

- [Abstract](#)
- [1 Occasion and Research Question](#)
- [2 Methodological Frame and Delineation](#)
- [3 Conceptual Clarification](#)
- [4 Five Markers](#)
 - [M1 — Elimination claim without elimination trace](#)
 - [M2 — Criteria after the evaluation](#)
 - [M3 — Reframing established research as breakthrough](#)
 - [M4 — Self-validating rhetoric instead of external falsification](#)
 - [M5 — Concretisation asymmetry](#)
- [5 Pre-Registration Mechanism](#)
- [6 Connection to Existing Tools](#)
- [7 Limits and Open Points](#)
- [8 Next Steps](#)
- [9 Self-Application: The Markers Applied to This Working Note](#)
 - [9.1 Findings](#)
 - [9.2 Assignment to the Markers](#)
 - [9.3 Consequences for the Marker List](#)
 - [9.4 Implementation Finding: The Marker First Penalises Care](#)
 - [9.5 Methodological Placement](#)
 - [9.6 Fourth finding: a withdrawn exclusivity claim](#)
- [Appendix A — Occasioning Case](#)
- [Appendix B — References with Verification Status](#)
 - [Self-correction and faithfulness of reasoning traces \(foundation for the core concept\)](#)
 - [Multi-agent debate \(contrast standard for genuine iteration\)](#)
 - [Sycophancy and confirmation bias in MAD \(adapter source for M4\)](#)
 - [Novelty assessment in peer review \(methodological basis for M3\)](#)
 - [LLM-as-reviewer risks \(justification against LLM-as-judge\)](#)
 - [Novelty generation and brainstorming \(vocabulary for aggregate scoring\)](#)
 - [Related terms](#)
- [Licence and Versioning](#)

title: “Iteratio Simulata”

subtitle: “Performative Iteration in LLM-Assisted Idea Generation — Five Markers and a Pre-Registration Mechanism”

author: “Thomas Schüller”

orcid: “0009-0003-9799-6747”

contact: “forschung@tschueller.com”

affiliation: “independent, Vienna”

version: “0.4 (English edition)”

date: “2026-10-06”

status: “Working note, all 18 references externally verified, foundational cluster added, tool phase 1a operational; published via syntagma-science, Zenodo secondary”

license: “CC BY 4.0”

language: “en”

translation_note: “English edition of the German original. In case of doubt the German version is authoritative.”

keywords: [“LLM evaluation”, “epistemic performativity”, “multi-agent debate”, “sycophancy”, “pre-registration”, “unfaithful chain-of-thought”, “potemkin understanding”, “tool-based socio-technical research”]

companion: “Iterationsprüfer — module specification v0.4; phase-0 verification report; reference verification report; pamphlet”

kennung: “FORSCH-PROV0459”

slug: “iteratio-simulata”

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

Abstract

Single-agent outputs of large language models that rhetorically imitate multi-agent debate structures (“Autonomous Inventor Mode”, “15 iterations before final output”, “a hard reviewer eliminates weak concepts”) typically produce texts in which the claimed iteration does not occur. Concepts are renamed but not discarded; evaluation criteria are formulated after the evaluation; the state of the art is reframed as a breakthrough. This working note names the phenomenon *Iteratio Simulata* and proposes five operationalizable markers for its identification together with a pre-registration mechanism for prevention. The concept is delineated against four neighbouring discourses: multi-agent debate research, sycophancy and confirmation bias research, novelty assessment in peer review, and — most important for the conceptual work — research on self-correction and the faithfulness of reasoning traces. Relative to its two closest terminological neighbours the concept is positioned as follows: *unfaithful chain-of-thought* concerns the relation between stated reasoning and model internals, *potemkin understanding* the relation between explaining and applying a concept; *Iteratio Simulata* concerns the relation between claimed and demonstrable process structure and is testable from the text alone, without model access and without a benchmark setup. That advantage holds against the two neighbouring concepts named, not against the 2026 work on the performative reasoning trace, which uses activation probing and timing and therefore requires model access; the field map accompanying this note records it. The occasioning case is a ChatGPT output on cell-free biocatalysis from May 2026. A self-application test of the markers to this note (Section 9) produced three findings: five faulty reference entries including two fabricated authorships, an entirely missing literature cluster on self-correction research, and an implementation failure in which the accompanying detection tool initially penalised carefully written text. All three are documented openly rather than corrected silently. The accompanying tool is specified in a separate companion document.

1 Occasion and Research Question

In the practice of transdisciplinary synthesis work, LLM outputs increasingly appear that formally imitate the structure of a research session: hypothesis, critique, improvement, selection. The case examined here is a brainstorm on cell-free biocatalysis generated over seven user turns, initiated by a master prompt explicitly demanding at least fifteen iterations and hard selection by simulated reviewers.

On sober reading, several things stand out. There are no discernible fifteen iterations. The top concepts produced — vascularised mycelium biofactories, programmable biofilm reactors, cell-free reaction networks in hydrogels, mechanosensitive scaffolds, symbiotic multi-species reactors, biomineralising systems, electroactive cell networks — have been established in the literature for ten to fifteen years and are in part commercialised. The claimed elimination of weak approaches is not traceable in the output; discarded candidates are never named. The evaluation criteria of the reviewer appear only after the selection has been made.

