Constraining Bias and Self-Reference in N=1 Cognitive Workflow Research: A Protocol-First Framework

Edervaldo José de Souza Melo
edersouzamelo@gmail.com
ORCID: 0009-0003-6835-135X

February 2026

Abstract

Background. Personalized cognitive systems and human–AI cognitive workflows are increasingly evaluated through single-case (N=1) studies. While such evaluations are common in practice, they are rarely accompanied by standardized protocols that ensure transparency, reproducibility, and auditability.

Problem. N=1 research in highly self-referential contexts—particularly when the researcher is also the system operator—faces elevated risks of bias, retrospective metric adjustment, undocumented configuration changes, and narrative reinterpretation. Existing methodological guidance addresses either clinical N-of-1 trials or qualitative case studies, but offers limited operational safeguards for evaluating cognitive workflows as systems.

Contribution. This manuscript introduces a protocol-first, pre-registered framework for N=1 evaluation of human–AI cognitive workflows. The contribution is meta-methodological: a fully specified protocol that defines units of analysis, controlled inputs, logging requirements, versioning rules, failure modes, audit criteria, and a transparency checklist. The framework explicitly separates system evaluation from population inference and formalizes reproducibility as repeatability under fixed conditions, not external generalization.

Transparency Statement. This manuscript reports no empirical results. No claims of efficacy, performance, or outcome improvement are made.

Relevance. By publishing the protocol prior to data collection and inviting community review through OSF Preprints and PCI Meta-Research, the manuscript aims to reduce researcher degrees of freedom, mitigate self-confirmation risks, and provide a reusable reporting standard for auditable N=1 cognitive-system research.

Keywords: N=1 research; meta-research; reproducibility; preregistration; auditability; human–AI interaction; cognitive workflows

1 Introduction

Concerns regarding reproducibility, researcher bias, and undisclosed analytic flexibility have become central to contemporary meta-research debates [1, 2, 3, 4, 5]. Practices such as HARKing, p-hacking, and selective reporting have been documented across multiple disciplines, motivating the adoption of preregistration, registered reports, and transparency standards [6, 7, 8, 9].

Concurrently, the emergence of personalized cognitive systems, including human–AI cognitive workflows, has shifted evaluation practices toward individualized, iterative, and self-referential experimentation. In such contexts, N=1 studies are often the only feasible evaluation mode. However, without explicit methodological safeguards, these studies risk conflating exploratory tinkering with scientific evaluation [10, 11].

This manuscript addresses a specific gap: how to design an auditable, reproducible N=1 protocol for evaluating cognitive workflows as systems, rather than as sources of population-level inference. The focus is on reproducibility, defined as repeatability of outcomes under fixed conditions, not on external replication or generalization [5, 12].

Beyond well-documented statistical malpractices, a central concern in contemporary meta-research involves researcher degrees of freedom in settings where hypotheses, metrics, and analytical boundaries remain weakly constrained. These risks are amplified in highly iterative and self-referential research contexts, where exploratory adjustments may silently migrate into confirmatory claims [6, 10, 11].

Human–AI cognitive workflows represent an extreme instance of this challenge. The researcher often acts simultaneously as system designer, operator, observer, and interpreter of outcomes. Without explicit procedural constraints, this role convergence creates conditions favorable to retrospective metric redefinition, undocumented configuration changes, and narrative reinterpretation of outputs [13, 14].

This manuscript adopts a protocol-first perspective in response to these pressures. Rather than proposing new analytical metrics or outcome measures, it focuses on formalizing the conditions under which an N=1 cognitive workflow evaluation may be considered transparent, reproducible, and auditable. By explicitly constraining procedural flexibility before data collection, the proposed framework aims to reduce researcher degrees of freedom and to distinguish exploratory system development from methodologically accountable evaluation [9, 8].

The protocol proposed in this manuscript is motivated by the growing prevalence of exploratory self-monitoring practices involving LLM-based prompt-agents. In such contexts, users routinely interact with agents designed to observe, summarize, and generate metrics about their own cognitive or behavioral patterns. These practices are explicitly treated here as exploratory and pre-scientific: they do not constitute empirical evaluation, do not satisfy reproducibility criteria, and are not interpreted as evidence of system performance. Rather, they function as a motivating context that reveals the need for explicit protocolization prior to any scientific claim. The present framework does not retroactively legitimize such exploratory uses, but instead defines the conditions under which a transition from exploratory interaction to auditable N=1 evaluation could occur.

