Falscher Konsens in Multi-LLM-Debatten: Messprobleme, Prüfverfahren und offene Evidenzlücken

KI-generierter Entwurf der Forschungsplattform Disputatio – zur Begutachtung durch Fachleute

Entwurf d2fc854b · Version 2 · Fertiggestellt 2026-10-25

> ***Forschungsfrage:*** *Was zeigt die empirische Forschung darüber, unter welchen Bedingungen Debatten zwischen mehreren Sprachmodellen zu falschem Konsens führen, und welche Prüfmechanismen können solche Fehler nachweislich erkennen oder verringern?*

Zusammenfassung

Die untersuchten Studien berichten, dass Peer-Austausch richtige Antworten verdrängen und Mehrheitsaggregation vorhandene richtige Minderheitslösungen verwerfen kann; Übereinstimmung und Antwortwechsel allein zeigen jedoch weder falschen Konsens noch schädlichen Peer-Einfluss an. [1, 2][2] Eine kritische Literatursynthese überwiegend aus Abstracts und ausgewählten Volltextstellen verbindet unabhängige Fachpositionen, eine Prüfung ihrer Überschneidungen, strukturierte Kritik und Revision sowie getrennte Bewertungen empirischer Berichte und unbestätigter Hypothesen.[3] Untersucht wurden unter anderem strukturbewusste Aggregation, korrekt argumentierender Dissens, Informationen über Peer-Zuverlässigkeit und abgeschirmte Verifikation; ihre settingspezifischen Verbesserungen begründen weder eine allgemeine Rangordnung noch isolierte Wirkungen sämtlicher Komponenten. [3, 4, 5, 6][4]

UNGELÖSTER KONFLIKT 1

Offen bleiben die Abgrenzung tatsächlicher Bedingungsabhängigkeit von inkommensurablen Messgrößen und die Spannung zwischen adaptiver Nutzenvorhersage und nicht unmittelbar beobachtbaren kontrafaktischen Eingriffseffekten.[5]

Fehlende Primärdatenprüfung, nicht isolierte Komponentenwirkungen und ungeprüfte Quellenunabhängigkeit begrenzen die Schlussfolgerungen; erforderlich sind kontrollierte, budgetangepasste Vergleiche mit unabhängigen Referenzkriterien, getrennten Kalibrierungs- und Testdaten sowie gemeinsamer Bilanzierung hilfreicher und schädlicher Revisionen.[6]

Dieser Text wurde von KI-Agenten aus einem eingefrorenen Forschungs-Snapshot erzeugt. Keine Aussage gilt dadurch als wahr oder akzeptiert. Typ und Herkunft jeder Aussage sind im Anhang dokumentiert (hochgestellte Nummern).

1 Einleitung

Die Bewertung von Multi-LLM-Debatten verlangt eine Trennung zwischen Übereinstimmung, Korrektheit und dem Beitrag des Austauschs. Falscher Konsens bezeichnet hier die zu untersuchende falsche Übereinstimmung, nicht eine bereits einheitlich operationalisierte Kennzahl. Eine falsche Mehrheit ist enger auf das Ergebnis einer Stimmenaggregation bezogen und setzt keine Einstimmigkeit voraus. Ebenso kann eine adjudizierte Einzelentscheidung falsch sein, obwohl die zuvor erzeugten Antworten weiterhin auseinandergehen. Ein Antwortwechsel beschreibt zunächst lediglich eine Veränderung gegenüber der Erstantwort. Konformität ist demgegenüber eine Interpretation der Anpassung an andere Positionen, die durch den Wechsel allein nicht identifiziert wird. Schädlicher Peer-Einfluss betrifft schließlich eine Verschlechterung nach fremder Information, die gegen geeignete Kontrollbedingungen und ein unabhängiges Korrektheitskriterium bewertet werden muss. Diese Begriffe sind deshalb weder

synonym noch durch dieselbe Endkennzahl hinreichend erfasst. Ihre Unterscheidung verhindert, dass eine sichtbare Annäherung bereits als Fehler, ein unverändertes Urteil als Robustheit oder eine richtige Endantwort als spezifischer Debattennutzen eingeordnet wird. [2][7] Eine anschauliche Grundlage dieser Begriffsarbeit bietet Not All Flips Are Conformity. Der Abstract vergleicht bei fünf GPT-4o-Modellen auf MMLU-Pro-Fragen mit anfänglichem Dissens Selbstreflexion, bloße Sichtbarkeit fremder Positionen und vollständige Begründungen. Bereits ohne Peer-Information wechseln 37 Prozent der Modell-Frage-Paare ihre Antwort. Unter den in dieser Kontrolle stabilen Paaren übernehmen anschließend 29,3 Prozent eine Peer-Antwort, nachdem sie fremde Positionen gesehen haben. Der Nenner der zweiten Angabe ist damit eine bedingt ausgewählte Teilmenge, nicht die Gesamtheit des ersten Vergleichs. Zudem umfassen diese Übernahmen Korrekturen, neue Fehler und Wechsel zwischen falschen Antworten. Die Angaben dürfen folglich weder addiert noch als zwei Schätzungen derselben Konformitätsrate behandelt werden. Ihr methodischer Gehalt liegt in der Zerlegung von Übergängen und der Kontrolle spontaner Revisionen. Der beobachtete Verlauf muss darauf bezogen werden, ob eine richtige Antwort verfügbar war, welche Information sichtbar wurde und ob sich die individuelle oder die aggregierte Antwort verbesserte. [2][8] The Cost of Consensus beschreibt dagegen problematische Verläufe homogener Gruppen kleinerer Modelle auf GSM-Hard und MMLU-Hard. Die dort berichteten Größen betreffen Mehrheitsübernahme, Destabilisierung zuvor richtiger Antworten und das Verwerfen vorhandener richtiger Kandidaten durch die Aggregation. Eine Übernahmerate zählt andere Ereignisse als eine Vulnerabilitätsrate; die Oracle-Lücke beschreibt wiederum eine Differenz zur Nutzung vorhandener richtiger Kandidaten und keine individuelle Anpassungswahrscheinlichkeit. Selbst hohe Werte dieser Größen ergeben deshalb keine gemeinsame Prävalenz falschen Konsenses. Wichtig ist außerdem die zeitliche Unterscheidung: Mehrere Modelle können schon vor dem Austausch denselben Fehler erzeugen, während andere Fehlentscheidungen erst nach Peer-Exposition auftreten. Diese Arbeitsunterscheidung ordnet beobachtbare Verläufe, identifiziert jedoch keine vollständige Ursachenstruktur. Insbesondere belegen gleiche falsche Erstantworten für sich genommen weder gemeinsame Wissenslücken noch überlappende Trainingsdaten. Dafür wären zusätzliche Informationen oder kontrollierte Variationen notwendig. Die Synthese verwendet die Fehlerpfade daher als analytische Vergleichspunkte und nicht als bereits unabhängig bestätigte Mechanismen. [1][9] Die Gegenmaßnahmen im ausgewerteten Material greifen an unterschiedlichen Stellen an. Strukturbewusste Aggregation prüft Verzweigungen vorhandener Begründungen, korrekt argumentierender Dissens stellt Gegeninformation bereit, Zuverlässigkeitsinformationen verändern die Gewichtung fremder Antworten, und abgeschirmte Verifikation begrenzt den Zugriff auf die ursprüngliche Antwort. Ein Gewinn bei einer dieser Aufgaben ist nicht automatisch ein Gewinn bei den anderen. So betrifft die Wiedergewinnung einer vorhandenen richtigen Minderheitslösung zunächst die Kandidatenauswahl, während Retrieval zusätzliche Information zugänglich machen kann. Organisatorische Trennung zwischen Antworterzeugung und Prüfung garantiert dabei weder unabhängige Wissensbestände noch voneinander unabhängige Quellen. Auch eine bessere Endgenauigkeit ist von einer gemessenen Erkennungsrate zu unterscheiden: Ein Verfahren kann insgesamt mehr Aufgaben lösen, ohne falsche Übereinstimmung ausdrücklich zu diagnostizieren. Klinische Terminologiefehler, Halluzinationen im Retrieval-Kontext und mathematische Antwortaggregation werden deshalb nicht als austauschbare Evaluationsgegenstände behandelt. Die Frage nach Schutz vor falschem Konsens verlangt für jeden Ansatz eine genaue Zuordnung von Zielgröße, Informationszugang und Vergleichsbedingung. [3, 6, 7][10] Die Leitfrage lautet entsprechend, unter welchen methodischen Voraussetzungen falsche Gruppenentscheidungen von korrekter Konvergenz getrennt und ihre Veränderungen bestimmten Eingriffen zugerechnet werden können. Der Artikel konzentriert sich nicht auf einen allgemeinen Sieger unter Debattenverfahren, sondern auf die Tragfähigkeit der

verfügbaren Vergleiche. Dazu werden drei Anforderungen auseinandergehalten: ein Zusammenhang mit riskanten Verläufen, eine hinreichende Klassifikation von Fehlern und eine Vorhersage des Nutzens eines konkreten Eingriffs. Ein Merkmal kann die erste Anforderung erfüllen, ohne die beiden anderen zu erfüllen. Die bereits formulierten Hypothesen werden entlang dieser Trennung gebündelt und unter ihren zentralen Gegenargumenten dargestellt. Ihre Revision verbessert die Bestimmtheit der Forschungsfragen, liefert aber keine neuen Erfolgsraten. Im Mittelpunkt stehen unabhängige Referenzkriterien, kontrollierte Peer-Exposition, fähigkeitsbezogene Vergleichbarkeit und eine vollständige Nutzen-Kosten-Bilanz. Die abschließend getrennt nummerierten Forschungslücken beziehen sich auf diese Voraussetzungen und sollen verhindern, dass aus methodischer Plausibilität eine noch nicht geprüfte Anwendungsempfehlung entsteht.[11]

