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

Fabian Retkowski[1], **Andreas Sudmann**[2], **Alexander Waibel**[1,3]
[1]Karlsruhe Institute of Technology, Germany
[2]University of Bonn, Germany
[3]Carnegie Mellon University, USA
{retkowski,waibel}@kit.edu / asudmann@uni-bonn.de

Abstract

Qualitative research often involves labor-intensive processes that are difficult to scale while preserving analytical depth. This paper introduces The AI Co-Ethnographer (AICoE), a novel end-to-end pipeline developed for qualitative research and designed to move beyond the limitations of simply automating code assignments, offering a more integrated approach. AICoE organizes the entire process, encompassing open coding, code consolidation, code application, and even pattern discovery, leading to a comprehensive analysis of qualitative data.

1 Introduction

Qualitative data analysis is a crucial research approach in the humanities, cultural studies, and social sciences, focusing on the synchronic and diachronic analysis and interpretation of non-numerical data such as texts, images, or audio files to gain insights into complex social phenomena, cultural expressions, and individual experiences (Creswell and Poth, 2017; Denzin et al., 2023). Coding is central to this process, structuring and interpreting research materials such as interviews, field notes, or group discussions by systematically assigning analytically relevant concepts to text segments or other data forms (Holton, 2007; Bernard, 2011; Harding, 2013; Bernard et al., 2016).

Although coding offers a formalized structure for data analysis, its application remains context-specific and flexible, adapting to the nuances of the research question and subject matter (Elliott, 2018). In many contexts, specifically in ethnographic approaches, coding is inherently iterative and closely tied to an ongoing process of collecting and reflecting on data. Codes evolve dynamically through an iterative process where they are merged, adjusted, added, or replaced as researchers engage with the data, identify patterns, and refine their conceptual understanding. This process may involve open or axial coding, deductively or inductively, quantitatively or qualitatively, and can be centered on interpretation or description. (Ritchie et al., 2014; Creswell, 2015; Saldana, 2015).

However, manual coding faces significant limitations. Scalability remains a critical challenge when researchers encounter larger datasets that require extensive time and resources to code effectively (Miles et al., 2019). It also increases the risk of intra- and intercoder unreliability, just to mention a few typical challenges. These constraints have spurred interdisciplinary efforts to automate the coding process over the past decade. Automated speech recognition (ASR) has emerged as a significant enabler in this landscape, allowing researchers to efficiently transcribe large volumes of interview data and prepare them for further analysis and processing (Nguyen et al., 2021). Related qualitative data processing tasks such as text summarization (Hori et al., 2002; Retkowski and Waibel, 2024b; Zhang et al., 2024), question answering (Singhal et al., 2025), and topic segmentation (Zechner and Waibel, 2000; Retkowski and Waibel, 2024a) have similarly benefited from computational advancements, providing researchers with tools to condense information and identify thematic boundaries.

Recently, large language models (LLMs) have demonstrated new epistemic capabilities to annotate research data, yet with certain limitations, such as understanding the broader context of codes (Tuschling et al., 2023; Fischer and Biemann, 2024; Rasheed et al., 2024; Ziems et al., 2024). In parallel, the concept of Agentic LLMs has emerged, designed to operate autonomously with goal-directed behaviors (Xi et al., 2023). For example, the *AI Scientist* (Lu et al., 2024) showcased an end-to-end automated workflow for writing scientific papers, from hypothesis generation, experimental design and manuscript drafting. This work illustrates the potential for autonomous agents to manage complex, multi-stage research processes. Inspired by

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 73–90
May 3-4, 2025 ©2025 Association for Computational Linguistics

these advances, our approach seeks to explore similar automation in the domain of qualitative research, also as an alternative to AI-assisted data analysis with proprietary systems like MaxQDA.

With the AI Co-Ethnographer (AICoE), we introduce a novel end-to-end pipeline that extends beyond the conventional focus on code assignments. The AICoE is part of a broader infrastructure for AI-assisted knowledge production, integrating diverse qualitative analysis methods, from open coding to pattern discovery. Whereas prior research has largely concentrated on automating the mapping of codes to text segments, our approach encompasses a more comprehensive qualitative analysis process. The pipeline extends the capabilities beyond the deductive application of pre-defined codes. Crucially, it also enables inductive code development and application, a process where novel codes are developed directly from the data itself instead of being pre-defined.

2 Related Research

LLM development has spurred transdisciplinary efforts to automate scholarly work, especially qualitative textual analysis (Morgan, 2023; Petersen-Frey et al., 2023; Fischer and Biemann, 2024; Lu et al., 2024; Franken and Vepřek, 2025), including ethnographically focused research (Dippel and Sudmann, 2023). This builds on a rich history of computational methods in qualitative research, from early tools like the General Inquirer (Stone and Hunt, 1963) and Salton‘s vector space model (Salton et al., 1975), to machine learning-based annotation (Sebastiani, 2002), and open-source platforms like WordFreak (Morton and LaCivita, 2003) and WebAnno (Yimam et al., 2014). More recently, Spinoso-Di Piano et al. (2023) introduced the Qualitative Code Suggestion (QCS) task, which assists in coding by providing a ranked list of predefined codes for a given text passage. To evaluate QCS, the authors present CVDQuoding, an annotated dataset of interviews with women at risk of cardiovascular disease. Human evaluation shows that their system provides relevant suggestions, highlighting its potential as an assistive tool. However, limitations remain, including a focus on code assignment rather than full codebook development and a lack of evaluation in applied research settings. Similarly, Ziems et al. (2024) evaluated the potential of LLMs for automating social science tasks, focusing on their zero-shot capabilities. Their findings indicate that LLMs demonstrate proficiency in both classification and explanation, suggesting their ability to augment the social science research pipeline. However, the authors do not recommend LLMs as a replacement for traditional methods.

