Paper XII: Public Benchmark Rescoring

Michael Darius Eastwood
Independent AI alignment researcher, London · Author, *Infinite Architects* (2026)
ORCID 0009-0004-3222-7442 · michael@michaeldariuseastwood.com · michaeldariuseastwood.com
The ARC Theory · OSF osf.io/3tzp7 · every claim checkable

> *Within the ARC Theory: the external-validity protocol of the blinding law on public benchmarks.*

THE ARC THEORY (THE THEORY OF ARTIFICIAL RECURSIVE CREATION) · PAPER XII · PROTOCOL PAPER · FIRST PUBLISHED 14 AUGUST 2026 · REVISED 27 AUGUST 2026 · DOI · WORKING PAPER V1.9 **10.17605/OSF.IO/3TZP7**

Does the blinding effect survive on someone else's benchmark?

Reading key. Claims come in three sizes: results, laws, frames. This is a results-class instrument paper at protocol stage: its study is written, dated and prepared as a draft registration awaiting human submission, and no data has been collected. Nothing here reports a finding; everything here is what will count as one, fixed in advance.

Abstract. Paper IV.d established that blinding changes measured alignment outcomes on the programme's own instrument. The standing objection is external validity: the effect might be a property of ARC-Align rather than of AI evaluation generally. This protocol paper fixes, in advance, the one test the programme cannot be accused of designing to its own advantage: the same blinding manipulation run on an established public benchmark the programme did not build, curate or control. Blinded scoring removes the producing model's identity, provider and family markers, self-referential meta-commentary and length cues; unblinded scoring leaves them. The registered questions are whether blinding changes the score (H1, the per-response difference Delta_S, direction deliberately unregistered after IV.d's own sign reversal), whether rankings displace (H2, Kendall tau), which cue carries the effect (H3, a registered fractional factorial), and cross-family evaluator agreement (H4, a minimum three-model panel where fewer than three valid scores is missing, never zero). The registered prediction is a non-zero Delta_S smaller than IV.d observed with modest rank displacement, and the stated most likely outcome is a positive H1 beside a null H2, read as blinding mattering for measurement precision rather than leaderboard order. This is a protocol at registration stage: no data has been collected, and nothing here reports a finding.

1. Why this is the keystone

This is the keystone, because it is the one test the programme cannot be accused of designing to its own advantage. **Paper IV.d established that blinding changes measured alignment outcomes on the programme's own instrument. The standing objection to that result is external validity: the effect might be a property of ARC-Align rather than of AI evaluation generally, and no amount of internal replication answers that.**

The answer is to run the same manipulation on an established public benchmark the programme did not build, did not curate and does not control. **If the blinding effect reproduces there, it is a property of AI evaluation. If it does not, it is a property of the programme's instrument, and the blinding standard should be scoped to it.**

The design is deliberately unfavourable to the hypothesis**, and that is the point. Using someone else's benchmark forfeits every advantage of a home instrument: the items are not chosen, the scoring rubric is not the programme's, the difficulty distribution is fixed, and any ceiling or floor in the benchmark works against detecting an effect. A result obtained under those conditions is worth more than several obtained at home.**

What blinding means here is registered precisely**, because the word is doing heavy lifting. Blinded scoring removes from the evaluator's view: the identity of the model that produced the response, any provider or model-family marker, self-referential meta-commentary in which a response names its own origin, and response-length cues where the rubric permits length normalisation. Unblinded scoring leaves them.**

A positive result would establish that the blinding manipulation changes measured scores on an external benchmark under the registered conditions. It would not establish the size of the effect elsewhere, would not establish that any specific published leaderboard is wrong, and would not license restating other researchers' results. That last restraint is registered because the temptation to over-read this finding is the main risk to the programme's credibility.

2. The questions, fixed in advance

Primary, H1, blinding changes the score. **For each response, define Delta_S as blinded score minus unblinded score. H1 predicts a non-zero mean Delta_S, tested two-sided at alpha 0.05 with a 95 per cent interval.** The direction is not predicted**, because Paper IV.d's own sign-reversal finding makes a directional prediction unjustified, and registering a direction the programme has already seen reverse would be exactly the flexibility this document exists to remove.**

Co-primary, H2, rank displacement. **Scores aggregate to a ranking. H2 measures displacement between blinded and unblinded rankings by Kendall tau.** This is the contrast that matters practically**, because a change in scores that leaves rankings intact has no consequence for how benchmarks are used, and the registration refuses to report a score effect as though it were a ranking effect.**

Secondary, H3, which cue carries it. **The four blinded cues are removed in a registered fractional factorial rather than all at once, so the contribution of each is estimable without a full factorial. Resolution and aliasing structure are stated in the attached design.**

Secondary, H4, cross-family evaluator agreement. **A minimum three-model cross-family evaluator panel scores every response. Fewer than three valid scores yields** a missing value, never a zero**, because a panel that fails silently toward zero manufactures the effect it is meant to measure.**

Directional prediction, registered in advance. **The author predicts a non-zero Delta_S of smaller magnitude than Paper IV.d observed, and** expects rank displacement to be modest, because benchmark rankings are usually driven by large capability gaps that survive cue removal. The author states plainly that a null on H2 alongside a positive H1 is the most likely outcome, and that the correct reading of it is that blinding matters for measurement precision rather than for leaderboard order.

3. The design

Within-response, randomised, blinded-versus-unblinded rescoring on an external public benchmark, with a fractional factorial over the four cue types.

Every response is scored twice, once under each condition, by a cross-family evaluator panel, with call order randomised within block and condition coded opaquely. The unit of analysis is the response; models are a blocking factor.