2 Hypothesen

H1. Hohe peer-blinde Ausgangsübereinstimmung und spätere Antwortangleichung könnten als Merkmale einer Fehlerklassifikation dienen; ihr Status ist unbestätigt. Die stärkere ursprüngliche Behauptung, beide Merkmale trennten korrelierten Ausfall und Ansteckung und verhinderten budgetgleich mehr Fehler ohne zusätzliche Korrekturverluste, wird nicht fortgeführt. Der zentrale Gegeneinwand lautet, dass sofortige Einigkeit richtig und spätere Angleichung eine legitime Korrektur sein kann. Eine vorgeschlagene erste Prüfung würde deshalb gespeicherte Erstantworten und unabhängige Referenzlösungen verwenden, neue Evidenz kontrollieren und vorbestehende gemeinsame Fehler, spätere schädliche Übergänge, korrekte Konvergenz und richtige Revision unterscheiden. Zur Hypothesenfamilie gehört das Gegenmodell, dass ein nur auf spätere Angleichung reagierender Auslöser gemeinsame Ausgangsfehler unterprüft und budgetgleicher Zufallsprüfung unterliegen könnte. Auch eine Interaktion zwischen Erstantwort-Fehlerkorrelation und geeignetem Prüfverfahren bleibt offen. Weder eine feste Zuordnung der Merkmale zu Eingriffen noch eine gemessene Trennschärfe wird vorausgesetzt; insbesondere ist die angenommene Unabhängigkeit externer Referenzen noch abzusichern.[12] H2. Präinterventionelle Merkmale könnten den durchschnittlichen Nutzen bestimmter Prüfinterventionen auf zurückgehaltenen Aufgaben vorhersagen; der Status bleibt unbestätigt und die Unsicherheit hoch. Die ursprüngliche Konvergenzregel wird nur als vorgeschlagene Vergleichsbedingung beibehalten. Gegen die Nutzenvorhersage spricht insbesondere, dass individuelle kontrafaktische Effekte nicht direkt beobachtbar sind und Risikomerkmale ohne gesonderte Kalibrierung Rauschmuster reproduzieren können. Vorgeschlagen sind deshalb randomisierte Eingriffsvergleiche in Kalibrierungsdaten, ausschließlich vorher verfügbare Merkmale, getrennte Kalibrierung von Risiko und Nutzen sowie Tests auf anderen Aufgaben und Modellfamilien. Der Vergleich würde Konvergenzauslöser und budgetgleiche Zufallsprüfung einschließen. Der Nettonutzen müsste Korrekturen, neu verursachte Fehler und gegebenenfalls gegenüber fortgesetzter Debatte verlorene Korrekturen unter vorab festgelegten Gewichten bilanzieren. Die Bewahrung unabhängiger Ausgangsbegründungen, die Prüfung ungeklärter Prämissen und Stichproben sofortiger Einigkeit sind Bestandteile der vorhandenen Vorschläge, keine bereits wirksame Kombination. Ressourcen wären separat auszuweisen.[13] H3. Eine gegenüber den Debattenantworten abgeschirmte Evidenzprüfung könnte falsche Übereinstimmung stabilisieren, wenn verschiedene Belege dieselbe fehlerhafte Ursprungsquelle übernehmen; der Status dieser möglichen Fehlerbedingung ist unbestätigt. Die Kernaussage unterscheidet organisatorische Prüfertrennung von epistemisch unabhängiger Information. Als zentrale Einschränkung fehlen direkte Beobachtungen des postulierten Schadens und eine abgesicherte Operationalisierung der Quellenabhängigkeit. Das Gegenmodell setzt rekonstruierbare Informationsursprünge und ein externes Korrektheitskriterium voraus. Eine vorgeschlagene Untersuchung würde Quellenabhängigkeit und Quellenrichtigkeit getrennt variieren

und Authentizität, tatsächliche Belegfunktion und sachliche Richtigkeit separat bewerten. Als nicht beobachtete Konsequenz wird erwogen, dass die Korrekturleistung bei gleicher Dokumentenzahl und vergleichbarer Retrievalqualität stärker von unabhängigen Ursprüngen als von der Zahl bestätigender Dokumente abhängt. Daraus folgt weder eine generelle Wirkungslosigkeit mehrerer Quellen noch ein gemessener Nachteil abgeschirmter Verifikation. Unbekannt bleiben die praktische Häufigkeit dieser Konstellation und ihr möglicher Anteil an falschen Gruppenentscheidungen.[14] H4. Der Nettoeffekt strukturierten Dissenses könnte von der Qualität prüfbarer Gegeninformation abhängen; der Status bleibt unbestätigt. Der zentrale Gegeneinwand richtet sich gegen die Übertragung eines korrekt argumentierenden Dissenters auf eine bloß zugewiesene Oppositionsrolle. Aus dessen berichteter Wirkung folgt weder eine symmetrische Destabilisierung richtigen Konsenses durch falsche Einwände noch eine vergleichbare Qualität realistisch erzeugter Opposition. Die frühere Schadensannahme wird deshalb nur als mögliches Gegenmodell beibehalten; ebenso könnte reine Opposition Rechenfehler aufdecken. Der vorgeschlagene Test würde freie Prüfung und verpflichtende Opposition jeweils mit und ohne zusätzliche relevante Evidenz vergleichen und eine unabhängige Qualitätsprüfung der Einwände ergänzen. Richtige und falsche Revisionen sowie Erzeugungs- und Prüfaufwand wären getrennt zu erfassen. Vorgegebene korrekte Gegenargumente wären eine informationsprivilegierte Vergleichsbedingung, keine realistische Standardbaseline. Ob Gruppen Gegenargumente nach sachlicher Qualität gewichten und ob die zusätzlichen Prüfkosten vertretbar sind, bleibt offen.[15]

3 Methodik

Ausgewertet wurden überwiegend Abstracts und ausgewählte Volltextstellen; eine vollständige Prüfung aller Originalmethoden, Anhänge und Primärdaten war nicht Bestandteil der dokumentierten Untersuchung. Das Design ist eine kritische Literatursynthese mit unabhängig formulierten Fachpositionen und anschließender Gegenprüfung, keine eigene experimentelle Evaluation von Multi-LLM-Debatten. Die Ausgangsbeiträge wurden getrennt erstellt und danach hinsichtlich Gegenstand, Evidenzbezug, Annahmen und Unsicherheit verglichen. Eine zunächst ohne Quellen formulierte methodische Rahmung wurde dabei nicht nachträglich als empirischer Befund behandelt. Die Unabhängigkeit der Beiträge bezeichnet ihre getrennte Erstellung, nicht statistisch unabhängige Wissensgrundlagen oder Replikationen. Gegenstand der Auswertung waren sowohl berichtete Beobachtungen als auch deren Interpretation und die daraus bereits entwickelten Forschungsfragen. Diese Ebenen blieben getrennt, damit ein plausibler Prüfplan nicht als vorhandenes Ergebnis erscheint. Soweit Literaturpassagen nur die Leistung eines Gesamtverfahrens beschreiben, wurde daraus keine isolierte Wirkung einzelner Bestandteile abgeleitet.[16] Alle vorgesehenen Beiträge der dokumentierten Phasen gingen ein: vier unabhängige Ausgangspositionen, jeweils drei Beiträge zur Literatursuche und zur Erschließung ausgewählter Volltextstellen, zwei explorative Beiträge, sieben Überschneidungsprüfungen, zwei Vorschläge, ein Kritikbeitrag, zwei Revisionen und dreizehn Bewertungen. Ein Ausfall einer vorgesehenen Perspektive ist nicht dokumentiert. Diese Prozesszahlen beschreiben den Umfang der Auseinandersetzung und sind keine wissenschaftlichen Stichprobengrößen. Insbesondere dürfen mehrere Stellungnahmen zu derselben Passage nicht als mehrfach unabhängige empirische Unterstützung gezählt werden. Die explorativen Beiträge erschlossen vorhandene Fragen, Analogien und Hypothesen; für den Artikel wurden keine zusätzlichen Hypothesen eingeführt. Inhaltlich überlappende und revidierte Vorschläge wurden zu vier Hypothesenfamilien zusammengeführt. Dabei blieben zurückgenommene stärkere Annahmen als Entwicklungsschritte erkennbar, ohne als zusätzliche Vorhersagen fortgeführt zu werden. Die Zusammenführung dient somit der begrifflichen Ordnung und verhindert, dass Varianten desselben Vorschlags den Eindruck mehrerer unabhängiger Argumente erzeugen.[17] Die Prüfung möglicher

