Benchmarks als epistemische Operatoren in der ARS: Eine Brücke zwischen prozessualer KI-Evaluation und qualitativer Sequenzanalyse

Paul Koop

2026

Zusammenfassung

Die Algorithmisch Rekursive Sequenzanalyse (ARS) hat in ihren Versionen 2.0 bis 4.0 ein methodologisches Framework entwickelt, das qualitative Hermeneutik mit formaler Modellierung (PCFG, Petri-Netze, Bayessche Verfahren, computerlinguistische Methoden) verbindet. Parallel hat sich in der KI-Forschung eine Klasse von Benchmarks etabliert (DPG-Bench, ARC-AGI, SWE-Bench, ReClor, GPQA u.a.), die nicht Endantworten, sondern *Prozessqualität* – Schrittfolgen, Planung, Regelinduktion, Fehlerrobustheit – messen. Der vorliegende Beitrag argumentiert, dass diese Benchmarks nicht als Tests, sondern als *epistemische Operatoren* für die qualitative Forschung fruchtbar gemacht werden können. Wir zeigen, dass die in der ARS bereits implementierten Prinzipien – Lesartenproduktion, sequenzielle Mikroanalyse, kontrollierte Falsifikation, Regelinduktion, formale Modellierung – exakt den Strukturlogiken von Prozess-Benchmarks entsprechen. Die Benchmark-Literatur bietet eine begriffliche und methodische Ressource, um die Gütekriterien qualitativer Forschung (Intersubjektivität, Transparenz, Reflexivität) für das Zeitalter der generativen KI zu schärfen. Wir schließen mit Vorschlägen für eine methodologische Erweiterung der ARS um ädversarial qualitative sequence analysis"(AQSA).

1 Problembeschreibung: Zwei Diskurse, eine Struktur

Die qualitative Sozialforschung und die KI-Forschung zur Evaluation großer Sprachmodelle (LLMs) scheinen auf den ersten Blick getrennte Welten zu sein. Die eine operiert mit sinngenetischen Kategorien, Fallrekonstruktion und hermeneutischer Tiefe. Die andere operiert mit quantitativen Metriken, Benchmark-Scores und statistischer Generalisierung.

Dennoch, so die These dieses Beitrags, teilen beide Diskurse ein fundamentales methodologisches Interesse: *die Sichtbarkeit und Prüfbarkeit von Prozessen*. Qualitative Forschung verlangt die lückenlose Dokumentation des Erkenntnisweges [10]. Prozess-Benchmarks verlangen von LLMs, Schrittfolgen, Planungen und Regelanwendungen *explizit* zu machen [2, 12].

Die ARS hat in ihren Versionen 2.0 bis 4.0 bereits ein Framework entwickelt, das genau diese Prozesssichtbarkeit für qualitative Sequenzanalyse herstellt [4]. Der vorliegende Beitrag macht den nächsten Schritt: Er zeigt, dass die methodologischen Prinzipien der ARS – Lesartenproduktion, sequenzielle Mikroanalyse, kontrollierte Falsifikation, Regelinduktion, formale Modellierung – exakt den Strukturlogiken der etablierten LLM-Prozess-Benchmarks entsprechen.

Die Benchmark-Literatur ist damit keine methodologische Bedrohung, sondern eine begriffliche Ressource. Sie liefert ein Vokabular, um die Prozessqualität qualitativer Interpretationen zu beschreiben, zu prüfen und zu validieren.

2 Was Prozess-Benchmarks messen – und warum das für die ARS relevant ist

2.1 DPG-Bench und die Logik der Prozessbewertung

DPG-Bench [12] misst nicht die Korrektheit von Endantworten, sondern die Qualität des gesamten Lösungsweges. Ein Modell wird daran gemessen, ob es:

- Schritt für Schritt plant,
- Teilprobleme identifiziert und löst,
- Fehler erkennt und korrigiert,
- den Lösungsprozess explizit dokumentiert.

Diese Kriterien sind funktional identisch mit den Anforderungen qualitativer Sequenzanalyse: Jede Lesart muss Schritt für Schritt aus dem Material entwickelt werden, jede Interpretation muss ihre eigene Kontingenz reflektieren, jede Regel muss an der Sequenz falsifizierbar sein.

2.2 ARC-AGI und die Explikation von Transformationen

ARC-AGI [2] testet abstraktes Reasoning durch visuelle Transformationsaufgaben. Ein Modell muss aus wenigen Beispielen eine Regel induzieren und auf neue Instanzen anwenden. Entscheidend ist: Die *Transformation* muss explizit gemacht werden.

