The Singularity Gate

A benchmark for AI-driven paradigm-shifting scientific discovery

Abstract

The question of whether AI systems are approaching the capability to drive autonomous scientific discovery, the kind of generative paradigm-shifting science that defines a technological singularity, has so far been argued without a measurement instrument. We introduce **The Singularity Gate**, a benchmark designed to test a *necessary but not sufficient* prerequisite for such capability: the ability to predict the actual paradigm-breaking content of post-cutoff scientific discoveries from training data alone. The benchmark's defining methodological investment is an exhaustive search for contamination-free, paradigm-shifting items. Each candidate is admitted only if (i) its first public breadcrumb — across corporate press releases, specialty-conference talk decks, Zenodo / OSF / ResearchGate / ResearchSquare preprints, early-online publication ahead of nominal print date, bioRxiv author-search gaps, asymmetric model-cutoff sensitivity probes, and direct online-publication-date verification — falls *strictly after* the latest empirically located training cutoff in the respondent panel; (ii) the finding is paradigm-breaking rather than incremental or a replication; (iii) the published abstract names a specific, well-defined mechanism, magnitude, and direction; and (iv) a per-item *parallel-true audit* confirms that the prompt admits exactly one published finding rather than multiple adjacent-but-true findings in neighbouring sub-domains. The corpus is maintained as the frontier moves: whenever a new model is added to the respondent panel, its empirical training cutoff is located by probing on dated world events, the audit is re-run against the new latest cutoff in the panel, items overtaken by it are retired, and strictly post-cutoff replacements are admitted — holding the corpus contamination-free across model generations. All respondents are evaluated in their lab's own native agentic harness (Claude Code for Claude models, Codex for GPT-5.5, Gemini CLI for Gemini 3.1 Pro), with tool use enabled and web search disabled, measuring deployed-product capability rather than bare-LLM capability. We report rankings across five frontier models (Claude Opus 4.7, Claude Sonnet 4.6, Claude Opus 4.6, Gemini 3.1 Pro, and GPT-5.5) across five broad scientific fields: life sciences, chemistry and materials, physics and astronomy, earth and planetary, and mathematics and theoretical computer science. Failure on this benchmark falsifies singularity-readiness; success is a necessary but not sufficient signal that the gate is open.

1 Introduction

1.1 The singularity prerequisite

A widely-discussed claim about contemporary AI is that, scaled and refined further, current architectures will reach a threshold at which they autonomously produce scientific discoveries: generative paradigm shifts of the kind that historically have defined major scientific revolutions. The strong form of this claim is the *technological singularity* hypothesis, which holds that AI's

ability to extend the frontier of human knowledge will at some point exceed humans' ability to extend it themselves.

Whether this hypothesis is true is, in the end, an empirical question. But empirical work on it has been hampered by the absence of a measurement instrument that *targets the right capability*. Most existing benchmarks measure either (i) test-bank knowledge already present in the training data (MMLU [Hendrycks et al. 2021]; GPQA [Rein et al. 2023]), (ii) constrained reasoning over fully-specified inputs (FrontierMath [Glazer et al. 2024]; AIME / programming benchmarks), or (iii) recognition tasks where the answer is one of N supplied options. Even Humanity's Last Exam [Phan et al. 2025], the most recent frontier-knowledge benchmark, scores models on questions whose answers exist somewhere in the corpus they were trained on. None of these probe whether a model can generate, from training-data priors alone, a structurally novel scientific finding it has never encountered.

The thought experiment that organises this benchmark is the **GR-1911 question**. We know Einstein discovered general relativity: he produced it by reasoning over the priors available to him before 1915 — Mercury's perihelion anomaly, the equivalence principle, Riemannian geometry, the Lorentz transformation. The converse question is the one we want a measurement instrument for: if an AI system had its knowledge cutoff in 1911, given those same priors and nothing later, could it invent a useful first-pass version of general relativity? The empirical version of that question — the one we can actually run today — has the same shape: not whether contemporary AI systems can derive a paradigm-shifting theory de novo, but whether they can *predict* scientific findings whose constituent priors exist in their training data but whose synthesised result does not, because the result was published after their training cutoff. Demis Hassabis has independently identified this same direction as a candidate AGI test [Hassabis 2025, 2026]; the present benchmark operationalises that direction as a measurement instrument that can be run on currently-deployed frontier systems.

1.2 Necessary, not sufficient

The Singularity Gate measures a strict subset of what would be required for AI-driven autonomous discovery. Concretely, we test whether a model can predict the **specific content** of a paradigm-breaking scientific finding published *after* its training cutoff, given only an open-ended question stripped of any signal that would specify the answer. This is a prerequisite capability. Passing it does not by itself certify singularity-readiness, and many further capabilities (experimental design, instrument construction, multi-year project planning, paradigm articulation, literature curation) would be required for actual autonomous science. But *failing* it falsifies the strong form of the singularity hypothesis: a model that cannot predict known post-cutoff findings from priors cannot be expected to generate new ones.

We adopt this framing, **failure-falsifies, success-necessitates-but-not-sufficient**, deliberately. It avoids the recurring failure mode of singularity-adjacent benchmarks, which is to overclaim ("this benchmark measures AI's capability for autonomous scientific discovery") in ways that the underlying instrument cannot support. The Singularity Gate's claim is narrower and stronger: it is the *first* gate; many gates lie beyond it.

1.3 What this paper contributes

1. **A working benchmark.** A corpus of items grouped into five broad scientific fields, all admitted under a multi-source contamination audit and a per-model cutoff grid search, with

prompts written under a locked mode-neutral, parallel-true-audited format.

2. **A contamination-audit methodology.** A protocol that systematically covers the categories of pre-cutoff disclosure that escape title-only web search: corporate press releases, specialty-conference talk decks, Zenodo / OSF / ResearchGate / ResearchSquare preprints, early-online publication ahead of nominal print date, bioRxiv author-search gaps, and direct online-publication-date verification.
3. **A per-model cutoff grid search.** Each respondent's training cutoff is empirically located by probing it on dated world events; the latest cutoff in the panel is adopted as the corpus admission floor, and only items whose first public breadcrumb falls strictly after that floor are admitted. This neutralises both lab-cutoff imprecision and asymmetric-cutoff contamination across respondents.
4. **A parallel-true audit methodology.** A per-item literature audit identifying alternative-but-true answers in adjacent sub-domains, with scope anchors selected to admit the actual finding while excluding the alternatives. Audit fields are part of every item's corpus record.
5. **A scoring metric that resists retrieval and luck.** Per-item score is Reasoning $\times$ Outcome (each on 0–5) normalised so that a perfect response scores 100%. Reasoning is anchored to the abstract's specific mechanism rather than to generic competence, so a response that reaches the right outcome by retrieval, paraphrase, or lucky guess earns near-zero unless it also shows a reasoning path that anticipates the abstract's mechanism.
6. **A cross-lab non-respondent judging protocol.** Three judges from three different labs, each non-respondent (no respondent is also a judge). Self-preference is removed at the design level rather than averaged out post-hoc.
7. **Native-harness evaluation.** Respondents are run in their lab's own agentic harness (Claude Code, Codex, Gemini CLI) with tool use enabled and web search disabled, measuring deployed-product capability rather than bare-LLM capability.
8. **A maintained corpus.** A refresh policy that retires items overtaken by new training cutoffs and adds strictly post-cutoff replacements, holding the headline count steady across model generations.
9. **A 5-model headline.** Rankings for Claude Opus 4.7, Claude Sonnet 4.6, Claude Opus 4.6, Gemini 3.1 Pro, and GPT-5.5 on the n=104 headline corpus, with a per-field breakdown across the five broad fields.

The paper is structured around what we believe are the methodological contributions, with the model rankings appearing as a result of applying that methodology rather than as the headline. The methodology *is* the contribution: most of the items in this benchmark could not be substituted by a different group with a different audit and produce the same numbers, because the audits are precisely what define what this benchmark measures.

2 Related Work

2.1 Scientific knowledge and reasoning benchmarks

The dominant evaluation paradigm for scientific knowledge in LLMs is closed-form multiple-choice or short-answer over a fixed test bank. **MMLU** [Hendrycks et al. 2021] introduced 15,908 multiple-choice questions across 57 subjects; frontier models now exceed 88% and the benchmark is approaching saturation. **GPQA** and its hardest subset GPQA-Diamond [Rein et al. 2023] used PhD experts to author 448 (Diamond: 198) science MCQs designed to be Google-proof; frontier models

in late 2025 reach ~90% [Google DeepMind 2025]. **AGIEval** [Zhong et al. 2023] benchmarks against human standardised tests (SAT, LSAT, Gaokao). **FrontierMath** [Glazer et al. 2024] presents original research-level math problems where SOTA at release was below 2%; even at the time of writing, frontier scores remain in the low-to-mid double digits on the hardest tier. **Humanity's Last Exam** [Phan et al. 2025] assembles 2,500 frontier expert questions; SOTA at release was 10–15%, with current frontier (May 2026) at 33–57% depending on model and tool access. **BIG-Bench** [Srivastava et al. 2022] and **BIG-Bench Hard** [Suzgun et al. 2022] cover heterogeneous tasks beyond science specifically. **ARC-AGI** [Chollet 2019; Chollet et al. 2025] tests abstract pattern induction.