Neuheit beschränkte sich auf Überschneidungen mit den gesammelten Positionen. Die sieben geprüften Beiträge wurden als teilweise überlappend eingeordnet; daraus wurde weder globale Neuheit noch wissenschaftliche Bestätigung abgeleitet. Verglichen wurden insbesondere bestehende Vorschläge zur konvergenzgesteuerten Prüfung, zur Erfassung gemeinsamer Ausgangsfehler, zur Quellenabhängigkeit und zur Qualität strukturierten Dissenses. Zuspitzungen wie die Unterabdeckung sofortiger gemeinsamer Fehler oder mehrere Dokumente mit derselben falschen Ursprungsquelle wurden als unbestätigte Forschungsgegenstände weitergeführt. Der begrenzte Vergleichsbestand erlaubt jedoch keine Aussage darüber, ob entsprechende Ideen außerhalb der gesammelten Literatur bereits ausgearbeitet wurden. Auch eine präzisere Formulierung nach der Überschneidungsprüfung wurde nicht als empirischer Fortschritt verbucht. Für die Synthese war vielmehr entscheidend, welche Annahmen ein Vorschlag zusätzlich benötigt und ob diese Annahmen durch die tatsächlich gelesenen Passagen adressiert werden. Die Prüfung ordnete damit den Argumentationsbestand, ersetzte aber weder eine umfassende Literaturrecherche noch einen experimentellen Test.[18] Die strukturierte Kritik konzentrierte sich auf uneinheitliche Nenner, ungesicherte Mechanismenzuschreibungen, Fähigkeitskonfundierung und die Gleichsetzung von Risikodiagnose mit Eingriffsnutzen. In den Revisionen wurden stärkere Behauptungen zurückgenommen oder auf explizit offene Prüfaufgaben eingegrenzt. Dazu gehören die behauptete Trennschärfe des Zwei-Signal-Auslösers und die vermutete symmetrische Schädlichkeit verpflichtender Opposition. Die anschließenden Bewertungen unterschieden stützende, herausfordernde, einschränkende und lediglich kontextualisierende Bezüge. Ein Kontextbezug wurde nicht zur direkten Unterstützung aufgewertet; eine argumentative Herausforderung wurde nicht als empirische Falsifikation behandelt. Bewertungen mit unzureichender Grundlage blieben entsprechend offen, und gemischte Bewertungen wurden nicht durch Mehrheitszählung vereinheitlicht. Revisionen zählten ebenfalls nicht als neue unabhängige Bestätigungen. Dieses Vorgehen erlaubt, die Reichweite einer Literaturinterpretation von der Reichweite einer Hypothese zu unterscheiden. Die beiden ausdrücklich dokumentierten Konflikte werden deshalb mit ihren stützenden und herausfordernden Argumenten erhalten, auch wenn spätere Formulierungen einen Teil der Kritik übernehmen.[19] Der fachliche Vergleich folgt den im Analyseplan benannten Dimensionen Aufgabenart, Modellklasse, unabhängige Erstantworten, Peer-Manipulation, Kontrollverfahren, Referenzkriterium, Eingriffsbestandteile und Ressourcen. Falsche Endmehrheit, falsche Einstimmigkeit und falsche adjudizierte Entscheidung werden begrifflich getrennt; ebenso werden Erkennung, Korrektur und neu verursachter Schaden nicht zu einer undifferenzierten Erfolgsgröße verdichtet. Quantitative Angaben bleiben an ihren Aufgabenraum, ihre Bedingung und ihren Nenner gebunden. Fehlende Angaben werden als Lücken behandelt, nicht durch Vermutungen ergänzt. Prognostische Befunde, kontrollierte Eingriffsvergleiche und nachträgliche Wiederholungen gespeicherter Verläufe erhalten unterschiedliche Interpretationen. Der Analyseplan enthält außerdem die gesonderte Prüfung des Bewertungsinstruments und seiner Kontextabhängigkeit. Diese Ordnung begründet eine qualitative Gegenüberstellung der verfügbaren Auszüge; sie ist keine harmonisierte Neuauswertung und keine Metaanalyse. Eine vollständige Extraktion sämtlicher vorgeschlagener Merkmale aus allen Originalarbeiten wird nicht vorausgesetzt. Wo die vorliegenden Passagen eine solche Einordnung nicht tragen, bleibt der betreffende Vergleich ausdrücklich unbestimmt.[20] Die ausdrückliche menschliche Entscheidung zur Forschungsrichtung priorisierte methodische Grenzen und Evidenzlücken sowie die Trennung empirischer Berichte von theoretischen Annahmen und unbestätigten Hypothesen. Damit wurde der Schwerpunkt der weiteren Untersuchung gewählt, nicht die Gültigkeit einer Hypothese oder die Überlegenheit eines Verfahrens festgestellt. Die Diskussion ist entsprechend nach Messproblemen, Ursacheninterpretationen, Heterogenität, Prüfverfahren und Kosten geordnet. Die Limitationen bestimmen die Grenzen der verfügbaren Information; das Fazit trennt

den eingegrenzten Erkenntnisstand von sechs konkreten Prüfaufgaben. Diese Entscheidung verändert weder die Herkunft der Aussagen noch den offenen Charakter widersprechender Bewertungen. Insbesondere werden aus der Synthese keine eigenen Erkennungs-, Korrektur- oder Nettowirkungsraten gewonnen. Die abschließenden Empfehlungen sind Anforderungen an spätere kontrollierte Untersuchungen. Ihre Ausführbarkeit, die Güte der Referenzkriterien und der zusätzliche Ressourcenbedarf müssten jeweils erst geprüft werden. Die Schwerpunktsetzung erlaubt damit eine klare Forschungsagenda, ohne einen wissenschaftlichen Ausgang der noch offenen Vergleiche vorwegzunehmen.[21]

4 Diskussion

UNGELÖSTER KONFLIKT 1 (ERNEUT AUFGEGRIFFEN)

Der erste ausdrücklich dokumentierte Konflikt betrifft die Schlussfolgerung aus der Verschiedenheit der Studienbefunde. Die ursprüngliche Position deutete falschen Konsens als nicht einheitlichen, sondern bedingungsabhängigen Effekt. Stützend wurden die Literaturgegenüberstellung und hohe Modaladoption in homogenen kleinen Gruppen angeführt; Wechsel ohne Peer-Information schränkten die Interpretation zusätzlich ein. Der zentrale Gegeneinwand lautet jedoch, dass Modaladoption, Yield, Oracle-Lücke und Wechselraten unterschiedliche Konstrukte mit verschiedenen Nennern erfassen. Ihre Verschiedenheit kann Messinkommensurabilität anzeigen, ohne eine bestimmte Bedingungsabhängigkeit desselben Phänomens zu identifizieren. Die Bewertung bleibt deshalb gemischt und ausdrücklich konflikthaft. Die Revisionen übernehmen den Einwand teilweise und verengen die Aussage darauf, dass Konvergenz und Antwortwechsel allein weder Korrektheit noch schädlichen Peer-Einfluss bestimmen. Diese Eingrenzung löst die feldweite Vergleichbarkeitsfrage nicht. Weder ist gezeigt, dass alle Arbeiten dasselbe Phänomen messen, noch dass ihre Bezeichnungen getrennte kausale Prozesse abbilden. Ein gemeinsames Design mit Erstantworten, kontrollierter Peer-Exposition und vorab klassifizierten Fehlertypen bleibt eine offene Voraussetzung weitergehender Schlussfolgerungen. [1, 2][22]

Die Kontrolllogik von Not All Flips Are Conformity begrenzt insbesondere die Interpretation zeitlicher Antwortveränderungen. Selbstreflexion ohne fremde Information erfasst eine Vergleichsgröße, gegenüber der zusätzliche Übernahme nach Peer-Sichtbarkeit betrachtet werden kann. Da die Übernahmen ausdrücklich Korrekturen und schädliche Übergänge enthalten, ist ihre bloße Verringerung kein hinreichendes Erfolgskriterium. Im primären Setting überwiegen bei verfügbarer richtiger Peer-Antwort nach Berücksichtigung der Peer-Unterstützung die Korrekturen gegenüber schädlicher Übernahme. Zugleich dämpft eine antwortagnostische Intervention beide Übergangsrichtungen ohne nachweisbaren Genauigkeitsgewinn. Diese Kombination spricht gegen eine pauschale Gleichsetzung von Anpassung und Schaden, nicht gegen jede mögliche Schutzintervention. Vollständige Begründungen erhöhen laut Abstract die individuelle Genauigkeit, liefern aber gegenüber Selbstreflexion keinen zusätzlichen Mehrheitsgewinn. Damit müssen der Informationsbeitrag für einzelne Modelle und sein Beitrag zur Gruppenentscheidung getrennt bewertet werden. Für offene Aufgaben bleibt zusätzlich zu klären, ob Referenzurteile die Übergänge ähnlich zuverlässig unterscheiden können. Die vorliegende Untersuchung mit bekannter Lösung beantwortet diese Übertragungsfrage nicht. [2][23] Die Maximalwerte aus The Cost of Consensus besitzen einen anderen Aussagebereich. Der Abstract untersucht Gruppen von zehn homogenen Qwen2.5-7B-, Llama-3.1-8B- oder Ministral-3-8B-Modellen über drei Runden auf zwei schwierigen Benchmarks. Als Kontrollen dienen isolierte Selbstkorrektur und sachfremde Begründungen. Modaladoption bis 85,5 Prozent, Vulnerabilität zuvor richtiger Antworten bis 70 Prozent und eine Oracle-Lücke bis 32,3 Prozentpunkte beschreiben verschiedene Fehlerpfade. Weder ihre Ad-

dition noch ihr Vergleich mit den bedingten Übernahmeraten der GPT-4o-Untersuchung wäre eine sinnvolle Gesamtrate. Die ausgewählte BenchForm-Passage berichtet ergänzend einen Anstieg der als CRT bezeichneten Konformitätsrate von 33,9 auf 44,4 Prozent bei einer gegenüber fünf Runden. Auch dieser Vergleich begründet keine allgemeine Dosis-Wirkungs-Beziehung für Debattendauer. Er zeigt vielmehr, welche Durchführungsmerkmale ein späterer Vergleich dokumentieren müsste. Ausgangskorrektheit, Modellklasse, Rundenzahl und Aggregation sind gemeinsam zu betrachten, bevor aus einem beobachteten Endfehler auf schädliche Interaktion geschlossen wird. [1, 8][24] Ein zusätzliches Messproblem liegt auf der Seite der Bewertung. Die ausgewählte Passage aus The Evaluator Is Part of the Experiment beschreibt bei unverändertem Antworttext modellabhängige Reaktionen auf Peer-Kontext: Urteile verschieben sich zur unterstützten Position hin, von ihr weg oder ungefähr gar nicht. Die antisymmetrische Spezifikation erfasst dabei gerichtete Kontextsensitivität, nicht jede mögliche Differenz zwischen verblindeter und informierter Bewertung. Eine bedingungsunabhängige Änderung allgemeiner Bewertungsstrenge wird ausdrücklich nicht separat identifiziert. Auch die Ankerprüfung bleibt relevant: Ohne Referenz werden richtige Anker bei 73,3 Prozent der Fragen höher als falsche eingeordnet, mit Goldreferenz bei 97,5 Prozent; die mittlere Trennung steigt von 1,59 auf 3,30 Punkte. Die Interpretation ist entsprechend begrenzt. Referenzzugang verbessert den berichteten Vergleich, garantiert aber keine fehlerfreie Bewertung komplexer Debattenantworten. Wenn Peer-Information zugleich Antwortproduktion und Urteil beeinflusst, müssen beide Einflüsse auseinandergehalten werden. Andernfalls kann eine Veränderung des Messinstruments als Veränderung der sachlichen Antwortqualität erscheinen. [9][25] Bei den Ursacheninterpretationen wird zunächst die vereinfachende Zuschreibung an RLHF-induzierte Sykophanz herausgefordert. Not Just RLHF berichtet über vier Modellfamilien dasselbe Substitutionsmuster bei vortrainierten Basismodellen wie bei Instruct-Varianten. Das spricht im untersuchten Setting gegen eine ausschließlich nachgelagert durch Alignment erzeugte Vulnerabilität. Der Abstract und die ausgewählte Volltextstelle beschreiben zudem Aktivierungs-Patching, dessen mechanistische Interpretation an Llama-3.1-8B-Instruct gebunden bleibt. Diese lokale Untersuchung darf nicht ohne Weiteres auf andere Modelle oder ganze Debattenprotokolle übertragen werden. Ein weiterer Abstract zu Nachgeben unter Nutzerdruck verbindet das Verhalten stärker mit Prüfkosten und Guardrail-Abdeckung als mit Drucktaktiken. Das Entfernen des Aufgabenfaktors kostet dort 0,485 McFadden-R^2 gegenüber 0,139 für Modellfamilie und 0,009 für Drucktaktik. Diese Größen sind keine kausalen Anteile einer gemeinsamen Erklärung falschen Konsenses. Sie stammen aus einem anderen Aufgaben- und Interaktionsraum. Zusammen begrenzen die Berichte eine reine RLHF-Erzählung, ohne die relative Bedeutung aller möglichen Ursachen abschließend zu bestimmen. [4, 10][26] Die Studie zur Meinungsdynamik bietet eine weitere, nicht unmittelbar austauschbare Lesart. Ihr Abstract beschreibt Konvergenz „not through sycophancy or blind conformity“, also nicht durch Sykophanz oder blinde Anpassung, sondern durch strukturierte asymmetrische Überzeugung. Zugleich werden die selektive Übernahme anschlussfähiger Positionen und ein messbarer Einfluss logischer Fehlschlüsse berichtet. Selektivität ist damit nicht gleichbedeutend mit epistemischer Angemessenheit. Der Vergleich mit hohen Mehrheitsübernahmen bei lösbaren Faktenaufgaben bleibt aber begrenzt, weil Meinungsentwicklung andere Referenzkriterien verwendet. Die berichteten Unterschiede können deshalb weder als unmittelbarer Widerspruch identischer Experimente noch als bereits aufgeklärte Aufgabeneffekte gelten. Ähnlich vorsichtig ist die Unterscheidung gemeinsamer Ausgangsfehler von späteren Peer-Effekten zu verwenden. Sie macht zeitliche Verläufe sichtbar, weist jedoch keine Trainingsdatenursachen nach und schließt Wechselwirkungen nicht aus. Für eine Ursachenprüfung müssten Framing, Argumentqualität, Informationszugang und Korrektheitskriterium im selben Design getrennt werden. Die vorhandenen Bezeichnungen liefern hierfür Forschungsfragen, aber keine gemeinsame kausale

