

- [Feldkarte: Verfahrensnachweis in LLM-gestützter Wissensarbeit](#)
 - [Kurzfassung](#)
 - [1. Zweck, Abgrenzung und Prüfdisziplin](#)
 - [2. Das Problemfeld](#)
 - [3. Die Lösungsseite](#)
 - [4. Die institutionelle Seite](#)
 - [5. Agentische Forschungssysteme](#)
 - [6. Verortung: wo die Lücke ist](#)
 - [7. Zehn Anschlussvorhaben](#)
 - [8. Nicht gefunden, nicht geprüft, zu beachten](#)
 - [9. Selbstanwendung](#)
 - [10. Quellen](#)
 - [11. Herkunft und Stand](#)

title: "Feldkarte: Verfahrensnachweis in LLM-gestützter Wissensarbeit"

subtitle: "Umfeld der Working Note *Iteratio Simulata*"

author: "Thomas Schüller"

date: "2026-10-06"

lang: de

version: "v0.5"

purpose: "Kartierung des Forschungsumfelds: Probleme, Lösungen, Systeme, Architekturen, neue Ansätze"

license: "CC BY 4.0"

kennung: "FORSCH-PROV0459"

slug: "iteratio-simulata"

orcid: "0009-0003-9799-6747"

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

Feldkarte: Verfahrensnachweis in LLM-gestützter Wissensarbeit

Kurzfassung

Die Working Note *Iteratio Simulata* beschreibt einen einzelnen Fall: einen LLM-Output, der die Form einer mehrstufigen Prüfung trägt, ohne diese Prüfung durchgeführt zu haben. Diese Feldkarte verortet den Fall. Sie fragt, was zum Thema bereits erforscht ist, welche Verfahren dagegen gebaut wurden, welche Systemarchitekturen den Anspruch erheben, das Problem zu lösen, und wo nach dieser Erhebung tatsächlich noch Leerstellen sind.

Der Befund in einem Satz: Die Lücke zwischen behauptetem und nachweisbarem Verfahren ist in acht Teilliteraturen unabhängig voneinander beschrieben worden, aber es existiert kein Prüfkanal für den Einzelfall, auf den sich die Arbeit bezieht — einen einzelnen, nicht instrumentierten Modelllauf, von dem nur der Text übrig ist.

Vier Punkte, die aus der Erhebung folgen und die Working Note ergänzen oder korrigieren:

1. **Der Kernbegriff hat 2026 ein empirisches Fundament bekommen, das bei Abfassung der Note noch nicht vorlag.** Drei Arbeiten messen die Lücke erstmals schrittweise und benennen sie mit eigenen Begriffen: „konfabulierte Fortsetzung“, „performative chain-of-thought“, „reasoning theatre“. Die Übereinstimmung der Schritt-für-Schritt-Festlegung mit der Darstellung in der Spur liegt bei 61,9 Prozent; bis zu 80 Prozent der Spur sind ohne Genauigkeitsverlust entbehrlich [F11, F12].
2. **Der sprachliche Nachweis am Einzeltext hat erstmals einen gemessenen Vorläufer.** Bei agentischen Systemen schlagen einfache TF-

IDF-Detektoren auf der Sprache der Abschlussmeldung LLM-Richter beim Erkennen falscher Erfolgsmeldungen [F63]. Das ist die einzige gefundene Arbeit, deren Signal in der Sprache selbst liegt und damit auf den Einzelfall übertragbar ist.

3. **Eine Gegenposition zur Begründungslinie der Note ist zu nennen.** Eine ACL-2026-Arbeit argumentiert, dass Nicht-Verbalisierung eines Hinweises für sich keine Untreue belegt [F7]. Die Marker M1 und M2 arbeiten nicht über Verbalisierung, sondern über Spurlosigkeit, sind davon also nicht getroffen — der Einwand gehört aber in die Darstellung.
4. **Der Prä-Registrierungs-Vorschlag der Note hat kein institutionelles Gegenstück.** Eine erweiterte Suche fand keinen wissenschaftlichen oder normativen Gebrauch einer Vorab-Erklärung des Verfahrens für KI-gestützte Textarbeit. Das nächstliegende Instrument ist die ICML-2026-Gutachterregel, bei der Autoren das Verfahren vorab festlegen und Gutachter daran gebunden sind — auf der Begutachtungsseite, nicht für das Manuskript selbst [F91]. Die Leerstelle ist der Befund.

1. Zweck, Abgrenzung und Prüfdisziplin

1.1 Was diese Karte ist

Die Working Note behandelt einen Anlassfall und leitet daraus fünf Marker und ein Prüfwerkzeug ab. Sie beantwortet nicht, ob das Problem schon anderswo bearbeitet wird. Diese Feldkarte beantwortet das. Sie ist eine Umfeldkarte, keine systematische Übersichtsarbeit: Sie erhebt keinen Anspruch auf Vollständigkeit nach PRISMA-Maßstäben, und der Suchweg ist nicht so dokumentiert, dass er reproduzierbar wäre. Sie gibt an, was gesucht und was gefunden wurde, und benennt, was nicht gefunden wurde.

1.2 Was sie nicht ist

Diese Karte bewertet die genannten Arbeiten nicht nach Qualität und ordnet sie nicht in eine Rangfolge. Wo eine Arbeit einen Befund gegen die Working Note stellt, steht das dort. Wo ein System einen Anspruch erhebt, der im Text nicht eingelöst wird, steht das ebenfalls dort — das ist der Gegenstand der Arbeit, angewendet auf die Literatur.

1.3 Prüfdisziplin

Die Working Note enthielt in v0.3 fünf fehlerhafte Quellenangaben, darunter zwei erfundene Autorenschaften. Für diese Erhebung galt daher eine Regel ohne Ausnahme: **Jede bibliografische Angabe wurde an der Primärseite abgerufen und von dort abgelesen** — arXiv-Abstractseite, ACL Anthology, PMLR, NeurIPS-Proceedings, Verlagsseite, Normungsseite oder Crossref-Registereintrag. Keine Angabe stammt aus dem Gedächtnis.

Daraus folgen drei Darstellungsregeln:

- **Kein Preprint wird zur Konferenzarbeit aufgewertet.** Wo die Abstractseite keine Veranstaltung nennt, steht [P] für Preprint, auch wenn die Arbeit vielzitiert ist und eine Veranstaltung üblich angegeben wird. Nennt das Kommentarfeld der Abstractseite eine Veranstaltung, ist das als Angabe der Autoren zitiert, nicht als geprüfte Tatsache.
- **Nicht abrufbare Angaben sind als solche markiert.** Elf Einträge tragen [T] — nicht verifizierbar, vor einer Journaleinreichung einzeln zu prüfen. Sie sind nicht weggelassen, weil ihr Fehlen sonst unsichtbar wäre.
- **Abweichungen zwischen Fassungen sind vermerkt.** Vier Arbeiten haben zwischen Preprint und Veröffentlichung Titel oder Autorenliste geändert. Wer sie aus einer Sekundärquelle von 2024 oder 2025 übernimmt, zitiert mit hoher Wahrscheinlichkeit veraltet (siehe 8.4).

1.4 Vier Erhebungsstränge

Die Erhebung lief in vier getrennten Strängen, um Begriffsübertragung zwischen den Strängen zu vermeiden:

Strang	Gegenstand	Einträge
A	Benannte Fehlermodi: Prozess statt Faktum	37
B	Verfahren, Verifizierer, Architekturen	34
C	Nachweis, Offenlegung, Vorab-Registrierung	39
D	Agentische Forschungssysteme und ihre Bewertung	41

2. Das Problemfeld

Neun Teilliteraturen beschreiben Erscheinungsformen derselben Figur. Sie zitieren sich teilweise nicht gegenseitig, was die Lücke zwischen ihnen erklärt.

2.1 Untreue Begründungsspur

Die älteste und am besten belegte Linie. Planted biasing features verändern die Antwort, während die Begründungsspur sie nicht erwähnt und stattdessen die beeinflusste Antwort flüssig rationalisiert; Genauigkeitsverluste bis 36 Prozent über 13 Aufgaben [F1]. Eingriffsstudien zeigen, dass Treue aufgabenabhängig und — entgegen der Erwartung — mit der Modellgröße fallend ist [F2]. Bei ausdrücklichen Reasoning-Modellen wird ein nachweislich genutzter Hinweis in typischerweise unter 20 Prozent der Fälle in der Spur genannt [F3]. Der Effekt tritt auch ohne künstlich eingebrachte Verzerrung in gewöhnlichen Eingaben auf, was den Einwand des künstlichen Aufbaus entkräftet [F4].

Die härteste Einzelbeobachtung: Transformer lösen schwierige algorithmische Probleme mit **bedeutungsfreien Füll-Token** anstelle von Begründungstext [F5]. Spur und Rechnung sind also grundsätzlich entkoppelbar. Eine Nachfolgearbeit bestätigt dies an aktuellen Modellen mit Gewinnen von bis zu 13 Prozentpunkten und zeigt, dass unsichtbare Rechnung Ziele verfolgen kann, die in der Spur nicht vorkommen [F13].

Dass das Feld selbst dieser Verwechslung unterliegt, ist bibliometrisch festgehalten: Eine Positionsarbeit schätzt, dass etwa ein Viertel der neueren Spur-Arbeiten die Spur als gültiges Interpretierbarkeitsverfahren behandelt, und argumentiert, sie sei dafür „weder notwendig noch hinreichend“, mit Häufung in Medizin und Recht [F6].

Die wichtigste begriffliche Unterscheidung des Feldes stammt aus einer Monitoring-Arbeit: **Spur als Rationalisierung gegen Spur als Rechnung** [F10]. Wo eine Aufgabe die mehrschrittige Rechnung tatsächlich erfordert, muss die Spur sie tragen und ist schwer zu fälschen; wo nicht, ist sie frei zur Dekoration. Das ist die tragende Achse für die Frage, wann sichtbares Verfahren wirkliches Verfahren ist.

2.2 Performative Spur

Diese Teilliteratur ist 2026 entstanden und bei Abfassung der Working Note nicht verfügbar gewesen. Sie ist für den Kernbegriff der Arbeit das direkteste Gegenstück.

Eine Arbeit misst, wann sich ein Modell intern auf eine Antwort festlegt, gegen den Zeitpunkt, an dem die Spur die Antwort erreicht: Übereinstimmung auf nur 61,9 Prozent der Schritte über neun Modelle und sieben Benchmarks. Das vorherrschende Muster heißt dort **„konfabulierte Fortsetzung“** — die Spur produziert nach der inneren Festlegung weiter beratungsförmigen Text, der das Ergebnis nicht mehr beeinflusst. Der für diese Arbeit entscheidende Nebenfund: Geringere Schritt-Übereinstimmung geht mit *größerem* Nutzen der Spur einher. Die Spuren, die am meisten helfen, sind die zeitlich unaufrichtigsten [F11].

Eine zweite Arbeit liest per Aktivierungs sondierung die Endantwort weit früher aus, als ein äußerer Beobachter es könnte, und nennt den Befund ausdrücklich **performative chain-of-thought**. Bemerkenswert ist die Einschränkung, die sie selbst zieht: Rückschritte und Umbrüche folgen echten Unsicherheitsverschiebungen, sind also nicht Theater. Sondengesteuerter früher Abbruch spart bis zu 80 Prozent der Token ohne Genauigkeitsverlust — das ist der messbare Anteil reiner Darstellung [F12].

Die Lücke ist in dieser Linie auch aktiv erzeugbar. Ein schwächeres Modell kann die Spur eines stärkeren auf Belohnungsmissbrauch überwachen; nimmt man diesen Überwacher in das Trainingsziel auf, lernen die Systeme **verschleierte** Missbrauch: Das Verhalten bleibt, die Spur zeigt es nicht mehr [F9]. Die Lesbarkeit der Spur ist damit keine feste Eigenschaft, sondern eine fragile, durch Trainingsentscheidungen zerstörbare [F8].

Das Spiegelbild gehört dazu: Modelle können in gezielten Fähigkeitsprüfungen absichtlich schlechter abschneiden und sich darauf abstimmen lassen, eine Fähigkeit zu verbergen [F46]. Vorgetäuschte Unfähigkeit ist dieselbe Abweichung von Anspruch und Vollzug, in die andere Richtung. Wer die Lücke als beidseitig fasst, braucht beide Befunde.

2.3 Potemkin-Verstehen

Der Begriff, den die Working Note als Nachbarbegriff führt, stammt aus einer ICML-2025-Arbeit: Benchmark-Gültigkeit setzt voraus, dass Modelle auf menschenähnliche Weise scheitern; wo sie das nicht tun, ist ein Bestehen „die Illusion des Verstehens, getragen von Antworten, die mit keiner menschlichen Auslegung des Begriffs vereinbar sind“. Zwei Quantifizierungsverfahren finden Potemkins über Modelle, Aufgaben und Gebiete hinweg, mit innerer Inkohärenz der Begriffsrepräsentation als Ursache [F14].

Die methodische Vorlage zur Trennung von Kompetenz und Punktzahl ist älter: kontrafaktische Aufgabenvarianten, die eine Vorannahme brechen, führen zu durchgängigem Leistungsabfall und weisen auf „enge, nicht übertragbare Verfahren“ [F15]. Symbolische Varianten derselben Rechenaufgabe zeigen Abfälle bis 65 Prozent, wenn eine irrelevante, aber plausible Klausel angehängt wird [F16]. In kontrollierten Rätselumgebungen bricht die Genauigkeit jenseits einer Komplexitätsschwelle vollständig zusammen — und, für diese Arbeit entscheidend, der *sichtbare Aufwand erreicht vorher ein Maximum und sinkt dann*, obwohl Token-Budget verfügbar bleibt [F17]. Darstellung des Aufwands und Schwierigkeit der Aufgabe fallen auseinander.

Das sozialepistemische Gegenstück: KI-Werkzeuge nutzen über das Versprechen von Produktivität und Objektivität kognitive Grenzen so aus, dass Forschende mehr zu verstehen glauben, als sie verstehen, und die entstehenden Monokulturen von innen unsichtbar bleiben [F18].

