Is there already a standard ?

7 Sep 2026

Is there already a standard for this?

Before proposing an open standard for qualitative text processing workflows in evaluation, we went looking for one. This page reports what we found, so that anyone who thinks the ground is already occupied can check.

The short answer is that it is not. There is no open standard, and no draft of one, for specifying a qualitative text-processing workflow in evaluation. Several standards hold one of the properties such a thing would need. None holds two.

Work in progress. This is a survey done in September 2026, not a literature review. It says what we looked at and what we could not verify.

Three properties, and nothing that has more than one

A standard worth having would need to do three things at once.

- **Be executable.** The specification is what runs the analysis, rather than a form somebody fills in afterwards.
- **Carry a claim back to a quote.** An assertion in a report resolves to the span of source text that supports it.
- **Fix the rule before the results.** The threshold, the rubric and the decision rule are registered, dated and versioned before anything is read.

Each of these exists somewhere. The W3C Web Annotation Data Model, a Recommendation since February 2017, solves anchoring properly: a passage can be identified by the words it contains as well as by its character offsets, so it survives an edit to the document. Canada's Algorithmic Impact Assessment is a computable decision rule, with weights and thresholds encoded in the questionnaire itself. Workflow languages such as the Common Workflow Language, and packaging standards such as RO-Crate, describe a run and its provenance in a form other software can read.

But the anchoring standard says nothing about what to ask, the decision rule runs over unevidenced free text, and the workflow standards trace files rather than assertions. Nobody has joined them.

The pile the evaluation profession is standing in

Everything the profession itself publishes falls into one of three groups, and none of them is the thing.

- **Ethics and quality principles.** The Program Evaluation Standards, the American Evaluation Association's Guiding Principles, the OECD DAC quality standards, the UNEG norms. Prose, aimed at judgement, silent on analysis procedure.
- **Reporting checklists.** COREQ, SRQR, ENTREQ, RAMESES, PRISMA. These standardise what you must say you did.
- **Data formats and inference engines.** REFI-QDA for coded data, the QCA packages, Bayesian process-tracing tools. Machine-readable, and each of them starts downstream of the step where text becomes data.

The pattern is consistent enough to be worth naming. Every method formalises everything except the single operation that touches the words. Qualitative comparative analysis formalises Boolean minimisation and then describes calibration as an interpretative act. Formal process tracing gives a vocabulary for classifying evidence after the inference has been made, and no procedure for assigning a passage to a class (Ricks & Liu 2018). Outcome Harvesting specifies six steps and leaves who substantiates an outcome, and how many people to ask, to the credibility the use requires ({Wilson-Grau 2018). The Most Significant Change guide states that selection criteria should not be decided in advance but should emerge through discussion, and offers five incompatible ways of deciding, without naming a default and without requiring anyone to record which was used (Davies & Dart 2005).

This is not an oversight, and it would be unfair to read it as one. The profession has considered the question and declined. The Magenta Book's own guide to quality in qualitative evaluation says its framework was "devised to aid informed judgement, not mechanistic rule-following", and notes in passing that "in qualitative research, there are no 'validated' instruments or standardised methods". Its last appraisal question is auditability, "How adequately has the research process been documented?", and the indicators under it are prose. Bricolage, mixing and adapting methods to the case in hand, is what most evaluators actually do, and it is defended rather than apologised for (Aston & Apgar 2022; Apgar et al. 2024).

An open standard has to live with that. The point is not to replace judgement with a rule, but to make it possible to say where the judgement was, and what it was given to work with.

The one domain standard, and what it cannot say

Qualitative analysis does have a real interchange standard. REFI-QDA, published in 2019 under an MIT licence, moves a coded project between ATLAS.ti, MAXQDA, NVivo, Dedoose, Quirkos and others, and an independent tooling ecosystem has grown up around it. It holds the documents, the codebook, the coded segments and their character offsets.

It describes the result rather than the process. Across the sixty pages of the specification, the words *audit*, *provenance*, *workflow* and *history* appear no times at all. A coding can say who made it and when, and cannot say that it was ever revised. The codebook is not versioned at all.

And its definition of a user is "A person who has worked on a project." There is no agent type. A model that codes a corpus can appear in a REFI-QDA file only by being entered as a person. The specification is candid about its own scope, saying that round-tripping "may not work without data loss" and that "the focus in developing the standard was only on exchange".

That is the gap, stated by the standard itself.

What comes closest

Two things are worth knowing about, because they are nearer than anything in evaluation.

The **Reproducible Open Coding Kit** is the only standard we found that was designed for process rather than for data (Peters & Z{\"o 2026). Codes, utterance identifiers and a codebook are written as plain-text conventions inside the source file, so the whole analysis is legible to a person and to a machine, and because it is plain text under version control the coding history can be reconstructed. It is small, academic and unsupported by any commercial software, and it is the right idea.

The **joint position statement on AI in evidence synthesis**, issued in November 2025 by Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence, is the nearest thing to a testable rule that any body has published (Flemyng & {Noel-Storr 2025). Its requirement is that "Authors should declare when they have used AI if it makes or suggests judgements", and it names the judgement tasks: eligibility, risk of bias, data extraction, synthesis, certainty assessment. Nothing implements it.