2 Scope and Definitions

2.1 N=1 / Single-Case Research

An N=1 study refers to a single-case or n-of-1 design in which repeated observations are conducted within one system or subject. In this protocol, N=1 does not imply population inference or statistical generalization [15, 16, 17].

2.2 Transparency

The public availability of protocols, decision rules, logs, and documentation sufficient for third-party inspection.

2.3 Reproducibility

The ability to repeat the same procedure under identical conditions and obtain traceably comparable outputs.

2.4 Auditability

The capacity for an independent reviewer to reconstruct procedural steps, configuration states, and decision points from preserved records.

2.5 Cognitive Workflow

A structured sequence of inputs, transformations, and outputs involving a human operator and computational components, with explicit logging of interactions and decisions.

These definitions align with established distinctions in reproducible and computational research, where transparency, reproducibility, and auditability are treated as separable but interdependent properties [12, 13, 5].

To avoid category errors, it is necessary to clarify what the present framework does not claim. First, N=1 research as defined here does not support population-level generalization, statistical inference, or predictive validity beyond the evaluated system [18]. Second, reproducibility in this context does not imply robustness across different participants, environments, or implementations [5].

Similarly, transparency should not be conflated with unrestricted data disclosure. The framework distinguishes between procedural transparency and data accessibility, consistent with current open science guidance [12, 14].

These distinctions are essential to prevent misinterpretation of the protocol's scope and to ensure that its evaluation criteria remain aligned with its stated objectives.

3 Why N=1 Can Be Scientifically Legitimate

The scientific legitimacy of N=1 research depends fundamentally on the target of inference. In the present protocol, the target is system behavior under controlled conditions, not population-level effects or statistical generalization [17, 18].

Analogous to engineering validation, a cognitive workflow may be evaluated for internal consistency, stability, and failure behavior through repeated trials with fixed parameters. Criticism of such studies for lacking generalizability reflects a categorical error when generalization is not the stated objective. Legitimacy is preserved when (i) the evaluand is a system or protocol, (ii) conditions are explicitly controlled and versioned, and (iii) claims are limited to reproducibility and auditability [17, 18].

Historically, N=1 and self-experimentation designs have played a non-trivial role in scientific discovery when used to test feasibility, mechanism plausibility, or procedural viability. Canonical examples include Ebbinghaus's systematic self-experiments on memory, which later became foundational to experimental psychology; Marshall's self-inoculation to test the bacterial hypothesis of peptic ulcers, subsequently validated through independent clinical trials; and Forssmann's self-catheterization to demonstrate procedural feasibility in cardiology [19, 20, 21].

These precedents illustrate that N=1 studies can be scientifically legitimate when explicitly framed as system-level evaluations and when embedded within a broader validation trajectory. The present framework adopts this position while adding explicit procedural constraints designed to mitigate self-referential bias [15, 16, 17].

4 Threat Model: Failure Modes & Biases in N=1

Table 1: Failure modes and mitigation strategies in N=1 cognitive workflow research

Failure Mode	**Detection Signal**	**Mitigation**	**Invalidation Condition**
Retrospective metric redefinition	Metrics appear post-hoc	Metric lock-in via preregistration	Metrics altered after data collection
Criterion adjustment	Shifting success thresholds	Predefined failure criteria	Thresholds modified silently
Operator–observer conflation	Narrative reinterpretation	Role separation and decision logs	Unlogged interpretive changes
Narrative self-confirmation	Selective session reporting	Mandatory full session archive	Cherry-picked data
Configuration drift	Unversioned prompts or settings	Hash-based versioning	Missing version identifiers

The failure modes listed in Table 1 are not exhaustive, but they represent the most frequent sources of methodological invalidation observed in self-referential N=1 research contexts. Importantly, these failures are often subtle and may occur without explicit intent. Similar vulnerabilities have been documented across psychological and computational research when procedural constraints are weak or implicit [6, 11, 13].

5 The Protocol

5.1 Unit of Analysis

A single session, defined as one complete execution cycle of the cognitive workflow.

5.2 Inputs

Predefined prompts, tasks, or stimuli, fixed prior to execution.

5.3 Procedure

Execution follows a scripted template with no discretionary modifications during runtime.

5.4 Outputs

All outputs are stored verbatim, time-stamped, and linked to configuration hashes.

5.5 Versioning

Any change to prompts, scripts, or parameters requires a new version ID and changelog entry.