3 Methodology

The AI Co-Ethnographer is composed of a comprehensive pipeline underpinned by LLMs to automate key qualitative research processes while aiming to preserve the interpretative depth central to ethnography. Building on recent advances in LLMs, the system mirrors several stages of qualitative analysis (see Figure 1): open coding, code consolidation, code application, and pattern finding. This approach enables scalable and consistent analysis of large volumes of qualitative data while mimicking ethnographic research practices.

3.1 Open Coding

A first step can be called *open coding*, where individual interviews are processed separately by the LLM. By isolating analyses per interview, the chosen research design addresses both the context window limitations of LLMs and the ethnographic principle of maintaining close connection to primary data. The system may suggest up to N codes per interview, balancing descriptive and interpretive coding approaches and, in doing so, automating a time-consuming element of qualitative analysis.

3.2 Code Consolidation

The *code consolidation* stage transitions to a global perspective and synthesizes findings across all interviews into a unified codebook. The synthesis process analyzes code overlap and merges similar concepts, culminating in a maximum of up to M consolidated codes. This stage represents a crucial bridge between individual narratives and broader theoretical development, akin to manual axial coding but computationally scaled.

3.3 Code Application

The pipeline returns to a local perspective in the *code application* stage, where each consolidated code is systematically applied to individual interview transcripts. Unlike existing approaches that work with limited text fragments (Spinoso-Di Piano et al., 2023), our system processes the entire interview for each code[1], thereby ensuring that

[1] We note that this approach allows for prompt caching for a more efficient application of the codes.

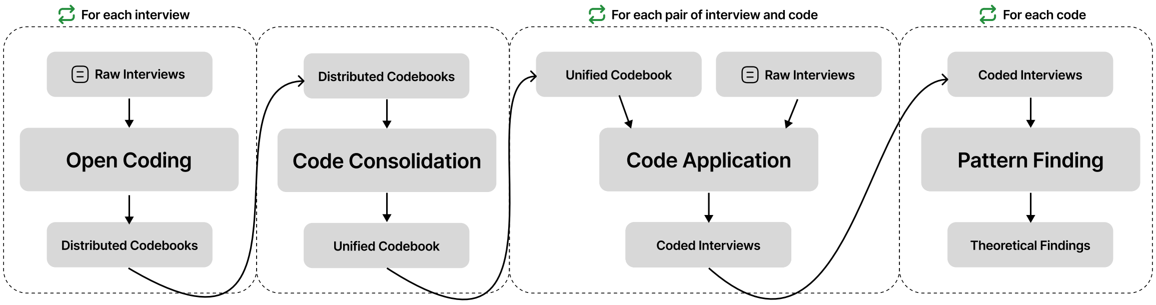


Figure 1: Conceptual Illustration of the AI Co-Ethnographer Pipeline

the full conversational context informs the identification of relevant passages. This preserves the crucial ethnographic context of when and where statements occur. The system maintains connections between codes and input data through extracted text segments that can be mapped back to the original interviews, primarily via unique exact matches or substring matches. In rarer cases when such a match is unavailable, we instead rely on a sufficiently large word overlap measure using ROUGE (Lin, 2004), addressing both the technical need for systematic analysis and the ethnographic requirement for contextual grounding.

3.4 Pattern Finding

Finally, the *pattern finding* stage shifts back to a holistic perspective, analyzing relationships between coded segments across the entire set of interviews to identify insights. This stage examines co-occurrence, contextual relationships, and thematic patterns, automating the transition from coding to broad theoretical and interpretative understanding.

3.5 Prompt Engineering

The developed prompts (see Appendix A) aim to emulate standard procedures in qualitative research, specifically in an ethnographic context. Each prompt corresponds to a phase of analysis and is structured to ensure methodological rigor. The scratchpad is important, as it allows the model to articulate its step-by-step reasoning, thereby making it transparent. By emphasizing verbatim text extraction and a strict correspondence between each extracted segment and the original interview line, we aim for high inter-rater reliability and transparency. Additionally, optional code descriptions during codebook development enhance clarity, and optional context helps guide the research direction. The `max_codes` parameter is a technical restriction to avoid overly lengthy prompts, but in practice can be adjusted according to factors such as the model's context length, its ability to maintain performance over long contexts, and the number of interviews. Although these prompts are illustrated with ethnographic interviews, the underlying principle of precise, code-based textual extraction readily extends to other qualitative research methodologies.

4 Experiments and Results

The system leverages `Llama-3.3-70B` (Dubey et al., 2024) as LLM, though the modular pipeline design permits integration with any modern LLM. We evaluate the model on three interviews each from the CVDQuoding and HiAICS datasets, the latter being our collection of interviews conducted as part of an ethnographic analysis with AI researchers. The study participants include both researchers who apply AI practically in their scientific disciplines and those who offer theoretical and critical analyses of AI's use in research. The interviews were transcribed using the speaker-attributed ASR system by Nguyen and Waibel (2025).[2]

4.1 Semantic Relatedness of Codebooks

To evaluate the semantic relatedness between different qualitative codebooks, we developed a novel framework for systematically comparing code taxonomies by specifying the following semantic relationships between codes:

- (M) **Match (1:1)** – Defines codes capturing broadly similar concepts across codebooks, though they may use different terminology
- (C) **Containment (1:n)** – Indicates when one code represents a broader concept encompassing one or more codes from the other scheme
- (P) **Partial Overlap (1:1)** – Denotes codes that share some aspects of their meaning while maintaining distinct elements

[2] We publish the HiAICS interviews under `https://codeberg.org/hiaics/interviews`.

