Coelacanth: A Benchmark of Cross-Field Rediscoveries for Evaluating Novelty and Precedent Checkers

Damir Iskakov

Independent researcher, Almaty, Kazakhstan
iskakov.dam@gmail.com

Abstract. Tools that check whether a scientific idea has been done before are now public, but we found no published measurement of how often they find earlier work stated in another field's vocabulary. We present Coelacanth v0, a benchmark for such measurements: 60 items (49 with adjudicated labels), 10 of them documented cross-field rediscoveries (`R`), plus attributed controls, AI-generated methods and 15 negatives. Systems see only a field-native claim and its date; each claim-candidate pair carries one of six typed relations. People chose every label, from three options a model proposed: five paid annotators answered 272 quiz questions (two readings of each of 136 pairs) in the author's presence, and the author settled all 51 disagreements and adjudicated the model-named works himself, 716 recorded human decisions in all; agreement is below the pre-registered thresholds. Models drafted the claim charts and quote pairs (35 of 135 candidates have no quote pair) and probed memory: nine of the ten documented rediscoveries are recalled closed-book by some of three Claude models, and for the tenth one model's safeguards declined every sample, so none qualifies as a closed-book failure (the tenth would be one if the declines were ignored) and a retrieval score on documented rediscoveries cannot be separated from memory. In a pilot scored against adjudicated gold (lower bounds), a web-search agent without and with a prior-art skill finds a cross-field gold work in the top 10 for 3 and 5 of the 10 items, a no-tools Sonnet 5.5 names one for 6, and both agents label almost every negative as related prior work. We also reproduce a published twin-retrieval baseline exactly. Code (MIT): https://github.com/IskakovDamir/coelacanth-benchmark; data (CC-BY-4.0, encrypted) and run logs: https://doi.org/10.17605/OSF.IO/672RJ.

Keywords: cross-field rediscovery · novelty checking · benchmark · memorization · closed-book probing

1 Introduction

In 1994 a paper in *Diabetes Care* proposed a model for the area under glucose tolerance curves [27]; a letter the same year was titled "Tai's formula is the trapezoidal rule" [18]. The paper's wording contains no term from numerical

analysis, so a search in its own vocabulary has nothing to match the rule on. We use this case as worked example E1; it is excluded from the benchmark.

Several tools now answer some form of "has anyone done this before?" [32, 24, 10, 5, 12, 3]. We found no published measurement of how often any of them retrieves earlier work stated in another field's vocabulary (checked again on 12 October 2026).

Existing novelty benchmarks cannot supply that measurement: they take ground truth from reviewers' overlap claims, survey co-citations, the papers a submission cites or publication order, or use no labels and manipulate a reference pool [31, 29, 23, 16, 17], so a cross-field twin that shares no citation with the claim cannot enter them. The closest resource, Kulikowski's 210 cross-field twin pairs [15], is a retrieval benchmark without verdict labels or memorization control.

Memorization is the second difficulty. A model may name the earlier work from memory, and in v0 it nearly always does: every documented cross-field rediscovery that could be probed is recalled by at least one of three models (Section 5). A benchmark for these tools therefore needs closed-book recall measured per item and results that report recalled and unrecalled items separately.

This paper describes Coelacanth v0 and what it measures:

- **(C1) The resource, with every label chosen by people:** field-native claims, typed pair labels chosen by paid annotators (272 quiz answers, two readings of each of 136 claim-candidate pairs) and adjudicated by the author, 716 recorded human decisions in all (Table 4) with their measured agreement, Claude-drafted claim charts with verbatim-checked quote pairs, and closed-book recall per item, model and prompt (Section 3).
- **(C2) Closed-book recall:** 9 of the 10 documented cross-field rediscoveries in `R` are recalled by some model, and none is a closed-book failure (Section 5).
- **(C3) A pilot measurement** of how often general-purpose checkers retrieve the cross-field earlier work and give the right typed verdict, scored against adjudicated gold as lower bounds (Section 7).
- **(C4) An exact reproduction** of Kulikowski's fingerprint TF-IDF rows and a closed-pool protocol for open-search systems (Section 6).
- **(C5) The release:** code, protocol and datasheet in a public repository; data and run logs in a password-protected archive (Section 10).

2 Related Work

Several published tools answer parts of "has this been done before?"; Table 1 sets the benchmark beside the closest resources and checkers.

Twin retrieval. Kulikowski strips field vocabulary into mechanism skeletons and retrieves cross-field twins with TF-IDF; his benchmark holds 109 papers, 5,812 cross-field pairs and 210 twins, with a ranked output, no verdict types and no closed-book control [15]. We reproduce his numbers (Section 6) and add typed labels, open-literature search and closed-book recall.

Checkers and reformulation. OpenNovelty requires verbatim quote pairs for every refutation [32]; Ai2's Idea Novelty Checker compares the facets of an idea with retrieved papers (accuracy 0.81 on 32 expert-labelled ideas) and counts a known facet in a new domain as new [24], and a later study finds it and the AI Scientist evaluator 8 to 29 macro-F1 points below a cheap prompted judge on ICLR 2026 novelty pairs [26]; Theorizer [10], Hypothetica [28, abstract only] and FutureHouse's Owl, now Precedent [5], label or retrieve prior work, and none reports accuracy on cross-field earlier work. Idea-Catalyst [12], SciAtlas [20] and the prior-art skill [3] translate or link vocabularies across fields, without a measured cross-field retrieval rate.