5.6 Storage

- OSF Project: protocol, preregistration, logs

- GitHub/Zenodo (optional): scripts, versioned artifacts

The protocol is designed to impose temporal and procedural constraints across three phases: pre-execution, execution, and post-execution. Prior to execution, all inputs, metrics, configuration parameters, and evaluation criteria must be finalized and versioned. During execution, discretionary intervention is prohibited, except where explicitly defined by the protocol and documented in the session log. Post-execution, outputs are preserved verbatim, and no retrospective filtering or reinterpretation is permitted.

A session constitutes the minimal unit of analysis and must be uniquely identifiable through a session identifier linked to the corresponding configuration hash. Repeated sessions are permitted only under identical conditions or under explicitly versioned protocol updates. Any deviation from this rule constitutes configuration drift and invalidates comparability [13, **?**].

Manual interventions, when unavoidable, must be explicitly logged and justified. Undocumented interventions are treated as protocol violations. This approach prioritizes procedural integrity over outcome interpretation and ensures that auditability is maintained regardless of the substantive content of the outputs [5].

6 Pre-Registration and Transparency Plan

Prior to data collection, the following will be frozen:

- Hypotheses (if any)
- Metrics and scoring rules
- Exclusion criteria
- Stopping rules
- Planned analyses

Preregistration will occur on OSF, with explicit public/private data declarations aligned with OSF Preprints fields.

Preregistration serves, in this framework, not as a guarantee of methodological quality, but as a public commitment mechanism. By freezing hypotheses, metrics, and evaluation rules prior to execution, the protocol reduces interpretive flexibility and provides external reviewers with a stable reference point against which deviations may be assessed [9, 8].

The preregistration record will explicitly declare which components are publicly available and which are restricted, together with justifications. This approach aligns with OSF Preprints practices and emphasizes that transparency is achieved through procedural clarity rather than unrestricted disclosure [12, 14].

The reporting checklist applies exclusively to protocolized N=1 evaluations conducted at the full auditability level; exploratory or partially constrained implementations fall outside its scope.

7 Reporting Standard / Checklist

This section formalizes the minimum reporting requirements necessary to ensure auditability and methodological transparency in N=1 cognitive workflow studies. Rather than serving as a post hoc

reporting aid, the checklist functions as an ex ante constraint on researcher degrees of freedom by specifying which artifacts, decisions, and documentation must be preserved and disclosed [22, 23].

The checklist emphasizes procedural completeness over outcome interpretation. Compliance is binary: items are either satisfied or constitute a protocol violation. Failure to meet mandatory reporting requirements invalidates the evaluative status of the study, regardless of the substantive content of the results [13].

Table 2 summarizes the required reporting elements and their corresponding transparency conditions.

Table 2: Reporting standard and transparency checklist for auditable N=1 studies

Item	**Requirement**
Protocol version	Explicitly stated
Configuration hash / ID	Required
Raw logs	Publicly available
Changelog	Required and versioned
Scripts / templates	Publicly available
Manual interventions	Explicitly justified
Failure criteria	Explicitly defined
Conflict of interest (COI) statement	Required
Ethics statement	Required

8 Path to Community Review (PCI Meta-Research)

Following dissemination as a preprint on MetaArXiv, the protocol will be submitted to PCI Meta-Research for community-based evaluation prior to any empirical execution. Community review at this stage is intended to assess the clarity, completeness, and enforceability of the protocol's constraints rather than to evaluate outcomes, aligning with emerging peer-review models focused on methodological robustness [24, 25].

Community review at this stage serves a distinct function from traditional peer review. Rather than evaluating results, reviewers are invited to assess the clarity, completeness, and enforceability of the protocol's constraints, including its failure modes, bias mitigation strategies, and audit criteria. Feedback is expected to inform refinements to the protocol while preserving version history and traceability.

By integrating PCI Meta-Research into the pre-experimental phase, the framework explicitly treats methodological robustness as a collective process rather than an individual assertion.

9 Limitations

This manuscript reports no empirical findings. The proposed framework does not demonstrate effectiveness, performance gains, or substantive benefits of any cognitive system or workflow. Its contribution is limited to defining procedural constraints that enable auditability and reproducibility in N=1 research contexts [18, 17].

The protocol does not support population-level inference, generalization, or claims of external validity. Additionally, compliance with the reporting checklist and preregistration plan cannot eliminate all sources of bias, particularly those arising from unanticipated interactions or undocumented contextual factors [5].