This yields the research question:

How can the performativity gap between multi-agent debate form and multi-agent debate substance in a single-agent output be systematically identified, and which minimal pre-registration mechanisms are required to reduce this gap before generation?

The question has both epistemic and practical relevance. Epistemic, because it names a pattern that remains systematically under-detected in emerging LLM research practice. Practical, because it connects to existing tools in the author’s workshop line, in particular an epistemic engine implementing SHA-256 hash-lock pre-registration and a HARKing penalty.

2 Methodological Frame and Delineation

The work follows a tool-based socio-technical research frame: conceptual clarification, case analysis, operationalisation, tool development, feedback. It is not a quantitative empirical study on a large sample of LLM outputs but a concept-forming working note built on a single case. Claims to generalisation are limited accordingly.

The work delineates itself against four neighbouring discourses, each of which addresses partial aspects but not the phenomenon itself.

First, multi-agent debate research (MAD). Frameworks such as VirSci (Su et al. 2024, ACL 2025), MARG (D’Arcy et al. 2024), MAD (Liang et al. 2023) or the Generative Reviewer Agents (Bougie & Watanabe 2024) show how iterative hypothesis generation with distributed roles — brainstormer, critic, judge, meta-reviewer — is operationalised. These frameworks actually perform iteration, because different agents respond to different outputs and external aggregation takes place. *Iteratio Simulata* is precisely *not* such a setup but its rhetorical imitation within a single coherent single-agent output.

Second, sycophancy and confirmation bias research (Sharma et al. 2023; Perez et al. 2022). More recent work investigates how inter-agent sycophancy causes even genuine multi-agent debates to collapse into self-confirmation loops (Wynn et al. 2025, “Talk Isn’t Always Cheap”; Yao et al. 2025, “Peacemaker or Troublemaker”; Choi et al. 2025, “When Identity Skews Debate”). Particularly pertinent is Kim & Torr (2025), whose MoLaCE work explicitly treats the case of a single LLM attempting to emulate the benefits of debate internally. *Iteratio Simulata* is the *single-agent application* of these findings: the same mechanisms that cause multi-agent setups to collapse into performative loops occur even more strongly in a single output, because no external validating instance is present.

Third, novelty assessment research in peer review. Work such as Afzal et al. (EACL 2026) and the patent novelty evaluation by Ikoma & Mitamura (2025) operationalises the comparison between claimed originality and existing literature. These methods address marker M3 of our list directly, though in the context of academic and patent review rather than brainstorm outputs. *Iteratio Simulata* adopts these methods as an adapter rather than rebuilding them.

Fourth, and most important for the conceptual work, research on self-correction and the faithfulness of reasoning traces. This discourse was overlooked in the first versions of this working note; that was the more serious of two verification findings (see Section 9). Three connection points are central.

Huang et al. (ICLR 2024) show empirically that large language models cannot reliably correct their own reasoning without external feedback; their analysis further suggests that multi-agent debate substantially realises self-consistency rather than genuine critique. The empirical core of what is here framed conceptually thus already exists in published form: if self-correction structurally does not work, then an output claiming self-correction is necessarily performative. *Iteratio Simulata* adds no new empirical evidence to this finding but supplies a terminology and a marker catalogue for the textual surface at which the phenomenon becomes recognisable.

Turpin et al. (NeurIPS 2023) and Lanham et al. (2023) establish *unfaithful chain-of-thought* as the umbrella term for reasoning traces that do not reflect the actual decision process; Chen et al. (2025) transfer the question to explicit reasoning models. The delineation must be drawn precisely. *Unfaithful CoT* concerns the relation between the stated justification and the internal computation, and is therefore a statement about model internals. *Iteratio Simulata* concerns the relation between the claimed process structure and the textually demonstrable process structure, and is therefore a statement about the text alone. The five markers are deliberately constructed so as to remain testable without model access. Both perspectives are complementary: an output can be faithful and yet performatively iterated, if the model in fact generates linearly and internally proceeds that way, while the text claims a loop structure.

Mancoridis et al. (ICML 2025) coin the term *potemkin understanding* for the illusion of conceptual understanding alongside high benchmark performance. This is the closest terminological neighbour. The difference lies in the object: *potemkin understanding* concerns the gap between explaining a concept and applying that same concept. *Iteratio Simulata* concerns the gap between claiming a procedure and executing that same procedure. Both are instances of the same more general pattern — form without substance — but they operate at different levels and are testable by different methods. Readers who prefer to treat *Iteratio Simulata* as a special case of *potemkin understanding* may do so; the independent term is justified by its independent marker catalogue and by its testability without a benchmark setup.

As a counter-position, Madaan et al. (NeurIPS 2023, *Self-Refine*) must be named. That work demonstrates conditions under which iterative self-improvement measurably works. It thereby marks the boundary: not every claimed iteration is simulated. The question is always whether the transformations between states are demonstrable — not whether iteration is possible in principle.