- (U) **Unmatched** – Codes representing entirely unique aspects absent in the other codebook

A visual demonstration of these relations can be found in Figure 3 in the Appendix. Based on these relationships, we also developed a scoring method to quantify them. We normalize n:1 containments into atomic 1:1 relationships and assign weights for semantic relevance: $w_m = 1.0$ for matches, $w_c = 0.7$ for containments, and $w_p = 0.5$ for overlaps. For each code x, let $R(x)$ denote its set of relationships. The individual score $s(x)$, codebook scores τ_i, and final score are calculated as:

$$s(x) = \max(\{w_r : r \in R(x)\} \cup \{0\}) \quad (1)$$

$$\tau_i = \frac{1}{|i|} \sum_{x \in i} s(x) \quad \text{for } i \in \{A, B\} \quad (2)$$

$$\tau_{sem} = \frac{\tau_A + \tau_B}{2} \quad (3)$$

where A and B represent the two complete sets of codes in codebooks.

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**

Table 1: Distribution of relationship types comparing codebooks derived from the HiAICS dataset. A visual side-by-side comparison is provided in Figure 5, and detailed results in Table 4 in the Appendix.

4.2 Relevance of Code Assignments

To assess code-to-text relevance independently of upstream stages, we provided the system with *human-curated codebooks* derived from prior manual analyses[3]. This controlled setup isolates the code application mechanism. Several experts assessed whether human-assigned and AI-assigned codes were *relevant* or *irrelevant* to corresponding text segments, blinded to origin.

4.3 Quality of Theoretical Findings

To assess the quality of the generated findings, we conducted a human evaluation using three criteria:

- (G) **Grounding** (*Data Grounding, Evidence Support & Accuracy*): Findings must be accurate, reliable, and well-supported by the interviews. Optimally, multiple coded segments are mentioned or provided.

Dataset	Human	AICoE
CVDQuoding	0.806	0.760
HiAICS	0.740	0.560
Overall Average	**0.773**	**0.660**

Table 2: Relevant code assignments averaged across interviews and evaluators, from human and AI coders; results for each evaluator are in Table 5 in the Appendix

- (R) **Relevance** (*Alignment with Code & Research Goals*): Findings should address the research objectives and the assigned code.
- (I) **Insight** (*Insightfulness, Novelty & Non-Triviality*): Findings should reveal deeper, non-obvious insights of intellectual value and avoid surface-level observations or trivialities.

For the HiAICS dataset, three experts who were asked to read the interviews before rated each finding on a 5-point Likert scale across these dimensions. The %HQ metric (percentage of high-quality findings) reflects the proportion of codes yielding at least one finding with an average rating of 4.00 or higher across experts and criteria.

	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**

Table 3: Evaluation scores for AICoE findings on HiAICS across 31 codes (151 total findings), detailed results for all findings are in Table 7 and exemplary, high-quality findings are in Figure 4, both in the Appendix

5 Discussion

Alignments, Gaps, and New Perspectives in Codebooks. The codebook alignments (Table 1) indicate that AICoE is not meaningfully more divergent from either human-coded schema than the two human codebooks are from each other. However, a closer manual inspection of the codebooks reveals that AICoE tends to prioritize thematic concepts, whereas human coders occasionally add codes reflecting individual interviewee experiences (e.g., "Biographical Context" or "Personal Work"). Notably, all three codebooks contained unique codes unmatched by the others, underscoring AICoE's potential to complement human analysis by offering alternative perspectives that can aid researchers in refining and expanding their codebooks.

[3] Specifically, for the CVDQuoding dataset, which was published with two codebooks, we utilized Coder 2's codebook. For our HiAICS dataset, we employed a codebook developed by one of our expert annotators (Coder 1).

Coding Performance Disparities. The observed performance gap between human and AI coding in HiAICS ($\Delta = 0.180$) compared to CVDQuoding ($\Delta = 0.046$) presumably stems from inherent data characteristics. First and most importantly, CVDQuoding consists of structured interviews with predefined questions, likely providing clearer thematic boundaries that facilitate more consistent coding. Second, an interview in HiAICS contains, on average, approximately twice the word count (10,663 versus 5,163 words), increasing the complexity for the model to maintain contextual coherence. This aligns with previous evidence showing that LLM performance generally degrades as the context length increases (Liu et al., 2024). Finally, the ASR-generated transcripts in HiAICS introduce linguistic noise through transcription artifacts and speech disfluencies.

Finding Meaning in Data. The results in Table 3 underscore that AICoE reliably identifies theoretically relevant patterns, achieving an overall quality score of 3.49 with 32.25% of codes with high-quality findings (≥ 4.00). Grounding (3.42) and relevance (3.76) outperformed insight (3.29), reflecting strength in anchoring findings in data and aligning them with research objectives while highlighting the difficulty of automating interpretative depth. Inter-rater correlations (see Appendix B.3.1) reveal more consistent assessments for grounding (E2–E3: $r = 0.6471$), but low agreement for relevance and insight (max $r = 0.1194$ and 0.2478), indicating more subjective judgments in evaluating thematic alignment and the novelty of findings.

