Judging Criteria

How we'll judge. We score submissions on the dimensions below. Strong submissions don't need to land well on every dimension — a deep contribution on a few can still win the top tier — but they should plausibly engage most of them.

1. Epistemic uplift. Does this actually help a thoughtful person reason better about the case?

- Meaningfully better than off-the-shelf deep research / top-of-range Claude Code investigation on the same sub-questions
- Faithful to evidence — doesn't sand off uncertainty or smuggle in unsupported confidence
- Makes load-bearing evidence visible: which pieces are actually driving the conclusion? Which claims and evidence are getting unearned, or not enough weighting in an analysis?
- Surfaces what matters: cruxes, missing perspectives, rhetorical-vs-evidential moves

2. Generalizability. Will the workflow travel?

- Works across cases of different shape (curated debate / confident answer w/ complex evidence / mundane-but-contested)
- Plausibly applicable to cases beyond the three we provide
- Not narrowly overfit to the provided case studies

3. Compounding & shareability. Do the artefacts produced help future investigators build on this work?

- Outputs are structured / interrogable, not just narrative summaries
- Another team could pick up the artefact and extend it
- Pieces could plausibly interoperate with other submissions' pieces

4. Scalability. Does the approach get better with more compute, better models, or more contributors?

- Not bottlenecked on any single hand-designed human step
- Benefits as base-model capability rises
- Benefits from increased resources: more (adversarial) scrutiny, incremental sources, more effort/compute spent on checks

5. Methodological transparency. Is the submission well-specified enough to evaluate, replicate, and critique?

- Spec/workflow written down, with the key decisions and tradeoffs called out
- Where the creator is uncertain, that uncertainty is named — not papered over
- A judge can tell why the methodology is shaped this way, not just what it does

6. Adversarial robustness. How well do the artefacts and methodology hold up when participants and consumers have differing views and priorities?

- Outputs withstand motivated reading and downstream-model interrogation
- The methodology resists being gamed by sources optimizing to mislead
- Failure modes and uncertainties are named and bounded, not hidden

7. **Insight contribution.** Does the submission shift how we think about the problem itself?
    - Surfaces sub-problems, framings, or considerations we'd missed
    - Critiques that force re-evaluation of promising approaches, especially with potent counterexamples
    - Comparative analyses that surface non-obvious tradeoffs or cutting dilemmas across methodologies

---

Prize tiers (guidance, not a formula — judges' discretion):

Tier	Range	Looks like
Transformative	$35k–$50k	Substantial advance on the SOTA — reshapes how we think the next generation of this tooling should be built. Could be a single deep contribution or a broader push across several dimensions.
Strong	$15k–$35k	Notable improvement on the SOTA. Either a meaningful gain across several dimensions, or a clearly impressive gain on one.
Promising	$5k–$15k	A real but mild improvement on the SOTA, or a partial/exploratory contribution containing insight or working components we'd want to build on.

Multiple prizes possible per tier; pool can expand for a wave of strong work; strong submissions may also lead to offers of further funded work with us.

Notes for judges.

1. Anchor against good baselines. Before scoring, check what off-the-shelf deep research or a careful Claude Code investigation produces on the same sub-question. The bar is "meaningfully better than that."
2. Read for the spec, not the polish. A clear workflow with a rough prototype usually beats a polished prototype with opaque methodology.
3. Run it, don't just read it. For tool and workflow submissions, exercise the methodology on a sub-question you're personally curious about — usefulness shows up in use, not in the writeup.

Strong examples

Abridged collection of strong examples

Transparent Replications (by Clearer Thinking) is an effort to spot-check scientific contributions in psychology. Not limited to perfunctory study replication per se, they've identified and catalogued 'importance hacking' practices by carefully scrutinizing what was actually measured alongside the ways it's communicated and framed.

Elisabeth Bik's 2021 John Maddox prize celebrated her record of exposing threats to research integrity — in particular a prodigious capacity for forensic analysis of image validity in scientific papers.

Nadel and Pritchett's 2016 discussion crystallized the construct-validity critique of development RCTs — that "the same" program is often a different thing across sites — drawing significant attention to a problem the field had been papering over.

Measurement Schmeasurement by Flake and Fried exposed 'questionable measurement practices' in psychology, a subtler and widespread complement to the more famous poor practice of p-hacking.

Data Colada's investigations into the origins and veritability of scientific datasets exposed several now-high-profile cases of scientific fraud, as well as several smaller-scale cases of sloppy scientific reporting.

Leveson's *Engineering a Safer World* and related publications build up a convincing 'systems theoretic' framework for safety engineering. Repeatedly encountering cases with specifications 'out of date as soon as published', she gives methods and mindsets fit for rapidly-changing and complex contexts.

Examine.com applies disciplined evidence-grading to supplement and nutrition research, parsing study quality and effect sizes claim-by-claim — a steady counterweight in a field saturated with industry-funded cherry-picking, and one that updates positions as new trials land.

Andrew Gelman's blog is a running catalog of houses of cards: shaky statistics, inappropriately confident conclusions, dubious extrapolations, and the like.

The Society Library's Diablo Canyon investigation ambitiously engaged with the whole range of perspectives on the fate of California's last nuclear power plant: a complex topic with deep chains of reasoning on many subtopics.

Dick Heuer, decorated CIA veteran, championed structured analytic techniques in intelligence analysis, including analysis of competing hypotheses; his insistence on rigor and open-mindedness broke through several commonly-accepted falsehoods — even in a highly adversarial and information-scarce environment.

Format & Length

Format & length

We're flexible on format. The cap we're really setting is on **judge attention budget**, not artifact size — submissions can include arbitrarily large code, knowledge bases, or formal specs, as long as you tell judges which slice deserves their attention first.

Reading budget: up to ~10 pages of end-to-end engagement, allocated however suits your submission — across written content, a knowledge-artifact slice, pseudocode, or a mix. Beyond this, judges skim and inspect particular areas they choose to dive into.

Written submissions (specs, protocols, analyses, critiques):

- **Core** (within the reading budget): ≤ ~10 pages
- **Supporting material**: up to ~20 more pages.
- **Appendices**: no limit. Formal type definitions, schemas, grammars, ontologies, worked examples, and reference material can live here.

Code / executable submissions:

- No length limit on the code itself.
- Include **either** readable pseudocode (≤ 2 pages capturing algorithm and key decision points) **or** a ~one-click runnable demo (single command via Docker / Colab / Modal / similar; fresh machine to running process in ~5 min). Both is better.
- Short README pointing judges at the most informative invocation(s) to try and the most interesting parts of the code to inspect.

Knowledge artifacts (structured outputs your method produces):

- No ‘length’ limit on the artifact itself.
- Designate **representative slices** that fit within the reading budget alongside any written core, perhaps formatted as a brief tour of a few select entry points.

AI-assisted reading is welcome on both sides. Include your own AI-generated digest of long artifacts if useful; judges may also use AI to triage.