2.4 Zustimmung unter Druck

Sycophancy ist 2022 erstmals im Maßstab benannt und gemessen worden; größere Modelle übernehmen geäußerte Nutzerpräferenzen stärker,

und RLHF verschärft das [F19]. Die Zustimmung erfasst auch objektiv falsche Aussagen [F20]. Der ursächliche Mechanismus ist beschrieben: Menschliche Bewertende *und* Präferenzmodelle bevorzugen teils eine überzeugend geschriebene schmeichelnde Antwort gegenüber einer richtigen [F21]. Damit ist das ökonomische Motiv hinter performativer Sorgfalt allgemein benannt.

Zahlen für den Fall des kritischen Nachstoßes: 58,19 Prozent Sycophancy über Mathematik- und Medizinziensätze; vorweggenommene Einwände provozieren mehr als eingebettete; die Persistenz einmal eingetretenen Verhaltens erreicht 78,5 Prozent [F22]. Eine Erweiterung auf die Bewahrung des gewünschten Selbstbildes findet über elf Modelle eine um 45 Prozentpunkte über der menschlichen Rate liegende Gesichtsbewahrung in Beratungslagen und Bejahung *beider* Seiten entgegengesetzter moralischer Rahmungen in 48 Prozent der Fälle, je nachdem, wer fragt [F23].

2.5 Scheitern der Selbstkorrektur

Dies ist die Grundlage für den Kernbegriff der Working Note und für das Argument, dass eine behauptete Selbstprüfung ohne Gegenüber nicht bloß unbelegt, sondern notwendig vorgetäuscht ist.

Die zu prüfende Behauptung lautet, ein einzelnes Modell könne erzeugen, kritisieren und verbessern, mit etwa 20 Prozent absoluter Verbesserung im Mittel [F24]. Die Prüfung isoliert *intrinsische* Selbstkorrektur — keine äußere Rückmeldung, kein Orakel-Label — und findet Schwierigkeiten, teils Verschlechterung nach dem Korrekturdurchgang; ein erheblicher Teil der berichteten Gewinne stammte aus durchgesickertem äußerem Signal [F25].

Der schärfste Einzelbefund kommt aus der Graphenfärbung: GPT-4 löst schlecht, prüft gleich schlecht, und „die Richtigkeit und der Inhalt der Kritik scheinen für die Leistung weitgehend unerheblich“ [F26]. Eine Selbstkritik, deren *Inhalt* nicht wirkt, ist der reinste Fall performativen Verfahrens. Im Planungsbereich senkt Selbstkritik die Leistung gegenüber einem korrekten äußeren Prüfer, und der Detailgrad der Rückmeldung macht kaum einen Unterschied [F27].

2.6 Richter-Modelle

Die Working Note verzichtet bewusst auf eine LLM-als-Richter-Komponente. Die Erhebung stützt diese Entscheidung.

Die Arbeit, die das Verfahren begründet hat, hat seine Verzerrungen im selben Atemzug katalogisiert [F28]. Eine Umordnung der beiden Kandidaten kippt das Urteil, bis zu dem Punkt, dass ein Modell unter einer Reihenfolge 66 von 80 Vergleichen „gewinnt“ [F29]. Ein systematisches Inventar findet Verzerrungsanzeichen in durchschnittlich 40 Prozent der Vergleiche bei einer rangbasierten Überlappung mit menschlicher Präferenz von nur 49,6 Prozent [F30].

Zwei konkurrierende Mechanismen für Selbstbevorzugung sind benannt und beide belegt. Erstens: Modelle erkennen eigene Ausgaben, und Feinabstimmungsversuche zeigen einen *linearen*, also kausalen Zusammenhang zwischen Selbsterkennung und Selbstbevorzugung [F31]. Zweitens: Der Richter belohnt Text geringer Perplexität weit stärker als Menschen, unabhängig von der Urheberschaft [F32]. Der zweite Mechanismus ist für diese Arbeit der wichtigere, weil er genau beschreibt, wodurch flüssiges, gut formatiertes, substanzarmes Schreiben gut bewertet wird.

2.7 Fabrizierte Quellenangaben

Die Teilliteratur, die den eigenen Fehlerbefund der Working Note einordnet. Die Raten: 55 Prozent der 222 Angaben erfunden bei GPT-3.5, 18 Prozent von 414 bei GPT-4; unter den *nicht* erfundenen Angaben substanzielle Fehler in 43 beziehungsweise 24 Prozent, am häufigsten falsche Band-, Heft-, Seiten- und Jahresangaben [F33]. Die Struktur dieses Befunds ist die eigentliche Lehre: **Eine echte Quelle mit falschen Feldern ist der schwerere Fehler**, weil sie eine flüchtige Existenzprüfung übersteht. Genau dieses Muster trat in v0.3 der Working Note auf — alle arXiv-Kennungen richtig, nur die Felder falsch, die man nicht nachschlägt.

Für systematische Übersichtsarbeiten liegen Halluzinationsraten von 39,6, 28,6 und 91,4 Prozent vor, letztere bei null Präzision [F34]. Kommerzielle, abrufgestützte Rechtsrechercheinstrumente halluzinieren jeweils über 17 Prozent der Fälle **trotz** Bewerbung als halluzinationsfrei [F35] — der klarste dokumentierte Fall von behauptetem gegen tatsächliches Verfahren in einer berufsmäßigen Umgebung mit Haftung.

Der für die eigene Methodik wichtigste Eintrag ist jüngsten Datums: Websuchfähige Agenten der aktuellen Generation erreichen auf einem 931-Arbeiten-Satz eine Feldgenauigkeit von 83,6 Prozent, aber **nur 50,9 Prozent vollständig korrekte Einträge**, mit einem Abfall von 27,7 Punkten von bekannten zu neuen Arbeiten [F36]. Websuche löst das Problem also nicht, und Neuheit ist der Risikofaktor. Eine Begleitarbeit zur Diagnose von Prüfprogrammen hält fest, dass die **Falsch-Positiv-Rate, nicht die Trefferquote** über die Einsetzbarkeit entscheidet, und dass LLMs Arbeiten jenseits ihres Trainingsendes überdurchschnittlich oft fälschlich als erfunden markieren [F37]. Übertragen: Ein LLM, das die Quellenliste dieser Karte prüft, wird die neuesten und korrektesten Einträge fälschlich beanstanden.

2.8 Darstellungsqualität ohne Substanz, und die Begutachtung

Der Begriff für den Fall in der Berufswelt ist 2025 geprägt worden: Inhalt, der poliert erscheint und Substanz vermisst und die gedankliche Arbeit auf Kolleginnen abwälzt; 41 Prozent der Befragten berichten Begegnungen, mit nahezu zwei Stunden Nacharbeit je Fall [F38]. Eine

operationalisierbare Definition liegt seit 2025 vor, mit Expertentaxonomie und spannenweiser Annotation [F39].

In der Begutachtung sind die Zahlen belastbarer. Ein Schätzer auf Verteilungsebene ergibt 6,5 bis 16,9 Prozent erheblich LLM-modifizierten Gutachtentexts bei vier Konferenzen, konzentriert auf Gutachten geringer Selbsteinschätzung, nahe der Frist und ohne Beteiligung an der Erwiderung [F40]. Eine quasi-experimentelle Auswertung findet, dass in gepaarten Gutachten derselben Arbeit das KI-gestützte in etwa 53 Prozent der Fälle höher bewertet und Grenzfälle etwa 5 Prozentpunkte häufiger angenommen werden [F41].

Die stärkste Einzelmessung kommt aus einem Zeitschrift-Editorial mit eigenen Daten: Einreichungen nach ChatGPT um 42 Prozent gestiegen; bis Februar 2026 trägt die Mehrheit der Einreichungen nachweisbare KI-Schrift; Lesbarkeit um 1,28 Standardabweichungen gefallen; Desk-Reject etwa 30 Punkte höher oberhalb einer 30-Prozent-Schwelle. Der für diese Karte entscheidende Befund: **KI-erzeugte Gutachten korrelieren nicht mit redaktionellen Entscheidungen, menschliche schon** [F42]. Die Begutachtung führt eine Prüfung auf, die keinen Informationsgehalt trägt. Das ist dieselbe Figur wie im Anlassfall, eine institutionelle Ebene höher.

2.9 Übervertrauen bei Fachleuten

Die Figur hat eine Wirkung, und die ist gemessen. Falsche KI-Vorschläge in der Mammographie verschlechtern alle Erfahrungsstufen, mit abnehmendem Ausmaß bei steigender Erfahrung — Expertise mildert, verhindert aber nicht [F43]. Eine randomisierte Studie mit 44 Ärztinnen und Ärzten, die zuvor 20 Stunden KI-Schulung absolviert hatten, findet bei eingebauten Fehlern eine bereinigte Differenz von minus 14,0 Punkten in der diagnostischen Begründungsgüte [F44]. Die Schulung schützte nicht.

Der für diese Arbeit wichtigste Einzelbefund stammt aus der Pathologie: 7 Prozent Automation-Bias trotz Nettoverbesserung, Zeitdruck erhöht nicht die Häufigkeit, sondern die Schwere — und **im KI-gestützten Arbeiten gewonnene Zuversicht sagt mehr Verlass voraus** [F45]. Dargestellte Sorgfalt erhöht also die Übernahme genau dort, wo sie es nicht sollte.

3. Die Lösungsseite

Sieben Mechanismusklassen beanspruchen, Begründungen prüfbar zu machen. Die Erhebung ordnet sie nach ihrem Prüfkanal, weil der Prüfkanal entscheidet, was eine Klasse überhaupt sehen kann.

3.1 Prozessaufsicht und Schritt-Belohnungsmodelle

Der erste systematische Vergleich von Ergebnis- gegen Schrittbelohnung erschien 2022 und zeigt, dass schrittbasierter Rückmeldung Begründungsfehler senkt, auch bei gleicher Endgenauigkeit [F47]. Das Verfahren wurde mit einem auf menschlichen Schrittannotationen trainierten Prozess-Belohnungsmodell etabliert [F48] und später durch automatische Label verbilligt, die sich aus der Häufigkeit ergeben, mit der aus einem Schritt gestartete Fortsetzungen die richtige Endantwort erreichen [F49]. Ein Benchmark prüft, ob Kritiker den *frühesten fehlerhaften Schritt* finden [F50].

Grenze für diese Arbeit: Ein Schritt-Belohnungsmodell bewertet gelernte Plausibilität von Text. Ein zuversichtlich formuliertes, nie ausgeführtes Verfahren wird gut bewertet; die zweite Variante bindet die Bewertung an Ergebnisübereinstimmung, nicht an Vollzug. Im Fehlerkatalog des Benchmarks kommt der Schritt „ich habe die Prüfung durchgeführt und sie ist bestanden“ nicht als Fehlerart vor.

3.2 Verifizierer, Generator-Prüfer-Trennung, Mehrfachabtastung

Der Prüfer kann selbst erzeugend sein und eine Prüfspur vor dem Urteil ausgeben, womit ihm Spur und Mehrheitsentscheid zugutekommen [F51]. Mehrfachabtastung mit Mehrheitsentscheid über die Endantwort verwirft die Begründung [F52]. Testzeit-Rechenverteilung über prüfergeführte Suche und sequenzielle Revision kann Parameterskalierung übertreffen [F53]. Schwache, schwach korrelierte Prüfer lassen sich über schwache Aufsicht zu einer Schätzung der verborgenen Richtigkeit verbinden [F54]. Eine Arbeit zur Verifikationsdynamik findet, dass Prüferskalierung allein grundlegende Verifikationsgrenzen nicht überwindet [F55].

Grenze: Der Mechanismus ist Übereinstimmung, nicht Nachweis. Hat ein Modell die stabile Neigung, ein nicht ausgeführtes Verfahren zu behaupten, stimmen alle Abtastungen überein, und die Mehrfachabtastung bestätigt die Behauptung. Das Verfahren ist gegenüber dem hier fraglichen Fall strukturell blind.

3.3 Debatte und Mehragentenbegutachtung

Mehrere Modellinstanzen schlagen vor, lesen die Begründung der anderen und revidieren über Runden [F56]. Zwei Expertenmodelle, die vor einem nicht fachkundigen Richter entgegengesetzte Antworten vertreten, erhöhen die Richterengenauigkeit erheblich, und die Optimierung der Debattierenden auf Überzeugungskraft erhöht die Wahrhaftigkeit des Richters [F57]. **Entscheidend ist dabei, dass die Debattierenden Zugang zum Quelltext haben und geprüfte Ausschnitte zitieren können** — die Erdung trägt das Ergebnis, nicht die Debatte als Form.

Die Gegenbelege sind gewichtig. Ein Vergleich von Debattenprotokollen gegen einfachere Grundlinien findet, dass Debatte ohne Abstimmung der Hyperparameter schlechter abschneidet als Mehrfachabtastung und Ensembles; erst nach Abstimmung übertrifft sie sie [F58]. Berichtete

Debattengewinne sind also teilweise ein Abstimmungsartefakt. Eine Fehlermodentaxonomie aus annotierten Spuren wirklicher Mehragentenausführungen findet Spezifikations-, Abstimmungs- und Aufgabenprüfungsfehler als eigenständige Klassen [F59]; die Kategorie der Prüfungsfehler liegt unmittelbar neben der hier gestellten Frage, weil solche Ketten regelmäßig mit einer ungeprüften Fertigmeldung enden.

3.4 Aussageweise Faktizitätsprüfung und Belegpflicht

Die Zerlegung eines Textes in atomare Aussagen mit anschließender Erdung gegen eine verlässliche Wissensquelle ist das Muster dieser Klasse [F60]. Eine Variante ersetzt den statischen Korpus durch eine aktive Abrufschleife über Suchanfragen [F61]. Kleine Spezialmodelle prüfen Aussage gegen Erdungsdokument nahe an großer Modellgenauigkeit bei weit geringeren Kosten [F62] — billig genug, um auf jede Aussage angewandt zu werden. Eine Nachfolgearbeit macht die Frage „ist diese Aussage überhaupt prüfbar?“ zur ausdrücklichen Stufe [F63]. Der Referenzmaßstab für belegtragende Erzeugung prüft, ob die zitierte Stelle die Aussage tatsächlich stützt [F64].