What unites these benchmarks is that the answer to every item already existed in some form before the model was trained. They measure recall, recognition, and constrained derivation, not synthesis of a finding the model could not have encountered. The Singularity Gate complements them by anchoring on findings whose first public appearance falls strictly after each respondent's empirically located cutoff (§5.3).

2.2 Training data contamination

Contamination has been a recurring concern for benchmark validity since the GPT-3 era. **Magar & Schwartz [2022]** distinguish memorisation from exploitation. **Sainz et al. [2023]** argue that per-benchmark contamination measurement is now mandatory. **Oren et al. [2024]** give a black-box statistical test for test-set contamination. **Roberts et al. [2024]** show, on Codeforces and Project Euler, that pass rates correlate with both GitHub popularity and pre-cutoff release date. **MMLU-CF** [Zhu et al. 2024] reconstructs MMLU under contamination control: GPT-4o drops from 88.0% on the original benchmark to 73.4% on the contamination-free reconstruction (a 14.6 pp drop), with similar 14–17.5 pp drops across other models. The seven structural blind spots we document in §5.1 are complementary to this literature rather than overlapping with it: where MMLU-CF reconstructs an existing benchmark, we describe the per-item audit categories required to admit a *new* paradigm-shift item without naive-audit contamination from the start. Our cutoff grid search (§5.3) and the same-lab-continuation filter (§3.1, criterion 1) are, to our knowledge, novel.

2.3 LLM-as-judge

The use of LLMs as automated judges was crystallised by **Zheng et al. [2023]** ("Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena"), which named the canonical biases (position, verbosity, self-enhancement) and reported that GPT-4 vs human-preference agreement matches inter-human agreement on conversational-quality judging. Subsequent surveys [Gu et al. 2024; Li et al. 2024] catalogue the field. **Panickssery et al. [2024]** demonstrate that LLM evaluators recognise and favour their own generations, with self-preference scaling linearly with self-recognition accuracy; this directly motivates the cross-lab non-respondent panel of §7. **Wataoka et al. [2024]** quantify GPT-4 self-preference under multiple judge-generator pairs. **Bavaresco et al. [2024]** ("Judge-Bench") measure inter-judge variance across 20 NLP tasks and find that variance is large in absolute terms but smaller in rank terms. **Shi et al. [2024]** quantify position bias in 15 LLM judges across 22 tasks, motivating the per-item label randomisation of §7.2. **JudgeBench** [Tan et al. 2025] reports the strongest LLM judge at 64% on hard items, an empirical anchor for the difficulty of LLM-judging at the frontier.

2.4 Where this paper sits

Existing scientific-knowledge benchmarks measure recall and constrained derivation over closed answer sets; existing contamination methodologies handle benchmark reconstruction but not new-item admission for post-cutoff paradigm shifts; existing LLM-as-judge methodology characterises self-preference and inter-judge variance but does not show how to remove the former from a multi-respondent panel by construction. The Singularity Gate combines an open-ended prompt format, a parallel-true scope-anchor audit, a seven-category contamination audit with empirical cutoff probing, an R × O scoring metric that resists retrieval and luck, and a cross-lab non-respondent judging panel that removes self-preference at the design level rather than averaging it out. The methodological combination is, to our knowledge, novel.

3 Benchmark Design

3.1 Goal and scope

The benchmark targets one capability: predicting the specific content of paradigm-breaking scientific findings published after a training cutoff, from priors alone. The unit of evaluation is an *item*: a single open-ended scientific question paired with a single published paper that supplies the ground-truth answer.

For an item to enter the corpus, it must satisfy six locked criteria:

1. **Conceptually derivable from priors.** The published finding's conceptual content — its mechanism, direction, structural characterization, theorem, or qualitative paradigm shape — must be derivable in principle from priors public at the model's training cutoff, by reasoning alone. The benchmark targets the GR-1911 capability of *deriving* paradigm shifts from priors. Items whose paper uses empirical confirmation are admissible: a model that arrives at the correct mechanism through reasoning still demonstrates the target capability, even if the paper itself runs the experiment. Items whose core answer is fundamentally an unpredictable specific numeric measurement (e.g., a clinical trial's specific primary-endpoint value to within a few percentage points, a detector's specific σ, a specific named compound chosen among many similar candidates by experimental screening alone) are excluded. The publishing group must also be distinct from the framework originators; same-lab continuations admit too much breadcrumb signal and are excluded by a hard filter.
2. **Strictly post-cutoff publication.** The paper's first public breadcrumb (online publication, preprint, press release, talk, thesis, or any other category in §5.1) must fall strictly after the panel cutoff floor: the latest empirical training cutoff among all respondents, located by the grid search of §5.3. The print date of the published paper is not used; the *online* date and the earliest *any* form of disclosure is.
3. **Paradigm-breaking, not incremental.** The finding overturns or substantially revises a prevailing default in its subfield. Incremental work, replication studies, and same-direction extensions of existing frameworks are excluded at admission.
4. **No prior public disclosure.** No preprint (arXiv, bioRxiv, ChemRxiv, medRxiv, Zenodo, OSF, ResearchGate, ResearchSquare), no corporate press release, no conference talk deck, no thesis chapter, no public lecture or news coverage, that discloses the *specific finding* before the panel cutoff floor. The audit categories that this criterion enforces are documented in §5.

5. **Single, well-defined answer.** The published abstract names a specific mechanism, magnitude, and direction. Vague or multi-part findings are excluded.
6. **Parallel-true-clean prompt.** A *parallel-true audit* (§6) confirms that the prompt's scope anchor admits exactly one published finding, not multiple alternative-but-true findings in adjacent sub-domains.

Items that pass all six are admissible.

3.2 Coverage across scientific fields

The Singularity Gate is positioned as a *broad-spectrum* benchmark rather than a within-discipline test. Items are sourced across all of natural and formal science, with deliberate effort to recover items in fields whose preprint culture would otherwise concentrate the corpus in a few subfields.

To keep per-field samples large enough to support useful disaggregated reporting in §9.3, the underlying sourcing subfields are aggregated into **five broad fields** for analysis. The aggregation follows standard taxonomic groupings (life sciences, physical sciences, formal sciences, earth sciences) rather than venue-specific cuts:

Field	Subfields aggregated
Life Sciences	biology (cellular/molecular/developmental), medicine and clinical, neuroscience and sleep biology, marine biology, paleontology and paleoanthropology
Physics and Astronomy	condensed-matter, statistical, quantum, astrophysics, cosmology
Mathematics and Theoretical CS	pure and applied math, theoretical CS, complexity theory, structural economic/labour-market modelling
Chemistry and Materials	organic, inorganic, materials, photovoltaics, catalysis
Earth and Planetary	geology, volcanology, climate, planetary, paleoclimate

Coverage rationale. No single field dominates: Life Sciences is largest, but the remaining four fields together account for the majority of the corpus, and three of the five (Chemistry, Physics, Math/TCS) each carry sample sizes large enough for stable per-field rankings. Two structural constraints shape the distribution.

First, preprint culture asymmetry. Physics, mathematics, and theoretical computer science publish almost universally on arXiv before peer review. Empirical median preprint-to-publication delay is approximately 21 months for mathematics and 6 months for physics [Wang et al. 2020]. The post-cutoff window after preprint exclusion is therefore narrow in these fields. Biology, chemistry, and medicine have weaker preprint culture in many subfields, opening larger windows. We deliberately invested research effort to recover items in the harder-to-source fields (Mathematics and TCS, theoretical physics) so that the corpus is not biased toward Life Sciences alone.

Second, same-lab continuation density. Some fields (synthetic chemistry, materials science) have many same-lab continuation papers that are predictable from a lab's recent trajectory; these are excluded by criterion (1) of §3.1. After this filter, raw publication volume in a field does not translate cleanly into corpus eligibility.

The current corpus is thus a deliberately broad-spectrum sample, not a demographic snapshot of contemporary publication. Passing the gate in one field is much weaker evidence than passing it

across fields, because intra-field retrieval-style memorisation could plausibly produce the former without the latter. The per-field analysis in §9.3 makes this gradient legible.

3.3 Corpus size

The strict-bar pool defined by the six criteria of §3.1 (paradigm-breaking findings published strictly after the panel cutoff floor, with no pre-cutoff disclosure across any of the categories audited in §5, with single well-defined answers admissible to a parallel-true-clean prompt) is small per unit time. The current headline corpus contains **n=104 items**. The per-item score (§7.4) is computed on this corpus, and the refresh policy in §5.4 maintains the headline count across model generations.

For comparison, GPQA-Diamond is n=198 [Rein et al. 2023] and FrontierMath is in the low hundreds [Glazer et al. 2024]; both, like ours, sacrifice item count for item quality.

4 Item Curation Pipeline

(To draft. Cover: source venues, the polish workflow that converged the corpus to CLEAN over 2.5 polish rounds, the ~26% verifier disagreement rate on first-pass polish (an important methodological finding that agent self-checks systematically under-detect leakage). Reference `polish/POLISH_BRIEF.md` and `methodology_review/REVIEWER_BRIEF.md`.)

5 Contamination Audit Methodology

Title-only web search is not a sufficient contamination audit for paradigm-shift prediction items. Empirical work across multiple LLM benchmarks shows that contamination is widespread and that removing it shifts measured performance by 14–17.5 percentage points in MMLU-CF [Zhu et al. 2024], with related findings on Codeforces and Project Euler [Roberts et al. 2024]. Detection methods such as black-box statistical contamination tests [Oren et al. 2024], canonical/shuffle tests [Magar & Schwartz 2022], and per-benchmark audit recommendations [Sainz et al. 2023] address contamination of benchmarks that already exist. The audit described in this section is complementary: it sets the conditions under which a *new* paradigm-shift item is admissible at all. Pre-cutoff disclosure of a specific scientific finding can take any of seven structurally distinct forms, each of which the locked audit covers explicitly.