Typed verdicts and novelty benchmarks. AViD returns equivalent, generalization, specialization or different for Lean-formalised mathematics, on 26 within-field withdrawals [19]; we adopt its four types. NovGauge, NovBench, RINoBench and RQ-Bench take ground truth from reviewer overlap, citations or author-anchored questions [31, 29, 23, 25], so a twin with no shared citation cannot be a positive in them. SchNovel takes 15,000 arXiv pairs from six fields and assumes the later paper is more novel [16]; an axiomatic benchmark checks whether ten novelty metrics, OpenNovelty among them, move the right way when a reference pool is manipulated, with a pool from a distant domain as a relevance probe [17]. Neither labels cross-field rediscoveries.

Rediscovery rates. Kim, Kojaku and Ahn detect same-year simultaneous discoveries from citation graphs and note that the method may fail across separate communities [13]; the Gradient Coupling Inquiry catalogues 18 cross-domain instances, 6 of them independent rediscoveries [8]. Within one field, 45.3% of agent-designed ML methods share an algorithmic coordinate with a human method [30], and 24% of 50 AI-generated research documents were paraphrased or significantly borrowed [7, 2].

Contamination. A naming probe can miss contamination that moves performance [21], although guided instructions detect more [6], and closed-book accuracy tracks the number of relevant pretraining documents [11]; Jacovi et al. argue for releasing test data encrypted [9], and Sainz et al. separate raw-text from annotation contamination [22].

3 Benchmark Construction

3.1 Sources

Declarations. Letters, comments, errata and withdrawal notes that state an equivalence name both works, the cleanest positives [19]. Two OpenAlex phrase searches (3,172 and 5,591 works), screened by Claude Haiku 4.5, and two blind Claude readings of sampled titles put the usable share at about 20 declarations (95% interval 8.5 to 43); annotators decide what enters v0. **Kulikowski's twins.** Of his papers, 72 have an earlier cross-field twin: 54 name the method they use (attributed imports; 8, drawn with a seed, form the control subset `C`, kept out of `R` and `H`), 14 name it partly or not at all (the rediscovery positives) and 4 are unmarked. **Other sources.** Six of the Gradient Coupling catalogue's 18 instances

Table 1. Related resources and checkers by property. Tags: [V] read on the primary page or full text; [S] secondary or abstract only. "Not reported" means the source reports nothing on the property, under the coverage of our review.

Resource or checker	Ground truth for prior work	Typed labels	Closed-book control	Non-comp. items	AI-origin items	Verbatim evidence
Kulikowski 2026 [V]	210 curated twins	no (ranked pairs)	not reported	excluded	no	no (human check)
OpenNovelty [V]	none quantitative	refute / cannot refute / unclear	not reported	untested	no	quote pairs checked in code
Ai2 INC [V]	32 expert-labelled ideas	novel / not novel	not reported	untested	generated ideas	no
Ai2 Theorizer [V]	label proportions only	established / derivable / new	not reported	any S2 field	generated theories	not documented [S]
Owl / Precedent [V]	no published accuracy	free text, found flag	not reported	biology-first	no	no (LLM summaries)
Idea-Catalyst [V]	LLM-judge win rates	coverage levels	not reported	as sources only	no	no
AViD [V]	26 within-field withdrawals	four equivalence types	not reported	no (mathematics)	no	no
NovGauge [V abstract]	reviewer overlap, co-citation	task / problem / method	not reported	no	no	no
Gradient Coupling [V]	18 instances, 6 independent	rediscovery or bridge code	no	physics of durability	no	no
Coelacanth v0	documented cross-field rediscoveries, labelled by annotators, adjudicated by the author	six pair labels, item verdict	two probes, three models	yes	10-item stratum	claim chart, quotes checked in code

are independent rediscoveries and about 4 are usable; two non-computational controls (group 2 innate lymphoid cells, RANKL) have a documented unification under several names; 10 of 54 methods that Claude Sonnet 5 and Opus 5.5 proposed as new form the descriptive stratum `A`. Negatives are Kulikowski's topical distractors, and `N` is defined by the adjudicated verdict `nothing_close`.

Composition. Batch b01 holds 30 items with 87 candidates and 161 quote pairs checked verbatim, and b02 holds 30 items, the 10 AI-origin items among them, with 49 candidates and 111 quote pairs. Table 2 gives the composition after labelling and the attribution round; batch b03 (cf-0068, cf-0069) is not counted.

One near-negative (cf-0064) has gold and so leaves `N`. `N` holds 15 items: the 9 near-negatives without gold and 6 mined positives that came out empty (a seventh, cf-0004, is excluded because its only candidate had no readable text). `R` holds 6 Kulikowski twins and 4 declarations. Three Kulikowski twins are recalled only through works the models named (Section 5.3), so `H` is empty, and one `R` item, cf-0044, is in neither column: Opus 5.5's safeguards declined every sample.

Table 2. Benchmark v0, batches b01 and b02. "With gold": at least one candidate rated `similar_mechanism` or stronger (any relation but `different`) by the quiz and the author's adjudication. `R`: adjudicated rediscovery positives with a cross-field gold work that the later work does not cite. Recalled: some closed-book sample names a cross-field gold work, or names a work the author judged cross-field gold. `H`: `R`, not recalled, and with a valid sample from every probed model.

Group	Source	Items	Computational	With gold	`R`	Recalled	`H`
positive	declaration	13	12	11	4	4	0
positive	Kulikowski	14	14	10	6	5	0
positive	catalogue	2	2	2	0	0	0
positive	post-cutoff lead	1	1	1	0	0	0
positive	non-computational control	2	0	1	0	0	0
attributed control (`C`)	Kulikowski	8	8	8	0	0	0
near-negative	Kulikowski	10	9	1	0	0	0
AI-origin (`A`)	generated	10	not counted	no candidates yet	0	0	0
Total		60		34	10	9	0

3.2 Field-native claims

Systems see a claim text and its date and nothing else. The text keeps the source field's own vocabulary, removes author names, titles, years, venues and the coined method name, and has at most 120 words; a script checks it. Claude drafted every claim text and neutral skeleton as pre-fill, and annotators checked the texts of 51 of the 60 items (the 9 near-negatives are unchecked). The skeleton is shown to systems only in the vocabulary-barrier ablation, because a neutral paraphrase as input would do the cross-field translation for the system.

