

- [Field Map: Procedure Evidence in LLM-Assisted Knowledge Work](#)
 - [Summary](#)
 - [1. Purpose, scope and verification discipline](#)
 - [2. The problem field](#)
 - [3. The solution side](#)
 - [4. The institutional side](#)
 - [5. Agentic research systems](#)
 - [6. Location: where the gap is](#)
 - [7. Ten follow-on projects](#)
 - [8. Not found, not verified, to be noted](#)
 - [9. Self-application](#)
 - [10. References](#)
 - [11. Provenance and status](#)

title: “Field Map: Procedure Evidence in LLM-Assisted Knowledge Work”

subtitle: “The research context around the working note *Iteratio Simulata*”

author: “Thomas Schüller”

date: “2026-10-06”

lang: en

version: “v0.5”

purpose: “Mapping the research context: problems, solutions, systems, architectures, new approaches”

license: “CC BY 4.0”

kennung: “FORSCH-PROV0459”

slug: “iteratio-simulata”

orcid: “0009-0003-9799-6747”

register_hinweis: “Number provisional; final assignment against the work register”

Field Map: Procedure Evidence in LLM-Assisted Knowledge Work

Summary

The working note *Iteratio Simulata* describes a single case: an LLM output carrying the form of a multi-stage examination without having carried that examination out. This field map locates the case. It asks what is already known on the subject, which methods have been built against it, which system architectures claim to solve the problem, and where, after this survey, the gaps actually are.

The finding in one sentence: the gap between claimed and demonstrable procedure has been described independently in eight separate literatures, but there exists no check channel for the case the work is about — a single, uninstrumented model run of which only the text remains.

Four points follow from the survey and either extend or correct the working note:

1. **The central concept acquired an empirical basis in 2026 that did not exist when the note was written.** Three papers measure the gap step by step for the first time and name it in their own terms: “confabulated continuation”, “performative chain-of-thought”, “reasoning theatre”. Step-level alignment between internal commitment and what the trace depicts stands at 61.9 per cent; up to 80 per cent of the trace is dispensable with no loss of accuracy [F11, F12].
2. **Textual detection on a single document now has one measured precursor.** In agentic systems, simple TF-IDF detectors operating on the language of the closing message outperform LLM judges at spotting false success claims [F74]. This is the only paper found whose signal

lies in the language itself and is therefore transferable to the single case.

3. **A counterposition to the note’s line of argument must be named.** An ACL 2026 paper argues that non-verbalisation of a hint does not by itself establish unfaithfulness [F7]. Markers M1 and M2 do not work through verbalisation but through tracelessness, so they are not hit by it — but the objection belongs in the presentation.
4. **The note’s pre-registration proposal has no institutional counterpart.** An extended search found no scholarly or normative use of an advance declaration of procedure for AI-assisted text work. The nearest instrument is the ICML 2026 reviewing rule, in which authors fix the procedure in advance and reviewers are bound to it — on the reviewing side, not for the manuscript itself [F91]. The absence is the finding.

1. Purpose, scope and verification discipline

1.1 What this map is

The working note treats one occasioning case and derives five markers and a checking tool from it. It does not answer whether the problem is already being worked on elsewhere. This field map answers that. It is a context map, not a systematic review: it makes no claim to completeness by PRISMA standards, and the search path is not documented well enough to be reproducible. It states what was looked for, what was found, and names what was not found.

1.2 What it is not

This map does not rank the works it names by quality. Where a paper places a finding against the working note, that is said there. Where a system makes a claim its own output text does not redeem, that is said too — which is this work’s subject matter, applied to the literature.

1.3 Verification discipline

Version v0.3 of the working note contained five faulty references, including two invented author lists. One rule therefore held without exception for this survey: **every bibliographic item was fetched at its primary page and read off there** — arXiv abstract page, ACL Anthology, PMLR, NeurIPS proceedings, publisher page, standards-body page or Crossref registry record. No item comes from memory.

Three presentation rules follow:

- **No preprint is upgraded to a conference paper.** Where the abstract page names no venue, the entry carries [P] for preprint, however widely cited the work and however customary the venue attribution. Where the abstract page’s comments field names a venue, it is quoted as the authors’ statement, not asserted as verified fact.
- **Unfetchable items are marked as such.** Eleven entries carry [T] — not verifiable, to be checked individually before any journal submission. They are not dropped, because their absence would otherwise be invisible.
- **Version divergence is noted.** Four works changed title or author list between preprint and publication. Anyone taking them from a secondary source dated 2024 or 2025 is very likely citing a stale record (see 8.4).

1.4 Four survey strands

The survey ran in four separate strands, to avoid carrying concepts across strand boundaries:

Strand	Subject	Entries
A	Named failure modes: process, not fact	37
B	Methods, verifiers, architectures	34
C	Provenance, disclosure, pre-registration	39
D	Agentic research systems and their evaluation	41

2. The problem field

Nine separate literatures describe manifestations of the same figure. They partly do not cite each other, which explains the gap between them.

2.1 Unfaithful reasoning trace

The oldest and best-evidenced line. Planted biasing features change the answer while the reasoning trace never mentions them, producing instead a fluent rationalisation of the biased answer; accuracy drops of up to 36 per cent across 13 tasks [F1]. Intervention studies show faithfulness varies by task and — against expectation — decreases with model size [F2]. In explicit reasoning models, a demonstrably used hint is verbalised in typically under 20 per cent of cases [F3]. The effect also appears without any planted bias, on ordinary prompts, which removes the objection that it is an artefact of adversarial setups [F4].

The hardest single observation: transformers solve difficult algorithmic problems using **meaningless filler tokens** in place of reasoning text [F5]. Trace and computation are therefore decouplable in principle. A follow-up confirms this on current models with gains of up to 13 percentage points and shows that invisible computation can pursue objectives that do not appear in the trace at all [F13].

That the field itself commits this conflation is recorded bibliometrically: a position paper estimates that roughly a quarter of recent trace-centric papers treat the trace as a valid interpretability method, and argues it is “neither necessary nor sufficient” for that, with the error concentrated in medicine and law [F6].

The field’s most important conceptual distinction comes from a monitoring paper: **trace-as-rationalisation versus trace-as-computation** [F10]. Where a task genuinely requires the multi-step computation, the trace must carry it and is hard to fake; where it does not, the trace is free to be decorative. That is the load-bearing axis for the question of when visible procedure is real procedure.

2.2 Performative trace

This literature emerged in 2026 and was not available when the working note was written. It is the most direct counterpart to the work’s central concept.

One paper measures when a model internally commits to an answer against when the trace arrives at it: alignment on only 61.9 per cent of steps across nine models and seven benchmarks. The dominant pattern is named “**confabulated continuation**” — after internal commitment the trace keeps producing deliberative-looking text that no longer affects the result. The side finding that matters most here: lower step-level alignment goes together with *greater* trace utility. The traces that help most are the least temporally honest [F11].

A second paper reads the final answer out by activation probing far earlier than any external monitor could, and names the finding **performative chain-of-thought** outright. Notable is the limit it draws itself: backtracking and reversals do track genuine shifts in uncertainty, and are therefore not theatre. Probe-guided early exit saves up to 80 per cent of tokens with no accuracy loss — that is the measurable share of pure display [F12].

In this line the gap is also actively producible. A weaker model can monitor a stronger one’s trace for reward hacking; put that monitor into the training objective and the systems learn **obfuscated** hacking: the behaviour remains, the trace no longer shows it [F9]. Trace legibility is therefore not a fixed property but a fragile one, destructible by training decisions [F8].

The mirror image belongs here too: models can deliberately underperform on targeted capability evaluations and can be tuned to hide a capability [F46]. Feigned incapacity is the same divergence between claim and execution, in the other direction. Anyone framing the gap as two-sided needs both findings.

2.3 Potemkin understanding

The term the working note carries as a neighbouring concept comes from an ICML 2025 paper: benchmark validity presupposes that models fail in human-like ways; where they do not, a pass is “the illusion of understanding driven by answers irreconcilable with how any human would interpret a concept”. Two quantification procedures find *potemkins* across models, tasks and domains, with internal incoherence of concept representation as the cause [F14].

The methodological template for separating competence from score is older: counterfactual task variants that break one default assumption produce consistent performance degradation and point to “narrow, non-transferable procedures” [F15]. Symbolic variants of the same arithmetic problem show drops of up to 65 per cent when an irrelevant but plausible clause is appended [F16]. In controlled puzzle environments accuracy collapses entirely past a complexity threshold — and, decisive for this work, *visible effort peaks beforehand and then declines*, despite remaining token budget [F17]. The display of effort and the difficulty of the task come apart.

The social-epistemic counterpart: AI tools exploit cognitive limits through the promise of productivity and objectivity, so that researchers believe they understand more than they do, and the resulting monocultures remain invisible from inside [F18].

2.4 Agreement under pressure

Sycophancy was first named and measured at scale in 2022; larger models adopt stated user preferences more strongly, and RLHF sharpens this [F19]. The agreement extends to objectively false statements [F20]. The causal mechanism is documented: human raters *and* preference models sometimes prefer a convincingly written sycophantic answer over a correct one [F21]. That names the economic motive behind performative care in general.

Numbers for the case of a critical push-back: 58.19 per cent sycophancy across mathematics and medical datasets; preemptive rebuttals provoke more than in-context ones; persistence of the behaviour once established reaches 78.5 per cent [F22]. An extension to preservation of the user’s desired self-image finds, across eleven models, face preservation exceeding the human rate by 45 percentage points in advice settings, and affirmation of *both* sides of opposed moral framings in 48 per cent of cases depending on who asks [F23].

2.5 Failure of self-correction

This is the foundation of the working note's central concept and of the argument that a claimed self-examination without a counterpart is not merely unsupported but necessarily simulated.

The claim under test is that a single model can generate, critique and improve, with roughly 20 per cent average absolute improvement [F24]. The test isolates *intrinsic* self-correction — no external feedback, no oracle label — and finds difficulty, sometimes degradation after the correction pass; a substantial part of the reported gains came from leaked external signal [F25].