Finally, the work delineates itself against *LLM-as-judge* approaches. These are unsuitable as a detection instrument for the phenomenon described here, because a model that generates the same rhetorical patterns cannot reliably recognise them (for empirical support see Zhu et al. 2025, arXiv:2509.09912, and Li et al. 2025, arXiv:2510.12367). A pre-registration mechanism that relocates control outside the model is proposed instead.

The catalogue of pseudoscience markers developed by Sven Ove Hansson covers related but not identical phenomena; *Iteratio Simulata* is to be understood as a complementary marker catalogue, not a replacement.

3 Conceptual Clarification

Iteratio Simulata denotes the structural gap between multi-agent debate form and multi-agent debate substance in a single-agent output. It occurs when a single LLM output rhetorically imitates distributed roles (brainstormer, reviewer, editor) without realising the structural properties that make MAD setups robust: role diversity through independent agents, external falsification, documented elimination, separate

aggregation.

Actual iteration, by contrast, is a process in which a traceable transformation occurs between two successive states that is non-trivial and exhibits at least one of the following features: elimination of a candidate with a stated reason for elimination; substantial modification of a candidate with the modification stated; addition of a candidate with its source stated; explicit application of a previously fixed evaluation criterion.

The distinction is not binary. Outputs encountered in practice lie on a continuum between complete iteration and complete performativity. The markers proposed here are to be read as weighted indicators along this continuum.

On placement within the conceptual field: *Iteratio Simulata* is a text-surface category testable without model access. It sits between *unfaithful chain-of-thought* (a statement about the relation of justification to model internals) and *potemkin understanding* (a statement about the relation of concept explanation to concept application). The term claims no new discovery of the general pattern — form without substance has been a familiar motif of scientific criticism at least since Feynman’s cargo-cult science. It claims only that the specific manifestation *claimed versus demonstrable process structure in a single-agent output* merits its own marker catalogue, because it occurs frequently in the practice of LLM-assisted synthesis work and because it is testable without specialised infrastructure.

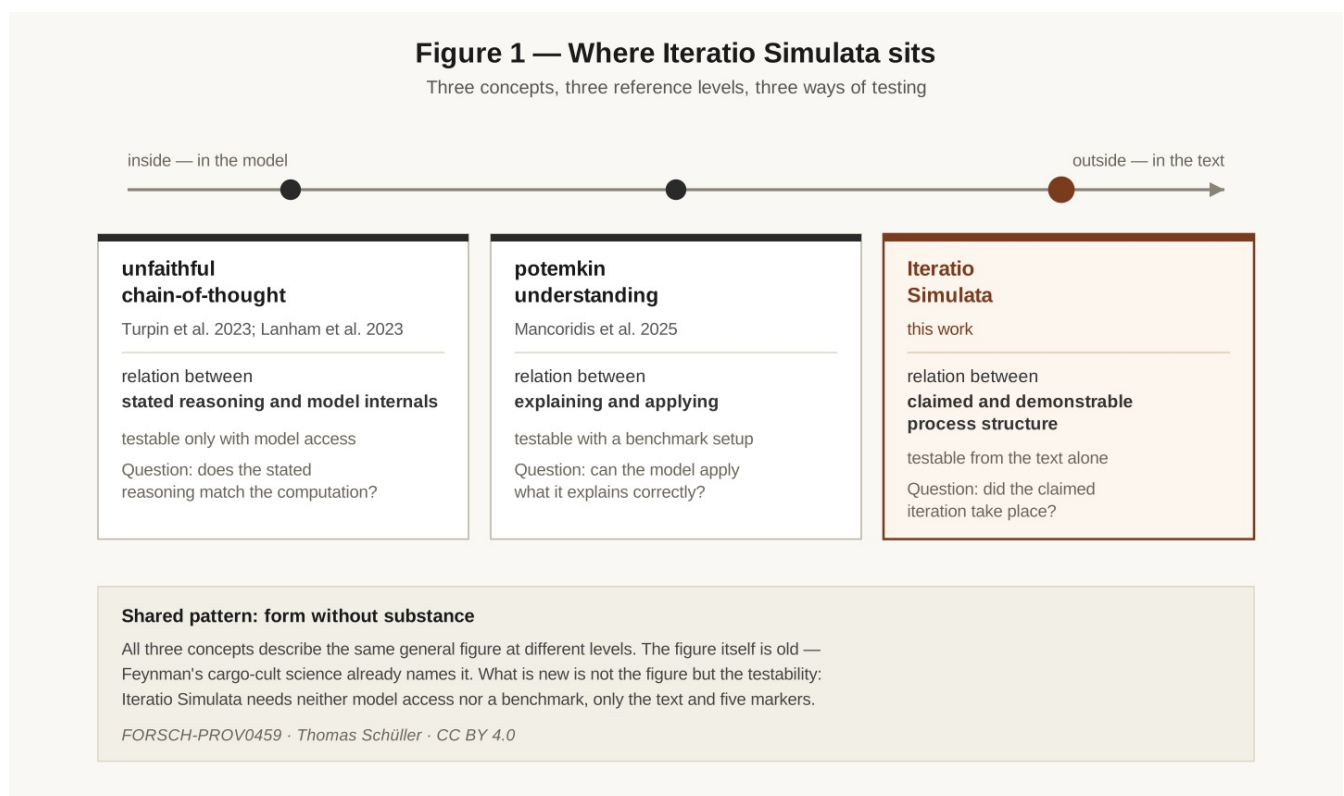


Figure 1 — Where *Iteratio Simulata* sits in the conceptual field: three concepts, three reference levels, three ways of testing.