5.1 Audit categories

1. **Corporate press releases.** Industry-sponsored trials routinely disclose primary endpoint outcomes via press release before peer-reviewed publication, sometimes 6–12 months earlier. The audit includes a manufacturer + lab + first-author press search for any industry-sponsored item.
2. **Conference talks at specialty meetings.** Specialist conferences (TAUP for dark matter, AD/PD for Parkinson's, physics-specialty Indico programmes) post slide decks and abstracts that often contain specific findings 4–8 months pre-publication, and are not indexed by Google Scholar. The audit includes Indico / programme search at every domain-specific specialty meeting of the prior 18 months.

3. **Zenodo / OSF / ResearchGate / ResearchSquare preprints.** These platforms host preprints that are not always indexed by Google Scholar or by the standard preprint server search interfaces. The audit checks each of them directly by site search.
4. **Early online publication ahead of nominal print date.** Nature and Science routinely post papers online 1–3 months before the nominal print date. The cutoff-relevant date is the *online* date, verified directly from the journal page rather than inferred from the print date.
5. **bioRxiv author-search coverage gap.** bioRxiv topic search frequently misses preprints that exist in the database under author-search. The audit performs corresponding-author preprint history checks in addition to topic searches.
6. **Pre-cutoff publication date verification.** The published paper's online date is verified directly against the panel cutoff floor rather than against any assumed model-stated cutoff.
7. **Asymmetric model-cutoff sensitivity.** Models in the same evaluation can have different empirical cutoffs. The audit floor is the *latest* such cutoff (§5.3), so items admissible under one respondent are admissible under all.

5.2 Audit protocol

Every item in the corpus has been audited against ALL seven categories:

- Direct site search at every domain-relevant preprint server: bioRxiv, medRxiv, chemRxiv, arXiv, EarthArXiv, PsyArXiv, plus Zenodo, OSF, ResearchSquare, SSRN;
- Corporate press release search if the work is industry-sponsored;
- Conference Indico / programme search at domain-specific specialty meetings of the prior 18 months;
- bioRxiv corresponding-author search in addition to topic search;
- Online publication date verified directly from the journal page;
- First public breadcrumb across all the above categories compared against the panel cutoff floor from §5.3.

An item is admitted only if all seven categories return clean.

5.3 Per-model cutoff grid search

Lab-stated training cutoffs lack the temporal precision needed for an admission floor: they are reported at month-or-coarser resolution, and the upper edge of training data effectively absorbed during late-stage fine-tuning is generally not the date the lab announces. We therefore perform an empirical cutoff grid search per model.

Protocol. For each respondent model, we probe with dated factual questions about *popular world events* (election outcomes, sports finals, named natural disasters, headline-grade scientific announcements, well-known deaths, government transitions, major product launches), sweeping a window from approximately three months before the lab-stated cutoff to approximately three months after. Probes are knowledge-only ("Who won the X? When did Y happen? Who is the current Z?"), not retrieval-of-prediction. We grade each probe as *known* (model produces the correct fact unprompted), *partial* (model knows there was an event but not specifics), or *unknown* (model gives no signal of the event existing or treats it as a future event).

The empirical cutoff is the latest date by which the *known* fraction has dropped to chance level.

The grid is dense (typically 2–4 probes per week-of-event) so the resulting cutoff has resolution of approximately 1–2 weeks.

Why probes need to be popular events, not scientific. Scientific publications are not in pretraining data uniformly. High-profile papers in Nature/Science get crawled and rehearsed dozens of times across paraphrasing, news coverage, and lay summaries; a niche specialty-journal paper may never enter the model's effective working knowledge even if technically present in scrape data. Popular world events behave differently: they are saturation-cited across web pages within days of occurrence, so a model's knowledge cutoff for popular events is a tight, interpretable lower bound on its effective training cutoff that scientific-paper recall would not provide.

Panel cutoff = max of respondent cutoffs. For corpus admission, we adopt the *latest* empirical cutoff among all respondents in the panel as the admission floor. An item is admissible only if its first public breadcrumb (preprint, press release, conference talk, thesis, ResearchGate post, or anything else in the §5.1 list, plus the published online date) falls *strictly after* this floor.

Why max, not min. The minimum cutoff would admit items contaminated for the latest-cutoff models. The maximum admits no item that could plausibly have leaked into any respondent's training, so resulting numbers are interpretable as ranking-on-novel-content for *all* models in the panel simultaneously, with no per-model asterisks.

5.4 Corpus refresh policy

The benchmark is not a fixed-corpus snapshot. As new models are released and the panel cutoff moves forward, items whose breadcrumbs fall before the new panel cutoff are retired, and strictly post-cutoff replacements are added. The corpus is maintained at the headline count across model generations under the following policy:

1. **Re-grid on every new release.** When a new respondent is added to the panel, the §5.3 grid search is run on it, the panel cutoff is recomputed (max over all respondents including the new one), and items whose breadcrumbs predate the new cutoff are flagged.
2. **Retire flagged items.** Such items move to an archived `retired/` partition with their original ID preserved, so historical model-generation comparisons remain reproducible.
3. **Source replacements.** New items satisfying all six locked criteria of §3.1 against the new panel cutoff floor are sourced and audited. The corpus refills to the headline count before the new respondent is reported in headline numbers.
4. **Field-balance preservation.** Replacements preferentially target fields that lost items in the retirement step, so the six-field breadth of §3.2 is preserved across releases.

The corpus state at each headline release is versioned in a `CORPUS_STATE_<date>.json`.

5.5 Three items rejected by the audit

Each of the three items below passed the parallel-true audit of §6 (the prompt is open-ended, non-leading, and admits exactly one published finding) and was rejected solely on the contamination floor. Each is from a different field.

(1) Physics and Astronomy — Olami-Feder-Christensen universality.

> *Prompt:* "Can the avalanche-size exponents of the Olami-Feder-Christensen self-organized-criticality model be shown to be universal? Why? If yes, describe in as much detail as possible. If no, describe why not."

Rejected on category 2 (conference talk). A workshop talk at a StatPhys-affiliated satellite meeting in summer 2025 presented the result with explicit exponent values and the proposed mechanism for non-universality, several months before the paper's online publication. The workshop slide deck was posted on the conference website and not indexed by Google Scholar; only direct programme search at the relevant theoretical-physics specialty meeting surfaced it.

(2) Mathematics and Theoretical CS — a quantum query complexity bound.

> *Prompt:* "What matching quantum query lower bound can be derived for the problem of detecting repeated patterns in strings, and what technique would establish it? Describe in as much detail as possible."

Rejected on category 3 (preprint blind-spot, arXiv variant). arXiv v1 was posted in October 2025 with the headline lower bound, the technique, and the matching upper-bound construction, three months before the panel cutoff. Median preprint-to-publication delay in mathematics is approximately 21 months [Wang et al. 2020] and arXiv coverage in this field is near-universal; this is the most common reason a paradigm-shift candidate from mathematics or TCS fails the audit.

(3) Social and Behavioural — sectoral-shift labour-market effects.

> *Prompt:* "From a structural life-cycle model with worker skill heterogeneity and job ladders, what does the pace of sectoral transitions imply about lifetime earnings losses for displaced workers? Describe in as much detail as possible."

Rejected on category 3 (preprint blind-spot, NBER variant). The same authors had circulated the result as an NBER working paper in September 2025 with the same headline finding and the same model specification. Economics rarely uses arXiv but uses NBER, SSRN, IZA, and CEPR working-paper series as the functional equivalent; the audit checks all four.

The shared property of these three rejections is that the prompt itself is admissible. Each was rejected because the *finding* had a public breadcrumb predating the panel cutoff floor, in a venue that title-only web search does not surface. The §5.1 categories are the locations the audit checks specifically because they are the locations where breadcrumbs actually live.

6 Parallel-True Audit Methodology

The parallel-true audit is the second core methodological contribution of this work. The motivating observation is simple: most paradigm-shifting scientific findings have *parallel-true* alternatives in adjacent sub-domains. A prompt that admits the parallel-true alternatives as valid answers would produce false positives, since models that "predict" any of several plausible-looking but wrong answers would score the same as models that predict the actual finding.

6.1 Per-item structure

Every item in the corpus has three audit fields filled in:

- `parallel_true_alternatives`: a list of the dominant alternative-true findings the prompt could admit if the scope anchor were too loose;
- `scope_anchor`: the minimum framing needed to exclude the parallel-true alternatives while still admitting the actual finding;
- `format_rationale`: justification for the chosen prompt format (Yes/No vs What/Where; see §6.3).

These fields are part of the corpus JSON, not metadata; they are load-bearing data that defines what the item actually measures.

6.2 Format selection: Yes/No vs What/Where

A second audit decision is the prompt format. We use two formats:

- **Yes/No.** The prompt frames the question in the *default* direction ("Will X behave as the textbook expects?") with a "Why? If yes, ... If no, describe what was found instead" branch. This is the sharper format when the field has a clean binary default that the paper inverts. It also produces a clean retrieval/synthesis discrimination: a model predicting "yes" is retrieving conventional wisdom; a model predicting "no" plus the correct finding is synthesizing against default.
- **What/Where.** The prompt frames the question as a wide open-ended one ("What is the structure?", "Where is the boundary?") without a clean binary default. Used when the answer space is wide enough that a Yes/No would lose information.