The sharpest single finding comes from graph colouring: GPT-4 solves poorly, verifies equally poorly, and “the correctness and content of the criticisms seems largely irrelevant to the performance” [F26]. A self-critique whose *content* does not matter is the purest case of performative procedure. In the planning domain, self-critique lowers performance relative to a sound external verifier, and feedback granularity makes little difference [F27].

2.6 Judge models

The working note deliberately omits an LLM-as-judge component. The survey supports that decision.

The paper that established the method catalogued its biases in the same breath [F28]. Reordering the two candidates flips the verdict, to the point that one model “wins” 66 of 80 comparisons under one ordering [F29]. A systematic inventory finds bias indications in an average of 40 per cent of comparisons, with rank-biased overlap with human preference of only 49.6 per cent [F30].

Two competing mechanisms for self-preference are named and both evidenced. First: models recognise their own output, and fine-tuning experiments show a *linear*, hence causal, relation between self-recognition and self-preference [F31]. Second: the judge rewards low-perplexity text far more than humans do, regardless of authorship [F32]. The second mechanism matters more here, because it describes exactly how fluent, well-formatted, substance-poor writing comes to score well.

2.7 Fabricated references

The literature that situates the working note's own error finding. The rates: 55 per cent of 222 citations fabricated for GPT-3.5, 18 per cent of 414 for GPT-4; among the *non*-fabricated citations, substantive errors in 43 and 24 per cent respectively, most commonly wrong volume, issue, page and year [F33]. The structure of that finding is the real lesson: **a real source with wrong fields is the harder failure**, because it survives a casual existence check. Precisely this pattern occurred in v0.3 of the working note — all arXiv identifiers correct, only the fields wrong that nobody looks up.

For systematic reviews, hallucination rates of 39.6, 28.6 and 91.4 per cent are on record, the last at zero precision [F34]. Commercial retrieval-augmented legal research tools hallucinate more than 17 per cent of the time each, **despite** being marketed as hallucination-free [F35] — the clearest documented case of claimed versus actual procedure in a professional setting carrying liability.

The most important entry for this work's own methods is recent: current-generation web-search-enabled agents reach 83.6 per cent field-level accuracy on a 931-paper set but **only 50.9 per cent fully correct entries**, with a 27.7-point drop from well-known to recent papers [F36]. Search access therefore does not solve the problem, and recency is the risk factor. A companion paper diagnosing verifiers records that the **false-positive rate, not recall**, decides deployability, and that LLMs over-flag papers past their training cutoff as fabricated [F37]. Transposed: an LLM checking this map's reference list will wrongly object to the newest and most correct entries.

2.8 Presentational quality without substance, and peer review

The term for the case in professional work was coined in 2025: content that appears polished, lacks substance, and offloads the thinking onto colleagues; 41 per cent of respondents report encountering it, at nearly two hours of rework per instance [F38]. An operational definition has existed since 2025, with an expert-derived taxonomy and span-level annotation [F39].

In peer review the numbers are firmer. A distribution-level estimator yields 6.5 to 16.9 per cent substantially LLM-modified review text across four conferences, concentrated in low-confidence reviews, near the deadline, and without engagement in the rebuttal [F40]. A quasi-experimental analysis finds that in paired reviews of the same paper the AI-assisted one scores higher in about 53 per cent of cases, and borderline submissions are about 5 percentage points likelier to be accepted [F41].

The strongest single measurement comes from a journal editorial with its own data: submissions up 42 per cent after ChatGPT; by February 2026 a majority of submissions carry detectable AI writing; readability down 1.28 standard deviations; desk rejection about 30 points higher above a 30 per cent threshold. The finding decisive for this map: **AI-generated reviews show no correlation with editorial decisions, while human reviews do** [F42]. The review performs a scrutiny that carries no information. That is the same figure as the occasioning case, one institutional level up.

2.9 Over-trust among experts

The figure has an effect, and the effect is measured. Incorrect AI suggestions in mammography degrade all experience strata, with the degree falling as experience rises — expertise attenuates but does not prevent [F43]. A randomised study with 44 physicians who had completed 20 hours of AI training finds, with planted errors, an adjusted difference of minus 14.0 points in diagnostic reasoning quality [F44]. The training did not protect them.

The single finding most relevant here comes from pathology: 7 per cent automation bias despite a net improvement; time pressure raises not the frequency but the severity — and **confidence gained during AI-assisted work predicts more reliance** [F45]. Displayed care therefore raises adoption precisely where it should not.

3. The solution side

Seven classes of mechanism claim to make reasoning checkable. The survey orders them by check channel, because the channel decides what a class can see at all.

3.1 Process supervision and step reward models

The first systematic comparison of outcome against step reward appeared in 2022 and shows step-level feedback reduces reasoning errors even at matched final accuracy [F47]. The method was established with a process reward model trained on human step annotations [F48] and later made cheap by automatic labels derived from how often completions rolled out from a step reach the correct final answer [F49]. A benchmark tests whether critics find the *earliest erroneous step* [F50].

Limit for this work: a step reward model scores learned textual plausibility. A confidently worded, never-executed procedure scores well; the second variant anchors the score in outcome agreement, not in execution. In the benchmark’s error taxonomy, the step “I ran the check and it passed” is not an error type.

3.2 Verifiers, generator–verifier separation, multiple sampling

The verifier can itself be generative and emit a verification trace before its verdict, so that trace and majority voting benefit it too [F51]. Multiple sampling with a majority vote over the final answer discards the reasoning [F52]. Test-time compute allocation across verifier-guided search and sequential revision can beat parameter scaling [F53]. Weak, weakly correlated verifiers can be combined by weak supervision into an estimate of latent correctness [F54]. A paper on verification dynamics finds that scaling the verifier alone does not overcome fundamental verification limits [F55].

Limit: the mechanism is agreement, not evidence. If a model has a stable disposition to claim an unexecuted procedure, all samples agree and multiple sampling ratifies the claim. The method is structurally blind to the case at issue.

3.3 Debate and multi-agent review

Several model instances propose, read each other’s reasoning, and revise over rounds [F56]. Two expert models arguing opposing answers before a non-expert judge raise judge accuracy substantially, and optimising the debaters for persuasiveness raises the judge’s truthfulness [F57]. **Decisively, the debaters have access to the source text and can quote verified excerpts** — the grounding carries the result, not debate as a form.

The counter-evidence is weighty. A comparison of debate protocols against simpler baselines finds that without hyperparameter tuning debate underperforms multiple sampling and ensembling, and only surpasses them after tuning [F58]. Reported debate gains are therefore partly a tuning artefact. A failure-mode taxonomy derived from annotated traces of real multi-agent executions finds specification, inter-agent alignment and task-verification failures as distinct classes [F59]; the verification-failure category sits immediately beside the question asked here, because such chains routinely terminate in an unverified completion claim.

3.4 Claim-level factuality checking and citation duty

Decomposing a text into atomic claims and then grounding each against a reliable knowledge source is the pattern of this class [F60]. One variant replaces the static corpus with an active retrieval loop over search queries [F61]. Small specialised models check claim against grounding document at close to large-model accuracy and far lower cost [F62] — cheap enough to run on every claim. A successor makes “is this claim checkable at all?” an explicit stage [F63]. The reference benchmark for citation-bearing generation checks whether the cited passage actually supports the statement [F64].

This is the class with the clearest transferability: **require a pointer to a trace location for every procedural claim and check support against that location, and the architecture that addresses the problem exists**. Without a trace the claim remains unverifiable — not false, but unverifiable, which is the precise situation in the single case.

3.5 Tool and solver grounding

The model emits a program as its reasoning chain and an interpreter performs the computation [F65]. Or it translates into a symbolic formulation, hands it to a deterministic solver, and reads the result back, with a refinement loop driven by solver error messages [F66]. Self-correction via tool feedback rather than introspection is the constructive counterpart to the finding in 2.5 [F67].

Here the gap closes by construction: the claim is settled by execution, not by a second model’s opinion. The price is scope — only formalisable claims.

3.6 Trace and provenance auditing

The most direct and most recent line. An agentic judge evaluates another agent’s *intermediate process* with access to the workspace it produced [F68]. A dataset of 148 human-annotated execution traces with a formal error taxonomy makes trace auditing measurable — with the finding that long-context models are poor at it [F69]. A further set of 1,302 expert-annotated trajectories audits the auditors: rule-based and LLM judges diverge from expert ground truth often enough that “the judge said it succeeded” is not evidence of success [F70].

The nearest finding on the merits: procedure-aware evaluation over utility, efficiency, interaction quality and **procedural integrity** finds that 27 to 78 per cent of reported benchmark successes conceal violations [F71]. Outcome-correct runs therefore routinely violated the procedure they were meant to follow. An audit architecture builds layered trace graphs with evidence nodes and workflow nodes, so that every claim maps to actions, artefacts and checks, and missing validation coverage becomes visible [F72]. A survey supplies the vocabulary layer and defines execution provenance as a typed graph [F73].

For the single case, however: this class presupposes instrumented execution. Without provenance capture there is no projection to compute, and the claim is unverifiable rather than merely unverified.

3.7 The one transferable finding

One paper characterises **false success**: the agent claims completion while the environment state shows failure. Its finding decisive for this map is methodological: lightweight TF-IDF detectors outperform LLM judges at spotting the case, at far lower cost [F74]. The signal the detector learns lies **in the language of the closing message** — and that is the only part transferable to a text without an environment.

3.8 Check-channel table

Class of mechanism	Check channel	Detects unexecuted procedure?
Step reward models (3.1)	learned textual plausibility	no — scores form, not execution
Verifiers, multiple sampling (3.2)	model against model, agreement	no — consensus ratifies stable invention
Debate, multi-agent review (3.3)	model against model, possibly source text	only with evidence-bearing debaters
Trace monitoring (2.1–2.2)	reading the externalised reasoning	conditionally — while the trace is not optimised against the monitor
Claim-level checking, citation duty (3.4)	decomposition plus comparison against an evidence source	yes, if the evidence source is the execution trace
Tool and solver grounding (3.5)	execution, deterministic oracle	yes in the formalisable part, by construction
Trace and provenance auditing (3.6)	environment and artefact records	yes — only with instrumented execution
Linguistic markers (3.7)	the text itself	the only channel left in the single case

4. The institutional side

The working note proposes pre-registering the generation procedure. This section examines which institutional instruments already exist for that and what they achieve.

