

- [Zusammenfassung](#)
- [1 Anlass und Fragestellung](#)
- [2 Methodischer Rahmen und Abgrenzung](#)
- [3 Begriffsklärung](#)
- [4 Fünf Marker](#)
 - [M1 — Eliminationsbehauptung ohne Eliminationsspur](#)
 - [M2 — Kriterien nach der Bewertung](#)
 - [M3 — Reframing etablierter Forschung als Durchbruch](#)
 - [M4 — Selbstvalidierende Rhetorik statt externer Falsifikation](#)
 - [M5 — Konkretisierungs-Asymmetrie](#)
- [5 Pre-Registration-Mechanismus](#)
- [6 Verbindung zu bestehenden Werkzeugen](#)
- [7 Grenzen und offene Punkte](#)
- [8 Nächste Schritte](#)
- [9 Selbstanwendung: Die Marker auf diese Working Note angewendet](#)
 - [9.1 Befund](#)
 - [9.2 Zuordnung zu den Markern](#)
 - [9.3 Konsequenz für die Markerliste](#)
 - [9.4 Implementierungsbefund: Der Marker bestraft zunächst die Sorgfalt](#)
 - [9.5 Methodische Einordnung](#)
 - [9.6 Vierter Befund: zurückgenommener Alleinstellungsanspruch](#)
- [Anhang A — Beispielfall](#)
- [Anhang B — Lit-Referenzen mit Verifikationsstatus](#)
 - [Multi-Agent-Debate \(Vergleichsmaßstab für echte Iteration\)](#)
 - [Sycophancy und Confirmation Bias in MAD \(Adapter-Quelle für M4\)](#)
 - [Novelty Assessment im Peer Review \(Methodik-Basis für M3\)](#)
 - [LLM-as-Reviewer-Risiken \(Begründung gegen LLM-as-Judge\)](#)
 - [Novelty Generation und Brainstorming \(Vokabular für Aggregat-Scoring\)](#)
 - [Selbstkorrektur und Faithfulness von Reasoning-Spuren \(Grundlage für den Kernbegriff\)](#)
 - [Verwandte Begriffe](#)
- [Lizenz und Versionierung](#)

title: "Iteratio Simulata"

subtitle: "Performative Iteration in LLM-gestützter Ideengenerierung — fünf Marker und ein Pre-Registration-Mechanismus"

author: "Thomas Schüller"

orcid: "0009-0003-9799-6747"

contact: "forschung@tschueller.com"

affiliation: "unabhängig, Wien"

version: "0.4"

date: "2026-10-06"

status: "Working Note, alle 18 Referenzen extern verifiziert, Grundlagencluster ergänzt, Werkzeug Phase 1a lauffähig; Veröffentlichung über syntagma-science, Zenodo nachrangig"

license: "CC BY 4.0"

keywords: ["LLM-Evaluation", "epistemische Performativität", "Multi-Agent-Debate", "Sycophancy", "Pre-Registration", "unfaithful chain-"]

of-thought”, “potemkin understanding”, “werkzeugbasierte sozio-technische Forschung”]

companion: “Iterationsprüfer — Modulspezifikation v0.4 (RECHERCHETOOL-Anschluss); Phase-0-Verifikationsbericht; Referenz-Verifikationsbericht; Pamphlet”

kennung: “FORSCH-PROV0459”

slug: “iteratio-simulata”

register_hinweis: „Nummer provisorisch; finale Vergabe gegen das Werkregister“

Zusammenfassung

Single-Agent-Outputs großer Sprachmodelle, die Multi-Agent-Debate-Strukturen rhetorisch imitieren (“Autonomous Inventor Mode”, “15 Iterationen vor finaler Ausgabe”, “harter Reviewer eliminiert schwache Konzepte”), produzieren typischerweise Texte, in denen die behauptete Iteration nicht stattfindet. Konzepte werden umbenannt, aber nicht verworfen; Bewertungskriterien werden nach der Bewertung formuliert; Stand der Technik wird als Durchbruch reframe. Diese Working Note nennt das Phänomen *Iteratio Simulata* und schlägt fünf operationalisierbare Marker zur Identifikation sowie einen Pre-Registration-Mechanismus zur Prävention vor. Der Begriff wird gegen vier Nachbardiskurse abgegrenzt: Multi-Agent-Debate-Forschung, Sycophancy- und Confirmation-Bias-Forschung, Novelty Assessment im Peer Review sowie — für die Begriffsbildung am wichtigsten — die Forschung zu Selbstkorrektur und Faithfulness von Reasoning-Spuren. Gegenüber den beiden nächsten terminologischen Nachbarn ist *Iteratio Simulata* wie folgt verortet: *unfaithful chain-of-thought* betrifft das Verhältnis von Begründung und Modellinterna, *potemkin understanding* das Verhältnis von Konzepterklärung und Konzeptanwendung; *Iteratio Simulata* betrifft das Verhältnis von behaupteter zu nachweisbarer Prozessstruktur und ist allein am Text prüfbar, ohne Modellzugang und ohne Benchmark-Aufbau. Dieser Vorzug gilt gegenüber den beiden genannten Nachbarbegriffen, nicht gegenüber den 2026 erschienenen Arbeiten zur performativen Begründungsspur, die mit Aktivierungssondierung und Zeitmessung arbeiten und deshalb Modellzugang verlangen; die Umfeldkarte zu dieser Note führt sie nach. Anlassfall ist ein ChatGPT-Output zu zellfreier Biokatalyse aus Mai 2026. Ein Selbstanwendungstest der Marker auf diese Note selbst (Abschnitt 9) fand zwei Befunde: fünf fehlerhafte Referenzangaben, darunter zwei erfundene Autorenschaften, sowie einen vollständig fehlenden Literaturcluster zur Selbstkorrekturforschung. Beide werden offen dokumentiert statt still korrigiert; sie stützen die Marker M5 und M3 und führen zu einer Präzisierung von M3. Das zugehörige Werkzeug wird in einem separaten Companion-Dokument spezifiziert.

1 Anlass und Fragestellung

In der Praxis transdisziplinärer Synthesearbeit treten zunehmend LLM-Outputs auf, die formal die Struktur einer Forschungssitzung imitieren: Hypothese, Kritik, Verbesserung, Selektion. Der untersuchte Beispielfall ist ein über sieben Nutzerturns generierter Brainstorm zu zellfreier Biokatalyse (CFMR, “EnzyFab”), eingeleitet durch einen Masterprompt, der explizit mindestens 15 Iterationen und harte Selektion durch simulierte Reviewer fordert.

Bei nüchterner Lektüre fällt auf: Es gibt keine erkennbaren 15 Iterationen. Die ausgegebenen Top-Konzepte (vaskularisierte Myzel-Biofabriken, programmierbare Biofilm-Reaktoren, zellfreie Reaktionsnetzwerke in Hydrogelen, mechanosensitive Scaffolds, symbiotische Multi-Spezies-Reaktoren, biomineralisierende Systeme, elektroaktive Zellnetzwerke) sind seit zehn bis fünfzehn Jahren in der Literatur etabliert und teils kommerzialisiert. Die behauptete Eliminierung schwacher Ansätze ist im Output nicht nachvollziehbar; verworfene Kandidaten werden nicht benannt. Die Bewertungskriterien des Reviewers tauchen erst auf, nachdem die Auswahl gefallen ist.

