a journey through Gelman and Loken's garden of forking paths [10].

The two preregistered evaluations documented here demonstrate both the necessity and the limits of prospective commitment. On one hand, locked analysis plans and fixed reliability gates fundamentally changed what could be reported. This prevented the post-hoc lowering of standards when judge agreement fell short of its preregistered threshold, and forced a candid correction when a second judge, added to a supplementary analysis, changed a result that had already been written up. On the other hand, these studies reveal that preregistration is not a panacea. A timestamped plan cannot validate a rubric or automatically catch assumptions in the code. Methodological validation, code review, and rigorous manual checks remain human responsibilities.

Timestamped preregistration does not guarantee that an experimental design is correct. Rather, its power lies in establishing an auditable chronology. By separating choices made before observing evaluation data from exploratory adaptations made after, researchers can preserve the evidential integrity of confirmatory claims. Adopting prospective disclosure and transparently reporting timing is a low-friction addition to current publisher guidelines. It represents a vital step toward maintaining trust, rigor, and reproducibility in an increasingly automated research landscape.

Disclosure

Claude Opus 5.5 (Anthropic) interviewed the author, searched the literature, checked facts and drafted the text, which the author reviewed and revised. ChatGPT (OpenAI, GPT-6) subsequently assisted with consistency review and editorial revision. The papers on the two studies did not report that their code was written by an AI coding assistant.

References

[1] Association for Computing Machinery. *ACM Policy on Authorship*, updated 14 May 2026.

[2] International Committee of Medical Journal Editors. *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*, updated January 2026. `https://www.icmje.org/icmje-recommendations.pdf`

[3] Elsevier. *Generative AI policies for journals*, updated June 2026. `https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals`

[4] Wiley. *Using AI tools in your research.* Undated web page, accessed 3 October 2026. `https://www.wiley.com/en-us/publishing/article/ai-guidelines`

[5] T. O. Cochran. *Progressive Disclosure for LLM-Maintained Wiki Knowledge Bases: a Preregistered Ablation.* arXiv:2607.04576v2. Preregistration: OSF, DOI 10.17605/OSF.IO/FEKA7.

[6] T. O. Cochran. *Single-Round Vector RAG vs an LLM-Compiled Wiki: A Preregistered Comparison on a Small Multi-Domain Research Corpus.* arXiv:2605.18490v3. Preregistration: OSF, DOI 10.17605/OSF.IO/ZEMHP (registered 2026-05-06).

[7] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.* In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023, 46595-46623.

[8] S. Roy, R. Pujari, T. Kumarage, C. Peris, et al. *PReMISE: Policy Rubrics as Measurement Specifications for LLM Judges.* arXiv:2605.30803.

[9] J. Baumann, P. Röttger, A. Urman, A. Wendsjö, F. M. Plaza-del-Arco, J. B. Gruber, and D. Hovy. *Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation.* arXiv:2509.08825, 2025.

[10] A. Gelman and E. Loken. *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Department of Statistics, Columbia University, 2013.

[11] L. Bertinetto, J. F. Henriques, S. Albanie, M. Paganini, and G. Varol, editors. *NeurIPS 2020 Workshop on Pre-registration in Machine Learning.* Proceedings of Machine Learning Research 148, 2021. S. Albanie, J. F. Henriques, L. Bertinetto, A. Hernández-García, H. Doughty, and G. Varol, editors. *NeurIPS 2021 Workshop on Pre-registration in Machine Learning.* Proceedings of Machine Learning Research 181, 2022.

[12] E. van Miltenburg, C. van der Lee, and E. Krahmer. *Preregistering NLP research.* In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, 613-623. DOI 10.18653/v1/2021.naacl-main.51.

[13] A. Søgaard, D. Hershcovich, and M. de Lhoneux. *A Two-Sided Discussion of Preregistration of NLP Research.* EACL 2023. arXiv:2302.10086.

[14] G. Leech, J. J. Vazquez, N. Kupper, M. Yagudin, and L. Aitchison. *Questionable practices in machine learning.* arXiv:2407.12220, 2024.

[15] M. Vaccaro. *Preregistration for Experiments with AI Agents.* ICML 2026, spotlight position paper. arXiv:2606.11217, 2026.

[16] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. *The preregistration revolution.* Proceedings of the National Academy of Sciences 115, 11 (2018), 2600-2606.

[17] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo. *Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences.* arXiv:2404.12272, 2024.