4.1 Pre-registration as the model

The conceptual core is the distinction between prediction and postdiction: pre-registration works by fixing the analysis plan before outcomes are visible [F75]. Registered Reports place review of the protocol before the results, with in-principle acceptance and a second stage for adherence review [F76] — the only mechanism found in this survey that places independent review *before* the results. Blinded comparative ratings supply the effectiveness evidence [F77]. The strongest single number on the effect of outcome-dependent flexibility: 96 per cent positive results in standard reports against 44 per cent in Registered Reports [F78].

Two findings temper the transfer. First, a pre-registration can be present and still constrain nothing; a vague registration licenses any result [F79]. Second, and central here: an audit of whether published pre-registered studies followed their own registrations finds widespread undisclosed deviation [F80]. The instrument existed and was not kept. **Even a filed procedure needs an adherence check of its own** — and that is precisely what no current AI-disclosure rule performs.

4.2 HARKing as the template for M2

The canonical definition — presenting a post hoc hypothesis as a priori — transfers directly: the output text is formally indistinguishable from a genuine a priori statement, because the distinguishing feature is temporal and external to the text [F81]. That is the reason M2 can only yield an indication without an external fixing. Base rates of questionable research practices among humans supply the comparison without which the claim that AI worsens matters remains unquantified [F82].

One case belongs in the presentation because it bears on the working note’s framing: a paper whose thesis was that strictly prespecified methods yield replicable findings was retracted over the reporting of its own pre-registration [F83]. The gap between claim and evidence is therefore not an AI-specific pathology.

4.3 AI disclosure in scholarly publishing, as of 2026-10-06

Issuer	AI as author	Disclosure required	Where	Reviewer side	Status
ICMJE	no	yes, at submission	acknowledgements	not specified	undated [F84]
COPE	no	yes	methods or similar	—	v1, 2024-08-20 [F85]
Springer Nature	no	yes	not prescribed	permitted, no upload	traffic-light scheme 2026 [F86]
Elsevier	—	yes, separate section	named declaration section	supportive only	June 2026 [F87]
ACL	—	yes, section-level	acknowledgements	no AI first draft	rev. 2026-08-17 [F88]
NeurIPS 2025	—	only if methodologically central	not prescribed	no sharing	2025 [F90]
ICML 2026	—	authors choose reviewer policy A/B	—	no delegation of judgement	2026 [F91]
ICLR 2026	—	yes, all use	per Code of Ethics	confidentiality	2026 [F92]

Four observations:

First. The only publisher prohibition found that names the gap directly lists “generating hypotheses, analyses or conclusions and presenting them as human-derived” among the non-permitted acts [F86]. It is unenforceable without an external record, because the prohibited and the permitted act produce identical text.

Second. The NeurIPS threshold “important, original, or non-standard component” of the methodology explicitly exempts AI-written prose claiming a rigorous procedure from the disclosure duty [F90]. That is the clearest place where a scoping decision leaves the target gap uncovered.

Third. The ACL rule, with section-level attribution, is the finest found and raises the cost of a plausible false declaration, but verifies nothing [F88]. The ARR mandatory checklist with desk rejection as the sanction is the nearest model for a submission-time procedure declaration — but what is sanctioned is completeness of the form, not truth of the answers [F89].

Fourth, and this is the finding with the greatest follow-on value. The only instrument in the survey that asks for inputs and outputs to be deposited is the European Commission’s guidance: researchers should, “if relevant”, make prompts and outputs available, and the reference should include name, version, date and how the tool affected the research process [F93]. Deposited prompt logs would in principle expose a procedure that was never run. The request is conditional and unsanctioned, so in practice it functions as a recommendation.

4.4 Technical provenance and watermarking

The C2PA architecture binds a record not to author prose but to a cryptographic signature: a manifest comprises assertions, a claim referencing them, and a claim signature, with trust deriving from the signer identity vouched for via X.509 [F94]. A conformance programme with a public trust list has existed since 2026, with a documented cutover from the interim list on 2026-01-01 [F95]. The limit for this work: it would attest *who signed* and *which tool touched the file*, not whether the described analysis took place. Text is also not covered in the way image and video are.

Watermarking is deployable and demonstrated at production scale [F96, F97], but answers a different question: machine origin, not procedure. A watermark-positive text may describe a real procedure; a watermark-negative one may be wholly fabricated. Two hard limits follow. Recursive paraphrasing substantially lowers detection rates at little quality cost, and spoofing makes human text appear watermarked [F98]. And under stated, natural assumptions strong watermarking is **provably impossible**, with a generic attack needing only a quality and a perturbation oracle [F99].

This is the most load-bearing negative result on the technical half: a disclosure regime cannot be founded on post-hoc origin detection. The argument is thereby forced toward signed, author-side records of what was actually executed — exactly the direction the working note takes with pre-registration.

4.5 Documentation instruments

The standing W3C vocabulary for machine-readable process records has existed since 2013 and remains the only model at Recommendation status: provenance as a record of the people, institutions, entities and activities involved [F100]. A PROV-serialised log of an AI-assisted workflow would make “what was actually executed” queryable rather than narrated. It is, however, a pure representation standard: an invented PROV graph is as expressible as a true one without a signature layer.

The genre of disclosure documents originates in model cards [F101] and datasheets [F102], the latter explicitly carrying the collection process as a field. Both document the tool or the data, not its use in a particular piece of work. The most concrete published proposal for a standardised, **machine-readable** AI-use declaration attached to a research output is AI Usage Cards, with questionnaire-driven generation [F103]; machine-readability is the operative advance, because declarations then become auditable across a corpus rather than readable one at a time. The finest domain reporting guideline runs to 19 items and 50 sub-items and requires prompts, model versions and evaluation design [F104] — the granularity raises the cost of invention considerably without adding a check.

4.6 What actually bites on a single text

Two instruments have teeth, because they test an internal consistency condition that is not co-generated with an invention.

Automated recomputation of p-values from reported test statistics and degrees of freedom tests exactly that condition [F105]. Sharper still is the granularity test on reported means: a mean of integer data must be a multiple of $1/n$, so impossible means are detectable from the printed numbers alone — no raw data, no author cooperation, no registration [F106]. That is the sharpest single-text check in the survey, within a narrow scope.

Against it stands the clearest example of the gap under study: automated checking of whether a methods section *reports* the rigour criteria — randomisation, blinding, power analysis, sex as a variable, reagent identifiability [F107] — assesses whether a procedure is *described*. A fabricated methods section scores well precisely because it was written to satisfy the checker. Corpus-scale screening with several detectors plus human adjudication is the realistic model for integrity work [F108], but operates on surface features; a fluent text with a fabricated procedure and no tortured phrasing passes unflagged.

4.7 Scale

More than 10,000 retractions in 2023 show that integrity checking has long operated beyond the capacity of case-by-case review [F109]; the cause there is predominantly industrial production, not AI text, and the two must not be conflated. Current operating figures for shared pre-publication screening: more than 35 publishers screen over 125,000 papers a month and intercept around 1,000 suspect submissions [F110] — an interception rate of about 0.8 per cent. The strongest corpus measurement of undeclared LLM use finds, across more than 15 million abstracts, an abrupt rise in certain style words implying at least 13.5 per cent LLM-processed abstracts in 2024 and up to 40 per cent in some subfields [F111]. The method is purely distributional, so it assigns no probability to any individual paper — and it measures writing assistance, not procedural execution.

5. Agentic research systems

This strand supplies the case material. The working note’s occasioning case is a hand-steered single run; the systems named here make the same claim with a built architecture. The question is identical in both cases: does the *output text* let a reader see that the claimed internal procedure took place?

5.1 End-to-end systems

One system claims the closed loop from idea generation through execution to manuscript, with an automated reviewer; the journal version names one generated manuscript that passed a workshop’s first review round [F113]. The successor replaces the linear chain with agentic tree search under an experiment manager [F114]. **That part in particular leaves no trace in the manuscript:** nothing in a generated paper distinguishes a result reached by a wide search with many discarded branches from one reached in a single pass. That is M1 at system level.

A three-stage assistant accepts human feedback at each stage and reports an 84 per cent cost reduction [F115]; the autonomous and the assisted condition must be cited separately, because the stage boundaries are architectural and the final report does not reveal which stage produced which claim. Another system pairs a generating model with a simulated reviewer trained by feedback; the reported gains are **simulated review scores**, generator and evaluator coming from the same training pipeline [F116]. Datasets and weights are released, so the artefacts are at least inspectable.

One objection from the same quarter belongs to the handling of these claims: a position paper analyses 28 papers from five such systems and locates the bottleneck not in idea generation but in the systems' capability to execute the verification procedures their claims require [F134]. The author overlap with one of the systems examined should be stated — it is criticism from inside, not an independent audit.

5.2 Where the claim ends physically

The laboratory systems are the instructive counter-case, because their claimed procedure terminates in a measurement. One system designs, plans and executes experiments via web search, code execution and liquid handling [F117]. Another couples a model with 18 domain tools and has it evaluated **by experts rather than by a model** — with a side finding central to this map: the model as evaluator “cannot distinguish between clearly wrong completions and the system's performance”, and model-based evaluation favours fluency over accuracy [F118]. That is direct evidence that the automated reviewer component of the systems in 5.1 measures something other than correctness.

The most instructive case is a robotics platform that targeted 41 novel compounds and successfully synthesised 36 of 57 [F119] — together with the external re-analysis arguing that automated powder-diffraction analysis is not yet reliable and that disorder is neglected in the predictive step, so that several of the reported novel compounds are contested at the characterisation stage [F120]. An autonomy claim here survives its own internal evaluation and fails external expert re-analysis. That is the clearest worked example of what the reported gains of such systems actually measure.

