

FORSCH-PROV0459_pamphlet-en_v0.5_2026-10-06_werk

title: "Iteratio Simulata — Pamphlet"
subtitle: "Detecting performative iteration in LLM outputs"
author: "Thomas Schüller"
orcid: "0009-0003-9799-6747"
contact: "forschung@tschueller.com"
version: "Pamphlet v0.4 (English edition)"
date: "2026-10-06"
status: "Short version of the working note for external reading"
license: "CC BY 4.0"
language: "en"
kennung: "FORSCH-PROV0459"
slug: "iteratio-simulata"
register_hinweis: "Number provisional; final assignment against the work register"

Iteratio Simulata

When a single LLM output imitates the structure of a research session without executing it.

What this is about

A pattern increasingly appears in LLM outputs: a prompt demands "15 iterations", "hard selection", "a critical reviewer who discards weak concepts" — and the model returns a text that imitates this form rhetorically perfectly, without the iteration actually taking place. Concepts are renamed but not eliminated. Evaluation criteria appear after the evaluation. Established research is reframed as a breakthrough. Across iteration steps, figures, prices and margins become ever more precise without a single one of them being supported.

The phenomenon is **not a factual error** (individual statements may be correct) and **not mere sycophancy** (the tone may be blunt and confrontational). It is a class of its own: the structural gap between *form* and *substance* of iterative research in a single-agent output.

We call it **Iteratio Simulata**.

Why it matters

Multi-agent debate frameworks (VirSci, MARG, MAD) show what genuine iteration with distributed roles looks like: separate agents, documented eliminations, an external judge, falsification pressure. Single-agent prompts such as the "Autonomous Inventor Mode" *imitate* this structure within one coherent output — and that is exactly what makes them attractive and dangerous. To non-specialists they read as structured research; they are linearly narrated brainstorming with ornamental framing.

The empirical basis for this claim already exists. Huang et al. (ICLR 2024) show that large language models cannot reliably correct their own reasoning without external feedback. If self-correction structurally does not work, then an output claiming self-correction is necessarily performative.

Five markers

M1 — Elimination claim without elimination trace. The output claims to have discarded candidates but does not name them and gives no reasons.

M2 — Criteria after the evaluation. Evaluation criteria are presented after the selection has already happened, rather than fixed beforehand and applied equally.

M3 — Reframing established research as breakthrough. Concepts established in the literature for years are presented as new discoveries. Measurable by comparison with publication data — methodology adaptable from Afzal et al. (EACL 2026) and patent novelty research.

M4 — Self-validating rhetoric instead of external falsification. High-confidence adjectives without external sources or data against which the evaluation could fail.

M5 — Concretisation asymmetry. Across iteration steps, figures become more precise while the depth of evidence stays low. Precision looks well-founded but is freely set.

Pre-registration as remedy

The central methodological question is not how to detect performative iteration after the fact, but how to make it harder *before* generation. Four minimal mechanisms:

- *Pre-commitment of the candidate set*: fix the complete candidate list before selection and hash it (SHA-256)
- *Pre-commitment of criteria*: fix criteria before selection; mark later changes
- *External anchoring*: at least one indexed source per concept class (DOI, PubMed ID, patent number)
- *Asymmetry audit*: report the ratio of quantitative claims to supported claims

These are not complete protection, but they raise the threshold at which performativity can be produced cheaply.

Position in the research field

The phenomenon sits in the foreground of four neighbouring discourses, fully overlapping with none:

- **Multi-agent debate research** provides the contrast standard — genuine versus simulated iteration
- **Sycophancy and confirmation bias research in MAD** provides related detection metrics for M4
- **Novelty assessment in peer review** provides the methodology for M3
- **Self-correction and faithfulness research** provides the empirical basis

The two closest terminological neighbours are *unfaithful chain-of-thought* (Turpin et al., NeurIPS 2023) and *potemkin understanding* (Mancoridis et al., ICML 2025). The delineation is precise: the former concerns the relation between stated reasoning and model internals, the latter between explaining and applying a concept. *Iteratio Simulata* concerns the relation between claimed and demonstrable process structure — and is the only one of these three testable from the text alone, without model access and without a benchmark setup. This does not hold for the 2026 work on the performative reasoning trace: it measures the same gap but needs model access, and so stands beside the markers rather than ahead of them.

Tool

The markers are operationalised in a standalone module, *Iterationsprüfer*. M1, M2 and M5 are implemented as structural heuristics; M3, M3b and M4 are adapters planned for later phases. The implementation deliberately contains no LLM-as-judge component: a model that generates the same rhetorical patterns cannot reliably recognise them.

What Iteratio Simulata is not

Not every structured brainstorm is *Iteratio Simulata*. Textbooks simplify deliberately. Strategy papers set priorities. Pedagogical texts legitimately reframe established research. The difference lies in **explicit marking as instructional text** versus **masquerade as a research process**. The markers are indications, not verdicts.

Self-application

A work that formulates criteria for performative iteration must submit itself to them. Applying the markers to this working note 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 tool initially penalised carefully written text, scoring a clean control text at maximum on two markers.

The third finding yielded a methodological rule: **negative tests before positive tests**. A detector for performative iteration must first demonstrate that it does not mark genuine iteration as performative.

All three findings are documented openly rather than corrected silently. They are the only empirical observations of this work that do not derive from the occasioning case but from its own genesis — including the observation that the described effect occurs even under deliberate attention to it.

Limits

This short version rests on a single case study (ChatGPT brainstorm on cell-free biocatalysis, May 2026). The markers M1–M5 are proposed as hypotheses, to be validated on further cases. Model dependence is open. The pre-registration mechanism is outlined, not implemented.

References

Full working note *Iteratio Simulata v0.4*, module specification *Iterationsprüfer v0.4*, phase-0 verification report and reference verification report: same author, same date. Licensed CC BY 4.0. Privately funded, no institutional affiliation.

Thomas Schüller — ORCID 0009-0003-9799-6747 — forschung@tschueller.com — Vienna, 2026