Zerlegung. [11][27] Die Analogie von Ansteckung und korreliertem Ausfall hilft bei der Formulierung möglicher Messgrößen, bleibt jedoch heuristisch. Ihr Nutzen besteht darin, gespeicherte Erstantworten von Veränderungen nach Kontakt zu unterscheiden und Reihenfolge, Sichtbarkeit sowie Rundenzahl als Variationen zu benennen. Die Gegenkritik betont, dass Antwortänderungen von Prompt, Argumentqualität und Kontext abhängen und nicht durch eine stabile Übertragungswahrscheinlichkeit beschrieben werden müssen. Zudem können über denselben Austausch Korrekturen und Fehler vermittelt werden. Das Netzwerkpapier im Vergleichsbestand behandelt diese Problematik unter expliziten Modell- und Budgetannahmen. Bei fester Exposition je Kommunikationskante ergeben sich andere Dichtefolgen als bei festem Senderbudget; unter letzterem ist die homogene Schwelle erster Ordnung unabhängig von der Netzwerkdichte. Simulationen und das begrenzte Experiment mit sechs Positionen erweitern den Kontext, validieren aber nicht die gesamte Analogie für beliebige Debatten. Eine Aussage wie mehr Vernetzung verursache generell mehr falschen Konsens würde daher die entscheidende Budgetkonvention und die Korrekturmöglichkeit ausblenden. [12][28] Bei der Modellheterogenität stehen unterschiedliche Aufgaben, Kontrollen und Zielgrößen besonders deutlich nebeneinander. Stop Overvaluing untersucht laut Abstract fünf Debattenverfahren, neun Benchmarks und vier Basismodelle und bezeichnet Heterogenität als „universal antidote“, sinngemäß als universelles Gegenmittel. Beyond Symmetric Agents variiert dagegen bei 23 kleinen offenen Modellen aus elf Anbieterfamilien auf fünf Aufgaben Personas, Temperatur und Modellidentität. Jede Konfiguration wird einer generierungsbudgetangepassten Mehrheitskontrolle gegenübergestellt. Gemischte Teams unterliegen dort Mehrheitsentscheidungen derselben Besetzung. Die kritische Lesart begrenzt beide Schlussfolgerungen: Der negative Vergleich betrifft zunächst den Zusatznutzen des untersuchten Debattenprotokolls gegenüber seiner Aggregationskontrolle, nicht kausale Wirkungslosigkeit heterogener Zusammensetzung. Umgekehrt trägt die positive Universalitätsformulierung keine Garantie über Aufgaben, Fähigkeiten und Budgets hinweg. Die Revisionen lassen den Gegensatz ausdrücklich offen. Fähigkeitsgematchte Zusammenstellungen mit gezielt variierter Modellheterogenität wären eine Anforderung an spätere Tests, nicht ein bereits vorliegendes Ergebnis. Ohne diese Kontrolle können Verschiedenheit der Modelle und Unterschiede ihres Leistungsniveaus nicht sicher auseinandergehalten werden. [13, 14][29] Die Knight–Knave–Spy-Untersuchung bezeichnet individuelle Schlussfolgerungsstärke und Gruppenvielfalt als dominante Erfolgsfaktoren in einem Logiksetting mit verifizierbarer Lösung. Der Abstract erlaubt jedoch keine sichere Rekonstruktion eines isolierten Diversitätsbeitrags gegenüber Fähigkeit und budgetgleichem Sampling. Ein separates Audit von Mehrheitsensembles prüft Diversitätsmaße über 31.900 Modellteilmengen unter expliziten Fähigkeitskontrollen. Rohe Zusammenhänge zwischen Diversität und Gewinn sind dort stark mit Fähigkeit verflochten und unter Kontrolle überwiegend instabil; als begrenzter Rest bleibt eine Assoziation gemeinsamen Fehlverhaltens mit geringerem Gewinn. Da dieses Audit keine Debattenmechanismen untersucht, ersetzt es den benötigten Debattenversuch nicht. Es zeigt aber, warum Modellidentität, Antwortverschiedenheit, Fehlerkomplementarität und Mitgliederfähigkeit nicht austauschbar sind. Der Bericht zu Entscheidungsprotokollen nennt zusätzlich Vorteile unabhängiger Lösungserzeugung. Auch daraus folgt keine universelle Wirkung heterogener Modellfamilien. Für spätere Studien sollten diese Achsen getrennt erfasst werden. Andernfalls kann eine günstige Zusammensetzung als Diversitätseffekt erscheinen, obwohl der konkrete Vergleich vor allem andere Mitgliederfähigkeit oder andere Aggregation widerspiegelt. [15, 16, 17][30] Unter den Prüfverfahren adressiert AgentAuditor besonders direkt die Auswahl aus vorhandenen Kandidaten. Begründungen werden in einem Reasoning Tree angeordnet und an kritischen Verzweigungen evidenzbezogen geprüft, statt allein Antwortfrequenzen zu zählen. Der Abstract nennt Gewinne bis zu fünf absoluten Genauigkeitspunkten. Davon zu unterscheiden sind die Volltextwerte

für MinC: 65,35 Prozent korrekte Entscheidungen auf GSM8K und 81,82 Prozent auf AMC. Dieses Teilset enthält definitionsgemäß eine falsche Mehrheit und mindestens eine richtige Minderheitslösung. Die null Prozent der Mehrheitswahl folgen daher aus der Auswahlbedingung und sind kein allgemeiner Baselinewert. Ebenso beschreiben die hohen Wiedergewinnungswerte keine allgemeine Fehlerkorrekturrate. Über Kandidatenpools, in denen alle Antworten falsch sind, sagen sie nichts aus. Der Befund ist dennoch für das Design relevant: Verfügbarkeit einer richtigen Lösung und Fähigkeit zur Auswahl müssen separat geprüft werden. Eine Aggregationsverbesserung kann vorhandene Information besser nutzen, ohne die zusätzliche Wissensgewinnung zu leisten, die bei vollständig falschen Ausgangsantworten erforderlich wäre. [3][31] MARCH und Chain-of-Verification verändern dagegen den Informationszugang einzelner Prüfschritte. MARCH zerlegt eine erzeugte Antwort in überprüfbare atomare Propositionen; der Checker bewertet diese anhand abgerufener Evidenz ohne Zugriff auf den ursprünglichen Solver-Text. Der Abstract verbindet diese Informationsasymmetrie mit gemeinsamem Multi-Agent-Reinforcement-Learning und berichtet verringerte Halluzinationen. Die Aussage, dadurch werde der Selbstbestätigungszyklus gebrochen, ist eine Interpretation der Quelle; der isolierte kausale Anteil der Textabschirmung wird im Auszug nicht ausgewiesen. Chain-of-Verification lässt geplante Verifikationsfragen unabhängig beantworten und berichtet weniger Halluzinationen auf mehreren Aufgabenformaten. Beide Ansätze motivieren Komponentenprüfungen von Verblindung und unabhängiger Bearbeitung. Sie liefern in den gelesenen Passagen aber keine allgemeine Messung falscher Mehrheitsentscheidungen in offenen heterogenen Debatten. Besonders wichtig bleibt die begriffliche Grenze: Kein Zugriff auf eine Ausgangsantwort bedeutet nicht automatisch unabhängiges Modellwissen oder authentische und richtige Retrievalquellen. Der Transfer von Halluzinationsreduktion zu Konsensfehlerkontrolle benötigt deshalb zusätzliche, auf diese Zielgröße zugeschnittene Vergleiche. [6, 18][32] BLUEmed kombiniert nochmals andere Elemente: Teilfragen zu klinischen Notizen, quellenpartitioniertes Retrieval, getrennte Wissensbasen, unabhängige Fachanalysen, Gegenargumentation bei Dissens, quellenübergreifende Entscheidung und Sicherheitsfilter. Unter Few-shot-Prompting nennt der Abstract 69,13 Prozent Genauigkeit, 74,45 Prozent ROC-AUC und 72,44 Prozent PR-AUC auf einem Terminologiesubstitutionsbenchmark. Diese Kennzahlen betreffen die Erkennung sprachlich plausibler, klinisch aber unterschiedlicher Ersetzungen. Sie sind weder mit MinC-Genauigkeiten noch mit Yield-Senkungen zu einer gemeinsamen Schutzrate zusammenzuführen. Die Vorteile gegenüber Einzelmodell-RAG und Debatte ohne Retrieval gelten dem berichteten Gesamtverfahren. Aus dem Auszug lässt sich keine Rangordnung der Einzelkomponenten ableiten. Quellenübergreifende Entscheidung ist außerdem nicht gleichbedeutend mit einer Prüfung gemeinsamer Informationsursprünge. Das in H3 formulierte Gegenmodell bleibt daher ein offener Prüfgegenstand, ohne den klinischen Gesamtbefund zu entwerten. Eine spätere Untersuchung müsste Dokumentenzahl, Herkunftsabhängigkeit und Richtigkeit getrennt behandeln. Nur so ließe sich klären, ob zusätzliche Dokumente tatsächlich unabhängige Gegeninformation liefern oder lediglich vorhandene Information wiederholen. [7][33] Für strukturierten Dissens ist die Qualität des Einwands die entscheidende noch offene Übertragungsbedingung. Not Just RLHF berichtet, dass ein einzelner korrekt argumentierender Dissenter Yield über die getesteten Framings um 54 bis 73 Prozentpunkte senkt; die ausgewählte Volltextstelle bestätigt diese beschränkte Formulierung. Zugleich wird unzureichender Transfer der stärksten Prompt-Verteidigung außerhalb ihrer Designfläche beschrieben. Der Befund trägt keine generische Aussage, jede Oppositionsrolle sei wirksam. Ebenso wenig begründet er die ursprüngliche Vermutung symmetrischen Schadens durch falsche Gegenargumente. Die Kritik trennt deshalb den beobachteten Effekt korrekten Dissenses von der ungeprüften Qualität realistisch erzeugter Opposition. Die revidierte Hypothese belässt Destabilisierung richtigen Konsenses als mögliches Gegenmodell und hält auch Fehleraufdeckung ohne zusätz-

