# Benchmarks as Epistemic Operators in ARS: Bridging Processual KI Evaluation and Qualitative Sequence Analysis

Paul Koop

2026

**Abstract**

The Algorithmic Recursive Sequence Analysis (ARS), in its versions 2.0 to 4.0, has developed a methodological framework that bridges qualitative hermeneutics and formal modeling (PCFG, Petri nets, Bayesian methods, computational linguistics). In parallel, the KI research community has established a class of benchmarks (DPG-Bench, ARC-AGI, SWE-Bench, ReClor, GPQA, among others) that measure not final answers but *process quality* – step sequences, planning, rule induction, robustness. This paper argues that these benchmarks can be fruitfully appropriated not as tests but as *epistemic operators* for qualitative research. We show that the principles already implemented in ARS – interpretation production, sequential micro-analysis, controlled falsification, rule induction, formal modeling – correspond exactly to the structural logics of process benchmarks. The benchmark literature provides a conceptual and methodological resource for sharpening the quality criteria of qualitative research (intersubjectivity, transparency, reflexivity) for the age of generative KI. We conclude with proposals for a methodological extension of ARS toward *adversarial qualitative sequence analysis (AQSA)*.

## 1 Problem Statement: Two Discourses, One Structure

Qualitative social research and KI research on large language model (LLM) evaluation appear, at first glance, to be separate worlds. The former operates with meaning-genetic categories, case reconstruction, and hermeneutic

depth. The latter operates with quantitative metrics, benchmark scores, and statistical generalization.

Nevertheless, as this paper argues, both discourses share a fundamental methodological interest: *the visibility and testability of processes*. Qualitative research demands complete documentation of the path of knowledge [10]. Process benchmarks demand that LLMs make step sequences, plans, and rule applications *explicit* [2, 12].

ARS has already developed, in versions 2.0 through 4.0, a framework that provides precisely this process visibility for qualitative sequence analysis [4]. This paper takes the next step: It shows that the methodological principles of ARS – interpretation production, sequential micro-analysis, controlled falsification, rule induction, formal modeling – correspond exactly to the structural logics of established LLM process benchmarks.

The benchmark literature is thus not a methodological threat but a conceptual resource. It provides a vocabulary for describing, testing, and validating the process quality of qualitative interpretations.

# 2 What Process Benchmarks Measure – and Why This Matters for ARS

## 2.1 DPG-Bench and the Logic of Process Evaluation

DPG-Bench [12] measures not the correctness of final answers but the quality of the entire solution path. A model is evaluated on whether it:

- plans step by step,
- identifies and solves subproblems,
- detects and corrects errors,
- documents the solution process explicitly.

These criteria are functionally identical to the requirements of qualitative sequence analysis: Every interpretation must be developed step by step from the material, every interpretation must reflect its own contingency, every rule must be falsifiable against the sequence.

## 2.2 ARC-AGI and the Explication of Transformations

ARC-AGI [2] tests abstract reasoning through visual transformation tasks. A model must induce a rule from few examples and apply it to new instances. Crucially, the *transformation* must be made explicit.

For ARS, this means: Every interpretation production is a transformation of sequence material into interpretive categories. The benchmark logic of ARC-AGI reminds us that these transformations must be explicit, traceable, and rule-based – precisely what the hierarchical grammar induction of ARS 3.0 accomplishes [5].

## 2.3 SWE-Bench and the Logic of Rule Induction

SWE-Bench Verified [3] tests the ability of LLMs to solve real-world GitHub issues. A model must generate code patches that meet specified requirements. The process logic is:

1. Problem understanding,
2. Rule induction from the codebase,
3. Patch generation,
4. Self-correction through test feedback.

This logic corresponds exactly to rule induction in ARS: From observed sequences, rules are extracted (nonterminals), formally modeled (PCFG, Petri net), and validated against new sequences.

## 2.4 ReClor, GPQA and the Logic of Adversarial Falsification

ReClor [11] tests logical reasoning under adversarial conditions – tasks are constructed so that superficial patterns lead astray. GPQA [9] presents extremely difficult scientific questions that are "Google-proof."