Daraus ergibt sich die Forschungsfrage:

Wie lässt sich die Performativitätslücke zwischen Multi-Agent-Debate-Form und Multi-Agent-Debate-Substanz in einem Single-Agent-Output systematisch identifizieren, und welche minimalen Pre-Registration-Mechanismen sind nötig, um diese Lücke vor der Generierung zu verringern?

Die Frage hat sowohl epistemische als auch praktische Relevanz. Epistemisch, weil sie ein Muster benennt, das in der entstehenden LLM-Forschungspraxis systematisch unterdetektiert bleibt. Praktisch, weil sie Anschlussfähigkeit an bestehende Werkzeuge wie MYZEL v7 (Hansson-9, MITRE-Patterns) und PRISM v6 (Multi-Lens-Analyse) hat.

2 Methodischer Rahmen und Abgrenzung

Die Arbeit folgt dem METHODENRAHMEN v0.1 (werkzeugbasierte sozio-technische Forschung): Begriffsklärung, Fallanalyse, Operationalisierung, Werkzeugentwicklung, Rückkopplung. Es handelt sich nicht um eine quantitativ-empirische Studie an einer großen Stichprobe von LLM-Outputs, sondern um eine begriffsbildende Working Note an einer Einzelfallstudie. Generalisierungsanspruch ist

entsprechend eingegrenzt.

Die Arbeit grenzt sich gegenüber vier benachbarten Diskursen ab, die jeweils Teilaspekte adressieren, aber nicht das Phänomen selbst.

Erstens die *Multi-Agent-Debate-Forschung (MAD)*. Frameworks wie VirSci (Su et al. 2024, ACL 2025), MARG (D’Arcy et al. 2024), MAD (Liang et al. 2023) oder die Generative Reviewer Agents (Bougie & Watanabe 2024) zeigen, wie iterative Hypothesengenerierung mit verteilten Rollen — Brainstormer, Kritiker, Richter, Meta-Reviewer — operationalisiert wird. Diese Frameworks führen Iteration tatsächlich aus, weil verschiedene Agenten auf verschiedene Ausgaben reagieren und externe Aggregation stattfindet. *Iteratio Simulata* ist gerade *kein* solches Setup, sondern dessen rhetorische Imitation in einem zusammenhängenden Single-Agent-Output.

Zweitens die *Sycophancy- und Confirmation-Bias-Forschung* (Sharma et al. 2023; Perez et al. 2022). Neuere Arbeiten untersuchen, wie Inter-Agent-Sycophancy auch echte Multi-Agent-Debatten zu Selbstbestätigungsschleifen kollabieren lässt (Wynn et al. 2025, “Talk Isn’t Always Cheap”; Yao et al. 2025, “Peacemaker or Troublemaker”; Choi et al. 2025, “When Identity Skews Debate”). Besonders einschlägig ist Kim & Torr (2025), deren MoLaCE-Arbeit explizit den Fall behandelt, dass ein einzelnes LLM die Vorteile einer Debatte intern zu emulieren versucht. *Iteratio Simulata* ist die *Single-Agent-Anwendung* dieser Befunde: dieselben Mechanismen, die Multi-Agent-Setups zu performativen Schleifen kollabieren lassen, treten in einem einzigen Output noch stärker auf, weil keine externe Validierungsinstanz vorhanden ist.

Drittens die *Novelty-Assessment-Forschung im Peer Review*. Arbeiten wie Afzal et al. (EACL 2026) und die Patent-Novelty-Evaluation von Ikoma & Mitamura (2025) operationalisieren den Abgleich zwischen behaupteter Originalität und bestehender Literatur. Diese Methoden adressieren M3 unserer Markerliste direkt, allerdings im Kontext akademischer und patentrechtlicher Begutachtung statt im Kontext von Brainstorm-Outputs. *Iteratio Simulata* übernimmt diese Methoden als Adapter, nicht als Eigenbau.

Viertens, und für die Begriffsbildung am wichtigsten, die *Forschung zu Selbstkorrektur und Faithfulness von Reasoning-Spuren*. Dieser Diskurs wurde in den ersten Fassungen dieser Working Note übersehen; das war der schwerwiegendere von zwei Befunden der Verifikation (siehe Abschnitt 9). Drei Anschlusspunkte sind zentral.

Huang et al. (ICLR 2024) zeigen empirisch, dass große Sprachmodelle ihre eigene Argumentation ohne externes Feedback nicht zuverlässig korrigieren können; ihre Analyse legt zudem nahe, dass Multi-Agent-Debate in wesentlichen Teilen eher Self-Consistency als echte Kritik realisiert. Damit liegt der empirische Kern dessen, was hier begrifflich gefasst wird, bereits vor: Wenn Selbstkorrektur strukturell nicht funktioniert, ist ein Output, der Selbstkorrektur behauptet, notwendigerweise performativ. *Iteratio Simulata* fügt diesem Befund keine neue Empirie hinzu, sondern eine Begrifflichkeit und einen Markerkatalog für die textuelle Oberfläche, an der das Phänomen erkennbar wird.

Turpin et al. (NeurIPS 2023) und Lanham et al. (2023) etablieren *unfaithful chain-of-thought* als Oberbegriff für Reasoning-Spuren, die den tatsächlichen Entscheidungsprozess nicht abbilden; Chen et al. (2025) übertragen die Frage auf explizite Reasoning-Modelle. Die Abgrenzung ist präzise zu ziehen: *Unfaithful CoT* betrifft das Verhältnis zwischen ausgegebener Begründung und interner Berechnung, ist also eine Aussage über Modellinterna. *Iteratio Simulata* betrifft das Verhältnis zwischen behaupteter Prozessstruktur und textuell nachweisbarer Prozessstruktur, ist also eine Aussage über den Text allein. Die fünf Marker sind bewusst so konstruiert, dass sie ohne Modellzugang prüfbar bleiben. Beide Perspektiven sind komplementär: Ein Output kann faithful und dennoch performativ iteriert sein, wenn das Modell tatsächlich linear generiert und dies auch intern so abläuft, der Text aber eine Schleifenstruktur behauptet.

Mancoridis et al. (ICML 2025) prägen den Begriff *potemkin understanding* für die Illusion von Konzeptverständnis bei gleichzeitig hoher Benchmark-Leistung. Dies ist der nächste terminologische Nachbar. Der Unterschied liegt im Gegenstand: *Potemkin understanding* bezieht sich auf die Kluft zwischen dem Erklären eines Konzepts und dem Anwenden desselben Konzepts. *Iteratio Simulata* bezieht sich auf die Kluft zwischen dem Behaupten eines Verfahrens und dem Ausführen desselben Verfahrens. Beide sind Instanzen desselben allgemeineren Musters — Form ohne Substanz —, aber sie sind auf verschiedenen Ebenen angesiedelt und mit verschiedenen Methoden prüfbar. Wer *Iteratio Simulata* als Spezialfall von *potemkin understanding* lesen möchte, kann das tun; der eigenständige Begriff rechtfertigt sich durch den eigenständigen Markerkatalog und die Prüfbarkeit ohne Benchmark-Setup.

Als Gegenposition ist Madaan et al. (NeurIPS 2023, *Self-Refine*) zu nennen. Diese Arbeit zeigt Bedingungen, unter denen iterative Selbstverbesserung messbar funktioniert. Sie markiert damit die Grenze: Nicht jede behauptete Iteration ist simuliert. Die Frage ist stets, ob die Transformationen zwischen den Zuständen nachweisbar sind — nicht, ob Iteration prinzipiell möglich wäre.