AI-Augmented Ethnography. While our approach presents a systematic pipeline for qualitative analysis, it should not be viewed solely through the lens of automation. Rather, the framework embraces human expertise and allows for critical intervention at every stage. The *unified codebook*, in particular, serves as a 'checkpoint' where researchers can review, refine, and adjust consolidated codes before proceeding to code application and pattern finding. Importantly, our framework also supports *deductive coding* approaches, allowing researchers to bypass the open coding and code consolidation stages by directly applying a pre-existing or theory-driven codebook. This flexibility extends throughout the pipeline – researchers can iterate through stages multiple times, run parallel samples, or modify intermediate outputs as needed. The pattern finding stage, as a final step, exemplifies this collaboration, where computational analysis assists human insight rather than replaces it.

Tool, Partner, or Epistemic Medium? Based on these considerations, it is imperative to clarify that the AI Co-Ethnographer is conceptualized neither as a mere instrument nor as a quasi-human agent. We must conscientiously avoid both anthropomorphic and anthropocentric framings, and equally guard against its reduction to a static, predetermined technological artifact. Rather, we posit the AI Co-Ethnographer as an epistemic medium, one that facilitates and supports the generation of knowledge, while remaining subject to critical reflection. Serving as such a medium, the AI Co-Ethnographer enriches the research infrastructure that underpins ethnographic and, more comprehensively, qualitative research.

Multimodality and Data Heterogeneity. Future research must address the inherent multimodality and data heterogeneity of scientific processes related to the analysis of qualitative data. While our pipeline focuses on textual data (interview transcripts), scientific activity extends far beyond text. It encompasses diverse multimodal inputs or media: spoken language (interviews, lectures, meetings), visual elements (slides, graphics, videos), and discipline-specific sensor data (Yang et al., 1998; Bett et al., 2000). Scientific discussions, for instance, exemplify this multimodality, integrating spoken interaction, nonverbal cues like gesture and gaze, or the presentation of visual materials. Achieving a broader, faster, and more contextualized understanding of scientific processes requires developing methods to process, interpret, and synthesize these diverse, cross-modal signals.

6 Conclusion

The AI Co-Ethnographer demonstrates both the potential and limitations of AI-supported qualitative research. Our evaluation reveals robust codebook development, reasonable code assignments, and the ability to generate meaningful findings. This represents a promising direction for qualitative research, enabling the processing of large volumes of data while maintaining analytical depth. Beyond functioning as a mere tool, AICoE serves as an epistemic medium in the research process.

Limitations

Debates continue over the extent to which ethnographic approaches to qualitative research can be automated or delegated to AI systems. However, larger amounts of ethnographic data can only be analyzed with the support of corresponding systems. In the context of our research, every phase of qualitative data analysis remains intrinsically tied to ethnographic experience and observation of human subjects. Future refinements to our framework could prioritize the specificities inherent in ethnographic data analysis, placing them at the core of this epistemic conduit. For instance, we might contemplate a more nuanced synthesis of interview transcripts and observational records, such as field notes. However, we consider it an asset, rather than a liability, that this proposed epistemic conduit offers flexible support for the annotation and interpretation of qualitative research data beyond solely ethnographic contexts. Consequently, it has the potential to reshape how AI supports transdisciplinary qualitative research in the future.

Ethics

The use of LLMs for automatic coding and qualitative analysis of research materials involves ethical challenges related to data privacy, algorithmic biases, and transparency. Researchers should ensure that participant data is adequately protected and obtain their informed consent for AI-assisted analysis. It is essential to critically evaluate potential biases in LLM-generated annotations and interpretations and to ensure transparency in AI's role in the analytical process. Clear authorship and accountability guidelines are necessary for LLM-assisted qualitative analysis. Finally, it is important to balance leveraging AI's ability to handle massive datasets with maintaining rigorous ethical research standards.

Acknowledgments

This research is supported by the project "How is AI Changing Science? Research in the Era of Learning Algorithms" (HiAICS), funded by the Volkswagen Foundation. We thank Matthias Ernst, Christine Hämmerling, Birte Luisa Kuhle, Markus Ramsauer, Johanna Maria Toussaint, and Charmaine Voigt for their contributions to annotation and evaluation.

References

H. Russell Bernard, Amber Wutich, and Gery W. Ryan. 2016. *Analyzing Qualitative Data: Systematic Approaches*. SAGE Publications. Google-Books-ID: yAi1DAAAQBAJ.

Harvey Russell Bernard. 2011. *Research Methods in Anthropology: Qualitative and Quantitative Approaches*. Rowman Altamira.

Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Multimodal Meeting Tracker. In *RIAO*, pages 32–45. Paris, France.

John W. Creswell. 2015. *30 Essential Skills for the Qualitative Researcher*. SAGE Publications. Google-Books-ID: fkJsCgAAQBAJ.

John W. Creswell and Cheryl N. Poth. 2017. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, 4 edition. SAGE Publications, Inc.

Norman K. Denzin, Yvonna S. Lincoln, Michael Donald Giardina, and Gaile S. Cannella. 2023. *The SAGE Handbook of Qualitative Research*, 6 edition. SAGE Publications, Inc, Los Angeles London New Delhi Singapore Washington DC Melbourne.