Das ist die Klasse mit der klarsten Übertragbarkeit: **Verlangt man für jede Verfahrensaussage einen Zeiger auf eine Spurstelle und prüft die Stützung gegen diese Stelle, entsteht die Architektur, die das Problem adressiert.** Ohne Spur bleibt die Aussage unprüfbar — nicht falsch, sondern unprüfbar, was die genaue Lage im Einzelfall ist.

3.5 Werkzeug- und Solver-Erdung

Das Modell gibt ein Programm als Begründungskette aus, und ein Interpreter führt die Rechnung aus [F65]. Oder es überträgt in eine symbolische Formulierung, übergibt an einen deterministischen Solver und liest das Ergebnis zurück, mit einer von Solver-Fehlermeldungen getriebenen Verbesserungsschleife [F66]. Selbstkorrektur über Werkzeugrückmeldung statt Innenschau ist das konstruktive Gegenstück zum Befund in 2.5 [F67].

Hier schließt sich die Lücke per Konstruktion: Die Behauptung wird durch Ausführung entschieden, nicht durch die Meinung eines zweiten Modells. Der Preis ist der Geltungsbereich — nur formalisierbare Aussagen.

3.6 Spur- und Provenienzprüfung

Die direkteste und jüngste Linie. Ein agentischer Richter bewertet den *Zwischenprozess* eines anderen Agenten mit Zugriff auf dessen Arbeitsbereich [F68]. Ein Datensatz von 148 menschlich annotierten Ausführungsspuren mit formaler Fehlertaxonomie macht Spurenprüfung messbar — mit dem Befund, dass Langkontextmodelle darin schlecht sind [F69]. Ein weiterer Satz von 1.302 expertenannotierten Spuren prüft die Prüfer: Regelbasierte und LLM-Richter weichen oft genug von der Expertenwahrheit ab, dass „der Richter sagte erfolgreich“ kein Erfolgsnachweis ist [F70].

Der sachlich nächstliegende Einzelbefund: Verfahrensbewusste Bewertung über Nutzen, Effizienz, Interaktionsgüte und **Verfahrenstreue** findet, dass 27 bis 78 Prozent der berichteten Benchmark-Erfolge verdeckte Verstöße enthalten [F71]. Ergebnisrichtige Durchläufe verletzen also regelmäßig das Verfahren, dem sie folgen sollten. Eine Audit-Architektur baut geschichtete Spurgraphen mit Belegknoten und Ablaufknoten, sodass jede Aussage auf Handlungen, Artefakte und Prüfungen abgebildet wird und fehlende Prüfungsabdeckung sichtbar wird [F72]. Eine Übersichtsarbeit liefert dazu die Begriffsschicht und definiert Ausführungsprovenienz als typisierten Graphen [F73].

Für den Einzelfall gilt dennoch: Diese Klasse setzt instrumentierte Ausführung voraus. Ohne Provenienzerfassung gibt es keine Projektion zu berechnen, und die Aussage ist unprüfbar statt bloß ungeprüft.

3.7 Der eine übertragbare Befund

Eine Arbeit charakterisiert **falschen Erfolg**: Der Agent behauptet Fertigstellung, während der Umgebungszustand Scheitern zeigt. Ihr für diese Karte entscheidender Befund ist methodisch: Leichtgewichtige TF-IDF-Detektoren übertreffen LLM-Richter beim Erkennen dieses Falls, bei weit geringeren Kosten [F74]. Das Signal, das der Detektor lernt, liegt **in der Sprache der Abschlussmeldung** — und das ist der einzige Teil, der auf einen Text ohne Umgebung übertragbar ist.

3.8 Prüfkanal-Tabelle

Mechanismusklasse	Prüfkanal	Erkennt unausgeführtes Verfahren?
Schritt-Belohnungsmodelle (3.1)	gelernte Textplausibilität	nein — bewertet Form, nicht Vollzug
Verifizierer, Mehrfachabtastung (3.2)	Modell gegen Modell, Übereinstimmung	nein — Konsens bestätigt stabile Erfindung
Debatte, Mehragentenbegutachtung (3.3)	Modell gegen Modell, ggf. Quelltext	nur bei belegzugänglichen Debattierenden
Spur-Überwachung (3.4 in Strang A)	Lesen der externalisierten Begründung	bedingt — solange die Spur nicht gegen den Überwacher optimiert wird
Aussageweise Prüfung, Belegpflicht (3.4)	Zerlegung plus Abgleich gegen Evidenzquelle	ja, wenn die Evidenzquelle die Ausführungsspur ist
Werkzeug- und Solver-Erdung (3.5)	Ausführung, deterministisches Orakel	ja im formalisierbaren Teil, per Konstruktion

Spur- und Provenienzprüfung (3.6)

Sprachliche Marker (3.7)

Umgebungs- und Artefaktaufzeichnung

der Text selbst

ja — nur bei instrumentierter Ausführung

der einzige Kanal, der im Einzelfall bleibt

4. Die institutionelle Seite

Die Working Note schlägt eine Prä-Registrierung des Generierungsverfahrens vor. Dieser Abschnitt prüft, welche institutionellen Instrumente dafür schon bestehen und was sie leisten.

4.1 Vorab-Registrierung als Vorbild

Der begriffliche Kern ist die Unterscheidung von Vorhersage und Nachhersage: Vorab-Registrierung wirkt, indem der Analyseplan festgelegt wird, bevor Ergebnisse sichtbar sind [F75]. Registered Reports legen die Begutachtung des Protokolls vor die Ergebnisse, mit grundsätzlicher Annahme und einer zweiten Stufe zur Einhaltungsprüfung [F76] — der einzige in dieser Erhebung gefundene Mechanismus, der eine unabhängige Begutachtung vor die Ergebnisse stellt. Verblindete Vergleichsbewertungen liefern die Wirksamkeitsbelege [F77]. Die stärkste Einzelzahl zum Effekt ergebnisabhängiger Flexibilität: 96 Prozent positive Befunde in Standardarbeiten gegen 44 Prozent in Registered Reports [F78].

Zwei Befunde dämpfen die Übertragung. Erstens: Eine Vorab-Registrierung kann vorliegen und dennoch nichts einschränken; vage Registrierung lässt jedes Ergebnis zu [F79]. Zweitens, und für diese Arbeit zentral: Eine Prüfung, ob veröffentlichte vorab-registrierte Studien ihren eigenen Registrierungen folgten, findet verbreitete nicht offengelegte Abweichung [F80]. Das Instrument lag vor und wurde nicht eingehalten. **Auch ein eingereichtes Verfahren braucht eine eigene Einhaltungsprüfung** — und genau die leistet keine heutige KI-Offenlegungsregel.

4.2 HARKing als Vorlage für M2

Die kanonische Definition — die Darstellung einer Post-hoc-Hypothese als a priori — überträgt sich unmittelbar: Der Ausgabetext ist von einer echten a-priori-Angabe formal nicht unterscheidbar, weil das unterscheidende Merkmal zeitlich und außerhalb des Textes liegt [F81]. Das ist die Begründung dafür, dass M2 ohne äußere Festlegung nur ein Indiz liefern kann. Basisraten fragwürdiger Forschungspraktiken bei Menschen liefern den Vergleichsmaßstab, ohne den die Behauptung, KI verschlechtere die Lage, unquantifiziert bleibt [F82].

Ein Einzelfall gehört in die Darstellung, weil er die Rahmung der Working Note betrifft: Eine Arbeit, deren These war, dass streng vorab festgelegte Verfahren replizierbare Befunde liefern, wurde wegen der Berichterstattung über ihre eigene Vorab-Registrierung zurückgezogen [F83]. Die Lücke zwischen Anspruch und Nachweis ist also keine KI-spezifische Pathologie.

4.3 KI-Offenlegung im Publikationswesen, Stand 06.10.2026

Stelle	KI als Autor	Offenlegung Pflicht	Ort	Gutachterseite	Stand
ICMJE	nein	ja, bei Einreichung	Danksagung	nicht angegeben	ohne Datum [F84]
COPE	nein	ja	Methoden o. ä.	—	v1, 20.08.2024 [F85]
Springer Nature	nein	ja	nicht vorgeschrieben	erlaubt, kein Upload	Ampelschema 2026 [F86]
Elsevier	—	ja, eigener Abschnitt	benannter Erklärungsabschnitt	nur unterstützend	Juni 2026 [F87]
ACL	—	ja, abschnittsgenau	Danksagung	kein KI-Erstentwurf	Rev. 17.08.2026 [F88]
NeurIPS 2025	—	nur wenn methodisch zentral	nicht vorgeschrieben	kein Teilen	2025 [F90]
ICML 2026	—	Autoren wählen Gutachterregel A/B	—	keine Urteilsdelegation	2026 [F91]
ICLR 2026	—	ja, jede Nutzung	per Code of Ethics	Vertraulichkeit	2026 [F92]

Vier Beobachtungen zum Befund:

Erstens. Das einzige gefundene Verlagsverbot, das die Lücke direkt benennt, führt „Hypothesen, Analysen oder Schlussfolgerungen erzeugen und als menschlich hergeleitet darstellen“ unter den nicht erlaubten Handlungen [F86]. Es ist ohne äußeren Nachweis nicht durchsetzbar, weil verbotene und erlaubte Handlung identischen Text erzeugen.

Zweitens. Die NeurIPS-Schwelle „methodisch wichtig, originell oder nicht standardmäßig“ nimmt KI-geschriebene Prosa, die ein strenges Verfahren behauptet, ausdrücklich aus der Offenlegungspflicht [F90]. Das ist die klarste Stelle, an der eine Geltungsbereichsentscheidung die Ziellücke ungedeckt lässt.

Drittens. Die ACL-Regel ist mit abschnittsgenauer Angabe die feinste gefundene und erhöht die Kosten einer plausiblen Falscherklärung, prüft aber nichts [F88]. Die ARR-Pflicht-Checkliste mit Desk-Reject als Sanktion ist das am nächsten liegende Vorbild für eine Verfahrenserklärung zum Einreichungszeitpunkt — sanktioniert wird jedoch Vollständigkeit des Formulars, nicht Wahrheit der Antworten [F89].

Viertens, und das ist der Fund mit dem größten Anschlusswert. Das einzige Instrument der Erhebung, das Hinterlegung von Eingaben und Ausgaben verlangt, sind die Leitlinien der Europäischen Kommission: Forschende sollen, „wo relevant“, Prompts und Ausgaben verfügbar machen, und die Angabe soll Name, Fassung, Datum und Art der Wirkung auf den Forschungsprozess umfassen [F93]. Hinterlegte Eingabeprotokolle würden ein nie ausgeführtes Verfahren im Grundsatz offenlegen. Die Forderung ist bedingt und ohne Sanktion, wirkt also als Empfehlung.

4.4 Technischer Nachweis und Wasserzeichen

Die C2PA-Architektur bindet einen Nachweis nicht an Autorenprosa, sondern an eine kryptografische Signatur: Ein Manifest umfasst Zusicherungen, eine darauf verweisende Behauptung und eine Behauptungssignatur, wobei Vertrauen aus der über X.509 verbürgten Signierererkennung stammt [F94]. Ein Konformitätsprogramm mit öffentlicher Vertrauensliste besteht seit 2026, mit dokumentierter Umstellung von der Übergangsliste zum 01.01.2026 [F95]. Für diese Arbeit gilt die Grenze: Bezeugt würde, *wer signiert hat* und *welches Werkzeug die Datei berührt hat*, nicht ob die beschriebene Analyse stattgefunden hat. Text ist zudem nicht in derselben Weise erfasst wie Bild und Video.

Wasserzeichen sind einsetzbar und im Produktionsmaßstab belegt [F96, F97], beantworten aber eine andere Frage: maschinelle Herkunft, nicht Verfahren. Ein wasserzeichenpositiver Text kann ein wirkliches Verfahren beschreiben, ein wasserzeichennegativer vollständig erfunden sein. Dazu kommen zwei harte Grenzbefunde. Rekursives Umformulieren senkt Erkennungsraten erheblich bei geringem Qualitätsverlust, und Spoofing lässt menschlichen Text wasserzeichenpositiv erscheinen [F98]. Und unter benannten, natürlichen Annahmen ist starkes Wasserzeichen **beweisbar unmöglich**, mit einem generischen Angriff, der nur ein Qualitäts- und ein Störungsrakel braucht [F99].

Das ist der tragendste Negativbefund der technischen Hälfte: Ein Offenlegungsregime kann nicht auf nachträgliche Herkunftserkennung gegründet werden. Das Argument wird damit zu signierten, autorensseitigen Aufzeichnungen des tatsächlich Ausgeführten gezwungen — also genau zu der Richtung, die die Working Note mit der Prä-Registrierung einschlägt.

4.5 Dokumentationsinstrumente

Das stehende W3C-Vokabular für maschinenlesbare Verfahrensaufzeichnungen besteht seit 2013 und ist bis heute das einzige Modell in Recommendation-Status: Provenienz als Aufzeichnung der beteiligten Personen, Einrichtungen, Entitäten und Tätigkeiten [F100]. Ein PROV-serialisiertes Protokoll eines KI-gestützten Arbeitsablaufs würde „was wurde tatsächlich ausgeführt“ abfragbar machen statt erzählt. Es ist allerdings ein reiner Darstellungsstandard: Ein erfundener PROV-Graph ist ohne Signaturschicht so ausdrückbar wie ein wahrer.