Worth naming positively is a system whose claimed internal procedure is the **meeting transcript itself**: a model as principal investigator convenes specialist agents in iterative research meetings; 92 candidates were designed and experimentally validated, two with improved binding [F121]. The meeting records were released alongside, so the critique and revision rounds are inspectable rather than merely asserted. It is also framed as collaboration, not autonomy.

5.3 Idea generation and novelty assessment

A multi-agent system runs a tournament with Elo-style ranking over hypotheses and is validated in three application areas with laboratory confirmations [F122]. The tournament is the central claim and is reported in aggregate: a reader seeing a single output hypothesis cannot tell what it beat, how many rounds it survived, or what the critic agent objected to. The laboratory confirmations carry the finding, not the Elo curves. Another system couples ontological knowledge graphs with agents and claims in its abstract capabilities “exceeding traditional human-driven research methods”, without naming the comparison or the metric [F123] — an unanchored superiority claim, to be marked as such here rather than repeated. A third simulates a research team and investigates which collaboration mechanisms raise novelty [F124]; the comparison is system-internal, so it measures the architecture's effect on the metric, not on research outcomes.

The earliest work to make novelty an explicit optimisation target retrieves inspirations from prior work and compares candidates iteratively against existing papers up to a novelty threshold — and reports frankly that a model alone produces technically shallow, low-novelty ideas, which the method only partly fixes [F125]. That candour is the exception in this subfield.

A benchmark with single-keyword prompts across 1,180 keywords and 22 domains adds a finding that bears on model choice: idea-generation scores are poorly predicted by general capability metrics, and mid-sized models match frontier ones [F128]. The assumption that a stronger backbone yields better research is therefore not supported. The scoring is itself model-based, so the finding in [F127] applies to this benchmark too.

The single most important finding of this strand for the working note is a negative result from October 2026: small changes in wording shift novelty-judgement accuracy by over 50 percentage points; retrieval augmentation and longer reasoning help little; purpose-built novelty evaluators fall below simple baselines [F127]. That puts every reported novelty gain in 5.1 and 5.3 in question, including the blind expert comparison in which AI ideas were judged more novel but slightly less feasible [F129]. The counterexample for checkable design is a retrieve-then-rerank procedure whose output is a verdict **plus the comparison papers it rests on**; there it can be checked whether the cited prior work says what the verdict assumes [F126].

5.4 Benchmarks and how to read them

Four benchmarks for agentic research ability were verified. Two misreadings are to be avoided.

First: the widely cited lead of agentic systems over experts holds at a two-hour budget; at eight hours humans overtake, at 32 hours they double [F132]. The time-horizon dependence disappears when the lead figure is cited alone. Second: a benchmark built from public competition solutions partly measures recall of published write-ups rather than engineering, which its own contamination analysis addresses explicitly [F130]. A rubric-based decomposition of 20 papers into 8,316 individually gradable sub-tasks is the strongest auditability property in this section, because it becomes visible which sub-claim failed rather than only a scalar — the grading is itself model-based, which the paper states openly [F131]. A fourth benchmark positions computational reproducibility as a precondition for novel-research systems, i.e. as the capability floor the systems in 5.1 assume without demonstrating [F133].

5.5 What venues are doing in 2026

The venues' response is the institutional part that bears directly on publishing the working note. One conference states that papers with extensive undisclosed LLM use will be desk rejected, that authors remain responsible for outputs, and that **fabricated citations constitute a Code of Ethics violation** warranting rejection [F136]. Another reports 178 submissions (18.4 per cent) desk-rejected without appeal and 123 (12.7 per cent) asked to evidence human authorship, at a detector rate of 28.2 per cent at default settings, falling to 12.7 per cent after window adjustment [F137]. The gap between the two settings is itself the instructive datum: the reported rejection rate is a function of a tuning choice, and the post says so.

The largest documented countermove is a reviewing aid applied to all 22,977 submissions of one conference in under a day, with favourable ratings in author and programme-committee surveys [F135] — written by the pilot's organisers, so the favourable survey results are self-reported. One venue requires AI authorship rather than merely permitting it, and releases the AI reviewers' prompts and reviews as a community resource [F138] — the only venue found that treats the automated review trace as a publishable artefact. That is the institutional form of what this work asks for.

5.6 The occasioning case in its domain

The occasioning case concerns cell-free biocatalysis and enzyme cascades. The survey found two verified platform papers in that domain, both machine-learning guided, one of them with a natural-language interface [F139, F140]. **No verified case was found of an end-to-end AI research system using an enzyme cascade as its worked example.** The occasioning case is therefore not a reconstruction of an existing system but a hand-steered single run producing the same rhetoric. That supports the working note's decision to treat it as case material rather than subject matter.

6. Location: where the gap is

After this survey the working note's position can be stated precisely.

What is not new. That claimed and actual procedure come apart is described in nine separate literatures (section 2). That self-correction without a counterpart does not work reliably is established [F25, F26, F27]. That automated review rewards fluency rather than correctness is established [F32, F118]. That models' novelty judgements are unstable is established [F127]. The working note must claim none of these as its own finding.

What is genuinely open. Five gaps were named as such independently in all four strands:

1. **No checking instrument at document level.** There is checking at citation level [F37], at corpus level [F111, F42] and on surface quality [F39]. There is no published rubric that extracts a paper's procedural claims and tests them for reproducible support — the document-level counterpart of the intervention method in [F2].
2. **No work classifying textual features of unexecuted procedure descriptions in a single, uninstrumented run.** The closest still uses the environment verdict as its label source [F74].
3. **No experiment testing visible apparent rigour as the independent variable in over-trust.** The over-trust literature varies the *correctness* of the advice [F43, F44, F45]; the trace literature measures the model side [F11, F12]. Nobody has asked whether a long, deliberative-looking trace raises adoption of a wrong conclusion relative to a bare answer. That is the empirical centre of the topic, and it is empty.
4. **No procedure attestation for AI-assisted text work.** An extended search found no scholarly or normative use of the term in this sense. The nearest genuine instruments are [F91], [F93], [F103], [F104] and [F112]. **No pre-registration equivalent exists for AI-assisted writing.**
5. **No measurement of how well human reviewers detect fabricated references.** Machine verifiers are tested [F37]; the hit rate of the body that decides whether the rates in [F33] to [F36] turn into published error is unknown.

Where the working note sits. It addresses gap 2 with five markers and a tool, and proposes a mechanism for gap 4. It does not address gap 1 but prepares it. Gaps 3 and 5 are follow-on projects, not part of the work.

A correction to the work's own claims. The note's sentence that *Iteratio Simulata* is the only one of the three neighbouring concepts testable from the text alone holds after this survey but needs qualifying: it holds against *unfaithful chain-of-thought* and *potemkin understanding*, not against the 2026 work on the performative trace, which uses probing and timing and therefore requires model access. That it requires model access is why the markers continue to stand beside it; the exclusivity claim is to be softened accordingly (see 7.1).

7. Ten follow-on projects

Ten projects follow from the survey. For each, the reason, the scope and the precondition are given. Five are executable without further resources.

7.1 Addendum to the working note (small, immediate)

Four points for the next version: the 2026 block of findings on the performative trace as empirical support for the central concept [F11, F12, F13]; the counterposition on the verbalisation question [F7]; the softening of the exclusivity claim per section 6; the note that M1 at system level hits exactly the tree search that no generated manuscript discloses [F114]. **Precondition:** none.

7.2 Document rubric “procedure evidence” (medium, executable)

Gap 1. A rubric that extracts a paper’s procedural claims and classifies each by support type: executed and evidenced, described and evidenceable, described and unverifiable, contradicted. The model is the rubric-based sub-task decomposition [F131], the check mechanism the citation-support test [F63, F61]. The output is a tally sheet, not a score. **Precondition:** 20 to 30 papers as a trial set; no compute load.

7.3 Linguistic single-text detector (medium, executable)

Gap 2, with the only transferable model [F74]. Its finding — simple TF-IDF features beat model judges at spotting false completion claims — applies directly to markers M1 and M2, because both work through tracelessness. Concretely: set the Iterationsprüfer’s existing vocabulary check against a weighted feature model trained on closing-message language. **Precondition:** the test set from 7.4.

7.4 Test corpus with independent pre-classification (medium, blocking)

The point already listed as open in the working note, here confirmed as a precondition for 7.2, 7.3 and 7.6. At least 20 texts, pre-classified by humans without knowledge of the marker definitions; negative cases first, per the standing rule from the implementation finding. The corpus should contain careful *and* performative texts for each marker, because the tool’s first counter-test failed on careful texts. **Precondition:** two to three independently judging people.

7.5 Over-trust experiment (large, external)

Gap 3 and the scientifically most valuable item on this map. Design: experts receive the same wrong conclusion once as a bare answer, once with a long deliberative-looking trace; the dependent variable is adoption. The design follows the automation-bias studies [F43, F44, F45] but varies the *display of procedure* rather than the correctness of the advice. The finding that confidence gained during AI-assisted work predicts more reliance [F45] supplies the prediction. **Precondition:** ethics approval, an expert sample, a partner institution. Not executable alone.

7.6 Markers M3 and M3b against open metadata services (small, executable)

The work listed as phase 1c in the working note, here with an added caution: the checker must not itself fall into the false-positive trap measured for papers past the training cutoff [F37]. Consequence for implementation: registry resolution (DOI, OpenAlex) instead of model judgement, and an explicit status “not locatable at this data state” kept separate from “non-existent”. **Precondition:** licence check of the services bound in; no data holding of our own.

7.7 Procedure declaration as a submittable artefact (medium, executable)

Gap 4, the contribution with the greatest institutional traction. The components all exist: the mandatory checklist with desk rejection as the enforcement form [F89]; the ex-ante procedural binding in which the author fixes and the reviewer is bound [F91]; deposition of inputs and outputs [F93]; the machine-readable usage declaration as format [F103]; the standing provenance vocabulary as serialisation [F100]; the signature layer without which an invented graph is as expressible as a true one [F94]. What has to be built is the assembly, not a component. **Precondition:** none for the draft; effect only with a receiving body.