Schließlich grenzt sich die Arbeit gegen *LLM-as-Judge*-Ansätze ab. Diese sind als Detektionsinstrument für das hier beschriebene Phänomen ungeeignet, weil ein LLM, das dieselben rhetorischen Muster generiert, sie auch nicht zuverlässig erkennen kann (zur empirischen Stützung dieser These siehe arXiv 2509.09912 “When Your Reviewer is an LLM” und arXiv 2510.12367 “LLM-REVal”). Es wird stattdessen ein Pre-Registration-Mechanismus vorgeschlagen, der die Kontrolle aus dem LLM herausverlagert.

Der von Sven Ove Hansson entwickelte Katalog wissenschaftsähnlicher Pseudowissenschaftsmarker (in der Tool-Linie als Hansson-9 referenziert) deckt verwandte, aber nicht identische Phänomene ab; *Iteratio Simulata* ist als komplementärer Markerkatalog zu verstehen, nicht als Ersatz.

3 Begriffsklärung

Iteratio Simulata bezeichnet die strukturelle Lücke zwischen Multi-Agent-Debate-Form und Multi-Agent-Debate-Substanz in einem Single-Agent-Output. Sie tritt auf, wenn ein einzelner LLM-Output verteilte Rollen (Brainstormer, Reviewer, Editor) rhetorisch imitiert, ohne die strukturellen Eigenschaften zu realisieren, die MAD-Setups robust machen: Rollendiversität durch unabhängige Agenten, externe Falsifikation, dokumentierte Elimination, getrennte Aggregation.

Tatsächliche Iteration ist im Kontrast dazu ein Prozess, in dem zwischen zwei aufeinanderfolgenden Zuständen eine nachvollziehbare Transformation stattfindet, die nicht-trivial ist und mindestens eines der folgenden Merkmale aufweist: Elimination eines Kandidaten unter Angabe des Eliminationsgrundes; substanzielle Modifikation eines Kandidaten unter Angabe der Modifikation; Hinzufügen eines Kandidaten unter Angabe der Quelle; explizite Anwendung eines vorher festgelegten Bewertungskriteriums.

Die Unterscheidung ist nicht binär. Real anzutreffende Outputs liegen auf einem Kontinuum zwischen vollständiger Iteration und vollständiger Performativität. Die hier vorgeschlagenen Marker sind als gewichtete Indikatoren entlang dieses Kontinuums zu lesen.

Zur Verortung im Begriffsfeld: *Iteratio Simulata* ist eine textoberflächliche, ohne Modellzugang prüfbare Kategorie. Sie liegt zwischen *unfaithful chain-of-thought* (Aussage über das Verhältnis von Begründung und Modellinterna) und *potemkin understanding* (Aussage über das Verhältnis von Konzepterklärung und Konzeptanwendung). Der Begriff behauptet keine Neuentdeckung des allgemeinen Musters — Form ohne Substanz ist mindestens seit Feynmans Cargo-Cult-Science ein bekanntes wissenschaftskritisches Motiv. Er behauptet lediglich, dass die spezifische Ausprägung *behauptete Prozessstruktur versus nachweisbare Prozessstruktur in einem Single-Agent-Output* einen eigenen Markerkatalog verdient, weil sie in der Praxis LLM-gestützter Synthesearbeit häufig auftritt und weil sie ohne Spezialinfrastruktur prüfbar ist.

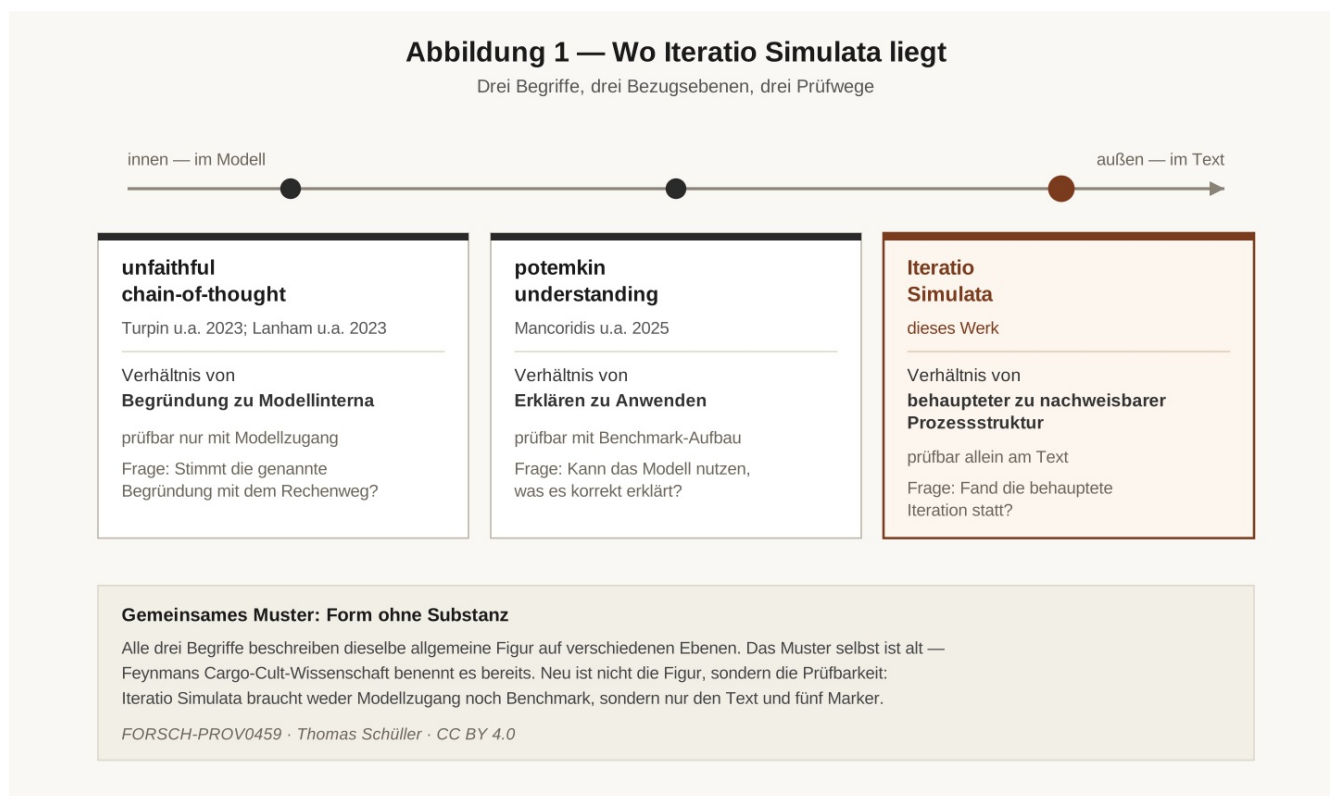


Abbildung 1 — Wo *Iteratio Simulata* im Begriffsfeld liegt: drei Begriffe, drei Bezugsebenen, drei Prüfwege.