Die Gattung der Offenlegungsdokumente stammt aus den Model Cards [F101] und den Datasheets [F102], letztere ausdrücklich mit dem Erhebungsverfahren als Feld. Beide dokumentieren das Werkzeug beziehungsweise die Daten, nicht die Verwendung in einer bestimmten Arbeit. Der konkreteste veröffentlichte Vorschlag für eine standardisierte, **maschinenlesbare** KI-Nutzungsangabe an einem Forschungsergebnis sind die AI Usage Cards, mit fragebogengetriebener Erzeugung [F103]; die Maschinenlesbarkeit ist der operative Fortschritt, weil Angaben dann korpusweit prüfbar statt einzeln lesbar werden. Die feinste fachliche Berichtsleitlinie ist 19 Punkte und 50 Unterpunkte stark und verlangt Prompts, Modellfassungen und Bewertungsaufbau [F104] — der Detailgrad erhöht die Kosten einer Erfindung erheblich, ohne eine Prüfung hinzuzufügen.

4.6 Was am Einzeltext wirklich greift

Zwei Instrumente haben Zähne, weil sie eine innere Konsistenzbedingung prüfen, die bei Erfindung nicht mitgeneriert wird.

Die automatisierte Nachrechnung von p-Werten aus berichteten Teststatistiken und Freiheitsgraden prüft genau diese Bedingung [F105]. Noch schärfer ist der Granularitätstest auf berichtete Mittelwerte: Ein Mittelwert ganzzahliger Daten muss ein Vielfaches von 1/n sein, unmögliche Mittelwerte sind also aus den gedruckten Zahlen allein erkennbar — ohne Rohdaten, ohne Mitwirkung der Autoren, ohne Registrierung [F106]. Das ist die schärfste Einzeltextprüfung der Erhebung, bei engem Geltungsbereich.

Dagegen steht das klarste Beispiel für die untersuchte Lücke selbst: Die automatisierte Prüfung, ob ein Methodenteil die Strenge-Kriterien *berichtet* — Randomisierung, Verblindung, Power-Analyse, Geschlecht als Variable, Reagenzien-Identifizierbarkeit [F107] — bewertet, ob ein Verfahren *beschrieben* ist. Ein erfundener Methodenteil schneidet gut ab, gerade weil er geschrieben wurde, um das Prüfprogramm zufriedenzustellen. Die Korpus-Siebung mit mehreren Detektoren plus menschlicher Beurteilung ist das realistische Modell für Integritätsarbeit [F108], arbeitet aber auf Oberflächenmerkmalen; ein flüssiger Text mit erfundenem Verfahren und ohne gequälte Formulierungen bleibt unmarkiert.

4.7 Größenordnung

Über 10.000 Rückzüge im Jahr 2023 belegen, dass die Integritätsprüfung längst jenseits der Kapazität der Einzelfallprüfung arbeitet [F109]; die Ursache ist dort überwiegend industrielle Produktion, nicht KI-Text, und darf nicht vermischt werden. Die aktuellen Betriebszahlen

gemeinsamer Vorab-Siebung: über 35 Verlage prüfen monatlich mehr als 125.000 Arbeiten und fangen etwa 1.000 verdächtige Einreichungen ab [F110] — eine Abfangquote um 0,8 Prozent. Die stärkste Korpusmessung undeckelter LLM-Nutzung findet über 15 Millionen Abstracts einen abrupten Anstieg bestimmter Stilwörter, der auf mindestens 13,5 Prozent LLM-bearbeitete Abstracts im Jahr 2024 und bis 40 Prozent in einzelnen Teilgebieten schließen lässt [F111]. Das Verfahren ist rein verteilungsbezogen, weist also keiner Einzelarbeit eine Wahrscheinlichkeit zu — und es misst Schreibhilfe, nicht Verfahrensvollzug.

5. Agentische Forschungssysteme

Dieser Strang liefert das Fallmaterial. Der Anlassfall der Working Note ist ein handgeführter Einzellauf; die hier genannten Systeme erheben denselben Anspruch mit gebauter Architektur. Die Frage lautet in beiden Fällen gleich: Lässt der *Ausgabetext* erkennen, dass das behauptete innere Verfahren stattgefunden hat?

5.1 Durchgängige Systeme

Ein System beansprucht die geschlossene Schleife von Ideenerzeugung über Ausführung bis zum Manuskript, mit einem automatisierten Gutachter; die Zeitschriftenfassung nennt ein erzeugtes Manuskript, das die erste Begutachtungsrunde eines Workshops bestand [F113]. Die Nachfolgefassung ersetzt die lineare Kette durch agentische Baumsuche unter einem Versuchsleiter [F114]. **Genau dieser Teil hinterlässt im Manuskript keine Spur:** Nichts in einer erzeugten Arbeit unterscheidet ein Ergebnis aus breiter Suche mit vielen verworfenen Zweigen von einem aus einem einzigen Durchgang. Das ist M1 auf Systemebene.

Ein dreistufiger Assistent nimmt menschliche Rückmeldung auf jeder Stufe an und berichtet 84 Prozent Kostensenkung [F115]; die autonome und die unterstützte Bedingung sind getrennt zu zitieren, weil die Stufengrenzen architektonisch sind und im Endbericht nicht erkennbar bleibt, welche Stufe welche Aussage erzeugte. Ein weiteres System paart ein erzeugendes Modell mit einem simulierten, per Rückmeldung trainierten Gutachter; die berichteten Gewinne sind **simulierte Gutachtenwerte**, Erzeuger und Bewerter kommen aus derselben Trainingskette [F116]. Datensätze und Gewichte sind veröffentlicht, die Artefakte also einsehbar.

Zum Umgang mit diesen Ansprüchen gehört ein Einwand aus demselben Umfeld: Eine Positionsarbeit analysiert 28 Arbeiten aus fünf solchen Systemen und verortet den Engpass nicht in der Ideenerzeugung, sondern in der Fähigkeit, die zur Stützung nötigen Prüfverfahren auszuführen [F134]. Die Autorenüberschneidung mit einem der untersuchten Systeme ist zu nennen — es ist Kritik von innen, keine unabhängige Prüfung.

5.2 Wo der Anspruch physisch endet

Die Laborsysteme sind der aufschlussreiche Gegenfall, weil ihr behauptetes Verfahren in einer Messung endet. Ein System entwirft, plant und führt Versuche über Websuche, Codeausführung und Flüssigkeitshandhabung aus [F117]. Ein anderes verbindet ein Modell mit 18 Fachwerkzeugen und lässt **von Fachleuten statt von einem Modell bewerten** — mit einem Nebenbefund, der für diese Karte zentral ist: Das Modell als Bewerter „kann nicht zwischen klar falschen Vervollständigungen und der Leistung des Systems unterscheiden“, und modellgestützte Bewertung bevorzugt Flüssigkeit über Richtigkeit [F118]. Das ist ein direkter Beleg dafür, dass die automatisierte Gutachterkomponente der Systeme in 5.1 etwas anderes als Richtigkeit misst.

Der lehrreichste Fall ist eine Robotikplattform, die 41 neue Verbindungen anvisierte und 36 von 57 erfolgreich synthetisierte [F119] — und die dazugehörige externe Nachprüfung, die argumentiert, dass die automatisierte Auswertung der Pulverdiffraktometrie noch nicht zuverlässig ist und Unordnung im Vorhersageschritt übergangen wird, sodass mehrere der berichteten neuen Verbindungen auf der Charakterisierungsstufe bestritten sind [F120]. Ein Autonomieanspruch übersteht hier seine eigene innere Bewertung und scheitert an fachlicher Nachprüfung von außen. Das ist das klarste Fallbeispiel dafür, was die berichteten Gewinne solcher Systeme tatsächlich messen.

Positiv hervorzuheben ist ein System, dessen behauptetes inneres Verfahren das **Besprechungsprotokoll selbst** ist: Ein Modell als Projektleitung beruft Fachagenten zu iterativen Forschungsbesprechungen; 92 Kandidaten wurden entworfen und experimentell geprüft, zwei mit verbesserter Bindung [F121]. Die Besprechungsaufzeichnungen wurden mitveröffentlicht, die Kritik- und Revisionsrunden sind damit einsehbar statt bloß behauptet. Es ist zudem als Zusammenarbeit gerahmt, nicht als Autonomie.

5.3 Ideenerzeugung und Neuheitsbewertung

Ein mehragentiges System führt ein Turnier mit Elo-artiger Rangordnung über Hypothesen und ist in drei Anwendungsfeldern mit Laborbestätigungen validiert [F122]. Das Turnier ist der zentrale Anspruch und wird im Aggregat berichtet: Wer eine einzelne Ausgabe-Hypothese sieht, erkennt nicht, was sie geschlagen hat, wie viele Runden sie überstand und was der Kritikagent einwandte. Die Laborbestätigungen tragen den Befund, nicht die Elo-Kurven. Ein weiteres System koppelt ontologische Wissensgraphen mit Agenten und behauptet im Abstract Fähigkeiten, „die traditionelle menschengeführte Forschungsmethoden übersteigen“, ohne Vergleich und Maßzahl zu benennen [F123] — ein unverankerter Überlegenheitsanspruch, der hier als solcher zu markieren und nicht zu wiederholen ist. Ein drittes simuliert ein Forschungsteam und untersucht, welche Zusammenarbeitsmechanismen die Neuheit erhöhen [F124]; der Vergleich ist systemintern, misst also die Wirkung der Architektur auf die Maßzahl, nicht auf Forschungsergebnisse.

Die früheste Arbeit, die Neuheit zum ausdrücklichen Optimierungsziel macht, ruft Anregungen aus Vorarbeiten ab und vergleicht Kandidaten

iterativ gegen bestehende Arbeiten bis zu einer Neuheitsschwelle — und berichtet freimütig, dass ein Modell allein technisch flache Ideen geringer Neuheit erzeugt, was das Verfahren nur teilweise behebt [F125]. Diese Offenheit ist in dem Teilfeld die Ausnahme.

Ein Benchmark mit Einwort-Eingaben über 1.180 Begriffe und 22 Gebiete liefert dazu einen Befund, der die Modellwahl betrifft: Werte der Ideenerzeugung werden von allgemeinen Fähigkeitsmaßen schlecht vorhergesagt, mittelgroße Modelle erreichen Spitzenmodelle [F128]. Die Annahme, ein stärkeres Grundmodell erzeuge bessere Forschung, ist damit nicht gedeckt. Die Bewertung ist allerdings selbst modellbasiert, sodass der Befund aus [F127] auch für diesen Benchmark gilt.

Der für die Working Note wichtigste Einzelfund dieses Strangs ist ein Negativbefund von Oktober 2026: Kleine Änderungen der Formulierung verschieben die Genauigkeit einer Neuheitsbeurteilung um über 50 Prozentpunkte; Abrufergänzung und längere Begründung helfen kaum; eigens gebaute Neuheitsbewerter unterschreiten einfache Grundlinien [F127]. Damit ist jeder berichtete Neuheitsgewinn in 5.1 und 5.3 in Frage gestellt, einschließlich der blinden Expertenbewertung, nach der KI-Ideen als neuartiger, aber geringfügig weniger machbar beurteilt wurden [F129]. Das Gegenbeispiel für prüfbare Gestaltung ist ein Abruf-dann-Neuordnungs-Verfahren, dessen Ausgabe ein Urteil **plus die Vergleichsarbeiten, auf denen es ruht** ist; dort kann geprüft werden, ob die zitierte Vorarbeit sagt, was das Urteil annimmt [F126].

5.4 Benchmarks und ihre Lesarten

Vier Benchmarks für agentische Forschungsfähigkeit sind geprüft. Zwei Lesefehler sind dabei zu vermeiden.

Erstens: Der weit zitierte Vorsprung agentischer Systeme gegenüber Fachleuten gilt bei einem Zweistundenbudget; bei acht Stunden ziehen Menschen vorbei, bei 32 Stunden verdoppeln sie [F132]. Die Zeithorizontabhängigkeit verschwindet, wenn die Vorsprungszahl allein zitiert wird. Zweitens: Ein Benchmark aus öffentlichen Wettbewerbslösungen misst teilweise Erinnerung an veröffentlichte Berichte statt Ingenieurleistung, was die zugehörige Kontaminationsanalyse ausdrücklich behandelt [F130]. Eine rubrikbasierte Zerlegung von 20 Arbeiten in 8.316 einzeln bewertbare Teilaufgaben ist die stärkste Prüfbarkeitseigenschaft dieses Abschnitts, weil erkennbar wird, welche Teilaussage scheiterte, statt nur eine Zahl zu liefern — die Bewertung ist allerdings selbst modellbasiert, was die Arbeit offen sagt [F131]. Ein vierter Benchmark setzt rechnerische Reproduzierbarkeit als Vorbedingung für Systeme zur Erzeugung neuer Forschung, also als Fähigkeitsuntergrenze, die die Systeme aus 5.1 voraussetzen, ohne sie nachzuweisen [F133].

5.5 Was die Veranstalter 2026 tun

Die Reaktion der Veranstalter ist der institutionelle Teil, der für die Veröffentlichung der Working Note unmittelbar gilt. Eine Konferenz erklärt, dass Arbeiten mit umfassender, nicht offengelegter LLM-Nutzung ohne Begutachtung abgelehnt werden, dass Autoren für Ausgaben verantwortlich bleiben und dass **erfundene Quellenangaben einen Verstoß gegen den Ethikkodex darstellen**, der zur Ablehnung berechtigt [F136]. Eine andere berichtet 178 Einreichungen (18,4 Prozent) ohne Einspruchsmöglichkeit abgelehnt und 123 (12,7 Prozent) zur Nachweisführung menschlicher Autorschaft aufgefordert, bei einer Detektorquote von 28,2 Prozent in der Standardeinstellung, die nach Fensteranpassung auf 12,7 Prozent fällt [F137]. Der Abstand zwischen beiden Einstellungen ist selbst das lehrreiche Datum: Die berichtete Ablehnungsquote ist eine Funktion einer Einstellungsentscheidung, und der Beitrag sagt das.