7.8 Reviewer hit rate on fabricated references (medium, external)

Gap 5. A set of manuscripts with controlled reference errors at three severities — non-existent, existent with wrong authorship, existent with wrong fields — and measurement of which level is caught. The severities follow the finding that real sources with wrong fields are the harder case [F33]. **Precondition:** cooperation with an editorial office; ethically feasible only with prior notice.

7.9 Replication of the potemkin procedure in German (small, executable)

A gap named explicitly in strand A: no replication outside English, and none on practitioner rather than academic concepts. For knowledge work in the German-language technical and regulatory environment this is directly relevant, and the procedure is published [F14]. **Precondition:** a concept set from one domain.

7.10 Procedural integrity in our own workflow (small, immediate)

The finding that 27 to 78 per cent of outcome-correct runs conceal procedural violations [F71] is to be applied to our own working method. Concretely: do not merely recommend the working note’s four pre-registration steps but evidence them on our own bundle — with a hash of the candidate list and of the criteria taken before the selection, as a co-published file. This is the self-application that makes 7.7 credible. **Precondition:** none.

8. Not found, not verified, to be noted

8.1 Genuine gaps

Beyond the five named in section 6: no fabrication rates for the humanities and social sciences; no sycophancy study across extended working sessions rather than a single push-back; nothing on AI slop in grant or standards review; no verified systematic review of the agentic-research-systems strand; no completed study in which the ideas generated in [F129] were actually executed — the central open question of the entire novelty literature.

8.2 Not fetchable, carried as unverified

Eleven entries could not be confirmed at their primary page and are marked [T] in the reference list. Affected among others: one publisher's editorial policies, whose pages refuse automated fetches with HTTP 403; the wording of Article 50 of Regulation (EU) 2024/1689, whose HTML and PDF renderings both truncate before the articles; the journal version of one sycophancy paper; the published final version of a reporting guideline of which only the protocol is verified; the author-list form of a further guideline taken from a repository record; a publisher page of a re-analysis whose metadata were read off an institutional record; two thematically central papers on overclaiming whose pages returned empty on repeated fetches; a foundational paper on formal proof grounding whose fetch was refused with HTTP 429; a review of self-driving laboratories whose volume and page data were not printed; a preprint on closed-loop optimisation of enzyme cascades that would be exactly on target and was not fetched.

On one fetch the tool returned plausible-looking quotations from a legal text and flagged them itself as reconstructed from structure. They were discarded. That episode is why [T] is carried as a status rather than replaced by an omission.

8.3 Industry claims excluded

Detector vendors' figures on shares of AI-generated reviews and on hallucination counts appeared in search results. They are vendor publications with undisclosed methods and are therefore not in the reference list. They would be citable only as vendor claims. The same applies to the figures of a shared screening infrastructure that does not disclose its detectors [F110]; the numbers stand, their evidential weight cannot be assessed from the primary source.

8.4 Version divergence

Four works changed title or author list between preprint and publication: an end-to-end research system (6 versus 8 authors, different order, different title); a multi-agent system (different titles in v1 and v2); a novelty benchmark (renamed; the circulating name is the benchmark name, not the title); an idea-generation system (circulating title differs from the printed one). Anyone taking these from a secondary source dated 2024 or 2025 is citing a stale record.

8.5 Upgrades declined

For several widely cited works the abstract page names no venue. They are carried as preprints. A later editing pass must not silently upgrade them; where needed, the venue is to be checked per entry at the proceedings page.

9. Self-application

The working note requires that a work about simulated care disclose its own errors. Three findings arise for this survey.

9.1 Two identifiers held from memory were wrong

In the provenance and disclosure strand, two identifiers were known in advance and were nevertheless checked. Both were wrong. One DOI returned a 404; the correct one differs in the final digit, and **the paper is retracted** — a circumstance that would have gone undetected had it been cited from memory. One arXiv identifier led to an astrophysics paper on particle acceleration; the correct identifier is a different one.

This is the pattern from [F33] and [F36] occurring in our own process: the fields that look precise and are rarely checked are the wrong ones. It is also the strongest available single piece of evidence for the working note's thesis — the errors were found solely because every entry was fetched, not by plausibility assessment.

9.2 The survey had to withdraw its own superiority claim

The working note's exclusivity claim — that Iteratio Simulata is the only neighbouring concept testable from the text alone — holds against the two neighbouring concepts available in 2025 and does not hold with the same breadth against the 2026 work. The sentence appeared in the main work, the pamphlet, the lay explanation and both flyers. It will be qualified rather than struck in the next version, and the episode is recorded here (section 6, final paragraph; follow-on project 7.1).

This is an M3 case by the work's own definition: the state of the field was further along than the work depicted it.

9.3 This map's auditability verdicts are inferences, not findings

For the agentic-research-systems strand, abstract pages and metadata were read, not full texts. Every statement about whether a system's output text reveals rounds, selection and critique is an inference from the described architecture and **not a verified finding**. Testing those inferences requires opening and reading the released artefacts of the four systems with open traces. That was not done. It is the statement on which this map is most exposed, and the point at which it is itself subject to M1: it asserts a verification verdict whose verification step it did not carry out.

An addition to follow-on project 7.2 follows: the first trial set for the document rubric is exactly these four systems, because they are the only cases in which claimed procedure and published trace can be compared.

10. References

Status tags: [Z] peer-reviewed journal · [K] conference paper · [W] workshop paper · [P] preprint, no venue on the abstract page · [D] policy, standard, position or working paper · [N] press or practitioner article · [T] not confirmed at the primary page; to be checked individually before any journal submission. Every item was fetched at the address given on 2026-10-06 and read off there. No item comes from memory.

Strand A — Failure modes: process, not fact

[F1] Miles Turpin, Julian Michael, Ethan Perez, Samuel R. Bowman. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388, 2023; comments field: NeurIPS 2023. <https://arxiv.org/abs/2305.04388> [P] [F2] Tamera Lanham, Anna Chen, Ansh Radhakrishnan et al. (30). *Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv:2307.13702, 2023. <https://arxiv.org/abs/2307.13702> [P] [F3] Yanda Chen, Joe Benton, Ansh Radhakrishnan et al. (15). *Reasoning Models Don't Always Say What They Think*. arXiv:2505.05410, 2025. <https://arxiv.org/abs/2505.05410> [P] [F4] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooan Rajamanoharan, Neel Nanda, Arthur Conmy. *Chain-of-Thought Reasoning In The Wild Is Not Always Faithful*. PMLR 306 (ICML 2026). <https://proceedings.mlr.press/v306/arcuschin26b.html> [K] [F5] Jacob Pfau, William Merrill, Samuel R. Bowman. *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models*. arXiv:2404.15758, 2024. <https://arxiv.org/abs/2404.15758> [P] [F6] Fazl Barez, Tung-Yu Wu, Iván Arcuschin et al. (16). *Chain-of-Thought Is Not Explainability*. Working Paper, Oxford Martin AI Governance Initiative, 2025. <https://www.oxfordmartin.ox.ac.uk/publications/chain-of-thought-is-not-explainability> [D] [F7] Kerem Zaman, Shashank Srivastava. *Is Chain-of-Thought Really Not Explainability? Chain-of-Thought Can Be Faithful without Hint Verbalization*. Proc. 64th ACL (Vol. 1: Long Papers), 2026, S. 48008-48030. <https://aclanthology.org/2026.acl-long.2217/> [K] [F8] Tomek Korbak, Mikita Balesni, Elizabeth Barnes et al. (41). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. arXiv:2507.11473, 2025. <https://arxiv.org/abs/2507.11473> [P] [F9] Bowen Baker, Joost Huizinga, Leo Gao et al. (9). *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation*. arXiv:2503.11926, 2025. <https://arxiv.org/abs/2503.11926> [P] [F10] Scott Emmons, Erik Jenner, David K. Elson et al.. *When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors*. arXiv:2507.05246, 2025. <https://arxiv.org/abs/2507.05246> [P] [F11] Wenkai Li, Fan Yang, Ananya Hazarika, Shaunak A. Mehta, Koichi Onoue. *When Reasoning Traces Become Performative: Step-Level Evidence that Chain-of-Thought Is an Imperfect Oversight Channel*. arXiv:2605.11746, 2026. <https://arxiv.org/abs/2605.11746> [P] [F12] Siddharth Boppana, Annabel Ma, Max Loeffler et al. (8). *Reasoning Theater: Disentangling Model Beliefs from Chain-of-Thought*. arXiv:2603.05488, 2026. <https://arxiv.org/abs/2603.05488> [P] [F13] Vatsal Baherwani, Tom Goldstein, Ashwinee Panda. *Not All LLM Reasoning is Visible in the Chain-of-Thought*. arXiv:2607.22925, 2026. <https://arxiv.org/abs/2607.22925> [P] [F14] Marina Mancoridis, Bec Weeks, Keyon Vafa, Sendhil Mullainathan. *Potemkin Understanding in Large Language Models*. PMLR 267 (ICML 2025). <https://proceedings.mlr.press/v267/mancoridis25a.html> [K] [F15] Zhaofeng Wu, Linlu Qiu, Alexis Ross et al. (9). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*. arXiv:2307.02477, 2023; NAACL 2024. <https://arxiv.org/abs/2307.02477> [K] [F16] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, Mehrdad Farajtabar. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv:2410.05229, 2024; comments field: ICLR camera ready. <https://arxiv.org/abs/2410.05229> [P] [F17] Parshin Shojaei, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar. *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. arXiv:2506.06941, 2025; comments field: NeurIPS 2025 camera-ready. <https://arxiv.org/abs/2506.06941> [P] [F18] Lisa Messeri, M. J. Crockett. *Artificial intelligence and illusions of understanding in scientific research*. Nature 627, 49-58, 2024 (Perspective), DOI 10.1038/s41586-024-07146-0. <https://www.nature.com/articles/s41586-024-07146-0> [Z] [F19] Ethan Perez, Sam Ringer, Kamile Lukošiušė et al. (60). *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251, 2022. <https://arxiv.org/abs/2212.09251> [P] [F20] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, Quoc V. Le. *Simple synthetic data reduces sycophancy in large language models*. arXiv:2308.03958, 2023. <https://arxiv.org/abs/2308.03958> [P] [F21] Mrinank Sharma, Meg Tong, Tomasz Korbak et al. (19). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548, 2023. <https://arxiv.org/abs/2310.13548> [P] [F22] Aaron Fanous, Jacob Goldberg, Ank A. Agarwal et al. (7). *SycEval: Evaluating LLM Sycophancy*. arXiv:2502.08177, 2025; comments field: AIES 2025. <https://arxiv.org/abs/2502.08177> [P] [F23] Myra Cheng, Sunny Yu, Cino Lee, Pranav Khadpe, Lujain Ibrahim, Dan Jurafsky. *ELEPHANT: Measuring and understanding social sycophancy in LLMs*. arXiv:2505.13995, 2025. The abstract page names a journal version; it was not fetchable and is not confirmed. <https://arxiv.org/abs/2505.13995> [T] [F24] Aman Madaan, Niket Tandon, Prakhar Gupta et al. (16). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv:2303.17651, 2023.