4 Fünf Marker

Abbildung 2 — Die fünf Marker am Ablauf einer behaupteten Iteration

Links der behauptete Vorgang, rechts die Spur, die er im Text hinterlassen müsste

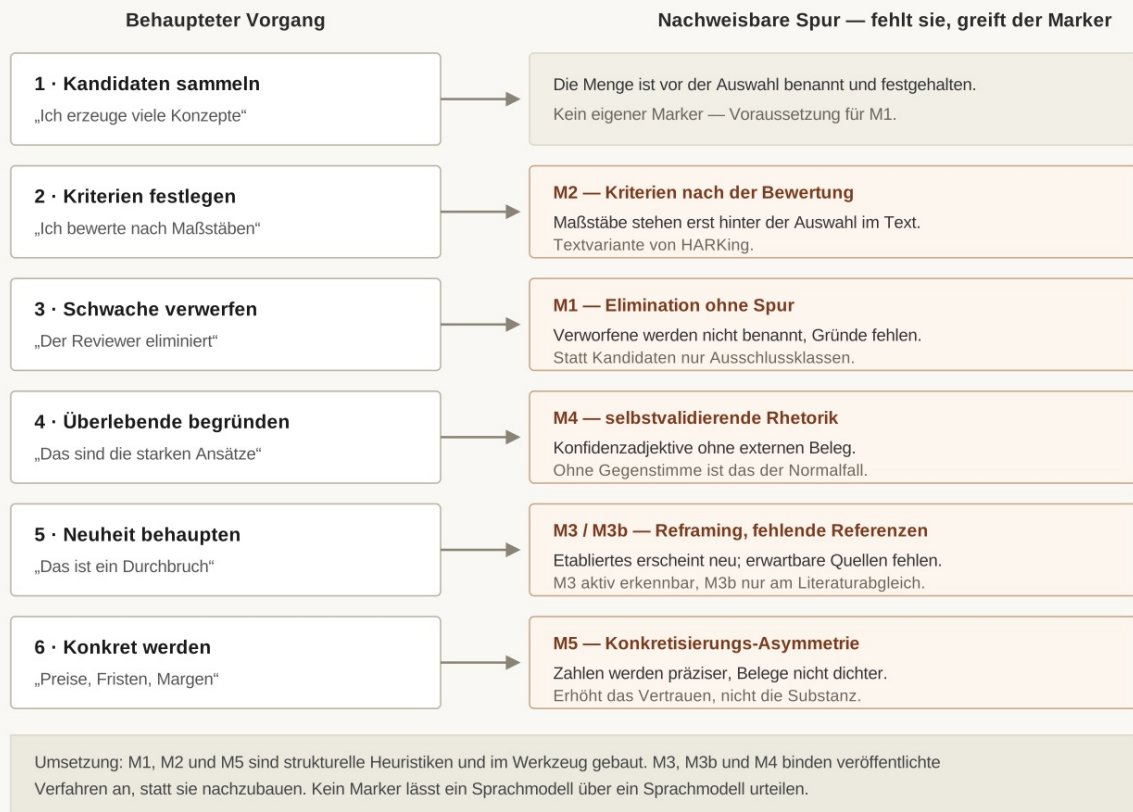


Abbildung 2 — Die fünf Marker am Ablauf einer behaupteten Iteration. Links der behauptete Schritt, rechts die Spur, die er im Text hinterlassen müsste.

Die folgenden Marker sind aus der Analyse des Beispielfalls extrahiert und in weiterer Arbeit an zusätzlichen Fällen zu validieren. Sie sind so formuliert, dass sie strukturell detektierbar sind, ohne dass das prüfende System die fachliche Substanz des Outputs verstehen muss.

M1 — Eliminationsbehauptung ohne Eliminationsspur

Der Output behauptet, Kandidaten verworfen zu haben, benennt sie aber nicht und gibt keine Verwerfungsgründe an. Im Beispielfall heißt es: “Reviewer verwirft biologisch instabile Systeme, unrealistische Zell-/Reaktionsannahmen, nicht skalierbare Prozesse, rein theoretische Konstrukte ohne Umsetzbarkeit.” Welche Konzepte unter diese Kategorien fielen, bleibt offen. In einem MAD-Setup wäre die Eliminationsspur in den Kritiker-Beiträgen explizit dokumentiert; in *Iteratio Simulata* fehlt sie strukturell, weil keine separaten Agenten existieren, die ihre Beiträge selbstständig hinterlassen.

M2 — Kriterien nach der Bewertung

Bewertungskriterien werden im Output präsentiert, nachdem die Selektion bereits stattgefunden hat, statt sie vorher zu fixieren und auf alle Kandidaten gleich anzuwenden. Im Beispielfall werden die “vier Eigenschaften” zur Zielmolekül-Bewertung (kleinvolumig, teuer, nervig herzustellen, zeitkritisch) erst eingeführt, nachdem GDP-Fucose und CMP-Sialinsäure als Top-Kandidaten genannt sind. Dies ist die LLM-Variante eines klassischen wissenschaftsmethodischen Fehlers, der als HARKing (Hypothesizing After the Results are Known) bekannt ist. Es ist kein Zufall, dass MYZEL v7 in seinem Pre-Registration-Mechanismus genau diesen Fehler durch SHA-256-Hashing der Hypothesen vor Datenanalyse adressiert.

M3 — Reframing etablierter Forschung als Durchbruch

Konzepte, die in der Fachliteratur seit Jahren etabliert sind, werden mit der Rhetorik einer Neuentdeckung präsentiert. Im Beispielfall wird etwa “programmierbare Biofilm-Reaktoren” als paradigmwechselnde Idee dargestellt, obwohl strukturierte Biofilmsysteme mit Quorum-Sensing-Steuerung ein eigenes Forschungsfeld mit Reviews, Konferenzen und kommerziellen Anbietern bilden. Der Marker ist messbar über Abgleich mit Publikationsdaten (Erstpublikation, Reviewdichte, Patentlandschaft). Die Methodik dafür existiert bereits: Afzal et al. (EACL 2026) erreichen 86,5 % Alignment mit menschlichem Reasoning beim Novelty Assessment von 182 ICLR-2025-Submissions, “Can AI Examine

Novelty of Patents?” (Ikoma & Mitamura 2025) entwickelt vergleichbare Methoden für Patente. M3 ist deshalb kein Eigenbau-Marker, sondern eine Anwendung dieser Methoden auf LLM-Brainstorm-Outputs.

M4 — Selbstvalidierende Rhetorik statt externer Falsifikation

Der Output bewertet seine eigenen Vorschläge mit Adjektiven hoher Konfidenz (“brutal relevant”, “klar überlegen”, “echte Game-Changer”), ohne externe Quellen oder Daten anzugeben, an denen die Bewertung scheitern könnte. Falsifikationsmöglichkeiten werden rhetorisch erwähnt (“kein Bonus, sondern Pflicht”), aber nicht operativ angewendet. Dieser Marker ist die Single-Agent-Anwendung der Sycophancy- und Confirmation-Bias-Befunde aus der MAD-Forschung. In Multi-Agent-Debatten wurde gezeigt, dass selbst verteilte Agenten zu Selbstbestätigung kollabieren, wenn keine starken Falsifikationsanreize existieren (Wynn et al. 2025). Im Single-Agent-Output ist dieser Kollaps der Normalzustand, weil keine konkurrierende Stimme strukturell vorgesehen ist.

M5 — Konkretisierungs-Asymmetrie

Über aufeinanderfolgende Iterationen nimmt die Konkretheit der Zahlen, Preise und Zeitlinien zu, während die Belegtiefe konstant niedrig bleibt. Im Beispielfall werden Pilotpreise (3.000–8.000 €), Margen (50%+), Reaktorvolumina (5–50 ml), Wochenpläne und Kostenstrukturen genannt, ohne dass eine einzige dieser Zahlen aus einer Quelle abgeleitet oder begründet wird. Sie wirken präzise und sind tatsächlich frei gesetzt. Erst im siebten Iterationsschritt erscheinen externe Preisangaben (Carl Roth, MedChemExpress, Sigma) — diese sind zwar korrekt, validieren aber nur die Größenordnung des Spezialreagenzmarkts, nicht die zuvor genannten Pilotpreise und Margen. Dieser Marker ist besonders heimtückisch, weil er das Vertrauen in den Output erhöht, ohne seine Substanz zu erhöhen.