The split is determined item-by-item: an item takes Yes/No when its field has a clean binary default the paper inverts, and What/Where when the answer space is wide enough that a binary would lose information. Each item carries a `format_rationale` field documenting the choice.

6.3 Mode-neutral prompt design

The per-item prompt is the *only* instruction the model receives. There is no system prompt. There is no scaffolding ("CONTEXT: ... QUESTION: ... INSTRUCTIONS: ..."). There are no headings, no domain tags, no dates, no venue names. The vocabulary that biases models into recall mode or creative-prediction mode is banned. Specifically forbidden: *predict, infer, analyze, informed, knowledge, based on what you know, post-cutoff, future paper, describe what you know about.*

The motivation is to make the model's output format an *endogenous, measurable property* rather than a prompted artifact. Hedging models score lower because the judge excludes hedged ranges and listed-possibility answers; committing models score on outcome match. We do not prescribe an output format, but the judge does penalise hedging; this exposes hedge-style as a model property.

6.4 Three failure modes the audit catches

The audit catches three structurally distinct kinds of prompt failure. Each is illustrated by a real corpus item in the form *bad* → *locked* → *why*. Locked prompts are kept short to match the bare-prompt format actually used in the corpus (typical length 25–40 words including the "Why? If yes / If no" scaffold).

(1) Leading prompt — Mathematics (Berger-Coburn conjecture).

> *Bad:* "What surprising counterexample disproves the long-standing Berger-Coburn endpoint boundedness conjecture for Toeplitz operators on the Bargmann-Fock space?"
>
> *Locked:* "Can the Berger-Coburn endpoint conjecture for Toeplitz operators on the Bargmann-Fock space be proved? Why? If yes, describe in as much detail as possible. If no, describe why not."

The bad prompt's *surprising* and *disproves* tell the model the conjecture fails before asking the question. The locked prompt presents the conjecture as the question being tested; the model has to predict whether it holds.

(2) Information leakage in the prompt — Physics (2D Fermi-liquid hydrodynamic pole).

> *Bad:* "How does the hydrodynamic pole of σ(q,ω) in the tomographic 2D Fermi liquid require two independent scaling exponents instead of a single z?"
>
> *Locked:* "In a clean 2D Fermi liquid in the tomographic regime, can the hydrodynamic pole of σ(q,ω) be characterised by a single dynamical exponent? Why? If yes, describe in as much detail as possible. If no, describe why not."

The bad prompt names the answer (two independent exponents) before asking the question. The locked prompt presents the open yes/no question without revealing the answer.

(3) Parallel-true alternative answers — Mathematics (graph-minor characterisation).

> *Bad:* "What structural characterisation does every 4-connected non-planar graph with sufficiently large minimum degree admit?"
>
> *Locked:* "Can a complete structural characterisation be established that settles the Kawarabayashi-Maharry conjecture? Describe in as much detail as possible."

The bad prompt admits several alternative-true answers from graph-minor literature: a Tutte-style "contains K_5 or K_{3,3} as a minor" answer (true for any non-planar graph), a density-based characterisation invoking the high minimum degree, or a planarity-via-bounded-treewidth answer. A model that confidently predicts one of these would name a real result from adjacent literature, not the one this paper reports. The locked prompt's scope anchor *Kawarabayashi-Maharry conjecture* together with the implication subquestion (Hamilton properties, torus embeddings) constrains the answer to the specific characterisation this paper proves.

The three failure modes are independent: a prompt can leak information (mode 2) without being leading (mode 1), or admit parallel-true alternatives (mode 3) without doing either. Every prompt in the corpus is audited against all three.

7 Judging Protocol

The benchmark uses an LLM-as-judge protocol with three design constraints, each of which addresses a documented failure mode of LLM-as-judge systems: (i) no respondent is also a judge, eliminating direct self-preference [Panickssery et al. 2024]; (ii) the panel is composed of one frontier model per major lab so that residual within-family scoring sympathy is, in principle, symmetric across the

labs producing the respondents; and (iii) per-item label randomisation combined with a single-call all-responses-together prompt anchors the 0–5 scoring scale within each item rather than across items.

7.1 Why not human judges

A natural alternative to LLM judging is a human-expert panel. For a benchmark restricted to a single subfield, that would be the preferred protocol. The Singularity Gate is not such a benchmark: it spans natural and formal science across five broad fields — life sciences, chemistry and materials, physics and astronomy, earth and planetary, and mathematics and theoretical computer science. The expert-judge alternative therefore requires either (a) a panel of breadth-experts capable of scoring at the frontier across all five fields, which essentially does not exist, or (b) one specialist per field per item, which costs an order of magnitude more than the entire respondent evaluation and introduces its own coordination overhead. A "couple of smart people" generalist panel substituted for either of the above would, on the items it cannot evaluate at the frontier, contribute primarily its own opaque biases rather than any scoring signal an LLM panel does not also contain.

Three further properties make LLM judges better-suited to this specific scoring task than a small human panel would be:

1. **Reproducibility.** An LLM judge with a fixed prompt and seed produces a stable ranking that any reader can re-run from the released per-(item, model, judge) data using the arithmetic chain in §7.4. A human panel's scoring drifts as panellists fatigue, recalibrate, and learn the items; the released numbers cannot be reproduced without re-running the panel.
2. **Anchored scoring scale.** Because all responses for an item are scored in a single judge call (§7.2), the 0–5 scale is anchored within that call rather than across calls. Human panellists, judging items sequentially over hours or days, recalibrate their scale between items in ways that introduce drift not present in the single-call LLM protocol.
3. **Documented and mitigable biases.** The biases of LLM judges — position [Shi et al. 2024], verbosity [Zheng et al. 2023], self-preference [Panickssery et al. 2024], family loyalty [Wataoka et al. 2024] — are documented at fine enough resolution to be mitigated by auditable design choices: per-item label randomisation (§7.2), cross-lab non-respondent composition (§7.3), and an empirical family-bias diagnostic (§7.5). The biases of a small human panel — which include the panellists' personal relationships and prior interactions with the labs producing the respondents — are not documented at the same granularity and are difficult to mitigate by construction.

What LLM judging cannot do that human expert judging can is provide ground-truth scoring on items requiring frontier-level expertise the judge itself does not possess. The scoring metric (§7.4) and per-item judge prompt are designed around this constraint: rather than asking the judge to grade the *correctness* of a frontier-science response from scratch, we provide the published abstract as the ground-truth document and ask the judge to score each response against the abstract's specific mechanism. The judge's role is comparison-against-anchor, not expertise-from-scratch — which is what allows a frontier LLM judge to score reliably on items whose answer the judge could not have produced.

7.2 Per-item judging: all responses together, abstract as ground truth

The judging protocol gives the judge, in one call, the abstract, the question, and *all panel responses for that item*. The scoring scale is anchored within the call rather than across calls: two responses of equivalent quality, judged in separate calls, can receive different R or O integers because LLM judging is non-deterministic and the judge re-anchors its 0–5 scale to whatever it sees at inference time. Within a single call, all responses are scored against the *same* anchored scale, so any per-call strictness drift affects all respondents identically and is symmetric at the per-item level. Cross-item judge variance still exists, but it washes out under per-item averaging (§7.4) and panel averaging (§7.3).

Abstract-only ground truth (rather than full paper) is used because the combined judge prompt — abstract + question + all panel responses + rubric — already consumes a substantial fraction of the context window in which long-context attention degrades [Liu et al. 2024; Hsieh et al. 2024]. Adding a full paper on top moves the prompt into the regime where attention is unevenly distributed across the document, increasing scoring variance across responses depending on their position in the prompt; this undermines the per-item scale-anchoring purpose of the single-call design. Abstracts foreground a paper's load-bearing findings and provide a self-contained summary, so abstract-only ground truth preserves judge accuracy at low scope cost.

Per item, the response labels (A, B, C, D, E for 5-way) are assigned to model identities by random permutation under a deterministic per-round seed (e.g., 20260508). The judge sees the labelled responses but does not see which model produced which response. Per-item label randomisation reduces position bias and contrast effects in N-way comparison [Shi et al. 2024]: a response that looks weak relative to the other four under one labelling can look stronger under a different labelling. Label mappings are saved per round; unblinding is only performed after scoring.

The judge prompt itself went through seven iterations driven by specific failure modes observed in the previous version (initial 4-element rubric; pole-anchored 0–5; removed FINAL COMMITMENT requirement; added anti-parallel-true clause; finally an anchored-Reasoning clause that scores Reasoning Quality on whether the response's reasoning *anticipates the abstract's mechanism* rather than on generic competence). The v6→v7 swap dropped mean R scores by ~2 points across the calibration dataset (Sonnet 4.6: 4.18→2.18; Opus 4.7: 4.60→2.70; Opus 4.6: 4.14→2.23) while Outcome scores barely moved, confirming that v6 had been inflating R by crediting generic competence. The full v7 prompt is reproduced in Appendix B.