Why the question is not academic

Two measurements make the case better than any argument about principle.

The first is what configuration choices do to conclusions. Replicating 37 annotation tasks from 21 published social science studies, across 18 models and 13 million labels, one study found incorrect conclusions in about 31% of hypotheses even for the best models, and showed that "with just a handful of prompt paraphrases, virtually anything can be presented as statistically significant" (Baumann & R{\"o 2025). The choices that produce that variation are exactly the ones no current standard requires anybody to write down.

The second is what people currently report. Across 75 studies using large language models in qualitative research between 2020 and 2025, 75% reported no parameter settings at all, 13 gave a temperature and 4 gave a top_p value, and agreement between model and human coders ranged from 36% to 99% (Kempny et al. 2026). A reader cannot tell the good end of that range from the bad end, because the reports do not contain what would distinguish them.

This is the older problem in a faster form. The case for making qualitative claims traceable to the passages behind them was made well before any of this (Moravcsik 2014), and the argument that the technology strengthens rather than threatens the commitments of qualitative inquiry is being made now (Wise et al. 2026). What has changed is that a machine which reads can be made to show its working, so the reason for not showing it has weakened.

Where the opening is

Two positions are unoccupied.

The first is the join. Anchoring is solved and says nothing about method; decision rules are expressible and float free of any evidence; run provenance is standardised and reaches files rather than claims. Putting the three together is the whole of the gap.

The second is the domain. The only registered reporting guideline aimed at qualitative work with language models is an extension of COREQ, whose protocol was published in September 2025 with a target of December that year and which had not appeared by September 2026. It is a prose checklist, and it addresses neither traceability nor pre-registration. A specification for evaluation rather than health, executable rather than declarative, would not collide with it.

Rubicon is our attempt at the second, built so that the first is possible. What it does, and the principles behind it, are on the Rubicon page.

What we checked, and what we could not

We searched the evaluation profession's standards bodies and method manuals, the qualitative data and annotation formats, the workflow and provenance standards, the reporting-guideline family indexed by EQUATOR, and the AI transparency and audit instruments including the EU AI Act, the ISO/IEC 42000 series and the national algorithm registers.

Quotations from the REFI-QDA specification, the Magenta Book guide, the Cochrane and partners position statement, and the two studies cited for figures were each read in the original rather than taken from a summary. Where this page says that something does not exist, that is a report of what our search found rather than a proof of absence.

One claim we wanted to make had to be dropped. The Dutch national algorithm register is the best-specified public register we saw, with a published schema and bulk export, and it is reported to have a large majority of entries with the impact-assessment field left empty. We could not reach the data to check the proportion, so the figure is not quoted here. The point it would have made is worth keeping anyway: a schema alone changes nothing, and a standard nobody has a reason to complete is a standard in name.

References

Apgar, Bradburn, Rohrbach, Wingender, Cubillos Rodriguez, & Arias (2024). *Rethinking Rigour to Embrace Complexity in Peacebuilding Evaluation*. Evaluation, 13563890241232405. https://doi.org/10.1177/13563890241232405.

Aston, & Apgar (2022). *The Art and Craft of Bricolage in Evaluation, CDI Practice Paper 24*. https://doi.org/10.19088/IDS.2022.068.

Baumann, & R{\"o (2025). *Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation*. arXiv.

Davies, & Dart (2005). *The Most Significant Change Technique, A Guide to Its Use*.

Flemyng, & {Noel-Storr (2025). *Position Statement on Artificial Intelligence (AI) Use in Evidence Synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025*. Cochrane Database of Systematic Reviews. https://doi.org/10.1002/14651858.ED000178.

Kempny, Frings, Rust, Meister, & Fehring (2026). *The Use and Methodological Reporting of Large Language Models in Qualitative Research: A Scoping Review*. BMC Medical Research Methodology, *26*. https://doi.org/10.1186/s12874-026-02913-1.

Moravcsik (2014). *Transparency: The Revolution in Qualitative Research*. APSCC Newsl., *47*, 48--53. https://doi.org/10.1017/S1049096513001789.

Peters, & Z{\"o (2026). *The Reproducible Open Coding Kit: A Human and Machine Readable Standard for Coding Qualitative Data*. PsyArXiv. https://doi.org/10.31234/osf.io/4q8bs_v2.

Ricks, & Liu (2018). *Process-Tracing Research Designs: A Practical Guide*. PS - Political Science and Politics, *51*, 842--846. https://doi.org/10.1017/S1049096518000975.

Wise, Gresalfi, & {Spencer-Smith (2026). *Why AI Is Not the Enemy: Opportunities to Strengthen Core Commitments of Qualitative Inquiry Through Trustworthy AI-in-the-Loop Analysis*. International Journal of Qualitative Methods, *25*, 16094069261435579. https://doi.org/10.1177/16094069261435579.

{Wilson-Grau (2018). *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.