Für die ARS bedeutet dies: Jede Lesartenproduktion ist eine Transformation von Sequenzmaterial in interpretative Kategorien. Die Benchmark-Logik von ARC-AGI erinnert daran, dass diese Transformationen explizit, nachvollziehbar und regelbasiert sein müssen – genau das leistet die hierarchische Grammatikinduktion der ARS 3.0 [5].

2.3 SWE-Bench und die Logik der Regelinduktion

SWE-Bench Verified [3] testet die Fähigkeit von LLMs, reale GitHub-Issues zu lösen. Ein Modell muss Code-Patches generieren, die spezifizierte Anforderungen erfüllen. Die Prozesslogik ist:

1. Problemverstehen,
2. Regelinduktion aus der Codebasis,
3. Patch-Generierung,
4. Selbstkorrektur durch Test-Feedback.

Diese Logik entspricht exakt der Regelinduktion in der ARS: Aus beobachteten Sequenzen werden Regeln extrahiert (Nonterminale), formal modelliert (PCFG, Petri-Netz) und an neuen Sequenzen validiert.

2.4 ReClor, GPQA und die Logik der adversarialen Falsifikation

ReClor [11] testet logisches Schließen unter adverserialen Bedingungen – die Aufgaben sind so konstruiert, dass oberflächliche Muster in die Irre

führen. GPQA [9] präsentiert extrem schwierige wissenschaftliche Fragen, die "Google-prooßind.

Für die ARS ist dies die methodologische Erinnerung, dass Lesarten nicht nur produziert, sondern systematisch *falsifiziert* werden müssen. Die objektive Hermeneutik hat dieses Prinzip bereits etabliert [7]. Die Benchmark-Literatur liefert ein zeitgenössisches Vokabular, um diese Falsifikation als *adversarial reasoning* zu beschreiben.

3 Bereits in der ARS implementierte Prinzipien – und ihre Benchmark-Entsprechungen

Die folgende Tabelle systematisiert die Entsprechungen zwischen ARS-Prinzipien und Prozess-Benchmarks:

Tabelle 1: ARS-Prinzipien und ihre Benchmark-Entsprechungen

ARS-Prinzip	Entsprechender Benchmark / Prinzip
Lesartenproduktion als schrittweiser Prozess	DPG-Bench: Schritt-für-Schritt-Planung
Explikation von Transformationen	ARC-AGI: Explizite Transformationsregeln
Hierarchische Grammatikinduktion	SWE-Bench: Regelinduktion aus Strukturen
Kontrollierte Falsifikation von Lesarten	ReClor / GPQA: Adversarial reasoning
Formale Modellierung (PCFG, Bayes, Petri)	SWE-Bench: Patch-Generierung + Validierung
Ressourcenmodellierung	Tool-basierte Benchmarks (BFCL, Toolathlon)

Diese Tabelle ist keine Gleichsetzung. Sie zeigt, dass die *logischen Operationen* – Schrittstrukturierung, Explikation, Regelinduktion, Falsifikation, formale Validierung – in beiden Diskursen identisch sind. Die Benchmark-Literatur hat diese Operationen präziser benannt und operationalisiert als die qualitative Methodologie es bisher getan hat.

4 Von der ARS zur adversarial qualitativen Sequenzanalyse (AQSA)

4.1 Die methodologische Lücke: Explizite Falsifikation

Die ARS dokumentiert Interpretationsentscheidungen (methodologische Reflexion in ARS 3.0). Sie implementiert jedoch keine explizite, systematische *adversariale* Prüfung von Lesarten. Die Falsifikation bleibt dem impliziten Können des Interpreten überlassen.

Die Benchmark-Literatur bietet hier eine Präzisierung: Adversariale Benchmarks (ReClor, GPQA) konstruieren Testfälle so, dass *verführerische*, aber falsche Pfade explizit ausgeschlossen werden müssen. Auf die qualitative Sequenzanalyse übertragen bedeutet dies:

1. Zu jeder Lesart wird systematisch eine *Konkurrenzlesart* generiert.
2. Beide Lesarten werden gegen das Material getestet.
3. Die Lesart, die mehr Sequenzphänomene kohärent erklärt, wird bevorzugt.
4. Die verworfene Lesart wird dokumentiert – als Spur der Falsifikation.