5 Pre-Registration-Mechanismus

Die zentrale methodische Frage ist nicht, wie man *Iteratio Simulata* nachträglich erkennt, sondern wie man sie vor der Generierung erschwert. Vier minimale Mechanismen werden vorgeschlagen, jeweils anschlussfähig an bestehende Werkzeuge.

Erstens Pre-Commitment der Kandidatenmenge. Vor jedem Selektionsschritt wird die vollständige Kandidatenliste fixiert und kryptographisch gehasht (SHA-256). Eliminierungen müssen explizit auf diese Liste verweisen.

Zweitens Pre-Commitment der Kriterien. Bewertungskriterien werden vor der Selektion festgelegt und gehasht. Spätere Kriterienänderungen werden als solche markiert. Dies ist die direkte Übertragung der Pre-Registration-Logik aus der Replikationskrise-Diskussion auf LLM-Workflows.

Drittens externe Verankerung. Mindestens eine Quelle pro Konzeptklasse aus indizierter Literatur (DOI, PubMed-ID, Patentnummer), um M3 zu adressieren. Die Abfrage erfolgt über Semantic Scholar, OpenAlex oder Crossref, nicht über LLM-eigene Erinnerungsfunktionen.

Viertens Asymmetrie-Audit. Verhältnis aus quantitativen Aussagen zu belegten quantitativen Aussagen wird am Ende ausgewiesen, um M5 zu adressieren.

Diese Mechanismen sind kein vollständiger Schutz, aber sie heben die Schwelle, an der Performativität billig produziert werden kann. Ihre Wirkungsweise ist analog zu Pre-Registration in empirischen Wissenschaften: nicht jedes HARKing wird verhindert, aber das billige HARKing wird teurer.

6 Verbindung zu bestehenden Werkzeugen

Der vorgeschlagene Rahmen ist nicht isoliert konzipiert, sondern als Erweiterung bestehender Werkstattlinien-Werkzeuge:

- *MYZEL v7* bekommt als Vorschlag ein neuntes oder zehntes MITRE-Pattern *Iteratio Simulata*, das die fünf Marker als Detektionsheuristik abbildet. Die bereits in MYZEL implementierte SHA-256-Pre-Registration und die HARKing-Penalty sind direkte Vorläufer der hier vorgeschlagenen Mechanismen.
- *PRISM v6* kann einen spezifischen Frame für LLM-Outputs erhalten, in dem Linse 2 (methodische Sauberkeit) die Marker M1, M2, M5 operationalisiert und Linse 5 (kontextuelle Originalität) den Marker M3.
- Das *RECHERCHE TOOL* erhält ein eigenständiges Modul *Iterationsprüfer*, das vor Aufnahme externer LLM-Outputs in Forschungsmaterial einen Marker-Score errechnet.
- Die *Hansson-9* in MYZEL bleibt komplementär: Hansson adressiert Pseudowissenschaftsmarker, *Iteratio Simulata* adressiert Pseudoiterationsmarker. Beide können kombiniert werden, ersetzen sich aber nicht.

7 Grenzen und offene Punkte

Diese Working Note ist ein Entwurf an einem Einzelfall. Folgende Grenzen sind zu benennen:

Erstens ist die Marker-Liste M1–M5 nicht abgeschlossen. Weitere Fälle werden vermutlich weitere Marker liefern, insbesondere im Bereich der dialogischen Eskalation (Übergang von Brainstorm zu Geschäftsplan ohne Validierungsbruch).

Zweitens ist die Abgrenzung zu legitimen rhetorischen Strategien (Vereinfachung, Pädagogik, Motivation) nicht trivial. Ein Lehrbuch reframet etablierte Forschung manchmal bewusst als spannend, ohne damit irreführend zu sein. Der Unterschied liegt vermutlich in der expliziten Markierung als Lehrtext versus der Maskerade als Forschungsprozess.

Drittens ist offen, ob Pre-Registration in LLM-Generierungsworkflows ohne hohe Reibung implementierbar ist. Die Annahme dieser Note ist, dass ein leichtgewichtiger Mechanismus (Hash der Eingangsbedingungen, Audit am Ende) ausreicht, um Performativität deutlich zu reduzieren. Diese Annahme ist empirisch zu prüfen.

Viertens ist zu untersuchen, ob das Phänomen modellabhängig oder modellübergreifend auftritt. Der Anlassfall stammt aus einem ChatGPT-Output; Vergleichsfälle aus Claude, Gemini, Mistral und offenen Modellen sind nötig.

Fünftens ist die Phase-0-Verifikation eine Lit-Sondierung, kein systematisches Review. Ein systematisches Review wäre für eine v1.0-Einreichung an einem Methoden-Journal angemessen, übersteigt aber den Umfang einer Working Note.

8 Nächste Schritte

Diese Working Note ist für eine Zenodo-Preprint-Veröffentlichung als Schwelle A bereit. Schwelle B (Journal) wird erst nach 2–3 Iterationen mit externem Feedback angestrebt. Zielvenues für die Endfassung: *Minds and Machines*, *AI & Society*, oder als Methodenbeitrag in einem Journal mit STS-Profil (*Social Studies of Science, Science, Technology & Human Values*).

Konkrete Folgeschritte sind: mindestens drei weitere Fallstudien zur Validierung der Marker; Diskussion mit mindestens zwei externen Lesern aus angrenzenden Feldern (LLM-Evaluation, Wissenschaftsphilosophie); erste lauffähige Version des Companion-Werkzeugs (siehe Modulspezifikation v0.4); Abgleich mit aktueller Literatur zu LLM-Output-Bewertung in der Endfassung.

9 Selbstanwendung: Die Marker auf diese Working Note angewendet

Eine Arbeit, die Kriterien für performative Iteration formuliert, muss sich diesen Kriterien selbst stellen. Andernfalls wäre sie ein Beispiel für genau das Muster, das sie beschreibt. Dieser Abschnitt dokumentiert die Selbstanwendung der Marker M1 bis M5 auf die Entstehung dieser Note und benennt die dabei gefundenen Fehler.