7.3 Cross-lab non-respondent panel

LLM-as-judge protocols admit a structural bias when a respondent is also a judge: Panickssery et al. [2024] show that LLM evaluators recognise and favour their own generations, with self-preference scaling linearly with self-recognition accuracy; Wataoka et al. [2024] quantify the effect across multiple judge-generator pairs; Zheng et al. [2023] name "self-enhancement" as one of the canonical biases of LLM-as-judge. The Singularity Gate eliminates direct self-preference by construction (no respondent is a judge) and addresses residual family-loyalty bias through panel composition rather than averaging it out post-hoc.

The panel was initially designed with three judges from three different labs, each non-respondent:

- **Opus 4.5 (R1)** (Anthropic; non-respondent. The Anthropic respondents are Opus 4.7 and Opus 4.6.)

- **GPT-5.4** (OpenAI; non-respondent. The OpenAI respondent is GPT-5.5.)
- **Gemini 3 Flash** (Google; non-respondent. The Google respondent is Gemini 3.1 Pro.)

Each is a deployed frontier model from its respective lab and remains at the capability frontier of LLM-as-judge work in the published literature. Two design properties follow:

1. **No respondent is a judge.** Every direct self-preference path is eliminated by construction.
2. **Each respondent's lab is represented by exactly one family-cousin judge** (Anthropic $\leftrightarrow$ Opus 4.5, OpenAI $\leftrightarrow$ GPT-5.4, Google $\leftrightarrow$ Gemini 3 Flash). Whatever residual within-family scoring sympathy exists is, in principle, symmetric across the three labs.

After running all three judges on the full corpus, a per-judge family-bias diagnostic (§7.5) revealed that the symmetry argument fails empirically for one of the three judges. The headline panel therefore uses two of the three judges; the full three-judge results are reported as a sensitivity check in §7.6.

7.4 Scoring metric: outcome backed by reasoning

The benchmark targets a single quantity per response: *outcome backed by reasoning*. A response is credited only to the extent that it predicts the abstract's finding (the *outcome*) AND shows the reasoning path that anticipates the abstract's mechanism. Neither half alone is acceptable. The scoring metric is the formal expression of that requirement.

Per item, the judge produces two integer scores Reasoning $R \in \{0, 1, \ldots, R_{\max}\}$ and Outcome $O \in \{0, 1, \ldots, O_{\max}\}$, with $R_{\max} = O_{\max} = 5$. The per-item *score* is the product $R \times O$ passed through a linear normalisation onto the percentage scale, so that a perfect response ($R = O = R_{\max} = O_{\max} = 5$) receives the maximum of 100%:

$$\text{score}(i) \;=\; \text{norm}(R_i \times O_i)\,, \qquad \text{norm}(x) \;=\; \frac{100 \cdot x}{R_{\max} \cdot O_{\max}}$$

The benchmark score for a model is the simple mean of score(i) over all items the model answered. A model that produces $R = O = R_{\max}$ on every item scores exactly 100%; a model that produces $R = O = 0$ on every item scores 0%. The full theoretical range $[0, 100]$ is achievable in principle, though in practice frontier-model scores cluster in the 10–25% range on the current corpus. This range is consistent with frontier-knowledge benchmarks at their release: Humanity's Last Exam SOTA at January 2025 release was 10–15% [Phan et al. 2025], and FrontierMath SOTA at November 2024 release was below 2% [Glazer et al. 2024]. Paradigm-shift prediction is *hard*.

norm($\cdot$) is the standard percent-of-maximum normalisation: it is monotone increasing on $[0, R_{\max} \cdot O_{\max}]$, maps $0 \to 0$ and $R_{\max} \cdot O_{\max} \to 100$, and is otherwise structurally inert (it does not change which of two responses scores higher). All of the formula's behavioural content lives in the product $R \times O$; the normalisation only fixes the reporting scale.

From raw judge integers to the headline. A reader who wishes to reproduce the §9 numbers from the per-(item, model, judge) raw integers (released as JSON in the data download) applies the following arithmetic chain. Let J be the panel of judges, n the number of items in the headline corpus, and $R_{i,m,j}, O_{i,m,j}$ the raw integer scores returned by judge j for model m on item i.

1. **Per-(item, model, judge) score.** Apply the normalisation above:

$$s_{i,m,j} \;=\; \mathrm{norm}(R_{i,m,j} \times O_{i,m,j}).$$

2. **Per-(item, model) panel cell.** Take the unweighted mean across the judges in J:

$$\bar{s}_{i,m} \;=\; \frac{1}{|J|} \sum_{j \in J} s_{i,m,j}.$$

3. **Per-model headline score.** Take the unweighted mean across the n headline items:

$$\mathrm{Score}_m \;=\; \frac{1}{n} \sum_{i=1}^{n} \bar{s}_{i,m}.$$

4. **Standard error and confidence interval.** The ± SE reported in the §9.1 table is the standard error of the mean of the per-item panel cells:

$$\mathrm{SE}_m \;=\; \mathrm{sd}_i(\bar{s}_{i,m}) \,/ \sqrt{n}.$$

   Error bars in the figures throughout this paper (and on the public leaderboard) show the **50% confidence interval of the mean**, $\pm\, z_{0.75} \cdot \mathrm{SE}_m \;=\; \pm\, 0.6745 \cdot \mathrm{SE}_m$. This is the IQR-equivalent two-sided coverage interval for a normal sampling distribution: tighter than the conventional $\pm 1\,\mathrm{SE}$ ($\approx 68\%$ CI), but still a real, named interval — chosen so that the visual width of the bars is not larger than half of the score-range spread between adjacent respondents on this corpus.

For the headline panel of §7.5, $J = \{\text{Opus 4.5 (R1), GPT-5.4}\}$, and $n = 104$. For the 3-judge sensitivity check of §7.6, the same chain is applied with $J = \{\text{Opus 4.5 (R1), GPT-5.4, Gemini 3 Flash}\}$. Per-field rows of §9.3 replace step 3's sum over n items with a sum over the items in each field. The chain is arithmetic; there is no fitting, no per-item weighting, and no per-item normalisation beyond the per-cell norm(·) already defined.

Why Reasoning is in the formula at all. This is the central design choice of the scoring metric, and the answer is that *Outcome alone is not safe to credit*. A response that names the abstract's finding can have arrived at it through any of three paths:

1. **Synthesis.** The model reasoned over priors and anticipated the mechanism. This is the target capability the benchmark is built to measure.
2. **Retrieval.** The model has the finding effectively in its training data via paraphrase, summary-of-summary diffusion in the indexed web, or any near-paraphrase of the abstract that the model encountered without the §5 audit recognising it as the *specific* paper.
3. **Lucky guess.** On items with a small effective answer space (e.g., a Yes/No item where one direction matches the paper), a model with no relevant reasoning can land on the correct answer at a non-trivial probability by chance alone.

Outcome cannot distinguish these three. Reasoning can. The judge is instructed (v7) to score Reasoning on whether the response's reasoning *anticipates the abstract's mechanism specifically*, and not on generic competence and not on whether it merely matches the final answer. A retrieved or guessed Outcome-5 response will typically not produce reasoning that closely tracks the abstract's mechanism, because the work isn't there; the response didn't actually go through the work. A

synthesised Outcome-5 response will. Multiplying by R therefore *zeroes out* the credit for any item where the outcome appears to have been reached without relevant reasoning (R=0 ⇒ score 0 regardless of O), and proportionally discounts items where the reasoning is partial.

This is the corpus-internal contamination guard, complementing the corpus-construction-level guards of §5: even if a contaminated item slips through the cutoff grid search and the seven-blind-spot audit, a retrieval-style response on that item earns near-zero from the R × O formula because R will be low. A model that scores well on this benchmark cannot do so by retrieval alone; it must show the reasoning path.

Why product, not sum. `R + O` is bounded above by 10 and scales linearly. `R × O` is bounded above by 25 and is supermodular: improvement from (R=2, O=3) to (R=3, O=4) gains 6 points (12 to 18) of the multiplicative metric versus 2 points of the additive metric. The supermodularity matters because of the contamination-guard logic above: an additive metric would still hand out half-credit to a high-O / low-R retrieval response (5 + 0 = 5), whereas the product zeroes it (5 × 0 = 0). The product is the formal expression of "neither half alone is acceptable; the model must do both."

Why per-item average, not aggregate. A common alternative would be `mean(R) × mean(O) / 25`. This is *different* from per-item averaging (by Jensen's inequality) and is wrong here. R and O are positively correlated within items: a response that reasons well typically also lands on the right answer. The per-item form preserves that covariance as legitimate signal; the aggregate form discards it. Worse, the aggregate form would partially undo the contamination guard: a model with high mean-O from a few retrieved items and high mean-R from a few different synthesis-attempt items would score similarly to a model that actually produced both R and O on the *same* items. Per-item multiplication enforces the "both, on the same item" requirement.

Diagnostic value of separating R and O. Mismatches between R and O expose two distinct failure modes that a pure outcome score would conflate:

- *Low R, high O*: retrieval / lucky-guess signature. The model lands on the answer but cannot show how (the bare-fact-recall fingerprint of contaminated training data, or a fortunate guess on a low-entropy item).
- *High R, low O*: synthesis-without-landing. The model reasoned well from priors but landed on a parallel-true alternative rather than the actual finding.
- *Low R, low O*: no useful signal.
- *High R, high O*: the target capability, synthesis that lands.

