

The AI Co-Ethnographer: How Far Can Automation Take Qualitative Research?

Fabian Retkowski, Andreas Sudmann, Alexander Waibel

The Triple Challenge of Qualitative Research

Scale Limitations

- Typically limited to dozens of interviews/observations

Inter-Annotator Reliability

- Same data, different conclusions

Intra-Annotator Variability

- Monday you  ≠ Friday you 

The AI Co-Ethnographer 🤖

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication. I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication.

I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

Medication, Patient's Commitments

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication. I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

Medication, Patient's Commitments

Lifestyle, Patient's Commitments

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication. I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

Medication, Patient's Commitments

Lifestyle, Patient's Commitments

Decision Making

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication. I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

Medication, Patient's Commitments

Lifestyle, Patient's Commitments

Decision Making



Multilabel classification task

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Well, I'm committed to taking the medication. I don't see a cardiologist anymore, he told me I didn't have to, he referred me to my GP. I'm committed to exercise and manage my weight. The only thing I could do better at is drink less wine, because I do like wine. It's one of my objectives. I'll drink about a couple of glasses a day. The decision to stay in shape and manage my weight and try to improve my lifestyle it's certainly something I take into account given my cardiovascular health.

Medication, Patient's Commitments

Lifestyle, Patient's Commitments

Decision Making



Classes that are totally made up

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Inductive Coding versus **Deductive Coding**

→ *In practice, typically data-driven (inductive) coding*

The *Qualitative Coding* Task



The central bridge between raw data and interpretation.

Inductive Coding versus **Deductive Coding**

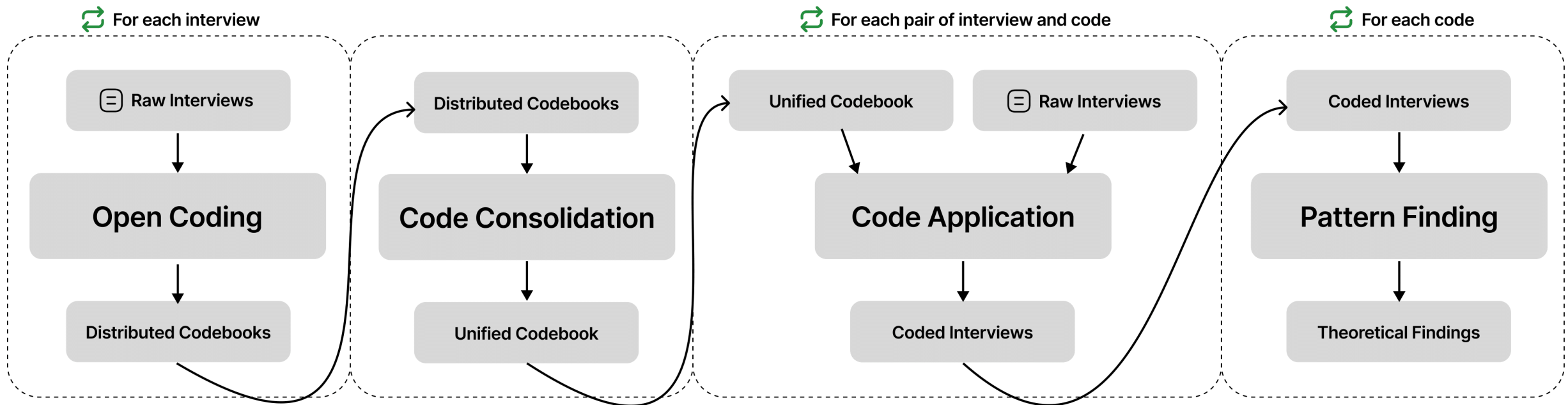
Open Coding and **Axial Coding**

→ *In practice often messy, subjective, ...*



The AI Co-Ethnographer (AICoE)