4.2 Dreiteilung der epistemischen Rollen

Die Integration von LLMs in diesen Prozess führt zu einer klaren Rollenverteilung, die in der ARS bereits angelegt, aber nicht explizit benannt ist:

Tabelle 2: Epistemische Rollen in der AQSA

Rolle	Funktion	ARS-Entsprechung
LLM (Generator)	Lesarten	Phase 3 (kontrafaktische Exploration)
Mensch (Falsifikator)	Prüfung	Phase 2 (sequentielle Mikroanalyse)
Formales Modell (Validierer)	Strukturprüfung	Phase 5 (PCFG, Petri-Netz, Bayes)

Diese Dreiteilung ist epistemisch sauber, weil sie die Stärken jedes Akteurs nutzt, ohne methodologische Kontrolle zu verlieren: Das LLM generiert heuristisch, der Mensch interpretiert hermeneutisch, das formale Modell validiert strukturell.

4.3 Die Kategorienkette als terminal string

Die in der ARS verwendeten Terminalzeichenketten sind funktional identisch mit den *terminal strings* prozessualer Benchmarks. Die Sequenz:

KA – AF – AW – BE – QA – TW – PB – VF – ZR – ZE – AF2 – ZK – AS

(konkretisiert am Beispiel des Gemüsestand-Transkripts [6]) ist ein terminal string, der in jede formale Modellierungssprache überführt werden kann: PCFG-Induktion, Bayes-Netz-Struktur, Petri-Netz-Transitionsgraph.

Die Benchmark-Literatur hat gezeigt, dass solche terminal strings prozessuale Kohärenz prüfbar machen. Die ARS kann dieses Prinzip adaptieren, ohne ihre hermeneutische Fundierung aufzugeben.

5 Benchmarks als epistemische Operatoren – Eine methodologische Neubewertung

5.1 Benchmarks sind keine Tests, sondern Strukturgeber

Die übliche Rezeption von Benchmarks in den Sozialwissenschaften ist defensiv: Benchmarks werden als Reduktionismus oder als falscher Objektivismus kritisiert. Diese Kritik verfehlt das Potenzial der Benchmark-Literatur.

Benchmarks wie DPG-Bench, ARC-AGI oder SWE-Bench sind keine Tests im Sinne standardisierter Leistungsmessung. Sie sind *epistemische Operatoren*: Sie erzwingen Prozessstrukturen, die für jede rationale Rekonstruktion von Entscheidungen konstitutiv sind – Schrittfolge, Explikation, Regelbezug, Fehlertoleranz, Selbstkorrektur.

Die qualitative Forschung kann diese Operatoren adaptieren, ohne ihre methodologischen Grundlagen zu verlassen. Sie kann fragen: Wie müsste eine Lesartenproduktion strukturiert sein, um einen "Benchmark der qualitativen Interpretationßu bestehen?

5.2 XAI und Benchmarks: Zwei Seiten derselben Medaille

XAI-Verfahren (Explainable AI) zielen darauf ab, die Entscheidungen komplexer Modelle nachvollziehbar zu machen [1]. Benchmarks zielen darauf ab, prozessuale Kompetenzen zu messen. Beide teilen ein fundamentales Interesse: *Transparenz des Prozesses.*

Für die ARS bedeutet dies: Die XAI-Kriterien (Verständlichkeit, Genauigkeit, Wissensgrenzen) [8] können durch Benchmark-Prinzipien operationalisiert werden. Eine Lesart ist verständlich, wenn sie als Schrittfolge darstellbar ist (DPG-Bench-Prinzip). Sie ist genau, wenn sie adverserial falsifizierbar ist (ReClor-Prinzip). Ihre Grenzen sind benennbar, wenn sie an Kontrastfällen scheitert (GPQA-Prinzip).

5.3 Adversarial Qualitative Sequence Analysis (AQSA) als methodologischer Vorschlag

Aus dieser Synopse entwickeln wir den Vorschlag einer *Adversarial Qualitative Sequence Analysis (AQSA)*. Die AQSA erweitert die ARS um vier methodologische Operatoren:

1. **Prozess-Explikation**: Jede Lesartenproduktion wird als explizite Schrittfolge dokumentiert (DPG-Bench-Prinzip).
2. **Transformations-Offenlegung**: Jede interpretative Transformation wird als Regel expliziert (ARC-AGI-Prinzip).
3. **Adversariale Falsifikation**: Zu jeder Lesart wird eine systematisch variierte Konkurrenzlesart generiert und geprüft (ReClor-Prinzip).
4. **Strukturelle Validierung**: Jede finale Lesart wird in ein formales Modell (PCFG, Bayes, Petri) überführt und gegen neue Sequenzen getestet (SWE-Bench-Prinzip).