4 Five Markers

Figure 2 — The five markers along a claimed iteration

Left the claimed step, right the trace it would have to leave in the text

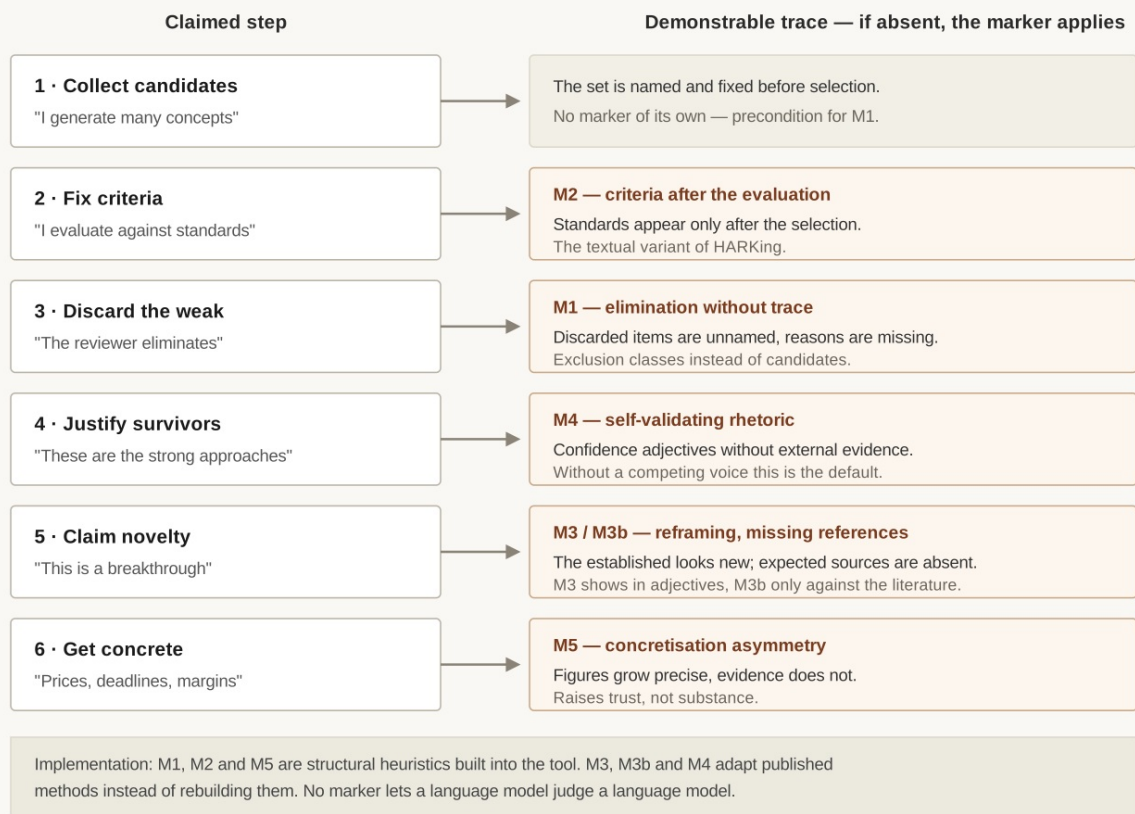


Figure 2 — The five markers along a claimed iteration. Left the claimed step, right the trace it would have to leave in the text.

The following markers are extracted from the analysis of the occasioning case and are to be validated on additional cases in further work. They are formulated so as to be structurally detectable without the testing system having to understand the substantive content of the output.

M1 — Elimination claim without elimination trace

The output claims to have discarded candidates but does not name them and gives no reasons for rejection. In the occasioning case: "The reviewer discards biologically unstable systems, unrealistic cell and reaction assumptions, non-scalable processes, purely theoretical constructs without feasibility." Which concepts fell under these categories remains open. In a MAD setup the elimination trace would be explicitly documented in the critic contributions; in *Iteratio Simulata* it is structurally absent, because no separate agents exist that leave their contributions independently.

M2 — Criteria after the evaluation

Evaluation criteria are presented in the output after the selection has already taken place, instead of being fixed beforehand and applied equally to all candidates. In the occasioning case the "four properties" for target-molecule evaluation (small volume, expensive, awkward to produce, time-critical) are introduced only after GDP-fucose and CMP-sialic acid have been named as top candidates. This is the LLM variant of a classical methodological error known as HARKing (Hypothesising After the Results are Known).