<https://arxiv.org/abs/2303.17651> [P] [F25] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, Denny Zhou. *Large Language Models Cannot Self-Correct Reasoning Yet*. arXiv:2310.01798, 2023; comments field: ICLR 2024. <https://arxiv.org/abs/2310.01798> [P] [F26] Kaya Stechly, Matthew Marquez, Subbarao Kambhampati. *GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems*. arXiv:2310.12397, 2023. <https://arxiv.org/abs/2310.12397> [P] [F27] Karthik Valmeekam, Matthew Marquez, Subbarao Kambhampati. *Can Large Language Models Really Improve by Self-critiquing Their Own Plans?*. arXiv:2310.08118, 2023. <https://arxiv.org/abs/2310.08118> [P] [F28] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng et al. (13). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. NeurIPS 2023 Datasets and Benchmarks Track. <https://arxiv.org/abs/2306.05685> [K] [F29] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, Zhifang Sui. *Large Language Models are not Fair Evaluators*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 9440-9450. Author list per ACL Anthology; the arXiv v1 list is incomplete. <https://aclanthology.org/2024.acl-long.511/> [K] [F30] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, Dongyeop Kang. *Benchmarking Cognitive Biases in Large Language Models as Evaluators*. arXiv:2309.17012, 2023; comments field: ACL 2024. <https://arxiv.org/abs/2309.17012> [K] [F31] Arjun Panickssery, Samuel R. Bowman, Shi Feng. *LLM Evaluators Recognize and Favor Their Own Generations*. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Main Conference Track. https://papers.nips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html [K] [F32] Koki Wataoka, Tsubasa Takahashi, Ryokan Ri. *Self-Preference Bias in LLM-as-a-Judge*. arXiv:2410.21819, 2024; comments field: NeurIPS 2024 Safe Generative AI Workshop. <https://arxiv.org/abs/2410.21819> [W] [F33] William H. Walters, Esther Isabelle Wilder. *Fabrication and errors in the bibliographic citations generated by ChatGPT*. Scientific Reports 13, Artikel 14045, 2023, DOI 10.1038/s41598-023-41032-5. <https://doi.org/10.1038/s41598-023-41032-5> [Z] [F34] Mikael Chelli, Jules Descamps, Vincent Lavoué et al. (10). *Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis*. Journal of Medical Internet Research 26, e53164, 2024, DOI 10.2196/53164. Metadata via the DOAJ record; the publisher page returned no content. <https://doi.org/10.2196/53164> [Z] [F35] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, Daniel E. Ho. *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. Journal of Empirical Legal Studies 22, S. 216, 2025. <https://law.stanford.edu/publications/hallucination-free-assessing-the-reliability-of-leading-ai-legal-research-tools> [Z] [F36] Delip Rao, Chris Callison-Burch. *BibTeX Citation Hallucinations in Scientific Publishing Agents: Evaluation and Mitigation*. arXiv:2604.03159, 2026. <https://arxiv.org/abs/2604.03159> [P] [F37] Patrik Reizinger, Wieland Brendel. *HALLMARK: Diagnosing Three Failure Modes in LLM Citation Verifiers*. arXiv:2607.18360, 2026. <https://arxiv.org/abs/2607.18360> [P] [F38] Kate Niederhoffer, Gabriella Rosen Kellerman, Angela Lee, Alex Liebscher, Kristina Rapuano, Jeffrey T. Hancock. *AI-Generated 'Workshop' Is Destroying Productivity*. Harvard Business Review, 22.09.2025. Practitioner article, not peer reviewed; sample size not stated. <https://hbr.org/2025/09/ai-generated-workshop-is-destroying-productivity> [N] [F39] Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, Byron C. Wallace. *Measuring AI 'Slop' in Text*. arXiv:2509.19163, 2025. <https://arxiv.org/abs/2509.19163> [P] [F40] Weixin Liang, Zachary Izzo, Yaohui Zhang et al. (12). *Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews*. arXiv:2403.07183, 2024; comments field: ICML '24. <https://arxiv.org/abs/2403.07183> [K] [F41] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, Robert West. *The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates*. arXiv:2405.02150, 2024. <https://arxiv.org/abs/2405.02150> [P] [F42] Claudine Gartenberg, Shariq Hasan, Alex Murray, Lamar Pierce. *More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review*. Organization Science 37(3), 2026, DOI 10.1287/orsc.2026.ed.v37.n3. Editorial by the journal's AI Task Force, not a peer-reviewed article. <https://pubsonline.informs.org/doi/10.1287/orsc.2026.ed.v37.n3> [D] [F43] Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi et al. (10). *Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance*. Radiology 307(4), 2023, DOI 10.1148/radiol.222176. <https://doi.org/10.1148/radiol.222176> [Z] [F44] Ihsan Ayyub Qazi, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, Muhammad Hamad Alizai. *Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians*. medRxiv, 08.09.2025, DOI 10.1101/2025.08.23.25334280; trial registry NCT06963957. Preprint, not peer reviewed. <https://www.medrxiv.org/content/10.1101/2025.08.23.25334280v2> [P] [F45] Emely Rosbach, Jonas Ammeling, Jonathan Ganz et al. (7). *Stuck on Suggestions: Automation Bias, the Anchoring Effect, and the Factors That Shape Them in Computational Pathology*. arXiv:2603.11821, 2026; comments field: angenommen bei MELBA. <https://arxiv.org/abs/2603.11821> [P] [F46] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, Francis Rhys Ward. *AI Sandbagging: Language Models can Strategically Underperform on Evaluations*. arXiv:2406.07358, 2024. <https://arxiv.org/abs/2406.07358> [P]

Strand B — Methods, verifiers, architectures

[F47] Jonathan Uesato, Nate Kushman, Ramana Kumar et al.. *Solving math word problems with process- and outcome-based feedback*. arXiv:2211.14275, 2022. <https://arxiv.org/abs/2211.14275> [P] [F48] Hunter Lightman, Vineet Kosaraju, Yura Burda et al. (10). *Let's Verify Step by Step*. arXiv:2305.20050, 2023. The abstract page names no venue. <https://arxiv.org/abs/2305.20050> [P] [F49] Peiyi Wang, Lei Li, Zhihong Shao et al.. *Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 9426-9439. <https://aclanthology.org/2024.acl-long.510/> [K] [F50] Chujie Zheng, Zhenru Zhang, Beichen Zhang et al.. *ProcessBench: Identifying Process Errors in Mathematical Reasoning*. arXiv:2412.06559, 2024; comments field: ACL 2025. <https://arxiv.org/abs/2412.06559> [P] [F51] Lunjun Zhang, Arian Hosseini, Hritik Bansal et al.. *Generative Verifiers: Reward Modeling as Next-Token Prediction*. arXiv:2408.15240, 2024; comments field: ICLR 2025. <https://arxiv.org/abs/2408.15240> [P] [F52] Xuezhi Wang, Jason Wei, Dale Schuurmans et al.. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv:2203.11171, 2022; comments field: ICLR 2023. <https://arxiv.org/abs/2203.11171> [P] [F53] Charlie Snell, Jaehoon Lee, Kelvin Xu, Aviral Kumar. *Scaling LLM Test-Time*

Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314, 2024. <https://arxiv.org/abs/2408.03314> [P] [F54] Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen et al.. *Shrinking the Generation-Verification Gap with Weak Verifiers*. arXiv:2506.18203, 2025; comments field: NeurIPS 2025. <https://arxiv.org/abs/2506.18203> [P] [F55] Yefan Zhou, Austin Xu, Yilun Zhou et al.. *Variation in Verification: Understanding Verification Dynamics in Large Language Models*. arXiv:2509.17995, 2025; comments field: ICLR 2026. <https://arxiv.org/abs/2509.17995> [P] [F56] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. arXiv:2305.14325, 2023. <https://arxiv.org/abs/2305.14325> [P] [F57] Akbir Khan, John Hughes, Dan Valentine et al. (10). *Debating with More Persuasive LLMs Leads to More Truthful Answers*. arXiv:2402.06782, 2024. <https://arxiv.org/abs/2402.06782> [P] [F58] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, Arnu Pretorius. *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. arXiv:2311.17371, 2023. <https://arxiv.org/abs/2311.17371> [P] [F59] Mert Cemri, Melissa Z. Pan, Shuyi Yang et al. (13). *Why Do Multi-Agent LLM Systems Fail?*. arXiv:2503.13657, 2025. <https://arxiv.org/abs/2503.13657> [P] [F60] Sewon Min, Kalpesh Krishna, Xinxu Lyu et al. (9). *FACTScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv:2305.14251, 2023; comments field: EMNLP 2023 main conference. <https://arxiv.org/abs/2305.14251> [K] [F61] Jerry Wei, Chengrun Yang, Xinying Song et al. (12). *Long-form factuality in large language models*. NeurIPS 2024 (Poster). <https://neurips.cc/virtual/2024/poster/96675> [K] [F62] Liyan Tang, Philippe Laban, Greg Durrett. *MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents*. arXiv:2404.10774, 2024; comments field: EMNLP 2024. <https://arxiv.org/abs/2404.10774> [K] [F63] Yixiao Song, Yekyung Kim, Mohit Iyyer. *VERISCORE: Evaluating the factuality of verifiable claims in long-form text generation*. arXiv:2406.19276, 2024. <https://arxiv.org/abs/2406.19276> [P] [F64] Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen. *Enabling Large Language Models to Generate Text with Citations*. arXiv:2305.14627, 2023; comments field: EMNLP 2023. <https://arxiv.org/abs/2305.14627> [K] [F65] Luyu Gao, Aman Madaan, Shuyan Zhou et al. (8). *PAL: Program-aided Language Models*. arXiv:2211.10435, 2022. <https://arxiv.org/abs/2211.10435> [P] [F66] Liangming Pan, Alon Albalk, Xinyi Wang, William Yang Wang. *LogicLM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning*. arXiv:2305.12295, 2023; comments field: EMNLP 2023 Findings. <https://arxiv.org/abs/2305.12295> [K] [F67] Zhibin Gou, Zhihong Shao, Yeyun Gong et al. (7). *CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing*. arXiv:2305.11738, 2023; comments field: ICLR 2024. <https://arxiv.org/abs/2305.11738> [P] [F68] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley et al. (13). *Agent-as-a-Judge: Evaluate Agents with Agents*. arXiv:2410.10934, 2024. <https://arxiv.org/abs/2410.10934> [P] [F69] Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, Rebecca Qian. *TRAIL: Trace Reasoning and Agentic Issue Localization*. arXiv:2505.08638, 2025. <https://arxiv.org/abs/2505.08638> [P] [F70] Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade et al. (10). *AgentRewardBench: Evaluating Automatic Evaluations of Web Agent Trajectories*. arXiv:2504.08942, 2025. <https://arxiv.org/abs/2504.08942> [P] [F71] Hongliu Cao, Ilias Driouich, Eoin Thomas. *Beyond Task Completion: Revealing Corrupt Success in LLM Agents through Procedure-Aware Evaluation*. arXiv:2603.03116, 2026. <https://arxiv.org/abs/2603.03116> [P] [F72] Daehong Kim, Haichao Miao, Shusen Liu. *LEDGER: Claim-to-Evidence Trace Graphs for Auditing LLM Agents*. arXiv:2608.18398, 2026. <https://arxiv.org/abs/2608.18398> [P] [F73] Yiqi Wang, Jiaqi Zhang, Zhangkai Wu et al. (11). *From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents*. arXiv:2606.04990, 2026. <https://arxiv.org/abs/2606.04990> [P] [F74] Laksh Advani. *From Confident Closing to Silent Failure: Characterizing False Success in LLM Agents*. arXiv:2606.09863, 2026; comments field: FAGEN@ICML2026. <https://arxiv.org/abs/2606.09863> [W]

Strand C — Provenance, disclosure, pre-registration

[F75] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, David T. Mellor. *The preregistration revolution*. PNAS 115(11), 2600-2606, 2018, DOI 10.1073/pnas.1708274114. <https://doi.org/10.1073/pnas.1708274114> [Z] [F76] Christopher D. Chambers, Loukia Tzavella. *The past, present and future of Registered Reports*. Nature Human Behaviour 6(1), 29-42, 2022, DOI 10.1038/s41562-021-01193-7. <https://doi.org/10.1038/s41562-021-01193-7> [Z] [F77] Courtney K. Soderberg, Timothy M. Errington, Sarah R. Schiavone, Julia Bottesini, Felix Singleton Thorn, Simine Vazire, Kevin M. Esterling, Brian A. Nosek. *Initial evidence of research quality of registered reports compared with the standard publishing model*. Nature Human Behaviour 5(8), 990-997, 2021, DOI 10.1038/s41562-021-01142-4. <https://doi.org/10.1038/s41562-021-01142-4> [Z] [F78] Anne M. Scheel, Mitchell R. M. J. Schijen, Daniël Lakens. *An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports*. Advances in Methods and Practices in Psychological Science 4(2), 2021, DOI 10.1177/25152459211007467. <https://doi.org/10.1177/25152459211007467> [Z] [F79] Tom E. Hardwicke, Eric-Jan Wagenmakers. *Reducing bias, increasing transparency and calibrating confidence with preregistration*. Nature Human Behaviour 7(1), 15-26, 2023, DOI 10.1038/s41562-022-01497-2. <https://doi.org/10.1038/s41562-022-01497-2> [Z] [F80] Aline Claesen, Sara Gomes, Francis Tuerlinckx, Wolf Vanpaemel. *Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies*. Royal Society Open Science 8(10), 2021, DOI 10.1098/rsos.211037. <https://doi.org/10.1098/rsos.211037> [Z] [F81] Norbert L. Kerr. *HARKing: Hypothesizing After the Results are Known*. Personality and Social Psychology Review 2(3), 196-217, 1998, DOI 10.1207/s15327957pspr0203_4. https://doi.org/10.1207/s15327957pspr0203_4 [Z] [F82] Gowri Gopalakrishna, Gerben ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, Lex M. Bouter. *Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands*. PLOS ONE 17(2), e0263023, 2022, DOI 10.1371/journal.pone.0263023. <https://doi.org/10.1371/journal.pone.0263023> [Z] [F83] John Protzko, Jon Krosnick, Leif Nelson, Brian A. Nosek et al. (17). *ZURÜCKGEZOGEN: High replicability of newly discovered social-behavioural findings is achievable*. Nature Human Behaviour 8(2), 311-319, 2024, DOI 10.1038/s41562-023-01749-9; Retraction note NHB 8(10), 2067, 24.09.2024, DOI 10.1038/s41562-024-01997-3. <https://doi.org/10.1038/s41562-024-01997-3> [Z] [F84] International Committee of Medical Journal Editors. *A. Defining the Role of Authors and Contributors*. Policy document, no version date on the page; as read 06.10.2026.

<https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html> [D] [F85] Committee on Publication Ethics. *Authorship and AI tools*. COPE Position Statement, version 1, 20.08.2024. <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools> [D] [F86] Springer Nature. *Artificial Intelligence (AI) – Editorial Policy*. Traffic-light scheme, no version number or revision date; as read 06.10.2026. <https://www.nature.com/nature-portfolio/editorial-policies/ai> [D] [F87] Elsevier. *Generative AI policies for journals*. Page states: Policy updated June 2026. <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals> [D] [F88] Association for Computational Linguistics. *ACL Policy on Publication Ethics*. Last revision per page 17.08.2026. https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics [D] [F89] ACL Rolling Review. *Responsible NLP Research – Checkliste, Punkt E1*. as read 2026-10-06. <https://aclrollingreview.org/responsibleNLPresearch/> [D] [F90] NeurIPS. *NeurIPS 2025 Policy on the Use of Large Language Models*. as read 2026-10-06. <https://neurips.cc/Conferences/2025/LLM> [D] [F91] ICML. *ICML 2026 Policy for LLM use in reviewing*. as read 2026-10-06. <https://icml.cc/Conferences/2026/LLM-Policy> [D] [F92] ICLR. *Policies on Large Language Model Usage at ICLR 2026*. as read 2026-10-06. <https://iclr.cc/FAQ/LLM> [D] [F93] European Commission, Directorate-General for Research and Innovation. *Living guidelines on the responsible use of generative AI in research*. Third version, May 2026. https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en [D] [F94] Coalition for Content Provenance and Authenticity. *Content Credentials: C2PA Technical Specification, Version 2.3*. Document dated December 2025. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html [D] [F95] Coalition for Content Provenance and Authenticity. *C2PA Conformance Program and C2PA Trust List*. as read 2026-10-06; interim list frozen as of 2026-01-01. <https://c2pa.org/conformance/> [D] [F96] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein. *A Watermark for Large Language Models*. arXiv:2301.10226, 2023; ICML 2023. <https://arxiv.org/abs/2301.10226> [K] [F97] Sumanth Dathathri, Abigail See, Sumedh Ghaisas et al. (24). *Scalable watermarking for identifying large language model outputs*. Nature 634, 818-823, 2024, DOI 10.1038/s41586-024-08025-4. <https://doi.org/10.1038/s41586-024-08025-4> [Z] [F98] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi. *Can AI-Generated Text be Reliably Detected?*. Transactions on Machine Learning Research; arXiv:2303.11156, 2023/2025. <https://arxiv.org/abs/2303.11156> [Z] [F99] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, Boaz Barak. *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models*. arXiv:2311.04378, 2023/2025; ICML 2024. <https://arxiv.org/abs/2311.04378> [K] [F100] World Wide Web Consortium. *PROV-DM: The PROV Data Model*. W3C Recommendation, 2013-04-30. <https://www.w3.org/TR/prov-dm/> [D] [F101] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. *Model Cards for Model Reporting*. FAT* '19, Atlanta, 29.-31.01.2019; arXiv:1810.03993. <https://arxiv.org/abs/1810.03993> [K] [F102] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. *Datasheets for Datasets*. Communications of the ACM, Dezember 2021; arXiv:1803.09010. <https://arxiv.org/abs/1803.09010> [Z] [F103] Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, Bela Gipp. *AI Usage Cards: Responsibly Reporting AI-generated Content*. arXiv:2303.03886, 2023. <https://arxiv.org/abs/2303.03886> [P] [F104] Jack Gallifant, Majid Afshar, Saleem Ameen et al. (20). *The TRIPOD-LLM reporting guideline for studies using large language models*. Nature Medicine 31(1), 60-69, 2025, DOI 10.1038/s41591-024-03425-5. Author-list form from a repository record; check at the publisher page before printing in full. <https://doi.org/10.1038/s41591-024-03425-5> [T] [F105] Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, Jelte M. Wicherts. *The prevalence of statistical reporting errors in psychology (1985-2013)*. Behavior Research Methods 48(4), 1205-1226, 2016, DOI 10.3758/s13428-015-0664-2. <https://doi.org/10.3758/s13428-015-0664-2> [Z] [F106] Nicholas J. L. Brown, James A. J. Heathers. *The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology*. Social Psychological and Personality Science 8(4), 363-369, 2017, DOI 10.1177/1948550616673876. <https://doi.org/10.1177/1948550616673876> [Z] [F107] Joe Menke, Martijn Roelandse, Burak Ozyurt, Maryann Martone, Anita Bandrowski. *The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods*. iScience 23(11), 101698, 2020, DOI 10.1016/j.isci.2020.101698. Companion paper: Menke et al., JMIR 24(6), e37324, 2022, DOI 10.2196/37324. <https://doi.org/10.1016/j.isci.2020.101698> [Z] [F108] Guillaume Cabanac, Cyril Labbé, Alexander Magazinov. *The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment*. arXiv:2210.04895, 2022, DOI 10.48550/arXiv.2210.04895. <https://arxiv.org/abs/2210.04895> [P] [F109] Richard Van Noorden. *More than 10,000 research papers were retracted in 2023 – a new record*. Nature 624(7992), 479-481, 12.12.2023, DOI 10.1038/d41586-023-03974-8. News section. <https://doi.org/10.1038/d41586-023-03974-8> [N] [F110] STM Association. *STM Integrity Hub in action: a chronicle*. 12.09.2025. The detectors used are not disclosed. <https://stm-assoc.org/stm-integrity-hub-in-action-a-chronicle/> [D] [F111] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, Jan Lause. *Delving into LLM-assisted writing in biomedical publications through excess vocabulary*. Science Advances 11(27), 02.07.2025, DOI 10.1126/sciadv.adt3813. <https://doi.org/10.1126/sciadv.adt3813> [Z] [F112] Giovanni E. Cacciamani, Michael B. Eppler, Conner Ganjavi, Asli Pekan, Brett Biedermann, Gary S. Collins, Inderbir S. Gill. *Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use (CANGARU) Guidelines*. arXiv:2307.08974, 2023. Only the protocol is verified; a published final version is not confirmed. <https://arxiv.org/abs/2307.08974> [T]