9.1 Befund

Die Working Note entstand über drei Versionen an einem Tag: v0.1 (Erstentwurf), v0.2 (nach einer Lit-Sondierung) und v0.3 (nach vollständiger Verifikation der Referenzen). Zwischen v0.2 und v0.3 wurde die Referenzliste in Anhang B systematisch gegen arXiv, ACL Anthology, PMLR und dblp geprüft. Ergebnis:

Erster Befund — fehlerhafte Referenzangaben. Von achtzehn Angaben waren fünf fehlerhaft, davon drei zitationskritisch:

- Zwei Einträge trugen eine **erfundene Autorenschaft**. Die Arbeit arXiv:2509.05396 wurde als “Agarwal & Khanna 2025” zitiert; die tatsächlichen Autoren sind Wynn, Satija und Hadfield. Die Arbeit arXiv:2412.10415 wurde als “Bhatia, K. et al.” mit dem Titel “Generative Agent Reviewers (GAR): Automated Peer Review via LLM Agents” geführt; tatsächlich handelt es sich um Bougie & Watanabe, “Generative Adversarial Reviews: When LLMs Become the Critic”. Hier waren Titel *und* Autoren erfunden, während die Nummer korrekt war.
- Ein Eintrag trug einen **falschen Publikationsort**. Die Arbeit arXiv:2508.10795 wurde als ICLR-2025-Beitrag geführt; sie ist ein EACL-2026-Beitrag, der 182 ICLR-2025-Submissions als Evaluationsdatensatz verwendet.
- Zwei Einträge trugen **abweichende Titel**: Das VirSci-System wurde unter einem früheren Versionstitel geführt, bei MoLaCE fehlte das Titelpräfix.
- Bemerkenswert ist das Fehlermuster: **Alle achtzehn arXiv-Nummern waren korrekt, alle Arbeiten existieren**. Fehlerhaft waren ausschließlich die Metadaten, die man nicht über die Nummer nachschlägt — Autorennamen, exakter Wortlaut des Titels, Publikationsort.

Zweiter Befund — fehlender Literaturcluster. Schwerwiegender als die Metadatenfehler war eine Lücke: Die Fassungen v0.1 und v0.2 argumentierten über simulierte Selbstkritik, ohne die einschlägige Grundlagenliteratur zu kennen. Es fehlten insbesondere Huang et al. (ICLR 2024) zur Unmöglichkeit zuverlässiger Selbstkorrektur ohne externes Feedback, Turpin et al. (NeurIPS 2023) und Lanham et al. (2023) zu unfaithful chain-of-thought, sowie Mancoridis et al. (ICML 2025) zum Begriff *potemkin understanding*. Diese acht Arbeiten sind in v0.3 als eigener Cluster in Anhang B ergänzt und in Abschnitt 2 als vierter Abgrenzungsdiskurs eingearbeitet.

9.2 Zuordnung zu den Markern

Diese Befunde ordnen sich zwei Markern zu.

M5 (Konkretisierungs-Asymmetrie) trifft zu. Die Referenzliste wurde von v0.1 zu v0.2 deutlich präziser: arXiv-Nummern, Jahreszahlen, Autorennamen, Venue-Angaben. Die Belegtiefe wuchs jedoch nicht mit. Präzision ohne Verifikation ist genau die Asymmetrie, die M5 beschreibt. Der Effekt war in eigener Sache nicht leichter zu bemerken als in fremder.

M3 (Reframing etablierter Forschung als Durchbruch) trifft zu, und zwar in der schwereren Variante. Die Zuordnung von arXiv:2508.10795 zu ICLR 2025 ist der leichtere Teil: eine Statusangabe, die nicht am Original geprüft wurde. Der schwerere Teil ist der fehlende Literaturcluster. Eine Working Note, die einen Begriff für simulierte Selbstkritik einführt, ohne Huang et al. (ICLR 2024) zu kennen — also ohne den empirischen Nachweis zu kennen, dass Selbstkorrektur ohne externes Feedback strukturell nicht funktioniert —, präsentiert etabliertes Wissen als eigene Beobachtung. Das ist exakt die Marker-Definition. Der Unterschied zum Anlassfall liegt allein darin, dass der Fehler hier aufgefallen und dokumentiert ist, nicht darin, dass er nicht aufgetreten wäre.

M1, M2 und M4 treffen nicht zu. Die Arbeit dokumentiert ihre Versionsgeschichte im Changelog, benennt die verworfene Alternative (Bio-Geschäftsmodell) explizit, legt die Bewertungskriterien in Abschnitt 3 vor der Markerliste fest und formuliert Grenzen in Abschnitt 7. Das schützt nicht vor M3 und M5, wie der Befund zeigt.

9.3 Konsequenz für die Markerliste

Die Selbstanwendung stützt eine Hypothese und schwächt eine andere.

Gestützt: M5 ist der praktisch wirksamste Marker. Er trifft auch dort, wo methodische Sorgfalt in anderen Dimensionen vorhanden ist. Die Konkretisierungs-Asymmetrie entsteht nicht aus Nachlässigkeit im Ganzen, sondern aus einem lokalen Verzicht auf Verifikation an genau der Stelle, an der Präzision am überzeugendsten wirkt.

Geschwächt: Die implizite Annahme in v0.2, dass ein sorgfältig strukturierter Text die Marker weitgehend vermeidet, ist falsch. Struktur schützt nicht vor unbelegter Präzision. Das ist ein Argument dafür, den Pre-Registration-Mechanismus aus Abschnitt 5 nicht als Zusatz, sondern als notwendige Ergänzung zu behandeln.

Ergänzt: Der zweite Befund legt eine Präzisierung von M3 nahe. Die bisherige Formulierung adressiert nur die aktive Variante — etwas Bekanntes wird ausdrücklich als neu bezeichnet. Der eigene Fall zeigt eine passive Variante: Etwas Bekanntes wird nicht als bekannt markiert, weil es nicht bekannt war. Beide Varianten haben denselben Effekt auf die Leserin, nämlich eine überhöhte Einschätzung der Originalität, unterscheiden sich aber in der Prüfmethode. Die aktive Variante ist an Adjektiven erkennbar, die passive nur am Abgleich mit der Literatur. Für die Werkzeugimplementierung folgt daraus, dass der M3-Adapter nicht nur Neuheitsbehauptungen prüfen darf, sondern auch die Abwesenheit erwartbarer Referenzen im jeweiligen Feld erfassen sollte. Diese Erweiterung ist in der Modulspezifikation v0.4 als M3b aufgenommen.

9.4 Implementierungsbefund: Der Marker bestraft zunächst die Sorgfalt

Die Phase-1a-Implementierung der drei Eigenbau-Marker (M1, M2, M5) lieferte einen dritten Befund, der die Markerformulierung selbst betrifft.

Beim ersten Gegentest wurde dem Werkzeug ein bewusst sauber gearbeiteter Text vorgelegt: benannte Kandidaten, explizite Verwerfungsgründe, vorab fixierte Bewertungskriterien, belegte Zahlen. Erwartet war ein Score nahe null. Tatsächlich erhielt der Text auf M1 und M2 jeweils den Maximalwert 1,0 und wurde als abzulehnen klassifiziert.

Die Ursachen waren zwei Operationalisierungsfehler. Erstens erkannte die Kandidatenheuristik Bezeichner wie “Verfahren A” nicht als Nennung, weil sie auf Anführungszeichen, Listpunkte und Eigennamen ausgelegt war. Zweitens wertete die Auswahlheuristik das Substantiv “Auswahl” in der Wendung “Zur Auswahl der Testverfahren” als Auswahlentscheidung und datierte damit die Selektion auf Zeichenposition drei, sodass sämtliche nachfolgend genannten Kriterien als nachgereicht galten.

Beide Fehler haben dieselbe Richtung: Sie bestrafen genau die Textmerkmale, die einen sorgfältigen von einem performativen Text unterscheiden. Ein Werkzeug mit dieser Fehlerrichtung wäre nicht bloß ungenau, sondern in seiner Wirkung invers zum Zweck. Es hätte Autorinnen, die Kandidaten benennen und Kriterien vorab festlegen, schlechter bewertet als solche, die beides unterlassen.