7.5 Family-bias diagnostic and the 2-judge decision

For each judge, we compute the **per-cell family advantage**: the difference between (mean R × O across all Claude-family cells judged) and (mean R × O across all non-Claude-family cells judged). A judge with no family preference scores both families equally on aggregate; a judge with family loyalty scores its own family higher per cell.

On the headline corpus the per-judge family advantages are:

Judge	Claude mean R×O	non-Claude mean R×O	**Claude advantage**
Opus 4.5 (Anthropic-family judge)	4.50	4.06	**+0.44**
GPT-5.4 (GPT-family judge)	3.26	3.57	**−0.31**

Judge	Claude mean R×O	non-Claude mean R×O	**Claude advantage**
Gemini 3 Flash (Google-family judge)	5.39	3.61	**+1.78**

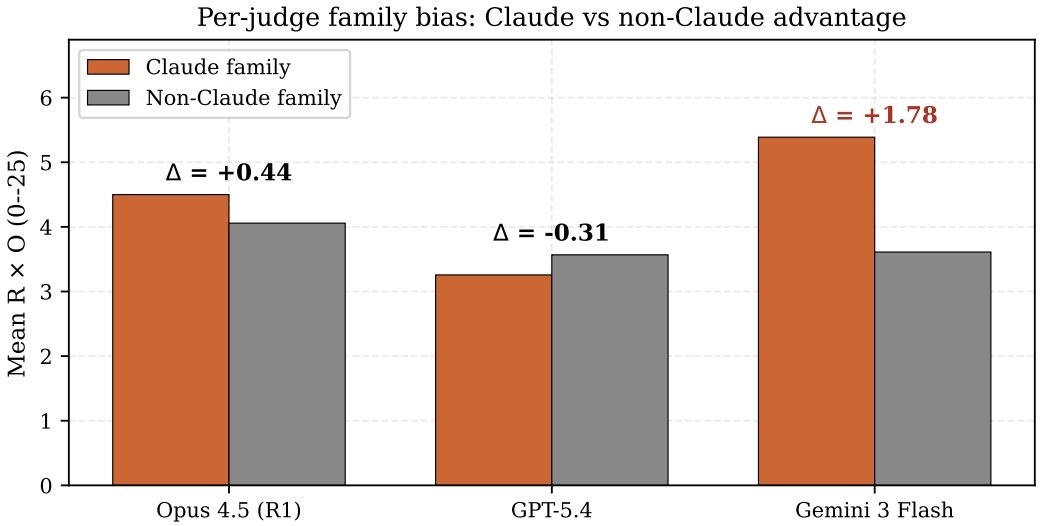

Figure 1: Per-judge family-bias diagnostic on the n=104 corpus. Each judge's mean $R \times O$ on Claude-family cells (orange) vs. non-Claude-family cells (grey). The same-family tilt is mild and comparable for Opus 4.5 and GPT-5.4; Gemini 3 Flash exhibits a pro-Claude advantage about 4× the magnitude and in the opposite direction from within-family loyalty, motivating its exclusion from the headline panel.

Opus 4.5's tilt toward its own family (+0.44) is within the magnitude that within-family scoring sympathy plausibly produces, and the same-direction tilt of comparable magnitude appears in GPT-5.4 (toward GPT models: +0.31 reading the sign as pro-family). Gemini 3 Flash exhibits a Claude-family advantage of **+1.78** — approximately four times the magnitude of the same-family judge's own tilt, and in the opposite direction from what within-family loyalty would predict (a pro-Gemini judge should favour Gemini-family responses, not Claude-family responses). Per-(judge, model) decomposition shows the effect concentrates uniformly on the three Claude respondents (per-cell inflation +1.07 to +1.52) with a much weaker effect on the same-family respondent Gemini 3.1 Pro (+0.56) and a negative effect on GPT-5.5 (−1.16).

A magnitude and direction of this kind is not consistent with the within-family scoring sympathy that the panel-composition symmetry was designed to absorb. The most parsimonious account is training-data exposure to Claude-graded or Claude-generated content during Gemini 3 Flash's training pipeline; we report the diagnostic without committing to causal mechanism.

Why a 3-judge mean does not fix this. With two of three judges (Opus 4.5, Gemini 3 Flash) tilted in the same direction at unequal magnitudes, the panel mean's net family tilt is $(+0.44 - 0.31 + 1.78)/3 = \mathbf{+0.64}$ per-cell R × O in favour of the Claude family. Standard aggregation methods (mean, median, Borda count, family-bias calibration relative to panel mean) all preserve this asymmetry, because all of them assume the panel composition is family-balanced; with the diagnostic above, it is not.

Why dropping Gemini 3 Flash is the principled fix. Removing the strongly-tilted judge restores the intended panel symmetry: one Anthropic-family judge with mild Claude tilt (+0.44), one GPT-family judge with mild GPT tilt (read off the non-Claude advantage). The net family tilt

of the 2-judge mean is $(+0.44 - 0.31)/2 = \mathbf{+0.065}$ — a tenfold reduction in residual family bias. The remaining per-judge biases are of comparable magnitude and point in opposite directions, so they approximately cancel at the per-cell level for cross-family comparisons.

Headline aggregation. We use the 2-judge mean of Opus 4.5 + GPT-5.4 as the headline *score* for each (respondent, item) cell, with per-cell averaging followed by per-model averaging across the headline corpus.

7.6 3-judge sensitivity check

We report the full 3-judge score (R1 + GPT-5.4 + Gemini 3 Flash, per-cell mean) as a sensitivity check, both for transparency and to verify that the decision to drop Gemini 3 Flash does not selectively penalise its family-cousin respondent Gemini 3.1 Pro:

Rank	Model	3-judge mean score	2-judge mean score	Δ (2-judge − 3-judge)
—	Claude Opus 4.7 (max)	20.60	17.75	−2.85
—	Claude Opus 4.6 (max)	16.54	15.11	−1.43
—	Claude Sonnet 4.6 (max)	15.44	13.67	−1.77
—	Gemini 3.1 Pro (high)	15.17	14.42	**−0.75**
—	GPT-5.5 (xhigh)	14.80	16.08	+1.28

Dropping Gemini 3 Flash from the panel reduces the absolute score of every Claude respondent by 1.4–2.9 points (Gemini 3 Flash was contributing a per-cell pro-Claude inflation of +1.07 to +1.52 to those scores). The reduction for Gemini 3.1 Pro is the **smallest in the panel** (−0.75), and GPT-5.5 is the only respondent whose score increases (+1.28, reflecting that Gemini 3 Flash's per-cell scoring of GPT-5.5 was −1.16 below the other two judges'). Gemini 3.1 Pro's rank position (#4) is unchanged across both aggregations. Removing Gemini 3 Flash therefore removes a Claude-favouring distortion from the panel without penalising the same-family Google respondent in any meaningful way.

The full per-(judge, model) score breakdown is reproduced in §9.2 as part of the robustness analysis.

8 Harness and Tool Use

8.1 Native-harness evaluation

Each respondent is evaluated in its lab's own native agentic harness (the configuration in which the model is realistically deployed for serious work) rather than as a bare API completion endpoint:

Model	Harness	Reasoning effort
Claude Opus 4.7	**Claude Code**	max
Claude Opus 4.6	**Claude Code**	max
Claude Sonnet 4.6	**Claude Code**	max

Model	Harness	Reasoning effort
GPT-5.5	**Codex**	xhigh
Gemini 3.1 Pro	**Gemini CLI**	high

Bare-LLM evaluation systematically understates capability that is realised through the harness: the planning loop, the iterative refinement, and the tool-mediated workspace that real-world deployment provides. Conversely, agentic-harness evaluation with web search enabled would systematically overstate capability by admitting retrieval. We aim to measure the model *as it is used*, not the model in either of the two artificially simplified configurations. Reasoning-effort is set per respondent at the highest documented setting offered by each lab's harness, so the headline reflects each model's deployed peak rather than a calibration-pruned default.

8.2 Tool use enabled, web search disabled

Within each harness, **tool use is enabled** (code execution, structured-output formatting, the agent's own scratchpad operations) but **web search and fetch are disabled** at the harness level. The respondent is free to think with code, write intermediate notes, run sanity calculations, and refine its own draft, all of which a working scientist might do. It is not free to retrieve the answer.

Web-search disabling is enforced by harness configuration, not by trust. For Claude Code we configure subagent dispatches with an explicit `tools:` whitelist that excludes WebSearch and WebFetch; for Codex and Gemini CLI we run with the documented "no internet" mode. Each harness is also probed with a diagnostic prompt asking the model to attempt a web search and report the result; in all cases the model correctly reports that no web tool is available.

The headline scores in §9 are therefore *deployment-realistic capability scores*: what each lab's deployed product produces when given these questions, with the strict constraint that the answer cannot be looked up. The benchmark measures the lab's deployed system, which is the unit of capability that real users encounter; bare-API completion-only evaluations would systematically understate that capability by stripping out the planning loop and tool-mediated refinement that the deployed harness provides.

9 Results

The headline numbers below reflect the full 5-way evaluation on the n=104 headline corpus, judged by the 2-judge panel (Opus 4.5 R1 + GPT-5.4) selected per §7.5. The full per-(judge, model) data and the 3-judge sensitivity check are in §7.6.