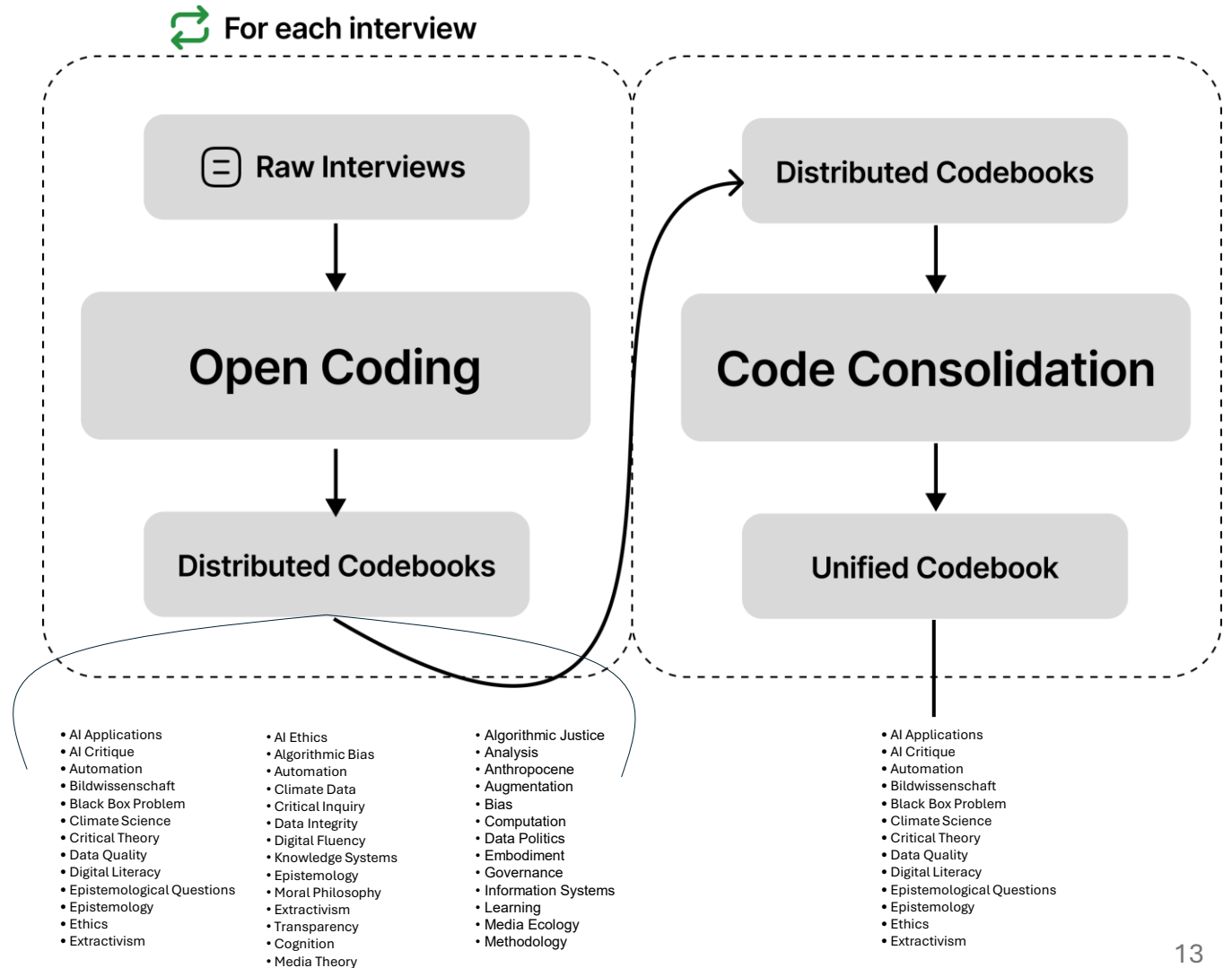
LLM-based pipeline mimicking several key qualitative research processes