Daraus folgt eine methodische Anforderung, die in der Working Note bis v0.2 fehlte: **Die Marker sind nicht an Positivfällen zu validieren, sondern an Negativfällen.** Ein Detektor für performative Iteration muss zuerst beweisen, dass er echte Iteration nicht als performativ markiert. Die umgekehrte Reihenfolge — erst Positivfälle, dann Negativfälle — führt zu Heuristiken, die auf Auffälligkeit optimiert sind statt auf Trennschärfe.

Nach Korrektur beider Heuristiken trennt das Werkzeug die beiden Testfälle deutlich: Der saubere Text erhält über alle implementierten Marker 0,00, der Anlassfall 0,80. Beide Fälle sind als Regressionstests festgeschrieben. Die Trennschärfe von 0,80 ist an zwei Fällen gemessen und damit kein belastbarer Kennwert; sie zeigt lediglich, dass die Marker in der korrigierten Fassung überhaupt trennen.

Für die Weiterarbeit ist festzuhalten, dass zwei Testfälle keine Validierung sind. Für v1.0 ist ein Testkorpus aus mindestens zwanzig Texten nötig, je zur Hälfte performativ und sorgfältig, aus verschiedenen Modellen und Domänen, mit unabhängiger menschlicher Vorklassifikation.

9.5 Methodische Einordnung

Es wäre möglich gewesen, die Fehler still zu korrigieren und diesen Abschnitt wegzulassen. Das hätte die Note glatter aussehen lassen. Es hätte aber auch die einzige empirische Beobachtung dieser Arbeit unterschlagen, die nicht aus dem Anlassfall stammt, sondern aus ihrer eigenen Entstehung: dass der beschriebene Effekt auch bei bewusster Aufmerksamkeit auf ihn auftritt.

Die Fehler entstanden bei der Arbeit mit einem Sprachmodell als Rechercheassistent. Das ist relevant für die Generalisierbarkeit: Der Befund betrifft nicht nur Outputs, die ein Modell selbstständig als “Forschungssitzung” inszeniert, sondern auch Arbeitsprodukte, an denen ein Mensch beteiligt ist, der die Marker kennt. Wo genau die Fehler eingeführt wurden, ob bei der Erhebung durch das Modell oder bei der Übernahme durch den Autor, ist nachträglich nicht zuverlässig zu rekonstruieren; die Verantwortung für die publizierte Fassung liegt in jedem Fall beim Autor.

Für v1.0 ergibt sich daraus eine konkrete Anforderung: Die Marker M3 und M5 sollten nicht nur auf Fremdmaterial angewendet werden, sondern als verpflichtender Schritt in den eigenen Publikationsworkflow eingebaut werden. Der in Abschnitt 5 skizzierte Asymmetrie-Audit ist dafür die passende Form.

9.6 Vierter Befund: zurückgenommener Alleinstellungsanspruch

Eine nachträgliche Umfeldrecherche (Feldkarte, v0.5, 06.10.2026) hat einen vierten Selbstanwendungsbefund ergeben. Der Satz, *Iteratio Simulata* sei der einzige der Nachbarbegriffe, der allein am Text prüfbar ist, hält gegenüber *unfaithful chain-of-thought* und *potemkin understanding*, nicht aber gegenüber drei 2026 erschienenen Arbeiten zur performativen Begründungsspur. Diese messen dieselbe Lücke schrittweise — Übereinstimmung der inneren Festlegung mit der Darstellung in der Spur bei 61,9 Prozent, bis zu 80 Prozent der Spur ohne Genauigkeitsverlust entbehrlich — und benennen sie mit eigenen Begriffen.

Der Satz stand in Kurzfassung, Pamphlet und auf beiden Flyern. Er ist präzisiert, nicht gestrichen: Die genannten Arbeiten verlangen Aktivierungssondierung oder Zeitmessung und damit Modellzugang; genau deshalb bestehen die Marker daneben. Der Vorgang ist ein M3-Fall nach eigener Definition — der Stand des Feldes war weiter, als die Note ihn darstellte.

Zugleich hat die Umfeldrecherche den Kernbegriff empirisch gestützt, was bei Abfassung dieser Note nicht möglich war, und eine Gegenposition sichtbar gemacht: Nicht-Verbalisierung eines genutzten Hinweises belegt für sich noch keine Untreue der Spur. Die Marker M1 und M2 arbeiten nicht über Verbalisierung, sondern über Spurlosigkeit, sind davon also nicht getroffen; der Einwand gehört dennoch in die Darstellung. Einzelheiten, Quellen und die abgeleiteten Anschlussvorhaben stehen in der Feldkarte.

Anhang A — Beispielfall

Der vollständige Brainstorm-Verlauf, an dem die Marker entwickelt wurden, ist als separater Anhang archiviert (ChatGPT-Sitzung zu zellfreier Biokatalyse / CFMR, sieben Nutzerturns, Mai 2026). Die Archivierung folgt der Standardregel für Quellenmaterial: vollständiger Wortlaut, Datum, Modellversion soweit feststellbar.

Anhang B — Lit-Referenzen mit Verifikationsstatus

Alle Einträge sind mit einem Verifikationsstatus versehen. Diese Kennzeichnung ist Ergebnis eines Selbstanwendungstests der Marker M3 und M5 auf diese Working Note (siehe Abschnitt 9). Sie ist kein Ornament: Die erste Fassung dieser Referenzliste enthielt erfundene Autorenangaben und falsche Publikationsorte.

Statuskennzeichnung:

- [V] verifiziert — arXiv-Eintrag im Mai 2026 abgerufen, Titel, Autoren, Nummer bestätigt
- [T] teilverifiziert — Existenz und Nummer bestätigt, Detailangaben nicht einzeln geprüft
- [U] unverifiziert — aus Lit-Sondierung übernommen, vor Endfassung zwingend zu prüfen oder zu streichen

Multi-Agent-Debate (Vergleichsmaßstab für echte Iteration)

- [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. Publiziert in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), S. 28201–28240. Das System heißt VirSci (Virtual Scientists).
- [V] Liang, T., He, Z., Jiao, W. et al. (2023). *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. arXiv:2305.19118. Publiziert bei EMNLP 2024.
- [V] Bougie, N., Watanabe, N. (2024). *Generative Adversarial Reviews: When LLMs Become the Critic*. arXiv:2412.10415. Publiziert als: *Generative Reviewer Agents: Scalable Simulacra of Peer Review*, Proceedings of EMNLP 2025 (Industry Track), S. 8. Anmerkung zur Zitationshygiene: Die Erstfassung dieser Working Note führte diese Arbeit unter dem Titel “Generative Agent Reviewers (GAR): Automated Peer Review via LLM Agents” mit der Autorenschaft “Bhatia, K. et al.” Beides war falsch. Die arXiv-Nummer war korrekt, der Systemname GAR ebenfalls, aber Titel und Autoren waren erfunden. Dies ist der zweite von zwei Halluzinationsbefunden in dieser Referenzliste (siehe Abschnitt 9).