For ARS, this is the methodological reminder that interpretations must not only be produced but systematically *falsified*. Objective hermeneutics has already established this principle [7]. The benchmark literature provides a contemporary vocabulary for describing this falsification as *adversarial reasoning*.

# 3 Principles Already Implemented in ARS – and Their Benchmark Correspondences

The following table systematizes the correspondences between ARS principles and process benchmarks:

Table 1: ARS Principles and Their Benchmark Correspondences

| ARS Principle | Corresponding Benchmark / Principle |
|---|---|
| Interpretation as stepwise process | DPG-Bench: Step-by-step planning |
| Explication of transformations | ARC-AGI: Explicit transformation rules |
| Hierarchical grammar induction | SWE-Bench: Rule induction from structures |
| Controlled falsification of interpretations | ReClor / GPQA: Adversarial reasoning |
| Formal modeling (PCFG, Bayes, Petri) | SWE-Bench: Patch generation + validation |
| Resource modeling | Tool-based benchmarks (BFCL, Toolathlon) |

This table is not an equation. It shows that the *logical operations* – step structuring, explication, rule induction, falsification, formal validation – are identical in both discourses. The benchmark literature has named and operationalized these operations more precisely than qualitative methodology has done to date.

# 4 From ARS to Adversarial Qualitative Sequence Analysis (AQSA)

## 4.1 The Methodological Gap: Explicit Falsification

ARS documents interpretive decisions (methodological reflection in ARS 3.0). However, it does not implement an explicit, systematic *adversarial* testing of interpretations. Falsification is left to the implicit skill of the interpreter.

The benchmark literature offers a precision here: Adversarial benchmarks (ReClor, GPQA) construct test cases such that *tempting but false* paths

must be explicitly excluded. Translated to qualitative sequence analysis, this means:

1. For each interpretation, a systematic *competing interpretation* is generated.
2. Both interpretations are tested against the material.
3. The interpretation that coherently explains more sequence phenomena is preferred.
4. The rejected interpretation is documented – as a trace of falsification.

## 4.2 Threefold Division of Epistemic Roles

The integration of LLMs into this process leads to a clear division of roles, already present in ARS but not explicitly named:

Table 2: Epistemic Roles in AQSA

| Role | Function | ARS Correspondence |
|---|---|---|
| LLM (Generator) | Production | Phase 3 (counterfactual exploration) |
| Human (Falsifier) | Testing | Phase 2 (sequential micro-analysis) |
| Formal Model (Validator) | Structural testing | Phase 5 (PCFG, Petri net, Bayes) |

This threefold division is epistemically clean because it uses the strengths of each actor without losing methodological control: The LLM generates heuristically, the human interprets hermeneutically, the formal model validates structurally.

## 4.3 The Category Chain as Terminal String

The terminal symbol chains used in ARS are functionally identical to the *terminal strings* of processual benchmarks. The sequence:

CA – AQ – SA – CO – QA – TS – PE – UQ – PR – PR2 – AQ2 – AP – CF

(concretized from the vegetable stand transcript [6]) is a terminal string that can be translated into any formal modeling language: PCFG induction, Bayesian network structure, Petri net transition graph.

The benchmark literature has shown that such terminal strings make processual coherence testable. ARS can adapt this principle without abandoning its hermeneutic foundation.

# 5 Benchmarks as Epistemic Operators – A Methodological Reassessment

## 5.1 Benchmarks Are Not Tests but Structuring Devices

The typical reception of benchmarks in the social sciences is defensive: Benchmarks are criticized as reductionism or false objectivism. This critique misses the potential of the benchmark literature.

Benchmarks such as DPG-Bench, ARC-AGI, or SWE-Bench are not tests in the sense of standardized performance measurement. They are *epistemic operators*: They enforce process structures that are constitutive for any rational reconstruction of decisions – step sequence, explication, rule reference, error tolerance, self-correction.