3.3 Claim charts and evidence

Claude drafted each claim's chart of two to six elements and the quote pairs (at most 40 words per side) that show an element covered by a candidate; every quote pair is checked verbatim against the cached texts after normalisation (`src/coelacanth/quotes.py`). Coverage is partial: 35 of the 135 candidates, and 19 of the 70 gold pairs, have no quote pair, and no key flags or covered elements were collected, so the frozen items fail the loader's chart-field validation rules (60 of 60). The labels rest on the annotators' judgement of the pre-filled packet, not on the charts.

3.4 Typed labels

Each claim-candidate pair gets one of six relations (Table 3), with AViD's four types and direction. The item verdict is the strongest relation among the gold candidates, or `nothing_close`, a statement about the search that carries a coverage note. `equivalent` and `specialization` are the "already done" verdicts,

whose precision is reported separately because a false "already done" is the costliest error. The labelling guide has ordered decision rules, 14 worked examples (all excluded) and nine clarifications fixed before any human label; when unsure between two labels, the annotator takes the one that claims less prior art.

3.5 Memorization control

Each positive is probed without tools on Claude Opus 5.5, Sonnet 5.5 and Haiku 4.5, on the field-native text and date: a frozen prompt asks for at most five earliest publications that state the idea (five samples per model), and a cross-field prompt asks whether the method or result appeared earlier in a different field (three samples), which under a rule fixed before its first call caught 8 of 9 pilot items the frozen prompt missed (Section 5.1). Declined samples are retried once and never counted as misses; a person decides borderline matches blind to model and prompt by a fixed rubric, and named works that resolve and predate the claim are judged as candidates.

`R` holds the adjudicated rediscovery positives (not `C`, not `A`) with at least one person-confirmed cross-field gold work that the later work does not cite. An `R` item is *recalled* when any sample of any probed model under either prompt names a cross-field gold work, or a work the author judged cross-field gold. `H`, the `R` items not recalled, is reported first, as k of n, described as "no closed-book recall detected under two probes"; `H-any` and `H-post` (also dated after the models' training cutoffs) are variants.

3.6 Leakage

Anything a system retrieves that is dated on or after the claim date is a leakage hit, web pages included (dated by the version read); a leakage hit never counts as a find, and the leakage rate is reported per system. A date filter on publication dates alone would hide 1 of the 156 gold works we checked, a reprint dated 2004 for a claim of 1984.

3.7 Annotation: what people decided

People made every relation and cross-field decision in v0; a model decided only 4 attribution pairs, noted below. Table 4 counts the recorded human decisions: 716 in all, 408 by paid annotators, 95 by the author and 213 by people whose split is not recorded. A count is of answers, not of people: five annotators worked under three IDs (a1 and a2 each used by two people, a3 by one, no person under two IDs), and the repository does not record which of the two people behind a1 or a2 gave an answer.

Five paid annotators with graduate STEM training (Section 9) labelled the 136 claim-candidate pairs of b01 and b02 in a quiz (the 10 AI-origin items have no candidates yet). For each pair they chose the relation from three options Claude proposed from the pre-filled packet, in fixed precedence order, or any

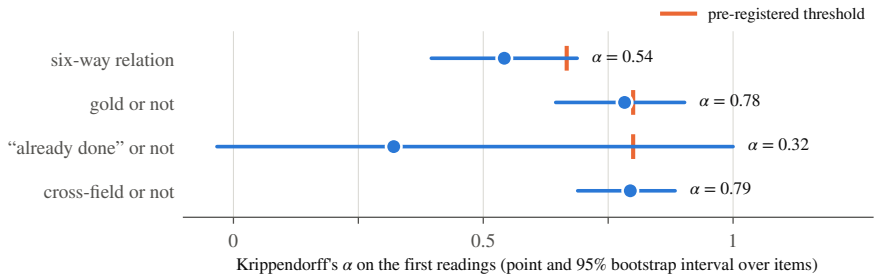


Fig. 1. Agreement of the quiz on the 136 pairs of b01 and b02. Points are Krippendorff's alpha on the first readings and bars the 95% percentile bootstrap intervals over items (2,000 resamples); orange ticks are the thresholds of protocol decision 18 (none is set for cross-field status).

other relation, and said whether the earlier work is from another field; every pair went to two IDs, 272 answers in all, so every pair was read by two different people, and which option Claude ranked first was never shown (it was always one of the three). They worked on 10 October 2026 in the author's presence and pasted in comments that were written outside the quiz page, which we do not use as evidence of reasoning. Median active page time was 12 to 14.5 s per question, and that the choices are the annotators' own rests on the author's presence. The author settled all 51 disagreements with a one-line reason.

Krippendorff's alpha on the first readings [14] (Figure 1) is 0.542 on the six-way relation (95% bootstrap interval over items 0.396 to 0.688), 0.783 on gold, 0.321 on "already done" (unstable: only 6 of 272 readings say it) and 0.794 on cross-field status, against pre-registered thresholds of 0.667 and 0.80: the point estimates miss all three, and the guide revision and second reliability measurement that decision 18 then prescribes were not done, since no annotators remain. By ID pair the six-way alpha is 0.351 (a1 and a2, 90 pairs), 0.936 (a1 and a3, 28) and 1.0 (a2 and a3, 18). After adjudication 70 pairs on 34 items are gold, 36 of them cross-field. Because model-proposed options may anchor annotators [4, 1], note that 189 of 272 answers (0.695; chance 0.333) equal the option Claude ranked first.