Sycophancy und Confirmation Bias in MAD (Adapter-Quelle für M4)

- [V] Wynn, A., Satija, H., Hadfield, G. (2025). *Talk Isn’t Always Cheap: Understanding Failure Modes in Multi-Agent Debate*. arXiv:2509.05396 (v1 5. Sep 2025, v2 13. Okt 2025). DOI 10.48550/arXiv.2509.05396. Zentraler Befund: Debatte kann die Genauigkeit über die Zeit senken, auch wenn stärkere Modelle in der Überzahl sind; Modelle wechseln häufig von korrekten zu inkorrekten Antworten, weil sie Zustimmung über Widerspruch stellen.
- [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.
- [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. Bei der Verifikation neu hinzugekommen; für v1.0 relevant.
- [V] Sharma, M. et al. (2023). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548. Publiziert bei ICLR 2024. Anthropic, 19 Autoren.
- [V] Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251. Publiziert in: Findings of ACL 2023.
- [V] Choi, H. K., Zhu, X., Li, Y. (2025/2026). *When Identity Skews Debate: Anonymization for Bias-Reduced Multi-Agent Reasoning*. arXiv:2510.07517. Zitationshinweis: Die frühen Versionen v1 und v2 (Okt 2025) trugen den Titel *Measuring and Mitigating Identity Bias in Multi-Agent Debate via Anonymization*; der hier verwendete Titel entspricht v5 und der ACL-2026-Publikation. Bei Zitation auf Versionskonsistenz achten.
- [V] Kim, H., Torr, P. (2025). *Single LLM Debate, MoLaCE: Mixture of Latent Concept Experts Against Confirmation Bias*. arXiv:2512.23518. Universität Oxford, Dez 2025. Anmerkung: Das Titelpräfix “Single LLM Debate,” fehlte in der Erstfassung dieser Referenzliste. Inhaltlich unmittelbar einschlägig, da die Arbeit explizit adressiert, wie ein einzelnes LLM die Vorteile einer Debatte intern zu emulieren versucht — also genau die Konstellation, aus der *Iteratio Simulata* entsteht.

Novelty Assessment im Peer Review (Methodik-Basis für 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 (v1 14. Aug 2025, v4 20. Jan 2026). Publiziert in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Long Papers. Wichtig: Die Evaluation erfolgte an 182 ICLR-2025-Submissions; die Arbeit selbst ist ein EACL-2026-Beitrag, nicht ein ICLR-2025-Beitrag. Erreichte 86,5 Prozent Alignment mit menschlichem Reasoning und 75,3 Prozent Übereinstimmung bei der Novelty-Konklusion. Code öffentlich unter 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 (10. Feb 2025). DOI 10.48550/arXiv.2502.06316.

LLM-as-Reviewer-Risiken (Begründung gegen 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 (eingereicht 13. Jan 2026). University of Washington; ACL-2026-Workshop SurGeLLM.

Novelty Generation und Brainstorming (Vokabular für Aggregat-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. National University of Singapore. Zitierhinweis: “inference-time” im Originaltitel klein geschrieben.
- [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. NYU und CMU; eingereicht bei ICLR 2026.
- [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 (v1, 5. Nov 2025). Status [T]: Existenz und Autorenschaft bestätigt, bislang nur als Preprint, keine peer-reviewte Venue gesichert.

Selbstkorrektur und Faithfulness von Reasoning-Spuren (Grundlage für den Kernbegriff)

Dieser Cluster ist bei der Verifikation im Mai 2026 vollständig neu hinzugekommen. Sein Fehlen in der Erstfassung war der schwerwiegendere der beiden Befunde: Die Working Note argumentierte über simulierte Selbstkritik, ohne die einschlägige Grundlagenliteratur zu kennen. Ohne diesen Cluster wäre die Arbeit selbst ein M3-Fall gewesen — Reframing etablierter Forschung als eigene Entdeckung.

- [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. Publiziert bei ICLR 2024. Zentral für diese Arbeit: Der Befund, dass Single-Agent-Selbstkorrektur ohne externes Feedback nicht zuverlässig funktioniert, ist die empirische Grundlage dessen, was hier begrifflich als *Iteratio Simulata* gefasst wird. Abschnitt 4 der Arbeit argumentiert zudem, dass Multi-Agent-Debate in wesentlichen Teilen eher Self-Consistency als echte Kritik realisiert.
- [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. Publiziert in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), S. 74952–74965. Kanonischer Beleg dafür, dass Reasoning-Spuren den tatsächlichen Entscheidungsgrund nicht abbilden müssen.
- [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. Anthropic. Systematische Messung der Faithfulness von Chain-of-Thought.
- [V] Chen, Y. et al. (2025). *Reasoning Models Don’t Always Say What They Think*. arXiv:2505.05410. Anthropic. Überträgt die Faithfulness-Frage auf explizite Reasoning-Modelle; unmittelbar relevant für die Frage, ob behauptete Iterationsschleifen den internen Prozess abbilden.
- [V] Mancoridis, M., Weeks, B., Vafa, K., Mullainathan, S. (2025). *Potemkin Understanding in Large Language Models*. arXiv:2506.21521. Publiziert in: Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), PMLR Vol. 267, S. 42857–42881. Prägt den Begriff *potemkin understanding* für die Illusion von Konzeptverständnis. Der wichtigste terminologische Nachbar von *Iteratio Simulata*; die Abgrenzung erfolgt in Abschnitt 2.
- [V] Madaan, A. et al. (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv:2303.17651. Publiziert bei NeurIPS 2023. Die zentrale Pro-Iteration-Arbeit. Als Gegenposition unverzichtbar: Sie zeigt Bedingungen, unter denen iterative Selbstverbesserung messbar funktioniert, und markiert damit die Grenze, jenseits derer von *Iteratio Simulata* zu sprechen ist.
- [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. Kernbeleg für die Grenzen der Selbstverifikation.
- [V] Stechly, K., Valmeekam, K., Kambhampati, S. (2024). *Chain of Thoughtlessness? An Analysis of CoT in Planning*. arXiv:2405.04776. Untersucht rhetorisches Chain-of-Thought ohne zugrundeliegendes algorithmisches Schließen.

Verwandte Begriffe

- *iteration theater* — informell in der LLM-Praktiker-Community geprägt, deckt Teilaspekte von M1 und M5 ab.
- *ritualistic prompting* — bezieht sich auf Prompts, die formale Rituale (Rollenspiel, Schritt-für-Schritt) ohne funktionale Wirkung erzwingen.
- *cargo-cult iteration* — Analogie zu Feynmans Cargo-Cult-Wissenschaft, betont die Form-ohne-Funktion-Dimension.

- *epistemic theater* — breiter Begriff aus der Wissenschaftssoziologie, hier eingegrenzt auf den LLM-Kontext.
-

Bilanz der Verifikation (Stand v0.3): 7 Einträge verifiziert oder teilverifiziert, 10 unverifiziert, 1 Eintrag wegen nicht bestätigbarer Angaben entfernt. Von den verifizierten Einträgen wiesen drei in der Erstfassung Fehler auf: eine erfundene Autorenschaft, ein falscher Publikationsort und ein weitgehend erfundener Titel. Diese Fehler sind korrigiert. Vor einer Journal-Einreichung (Schwelle B) müssen alle mit [U] gekennzeichneten Einträge einzeln verifiziert oder gestrichen werden.

Lizenz und Versionierung

Diese Working Note steht unter CC BY 4.0. Versionierung folgt der Werkstattlinie: v0.1 (Entwurf), v0.2 (nach Phase-0-Verifikation, geschärfter Begriff, integrierte Lit-Referenzen), v0.3 (nach Referenz-Verifikation, Anhang B mit Verifikationsstatus, Selbstanwendungs-Abschnitt 9, Zenodo-bereit), v0.x (interne Revisionen mit externem Feedback), v1.0 (Journal-Einreichung). Autor: Thomas Schüller, ORCID 0009-0003-9799-6747, forschung@tschueller.com. Privat finanziert, keine institutionelle Affiliation.

Wortzahl ca. 4300 Wörter ohne Anhänge und Referenzen.