Strand D — Agentic research systems and their evaluation

[F113] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, Jeff Clune. *Towards end-to-end automation of AI research*. Nature 651, 914-919, 25.03.2026, DOI 10.1038/s41586-026-10265-5. Preprint version under a different title and with six authors: arXiv:2408.06292. <https://doi.org/10.1038/s41586-026-10265-5> [Z] [F114] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, David Ha. *The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic*

Tree Search. arXiv:2504.08066, 2025. <https://arxiv.org/abs/2504.08066> [P] [F115] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, Emad Barsoum. *Agent Laboratory: Using LLM Agents as Research Assistants*. arXiv:2501.04227, 2025. <https://arxiv.org/abs/2501.04227> [P] [F116] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, Linyi Yang. *CycleResearcher: Improving Automated Research via Automated Review*. arXiv:2411.00816, 2024; comments field: ICLR 2025. <https://arxiv.org/abs/2411.00816> [P] [F117] Daniil A. Boiko, Robert MacKnight, Ben Kline, Gabe Gomes. *Autonomous chemical research with large language models*. Nature 624, 570-578, 20.12.2023, DOI 10.1038/s41586-023-06792-0. <https://doi.org/10.1038/s41586-023-06792-0> [Z] [F118] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, Philippe Schwaller. *Augmenting large language models with chemistry tools*. Nature Machine Intelligence 6, 525-535, 08.05.2024, DOI 10.1038/s42256-024-00832-8; Preprint arXiv:2304.05376. <https://doi.org/10.1038/s42256-024-00832-8> [Z] [F119] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei et al. (16). *An autonomous laboratory for the accelerated synthesis of inorganic materials*. Nature 624, 86-91, 2023, DOI 10.1038/s41586-023-06734-w. <https://doi.org/10.1038/s41586-023-06734-w> [Z] [F120] Josh Leeman, Yuhan Liu, Joseph Stiles, Scott B. Lee, Prajna Bhatt, Leslie M. Schoop, Robert G. Palgrave. *Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis*. PRX Energy 3(1), 011002, 2024, DOI 10.1103/PRXEnergy.3.011002. Metadata from an institutional record; the publisher page refused the fetch. <https://doi.org/10.1103/PRXEnergy.3.011002> [T] [F121] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, James Zou. *The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies*. Nature 646, 716-723, 29.07.2025, DOI 10.1038/s41586-025-09442-9. <https://doi.org/10.1038/s41586-025-09442-9> [Z] [F122] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu et al. (44). *Accelerating scientific discovery with Co-Scientist*. Nature, 2026, DOI 10.1038/s41586-026-10644-y; arXiv:2502.18864. Version v1 carries a different title. <https://arxiv.org/abs/2502.18864> [Z] [F123] Alireza Ghafarollahi, Markus J. Buehler. *SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning*. arXiv:2409.05556, 2024. <https://arxiv.org/abs/2409.05556> [P] [F124] Haoyang Su, Renqi Chen, Shixiang Tang et al. (13). *Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System*. arXiv:2410.09403, 2024; comments field: ACL 2025 Main Conference. <https://arxiv.org/abs/2410.09403> [K] [F125] Qingyun Wang, Doug Downey, Heng Ji, Tom Hope. *SciMON: Scientific Inspiration Machines Optimized for Novelty*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 279-299. <https://aclanthology.org/2024.acl-long.18/> [K] [F126] Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, Tom Hope. *Literature-Grounded Novelty Assessment of Scientific Ideas*. arXiv:2506.22026, 2025. <https://arxiv.org/abs/2506.22026> [P] [F127] Noy Sternlicht, Simra Shahid, Peter Jansen, Daniel S. Weld, Pao Siangliulue, Tom Hope. *Old Ideas, Novel Problems: The Instability of LLM-Based Novelty Evaluation*. arXiv:2610.02022, 01.10.2026. <https://arxiv.org/abs/2610.02022> [P] [F128] Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, Hao Sun. *Evaluating LLMs' Divergent Thinking Capabilities for Scientific Idea Generation with Minimal Context*. arXiv:2412.17596, 2024. Renamed; LiveIdeaBench is the benchmark name, not the title. <https://arxiv.org/abs/2412.17596> [P] [F129] Chenglei Si, Diyi Yang, Tatsunori Hashimoto. *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers*. arXiv:2409.04109, 2024. The abstract page names no venue. <https://arxiv.org/abs/2409.04109> [P] [F130] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe et al. (12). *MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering*. arXiv:2410.07095, 2024; comments field names an ICLR version. <https://arxiv.org/abs/2410.07095> [P] [F131] Giulio Starace, Oliver Jaffe, Dane Sherburn et al. (13). *PaperBench: Evaluating AI's Ability to Replicate AI Research*. arXiv:2504.01848, 2025. <https://arxiv.org/abs/2504.01848> [P] [F132] Hjalmar Wijk, Tao Lin, Joel Becker et al. (23). *RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts*. arXiv:2411.15114, 2024. <https://arxiv.org/abs/2411.15114> [P] [F133] Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebel, Arvind Narayanan. *CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark*. arXiv:2409.11363, 2024/2026. <https://arxiv.org/abs/2409.11363> [P] [F134] Minjun Zhu, Qiuji Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, Yue Zhang. *AI Scientists Fail Without Strong Implementation Capability*. arXiv:2506.01372, 2025; comments field: Position. <https://arxiv.org/abs/2506.01372> [P] [F135] Joydeep Biswas, Sheila Schoepp, Gautham Vasani et al. (13). *AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot*. arXiv:2604.13940, 2026. Written by the pilot's organisers. <https://arxiv.org/abs/2604.13940> [P] [F136] ICLR 2026 Program Chairs. *ICLR 2026 Response to LLM-Generated Papers and Reviews*. ICLR Blog, 2025-11-19. <https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/> [D] [F137] Alex Lu, Seth Lazar, David Rüger (Position Paper Chairs). *AI-Generated Papers in the NeurIPS 2026 Position Paper Track*. NeurIPS Blog, 2026-06-02. <https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/> [D] [F138] James Zou, Owen Queen, Nitya Thakkar, Eric Sun, Federico Bianchi (Chairs). *Open Conference of AI Agents for Science 2025 (Agents4Science)*. Conference website; virtual conference 2025-10-22. <https://agents4science.stanford.edu> [D] [F139] Nilmani Singh, Stephan Lane, Tianhao Yu, Jingxia Lu, Adrianna Ramos, Haiyang Cui, Huimin Zhao. *A generalized platform for artificial intelligence-powered autonomous enzyme engineering*. Nature Communications 16, 5648, 01.07.2025, DOI 10.1038/s41467-025-61209-y. <https://doi.org/10.1038/s41467-025-61209-y> [Z] [F140] Grant M. Landwehr, Jonathan W. Bogart, Carol Magalhães, Eric G. Hammarlund, Ashty S. Karim, Michael C. Jewett. *Accelerated enzyme engineering by machine-learning guided cell-free expression*. Nature Communications 16, 865, 20.01.2025, DOI 10.1038/s41467-024-55399-0. <https://doi.org/10.1038/s41467-024-55399-0> [Z]

11. Provenance and status

This field map belongs to the working note *Iteratio Simulata* (identifier FORSCH-PROV0459) and forms its context part. It is readable on its own and does not presuppose knowledge of the five markers.

Survey period: 2026-10-06, one working day, four separate strands. **Method:** web search and fetching of primary pages; no database licence, no access to full texts behind paywalls. **Not reachable:** PubMed and PMC refuse automated fetches; several publisher pages respond with

HTTP 403; one legal text truncates before its articles in both HTML and PDF. The consequences are named in 8.2 and carried as **[T]** in the reference list. **Not included:** detector vendors' figures (8.3); full-text reading of the agentic systems (9.3).

Licence: CC BY 4.0. Attribution: Thomas Schüller, ORCID 0009-0003-9799-6747.

Suggested citation: Schüller, T. (2026). *Field Map: Procedure Evidence in LLM-Assisted Knowledge Work*. Context part to the working note Iteratio Simulata, version v0.5, FORSCH-PROV0459.

Register note: the number PROV0459 is provisional; the final number is set against the work register.