The quiz did not cover claim texts, key elements or attribution. The attribution of the 24 cross-field gold pairs and the claim texts of the 51 probed items were checked in separate rounds in the author's presence, with files repackaged by Codex, a model, at his request; the two attribution readings are word for word identical, so no agreement is reported, and 4 pairs the annotators could not check were decided by Claude under the author's delegation from the publishers' reference lists in Crossref. Batch b03 (3 pairs) and the judgments of the 31 works named closed-book (Section 5.3) were answered by hand on printed sheets in the author's presence and transcribed by Codex; the files carry one export time and no page times, so we take the relation and cross-field values as the annotators' on the author's statement and keep the b03 pairs out of the

agreement figures. The author then adjudicated the 14 gold-rated named works himself on a decision page (Section 5.3).

4 Measurement A: Design

Systems and scoring. Four rows are scored: a single agent with web search, and the same agent with the prior-art skill [3], both Claude Code with Sonnet 5.5 at high effort and the same tools and caps ($0.40 per item); the memory reference, Sonnet 5.5 without tools under the frozen closed-book prompt, scored on `R` like a system; and Kulikowski's fingerprint in the closed pool (Section 6). The protocol planned ten systems; the others are not run, because the API credit ($10.87) does not cover them and no annotator remains to judge their pools, so "existing checkers" below means the two agents and the memory reference. Pooled judging is replaced by scoring against the gold already adjudicated: a retrieved work outside it is unjudged, not a miss, so every retrieval figure is a lower bound. The primary metric is cross-field success@10, the share of items with a cross-field gold work in the top 10 (Wilson intervals); verdicts are scored by three-class macro-F1 on `R` plus the 9 near-negatives. `R` holds 10 items, below the 20 that decision 23 requires of a confirmatory test, so every comparison is descriptive (Supplement S3).

Freeze. The rules freeze (tag `protocol-v1.0`, 11 October 2026) came after the quiz labelling and the closed-book runs, not before them as planned, and before any system ran on a benchmark item; each agent configuration was hashed before its first scored run. Later changes are logged as deviations, among them the single agent's skeleton run, decided after the first scored runs had been looked at, and the correction of the gold (Section 7). The SHA-256 of the protocol at the tag (commit 658b3fb of the private development repository) is `06331185afb8be891801f5431987ffcde2e40d25c0ba20ea60ecf58ab3867a88`, and that of `splits-v1.json` is `8527d1db9ebfd51559ba10709c6db22d419ec5ea2606b1ee0dbc8157b880637c`; both can be checked against the released files, `eval/protocol-v1.0.md` being the protocol exactly as tagged and `eval/protocol.md` the same text with the logged deviations added.

5 Closed-Book Recall of Documented Cross-Field Rediscoveries

A published equivalence declaration puts the link itself into the literature the models learn from, so documented rediscoveries are exposed by construction, and a retrieval score on them can measure memory. This section measures that exposure.

5.1 Pilots

Three pilots on Claude-drafted claim texts preceded the v0 runs, which replace them (Supplement S1). They gave the rule used here: the cross-field prompt

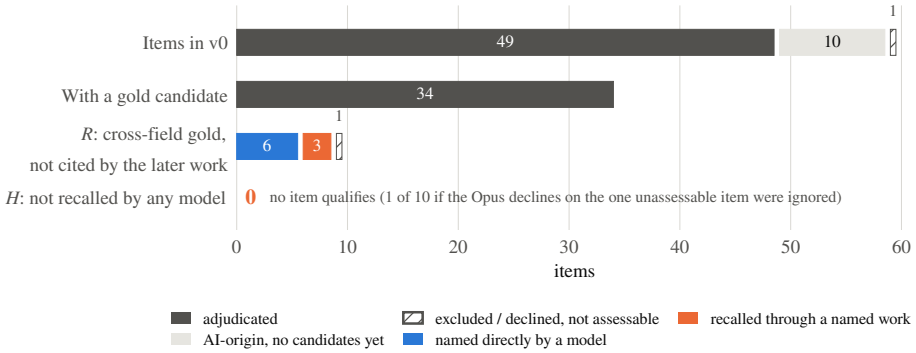


Fig. 2. From the items of v0 to `H`. Of 60 items, 49 carry adjudicated labels, 34 of them have a gold candidate, and 10 are in `R`. Nine of those ten are recalled by a model, six by naming a gold work directly and three through a work they named that the author judged cross-field gold; the tenth is not assessable because Opus 5.5 declined every sample. `H` is empty.

caught 8 of the 9 pilot items that the frozen prompt missed, and under either prompt all 16 Kulikowski items were named. What the cross-field prompt surfaces is mostly the method's origin in another field (for example statistics for population pharmacokinetics, linear prediction theory for kriging).

5.2 Supply and what it implies for `H`

The pilots projected `H` at about 5 to 7 items, under the 20 that a test needs, so on 9 Oct, before any human label, any system run on a benchmark item and the rules freeze, the protocol was updated: the union rule stays as pre-registered, `H`'s k of n comes first in every headline table, descriptively, and the memory reference is added. We rejected redefining gold after seeing what the models named. The paper reports no test on `H`, does not call any subset contamination-free, does not read source differences as a memorization effect (they are confounded with popularity) and makes no claim about undocumented rediscoveries.