9.1 Headline ranking (2-judge mean score)

Rank	Model	Mean R	Mean O	Score	± SE
1	Claude Opus 4.7 (max)	2.17	1.66	**17.75%**	1.51
2	GPT-5.5 (xhigh)	2.04	1.57	16.08%	1.48
3	Claude Opus 4.6 (max)	1.95	1.53	15.11%	1.38
4	Gemini 3.1 Pro (high)	1.87	1.52	14.42%	1.29

Rank	Model	Mean R	Mean O	Score	± SE
5	Claude Sonnet 4.6 (max)	1.83	1.43	13.67%	1.24

Opus 4.7 places first by 1.67 score points (≈0.79σ on the conservative independent-SE assumption); the remaining four respondents cluster within 2.4 points of each other.

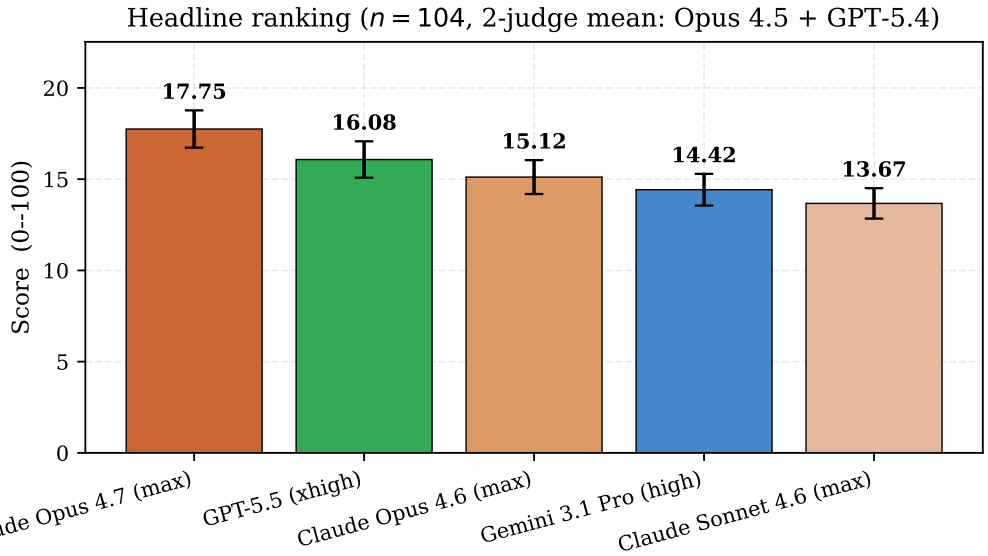

Figure 2: Headline ranking on the n=104 corpus, 2-judge mean score (Opus 4.5 R1 + GPT-5.4). Error bars: 50% confidence interval of the mean ($\pm 0.6745 \cdot$ SE, the IQR-equivalent two-sided interval for a normal sampling distribution).

GPT-5.5's position at #2 is partially attributable to a residual self-preference effect from GPT-5.4 on GPT-5.5-family cells (concentrated +0.72 per-cell, vs Opus 4.5's pro-Claude tilt of +0.44 distributed evenly across three Claude cells per item); the panel cancellation argument of §7.5 holds at the family level but does not perfectly cancel the per-model concentration of bias. The 3-judge sensitivity check in §7.6 places GPT-5.5 last under the alternative composition. **No respondent achieved a fully-correct prediction (R = O = R_max = 5) on any item in the corpus; reported scores reflect partial credit only.**

9.2 Robustness across panels

The full per-judge per-model scores and the two panel aggregations are reproduced here from §7.6 for direct comparison:

Model	Opus 4.5	GPT-5.4	Gemini 3 Flash	**2-judge mean**	3-judge mean
Opus 4.7 (max)	20.35	15.15	23.84	**17.75**	20.60
Opus 4.6 (max)	17.58	12.65	19.38	**15.11**	16.54
Sonnet 4.6 (max)	16.08	11.27	18.96	**13.67**	15.44

Model		Opus 4.5	GPT-5.4	Gemini 3 Flash	**2-judge mean**	3-judge mean
Gemini 3.1 Pro (high)	16.85	12.00	16.65	**14.42**	15.17	
GPT-5.5 (xhigh)	15.62	16.54	12.23	**16.08**	14.80	

Stable across both panels. Opus 4.7 is first in both aggregations; Gemini 3.1 Pro is fourth in both. The Opus 4.7 lead over the next-ranked respondent is 1.67 score points under the 2-judge headline and 4.06 score points under the 3-judge view (the latter inflated by Gemini 3 Flash's Claude tilt).

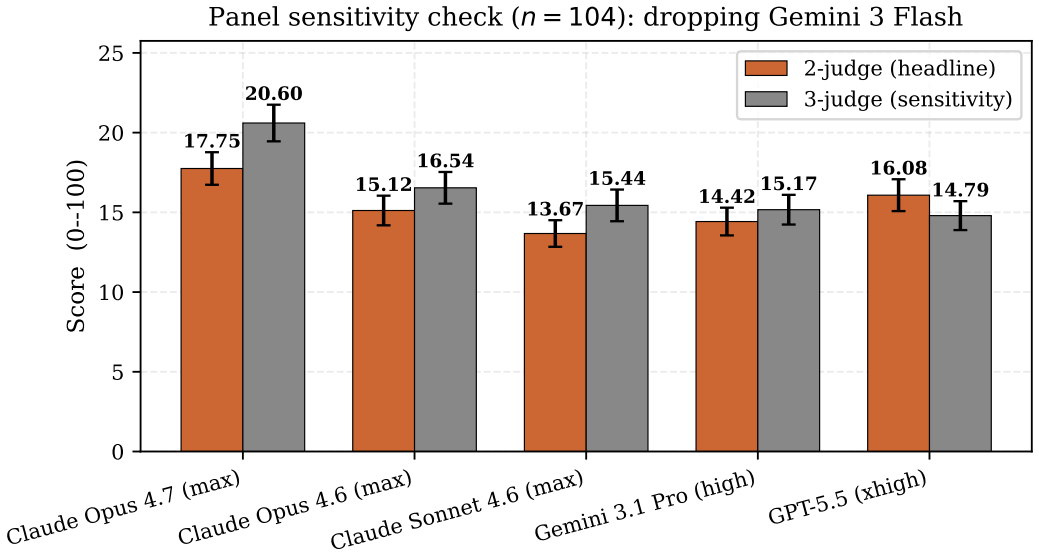

Figure 3: Panel sensitivity check: per-model score under the 2-judge headline panel (Opus 4.5 + GPT-5.4) versus the 3-judge panel that adds Gemini 3 Flash. Dropping Gemini 3 Flash reduces all Claude scores by 1.4–2.9 points, reduces Gemini 3.1 Pro's by only 0.75 (smallest in the panel), and *raises* GPT-5.5's by 1.28, leaving rank #1 (Opus 4.7) and rank #4 (Gemini 3.1 Pro) unchanged. Error bars: 50% confidence interval of the mean (same convention as Figure 2).

Position-2 swing. GPT-5.5 (2-judge: #2) and Sonnet 4.6 / Opus 4.6 (3-judge: #2-#3) trade positions across panels. Under each individual non-family-cousin judge, GPT-5.5 ranks #5 (Opus 4.5: #5) or #1 (GPT-5.4: #1 with concentrated self-preference); its 2-judge-mean position #2 falls between these two judge-specific extremes. The 3-judge mean places it at #5 only because Gemini 3 Flash's pro-Claude tilt suppresses its scores by an additional 1.28 points per cell. Both aggregations are arguable; we report the 2-judge mean as headline because its family-bias balance is empirically tenfold better than the 3-judge mean's (§7.5), and we report the 3-judge mean as the sensitivity check because it confirms the leader (Opus 4.7) and the Gemini-family position is unchanged.

9.3 Per-field breakdown

The headline corpus is aggregated into the five broad fields defined in §3.2. We report the per-model score within each field; per-field sample sizes are shown alongside.

Field (n)	Opus 4.7 (max)	Opus 4.6 (max)	Sonnet 4.6 (max)	Gemini 3.1 Pro (high)	GPT-5.5 (xhigh)
Life Sciences (n=30)	**17.53**	15.93	15.60	15.80	16.40
Chemistry and Materials (n=20)	**18.00**	15.70	13.60	14.20	13.70
Physics and Astronomy (n=18)	17.56	17.78	14.56	16.44	**25.22**
Earth and Planetary (n=19)	**21.26**	14.63	13.05	14.53	13.68
Mathematics and TCS (n=17)	**14.12**	10.71	10.12	10.00	11.29
Overall (n=104)	**17.75**	**15.11**	**13.67**	**14.42**	**16.08**