Anne Dippel and Andreas Sudmann. 2023. AI ethnography. In Simon Lindgren, editor, *Handbook of Critical Studies of Artificial Intelligence*, pages 826–844. Edward Elgar Publishing.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, ..., and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

V. Elliott. 2018. Thinking about the coding process in qualitative data analysis. *Qualitative Report*, 23(11). Publisher: Nova Southeastern University.

Tim Fischer and Chris Biemann. 2024. Exploring Large Language Models for Qualitative Data Analysis. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 423–437, Miami, USA. Association for Computational Linguistics.

Lina Franken and Libuše Hannah Vepřek. 2025. AI in and for Qualitative Research. In Heidrun Friese, Marcus Nolden, and Miriam Schreiter, editors, *Handbuch Soziale Praktiken und Digitale Alltagswelten*, pages 1–9. Springer Fachmedien, Wiesbaden.

Jamie Harding. 2013. *Qualitative Data Analysis from Start to Finish*. SAGE. Google-Books-ID: 9YUQAgAAQBAJ.

Judith A. Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory*, 3:265–289.

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. 2002. Automatic speech summarization applied to English broadcast news speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–9. IEEE.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. Place: Cambridge, MA Publisher: MIT Press.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint*. ArXiv:2408.06292 [cs].

Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2019. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications. Google-Books-ID: Bt0uuQEACAAJ.

David L. Morgan. 2023. Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods*, 22:16094069231211248. Publisher: SAGE Publications Inc.

Thomas Morton and Jeremy LaCivita. 2003. WordFreak: An Open Tool for Linguistic Annotation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Demonstrations*, pages 17–18.

Thai-Binh Nguyen and Alexander Waibel. 2025. MSA-ASR: Efficient Multilingual Speaker Attribution with frozen ASR Models. *arXiv preprint*. ArXiv:2411.18152 [cs].

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. 2021. Super-Human Performance in Online Low-Latency Recognition of Conversational Speech. In *Interspeech 2021*, pages 1762–1766. ISSN: 2958-1796.

Fynn Petersen-Frey, Tim Fischer, Florian Schneider, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023. From Qualitative to Quantitative Research: Semi-Automatic Annotation Scaling in the Digital Humanities. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 52–62, Ingolstadt, Germany. Association for Computational Lingustics.

Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Wang Xiaofeng, Anh Nguyen Duc, and Pekka Abrahamsson. 2024. Can Large Language Models Serve as Data Analysts? A Multi-Agent Assisted Approach for Qualitative Data Analysis. *arXiv preprint*. ArXiv:2402.01386 [cs].

Fabian Retkowski and Alexander Waibel. 2024a. From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419, St. Julian's, Malta. Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2024b. Zero-Shot Strategies for Length-Controllable Summarization. *arXiv preprint*. ArXiv:2501.00233 [cs].

Jane Ritchie, Jane Lewis, Carol McNaughton Nicholls, and Rachel Ormston. 2014. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. SAGE Publications. Google-Books-ID: zkITlwEACAAJ.

Johnny Saldana. 2015. *The Coding Manual for Qualitative Researchers*. SAGE. Google-Books-ID: ZhxiCgAAQBAJ.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, ..., and Vivek Natarajan. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8. Publisher: Nature Publishing Group.

Cesare Spinoso-Di Piano, Samira Rahimi, and Jackie Cheung. 2023. Qualitative Code Suggestion: A Human-Centric Approach to Qualitative Coding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14887–14909, Singapore. Association for Computational Linguistics.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference on - AFIPS '63 (Spring)*, page 241, Detroit, Michigan. ACM Press.

Anna Tuschling, Andreas Sudmann, and Bernhard J. Dotzler, editors. 2023. *ChatGPT und andere*

»Quatschmaschinen«: Gespräche mit Künstlicher Intelligenz. transcript Verlag. Accepted: 2024-02-02T16:04:26Z.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, ..., and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint*. ArXiv:2309.07864 [cs].

Jie Yang, Rainer Stiefelhagen, Uwe Meier, and Alex Waibel. 1998. Visual tracking for multimodal human computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 140–147.

Seid Muhie Yimam, Chris Biemann, Richard Eckart De Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96.

Klaus Zechner and Alex Waibel. 2000. DIASUMM: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. Place: Cambridge, MA Publisher: MIT Press.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A Pipeline Prompts

Open Coding

You are an AI assistant tasked with suggesting relevant codes for an ethnographic interview transcript. In ethnography, coding is the process of assigning labels or categories to segments of qualitative data to identify themes and patterns. This is a crucial step in analyzing interview data.

You will be presented with a transcript from an ethnographic interview. Your task is to suggest a set of codes that are relevant to this transcript. Remember, you are not assigning codes to specific sentences but rather proposing a list of codes that could be used to analyze this transcript.

Here is the transcript:

<transcript>
`<transcript>`
</transcript>

Please analyze this transcript and suggest a set of codes that could be used to categorize and understand the themes present in the interview. Follow these guidelines:

1. Codes should be concise, typically consisting of one to three words
2. Codes should capture key concepts, themes, or ideas present in the transcript
3. Aim for a mix of descriptive codes (what is happening) and interpretive codes (the underlying meaning)
4. Consider both explicit content and implicit meanings in the transcript
5. Avoid overly broad or vague codes
6. You are free to suggest up to `<max_codes>` codes, depending on the complexity and length of the transcript
7. Provide a brief description (up to 20 words) for each code to clarify its meaning and application that differentiates it from other codes.

