🌻 A workflow for collecting and understanding stories at scale – Summary (eval2025)

23 Dec 2025

(Powell et al. 2025)

Source: *Evaluation* 31(3), 394–411 (2025).

- **What problem this paper solves**
- Evaluations often start with a ToC and then collect evidence for each link (e.g. Contribution Analysis), but in many real settings the ToC is uncertain, contested, or incomplete.
- The paper proposes collecting evidence about **structure/theory** (what influences what) and **contribution** simultaneously, using a scalable workflow that stays open-ended.

- **Core idea: "AI-assisted causal mapping pipeline"**

- Treat causal mapping as **causal QDA**: each coded unit is an ordered pair **(influence → consequence)** with provenance, rather than a theme tag.
- Use AI as a **low-level assistant** for interviewing + exhaustive extraction, leaving high-level judgement (prompt design, clustering choices, interpretation) with the evaluator.

- **Pipeline (end-to-end)**

- **Step 1 — AI interviewer**: a single LLM "AI interviewer" conducts semi-structured, adaptive chat interviews at scale.
- **Step 2 — Autocoding causal claims**: an LLM is instructed (radical zero-shot) to list *each* causal link/chain and to ignore hypotheticals.
- **Step 2c — Clustering labels**: embed factor labels and cluster them; then label clusters and optionally do a second "deductive" assignment step to ensure cluster cohesion.
- **Step 3 — Analysis via maps/queries**: produce overview maps, trace evidence for (direct/indirect) contributions, compare subgroups/timepoints.

- **Demonstration study (proof-of-concept)**

- Respondents: online workers recruited via Amazon MTurk; topic: "problems facing the USA" (chosen to elicit causal narratives without a specific intervention).
- Data collection repeated across **three timepoints**; data pooled.
- This is an analogue demonstration; not intended to generalise substantively about "the USA".

- **Key results (reported metrics)**

- **AI interviewing acceptability (proxy)**: 78.5% of interviewees did not ask for changes to the AI's end-of-interview summary; 4.29% asked for changes; 15.3% had no summary (drop-off).
- **Autocoding effort/cost**: ~5 hours to write/test coding instructions; ~$20 API cost (in the reported experiment set-up).
- **Autocoding recall/precision**:
  - Ground-truth link count (authors' assessment): 1154 links.
  - AI-identified links: 1024 (≈ 89%) before precision screening.
  - Precision scoring (0–2 on four criteria: correct endpoints; true causal claim; not hypothetical; correct direction): 65% perfect; 72% dropped only one point.
- **Overview-map "coding coverage"**

  - An 11-factor overview map (plus filters) covered ~42% of raw coded claims while remaining readable.
  - Coarse clustering can collapse opposites/valence (e.g. "military strengthening" and "military weakening" both under "International conflict").
- **Interpretation claims**

- The approach is good for sketching **"causal landscapes"** and triaging hypotheses; it is not reliable enough for high-stakes single-link adjudication without human checking.
- Many outputs depend on **non-automated clustering decisions** (number of clusters, labelling intervention), analogous to researcher degrees of freedom in variable construction.

- **Caveats / ethics**

- Not suitable for **sensitive data** when using third-party LLM APIs; risks of bias and hegemonic worldviews are highlighted.
- Differential response/selection into AI interviewing may not be random.
- Causal mapping shows **strength of evidence**, not **effect size**; forcing magnitudes/polarity is risky.

References

Powell, Cabral, & Mishan (2025). *A Workflow for Collecting and Understanding Stories at Scale, Supported by Artificial Intelligence*. SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/13563890251328640.