5.3 v0 runs

All 51 positives of b01 and b02 were probed on 10 October 2026 under both prompts (1,549 calls for $22.34; 1,263 of 1,272 planned samples are done, the 9 missing all Opus 5.5 on AI-origin items). Opus 5.5 declined 75 of its 470 calls, almost all on biology, medicine and immunology claims, and after the retries gave no valid sample on five items: cf-0006 (a control), cf-0044 (a Kulikowski twin on viral genome classification) and three AI-origin items; those five cannot enter `H`. Sonnet 5.5 and Haiku 4.5 declined none. People decided the 213 borderline matches by a fixed rubric (Table 4), 37 of them as naming the gold work; the remaining bookkeeping, which changes the status of no `R` item, is in Supplement S2.

Of the 11 provisional `R` items, 7 were recalled; the attribution round on the 24 cross-field gold pairs then removed cf-0031, so `R` holds 10 items. Six are recalled directly. Three more, cf-0013, cf-0036 and cf-0046 (all Kulikowski twins), name no gold work directly but are recalled through works the models named outside the gold lists, which the author judged cross-field gold (below). The tenth, cf-0044, is in neither group: Opus 5.5's safeguards declined every sample on its checked text and decision 6 keeps such an item out of `H`; Sonnet 5.5 and Haiku 4.5 named no gold in a decided match, so if the declines were ignored `H` would be 1 of 10 (Wilson 0.02 to 0.40). All declaration items in `R` are recalled (four of four): a published declaration puts the link itself into the literature models are trained on. On `R` the cross-field prompt raises the number of items whose cross-field gold a model names directly from 4 to 6 for Opus 5.5 and Sonnet 5.5 and from 2 to 4 for Haiku 4.5 (Figure 4, left); Sonnet 5.5 under the frozen prompt, the memory reference, names one for 4 of the 10 items, and for 6 with the named works the author judged gold, which Section 7 uses.

For the three items that no model named a gold work for, the 31 distinct works the models named outside the gold lists were judged, one reading each by the item's annotator: 14 were rated gold and 6 of them cross-field gold. The author adjudicated all 14 on 11 October (13 as the annotator read them, one lowered from `equivalent` to `partial`) and confirmed every cross-field reading, so each of the three items counts as recalled and `H` is empty: 0 of 10 (Wilson 95% interval 0.00 to 0.28), as are `H-any`, `H-post` and `H_half`. The control thus shows the opposite of a benchmark with an unrecalled core: every documented cross-field rediscovery in v0 that could be probed is recalled by at least one of three models, directly or through a work it names, and a retrieval score on `R` cannot be separated from memory. A re-decision of 30 random borderline entries agreed on all 30, but the author made it, so it is not an independent second reading.

6 Kulikowski Reproduction and the Closed-Pool Protocol

We reproduced the TF-IDF rows of Kulikowski's benchmark from his package v1.0.0 with one command (2,181 files verified, about 40 seconds, no model call). Every fingerprint number reproduces exactly: whole-fingerprint AP 0.557 [0.407, 0.720], P@1 0.826 and MRR 0.870 under the unigram setting, skeleton P@1 0.815 (75 of 92); the abstract rows differ by at most one query, because 5 abstracts are not redistributable (Supplement S4). The frequently quoted pair "AP 0.557, P@1 0.82" mixes two representations and two TF-IDF settings; under the unigram setting AP is 0.441. The pool is also closed, so a system that searches open literature is penalised for correct earlier works outside the 109 papers: the closed-pool protocol has every system rank only the pool's cross-field candidates for each query, scored by P@1 and MRR per query and by AP over the 5,812 pairs. The families are well-known imports (all 16 Kulikowski items were named closed-book in the second probe), and the reference pipeline is not built in this version, so no pipeline is compared with the fingerprint.

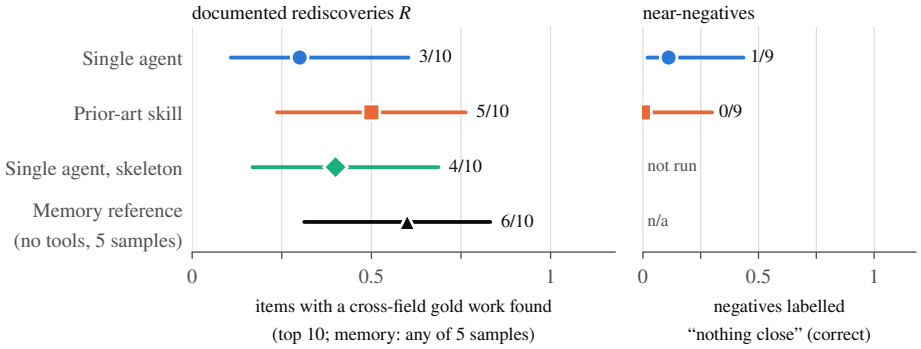


Fig. 3. Retrieval on `R` and labelling of the near-negatives. Left: items with a cross-field gold work found in the top 10 (memory reference: in any of five samples), with Wilson 95% intervals; gold includes the named works the author adjudicated. Right: near-negatives labelled `nothing close`, the correct label, for the two agents that were run on them.

7 Results of Measurement A

All figures are on `R` (10 items), so every comparison is descriptive. Gold is the quiz-labelled gold of the frozen items plus the 14 named works the author adjudicated as gold on 11 October, before any system ran; a retrieved work outside the gold is unjudged, not a miss. The first scoring used the frozen items file alone, which lacked those named works (an audit found the gap), and gave 2, 3, 3 and 4 successes instead of 3, 5, 4 and 6; the corrected figures below are the ones we report, and both scorings are kept. `H` is empty (Section 5.3).