You will be provided context that you can and should consider when suggesting codes.

<context>
`<context>`
</context>

Optional

Before providing your final list of codes, use the <scratchpad> to think through your process:

<scratchpad>

1. Identify the main topics discussed in the interview
2. Note any recurring themes or ideas
3. Consider the context and any underlying meanings
4. Think about the interviewee's experiences, attitudes, and behaviors
5. Reflect on how these elements could be categorized into codes

</scratchpad>

Now, please provide your suggested list of codes for this transcript. Present your codes in the following format:

<suggested_codes>

- Code 1 | Description that explains the meaning and context of Code 1 in up to 20 words
- Code 2 | Description that explains the meaning and context of Code 2 in up to 20 words

...

</suggested_codes>

Remember, these codes should be relevant to the given transcript and useful for further analysis in an ethnographic study. Do not write content outside <scratchpad> or <suggested_codes>.

Parameters

- `<transcript>`: The raw interview transcript to analyze
- `<context>`: Optional additional context to consider
- `<max_codes>`: Maximum number of codes to suggest

Code Consolidation

You are an AI assistant tasked with generating a comprehensive set of codes based on multiple ethnographic interviews. Your goal is to create a coherent and inclusive set of codes that covers the themes from all the interviews provided.

You will be provided context that you can and should consider when creating your final set of codes.

<context>
<context>
</context>

Optional

You will be presented with sets of codes generated from multiple interviews. These codes are contained in the following variable:

<interview_codes>
<interview_codes>
</interview_codes>

Analyze these sets of codes and create a single, comprehensive set that encompasses the themes from all interviews. Follow these guidelines:

1. Review all the code sets carefully, identifying common themes and unique concepts.
2. Combine similar codes across different interviews, choosing the most descriptive and clear wording.
3. Generalize codes when appropriate to capture broader themes that appear across multiple interviews.
4. Retain unique codes that represent important themes specific to individual interviews.
5. Ensure that the final set of codes is balanced, covering all major themes present in the original code sets.
6. Aim for clarity and conciseness in your final codes, typically using one to three words per code.
7. Provide a brief description (up to 20 words) for each code to clarify its meaning and application that differentiates it from other codes.

Before providing your final list of codes, use the <scratchpad> to think through your process:

<scratchpad>

1. Identify recurring codes and themes across all interviews
2. Note any unique codes that represent important individual perspectives
3. Consider how to merge similar codes without losing nuance
4. Reflect on potential broader categories that could encompass multiple codes
5. Ensure all major themes from the original code sets are represented

</scratchpad>

Now, please provide your comprehensive set of codes based on all the interviews. Present your codes in the following format:

<comprehensive_codes>

- Code 1 | Description that explains the meaning and context of Code 1 in up to 20 words
- Code 2 | Description that explains the meaning and context of Code 2 in up to 20 words
- Code 3 | Description that explains the meaning and context of Code 3 in up to 20 words

  ...

</comprehensive_codes>

Remember, your final set should have no more than <max_codes> codes. Ensure that these codes are relevant, clear, and useful for further analysis in an ethnographic study. Do not write content outside <scratchpad> or <comprehensive_codes>.

Parameters

- <interview_codes>: The sets of codes from multiple interviews
- <context>: Optional additional context to consider
- <max_codes>: Maximum number of codes to present in the final set

Code Application

You are an AI assistant tasked with analyzing an ethnographic interview and extracting relevant parts that correspond to a specific code from a given taxonomy. Follow these instructions carefully:

1. First, you will be presented with the full text of an interview:

<interview>
`<interview>`
</interview>

2. Next, you will be given a taxonomy of codes with its differentiating descriptions:

<taxonomy>
`<set_of_codes>`
</taxonomy>

3. You will be focusing on one specific code from this taxonomy:

<code>
`<specific_code>`
</code>

4. Your task is to carefully read through the interview text and identify parts that are most important or salient in relation to the specified code. These parts should justify assigning the code to those sections of the interview.

5. When you find relevant parts, list them in the following format:

- `- <part>exact text from the interview</part>`
- `- <part>another exact text from the interview</part>`
- (Continue this format for all relevant parts you find)

Important notes:

- Do not change the content of the extracted parts in any way.
- Include only the most relevant and important parts. Quality is more important than quantity.
- Ensure that each extracted part corresponds to exactly one line from the original interview. Do not merge multiple lines or extract partial lines.
- Ensure that the extracted parts, when taken together, provide a clear justification for assigning the specified code.

6. If you cannot find any parts of the interview that are relevant to the specified code, respond with:

`None`

Remember, your goal is to provide an accurate and focused analysis that helps understand how the specified code applies to this interview. Be thorough in your examination but selective in your choices of relevant parts. Present your findings without any additional commentary. Start your response with either the list of parts or "None" if no relevant parts are found.

Parameters

- `<interview>`: The full text of the ethnographic interview
- `<set_of_codes>`: The taxonomy of codes with differentiating descriptions
- `<specific_code>`: The single code from the taxonomy that you must focus on

Pattern Finding

You are an AI assistant tasked with the final stage of an automated ethnography pipeline: Pattern Finding. Your goal is to analyze coded segments from multiple interviews and generate theoretical findings based on this primary, coded data.