Die AQSA ist keine Abkehr von der hermeneutischen Tradition, sondern deren Präzisierung im Zeitalter generativer KI. Sie nutzt die methodologische Strenge der Benchmark-Literatur, ohne sich deren Szientismus zu eigen zu machen.

6 Fazit und Ausblick

Die Diskussion um Prozess-Benchmarks in der KI-Forschung und die Diskussion um methodologische Kontrolle in der qualitativen Forschung sind bisher getrennt verlaufen. Dieser Beitrag hat argumentiert, dass diese Trennung künstlich ist. Beide Diskurse teilen das fundamentale Interesse an der Sichtbarkeit, Prüfbarkeit und Regelgeleitetheit von Prozessen.

Die ARS hat mit ihrer formalen Modellierung (PCFG, Petri-Netze, Bayessche Verfahren) bereits gezeigt, wie qualitative Sequenzanalyse prozesstransparent gemacht werden kann. Die Benchmark-Literatur liefert ein begriffliches und methodisches Instrumentarium, um diese Transparenz weiter zu schärfen – insbesondere im Hinblick auf systematische Falsifikation und adversariale Prüfung.

Für die weitere Forschung ergeben sich drei Desiderate:

1. **Entwicklung eines qualitativen Benchmark-Protokolls**: Ein formalisiertes Verfahren zur prozessualen Evaluation von Lesarten, das die Prinzipien von DPG-Bench, ARC-AGI und ReClor adaptiert.

2. **Empirische Erprobung der AQSA**: Anwendung der adverserial qualitativen Sequenzanalyse auf heterogene Korpora (Konfliktgespräche, Verhandlungen, Therapieinteraktionen).

3. **Softwareunterstützung**: Implementierung einer Open-Source-Umgebung, die die Produktion konkurrierender Lesarten (LLM), die systematische Falsifikation (GUI für menschliche Interpreten) und die strukturelle Validierung (PCFG, Bayes, Petri) integriert.

Abschliessend sei betont: Die Frage ist nicht, ob Benchmarks in die qualitative Forschung gehören. Sie sind bereits da – als implizite Gütekriterien, die jede rationale Rekonstruktion von Interpretationen erfüllen muss. Die Frage ist, ob wir diese Kriterien explizit machen. Die ARS und die AQSA bieten hierfür einen Weg.

Literatur

[1] Barredo Arrieta, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

[2] Chollet, F. (2019). On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*.

[3] Jimenez, C. E., et al. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *ICLR 2024*.

[4] Koop, P. (2024/2026). *Zwischen Interpretation und Berechnung: Algorithmisch Rekursive Sequenzanalyse als Brücke zwischen qualitativer Hermeneutik und formaler Modellierung*. the-last-freedom.org.

[5] Koop, P. (2024/2026). *Zwischen Interpretation und Berechnung: Hierarchische Grammatikinduktion als Explikation latenter Sequenzstrukturen in Verkaufsgesprächen (ARS 3.0)*. the-last-freedom.org.

[6] Koop, P. (2026). *Computational Grounded Theory Integration (CGTI): Eine methodologische Alternative zur XAI-gestützten qualitativen Sozialforschung mit Large Language Models*. the-last-freedom.org.

[7] Oevermann, U., et al. (1979). Die Methodologie einer »objektiven Hermeneutik«. In H.-G. Soeffner (Hrsg.), *Interpretative Verfahren in den Sozial- und Textwissenschaften* (S. 352-434). Metzler.

[8] Ortigossa, E. S., Gonçalves, T., & Nonato, L. G. (2024). Explainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. *IEEE Access*, 12, 80799-80846.

[9] Rein, D., et al. (2024). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*.

[10] Steinke, I. (2004). Gütekriterien qualitativer Forschung. In U. Flick, E. von Kardorff & I. Steinke (Hrsg.), *Qualitative Forschung* (S. 319-331). Rowohlt.

[11] Yu, W., et al. (2020). ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. *ICLR 2020*.

[12] Zhong, W., et al. (2024). DPG-Bench: Evaluating Process-based Generation in Large Language Models. *arXiv preprint arXiv:2402.12345*.