Codebook Development

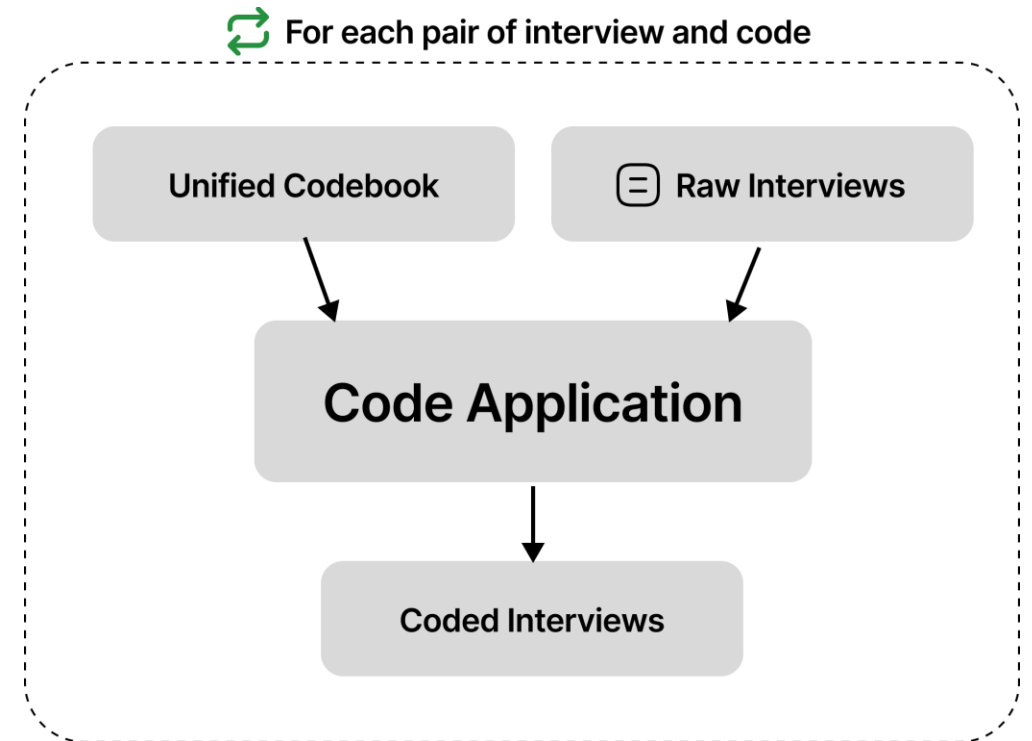
Two-Stage Process accommodates both

- Grounded Theory („Open“ & „Axial“)
- Context Window Limitations



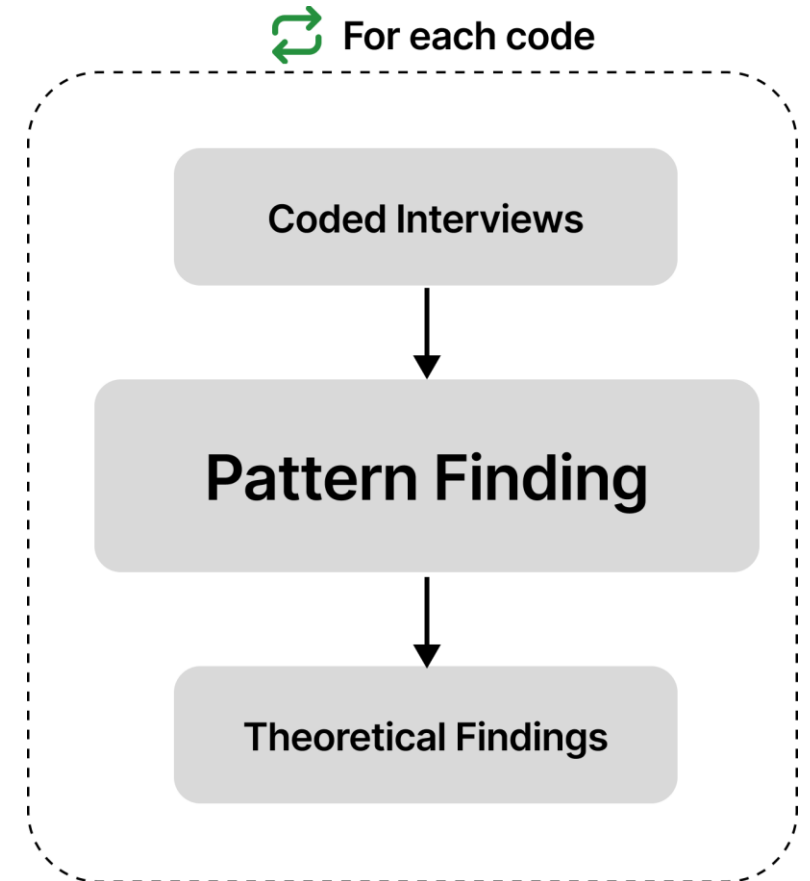
Code Application

- Full Context
- Prompt Caching
- Coding via the extraction of relevant passages





Pattern Finding

Informed by coded segments
across the entire set of interviews



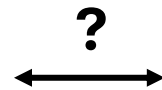
Experimental Setup

- Llama-3.3-70B 
- Interviews
 - CVDQuoading ([Spinoso et al.](#))
 - HiAICS Interviews (ours)
- Human Evaluation with Experts 

Semantic Relatedness of Codebooks

- Prior work focused on code assignment, not codebook development

- AI Critique
- AI for Science
- Algorithm
- Algorithmic Biases
- Autonomy & Agency
- Biographical Context
- Black Box
- Data
- Epistemic and Infrastructural of Media
- Expert Systems
- Expertise & Competence
- Facial Recognition
- First Encounters with AI
- (...)