You will be presented with coded segments for a specific code found across all interviews. These segments are contained in the following variable:

<coded_segments>
`<coded_segments>`
</coded_segments>

The specific code these segments relate to is:

<code>
`<code>`
</code>

Your task is to carefully analyze these coded segments and identify meaningful patterns, themes, or theoretical findings. Follow these guidelines:

1. Read through all the coded segments thoroughly, paying attention to recurring ideas, contradictions, and unique perspectives.
2. Look for connections between different segments that might reveal deeper insights or patterns.
3. Aim to generate 3–5 significant findings or patterns. Focus on quality over quantity.
4. Prioritize non-trivial findings that go beyond surface-level observations.
5. Each finding should be supported by evidence from multiple coded segments when possible.

Before presenting your final findings, use the <scratchpad> to think through your analysis:

<scratchpad>

1. Identify recurring themes or ideas across the coded segments
2. Note any contradictions or divergent perspectives
3. Consider how these segments relate to the specific code and the broader context of the study
4. Reflect on potential deeper meanings or implications of the data
5. Formulate initial ideas for findings or patterns

</scratchpad>

Now, present your findings in the following format:

<findings>

1. **Brief title of finding**
   [Detailed explanation of the finding, including supporting evidence from the coded segments]
2. **Brief title of finding**
   [Detailed explanation of the finding, including supporting evidence from the coded segments]
3. [Continue this format for all findings]

</findings>

Remember to focus on generating insightful, non-trivial findings that contribute to a deeper understanding of the research topic. Ensure that your findings are well-supported by the data and relevant to the specific code and overall research context.

Parameters

- `<coded_segments>`: The coded segments from multiple interviews that relate to the specific code
- `<code>`: The code under analysis for which the segments have been collected

B Detailed Evaluation Results

B.1 Semantic Relatedness of Codebooks

	# Rel.	Distribution of Relationships					
	N	M	C	P	U	τ_{sem}	Mean τ_{sem}
Coder A – Coder B							
Evaluator 1	28	0.154	0.352	0.185	0.308	0.493	0.584
Evaluator 2	49	0.154	0.386	0.445	0.015	0.647	
Evaluator 3	47	0.339	0.301	0.122	0.238	0.611	
Coder A – AI							
Evaluator 1	35	0.191	0.396	0.191	0.223	0.563	0.638
Evaluator 2	45	0.064	0.522	0.382	0.032	0.620	
Evaluator 3	101	0.364	0.523	0.000	0.113	0.730	
Coder B – AI							
Evaluator 1	36	0.152	0.517	0.136	0.195	0.582	0.545
Evaluator 2	72	0.030	0.688	0.222	0.060	0.623	
Evaluator 3	36	0.061	0.515	0.016	0.409	0.429	

Table 4: Relationship distributions between codebooks from human coders and AICoE, as evaluated by annotators

B.2 Relevance Scores of Code Assignments

	Int. ID	Human	AI
Evaluator 1	1	0.926	0.960
	2	0.984	0.967
	3	0.994	0.992
Evaluator 2	1	0.759	0.854
	2	0.875	0.797
	3	0.872	0.671
Evaluator 3	1	0.519	0.510
	2	0.661	0.625
	3	0.667	0.461
Overall Average		**0.806**	**0.760**

(a) Scores for the CVDQuoading dataset

	Int. ID	Human	AI
Evaluator 1	1	0.685	0.551
	2	0.881	0.643
	3	0.966	0.935
Evaluator 2	1	0.849	0.721
	2	0.944	0.599
	3	0.896	0.673
Evaluator 3	1	0.542	0.389
	2	0.457	0.224
	3	0.444	0.263
Overall Average		**0.740**	**0.560**

(b) Scores for the HiAICS dataset

Table 5: Relevant code assignments from human and AI coders for each interview and evaluator