liche Evidenz offen. Ein späterer Vergleich müsste Rollenauftrag, relevante Gegeninformation und unabhängige Qualitätsprüfung getrennt variieren. Dabei wären richtige und falsche Revisionen gesondert auszuweisen. Die vorgegeben korrekte Gegenposition wäre eine informationsprivilegierte Vergleichsbedingung, nicht eine ohne Weiteres verfügbare Standardlösung. [4][34] Informationen über Peer-Zuverlässigkeit bilden einen weiteren Ansatz, dessen Verhaltenswirkung nicht vorschnell einem einzigen Mechanismus zuzuschreiben ist. Too Polite to Disagree untersucht sechs offene Modelle mit statisch oder dynamisch geschätzten Rangfolgen der Sykophanzneigung. Der Abstract berichtet geringeren Einfluss entsprechend eingestufter Peers, weniger Fehlerkaskaden und eine absolute Verbesserung finaler Diskussionsgenauigkeit um 10,5 Prozentpunkte. Dies ist ein interventionsbezogener Hinweis auf veränderte Informationsgewichtung, isoliert anhand des Abstracts aber nicht Sykophanz gegenüber allen anderen Erklärungen. Die ausgewählte ECL-Passage beschreibt ergänzend Robustheitsgewinne durch eine separat erzeugte eigene Begründung bei ausschließlich falschen Peers; die Leistung bleibt deutlich unter den Standardbedingungen. Historische Vertrauensinformationen sind damit kein Ersatz für aktuelle Prüfung. Für den vorgeschlagenen adaptiven Ansatz bleibt offen, wie solche Gewichte auf neue Aufgaben übertragen und auf plötzliche Zuverlässigkeitswechsel reagieren. Auch ihre Erzeugungs- und Aktualisierungskosten gehören zum Vergleich. Der durchschnittliche berichtete Genauigkeitsgewinn beantwortet diese Fragen nicht und liefert noch keinen Test der in H2 gebündelten Strategie. [5, 19][35]

UNGELÖSTER KONFLIKT 2

Der zweite ausdrücklich dokumentierte Konflikt betrifft die Vorhersage interventionsspezifischen Nettonutzens. Stützende Argumente betonen die Trennung von Risikodiagnose und Korrigierbarkeit sowie die Bedeutung verfügbarer, verlässlicher Gegeninformation. Der stärkste Einwand richtet sich gegen die Zielgröße: Für dieselbe Entscheidung sind Verläufe mit und ohne Eingriff nicht zugleich direkt beobachtbar. Werden Risikomerkmale ungeprüft auch zur Nutzenvorhersage verwendet, kann eine Regel an Rauschmustern kalibriert werden. Die Passage zum Freeze-Replay illustriert die mögliche Diskrepanz: Ein probe-gesteuertes Einfrieren verhindert 29 Zusammenbrüche, verwirft jedoch 108 Korrekturen und erreicht bei gleichen Gewichten einen Nettonutzen von minus 79. Dieser Bericht kontextualisiert die Hypothese, bestätigt aber keinen alternativen Nutzenprädiktor. Die Revisionen übernehmen den kontrafaktischen Einwand und schlagen randomisierte Eingriffsvergleiche, präinterventionelle Merkmale, getrennte Kalibrierung und Tests auf zurückgehaltenen Aufgaben und Modellfamilien vor. Damit wird die Prüfaufgabe präziser, nicht empirisch gelöst. Die Bewertung bleibt gemischt; die Hypothese ist weiterhin unbestätigt und hoch unsicher. [20][36]

WEITERER DISSENS (NICHT ALS EXPLIZITER KONFLIKT ERFASST)

Eine zusätzliche dokumentierte Unstimmigkeit betrifft die Bewertung der ursprünglichen Zwei-Signal-Hypothese. Sie wird im ausgewerteten Kontext als herausgefordert beurteilt, während dieselbe Bewertung keinen ausdrücklichen Konflikt verzeichnet. Diese Abweichung bleibt bestehen und wird nicht nachträglich vereinheitlicht. Inhaltlich richtet sich die Herausforderung gegen die vorausgesetzte Trennbarkeit: Peer-blinde Einigkeit kann korrekt sein, spätere Angleichung kann legitime Revision abbilden, und die Unabhängigkeit externer Evidenz ist nicht abgesichert. Die Revisionen nehmen deshalb sowohl die behauptete Diagnosefähigkeit als auch die budgetgleiche Überlegenheit ohne zusätzliche Korrekturverluste zurück. Weitergeführt wird nur eine unkalibrierte Designhypothese. Die Herausforderung ist somit keine experimentelle Falsifikation, und die fehlende ausdrückliche Konfliktkennzeichnung ist keine Entkräftung des Einwands. Für das weitere Design bleibt entscheidend, Klassifikationsgüte und Eingriffsnutzen getrennt zu prüfen. Referenzlösungen

> dürften dabei die nachträgliche Bewertung ermöglichen, aber nicht verdeckt in einen operativen Auslöser eingehen. Die spätere eingeschränkte Bewertung liefert hierfür lediglich Kontext beziehungsweise Qualifizierung, keine direkte Bestätigung.[37]

Die Nutzengewichtung entscheidet mit darüber, welche Handlung überhaupt als vorteilhaft erscheint. Die negative Bilanz des probe-gesteuerten Freeze-Replays gilt unter gleicher Gewichtung verhinderter Zusammenbrüche und verlorener Korrekturen. Für eine andere, feste Regel anhand einer Mehrheitsänderung in der ersten Runde nennt die Passage eine annähernde Gewinnschwelle, wenn ein Zusammenbruch vorab mindestens 1,23-mal so stark gewichtet wird wie eine entgangene Korrektur. Diese Schwelle darf nicht auf die probe-gesteuerte Regel übertragen werden. Regeldefinition und Gewichtung müssen gemeinsam offengelegt werden; eine nachträglich gewählte günstige Gewichtung beantwortet nicht denselben Vergleich wie eine vorab begründete. Zudem handelt es sich um gespeicherte Verläufe. Eine laufende Intervention kann die spätere Kommunikation verändern und verlangt deshalb zusätzliche Annahmen oder direkte Tests. Risikoeinstufung vor Auswahl, Diagnose während der Debatte, Nutzengewichtung und Handlung bleiben getrennte Evaluationsgegenstände. H2 adressiert gerade diese Trennung, ohne bereits einen leistungsfähigen Nutzenprädiktor bereitzustellen. Das Material motiviert die Prüfung, bestimmt aber nicht deren Ausgang. [20][38] Schließlich muss zusätzlicher Aufwand von spezifischem Debattennutzen getrennt werden. Bei den homogenen kleinen Modellen benötigt Debatte laut Abstract 2,1- bis 3,4-mal mehr Tokens als Selbstkorrektur bei gleicher oder geringerer Genauigkeit. Not All Flips Are Conformity nennt 57 Prozent mehr Tokens für vollständige Begründungen ohne zusätzlichen Mehrheitsgewinn gegenüber Selbstreflexion. Beyond Symmetric Agents berichtet bei budgetangepassten Bedingungen Gleichstand oder Nachteile gegenüber Self-Consistency sowie das 1,6-Fache der Laufzeit und das 3,4-Fache der Tokens. Dort verändert zudem die Korrektur unbemerkter Kontextfensterüberschreitungen den Vergleich von minus 1,8 Punkten auf Gleichstand. Diese Angaben sind konfigurationsgebunden und keine allgemeinen Kostenfaktoren sämtlicher Debattenverfahren. Die ältere Vergleichsstudie Should we be going MAD? ergänzt fehlende zuverlässige Überlegenheit gegenüber Self-Consistency und Ensembling, aber auch Vorteile abgestimmter Konfigurationen und Hyperparametersensitivität. Spätere Vergleiche sollten deshalb unabhängiges Sampling, Selbstkorrektur und Debatte ohne zusätzliche Prüfung ernsthaft einbeziehen. Retrieval, Quellenprüfung und gegebenenfalls Training gehören getrennt in die Ressourcenbilanz, bevor adaptiver Prüfung ein Mehrwert zugeschrieben wird. [1, 2, 14, 21][39]