Resolution-IV fractional factorial over four cues in eight runs, so main effects are unaliased with each other. The aliasing structure is published rather than described.

4. The benchmark

Benchmark. MT-Bench, the established public multi-turn evaluation benchmark maintained by LMSYS (Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", 2023), 80 questions across eight categories with a published GPT-4 pairwise/single-answer rubric, publicly available model responses, and licence terms permitting rescoring. Rationale for the default: MT-Bench is the programme's identified keystone external benchmark (the one benchmark this programme cannot be accused of designing to its own advantage), its rubric is published verbatim, model responses are publicly hosted, and its licence permits rescoring. **The benchmark is named in the registration and cannot be swapped afterwards; swapping a benchmark after a disappointing result is the specific abuse this clause prevents.**

The benchmark citation, verified against the arXiv record: Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks; arXiv:2306.05685.

5. Status, honestly

The draft registration's full title is "Does the Blinding Effect Survive on Someone Else's Benchmark? A Preregistered Rescoring of an Established Public Benchmark Under Blinded and Unblinded Conditions". It is drafted, dated and prepared as a draft registration awaiting human submission; nothing has been submitted, nothing has been run, and no result is claimed. Three conditions remain open on the working unit and are recorded there rather than glossed: the concrete sample size is computed at packaging from the benchmark's own published variance data; the named benchmark defaults to MT-Bench per the programme's keystone plan, subject to the author's override before submission; and the unit's own re-verification is pending. The public OSF component for this paper is osf.io/3tzp7. The full listing of the programme's drafted trials is public at the drafted-trials page.

6. What a result would and would not mean

A positive result would establish that the blinding manipulation changes measured scores on an external benchmark under the registered conditions. It would not establish the size of the effect elsewhere, would not establish that any specific published leaderboard is wrong, and would not

license restating other researchers' results. **That last restraint is registered because the temptation to over-read this finding is the main risk to the programme's credibility.**

The restraint above is part of the design: the temptation to over-read an external-benchmark result is the main risk to the programme's credibility, and it is fenced in the registration itself.

Related papers of the theory

The blinding law this study externalises: Paper IV.d (10.17605/OSF.IO/2S3E6), the measurement-discipline paper whose blinding reversal produced the programme's one public retraction. The theory this instrument serves: the statement paper (10.17605/OSF.IO/GW5MX). The full catalogue is at the papers index.

How to cite this paper

DOI (this paper): 10.17605/OSF.IO/3TZP7 · Umbrella DOI: 10.17605/OSF.IO/6C5XB

Eastwood, M. D. (2026). Paper XII: Public Benchmark Rescoring. The ARC Theory · ARC/Eden experiments. https://doi.org/10.17605/OSF.IO/3TZP7

Declaration of AI-Assisted Human Authorship

The author of this work is Michael Darius Eastwood, a human being. Every core concept, claim and conclusion originates from human ideation; artificial-intelligence tools were used as instruments under continuous human direction for drafting, verification and formatting. All selection, coordination, arrangement and final editorial judgement are the author's, who takes full responsibility for the accuracy and integrity of the text.

PAPER XII: PUBLIC BENCHMARK RESCORING

Protocol paper | 14 August 2026

Michael Darius Eastwood · ORCID 0009-0004-3222-7442

The ARC Theory (the Theory of Artificial Recursive Creation) · study drafted awaiting human submission · this protocol paper first published 14 August 2026

Artificial-intelligence tools (Anthropic's Claude family and other large-language-model assistants) were used as instruments under continuous human direction, in the way a word processor, calculator or research assistant is used: for editing and prose refinement, literature search and summarisation (manually verified against primary sources), document structure, formatting, brainstorming against author-defined questions, and the acceleration of drafting to author-defined outlines and instructions. All selection, coordination, arrangement and final editorial

judgment are the author's. Every substantive output was reviewed, tested or verified by the author, who takes full responsibility for the accuracy and integrity of the final text. The tools increased the speed of the work; they were never relied upon as its source.

Epistemic status. What this programme names Laws are conjectures under registered adversarial test; every quantity in this paper is operationally defined, and established-law standing is claimed nowhere. The registered programme exists to earn that standing, or lose it, by measurement, replication and survived refutation.

Standing covenant. Prove this paper wrong, and I will publish the refutation myself. Falsification conditions are stated in this paper; the standing challenge: github.com/MichaelDariusEastwood/arc-scaling-challenge.

How to cite this paper

DOI: 10.17605/OSF.IO/6C5XB

APA (AUTHOR-DATE, 7TH EDITION)

```
Eastwood, M. D. (2026). Paper XII: Public Benchmark Rescoring. ARC and Eden
Research Programme. https://doi.org/10.17605/OSF.IO/6C5XB.
https://www.michaeldariuseastwood.com/research/papers/paper-xii-public-
benchmark-rescoring.html
```

BIBTEX

```
@article{paperpaperxiipublicbenchmarkrescoring,
  author  = {Michael Darius Eastwood},
  title   = {Paper XII: Public Benchmark Rescoring},
  year    = {2026},
  date    = {2026-08-14},
  journal = {ARC and Eden Research Programme},
  doi     = {10.17605/OSF.IO/6C5XB},
  url     = {https://www.michaeldariuseastwood.com/research/papers/paper-xii-
public-benchmark-rescoring.html}
}
```

INLINE

```
Eastwood, 2026
```

> See /how-to-cite/ for the canonical citation manual across all artefacts, or /citations.json for the machine-readable index.