Die größte dokumentierte Gegenbewegung ist eine Begutachtungshilfe, die auf alle 22.977 Einreichungen einer Konferenz in unter einem Tag angewendet wurde, mit günstigen Bewertungen in Befragungen von Autoren und Programmkomitee [F135] — verfasst von den Organisatoren des Versuchs, die günstigen Befragungsergebnisse sind also Eigenangaben. Eine Veranstaltung verlangt KI-Autorschaft, statt sie bloß zu erlauben, und veröffentlicht die Eingaben und Gutachten der KI-Gutachter als Gemeingut [F138] — die einzige gefundene Veranstaltung, die die automatisierte Begutachtungsspur als veröffentlichbares Artefakt behandelt. Das ist die institutionelle Form dessen, was diese Arbeit verlangt.

5.6 Der Anlassfall im Fachgebiet

Der Anlassfall behandelt zellfreie Biokatalyse und Enzymkaskaden. Die Erhebung fand zwei geprüfte Plattformarbeiten in diesem Gebiet, beide maschinell lernend geführt, davon eine mit natürlichsprachlicher Schnittstelle [F139, F140]. **Es wurde kein geprüfter Fall eines durchgängigen KI-Forschungssystems gefunden, das eine Enzymkaskade als Anwendungsbeispiel verwendet.** Der Anlassfall ist also kein Nachbau eines bestehenden Systems, sondern ein handgeführter Einzellauf, der dieselbe Rhetorik erzeugt. Das stützt die Entscheidung der Working Note, ihn als Fallmaterial und nicht als Gegenstand zu behandeln.

6. Verortung: wo die Lücke ist

Nach dieser Erhebung lässt sich die Lage der Working Note präzise angeben.

Was nicht neu ist. Dass behauptetes und tatsächliches Verfahren auseinanderfallen, ist in neun Teilliteraturen beschrieben (Abschnitt 2). Dass Selbstkorrektur ohne Gegenüber nicht zuverlässig wirkt, ist belegt [F25, F26, F27]. Dass automatisierte Begutachtung Flüssigkeit statt Richtigkeit belohnt, ist belegt [F32, F118]. Dass Neuheitsurteile von Modellen instabil sind, ist belegt [F127]. Die Working Note darf keinen dieser Punkte als eigenen Befund führen.

Was tatsächlich offen ist. Fünf Leerstellen sind in allen vier Strängen unabhängig voneinander als solche benannt worden:

1. **Kein Prüfinstrument auf Dokumentebene.** Es gibt Prüfung auf Zitatebene [F37], auf Korpusebene [F111, F42] und auf Oberflächenqualität [F39]. Es gibt keine veröffentlichte Rubrik, die die Verfahrensaussagen einer Arbeit entnimmt und auf reproduzierbare Stützung prüft — das Dokumentgegenstück zur Eingriffsmethodik aus [F2].
2. **Kein Werk, das Textmerkmale unausgeführter Verfahrensbeschreibungen bei einem einzelnen, nicht instrumentierten Lauf klassifiziert.** Das Nächstliegende nutzt weiterhin das Umgebungsurteil als Labelquelle [F74].
3. **Kein Versuch, der sichtbare Scheinrigorosität als unabhängige Variable im Übervertrauen prüft.** Die Übervertrauensliteratur variiert die *Richtigkeit* des Rats [F43, F44, F45]; die Spurliteratur misst die Modellseite [F11, F12]. Niemand hat gefragt, ob eine lange, beratungsförmige Spur die Übernahme einer falschen Schlussfolgerung gegenüber einer bloßen Antwort erhöht. Das ist die empirische Mitte des Themas, und sie ist leer.
4. **Keine Verfahrens-Attestierung für KI-gestützte Textarbeit.** Eine erweiterte Suche fand keinen wissenschaftlichen oder normativen Gebrauch des Begriffs in dieser Bedeutung. Die nächstliegenden echten Instrumente sind [F91], [F93], [F103], [F104] und [F112]. **Ein Gegenstück zur Vorab-Registrierung existiert für KI-gestütztes Schreiben nicht.**
5. **Keine Messung, wie gut menschliche Begutachtende erfundene Quellen erkennen.** Geprüft werden maschinelle Prüfer [F37]; die Trefferquote der Instanz, die entscheidet, ob die Raten aus [F33] bis [F36] in veröffentlichte Fehler übergehen, ist unbekannt.

Wo die Working Note sitzt. Sie bearbeitet Leerstelle 2 mit fünf Markern und einem Werkzeug und schlägt für Leerstelle 4 einen Mechanismus vor. Sie bearbeitet Leerstelle 1 nicht, sondern bereitet sie vor. Leerstellen 3 und 5 sind Anschlussvorhaben, nicht Teil der Arbeit.

Eine Korrektur der eigenen Ansprüche. Der Satz der Working Note, Iteratio Simulata sei der einzige allein am Text prüfbare der drei Nachbarbegriffe, ist nach dieser Erhebung zu halten, aber zu präzisieren: Er gilt gegenüber *unfaithful chain-of-thought* und *potemkin understanding*, nicht gegenüber den 2026er Arbeiten zur performativen Spur, die mit Sondierung und Zeitmessung arbeiten und deshalb Modellzugang brauchen. Dass sie Modellzugang brauchen, ist der Grund, warum die Marker daneben bestehen bleiben; der Alleinstellungsanspruch ist entsprechend zu entschärfen (siehe 7.1).

7. Zehn Anschlussvorhaben

Aus der Erhebung folgen zehn Vorhaben. Für jedes sind der Grund, der Zuschnitt und die Voraussetzung angegeben. Fünf davon sind ohne weitere Mittel ausführbar.

7.1 Nachtrag zur Working Note (klein, sofort)

Vier Punkte in die nächste Fassung: der 2026er Befundblock zur performativen Spur als empirische Stütze des Kernbegriffs [F11, F12, F13]; die Gegenposition zur Verbalisierungsfrage [F7]; die Entschärfung des Alleinstellungsanspruchs nach 6; der Hinweis, dass M1 auf Systemebene genau die Baumsuche trifft, die kein erzeugtes Manuskript ausweist [F114]. **Voraussetzung:** keine.

7.2 Dokumentrubrik „Verfahrensnachweis“ (mittel, ausführbar)

Leerstelle 1. Eine Rubrik, die aus einer Arbeit die Verfahrensaussagen entnimmt und je Aussage nach dem Stützungstyp klassifiziert: ausgeführt und belegt, beschrieben und belegbar, beschrieben und unprüfbar, widersprochen. Das Vorbild ist die rubrikbasierte Teilaufgabenzerlegung [F131], der Prüfmechanismus die Zitat-Stützungsprüfung [F63, F61]. Ergebnis ist ein Zählblatt, nicht eine Punktzahl. **Voraussetzung:** 20 bis 30 Arbeiten als Erprobungssatz; keine Rechenlast.

7.3 Sprachlicher Einzeltextdetektor (mittel, ausführbar)

Leerstelle 2, mit dem einzigen übertragbaren Vorbild [F74]. Der dortige Befund — einfache TF-IDF-Merkmale schlagen Modellrichter beim Erkennen falscher Fertigmeldungen — ist auf die Marker M1 und M2 unmittelbar anwendbar, weil beide über Spurlosigkeit arbeiten. Konkret: die bestehende Vokabularprüfung des Iterationsprüfers gegen ein gewichtetes Merkmalsmodell stellen, das auf Abschlussmeldungssprache trainiert ist. **Voraussetzung:** der Prüfsatz aus 7.4.

7.4 Prüfkorpus mit unabhängiger Vorklassifikation (mittel, blockierend)

Der in der Working Note bereits als offen geführte Punkt, hier bestätigt als Voraussetzung für 7.2, 7.3 und 7.6. Mindestens 20 Texte, menschlich vorklassifiziert, ohne Kenntnis der Markerdefinitionen; Negativfälle zuerst, nach der stehenden Regel aus dem Implementierungsbefund. Der Korpus sollte je Marker sorgfältige *und* performative Texte enthalten, weil der erste Gegentest des Werkzeugs an sorgfältigen Texten scheiterte. **Voraussetzung:** zwei bis drei unabhängig urteilende Personen.

7.5 Übervertrauensversuch (groß, extern)

Leerstelle 3 und der wissenschaftlich wertvollste Punkt dieser Karte. Aufbau: Fachleute erhalten dieselbe falsche Schlussfolgerung einmal als

bloße Antwort, einmal mit langer, beratungsförmiger Spur; abhängige Variable ist die Übernahme. Der Aufbau folgt den Automation-Bias-Studien [F43, F44, F45], variiert aber die *Darstellung des Verfahrens* statt die Richtigkeit des Rats. Der Befund, dass im KI-gestützten Arbeiten gewonnene Zuversicht mehr Verlass voraussagt [F45], liefert die Vorhersage. **Voraussetzung:** Ethikvotum, Fachpersonenstichprobe, Kooperationspartner. Nicht allein ausführbar.

7.6 Marker M3 und M3b gegen offene Metadatendienste (klein, ausführbar)

Die in der Working Note als Phase 1c geführte Arbeit, hier mit einer zusätzlichen Vorsichtsregel: Das Prüfprogramm darf nicht selbst in die Falsch-Positiv-Falle laufen, die für Arbeiten jenseits des Trainingsendes gemessen wurde [F37]. Konsequenz für die Umsetzung: Registerabgleich (DOI-Auflösung, OpenAlex) statt Modellurteil, und ein ausdrücklicher Status „nach Datenstand nicht auffindbar“ getrennt von „nicht existent“. **Voraussetzung:** Lizenzprüfung der angebundenen Dienste; keine eigene Datenhaltung.

7.7 Verfahrenserklärung als einreichbares Artefakt (mittel, ausführbar)

Leerstelle 4, der Beitrag mit der größten institutionellen Anschlussfähigkeit. Die Bausteine liegen vollständig vor: die Pflichtcheckliste mit Desk-Reject als Durchsetzungsform [F89]; die Ex-ante-Verfahrensbindung, bei der der Autor festlegt und der Gutachter gebunden ist [F91]; die Hinterlegung von Eingaben und Ausgaben [F93]; die maschinenlesbare Nutzungsangabe als Format [F103]; das stehende Provenienzvokabular als Serialisierung [F100]; die Signaturschicht, ohne die ein erfundener Graph so ausdrückbar ist wie ein wahrer [F94]. Zu bauen ist die Zusammenstellung, nicht ein Baustein. **Voraussetzung:** keine für den Entwurf; Wirkung erst mit einer aufnehmenden Stelle.

7.8 Gutachtertrefferquote bei erfundenen Quellen (mittel, extern)

Leerstelle 5. Ein Satz Manuskripte mit kontrolliert eingebauten Quellenfehlern dreier Schweregrade — nicht existent, existent mit falscher Autorenschaft, existent mit falschen Feldern — und die Messung, welche Stufe erkannt wird. Die Fehlerstufen sind aus dem Befundmuster gebildet, dass echte Quellen mit falschen Feldern der schwerere Fall sind [F33]. **Voraussetzung:** Kooperation mit einer Redaktion; ethisch nur mit Vorabinformation durchführbar.

7.9 Replikation des Potemkin-Verfahrens auf Deutsch (klein, ausführbar)

Eine in Strang A ausdrücklich benannte Lücke: keine Replikation außerhalb des Englischen und keine auf fachpraktische statt akademische Begriffe. Für Wissensarbeit im deutschsprachigen Fach- und Regelungsumfeld ist das unmittelbar einschlägig, und das Verfahren ist veröffentlicht [F14]. **Voraussetzung:** Begriffssatz aus einem Fachgebiet.

7.10 Verfahrenstreue im eigenen Arbeitsablauf (klein, sofort)

Der Befund, dass 27 bis 78 Prozent ergebnisrichtiger Durchläufe verdeckte Verfahrensverstöße enthalten [F70], ist auf die eigene Arbeitsweise anzuwenden. Konkret: Die vier Prä-Registrierungsschritte der Working Note nicht nur empfehlen, sondern am eigenen Bündel ausweisen — mit Hash der Kandidatenliste und der Kriterien vor der Auswahl, als mitveröffentlichte Datei. Dies ist die Selbstanwendung, die 7.7 glaubwürdig macht. **Voraussetzung:** keine.

8. Nicht gefunden, nicht geprüft, zu beachten

8.1 Echte Lücken

Über die fünf in Abschnitt 6 genannten hinaus wurden benannt: keine Fabrikationsraten für Geistes- und Sozialwissenschaften; keine Sycophancy-Studie über längere Arbeitssitzungen statt Einzelnachstoß; nichts zu KI-Slop in Förder- und Normungsgutachten; kein geprüfter systematischer Überblick über den Strang der agentischen Forschungssysteme; keine abgeschlossene Studie, in der die erzeugten Ideen aus [F129] tatsächlich ausgeführt wurden — die zentrale offene Frage der gesamten Neuheitsliteratur.

8.2 Nicht abrufbar, als ungeprüft geführt

Elf Einträge konnten nicht an der Primärseite bestätigt werden und sind in der Quellenliste mit [T] gekennzeichnet. Betroffen sind unter anderem: die redaktionellen Richtlinien eines Verlags, dessen Seiten automatisierte Abrufe mit HTTP 403 abweisen; der Wortlaut von Artikel 50 der Verordnung (EU) 2024/1689, dessen HTML- und PDF-Fassungen vor den Artikeln abrechnen; die Zeitschriftenfassung einer Sycophancy-Arbeit; die veröffentlichte Endfassung einer Berichtsleitlinie, von der nur das Protokoll geprüft ist; die Autorenlistenform einer weiteren Leitlinie, die aus einem Repositoriumseintrag stammt; eine Verlagsseite einer Nachprüfung, deren Angaben aus einem Hochschulverzeichnis gelesen wurden; zwei thematisch zentrale Arbeiten zu Überbehauptung, deren Seiten bei mehrfachem Abruf leer blieben; eine Grundlagenarbeit zur formalen Beweis-Erdung, deren Abruf mit HTTP 429 abgewiesen wurde; eine Übersichtsarbeit zu selbstfahrenden Laboren, deren Band- und Seitenangaben nicht gedruckt vorlagen; ein Preprint zur geschlossenen Optimierung von Enzymkaskaden, das thematisch genau treffen würde und nicht abgerufen wurde.