5 Limitationen

Die wichtigste Grenze betrifft die Tiefe der ausgewerteten Information. Überwiegend Abstracts und einzelne Volltextstellen erlauben keine vollständige Prüfung von Stichprobenauswahl, Unsicherheitsintervallen, Ausfällen, statistischen Annahmen und sämtlichen Kontrollbedingungen. Ausgewählte Passagen können einzelne Operationalisierungen präzisieren, ersetzen aber nicht die Originalmethodik und deren Nachprüfung. Fehlende Angaben im vorliegenden Material bedeuten deshalb nicht, dass die ursprünglichen Studien entsprechende Vorkehrungen unterlassen haben. Umgekehrt darf eine weitreichende Abstractformulierung nicht als bereits abgesicherte Reichweite übernommen werden. Die Synthese rekonstruiert berichtete Befunde und deren strittige Lesarten, ohne sie unabhängig zu replizieren. Auch der unterschiedliche Publikationsstand verlangt eine spätere gesonderte Bewertung. Eine qualitative Einordnung ist unter diesen Bedingungen möglich, eine verlässliche quantitative Gesamtschätzung jedoch nicht. Die erhaltenen Zahlen werden daher ausschließlich mit ihren jeweiligen Bedingungen beschrieben. Diese Einschränkung gilt auch dort, wo mehrere Beiträge dieselbe Passage ähn-

lich interpretieren: Wiederholte Interpretation erhöht nicht die Zahl unabhängiger Beobachtungen und gleicht fehlende Primärdaten nicht aus.[40] Die externe Validität bleibt durch enge Aufgaben- und Modellräume begrenzt. Mehrere zentrale Berichte betreffen kleine homogene Gruppen, bestimmte Familien oder Benchmarks mit bekannter Referenzlösung. Gerade die bekannte Lösung ermöglicht die Trennung richtiger und falscher Revisionen, macht die Übertragung auf offene wissenschaftliche Fragen aber nicht automatisch gültig. Klinische Terminologiesubstitution, mathematische Antwortauswahl und logische Rätsel besitzen unterschiedliche Referenzkriterien und Fehlerstrukturen. Ein Erfolg in einem dieser Räume begründet keine allgemeine Erkennungsleistung in den anderen. Ebenso bleibt die Verfügbarkeit eines korrekt argumentierenden Dissenters außerhalb informationsprivilegierter Versuche ungeklärt. Die Fähigkeit, vorgegebene richtige Gegeninformation zu nutzen, ist von ihrer eigenständigen Erzeugung zu unterscheiden. Der Vergleich erlaubt deshalb keine belastbare Rangordnung nach Modellgröße, Domäne, Sprache oder Offenheit der Aufgabe. Auch die Kritik an einem Nullbefund kleiner Modelle darf nicht in eine positive Behauptung über größere heterogene Modelle umgedeutet werden. Beide Übertragungsrichtungen benötigen zusätzliche Daten und dürfen nicht durch Plausibilität ersetzt werden.[41] Die Konstruktvalidität ist durch nicht harmonisierte Zielgrößen eingeschränkt. Individuelle Modaladoption, Yield, Oracle-Lücke und falsche Endmehrheit können nicht als dieselbe Messung behandelt werden. Ohne gemeinsame Nenner und vollständige Übergangstabellen bleibt eine quantitative Zusammenführung unzulässig. Auch die zeitliche Ordnung gemeinsamer Ausgangsfehler und später entstandener Fehlentscheidungen ist keine vollständige Ursachenklassifikation. Gleiche falsche Antworten identifizieren weder Wissenslücken noch Datenüberlappung; spätere Angleichung identifiziert nicht automatisch Ansteckung. Hinzu kommt die Unsicherheit des Bewertungsinstruments bei offenen Antworten. Referenzgestützte Ankerkalibrierung verbessert den berichteten Vergleich, sichert aber weder jede komplexe Bewertung noch die Angemessenheit aller Schadensrubriken ab. Verblindung und Kontextzugang der Urteile müssen daher unabhängig vom eigentlichen Debattenprotokoll geprüft werden. Für die vorgeschlagenen Fehlerklassen fehlen zudem eigene Daten zur Zuverlässigkeit der Annotation. Die inhaltliche Präzisierung dieser Klassen ist eine Voraussetzung eines späteren Tests, nicht dessen Ergebnis. Solange gemeinsame Beobachtungen fehlen, können widersprechende Interpretationen nicht allein durch begriffliche Vereinheitlichung entschieden werden. [9][42] Bei den Prüfverfahren fehlen insbesondere isolierte Komponentenvergleiche und direkte Tests von Quellenunabhängigkeit. Verbesserungen kombinierter Verfahren lassen sich anhand der Auszüge nicht sicher Verblindung, Retrieval, Gegenargumentation, Sicherheitsfiltern oder gemeinsamem Training zuschreiben. Mehrere Dokumente oder getrennte Wissensbasen sind außerdem keine bereits validierten Maße unabhängiger Informationsursprünge. Das Gegenmodell einer gemeinsam übernommenen falschen Ursprungsquelle bleibt unbestätigt; weder seine Häufigkeit noch sein Schaden wurde hier gemessen. Authentizität, Belegfunktion und sachliche Richtigkeit stellen verschiedene Anforderungen dar. Ein authentisches Dokument kann eine Behauptung unzureichend stützen, und passende Dokumente können voneinander abhängig sein. Daraus folgen Operationalisierungsaufgaben, keine gemessenen Versagensraten der besprochenen Systeme. Ebenso fehlen ausreichende Daten zur Qualität realistisch erzeugter Einwände und zum Aufwand ihrer unabhängigen Prüfung. Die vorliegenden Bewertungen können H3 und H4 deshalb nicht empirisch bestätigen oder zurückweisen. Auch schärfere Versuchsbedingungen beseitigen diese Lücke erst dann, wenn sie tatsächlich untersucht und ihre Vergleichsbedingungen transparent ausgewiesen werden.[43] Die Nutzenvorhersage adaptiver Prüfung ist durch das kontrafaktische Ziel und fehlende eigene Testdaten besonders begrenzt. Für die vorgeschlagenen Strategien liegen keine hier erhobenen Erkennungs-, Korrektur- oder Nettowirkungsraten vor. Eine Uplift-Schätzung bleibt ein Designvorschlag, solange kontrollierte Eingriffsvergleiche und ge-

trennte Prüfungen auf anderen Aufgaben und Modellfamilien fehlen. Der Replay-Befund begründet die Berücksichtigung verlorener Korrekturen, lässt sich aber nicht ohne zusätzliche Annahmen auf laufende Eingriffe übertragen. Auch ein einheitliches Ressourcenbudget ist nicht bestimmt. Tokens erfassen weder Latenz noch Retrieval-, Quellenprüfungs- oder Trainingsaufwand vollständig; unterschiedliche Budgetkonventionen können den Vergleich verändern. Zudem muss das Verhältnis vermiedener Fehler und entgangener Verbesserungen vorab begründet werden. Ohne diese Festlegungen bleibt die behauptete Budgetgleichheit mehrdeutig. Eine vollständig beschriebene Bewertungsregel ist deshalb noch kein Beleg ihres Nutzens. Die Unsicherheit betrifft sowohl die Vorhersagbarkeit durchschnittlicher Eingriffseffekte als auch ihre Übertragbarkeit und die Kosten, die nötig wären, um geeignete Merkmale überhaupt rechtzeitig zu erheben.[44] Auch die Zusammensetzung der kritischen Synthese begrenzt ihre Reichweite. Sämtliche vorgesehenen Beiträge wurden erhalten; eine wegen ausgefallener Beteiligung fehlende Perspektive ist nicht dokumentiert. Dennoch ersetzt ein einzelner strukturierter Kritikbeitrag keine breite unabhängige Begutachtung, und vollständige Beteiligung an einem begrenzten Design bedeutet keine vollständige Abdeckung möglicher Gegenpositionen. Die Überschneidungsprüfung betrifft ausschließlich die gesammelten Positionen und etabliert keine globale Neuheit. Die menschliche Schwerpunktsetzung auf methodische Grenzen bestimmt die Forschungsrichtung, nicht den Ausgang offener wissenschaftlicher Vergleiche. Ebenso wenig sind Bewertungen als unzureichend kontextgestützt oder unbestimmt negative experimentelle Ergebnisse. Eine im jeweiligen Kontext herausgeforderte Hypothese ist nicht allgemein falsifiziert. Die beiden ausdrücklich dokumentierten Konflikte bleiben deshalb offen, während weitere Einwände als Grenzen der jeweiligen Lesart erhalten werden. Zusätzliche Literaturprüfung, unabhängige Replikation, methodische Begutachtung und experimenteller Eingriffsvergleich sind unterschiedliche nächste Schritte. Keiner dieser Schritte kann allein durch die präzisere Formulierung einer Hypothese oder durch wiederholte Zustimmung zu derselben Literaturpassage ersetzt werden.[45]

6 Schlussfolgerung

Der eingegrenzte Erkenntnisstand betrifft vor allem Messung und Vergleichslogik. Die gelesenen Studien berichten schädliche Antwortverdrängung, Verlust richtiger Minderheitslösungen und settingspezifische Verbesserungen durch Prüfverfahren. Zugleich zeigen sie, dass Antwortwechsel sowohl Korrekturen als auch neue Fehler umfassen können. Übereinstimmung allein ist deshalb weder Qualitätsmaß noch ausreichender Interventionsauslöser. Die vier Hypothesenfamilien ordnen offene Prüfaufgaben, begründen aber keine allgemeine Überlegenheit adaptiver Prüfung, Quellenvalidierung oder strukturierten Dissenses. Der methodische Ertrag liegt in der Trennung bislang vermischter Fragen: Welche Entscheidung ist falsch, welche Information verändert sie, kann der konkrete Eingriff helfen, und welche Korrekturen oder Ressourcen gehen dabei verloren? Die folgenden sechs Forschungslücken bezeichnen konkrete nächste Untersuchungen und sind von diesem begrenzten Erkenntnisstand zu unterscheiden.[46]

1. Offen ist eine gemeinsame, unabhängig referenzierte Messung der Fehlerverläufe. Ein nächster Versuch sollte peer-blinde Erstantworten speichern, Peer-Sichtbarkeit und Reihenfolge randomisieren und eine Kontrolle ohne Peer-Information vorsehen. Korrekturen, neue Fehler und Wechsel zwischen falschen Antworten wären vorab getrennt zu definieren. Zu prüfen wäre insbesondere, ob die Unterscheidung gemeinsamer Ausgangsfehler und späterer schädlicher Übergänge zuverlässig annotiert werden kann.[47]