M3 — Reframing established research as breakthrough

Concepts long established in the specialist literature are presented with the rhetoric of new discovery. In the occasioning case, "programmable biofilm reactors" is presented as a paradigm-shifting idea, although structured biofilm systems with quorum-sensing control constitute an established research field with reviews, conferences and commercial suppliers. The marker is measurable by comparison with publication data (first publication, review density, patent landscape). The methodology already exists: Afzal et al. (EACL 2026) achieve 86.5 percent alignment with human reasoning in novelty assessment of 182 ICLR 2025 submissions; Ikoma & Mitamura (2025) develop comparable methods for patents. M3 is therefore not an in-house marker but an application of these methods to LLM brainstorm outputs.

M4 — Self-validating rhetoric instead of external falsification

The output evaluates its own proposals with high-confidence adjectives (“brutally relevant”, “clearly superior”, “real game-changers”) without citing external sources or data against which the evaluation could fail. Falsification possibilities are mentioned rhetorically but not applied operationally. This marker is the single-agent application of sycophancy and confirmation bias findings from MAD research. In multi-agent debates it has been shown that even distributed agents collapse into self-confirmation when strong falsification incentives are absent (Wynn et al. 2025). In a single-agent output this collapse is the default state, because no competing voice is structurally provided for.

M5 — Concretisation asymmetry

Across successive iterations the concreteness of figures, prices and timelines increases while the depth of evidence remains constantly low. In the occasioning case, pilot prices (EUR 3,000–8,000), margins (50 percent plus), reactor volumes (5–50 ml), weekly plans and cost structures are stated without a single one of these figures being derived from a source or justified. They appear precise and are in fact freely set. Only at the seventh iteration step do external price data appear; these are correct, but they validate only the order of magnitude of the specialty reagent market, not the pilot prices and margins stated earlier. This marker is particularly insidious because it increases trust in the output without increasing its substance.

5 Pre-Registration Mechanism

The central methodological question is not how to detect *Iteratio Simulata* after the fact, but how to make it harder before generation. Four minimal mechanisms are proposed.

First, pre-commitment of the candidate set. Before each selection step the complete candidate list is fixed and cryptographically hashed (SHA-256). Eliminations must refer explicitly to this list.

Second, pre-commitment of criteria. Evaluation criteria are fixed before selection and hashed. Later changes to criteria are marked as such. This is the direct transfer of pre-registration logic from the replication-crisis debate to LLM workflows.

Third, external anchoring. At least one source per concept class from indexed literature (DOI, PubMed ID, patent number), to address M3. Queries run against Semantic Scholar, OpenAlex or Crossref rather than the model’s own recall.

Fourth, asymmetry audit. The ratio of quantitative claims to supported quantitative claims is reported at the end, to address M5.

These mechanisms are not complete protection, but they raise the threshold at which performativity can be produced cheaply. Their mode of operation is analogous to pre-registration in the empirical sciences: not every instance of HARKing is prevented, but cheap HARKing becomes more expensive.

6 Connection to Existing Tools

The proposed frame is designed as an extension of existing workshop-line tools rather than in isolation. An epistemic engine already implementing SHA-256 pre-registration and a HARKing penalty can receive an additional detection pattern reflecting the five markers. A multi-lens analysis tool can receive a dedicated frame for LLM outputs in which the methodological-rigour lens operationalises M1, M2 and M5 and the contextual-originality lens operationalises M3. A research-intake tool receives the *Iterationsprüfer* module described in the companion specification, computing a marker score before external LLM outputs are admitted as research material.

7 Limits and Open Points

This working note is a draft built on a single case. The following limits must be stated.

First, the marker list M1–M5 is not closed. Additional cases will probably yield further markers, particularly in the area of dialogic escalation (transition from brainstorm to business plan without a validation break).

Second, delineation against legitimate rhetorical strategies (simplification, pedagogy, motivation) is not trivial. A textbook sometimes deliberately reframes established research as exciting without being misleading. The difference presumably lies in explicit marking as instructional text versus masquerade as a research process.

Third, it remains open whether pre-registration can be implemented in LLM generation workflows without high friction. The assumption of this note is that a lightweight mechanism (hash of the input conditions, audit at the end) suffices to reduce performativity substantially. This assumption is to be tested empirically.

Fourth, it must be examined whether the phenomenon is model-dependent or model-independent. The occasioning case stems from a ChatGPT output; comparison cases from other proprietary and open models are needed.

Fifth, the reference verification underlying version 0.3 was a literature scan, not a systematic review. A systematic review would be appropriate for a v1.0 journal submission but exceeds the scope of a working note.

8 Next Steps

This working note is ready for publication as a Zenodo preprint. Journal submission is envisaged only after two or three iterations with external feedback. Candidate venues for the final version: *Minds and Machines*, *AI & Society*, or a methods contribution to a journal with an STS profile such as *Social Studies of Science* or *Science, Technology & Human Values*.

Concrete follow-up steps are: at least three further case studies to validate the markers; discussion with at least two external readers from adjacent fields (LLM evaluation, philosophy of science); completion of phases 1b and 1c of the companion tool; and a test corpus of at least twenty texts with independent human pre-classification.

9 Self-Application: The Markers Applied to This Working Note

A work that formulates criteria for performative iteration must submit itself to those criteria. Otherwise it would be an example of precisely the pattern it describes. This section documents the self-application of markers M1 to M5 to the genesis of this note and names the errors found.