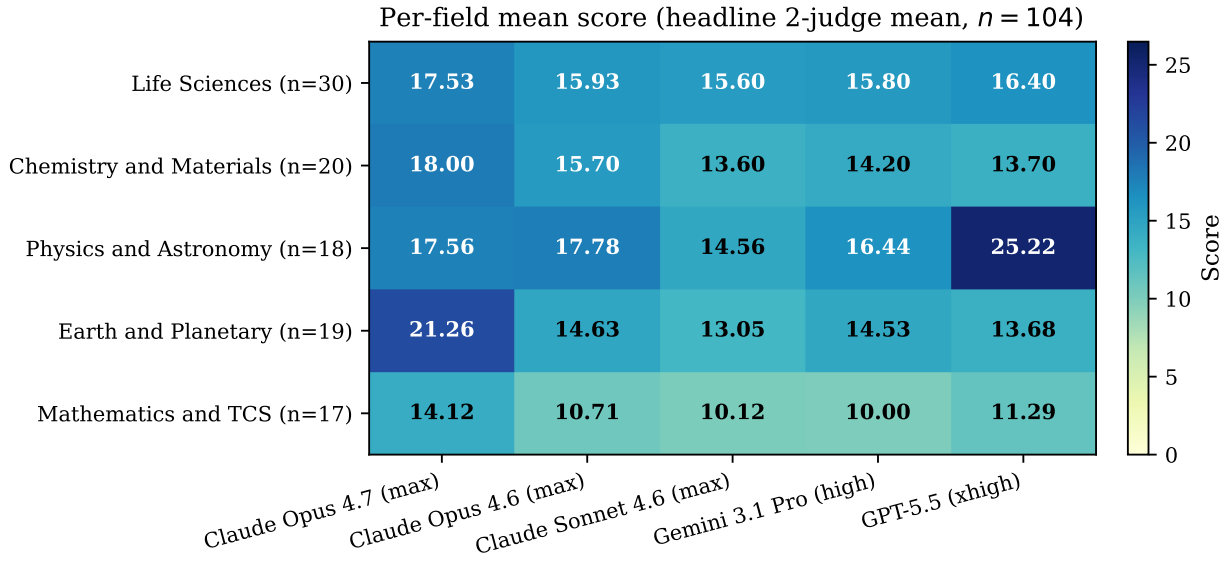

Figure 4: Per-field per-model mean score on the n=104 headline corpus, 2-judge panel. Per-field sample counts in row labels. Opus 4.7 leads in 4 of 6 fields; GPT-5.5's lead in Physics and Astronomy is the largest single-field gap.

Per-field cells reflect the headline 2-judge mean score within each field; overall row repeats the headline from §9.1. **Bold** marks the leader per row. The figure below shows the same data as a heatmap.

What the per-field gradient tells us. Three diagnostic questions are answered by the per-field table:

1. **Does the headline ranking hold within each field?** Opus 4.7 leads in four of the five fields (Life Sciences, Chemistry and Materials, Earth and Planetary, Mathematics and

TCS). In Physics and Astronomy GPT-5.5 leads by a wide margin (25.22 vs Opus 4.7's 17.56). The overall headline ranking therefore reflects genuine cross-field capability rather than concentration in one or two fields.

2. **Where do models *cluster* vs. *separate*?** Life Sciences shows the tightest cluster (5-model spread of 1.93 score points), consistent with widespread biology training-data coverage that lifts all models. Earth and Planetary and Physics and Astronomy show the widest spreads ($\geq$8 points between best and worst); these are the most discriminative fields. Mathematics and TCS produces uniformly low scores (5-model mean $\approx$ 11.2) consistent with how hard frontier mathematics is to derive from priors.
3. **Is the lead concentrated where training data was strongest?** The Opus 4.7 advantage is broad rather than concentrated: it leads in four of five fields, and its single field of clear weakness (Physics and Astronomy) is the one where GPT-5.5 has a marked advantage rather than a field where Opus 4.7 itself collapses.

Per-field sample sizes. All five fields (Life Sciences, Chemistry and Materials, Physics and Astronomy, Earth and Planetary, Mathematics and TCS) carry enough items (n=17 to n=30) for per-field rankings under the cross-judge variance estimated in §9.2. The refresh policy of §5.4 prioritises replenishment of the smaller fields when items retire.

9.4 Reproducibility check

GPT-5.5 was collected twice independently on a reproducibility subset under the headline conditions of §8. The two runs scored within 1–2 pp of each other across the panel cross-validation, confirming that the reported score reflects a stable distributional property of the model rather than a one-off draw.

10 Discussion

(To draft. Cover: - What the gradient across fields (§9.3) says about retrieval vs. synthesis: a model uniformly best across all five fields is doing something different from a model whose lead concentrates in fields where its training data was strongest. - How the per-field results constrain singularity-readiness claims: a model uniformly best across all five fields is more compelling evidence than a model whose lead concentrates in fields where its training data was strongest. - Why the cross-lab non-respondent panel design in §7 should become the standard configuration for LLM-as-judge benchmarks where ranking integrity is required. - The specific ways in which this benchmark is *not* the singularity gate but only *a* gate: the further capabilities (experimental design, instrument construction, multi-year project planning, paradigm articulation, literature curation) that an AI-driven autonomous-discovery pipeline would require, and what would be needed to instrument them.)

Appendix A: Per-item parallel-true audits

(Selected worked examples — statphys_001 (OFC universality), math_r2_004 (Berger-Coburn conjecture), math_r2_005 (graph-minor characterisation), tcs_r2_001 (quantum query complexity), econ_001 (sectoral transitions) — with full `parallel_true_alternatives`, `scope_anchor`, `format_rationale` fields shown.)

Appendix B: Judge prompt v7 (full text)

(Full v7 prompt template.)

Appendix C: Item curation provenance

(Per-item: source venue, polish round (v0/v1/v2/manual), final reviewer status, contamination-audit notes.)

References

Bavaresco, A., Bernardi, R., Bertolazzi, L., et al. (2024). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks (Judge-Bench). *arXiv preprint* arXiv:2406.18403.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint* arXiv:1911.01547.

Chollet, F., et al. (2025). ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint* arXiv:2505.11831.

Glazer, E., Erdil, E., Besiroglu, T., et al. (2024). FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint* arXiv:2411.04872.

Google DeepMind. (2025, December 17). *Introducing Gemini 3 Flash* [Blog post / model card]. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/ ; https://deepmind.google/models/gemini/flash/.

Gu, J., Jiang, X., Lin, Z., Cao, R., Yu, S., Zheng, Z., et al. (2024). A survey on LLM-as-a-judge. *arXiv preprint* arXiv:2411.15594.

Hassabis, D. (2025, July 23). Conversation with Lex Fridman, Lex Fridman Podcast #475: *Demis Hassabis: Future of AI, Simulating Reality, Physics and Video Games*. Transcript: https://lexfridman.com/demis-hassabis-2-transcript/. ("Have a cut-off of 1900 and then give the system everything that was written up to 1900 and then see if it could come up with special relativity and general relativity, right? Like Einstein did. That would be an interesting test.")

Hassabis, D. (2026, February 17). Address at the Indian Institute of Science, Bangalore / India AI Impact Summit 2026. Reported at https://officechai.com/ai/a-test-of-agi-could-be-if-a-system-trained-till-1911-data-could-discover-general-relativity-google-deepmind-ceo-demis-hassabis/. ("The kind of test I would be looking for is training an AI system with a knowledge cutoff of, say, 1911, and then seeing if it could come up with general relativity, like Einstein did in 1915. ... It's clear today's systems couldn't do that.")

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding (MMLU). In *Proc. International Conference on Learning Representations (ICLR)*. arXiv:2009.03300.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., & Ginsburg, B. (2024). RULER: What's the real context size of your long-context language models? *arXiv preprint* arXiv:2404.06654.

Li, D., Sun, Z., Tan, X., Wang, Y., Liu, J., Zhang, Y., et al. (2024). LLMs-as-judges: A

comprehensive survey on LLM-based evaluation methods. *arXiv preprint* arXiv:2412.05579.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. arXiv:2307.03172.

Magar, I., & Schwartz, R. (2022). Data contamination: From memorization to exploitation. In *Proc. ACL 2022*. arXiv:2203.08242.

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., & Hashimoto, T. (2024). Proving test set contamination in black-box language models. In *Proc. ICLR 2024*. arXiv:2310.17623.

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Proc. NeurIPS 2024*. arXiv:2404.13076.

Phan, L., Gatti, A., Han, Z., et al. (2025). Humanity's Last Exam. *Nature* (2025); arXiv:2501.14249. https://www.nature.com/articles/s41586-025-09962-4.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint* arXiv:2311.12022.

Roberts, M., Demir, O., Sabin, A., et al. (2024). To the cutoff... and beyond? A longitudinal perspective on LLM data contamination. In *Proc. ICLR 2024*. https://openreview.net/forum?id=m2NVG4Htxs.

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP 2023*. arXiv:2310.18018.

Shi, L., Sun, Y., Xiong, K., & Yu, P. S. (2024). Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint* arXiv:2406.07791.

Srivastava, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (BIG-Bench). *Transactions on Machine Learning Research*. arXiv:2206.04615.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). Challenging BIG-Bench tasks and whether chain-of-thought can solve them (BBH). *arXiv preprint* arXiv:2210.09261.

Tan, S., Ding, J., Sun, Z., Wang, J., Liu, Y., et al. (2025). JudgeBench: A benchmark for evaluating LLM-based judges. In *Proc. ICLR 2025*. arXiv:2410.12784.

Wang, Y., Wang, R., et al. (2020). Preprints as accelerator of scholarly communication: An empirical analysis in mathematics. *Journal of Informetrics*; arXiv:2011.11940.

Wataoka, K., Takahashi, M., et al. (2024). Self-preference bias in LLM-as-a-judge. *arXiv preprint* arXiv:2410.21819.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proc. NeurIPS 2023 Datasets and Benchmarks Track*. arXiv:2306.05685.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., & Duan, N. (2023). AGIEval: A human-centric benchmark for evaluating foundation models. *arXiv preprint* arXiv:2304.06364.

Zhu, Q., Pei, J., Hu, Z., et al. (2024). MMLU-CF: A contamination-free multi-task language

understanding benchmark. *arXiv preprint* arXiv:2412.15194.