- AI Applications
- AI Critique
- Automation
- Bildwissenschaft
- Black Box Problem
- Climate Science
- Critical Theory
- Data Quality
- Digital Literacy
- Epistemological Questions
- Epistemology
- Ethics
- Extractivism
- (...)

Semantic Relatedness of Codebooks

- Prior work focused on code assignment, not codebook structure

- **AI Critique**

- AI for Science
- Algorithm
- Algorithmic Biases
- Autonomy & Agency
- Biographical Context
- Black Box
- Data
- Epistemic and Infrastructural of Media
- Expert Systems
- Expertise & Competence
- Facial Recognition
- First Encounters with AI
- (...)



- AI Applications

- **AI Critique**
- Automation
- Bildwissenschaft
- Black Box Problem
- Climate Science
- Critical Theory
- Data Quality
- Digital Literacy
- Epistemological Questions
- Epistemology
- Ethics
- Extractivism
- (...)

Semantic Relatedness of Codebooks

- Prior work focused on code assignment, not codebook structure

- **AI Critique**

- AI for Science
- Algorithm
- Algorithmic Biases
- Autonomy & Agency
- Biographical Context
- Black Box

- **Data**

- Epistemic and Infrastructural of Media
- Expert Systems
- Expertise & Competence
- Facial Recognition
- First Encounters with AI
- (...)



- AI Applications

- **AI Critique**

- Automation
- Bildwissenschaft
- Black Box Problem
- Climate Science
- Critical Theory

- **Data Quality**

- Digital Literacy
- Epistemological Questions
- Epistemology
- Ethics
- Extractivism
- (...)

Semantic Relatedness of Codebooks

Relationship Types

- (M) Match
- (C) Containment
- (P) Partial Overlap
- (U) Unmatched

$$s(x) = \max(\{w_r : r \in R(x)\} \cup \{0\})$$

$$\tau_i = \frac{1}{|i|} \sum_{x \in i} s(x) \quad \text{for } i \in \{A, B\}$$

$$\tau_{sem} = \frac{\tau_A + \tau_B}{2}$$

Semantic Relatedness of Codebooks

Relationship Types

- (M) Match
- (C) Containment
- (P) Partial Overlap
- (U) Unmatched

$$s(x) = \max(\{w_r : r \in R(x)\} \cup \{0\})$$

$$\tau_i = \frac{1}{|i|} \sum_{x \in i} s(x) \quad \text{for } i \in \{A, B\}$$

$$\tau_{sem} = \frac{\tau_A + \tau_B}{2}$$

| Schema 1 | Schema 2 | M | C | P | U | τ_{sem} |
|----------|----------|-------|-------|-------|-------|--------------|
| Coder A | Coder B | 0.216 | 0.346 | 0.251 | 0.187 | 0.584 |
| Coder A | AICoE | 0.206 | 0.480 | 0.191 | 0.123 | 0.638 |
| Coder B | AICoE | 0.081 | 0.573 | 0.125 | 0.221 | 0.545 |

- ◆ AICoE is no more divergent from humans than humans are from each other
- ◆ Each codebook also contributed unique insights
→ AICoE can complement human coding