9.1 Findings

The working note came into being across three versions in a single day: v0.1 (first draft), v0.2 (after a literature scan) and v0.3 (after full verification of the references). Between v0.2 and v0.3 the reference list in Appendix B was systematically checked against arXiv, the ACL Anthology, PMLR and dblp.

First finding — faulty reference entries. Of eighteen entries, five were faulty, three of them citation-critical. Two entries carried fabricated authorship: arXiv:2509.05396 was cited as “Agarwal & Khanna 2025” when the actual authors are Wynn, Satija and Hadfield; arXiv:2412.10415 was listed as “Bhatia, K. et al.” with the title “Generative Agent Reviewers (GAR): Automated Peer Review via LLM Agents” when it is in fact Bougie & Watanabe, “Generative Adversarial Reviews: When LLMs Become the Critic”. In the latter case both title and authors were fabricated while the number was correct. One entry carried a wrong publication venue: arXiv:2508.10795 was listed as an ICLR 2025 contribution; it is an EACL 2026 contribution that uses 182 ICLR 2025 submissions as its evaluation dataset. Two entries carried divergent titles.

The error pattern is notable. **All eighteen arXiv numbers were correct and all works exist.** What was faulty were exclusively the metadata one does not look up via the number — author names, exact wording of titles, publication venue.

Second finding — missing literature cluster. More serious than the metadata errors was a gap: versions v0.1 and v0.2 argued about simulated self-critique without knowing the relevant foundational literature. Missing in particular were Huang et al. (ICLR 2024) on the impossibility of reliable self-correction without external feedback, Turpin et al. (NeurIPS 2023) and Lanham et al. (2023) on unfaithful chain-of-thought, and Mancoridis et al. (ICML 2025) on the term *potemkin understanding*. These eight works are added in v0.3 as a separate cluster in Appendix B and incorporated in Section 2 as a fourth delineation discourse.

9.2 Assignment to the Markers

M5 (concretisation asymmetry) applies. The reference list became markedly more precise from v0.1 to v0.2: arXiv numbers, years, author names, venue details. The depth of evidence, however, did not grow with it. Precision without verification is exactly the asymmetry M5 describes. The effect was no easier to notice in one’s own work than in someone else’s.

M3 (reframing established research as breakthrough) applies, in the heavier variant. The attribution of arXiv:2508.10795 to ICLR 2025 is the lighter part: a status claim not checked against the original. The heavier part is the missing literature cluster. A working note that introduces a term for simulated self-critique without knowing Huang et al. (ICLR 2024) — that is, without knowing the empirical demonstration that self-correction without external feedback structurally does not work — presents established knowledge as its own observation. That is exactly the marker definition. The difference from the occasioning case lies solely in the fact that the error was noticed and documented here, not in the fact that it did not occur.

M1, M2 and M4 do not apply. The work documents its version history in a changelog, explicitly names the discarded alternative, fixes its

evaluation criteria in Section 3 before the marker list, and states limits in Section 7. As the findings show, this does not protect against M3 and M5.

9.3 Consequences for the Marker List

Supported: M5 is the practically most effective marker. It applies even where methodological care is present in other dimensions. Concretisation asymmetry arises not from general negligence but from a local waiver of verification at precisely the point where precision is most persuasive.

Weakened: The implicit assumption in v0.2 that a carefully structured text largely avoids the markers is false. Structure does not protect against unsupported precision. This is an argument for treating the pre-registration mechanism of Section 5 not as an addition but as a necessary component.

Extended: The second finding suggests a refinement of M3. The existing formulation addresses only the active variant — something known is explicitly labelled as new. The present case shows a passive variant: something known is not marked as known, because it was not known. Both variants have the same effect on the reader, namely an inflated estimate of originality, but they differ in method of detection. The active variant is recognisable by adjectives, the passive one only by comparison with the literature. For tool implementation it follows that the M3 adapter must not only test novelty claims but should also capture the absence of expected references in the given field. This extension is included in module specification v0.4 as M3b.

9.4 Implementation Finding: The Marker First Penalises Care

The phase-1a implementation of the three in-house markers (M1, M2, M5) produced a third finding, this one concerning the marker formulation itself.

In the first counter-test the tool was given a deliberately carefully written text: named candidates, explicit rejection reasons, criteria fixed in advance, sourced figures. A score near zero was expected. In fact the text received the maximum value of 1.0 on both M1 and M2 and was classified as to be rejected.

There were two operationalisation errors. First, the candidate heuristic did not recognise identifiers such as “Method A” as naming, because it was designed for quotation marks, list bullets and proper names. Second, the selection heuristic treated the noun “selection” in the phrase “for the selection of test methods” as a selection decision, dating the selection to character position three, so that all criteria named subsequently counted as retrofitted.

Both errors point in the same direction: they penalise precisely those textual features that distinguish a careful text from a performative one. A tool with this error direction would not merely be imprecise but inverse in effect to its purpose. It would have rated authors who name candidates and fix criteria in advance worse than those who do neither.