Qualitative research can adapt these operators without abandoning its methodological foundations. It can ask: How would an interpretation production have to be structured to pass a "benchmark of qualitative interpretation"?

## 5.2 XAI and Benchmarks: Two Sides of the Same Coin

XAI methods (Explainable AI) aim to make the decisions of complex models traceable [1]. Benchmarks aim to measure processual competencies. Both share a fundamental interest: *transparency of process*.

For ARS, this means: The XAI criteria (meaningfulness, accuracy, knowledge limits) [8] can be operationalized through benchmark principles. An interpretation is meaningful if it can be represented as a step sequence (DPG-Bench principle). It is accurate if it is adversarially falsifiable (ReClor principle). Its limits are identifiable if it fails on contrast cases (GPQA principle).

## 5.3 Adversarial Qualitative Sequence Analysis (AQSA) as a Methodological Proposal

From this synopsis, we develop the proposal for an *Adversarial Qualitative Sequence Analysis (AQSA)*. AQSA extends ARS by four methodological operators:

1. **Process Explication**: Every interpretation production is documented as an explicit step sequence (DPG-Bench principle).

2. **Transparency of Transformations**: Every interpretive transformation is explicated as a rule (ARC-AGI principle).

3. **Adversarial Falsification**: For each interpretation, a systematically varied competing interpretation is generated and tested (ReClor principle).

4. **Structural Validation**: Every final interpretation is translated into a formal model (PCFG, Bayes, Petri) and tested against new sequences (SWE-Bench principle).

AQSA is not a departure from the hermeneutic tradition but its precisification in the age of generative KI. It uses the methodological rigor of the benchmark literature without adopting its scientism.

# 6 Conclusion and Outlook

The discussion of process benchmarks in KI research and the discussion of methodological control in qualitative research have so far proceeded separately. This paper has argued that this separation is artificial. Both discourses share the fundamental interest in the visibility, testability, and rule-guidedness of processes.

ARS, with its formal modeling (PCFG, Petri nets, Bayesian methods), has already shown how qualitative sequence analysis can be made process-transparent. The benchmark literature provides a conceptual and methodological toolkit to sharpen this transparency further – especially with regard to systematic falsification and adversarial testing.

For future research, three desiderata emerge:

1. **Development of a Qualitative Benchmark Protocol**: A formalized procedure for the processual evaluation of interpretations that adapts the principles of DPG-Bench, ARC-AGI, and ReClor.

2. **Empirical Testing of AQSA**: Application of adversarial qualitative sequence analysis to heterogeneous corpora (conflict conversations, negotiations, therapeutic interactions).

3. **Software Support**: Implementation of an open-source environment that integrates the production of competing interpretations (LLM), systematic falsification (GUI for human interpreters), and structural validation (PCFG, Bayes, Petri).

In conclusion: The question is not whether benchmarks belong in qualitative research. They are already there – as implicit quality criteria that any rational reconstruction of interpretations must satisfy. The question is whether we make these criteria explicit. ARS and AQSA offer a way forward.

# References


[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). *Between Interpretation and Computation: Algorithmic Recursive Sequence Analysis as a Bridge between Qualitative Hermeneutics and Formal Modeling*. the-last-freedom.org.

[5] Koop, P. (2024/2026). *Between Interpretation and Computation: Hierarchical Grammar Induction as Explication of Latent Sequence Structures in Sales Conversations (ARS 3.0)*. the-last-freedom.org.

[6] Koop, P. (2026). *Computational Grounded Theory Integration (CGTI): A Methodological Alternative to XAI-Supported Qualitative Social Research with Large Language Models*. the-last-freedom.org.

[7] Oevermann, U., et al. (1979). The Methodology of ›Objective Hermeneutics‹ and Its General Research-Logical Significance in the Social Sciences. In H.-G. Soeffner (Ed.), *Interpretative Procedures in the Social and Text Sciences* (pp. 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). Quality Criteria in Qualitative Research. In U. Flick, E. von Kardorff & I. Steinke (Eds.), *A Companion to Qualitative Research* (pp. 184-190). Sage.

[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*.