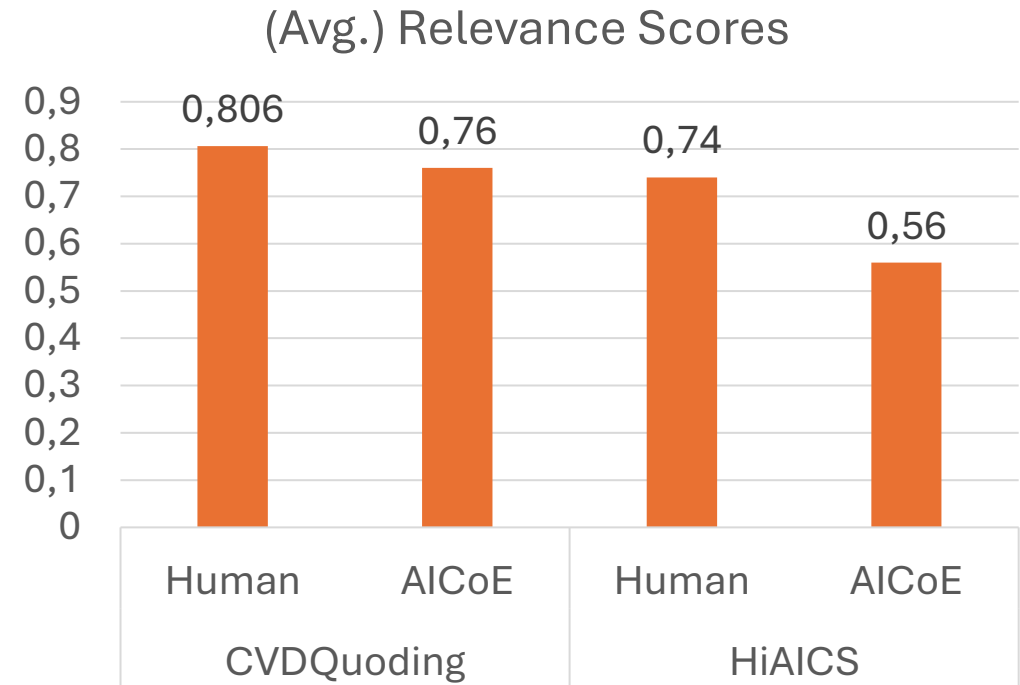
Code Assignments

Evaluation Protocol

- Used expert-curated codebooks for deductive coding
- Experts rated blinded code-to-text assignments (humans & AI)

! Performance Gap (HiAICS > CVDQuoding)

- Due to inherent data characteristics



Quality of Findings



Human Evaluation Criteria

- **Grounding (G):** Supported by data (accurate + referenced)
- **Relevance (R):** Matches research goals and code
- **Insight (I):** Non-trivial, surprising, intellectually valuable

- ◆ *AICoE can produce well-grounded, relevant findings*
- ◆ *Insight is hardest to automate*
- ◆ *1 in 3 codes yielded at least one expert-rated high-quality finding*

| | Mean | SD | % HQ |
|------------------------|-------------|-------------|--------------|
| Grounding | 3.42 | 0.61 | – |
| Relevance | 3.76 | 0.41 | – |
| Insight | 3.29 | 0.46 | – |
| Overall Quality | 3.49 | 0.38 | 32.25 |

| Criterion | E1-E2 | E1-E3 | E2-E3 |
|-----------|---------|---------|--------|
| Grounding | -0.0430 | 0.0269 | 0.6471 |
| Relevance | 0.0064 | 0.0603 | 0.1194 |
| Insight | 0.0846 | -0.0384 | 0.2478 |

Correlation Coefficients between Evaluators

AI-Augmented Workflows

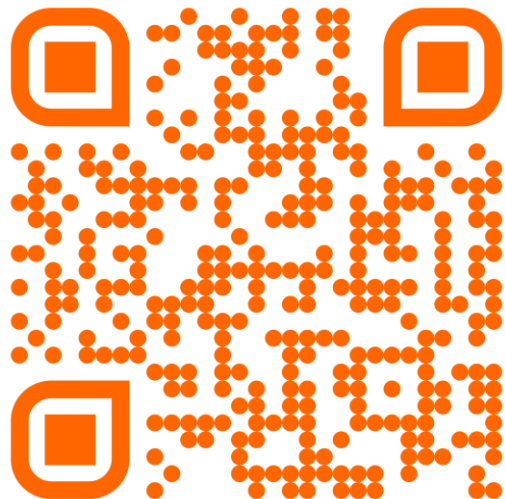
- Every stage allows intervention
- *Unified codebook* as critical checkpoint
- Also supports deductive coding by skipping codebook development stages
- Researchers can iterate through stages multiple times, run parallel samples, or modify intermediate outputs as needed
- Inspiration for your own codebooks and findings

Conclusion

AICoE shows robust codebook development, reasonable code assignments, and the ability to generate meaningful findings

A promising direction for qualitative research

HiAICS Project



Paper