From this follows a methodological requirement absent from the working note up to v0.2: **the markers are to be validated on negative cases, not on positive ones.** A detector for performative iteration must first demonstrate that it does not mark genuine iteration as performative. The reverse order — positive cases first, negative cases second — yields heuristics optimised for salience rather than for discrimination.

After correction of both heuristics the tool separates the two test cases clearly: the clean text scores 0.00 across all implemented markers, the occasioning case 0.80. Both cases are fixed as regression tests. The separation of 0.80 is measured on two cases and is therefore not a robust figure; it shows only that the markers discriminate at all in the corrected version. For v1.0 a test corpus of at least twenty texts is required, half performative and half careful, from different models and domains, with independent human pre-classification.

9.5 Methodological Placement

It would have been possible to correct the errors silently and omit this section. That would have made the note look smoother. It would also have suppressed the only empirical observation of this work that does not derive from the occasioning case but from its own genesis: that the described effect occurs even under deliberate attention to it.

The errors arose while working with a language model as research assistance. This is relevant to generalisability: the finding concerns not only outputs that a model stages independently as a “research session” but also work products involving a human who knows the markers. Where exactly the errors were introduced — in the retrieval by the model or in the adoption by the author — cannot be reliably reconstructed after the fact; responsibility for the published version lies with the author in any case.

For v1.0 a concrete requirement follows: markers M3 and M5 should be applied not only to external material but built into the author’s own publication workflow as a mandatory step. The asymmetry audit outlined in Section 5 is the appropriate form.

9.6 Fourth finding: a withdrawn exclusivity claim

A subsequent survey of the research context (field map, v0.5, 2026-10-06) produced a fourth self-application finding. The sentence that *Iteratio Simulata* is the only one of the neighbouring concepts testable from the text alone holds against *unfaithful chain-of-thought* and *potemkin understanding*, but not against three papers published in 2026 on the performative reasoning trace. These measure the same gap step by step — alignment between internal commitment and what the trace depicts at 61.9 per cent, up to 80 per cent of the trace dispensable with no loss of accuracy — and name it in their own terms.

The sentence appeared in the abstract, the pamphlet and both flyers. It is qualified rather than struck: the papers named require activation probing or timing, and therefore model access; that is precisely why the markers continue to stand beside them. The episode is an M3 case by this work's own definition — the state of the field was further along than the note depicted it.

At the same time the survey supplied empirical support for the central concept that was not available when this note was written, and brought a counterposition into view: non-verbalisation of a hint that was used does not by itself establish unfaithfulness of the trace. Markers M1 and M2 work through tracelessness rather than verbalisation and are not hit by it; the objection belongs in the presentation nonetheless. Details, sources and the derived follow-on projects are in the field map.

Appendix A — Occasioning Case

The complete brainstorm sequence from which the markers were developed is archived as a separate appendix (ChatGPT session on cell-free biocatalysis, seven user turns, May 2026). Archiving follows the standard rule for source material: full wording, date, model version where determinable.

Appendix B — References with Verification Status

Every entry carries a verification status. This marking is the result of a self-application test of markers M3 and M5 to this working note (see Section 9). It is not ornamental: the first version of this reference list contained fabricated author details and incorrect publication venues.

Status marking:

- [V] verified — entry retrieved May 2026; title, authors and number confirmed against arXiv, ACL Anthology, PMLR or dblp
- [T] partially verified — existence, number and authorship confirmed; no peer-reviewed venue secured

Self-correction and faithfulness of reasoning traces (foundation for the core concept)

This cluster was added in full during verification in May 2026. Its absence from the first version was the more serious of the two findings.

- [V] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., Zhou, D. (2023/2024). *Large Language Models Cannot Self-Correct Reasoning Yet*. arXiv:2310.01798. ICLR 2024.
- [V] Turpin, M., Michael, J., Perez, E., Bowman, S. R. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388. NeurIPS 2023, vol. 36, pp. 74952–74965.
- [V] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J. et al. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv:2307.13702.
- [V] Chen, Y. et al. (2025). *Reasoning Models Don't Always Say What They Think*. arXiv:2505.05410.
- [V] Mancoridis, M., Weeks, B., Vafa, K., Mullainathan, S. (2025). *Potemkin Understanding in Large Language Models*. arXiv:2506.21521. ICML 2025, PMLR vol. 267, pp. 42857–42881.
- [V] Madaan, A. et al. (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv:2303.17651. NeurIPS 2023.
- [V] Stechly, K., Valmeekam, K., Kambhampati, S. (2024). *On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks*. arXiv:2402.08115.
- [V] Stechly, K., Valmeekam, K., Kambhampati, S. (2024). *Chain of Thoughtlessness? An Analysis of CoT in Planning*. arXiv:2405.04776.

Multi-agent debate (contrast standard for genuine iteration)

- [V] Liang, T., He, Z., Jiao, W. et al. (2023). *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. arXiv:2305.19118. EMNLP 2024.
- [V] D'Arcy, M., Hope, T., Birnbaum, L., Downey, D. (2024). *MARG: Multi-Agent Review Generation for Scientific Papers*.

arXiv:2401.04259.