The agents find a cross-field gold work for 3, 5 and 4 of the 10 items (Figure 3). Every success is on an item for which the memory reference also names a gold work (Figure 4, right), and the memory reference, with no search at all, names one for 6. The prior-art skill finds 2 items that the single agent does not and none the other way round (exact McNemar test, discordant counts 2 and 0, $p = 0.5$); the skeleton condition, which hides the field's vocabulary, finds 2 items that the text condition misses and misses 1 that it finds (2 and 1, $p = 1.0$). With 10 items we read none of these differences as an effect. Of the 45, 66 and 57 cleaned candidates the agents return on `R`, 6, 13 and 8 are gold; the other works are unjudged and released for later judging, so the figures are lower bounds. Leakage hits (a top-10 or evidence item dated on or after the claim, which includes any web page fetched in 2026) occur on 50% to 70% of items and never count as finds.

Verdicts. On `R` plus the 9 near-negatives (19 items), the single agent has macro-F1 0.261 and the prior-art skill 0.198, because both label almost every negative as related prior work (`nothing_close` for 1 of 9 and 0 of 9; the near-negatives' claim texts were not checked by a person). "Already done" is predicted for 3 of 18 weaker items by the single agent and 1 of 18 by the prior-art skill, and the one item that is already done is called so by neither. The skeleton run



Fig. 4. The 10 items of `R`: closed-book recall (left) and what each system found (right). Left, by model and prompt (frozen or cross-field, "cross"): blue marks a decided match that names a cross-field gold work in at least one of the five (frozen) or three (cross-field) samples; the last column marks the three items recalled only through a named work the author judged gold; cf-0044 is hatched where Opus 5.5 declined every sample, and its Haiku 4.5 cells hold only undecided loose matches. Right: items for which a system has a cross-field gold work in its top 10 (memory reference: in any of five samples); markers follow Figure 3. Every item found by an agent is also named by the memory reference.

has no negatives, so its macro-F1 (0.600 on `R` alone) is not comparable. An agent run costs $0.07 to $0.14 per item; the five scored runs cost $4.88.

8 Limitations

v0 was planned at about 80 items and ships with what is adjudicated: 60 items in b01 and b02, 51 probed as positives, `R` with 10 items against a target of 30, so no confirmatory test runs, and `H` is empty. **Labels.** The pair labels come from a quiz over a pre-filled chart and model-proposed options, answered fast (median 12 to 14.5 s per question) by five annotators sharing three IDs whom the author supervised in person; 189 of 272 answers (0.695) equal the option the model ranked first, all three reliability thresholds are missed (six-way alpha 0.542 against 0.667, gold 0.783 and already done 0.321 against 0.80), the comments were pasted from outside the page, and several returns were transcribed by a model (Section 3.7). A reader who needs independent human labels should treat them as one supervised reading. **Pre-fill.** The claim charts and quote pairs are Claude's pre-fill (35 of the 135 candidates have no quote pair, no key flags or covered elements were collected), every claim text is a Claude draft that annotators checked for 51 of the 60 items, and attribution rests on two identical

readings, which count as one, and 4 model-decided pairs. **Closed-book probes.** Two prompts are a narrow instrument: a miss is weak evidence that a model has not memorised the link [21, 22], a hit does not show that a system used memory, the probes are Claude models only, and safeguard declines can leave an item without valid samples for one model (Section 5.3). **Scope.** The benchmark holds documented rediscoveries, which are exposed by construction, and says nothing about undocumented ones; each source is biased (declaration phrase search has low precision, Kulikowski's corpus is arXiv-weighted and computational, the catalogue covers one family of physical problems), and Claude models served as pre-fill, probes, pilot readers and the AI-origin generator. Measurement A covers two agents, a memory reference and Kulikowski's fingerprint, scored against adjudicated gold, so retrieval figures are lower bounds and the other tools of Table 1 are not measured; the first scoring missed adjudicated gold works and was corrected (Section 7). Paywalled texts limit some evidence to abstracts, and sources are English only.

9 Ethics, Licences and AI Use

Five annotators took part, briefed in writing and in a Russian kit and paid $12 to $14 per hour (the posted offer said $8 to $10; the pay is the author's statement): two women and three men, aged 26 to 37, with graduate training in STEM (three MSc, two PhD) in engineering, chemistry and biochemistry, computer and data science and mathematics, from universities in the United States, the United Kingdom and Kazakhstan. They appear in the data only as IDs a1 to a3 (a1 and a2 were each used by two of them, a3 by one; which of the two gave a given answer is not recorded), may be named in the acknowledgements if they wish, and keep item files confidential until release. The author reports that all five agreed in writing to take part and to have their answers released under an ID only. Items describe relations between published works; the wording never accuses, and `nothing_close` is a statement about the search.

Code is MIT and data CC-BY-4.0; OpenAlex metadata is CC0, and Kulikowski's derived data and the Gradient Coupling catalogue are CC-BY-4.0 and attributed. No Semantic Scholar or Asta records or full texts are shipped, and quotes are at most 40 words.

Disclosure of interests. The author wrote Hunch and co-founded AgentMash, whose lab publishes it; Hunch is not run in this version.

AI use. People chose every pair label and cross-field answer (272 quiz answers by paid annotators, from options a model proposed in fixed precedence order, Section 3.7), and the author settled the 51 disagreements and adjudicated the named works himself: 716 recorded human decisions in all (Table 4). Claude drafted claim texts, charts and evidence as pre-fill, screened the declarations, read the pilots' borderline matches and named works, generated the AI-origin items, ran the harness, wrote code and helped draft this paper; Claude decided the attribution of 4 pairs under the author's delegation, and Codex, a model, transcribed or repackaged the b03, attribution, claim-text and named-works re-

turns at the author's request. The author is responsible for all content of this paper. The figures are drawn from the repository's result files by scripts.