2. Ungeklärt ist der eigenständige Beitrag von Modellheterogenität nach Kontrolle der Mitgliederfähigkeit und des Kommunikationsbudgets. Erforderlich wären fähigkeitsgematchte Zusammenstellungen mit variierter Modellidentität und gemessener Fehlerkorrelation. Debatte wäre mit unabhängiger Mehrheitsaggregation derselben Besetzung und budgetangepasstem Sampling zu vergleichen. Damit ließen sich Modellverschiedenheit, Fehlerkomplementarität und zusätzlicher Debattennutzen getrennt untersuchen, ohne positive und negative Abstractbefunde vorschnell zu vereinheitlichen.[48]
3. Offen bleibt, wie Quellenabhängigkeit unabhängig von Quellenrichtigkeit erfasst werden kann. Ein späterer Vergleich sollte gemeinsame Informationsursprünge bei kontrollierter Dokumentenzahl und vergleichbarer Retrievalqualität variieren. Authentizität, tatsächliche Belegfunktion und sachliche Richtigkeit wären separat zu bewerten. Erst damit wäre prüfbar, ob organisatorisch getrennte Verifikation zusätzliche unabhängige Information erschließt oder dieselbe fehlerhafte Grundlage mehrfach bestätigt.[49]
4. Der Nutzen realistischer Dissenter ohne Wahrheitswissen ist ungeklärt. Freie Prüfung und verpflichtende Opposition sollten jeweils mit und ohne zusätzliche relevante Evidenz sowie mit unabhängiger Qualitätsprüfung der Einwände verglichen werden. Richtige und falsche Revisionen und der Aufwand ihrer Erzeugung wären getrennt auszuweisen. Vorgegebene korrekte Gegenargumente blieben eine informationsprivilegierte Vergleichsbedingung und dürften nicht als bereits verfügbare Standardlösung behandelt werden.[50]
5. Zu prüfen ist, ob präinterventionelle Merkmale den durchschnittlichen Nutzen eines Eingriffs außerhalb der Kalibrierungsdaten vorhersagen. Randomisierte Eingriffsvergleiche sollten getrennte Kalibrierungs- und Testaufgaben sowie zurückgehaltene Modellfamilien verwenden. Adaptive Prüfung wäre unter vorab festgelegter Nutzengewichtung mit Konvergenzauslösern und budgetgleicher Zufallsprüfung zu vergleichen. Auch sofortige Einigkeit müsste stichprobenartig erfasst werden, damit gemeinsame Ausgangsfehler nicht außerhalb der bewerteten Prüfungsabdeckung bleiben.[51]
6. Offen bleibt die Belastbarkeit unter unabhängiger Evaluatorprüfung und vollständiger Ressourcenbilanz. Erforderlich wären verblindete Urteile, geprüfte Referenzanker, dokumentierte Kontextfenster und getrennte Angaben zu Tokens, Laufzeit, Retrieval und gegebenenfalls Training. Verbleibende Fehler im Gesamtsystem, neu verursachte Schäden und verlorene Korrekturen sollten gemeinsam erfasst werden. Anschließende Replikationen über Aufgaben, Domänen und Sprachen müssten zeigen, welche Schlussfolgerungen über die ursprünglichen Bedingungen hinaus Bestand haben.[52]

Literatur

1. *The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate*. 10.1145/3786335.3813137
2. *Not All Flips Are Conformity: Decomposing Stance Convergence in Multi-Agent LLM Debate*. arXiv (Cornell University). 10.48550/arxiv.2606.00820
3. *Auditing Multi-Agent LLM Reasoning Trees Outperforms Majority Vote and LLM-as-Judge*. arXiv (Cornell University). 10.48550/arxiv.2602.09341
4. *Not Just RLHF: Why Alignment Alone Won't Fix Multi-Agent Sycophancy*. arXiv (Cornell University). 10.48550/arxiv.2605.12991
5. *Too Polite to Disagree: Understanding Sycophancy Propagation in Multi-Agent Systems*. Annual Meeting of the Special Interest Group on Discourse and Dialogue. 10.18653/v1/2026.sigdial-1.56
6. *MARCH: Multi-Agent Reinforced Self-Check for LLM Hallucination*. arXiv (Cornell University). 10.48550/arxiv.2603.24579

7. *BLUEmed: Retrieval-Augmented Multi-Agent Debate for Clinical Error Detection*. IEEE International Conference on Healthcare Informatics (ICHI). 10.1109/ichi69079.2026.00112
8. *Do as We Do, Not as You Think: the Conformity of Large Language Models*. arXiv (Cornell University). 10.48550/arxiv.2501.13381
9. *The Evaluator Is Part of the Experiment: Measuring Open-Ended LLM Conformity*. arXiv (Cornell University). 10.48550/arxiv.2608.04463
10. *Beyond the Sycophancy Score: How Task, Model, and Pressure Shape LLM Yielding*. arXiv (Cornell University). 10.48550/arxiv.2610.08840
11. *Selective agreement, not sycophancy: investigating opinion dynamics in LLM interactions*. EPJ Data Science. 10.1140/epjds/s13688-025-00579-1
12. *Reliability-Contagion Feasibility in LLM Multi-Agent Networks*. arXiv (Cornell University). 10.48550/arxiv.2607.21912
13. *Stop Overvaluing Multi-Agent Debate -- We Must Rethink Evaluation and Embrace Model Heterogeneity*. arXiv (Cornell University). 10.48550/arxiv.2502.08788
14. *Beyond Symmetric Agents: Cognitive Diversity and Multi-Agent Debate in Small Language Models*. arXiv (Cornell University). 10.48550/arxiv.2609.35875
15. *Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning*. arXiv (Cornell University). 10.48550/arxiv.2511.07784
16. *Are Diversity Metrics Measuring Diversity? A Capability-Controlled Audit of Majority-Vote Gain in LLM Ensembles*. arXiv (Cornell University). 10.48550/arxiv.2607.20768
17. *Decision Protocols in Multi-Agent Large Language Model Conversations*. arXiv (Cornell University). 10.48550/arxiv.2607.05477
18. *Chain-of-Verification Reduces Hallucination in Large Language Models*. 10.18653/v1/2024.findings-acl.212
19. *Epistemic Context Learning: Building Trust the Right Way in LLM-Based Multi-Agent Systems*. arXiv (Cornell University). 10.48550/arxiv.2601.21742
20. *Measuring Collapse and Correction in Homogeneous-Panel LLM Debate*. arXiv (Cornell University). 10.48550/arxiv.2609.35279
21. *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. arXiv (Cornell University). 10.48550/arxiv.2311.17371

Anhang: Provenienz und Prüfpfad

Jede nummerierte Aussage ist an Objekte des eingefrorenen Snapshots gebunden. IDs sind auf 8 Zeichen gekürzt.

Nr.	Aussagetyp	Bindungen
1	Generierte Interpretation `TITLE · T1`	node_ids: `d0cc19fb, 04caea1b`
2	Generierte Interpretation `ABSTRACT · A1`	node_ids: `48844bfb, 617d3d5c` evidence_ids: `ea4b76ea, e6803395` citation_reference_ids: `C12, C63, C75`
3	Methodenbeschreibung `ABSTRACT · A2`	node_ids: `88fdd801` workflow_ids: `5527bed5`
4	Generierte Interpretation `ABSTRACT · A3`	node_ids: `7b172f08, 04caea1b` evidence_ids: `ac478d0e, 4c5f970b, 85ab2809, 1a5b469a` citation_reference_ids: `C5, C58, C41, C1`
5	Konfliktbeschreibung `ABSTRACT · A4`	node_ids: `544ca568, 56f9b396, 40cef065, a3ca7b1c` evaluation_ids: `2b701234, e38a76ec` conflict_ids: `2b701234, e38a76ec`
6	Limitation `ABSTRACT · A5`	node_ids: `04caea1b, 4870586c, fb4bd755` evaluation_ids: `e38a76ec`
7	Generierte Interpretation `INTRODUCTION · I1`	node_ids: `fe1ec676, d0cc19fb, c5543595` evidence_ids: `e6803395` citation_reference_ids: `C10, C63`
8	Generierte Interpretation `INTRODUCTION · I2`	node_ids: `617d3d5c, 48844bfb` evidence_ids: `e6803395` citation_reference_ids: `C10, C75, C63` evaluation_ids: `4a9555cd`
9	Generierte Interpretation `INTRODUCTION · I3`	node_ids: `48844bfb, d0cc19fb` evidence_ids: `ea4b76ea` citation_reference_ids: `C12, C68` evaluation_ids: `2b701234`
10	Generierte Interpretation `INTRODUCTION · I4`	node_ids: `04caea1b, 7b172f08, 46cf7e0b` evidence_ids: `fd01bcea, 1a5b469a, 7c8bff99` citation_reference_ids: `C36, C43, C64`
11	Generierte Interpretation `INTRODUCTION · I5`	node_ids: `1d0acf6a, 0494d5b2, 4870586c, bc502a2a`
12	Generierte Hypothese `HYPOTHESES · H1`	node_ids: `b087e9d2, dd823bde, 0494d5b2, c8a498c5, efbeb580` evaluation_ids: `a29eff1f, 245dc431, 099bd8bd, 292f8ce0, b463de0f`
13	Generierte Hypothese `HYPOTHESES · H2`	node_ids: `b009b96e, 40cef065, 3a875760, 4870586c, 2c5e6ba9` evaluation_ids: `e38a76ec, a666fdaf, 32b9b745`
14	Generierte Hypothese `HYPOTHESES · H3`	node_ids: `9ae486fc` evaluation_ids: `bb2700d8`
15	Generierte Hypothese `HYPOTHESES · H4`	node_ids: `fe46929a, 6d18a3ae` evaluation_ids: `980ff83c, 740f03e0`
16	Methodenbeschreibung `METHODS · M1`	node_ids: `08ea2d13, 88fdd801, 04caea1b` workflow_ids: `5527bed5`
17	Methodenbeschreibung `METHODS · M2`	node_ids: `0494d5b2, 4870586c, 6d18a3ae` workflow_ids: `5527bed5`
18	Methodenbeschreibung `METHODS · M3`	novelty_audit_ids: `0071e453, 0552c6b3, 5d654f92, c250ca42, c4cf9633, d5618c5a, eeb5b351` workflow_ids: `5527bed5`
19	Methodenbeschreibung `METHODS · M4`	node_ids: `66f83bd9, 29b7ba31, 0494d5b2, 395f4b03` evaluation_ids: `2b701234, e38a76ec, 099bd8bd, 740f03e0` workflow_ids: `5527bed5`