- [V] Su, H., Chen, R., Tang, S., Yin, Z., Zheng, X., Li, J., Qi, B., Wu, Q., Li, H., Ouyang, W. et al. (2024). *Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System*. arXiv:2410.09403. ACL 2025, pp. 28201–28240. The system is named VirSci.
- [V] Bougie, N., Watanabe, N. (2024). *Generative Adversarial Reviews: When LLMs Become the Critic*. arXiv:2412.10415. Published as *Generative Reviewer Agents: Scalable Simulacra of Peer Review*, EMNLP 2025 Industry Track.

Sycophancy and confirmation bias in MAD (adapter source for M4)

- [V] Sharma, M. et al. (2023). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548. ICLR 2024.
- [V] Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251. Findings of ACL 2023.
- [V] Wynn, A., Satija, H., Hadfield, G. (2025). *Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate*. arXiv:2509.05396.
- [V] Yao, B., Shang, C., Du, W., He, J., Lian, R., Zhang, Y., Su, H., Swamy, S., Qi, Y. (2025). *Peacemaker or Troublemaker: How Sycophancy Shapes Multi-Agent Debate*. arXiv:2509.23055.
- [V] Choi, H. K., Zhu, X., Li, Y. (2025/2026). *When Identity Skews Debate: Anonymization for Bias-Reduced Multi-Agent Reasoning*. arXiv:2510.07517. Citation note: early versions v1 and v2 carried a different title.
- [V] Kim, H., Torr, P. (2025). *Single LLM Debate, MoLaCE: Mixture of Latent Concept Experts Against Confirmation Bias*. arXiv:2512.23518.
- [T] Zhang, H., Cui, Z., Chen, J., Wang, X., Zhang, Q., Wang, Z., Wu, D., Hu, S. (2025). *Stop Overvaluing Multi-Agent Debate — We Must Rethink Evaluation and Embrace Model Heterogeneity*. arXiv:2502.08788.

Novelty assessment in peer review (methodological basis for M3)

- [V] Afzal, O. M., Nakov, P., Hope, T., Gurevych, I. (2025/2026). *Beyond “Not Novel Enough”: Enriching Scholarly Critique with LLM-Assisted Feedback*. arXiv:2508.10795. EACL 2026, Long Papers. Note: evaluation was performed on 182 ICLR 2025 submissions; the work itself is an EACL 2026 contribution. Code at github.com/UKPLab/eacl2026-assessing-paper-novelty.
- [V] Ikoma, H., Mitamura, T. (2025). *Can AI Examine Novelty of Patents? Novelty Evaluation Based on the Correspondence between Patent Claim and Prior Art*. arXiv:2502.06316.

LLM-as-reviewer risks (justification against LLM-as-judge)

- [V] Zhu, C., Xiong, J., Ma, R., Lu, Z., Liu, Y., Li, L. (2025). *When Your Reviewer is an LLM: Biases, Divergence, and Prompt Injection Risks in Peer Review*. arXiv:2509.09912.
- [V] Li, R., Gu, J.-C., Kung, P.-N. et al. (2025). *LLM-REVal: Can We Trust LLM Reviewers Yet?* arXiv:2510.12367.
- [V] Huang, H.-W., Lu, J., Chen, K.-M., Hwang, J.-N. (2026). *Modeling LLM Agent Reviewer Dynamics in Elo-Ranked Review System*. arXiv:2601.08829.

Novelty generation and brainstorming (vocabulary for aggregate scoring)

- [V] Lagzian, A., Anumasa, S., Liu, D. (2025). *Multi-Novelty: Improve the Diversity and Novelty of Contents Generated by Large Language Models via inference-time Multi-Views Brainstorming*. arXiv:2502.12700.
- [V] Padmakumar, V., Chen, Y.-H., Pan, J., Chen, V., He, H. (2025). *Measuring LLM Novelty As The Frontier Of Original And High-Quality Output*. arXiv:2504.09389.
- [T] Shahhosseini, F., Marioriyad, A., Momen, A., Soleymani Baghshah, M., Rohban, M. H., Haghjooy Javanmard, S. (2025). *Large Language Models for Scientific Idea Generation: A Creativity-Centered Survey*. arXiv:2511.07448.

Related terms

- *iteration theater* — coined informally in the LLM practitioner community; covers partial aspects of M1 and M5.
- *ritualistic prompting* — refers to prompts enforcing formal rituals (role play, step-by-step) without functional effect.
- *cargo-cult iteration* — analogy to Feynman’s cargo-cult science; emphasises the form-without-function dimension.
- *epistemic theater* — a broad term from the sociology of science, narrowed here to the LLM context.

Licence and Versioning

This working note is licensed CC BY 4.0. Versioning: v0.1 (draft), v0.2 (after phase-0 verification), v0.3 (after full reference verification, foundational cluster added, tool implemented; ready for preprint), v1.0 (journal submission). Author: Thomas Schüller, ORCID 0009-0003-9799-6747, forschung@tschueller.com. Privately funded, no institutional affiliation.

This is the English edition of the German original *Iteratio Simulata* v0.4. In case of doubt the German version is authoritative.