The protocol defines an upper bound of methodological rigor rather than a mandatory implementation baseline. Given the operational and cognitive demands imposed by full auditability, it is both expected and acceptable that empirical implementations may adopt a staged approach. Exploratory interaction without protocol constraints (Level 0), minimal protocolized evaluation with fixed inputs and basic logging (Level 1), and fully specified audit-ready execution (Level 2) represent conceptually distinct modes of use. Crucially, only the latter qualifies as an auditable N=1 evaluation under the present framework. Lower levels are not treated as deficient, but as explicitly non-evaluative. This stratification is introduced to prevent category errors, reduce implicit overclaiming, and align methodological expectations with practical feasibility.

This manuscript does not commit to the execution of an empirical N=1 study. As a protocol-first contribution, its primary function is to specify methodological constraints, failure criteria, and reporting requirements prior to data collection. The absence of subsequent empirical execution does not invalidate the protocol, nor does it imply incomplete research. Protocols, registered reports, and methods papers commonly serve as normative references independent of downstream implementation. The present work is therefore evaluated on the clarity, enforceability, and transparency of its procedural specifications, not on the production of empirical outcomes.

Finally, the framework assumes good-faith adherence to declared procedures. While it reduces opportunities for self-referential bias, it cannot fully prevent deliberate misrepresentation. These limitations are inherent to protocol-based approaches and underscore the importance of community review and transparent versioning [2].

10 Conclusion

This manuscript proposed a protocol-first framework for conducting auditable and reproducible N=1 evaluations of human–AI cognitive workflows. Rather than advancing claims of effectiveness or performance, the framework prioritizes procedural constraints designed to reduce self-referential bias and limit researcher degrees of freedom [2, 3].

By formalizing failure modes, versioning rules, reporting requirements, and a pre-experimental community review pathway, the protocol reframes N=1 research as a legitimate system-level evaluative strategy when its scope is clearly defined and its claims appropriately constrained [5, 4].

In this sense, the framework advances a model of transparency that treats methodological robustness as a collective and procedural achievement rather than a post hoc narrative of results [9].

References

[1] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

[2] John P. A. Ioannidis. How to make more published research true. *PLoS Medicine*, 11(10):e1001747, 2014.

[3] Marcus R. Munafò et al. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021, 2017.

[4] Brian A. Nosek et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[5] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12, 2016.

[6] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology. *Psychological Science*, 22(11):1359–1366, 2011.

[7] Norbert L. Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.

[8] Christopher D. Chambers et al. Registered reports: A new publishing initiative. *Cortex*, 49(3):609–610, 2014.

[9] Brian A. Nosek et al. The preregistration revolution. *PNAS*, 115(11):2600–2606, 2018.

[10] Andrew Gelman and Eric Loken. The garden of forking paths. *Unpublished manuscript*, 2013.

[11] Jelte M. Wicherts et al. Degrees of freedom in planning, running, analyzing, and reporting psychological studies. *Frontiers in Psychology*, 7:1832, 2016.

[12] Roger D. Peng. Reproducible research in computational science. *Science*, 334:1226–1227, 2011.

[13] Geir K. Sandve et al. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 2013.

[14] Victoria Stodden et al. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016.

[15] Gordon H. Guyatt et al. Determining optimal therapy—randomized trials in individual patients. *New England Journal of Medicine*, 314:889–892, 1990.

[16] Richard L. Kravitz et al. Design and implementation of n-of-1 trials. *Journal of Clinical Epidemiology*, 67(4):343–350, 2014.

[17] Alan E. Kazdin. *Single-case research designs*. Oxford University Press, 2011.

[18] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and quasi-experimental designs*. Houghton Mifflin, 2002.

[19] Hermann Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers College, Columbia University, 1913.

[20] Barry J. Marshall and J. Robin Warren. Attempt to fulfil koch's postulates for pyloric campylobacter. *Medical Journal of Australia*, 142:436–439, 1985.

[21] Werner Forssmann. Die sondierung des rechten herzens. *Klinische Wochenschrift*, 8:2085–2087, 1929.

[22] Robyn L. Tate et al. The single-case reporting guideline in behavioural interventions (scribe 2016). *Archives of Scientific Psychology*, 4(1):1–9, 2016.

[23] CENT Group. The cent 2015 statement. *BMJ*, 350:h1738, 2015.

[24] Christophe Roux et al. Peer community in: a new model for peer review. *F1000Research*, 9:153, 2020.

[25] Jonathan P. Tennant et al. The state of the art in peer review. *FEMS Microbiology Letters*, 365(19):fny204, 2019.