10 Availability

A clean public repository, https://github.com/IskakovDamir/coelacanth-benchmark, holds the code (MIT), the protocol and its deviations log, the datasheet, the loader with its validation tests, the closed-book runner and scorer with both prompts, the harness and scorer of measurement A and the one-command Kulikowski reproduction (`eval/results/kulikowski-baseline/run.sh`). The items with their gold, the closed-book answers, the records of the scored runs with their per-call cost ledgers, the list of retrieved works outside the gold, the quiz returns with their provenance notes and the supplement ship in one password-protected archive with a canary string (protocol decision 26), deposited on OSF with the DOI 10.17605/OSF.IO/672RJ (https://osf.io/672rj/), with mirrors on Hugging Face (DamirIskakov/coelacanth-v0) and Figshare (10.6084/m9.figshare.34337127). The freeze is identified by the tag `protocol-v1.0` (commit 658b3fb) of the private development repository and by the SHA-256 values in Section 4.

References

1. Y. Berzak, Y. Huang, A. Barbu, A. Korhonen, B. Katz. Anchoring and Agreement in Syntactic Annotations. EMNLP 2016, pp. 2215 to 2224. https://aclanthology.org/D16-1239/
2. S. Schmidgall et al. (35 authors, from Schmidgall to Tao Tu; list on the arXiv page). Accelerating Scientific Research with Gemini in the Real-World. arXiv 2608.26701, 27 Aug 2026. https://arxiv.org/abs/2608.26701
3. R. Czerwinski. prior-art Agent Skill, v1.0, MIT, 4 Aug 2026. https://github.com/rianczerwinski/prior-art-skill
4. K. Fort, B. Sagot. Influence of Pre-annotation on POS-tagged Corpus Development. LAW IV, 2010, pp. 56 to 63. https://aclanthology.org/W10-1807/
5. A. White, S. Rodriques et al. FutureHouse Platform: Superintelligent AI Agents for Scientific Discovery, 1 May 2025. https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents ; Edison Platform documentation, Agents page, https://docs.edisonscientific.com/agents
6. S. Golchin, M. Surdeanu. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. ICLR 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/bc39a59c49b731c51398ad6b12f301d3-Abstract-Conference.html
7. T. Gupta, D. Pruthi. All That Glitters is Not Novel: Plagiarism in AI Generated Research. ACL 2025, pp. 25721 to 25738. https://aclanthology.org/2025.acl-long.1249/
8. L. Holden. Gradient Coupling Inquiry: a cross-domain gap-finding vault, 5 Sep 2026 (v0.2.0). https://github.com/deciduus/same-object ; concept doi:10.5281/zenodo.22334047 (resolves to record doi:10.5281/zenodo.22337575), CC BY 4.0

9. A. Jacovi, A. Caciularu, O. Goldman, Y. Goldberg. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. EMNLP 2023, pp. 5075 to 5084. https://aclanthology.org/2023.emnlp-main.308/
10. P. Jansen, P. Clark, D. Downey, D. S. Weld. Generating Literature-Driven Scientific Theories at Scale. arXiv 2601.16282, 22 Jan 2026; asta-plugins (skill generate-theories), Apache-2.0. https://arxiv.org/abs/2601.16282
11. N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. ICML 2023, PMLR 202:15696 to 15707. https://proceedings.mlr.press/v202/kandpal23a.html
12. P. Kargupta, S. Mehri, D. Hakkani-Tur, J. Han. Sparking Scientific Creativity via LLM-Driven Interdisciplinary Inspiration. arXiv 2603.12226, 12 Mar 2026. https://arxiv.org/abs/2603.12226
13. M. Kim, S. Kojaku, Y.-Y. Ahn. Uncovering simultaneous breakthroughs with a robust measure of disruptiveness. Science Advances 12(14):eadx3420, 2026. doi:10.1126/sciadv.adx3420
14. K. Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Human Communication Research 30(3):411 to 433, 2004. doi:10.1111/j.1468-2958.2004.tb00738.x
15. E. Kulikowski. Same Problem, Different Field: Cross-Domain Solution Import via Domain-Stripped Computational Fingerprints. JCDL '26; arXiv 2609.07595v2. doi:10.1145/3805696.3846016; package doi:10.48804/W3B9WC [V] Note: the ACM DOI 10.1145/3805696.3846016 did not resolve on 11 Oct 2026 (JCDL '26 starts 13 Oct); the arXiv ID and the KU Leuven package DOI 10.48804/W3B9WC resolve.
16. E. Lin, Z. Peng, Y. Fang. Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications (SchNovel). AISD 2025, pp. 46 to 57. doi:10.18653/v1/2025.aisd-main.5; arXiv 2409.16605. https://arxiv.org/abs/2409.16605
17. M. Liu, C. Zhai. An Axiomatic Benchmark for Evaluation of Scientific Novelty Metrics. arXiv 2604.15145v2, 5 Aug 2026. https://arxiv.org/abs/2604.15145
18. J. H. Monaco, R. L. Anderson. Tai's formula is the trapezoidal rule. Diabetes Care 17(10):1224 to 1225, 1994. doi:10.2337/diacare.17.10.1224
19. A. Porto. Beyond Correctness: Toward Automated Novelty Verification with Lean 4. arXiv 2608.14669v1, 2 Aug 2026. https://arxiv.org/abs/2608.14669 ; code https://github.com/ayrtonporto/avid-journal (MIT)
20. S. Qiao et al. (19 authors, from Qiao to H. Chen; list on the arXiv page). SciAtlas: A Computable Atlas of Science for Knowledge-Grounded AI Research. arXiv 2605.22878v2 (v1 20 May 2026, v2 29 Aug 2026). https://arxiv.org/abs/2605.22878v2
21. M. Roberts, H. Thakur, C. Herlihy, C. White, S. Dooley. To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination. ICLR 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/2d04d97593c8c33d415337f408ed0e1b-Abstract-Conference.html
22. O. Sainz, J. A. Campos, I. García-Ferrero, J. Etxaniz, O. Lopez de Lacalle, E. Agirre. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. Findings of EMNLP 2023, pp. 10776 to 10787. doi:10.18653/v1/2023.findings-emnlp.722
23. T. Schopf, M. Färber. Is this Idea Novel? An Automated Benchmark for Judgment of Research Ideas (RINoBench). LREC 2026; arXiv 2603.10303, 11 Mar 2026. https://arxiv.org/abs/2603.10303