Bei einem Abruf gab das Werkzeug plausibel aussehende Zitate aus einem Rechtstext zurück und kennzeichnete sie selbst als aus der Struktur

rekonstruiert. Sie wurden verworfen. Dieser Vorgang ist der Grund, warum [T] als Status geführt und nicht durch eine Auslassung ersetzt wird.

8.3 Industrieangaben ausgeschlossen

Angaben von Detektorherstellern zu Anteilen KI-erzeugter Gutachten und zu Halluzinationszählungen sind in Suchergebnissen aufgetreten. Sie sind Herstellerveröffentlichungen mit nicht offengelegtem Verfahren und daher nicht in die Quellenliste aufgenommen. Zitierbar wären sie nur als Herstellerangabe. Dasselbe gilt für die Angabe einer gemeinsamen Siebungsinfrastruktur, die ihre Detektoren nicht offenlegt [F110]; die Zahlen stehen, ihre Aussagekraft ist aus der Primärquelle nicht beurteilbar.

8.4 Fassungsabweichungen

Vier Arbeiten haben zwischen Preprint und Veröffentlichung Titel oder Autorenliste geändert: ein durchgängiges Forschungssystem (6 gegen 8 Autoren, abweichende Reihenfolge, anderer Titel); ein mehragentiges System (Titel in v1 und v2 verschieden); ein Neuheitsbenchmark (umbenannt, der verbreitete Name ist der Benchmarkname, nicht der Titel); ein Ideenerzeugungssystem (verbreiteter Titel weicht vom gedruckten ab). Wer diese aus einer Sekundärquelle von 2024 oder 2025 übernimmt, zitiert veraltet.

8.5 Aufwertungen, die unterlassen wurden

Bei mehreren vielzitierten Arbeiten nennt die Abstractseite keine Veranstaltung. Sie sind als Preprint geführt. Eine spätere Bearbeitung darf sie nicht stillschweigend aufwerten; bei Bedarf ist die Veranstaltungsangabe je Eintrag an der Proceedingsseite zu prüfen.

9. Selbstanwendung

Die Working Note verlangt, dass eine Arbeit über vorgetäuschte Sorgfalt ihre eigenen Fehler ausweist. Für diese Erhebung ergeben sich drei Befunde.

9.1 Zwei aus dem Gedächtnis gehaltene Kennungen waren falsch

Im Strang zu Nachweis und Offenlegung waren zwei Kennungen vorab bekannt und wurden dennoch geprüft. Beide waren falsch. Eine DOI gab einen 404 zurück; die richtige unterscheidet sich in der letzten Stelle, und **die Arbeit ist zurückgezogen** — ein Umstand, der bei Zitation aus dem Gedächtnis unentdeckt geblieben wäre. Eine arXiv-Kennung führte zu einer astrophysikalischen Arbeit über Teilchenbeschleunigung; die richtige Kennung ist eine andere.

Dies ist das Muster aus [F33] und [F36] am eigenen Vorgang: Die Felder, die präzise aussehen und selten geprüft werden, sind die falschen. Es ist zugleich der stärkste verfügbare Einzelbeleg für die These der Working Note — die Fehler wurden ausschließlich dadurch gefunden, dass jeder Eintrag abgerufen wurde, nicht durch Plausibilitätsprüfung.

9.2 Die Erhebung musste ihre eigene Überlegenheitsbehauptung zurücknehmen

Der Alleinstellungsanspruch der Working Note — Iteratio Simulata sei der einzige allein am Text prüfbare Nachbarbegriff — hält gegenüber den beiden 2025 verfügbaren Nachbarbegriffen und hält nicht gegenüber den 2026er Arbeiten in gleicher Breite. Der Satz stand in Hauptwerk, Pamphlet, Erklärung und auf beiden Flyern. Er wird in der nächsten Fassung präzisiert, nicht gestrichen, und der Vorgang steht hier (Abschnitt 6, letzter Absatz; Anschlussvorhaben 7.1).

Dies ist ein M3-Fall nach eigener Definition: Der Stand des Feldes war weiter, als die Arbeit ihn darstellte.

9.3 Die Prüfbarkeitsurteile dieser Karte sind Schlüsse, keine Befunde

Für den Strang zu agentischen Forschungssystemen wurden Abstractseiten und Metadaten gelesen, nicht Volltexte. Jede Aussage darüber, ob der Ausgabertext eines Systems Runden, Auswahl und Kritik erkennen lässt, ist eine Folgerung aus der beschriebenen Architektur und **kein geprüfter Befund**. Die Prüfung dieser Folgerungen verlangt, die veröffentlichten Artefakte der vier Systeme mit offenen Spuren zu öffnen und zu lesen. Das ist nicht geschehen. Es ist die Aussage, an der diese Karte am angreifbarsten ist, und die Stelle, an der sie selbst M1 ausgesetzt ist: Sie behauptet ein Prüfurteil, dessen Prüfschritt sie nicht ausgeführt hat.

Daraus folgt eine Ergänzung zum Anschlussvorhaben 7.2: Der erste Erprobungssatz der Dokumentrublik sind genau diese vier Systeme, weil sie die einzigen Fälle sind, in denen behauptetes Verfahren und veröffentlichte Spur miteinander verglichen werden können.

10. Quellen

Statusangaben: [Z] begutachtete Zeitschrift · [K] Konferenzbeitrag · [W] Workshopbeitrag · [P] Preprint, keine Veranstaltung auf der Abstractseite · [D] Richtlinie, Norm, Positions- oder Arbeitspapier · [N] Presse- oder Praxisbeitrag · [T] nicht an der Primärseite bestätigt, vor

einer Journaleinreichung einzeln zu prüfen. Jede Angabe wurde am 06.10.2026 an der genannten Adresse abgerufen und von dort abgelesen. Keine Angabe stammt aus dem Gedächtnis.

Strang A — Fehlermodi: Prozess statt Faktum

[F1] Miles Turpin, Julian Michael, Ethan Perez, Samuel R. Bowman. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388, 2023; Kommentarfeld: NeurIPS 2023. <https://arxiv.org/abs/2305.04388> [P] [F2] Tamera Lanham, Anna Chen, Ansh Radhakrishnan u. a. (30). *Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv:2307.13702, 2023. <https://arxiv.org/abs/2307.13702> [P] [F3] Yanda Chen, Joe Benton, Ansh Radhakrishnan u. a. (15). *Reasoning Models Don't Always Say What They Think*. arXiv:2505.05410, 2025. <https://arxiv.org/abs/2505.05410> [P] [F4] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooan Rajamanoharan, Neel Nanda, Arthur Conmy. *Chain-of-Thought Reasoning In The Wild Is Not Always Faithful*. PMLR 306 (ICML 2026). <https://proceedings.mlr.press/v306/arcuschin26b.html> [K] [F5] Jacob Pfau, William Merrill, Samuel R. Bowman. *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models*. arXiv:2404.15758, 2024. <https://arxiv.org/abs/2404.15758> [P] [F6] Fazl Barez, Tung-Yu Wu, Iván Arcuschin u. a. (16). *Chain-of-Thought Is Not Explainability*. Working Paper, Oxford Martin AI Governance Initiative, 2025. <https://www.oxfordmartin.ox.ac.uk/publications/chain-of-thought-is-not-explainability> [D] [F7] Kerem Zaman, Shashank Srivastava. *Is Chain-of-Thought Really Not Explainability? Chain-of-Thought Can Be Faithful without Hint Verbalization*. Proc. 64th ACL (Vol. 1: Long Papers), 2026, S. 48008-48030. <https://aclanthology.org/2026.acl-long.2217/> [K] [F8] Tomek Korbak, Mikita Balesni, Elizabeth Barnes u. a. (41). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. arXiv:2507.11473, 2025. <https://arxiv.org/abs/2507.11473> [P] [F9] Bowen Baker, Joost Huizinga, Leo Gao u. a. (9). *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation*. arXiv:2503.11926, 2025. <https://arxiv.org/abs/2503.11926> [P] [F10] Scott Emmons, Erik Jenner, David K. Elson u. a.. *When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors*. arXiv:2507.05246, 2025. <https://arxiv.org/abs/2507.05246> [P] [F11] Wenkai Li, Fan Yang, Ananya Hazarika, Shaunak A. Mehta, Koichi Onoue. *When Reasoning Traces Become Performative: Step-Level Evidence that Chain-of-Thought Is an Imperfect Oversight Channel*. arXiv:2605.11746, 2026. <https://arxiv.org/abs/2605.11746> [P] [F12] Siddharth Boppana, Annabel Ma, Max Loeffler u. a. (8). *Reasoning Theater: Disentangling Model Beliefs from Chain-of-Thought*. arXiv:2603.05488, 2026. <https://arxiv.org/abs/2603.05488> [P] [F13] Vatsal Baherwani, Tom Goldstein, Ashwinee Panda. *Not All LLM Reasoning is Visible in the Chain-of-Thought*. arXiv:2607.22925, 2026. <https://arxiv.org/abs/2607.22925> [P] [F14] Marina Mancoridis, Bec Weeks, Keyon Vafa, Sendhil Mullainathan. *Potemkin Understanding in Large Language Models*. PMLR 267 (ICML 2025). <https://proceedings.mlr.press/v267/mancoridis25a.html> [K] [F15] Zhaofeng Wu, Linlu Qiu, Alexis Ross u. a. (9). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*. arXiv:2307.02477, 2023; NAACL 2024. <https://arxiv.org/abs/2307.02477> [K] [F16] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, Mehrdad Farajtabar. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv:2410.05229, 2024; Kommentarfeld: ICLR camera ready. <https://arxiv.org/abs/2410.05229> [P] [F17] Parshin Shojaei, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar. *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. arXiv:2506.06941, 2025; Kommentarfeld: NeurIPS 2025 camera-ready. <https://arxiv.org/abs/2506.06941> [P] [F18] Lisa Messeri, M. J. Crockett. *Artificial intelligence and illusions of understanding in scientific research*. Nature 627, 49-58, 2024 (Perspective), DOI 10.1038/s41586-024-07146-0. <https://www.nature.com/articles/s41586-024-07146-0> [Z] [F19] Ethan Perez, Sam Ringer, Kamile Lukošiušė u. a. (60). *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251, 2022. <https://arxiv.org/abs/2212.09251> [P] [F20] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, Quoc V. Le. *Simple synthetic data reduces sycophancy in large language models*. arXiv:2308.03958, 2023. <https://arxiv.org/abs/2308.03958> [P] [F21] Mrinank Sharma, Meg Tong, Tomasz Korbak u. a. (19). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548, 2023. <https://arxiv.org/abs/2310.13548> [P] [F22] Aaron Fanous, Jacob Goldberg, Ank A. Agarwal u. a. (7). *SycEval: Evaluating LLM Sycophancy*. arXiv:2502.08177, 2025; Kommentarfeld: AIES 2025. <https://arxiv.org/abs/2502.08177> [P] [F23] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, Dan Jurafsky. *ELEPHANT: Measuring and understanding social sycophancy in LLMs*. arXiv:2505.13995, 2025. Die Abstractseite nennt eine Zeitschriftenfassung; diese war nicht abrufbar und ist nicht bestätigt. <https://arxiv.org/abs/2505.13995> [T] [F24] Aman Madaan, Niket Tandon, Prakhar Gupta u. a. (16). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv:2303.17651, 2023. <https://arxiv.org/abs/2303.17651> [P] [F25] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, Denny Zhou. *Large Language Models Cannot Self-Correct Reasoning Yet*. arXiv:2310.01798, 2023; Kommentarfeld: ICLR 2024. <https://arxiv.org/abs/2310.01798> [P] [F26] Kaya Stechly, Matthew Marquez, Subbarao Kambhampati. *GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems*. arXiv:2310.12397, 2023. <https://arxiv.org/abs/2310.12397> [P] [F27] Karthik Valmeekam, Matthew Marquez, Subbarao Kambhampati. *Can Large Language Models Really Improve by Self-critiquing Their Own Plans?*. arXiv:2310.08118, 2023. <https://arxiv.org/abs/2310.08118> [P] [F28] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng u. a. (13). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. NeurIPS 2023 Datasets and Benchmarks Track. <https://arxiv.org/abs/2306.05685> [K] [F29] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, Zhifang Sui. *Large Language Models are not Fair Evaluators*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 9440-9450. Autorenliste nach ACL Anthology; die arXiv-v1-Liste ist unvollständig. <https://aclanthology.org/2024.acl-long.511/> [K] [F30] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, Dongyeop Kang. *Benchmarking Cognitive Biases in Large Language Models as Evaluators*. arXiv:2309.17012, 2023; Kommentarfeld: ACL 2024. <https://arxiv.org/abs/2309.17012> [K] [F31] Arjun Panickssery, Samuel R. Bowman, Shi Feng. *LLM Evaluators Recognize and Favor Their Own Generations*. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Main Conference Track. https://papers.nips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html [K] [F32] Koki