Nr.	Aussagetyp	Bindungen
20	Methodenbeschreibung METHODS · M5	node_ids: fe1ec676, 31efb6a0 workflow_ids: 5527bed5
21	Methodenbeschreibung METHODS · M6	node_ids: bc502a2a, fe1ec676 human_decision_ids: 1e9adc54 workflow_ids: 5527bed5
22	Konfliktbeschreibung DISCUSSION · D1	node_ids: 544ca568, 2b060464, 56f9b396, 617d3d5c, d0cc19fb evidence_ids: ea4b76ea, e6803395 citation_reference_ids: C68, C7 evaluation_ids: 2b701234 conflict_ids: 2b701234
23	Generierte Interpretation DISCUSSION · D2	node_ids: 48844bfb, 617d3d5c, 0494d5b2 evidence_ids: e6803395 citation_reference_ids: C10, C63, C75, C29 evaluation_ids: 4a9555cd, 099bd8bd
24	Generierte Interpretation DISCUSSION · D3	node_ids: d9bf2477, 48844bfb evidence_ids: ea4b76ea, cd3e2891 citation_reference_ids: C44, C12, C47
25	Generierte Interpretation DISCUSSION · D4	node_ids: fe1ec676, fb4bd755 evidence_ids: 102542ca citation_reference_ids: C4, C2
26	Generierte Interpretation DISCUSSION · D5	node_ids: c9e445e5, 88fdd801, d9bf2477 evidence_ids: 4c5f970b, 720774ae, cfc9eb8d citation_reference_ids: C45, C23, C31
27	Generierte Interpretation DISCUSSION · D6	node_ids: c5543595, 2b060464, d0cc19fb evidence_ids: 9066bd0d citation_reference_ids: C70
28	Generierte Interpretation DISCUSSION · D7	node_ids: 3e2eb1b9, 60ccce2c, d0cc19fb evidence_ids: d06c7ea6 citation_reference_ids: C11, C17
29	Generierte Interpretation DISCUSSION · D8	node_ids: 7dd627ad, 16e69c6e, fc727633, 0a900353 evidence_ids: fc4b99c0, 7dbdc2ab citation_reference_ids: C42, C46, C72
30	Generierte Interpretation DISCUSSION · D9	node_ids: 6bb5d88d, 0a900353, 4ed9e3cd evidence_ids: ee37338c, 64cc69d2, 1ccfcee0 citation_reference_ids: C32, C40, C52
31	Generierte Interpretation DISCUSSION · D10	node_ids: 46cf7e0b, 04caea1b, 7b172f08 evidence_ids: ac478d0e, fd01bcea citation_reference_ids: C5, C33, C36
32	Generierte Interpretation DISCUSSION · D11	node_ids: 04caea1b, 0f374dab, 395f4b03 evidence_ids: 1a5b469a, e5c33cb6 citation_reference_ids: C48, C55
33	Generierte Interpretation DISCUSSION · D12	node_ids: 46cf7e0b, 04caea1b, 395f4b03, 9ae486fc evidence_ids: 7c8bff99 citation_reference_ids: C61 evaluation_ids: bb2700d8
34	Generierte Interpretation DISCUSSION · D13	node_ids: 29b7ba31, 395f4b03, 6d18a3ae, 25f92612 evidence_ids: 4c5f970b, 720774ae citation_reference_ids: C22, C23 evaluation_ids: 980ff83c, 740f03e0
35	Generierte Interpretation DISCUSSION · D14	node_ids: 8d5d8020, 2c5e6ba9 evidence_ids: 85ab2809, f2322f8a citation_reference_ids: C8, C15

Nr.	Aussagetyp	Bindungen
36	Konfliktbeschreibung DISCUSSION · D15	node_ids: 40cef065, 48844bfb, 04caea1b, a3ca7b1c, 3a875760, 4870586c evidence_ids: e9a6169a citation_reference_ids: C30 evaluation_ids: e38a76ec, a666fdaf, 32b9b745 conflict_ids: e38a76ec
37	Konfliktbeschreibung DISCUSSION · D16	node_ids: b087e9d2, 66f83bd9, dd823bde, 0494d5b2 evaluation_ids: a29eff1f, 245dc431, 099bd8bd
38	Generierte Interpretation DISCUSSION · D17	node_ids: fb4bd755, 4870586c, a3ca7b1c evidence_ids: e9a6169a citation_reference_ids: C28 evaluation_ids: 32b9b745, a666fdaf
39	Generierte Interpretation DISCUSSION · D18	node_ids: 4ed9e3cd, 0a900353, 25f92612, 4870586c evidence_ids: ea4b76ea, e6803395, 7dbdc2ab, e7bdf99e citation_reference_ids: C16, C10, C39, C69
40	Limitation LIMITATIONS · L1	node_ids: 88fdd801, 2b060464, fe1ec676 evaluation_ids: 2b701234, e38a76ec
41	Limitation LIMITATIONS · L2	node_ids: 25f92612, 395f4b03, 0a900353 evaluation_ids: 740f03e0, 2b701234
42	Limitation LIMITATIONS · L3	node_ids: d0cc19fb, fe1ec676, dd823bde evidence_ids: 102542ca citation_reference_ids: C4, C2 evaluation_ids: 245dc431, a29eff1f
43	Limitation LIMITATIONS · L4	node_ids: 04caea1b, 9ae486fc, 6d18a3ae, 395f4b03 evaluation_ids: bb2700d8, 740f03e0
44	Limitation LIMITATIONS · L5	node_ids: 4870586c, 3a875760, fb4bd755, bc502a2a evaluation_ids: 32b9b745, a666fdaf
45	Limitation LIMITATIONS · L6	node_ids: 099bd8bd, 740f03e0, a29eff1f evaluation_ids: 2b701234, e38a76ec novelty_audit_ids: 0552c6b3, d5618c5a, eeb5b351 human_decision_ids: 1e9adc54 workflow_ids: 5527bed5
46	Generierte Interpretation CONCLUSION · K1	node_ids: 48844bfb, 04caea1b, 0494d5b2, 4870586c
47	Offene Frage CONCLUSION · K2	node_ids: 27ebd4c2, dd823bde, 31efb6a0 evaluation_ids: 245dc431
48	Offene Frage CONCLUSION · K3	node_ids: 16e69c6e, 0a900353, 6bb5d88d
49	Offene Frage CONCLUSION · K4	node_ids: 9ae486fc, 31efb6a0 evaluation_ids: bb2700d8
50	Offene Frage CONCLUSION · K5	node_ids: 6d18a3ae, 29b7ba31 evaluation_ids: 740f03e0
51	Offene Frage CONCLUSION · K6	node_ids: 3a875760, 4870586c, c8a498c5 evaluation_ids: 32b9b745, a666fdaf
52	Offene Frage CONCLUSION · K7	node_ids: fe1ec676, fb4bd755, bc502a2a, 4870586c evaluation_ids: 32b9b745

Erhaltener wissenschaftlicher Kontext

Menschliche Richtungsentscheidungen	Methodische Grenzen und Evidenzlücken priorisieren
Explizite ungelöste Konflikte	2
Evaluationen	13
Novelty-Audits (nur intern, kein globaler Neuheitsnachweis)	7

Menschliche Eingriffe

Wörtlich wiedergegeben. Hinweise des Betreibers betreffen nur Darstellung und Gewichtung; sie sind keine Evidenz und erzeugen keine Aussagen.

Schwerpunktwahl am Gate	Methodische Grenzen und Evidenzlücken priorisieren
Kommentar am Gate	Bitte empirisch belegte Befunde klar von theoretischen Annahmen und unbestätigten Hypothesen trennen. Widersprüchliche Studien nach Methodik, Datengrundlage und Evaluationskriterien gegenüberstellen. Die offenen Forschungslücken, die für ein späteres experimentelles Design relevant sind, am Ende als nummerierte Liste zusammenfassen.
Hinweis zu dieser Version	Nur mit dem eingefrorenen Forschungsstand überarbeiten; keine neuen Quellen, Zahlen, Aussagen oder Ergebnisse; Quellenverweise und Bindungen erhalten. 1. Zusammenfassung: Forschungsfrage direkt beantworten – was die Studien über falschen Konsens zeigen, welche Gegenmaßnahmen untersucht wurden, was offen bleibt. Konfliktbeschreibung dort kürzen. 2. Einleitung: „falscher Konsens“, „falsche Mehrheit“, „Antwortwechsel“, „Konformität“ und „schädlicher Peer-Einfluss“ knapp voneinander abgrenzen; unterschiedliche Messgrößen und Nenner erklären. 3. Hypothesen H1–H4: je Kernaussage, zentraler Gegeneinwand, Status „unbestätigt“; kürzer; vorgeschlagene Tests nicht wie Ergebnisse formulieren. 4. Methodik: zuerst sagen, dass überwiegend Abstracts und ausgewählte Volltextstellen ausgewertet wurden; interne Prozesszahlen knapper (Details im Anhang). 5. Diskussion: in der Reihenfolge Messprobleme, Ursachen, Heterogenität, Prüfverfahren, Kosten ordnen; gegensätzliche Studien nach den genannten Aufgaben, Kontrollen und Zielgrößen vergleichen; nicht vergleichbare Prozentsätze nicht zusammenführen. 6. Die zwei dokumentierten Konflikte vollständig erhalten; weitere Kontroversen überwiegend im Fließtext statt als eigene Konfliktbeschreibung; Wiederholungen reduzieren. 7. Limitationen und Fazit: wiederholte Vorbehalte straffen; die sechs Forschungslücken als getrennte nummerierte Punkte; im Fazit Erkenntnisstand von offenen Prüfaufgaben trennen.
Überarbeitung von Version	1

Prüfdaten

Research Run	`46796a44-b447-4adf-bae3-5cda1385080a`
Synthese-Snapshot	`2d7dae31-d4b3-4875-ab9f-fd0fa2b67588`
Writer-Execution	`fd767534-39a8-42e8-be53-af9ece6289e2`
Draft-Digest (SHA-256)	`81173653a8fc26c4845c14d6723b9de98e2b91ea415fb6f0f40b3e0f6d3dde33`
Renderer-Version	`6`