24. S. Shahid, M. Radensky, R. Fok, P. Siangliulue, D. S. Weld, T. Hope. Literature-Grounded Novelty Assessment of Scientific Ideas. SDP 2025, pp. 96 to 113. doi:10.18653/v1/2025.sdp-1.9
25. S. Sinhahajari, N. Majumder, S. Poria. On the Limits of LLM-as-Judge for Scientific Novelty Assessment. arXiv 2606.12071v1, 10 Jun 2026. https://arxiv.org/abs/2606.12071
26. N. Sternlicht, S. Shahid, P. Jansen, D. S. Weld, P. Siangliulue, T. Hope. Old Ideas, Novel Problems: The Instability of LLM-Based Novelty Evaluation. arXiv 2610.02022v1, 1 Oct 2026. https://arxiv.org/abs/2610.02022
27. M. M. Tai. A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. Diabetes Care 17(2):152 to 154, 1994. doi:10.2337/diacare.17.2.152
28. H. H. Tan, A. A. Malkoç, K. Becerir, B. Erol, M. Karakaya. Hypothetica: A Multi-Agent System for AI-Powered Originality Assessment. IISEC 2026, Ankara, 5 to 6 Feb 2026, pp. 525 to 530. doi:10.1109/IISEC69317.2026.11418474
29. W. Wu, Y. Zhao, Y. Wang, S. Li, J. Shao, Y. Long, C. Zhang. NovBench: Evaluating Large Language Models on Academic Paper Novelty Assessment. Findings of ACL 2026, pp. 32103 to 32133, doi:10.18653/v1/2026.findings-acl.1607; arXiv 2604.11543
30. Y. Yang, Z. Yang, L. Peng, M. Luo, Y. Kan, W. Gao, J. Zhan. When AI Designs AI: Innovation or Imitation? arXiv 2608.17471v1, 18 Aug 2026. https://arxiv.org/abs/2608.17471
31. G. Zhang, K. Tan, M. Zhang, L. Ju, et al. (15 authors). NovGauge: A Fine-Grained Benchmark for Diagnosing LLMs' Capability in Paper Novelty Assessment. arXiv 2609.11234v1, 10 Sep 2026.
32. M. Zhang, K. Tan, Y. Huang, et al. (23 authors). OpenNovelty: An LLM-powered Agentic System for Verifiable Scholarly Novelty Assessment. arXiv 2601.01576v2, 2026.

Table 3. Pair labels. C is the claim, P the earlier candidate.

Label	Definition (abridged)	AViD type	Already done
`equivalent`	P states the same mechanism or result as C under the same hypotheses, up to renaming and routine reformulation; every key element covered with the same meaning	equivalent	yes
`specialization`	C is a strict particular case of P: restricting P yields C with the same meaning	specialization	yes
`generalization`	C is strictly stronger; P is obtained from C by fixing a parameter or adding a hypothesis	generalization	no
`partial`	P covers at least two key elements of C with the same meaning and leaves at least one uncovered	(none)	no
`similar_mechanism`	same functional mechanism or form in a different system or with a different meaning of the variables; every key element has a counterpart, at least one by role only	different	no
`different`	below `similar_mechanism`	different	no

Table 4. What people decided in v0, by task. Counts are recorded answers read from the repository files, not people; the re-decision repeats entries already counted in the row above it.

Task	Decided by	Answers	Covers
Pair label (relation, cross-field answer)	paid annotators	272	136 pairs of b01 and b02, two readings each
Disagreements settled	the author	51	one-line reason each
Attribution of cross-field gold pairs	paid annotators	48	24 pairs, two readings each
Claim-text check	paid annotators	51	the 51 probed items
Pair label, batch b03 (printed sheets)	paid annotators	6	3 pairs, two readings each
Works the models named closed-book	paid annotators	31	3 items (printed sheets)
Named works rated gold	the author	14	14 of the 31 works
Borderline closed-book matches	people, split not recorded	213	first run
Re-decision of random borderline matches	the author	30	agreed on all 30
Total		**716**	

Table 5. Retrieval on `R`, text condition unless stated. Cross-field success@10 is the share of the 10 items for which a cross-field gold work is in the system's cleaned top 10; the miss rate has a Wilson 95% interval. The memory reference is Sonnet 5.5 without tools under the frozen closed-book prompt: its figure counts an item when any of five samples, each listing up to five publications, names a cross-field gold work, so it is not a success@10.

System	Success (k of 10)	Miss rate [Wilson]	Recall@10	MRR@10	Success@1	Leakage rate	$ per item
Single agent	3	0.70 [0.40, 0.89]	0.153	0.167	1	0.60	0.070
Prior-art skill	5	0.50 [0.24, 0.76]	0.270	0.263	1	0.70	0.130
Single agent, skeleton	4	0.60 [0.31, 0.83]	0.200	0.264	2	0.50	0.084
Memory reference	6	0.40 [0.17, 0.69]	n/a	n/a	n/a	n/a	none