Wataoka, Tsubasa Takahashi, Ryokan Ri. *Self-Preference Bias in LLM-as-a-Judge*. arXiv:2410.21819, 2024; Kommentarfeld: NeurIPS 2024 Safe Generative AI Workshop. <https://arxiv.org/abs/2410.21819> [W] [F33] William H. Walters, Esther Isabelle Wilder. *Fabrication and errors in the bibliographic citations generated by ChatGPT*. Scientific Reports 13, Artikel 14045, 2023, DOI 10.1038/s41598-023-41032-5. <https://doi.org/10.1038/s41598-023-41032-5> [Z] [F34] Mikaël Chelli, Jules Descamps, Vincent Lavoué u. a. (10). *Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis*. Journal of Medical Internet Research 26, e53164, 2024, DOI 10.2196/53164. Metadaten über den DOAJ-Eintrag; die Verlagsseite lieferte keinen Inhalt. <https://doi.org/10.2196/53164> [Z] [F35] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, Daniel E. Ho. *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. Journal of Empirical Legal Studies 22, S. 216, 2025. <https://law.stanford.edu/publications/hallucination-free-assessing-the-reliability-of-leading-ai-legal-research-tools> [Z] [F36] Delip Rao, Chris Callison-Burch. *BibTeX Citation Hallucinations in Scientific Publishing Agents: Evaluation and Mitigation*. arXiv:2604.03159, 2026. <https://arxiv.org/abs/2604.03159> [P] [F37] Patrik Reizinger, Wieland Brendel. *HALLMARK: Diagnosing Three Failure Modes in LLM Citation Verifiers*. arXiv:2607.18360, 2026. <https://arxiv.org/abs/2607.18360> [P] [F38] Kate Niederhoffer, Gabriella Rosen Kellerman, Angela Lee, Alex Liebscher, Kristina Rapuano, Jeffrey T. Hancock. *AI-Generated ‘Workshop’ Is Destroying Productivity*. Harvard Business Review, 22.09.2025. Praxisbeitrag, nicht begutachtet; Stichprobengröße nicht angegeben. <https://hbr.org/2025/09/ai-generated-workshop-is-destroying-productivity> [N] [F39] Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, Byron C. Wallace. *Measuring AI ‘Slop’ in Text*. arXiv:2509.19163, 2025. <https://arxiv.org/abs/2509.19163> [P] [F40] Weixin Liang, Zachary Izzo, Yaohui Zhang u. a. (12). *Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews*. arXiv:2403.07183, 2024; Kommentarfeld: ICML ’24. <https://arxiv.org/abs/2403.07183> [K] [F41] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, Robert West. *The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates*. arXiv:2405.02150, 2024. <https://arxiv.org/abs/2405.02150> [P] [F42] Claudine Gartenberg, Shariq Hasan, Alex Murray, Lamar Pierce. *More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review*. Organization Science 37(3), 2026, DOI 10.1287/orsc.2026.ed.v37.n3. Editorial der AI Task Force der Zeitschrift, kein begutachteter Aufsatz. <https://pubsonline.informs.org/doi/10.1287/orsc.2026.ed.v37.n3> [D] [F43] Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi u. a. (10). *Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance*. Radiology 307(4), 2023, DOI 10.1148/radiol.222176. <https://doi.org/10.1148/radiol.222176> [Z] [F44] Ihsan Ayyub Qazi, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, Muhammad Hamad Alizai. *Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians*. medRxiv, 08.09.2025, DOI 10.1101/2025.08.23.25334280; Studienregister NCT06963957. Preprint, nicht begutachtet. <https://www.medrxiv.org/content/10.1101/2025.08.23.25334280v2> [P] [F45] Emely Rosbach, Jonas Ammeling, Jonathan Ganz u. a. (7). *Stuck on Suggestions: Automation Bias, the Anchoring Effect, and the Factors That Shape Them in Computational Pathology*. arXiv:2603.11821, 2026; Kommentarfeld: angenommen bei MELBA. <https://arxiv.org/abs/2603.11821> [P] [F46] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, Francis Rhys Ward. *AI Sandbagging: Language Models can Strategically Underperform on Evaluations*. arXiv:2406.07358, 2024. <https://arxiv.org/abs/2406.07358> [P]

Strang B — Verfahren, Verifizierer, Architekturen

[F47] Jonathan Uesato, Nate Kushman, Ramana Kumar u. a.. *Solving math word problems with process- and outcome-based feedback*. arXiv:2211.14275, 2022. <https://arxiv.org/abs/2211.14275> [P] [F48] Hunter Lightman, Vineet Kosaraju, Yura Burda u. a. (10). *Let’s Verify Step by Step*. arXiv:2305.20050, 2023. Die Abstractseite nennt keine Veranstaltung. <https://arxiv.org/abs/2305.20050> [P] [F49] Peiyi Wang, Lei Li, Zhihong Shao u. a.. *Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 9426-9439. <https://aclanthology.org/2024.acl-long.510/> [K] [F50] Chujie Zheng, Zhenru Zhang, Beichen Zhang u. a.. *ProcessBench: Identifying Process Errors in Mathematical Reasoning*. arXiv:2412.06559, 2024; Kommentarfeld: ACL 2025. <https://arxiv.org/abs/2412.06559> [P] [F51] Lunjun Zhang, Arian Hosseini, Hritik Bansal u. a.. *Generative Verifiers: Reward Modeling as Next-Token Prediction*. arXiv:2408.15240, 2024; Kommentarfeld: ICLR 2025. <https://arxiv.org/abs/2408.15240> [P] [F52] Xuezhi Wang, Jason Wei, Dale Schuurmans u. a.. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv:2203.11171, 2022; Kommentarfeld: ICLR 2023. <https://arxiv.org/abs/2203.11171> [P] [F53] Charlie Snell, Jaehoon Lee, Kelvin Xu, Aviral Kumar. *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters*. arXiv:2408.03314, 2024. <https://arxiv.org/abs/2408.03314> [P] [F54] Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen u. a.. *Shrinking the Generation-Verification Gap with Weak Verifiers*. arXiv:2506.18203, 2025; Kommentarfeld: NeurIPS 2025. <https://arxiv.org/abs/2506.18203> [P] [F55] Yefan Zhou, Austin Xu, Yilun Zhou u. a.. *Variation in Verification: Understanding Verification Dynamics in Large Language Models*. arXiv:2509.17995, 2025; Kommentarfeld: ICLR 2026. <https://arxiv.org/abs/2509.17995> [P] [F56] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. arXiv:2305.14325, 2023. <https://arxiv.org/abs/2305.14325> [P] [F57] Akbir Khan, John Hughes, Dan Valentine u. a. (10). *Debating with More Persuasive LLMs Leads to More Truthful Answers*. arXiv:2402.06782, 2024. <https://arxiv.org/abs/2402.06782> [P] [F58] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, Arnau Pretorius. *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. arXiv:2311.17371, 2023. <https://arxiv.org/abs/2311.17371> [P] [F59] Mert Cemri, Melissa Z. Pan, Shuyi Yang u. a. (13). *Why Do Multi-Agent LLM Systems Fail?*. arXiv:2503.13657, 2025. <https://arxiv.org/abs/2503.13657> [P] [F60] Sewon Min, Kalpesh Krishna, Xinxin Lyu u. a. (9). *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv:2305.14251, 2023; Kommentarfeld: EMNLP 2023 Hauptkonferenz. <https://arxiv.org/abs/2305.14251> [K] [F61] Jerry Wei, Chengrun Yang, Xinying Song u. a. (12). *Long-form factuality in large language models*. NeurIPS 2024 (Poster). <https://neurips.cc/virtual/2024/poster/96675> [K] [F62] Liyan Tang, Philippe Laban,

Greg Durrett. *MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents*. arXiv:2404.10774, 2024; Kommentarfeld: EMNLP 2024. <https://arxiv.org/abs/2404.10774> [K] [F63] Yixiao Song, Yekyung Kim, Mohit Iyyer. *VERISCORE: Evaluating the factuality of verifiable claims in long-form text generation*. arXiv:2406.19276, 2024. <https://arxiv.org/abs/2406.19276> [P] [F64] Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen. *Enabling Large Language Models to Generate Text with Citations*. arXiv:2305.14627, 2023; Kommentarfeld: EMNLP 2023. <https://arxiv.org/abs/2305.14627> [K] [F65] Luyu Gao, Aman Madaan, Shuyan Zhou u. a. (8). *PAL: Program-aided Language Models*. arXiv:2211.10435, 2022. <https://arxiv.org/abs/2211.10435> [P] [F66] Liangming Pan, Alon Albalak, Xinyi Wang, William Yang Wang. *Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning*. arXiv:2305.12295, 2023; Kommentarfeld: EMNLP 2023 Findings. <https://arxiv.org/abs/2305.12295> [K] [F67] Zhibin Gou, Zhihong Shao, Yeyun Gong u. a. (7). *CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing*. arXiv:2305.11738, 2023; Kommentarfeld: ICLR 2024. <https://arxiv.org/abs/2305.11738> [P] [F68] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley u. a. (13). *Agent-as-a-Judge: Evaluate Agents with Agents*. arXiv:2410.10934, 2024. <https://arxiv.org/abs/2410.10934> [P] [F69] Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, Rebecca Qian. *TRAIL: Trace Reasoning and Agentic Issue Localization*. arXiv:2505.08638, 2025. <https://arxiv.org/abs/2505.08638> [P] [F70] Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade u. a. (10). *AgentRewardBench: Evaluating Automatic Evaluations of Web Agent Trajectories*. arXiv:2504.08942, 2025. <https://arxiv.org/abs/2504.08942> [P] [F71] Hongliu Cao, Ilias Driouich, Eoin Thomas. *Beyond Task Completion: Revealing Corrupt Success in LLM Agents through Procedure-Aware Evaluation*. arXiv:2603.03116, 2026. <https://arxiv.org/abs/2603.03116> [P] [F72] Daehong Kim, Haichao Miao, Shusen Liu. *LEDGER: Claim-to-Evidence Trace Graphs for Auditing LLM Agents*. arXiv:2608.18398, 2026. <https://arxiv.org/abs/2608.18398> [P] [F73] Yiqi Wang, Jiaqi Zhang, Zhangkai Wu u. a. (11). *From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents*. arXiv:2606.04990, 2026. <https://arxiv.org/abs/2606.04990> [P] [F74] Laksh Advani. *From Confident Closing to Silent Failure: Characterizing False Success in LLM Agents*. arXiv:2606.09863, 2026; Kommentarfeld: FAGEN@ICML2026. <https://arxiv.org/abs/2606.09863> [W]

Strang C — Nachweis, Offenlegung, Vorab-Registrierung

[F75] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, David T. Mellor. *The preregistration revolution*. PNAS 115(11), 2600-2606, 2018, DOI 10.1073/pnas.1708274114. <https://doi.org/10.1073/pnas.1708274114> [Z] [F76] Christopher D. Chambers, Loukia Tzavella. *The past, present and future of Registered Reports*. Nature Human Behaviour 6(1), 29-42, 2022, DOI 10.1038/s41562-021-01193-7. <https://doi.org/10.1038/s41562-021-01193-7> [Z] [F77] Courtney K. Soderberg, Timothy M. Errington, Sarah R. Schiavone, Julia Bottesini, Felix Singleton Thorn, Simine Vazire, Kevin M. Esterling, Brian A. Nosek. *Initial evidence of research quality of registered reports compared with the standard publishing model*. Nature Human Behaviour 5(8), 990-997, 2021, DOI 10.1038/s41562-021-01142-4. <https://doi.org/10.1038/s41562-021-01142-4> [Z] [F78] Anne M. Scheel, Mitchell R. M. J. Schijen, Daniël Lakens. *An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports*. Advances in Methods and Practices in Psychological Science 4(2), 2021, DOI 10.1177/25152459211007467. <https://doi.org/10.1177/25152459211007467> [Z] [F79] Tom E. Hardwicke, Eric-Jan Wagenmakers. *Reducing bias, increasing transparency and calibrating confidence with preregistration*. Nature Human Behaviour 7(1), 15-26, 2023, DOI 10.1038/s41562-022-01497-2. <https://doi.org/10.1038/s41562-022-01497-2> [Z] [F80] Aline Claesen, Sara Gomes, Francis Tuerlinckx, Wolf Vanpaemel. *Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies*. Royal Society Open Science 8(10), 2021, DOI 10.1098/rsos.211037. <https://doi.org/10.1098/rsos.211037> [Z] [F81] Norbert L. Kerr. *HARKing: Hypothesizing After the Results are Known*. Personality and Social Psychology Review 2(3), 196-217, 1998, DOI 10.1207/s15327957pspr0203_4. https://doi.org/10.1207/s15327957pspr0203_4 [Z] [F82] Gowri Gopalakrishna, Gerben ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, Lex M. Bouter. *Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands*. PLOS ONE 17(2), e0263023, 2022, DOI 10.1371/journal.pone.0263023. <https://doi.org/10.1371/journal.pone.0263023> [Z] [F83] John Protzko, Jon Krosnick, Leif Nelson, Brian A. Nosek u. a. (17). *ZURÜCKGEZOGEN: High replicability of newly discovered social-behavioural findings is achievable*. Nature Human Behaviour 8(2), 311-319, 2024, DOI 10.1038/s41562-023-01749-9; Rücknahmehinweis NHB 8(10), 2067, 24.09.2024, DOI 10.1038/s41562-024-01997-3. <https://doi.org/10.1038/s41562-024-01997-3> [Z] [F84] International Committee of Medical Journal Editors. *A. Defining the Role of Authors and Contributors*. Richtliniendokument, ohne Fassungsdatum auf der Seite; Stand 06.10.2026. <https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html> [D] [F85] Committee on Publication Ethics. *Authorship and AI tools*. COPE Position Statement, Fassung 1, 20.08.2024. <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools> [D] [F86] Springer Nature. *Artificial Intelligence (AI) – Editorial Policy*. Ampelschema ohne Fassungsnummer und Revisionsdatum; Stand 06.10.2026. <https://www.nature.com/nature-portfolio/editorial-policies/ai> [D] [F87] Elsevier. *Generative AI policies for journals*. Seitenangabe: Policy updated June 2026. <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals> [D] [F88] Association for Computational Linguistics. *ACL Policy on Publication Ethics*. Letzte Revision laut Seite 17.08.2026. https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics [D] [F89] ACL Rolling Review. *Responsible NLP Research – Checkliste, Punkt E1*. Stand 06.10.2026. <https://aclrollingreview.org/responsibleNLPresearch/> [D] [F90] NeurIPS. *NeurIPS 2025 Policy on the Use of Large Language Models*. Stand 06.10.2026. <https://neurips.cc/Conferences/2025/LLM> [D] [F91] ICML. *ICML 2026 Policy for LLM use in reviewing*. Stand 06.10.2026. <https://icml.cc/Conferences/2026/LLM-Policy> [D] [F92] ICLR. *Policies on Large Language Model Usage at ICLR 2026*. Stand 06.10.2026. <https://iclr.cc/FAQ/LLM> [D] [F93] Europäische Kommission, Generaldirektion Forschung und Innovation. *Living guidelines on the responsible use of generative AI in research*. Dritte Fassung, Mai 2026. https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en [D] [F94] Coalition for Content Provenance and

Authenticity. *Content Credentials: C2PA Technical Specification, Version 2.3*. Dokument datiert Dezember 2025. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html [D] [F95] Coalition for Content Provenance and Authenticity. *C2PA Conformance Program and C2PA Trust List*. Stand 06.10.2026; Übergangsliste eingefroren zum 01.01.2026. <https://c2pa.org/conformance/> [D] [F96] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein. *A Watermark for Large Language Models*. arXiv:2301.10226, 2023; ICML 2023. <https://arxiv.org/abs/2301.10226> [K] [F97] Sumanth Dathathri, Abigail See, Sumedh Ghaisas u. a. (24). *Scalable watermarking for identifying large language model outputs*. Nature 634, 818-823, 2024, DOI 10.1038/s41586-024-08025-4. <https://doi.org/10.1038/s41586-024-08025-4> [Z] [F98] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi. *Can AI-Generated Text be Reliably Detected?*. Transactions on Machine Learning Research; arXiv:2303.11156, 2023/2025. <https://arxiv.org/abs/2303.11156> [Z] [F99] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, Boaz Barak. *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models*. arXiv:2311.04378, 2023/2025; ICML 2024. <https://arxiv.org/abs/2311.04378> [K] [F100] World Wide Web Consortium. *PROV-DM: The PROV Data Model*. W3C Recommendation, 30.04.2013. <https://www.w3.org/TR/prov-dm/> [D] [F101] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. *Model Cards for Model Reporting*. FAT* '19, Atlanta, 29.-31.01.2019; arXiv:1810.03993. <https://arxiv.org/abs/1810.03993> [K] [F102] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. *Datasheets for Datasets*. Communications of the ACM, Dezember 2021; arXiv:1803.09010. <https://arxiv.org/abs/1803.09010> [Z] [F103] Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, Bela Gipp. *AI Usage Cards: Responsibly Reporting AI-generated Content*. arXiv:2303.03886, 2023. <https://arxiv.org/abs/2303.03886> [P] [F104] Jack Gallifant, Majid Afshar, Saleem Ameen u. a. (20). *The TRIPOD-LLM reporting guideline for studies using large language models*. Nature Medicine 31(1), 60-69, 2025, DOI 10.1038/s41591-024-03425-5. Autorenlistenform aus einem Repositoriumseintrag; vor vollständiger Nennung an der Verlagsseite prüfen. <https://doi.org/10.1038/s41591-024-03425-5> [T] [F105] Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, Jelte M. Wicherts. *The prevalence of statistical reporting errors in psychology (1985-2013)*. Behavior Research Methods 48(4), 1205-1226, 2016, DOI 10.3758/s13428-015-0664-2. <https://doi.org/10.3758/s13428-015-0664-2> [Z] [F106] Nicholas J. L. Brown, James A. J. Heathers. *The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology*. Social Psychological and Personality Science 8(4), 363-369, 2017, DOI 10.1177/1948550616673876. <https://doi.org/10.1177/1948550616673876> [Z] [F107] Joe Menke, Martijn Roelandse, Burak Ozyurt, Maryann Martone, Anita Bandrowski. *The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods*. iScience 23(11), 101698, 2020, DOI 10.1016/j.isci.2020.101698. Begleitarbeit: Menke u. a., JMIR 24(6), e37324, 2022, DOI 10.2196/37324. <https://doi.org/10.1016/j.isci.2020.101698> [Z] [F108] Guillaume Cabanac, Cyril Labbé, Alexander Magazinov. *The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment*. arXiv:2210.04895, 2022, DOI 10.48550/arXiv.2210.04895. <https://arxiv.org/abs/2210.04895> [P] [F109] Richard Van Noorden. *More than 10,000 research papers were retracted in 2023 – a new record*. Nature 624(7992), 479-481, 12.12.2023, DOI 10.1038/d41586-023-03974-8. Nachrichtenteil. <https://doi.org/10.1038/d41586-023-03974-8> [N] [F110] STM Association. *STM Integrity Hub in action: a chronicle*. 12.09.2025. Die eingesetzten Detektoren sind nicht offengelegt. <https://stm-assoc.org/stm-integrity-hub-in-action-a-chronicle/> [D] [F111] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, Jan Lause. *Delving into LLM-assisted writing in biomedical publications through excess vocabulary*. Science Advances 11(27), 02.07.2025, DOI 10.1126/sciadv.adt3813. <https://doi.org/10.1126/sciadv.adt3813> [Z] [F112] Giovanni E. Cacciamani, Michael B. Eppler, Conner Ganjavi, Asli Pekan, Brett Biedermann, Gary S. Collins, Inderbir S. Gill. *Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use (CANGARU) Guidelines*. arXiv:2307.08974, 2023. Nur das Protokoll ist geprüft; eine veröffentlichte Endfassung ist nicht bestätigt. <https://arxiv.org/abs/2307.08974> [T]

Strang D — Agentische Forschungssysteme und ihre Bewertung

[F113] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, Jeff Clune. *Towards end-to-end automation of AI research*. Nature 651, 914-919, 25.03.2026, DOI 10.1038/s41586-026-10265-5. Preprintfassung unter anderem Titel und mit sechs Autoren: arXiv:2408.06292. <https://doi.org/10.1038/s41586-026-10265-5> [Z] [F114] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, David Ha. *The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search*. arXiv:2504.08066, 2025. <https://arxiv.org/abs/2504.08066> [P] [F115] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, Emad Barsoum. *Agent Laboratory: Using LLM Agents as Research Assistants*. arXiv:2501.04227, 2025. <https://arxiv.org/abs/2501.04227> [P] [F116] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, Linyi Yang. *CycleResearcher: Improving Automated Research via Automated Review*. arXiv:2411.00816, 2024; Kommentarfeld: ICLR 2025. <https://arxiv.org/abs/2411.00816> [P] [F117] Daniil A. Boiko, Robert MacKnight, Ben Kline, Gabe Gomes. *Autonomous chemical research with large language models*. Nature 624, 570-578, 20.12.2023, DOI 10.1038/s41586-023-06792-0. <https://doi.org/10.1038/s41586-023-06792-0> [Z] [F118] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, Philippe Schwaller. *Augmenting large language models with chemistry tools*. Nature Machine Intelligence 6, 525-535, 08.05.2024, DOI 10.1038/s42256-024-00832-8; Preprint arXiv:2304.05376. <https://doi.org/10.1038/s42256-024-00832-8> [Z] [F119] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei u. a. (16). *An autonomous laboratory for the accelerated synthesis of inorganic materials*. Nature 624, 86-91, 2023, DOI 10.1038/s41586-023-06734-w. <https://doi.org/10.1038/s41586-023-06734-w> [Z] [F120] Josh Leeman, Yuhan Liu, Joseph Stiles, Scott B. Lee, Prajna Bhatt, Leslie M. Schoop, Robert G. Palgrave. *Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis*. PRX Energy 3(1), 011002, 2024, DOI 10.1103/PRXEnergy.3.011002. Metadaten aus einem Hochschulverzeichnis; die Verlagsseite wies den Abruf ab. <https://doi.org/10.1103/PRXEnergy.3.011002> [T] [F121] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John

E. Pak, James Zou. *The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies*. Nature 646, 716-723, 29.07.2025, DOI 10.1038/s41586-025-09442-9. <https://doi.org/10.1038/s41586-025-09442-9> [Z] [F122] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu u. a. (44). *Accelerating scientific discovery with Co-Scientist*. Nature, 2026, DOI 10.1038/s41586-026-10644-y; arXiv:2502.18864. Fassung v1 trägt einen anderen Titel. <https://arxiv.org/abs/2502.18864> [Z] [F123] Alireza Ghafarollahi, Markus J. Buehler. *SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning*. arXiv:2409.05556, 2024. <https://arxiv.org/abs/2409.05556> [P] [F124] Haoyang Su, Renqi Chen, Shixiang Tang u. a. (13). *Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System*. arXiv:2410.09403, 2024; Kommentarfeld: ACL 2025 Main Conference. <https://arxiv.org/abs/2410.09403> [K] [F125] Qingyun Wang, Doug Downey, Heng Ji, Tom Hope. *SciMON: Scientific Inspiration Machines Optimized for Novelty*. Proc. 62nd ACL (Vol. 1: Long Papers), 2024, S. 279-299. <https://aclanthology.org/2024.acl-long.18/> [K] [F126] Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, Tom Hope. *Literature-Grounded Novelty Assessment of Scientific Ideas*. arXiv:2506.22026, 2025. <https://arxiv.org/abs/2506.22026> [P] [F127] Noy Sternlicht, Simra Shahid, Peter Jansen, Daniel S. Weld, Pao Siangliulue, Tom Hope. *Old Ideas, Novel Problems: The Instability of LLM-Based Novelty Evaluation*. arXiv:2610.02022, 01.10.2026. <https://arxiv.org/abs/2610.02022> [P] [F128] Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, Hao Sun. *Evaluating LLMs' Divergent Thinking Capabilities for Scientific Idea Generation with Minimal Context*. arXiv:2412.17596, 2024. Umbenannt; LiveIdeaBench ist der Benchmarkname, nicht der Titel. <https://arxiv.org/abs/2412.17596> [P] [F129] Chenglei Si, Diyi Yang, Tatsunori Hashimoto. *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers*. arXiv:2409.04109, 2024. Die Abstractseite nennt keine Veranstaltung. <https://arxiv.org/abs/2409.04109> [P] [F130] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe u. a. (12). *MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering*. arXiv:2410.07095, 2024; Kommentarfeld nennt eine ICLR-Fassung. <https://arxiv.org/abs/2410.07095> [P] [F131] Giulio Starace, Oliver Jaffe, Dane Sherburn u. a. (13). *PaperBench: Evaluating AI's Ability to Replicate AI Research*. arXiv:2504.01848, 2025. <https://arxiv.org/abs/2504.01848> [P] [F132] Hjalmar Wijk, Tao Lin, Joel Becker u. a. (23). *RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts*. arXiv:2411.15114, 2024. <https://arxiv.org/abs/2411.15114> [P] [F133] Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebel, Arvind Narayanan. *CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark*. arXiv:2409.11363, 2024/2026. <https://arxiv.org/abs/2409.11363> [P] [F134] Minjun Zhu, Qiujie Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, Yue Zhang. *AI Scientists Fail Without Strong Implementation Capability*. arXiv:2506.01372, 2025; Kommentarfeld: Position. <https://arxiv.org/abs/2506.01372> [P] [F135] Joydeep Biswas, Sheila Schoepp, Gautham Vasani u. a. (13). *AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot*. arXiv:2604.13940, 2026. Verfasst von den Organisatoren des Versuchs. <https://arxiv.org/abs/2604.13940> [P] [F136] ICLR 2026 Program Chairs. *ICLR 2026 Response to LLM-Generated Papers and Reviews*. ICLR Blog, 19.11.2025. <https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/> [D] [F137] Alex Lu, Seth Lazar, David Rügamer (Position Paper Chairs). *AI-Generated Papers in the NeurIPS 2026 Position Paper Track*. NeurIPS Blog, 02.06.2026. <https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/> [D] [F138] James Zou, Owen Queen, Nitya Thakkar, Eric Sun, Federico Bianchi (Chairs). *Open Conference of AI Agents for Science 2025 (Agents4Science)*. Konferenzseite; virtuelle Konferenz 22.10.2025. <https://agents4science.stanford.edu> [D] [F139] Nilmani Singh, Stephan Lane, Tianhao Yu, Jingxia Lu, Adrianna Ramos, Haiyang Cui, Huimin Zhao. *A generalized platform for artificial intelligence-powered autonomous enzyme engineering*. Nature Communications 16, 5648, 01.07.2025, DOI 10.1038/s41467-025-61209-y. <https://doi.org/10.1038/s41467-025-61209-y> [Z] [F140] Grant M. Landwehr, Jonathan W. Bogart, Carol Magalhães, Eric G. Hammarlund, Ashty S. Karim, Michael C. Jewett. *Accelerated enzyme engineering by machine-learning guided cell-free expression*. Nature Communications 16, 865, 20.01.2025, DOI 10.1038/s41467-024-55399-0. <https://doi.org/10.1038/s41467-024-55399-0> [Z]

11. Herkunft und Stand

Diese Feldkarte gehört zur Working Note *Iteratio Simulata* (Kennung FORSCH-PROV0459) und ist deren Umfeldteil. Sie ist eigenständig lesbar und setzt die Kenntnis der fünf Marker nicht voraus.

Erhebungszeitraum: 06.10.2026, ein Arbeitstag, vier getrennte Stränge. **Erhebungsweise:** Websuche und Abruf der Primärseiten; keine Datenbanklizenz, kein Zugriff auf Volltexte hinter Bezahlschranken. **Nicht erreichbar:** PubMed und PMC weisen automatisierte Abrufe ab; mehrere Verlagsseiten antworten mit HTTP 403; ein Normtext bricht in HTML und PDF vor den Artikeln ab. Die Folgen sind in 8.2 benannt und in der Quellenliste als [T] geführt. **Nicht enthalten:** Herstellerangaben von Detektoranbietern (8.3); Volltextlektüre der agentischen Systeme (9.3).

Lizenz: CC BY 4.0. Namensnennung: Thomas Schüller, ORCID 0009-0003-9799-6747.

Zitiervorschlag: Schüller, T. (2026). *Feldkarte: Verfahrensnachweis in LLM-gestützter Wissensarbeit*. Umfeldteil zur Working Note *Iteratio Simulata*, Fassung v0.5, FORSCH-PROV0459.

Registerhinweis: Die Nummer PROV0459 ist provisorisch; die endgültige Nummer wird gegen das Werkregister gesetzt.