B.3 Evaluation of Theoretical Findings

B.3.1 Correlation Coefficients

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

Table 6: Correlation Coefficients between Evaluators for Each Criterion

B.3.2 Quality of Theoretical Findings

Code	Grounding			Avg	Relevance			Avg	Insight			Avg	Avg
	E1	E2	E3		E1	E2	E3		E1	E2	E3		
AI Critique	5	4	5	4.67	4	4	3	3.67	3	3	3	3.00	3.78
	4	5	4	4.33	4	3	4	3.67	4	4	2	3.33	3.78
	4	4	5	4.33	4	5	4	4.33	3	5	4	4.00	4.22
	4	4	5	4.33	4	4	5	4.33	3	4	5	4.00	4.22
	3	2	1	2.00	4	2	4	3.33	4	3	4	3.67	3.00
AI for Science	4	4	3	3.67	4	5	4	4.33	4	3	4	3.67	3.89
	4	4	2	3.33	4	5	3	4.00	4	3	1	2.67	3.33
	4	5	5	4.67	3	5	4	4.00	4	3	3	3.33	4.00
	4	3	5	4.00	4	5	4	4.33	4	3	3	3.33	3.89
	3	4	4	3.67	4	5	4	4.33	4	4	3	3.67	3.89
Algorithm	3	4	5	4.00	3	4	3	3.33	3	3	2	2.67	3.33
	4	4	3	3.67	4	4	4	4.00	4	3	3	3.33	3.67
	4	3	3	3.33	4	3	4	3.67	4	3	4	3.67	3.56
	4	4	3	3.67	4	4	3	3.67	3	4	4	3.67	3.67
	4	3	3	3.33	4	4	4	4.00	4	3	4	3.67	3.67
Algorithmic Biases	4	5	5	4.67	4	5	4	4.33	4	5	3	4.00	4.33
	4	3	5	4.00	4	5	4	4.33	3	4	4	3.67	4.00
	4	3	3	3.33	5	5	4	4.67	4	4	3	3.67	3.89
	4	4	4	4.00	4	3	3	3.33	4	3	3	3.33	3.55
	4	3	3	3.33	4	3	3	3.33	4	4	4	4.00	3.55
Autonomy & Agency	3	3	4	3.33	4	2	3	3.00	4	2	2	2.67	3.00
	4	3	3	3.33	4	2	3	3.00	4	2	3	3.00	3.11
	4	2	3	3.00	4	2	4	3.33	4	2	4	3.33	3.22
	3	2	3	2.67	3	3	3	3.00	3	2	2	2.33	2.67
	4	4	3	3.67	4	4	4	4.00	4	3	2	3.00	3.56
Biographical Context	4	5	4	4.33	3	4	4	3.67	3	4	4	3.67	3.89
	4	4	4	4.00	4	3	5	4.00	4	4	3	3.67	3.89
	3	3	3	3.00	3	4	4	3.67	3	3	3	3.00	3.22
	3	4	3	3.33	4	3	4	3.67	3	3	3	3.00	3.33
	4	3	3	3.33	3	3	4	3.33	3	3	4	3.33	3.33
Black Box	4	3	4	3.67	4	4	4	4.00	3	3	2	2.67	3.45
	4	2	3	3.00	4	3	4	3.67	4	4	3	3.67	3.45
	4	2	3	3.00	4	3	4	3.67	3	2	2	2.33	3.00
	2	2	3	2.33	4	4	4	4.00	3	3	4	3.33	3.22
Data	4	4	4	4.00	3	3	4	3.33	2	3	3	2.67	3.33
	4	3	3	3.33	4	4	5	4.33	4	4	3	3.67	3.78
	4	3	3	3.33	3	4	4	3.67	2	4	5	3.67	3.56
	3	4	4	3.67	3	4	4	3.67	3	3	3	3.00	3.45
Epistemic and Infrastructural Media	3	5	5	4.33	3	4	4	3.67	3	5	3	3.67	3.89
	3	4	3	3.33	3	4	5	4.00	2	3	4	3.00	3.44
	4	3	3	3.33	4	4	5	4.33	3	4	4	3.67	3.78
	3	3	4	3.33	3	4	4	3.67	3	3	3	3.00	3.33
	3	2	2	2.33	3	4	5	4.00	3	3	5	3.67	3.33
Expert Systems	4	4	4	4.00	4	4	5	4.33	4	4	3	3.67	4.00
	3	3	4	3.33	4	5	4	4.33	3	3	3	3.00	3.55
	3	3	4	3.33	3	5	3	3.67	3	3	3	3.00	3.33
	4	3	3	3.33	4	4	5	4.33	4	3	4	3.67	3.78
	4	3	3	3.33	3	4	4	3.67	3	3	4	3.33	3.44
Expertise Competence	3	4	4	3.67	4	5	3	4.00	4	3	2	3.00	3.56
	4	3	4	3.67	4	4	4	4.00	4	2	3	3.00	3.56
	4	3	4	3.67	4	4	4	4.00	4	3	4	3.67	3.78
	4	1	3	2.67	4	3	3	3.33	3	2	3	2.67	2.89
	4	3	4	3.67	4	4	4	4.00	4	3	3	3.33	3.67
Facial Recognition	5	3	4	4.00	4	2	4	3.33	4	4	4	4.00	3.78
	4	3	2	3.00	4	4	4	4.00	4	3	3	3.33	3.44
	4	2	2	2.67	4	3	4	3.67	3	3	3	3.00	3.11
	3	2	2	2.33	3	4	3	3.33	3	2	2	2.33	2.66
	4	1	2	2.33	3	3	4	3.33	3	3	4	3.33	3.00
First Encounters with AI	4	5	4	4.33	4	4	4	4.00	4	4	4	4.00	4.11
	3	3	3	3.00	3	4	4	3.67	3	3	3	3.00	3.22
	2	4	4	3.33	2	4	4	3.33	2	2	3	2.33	3.00
	3	3	3	3.00	3	4	4	3.67	3	3	4	3.33	3.33
	4	3	3	3.33	4	4	4	4.00	4	4	3	3.67	3.67
Format	3	2	4	3.00	3	3	3	3.00	3	3	2	2.67	2.89
	3	2	3	2.67	4	2	3	3.00	3	2	4	3.00	2.89
	4	3	3	3.33	4	3	4	3.67	4	4	4	4.00	3.67
	4	2	3	3.00	4	3	4	3.67	3	3	4	3.33	3.33
	3	2	2	2.33	4	3	3	3.33	4	3	4	3.67	3.11
Generative AI	2	5	4	3.67	3	4	4	3.67	3	4	3	3.33	3.56
	4	3	3	3.33	3	4	3	3.33	3	3	4	3.33	3.33
	4	4	4	4.00	3	3	4	3.33	3	2	3	2.67	3.33
	3	2	1	2.00	3	4	3	3.33	3	3	2	2.67	2.67
	4	3	3	3.33	4	5	4	4.33	4	3	4	3.67	3.78
Historical Perspectives on AI, ML, ANN	3	3	4	3.33	3	4	4	3.67	3	4	3	3.33	3.44
	3	3	4	3.33	3	2	4	3.00	3	2	3	2.67	3.00
	3	3	3	3.00	3	3	4	3.33	3	2	3	2.67	3.00
	3	3	3	3.00	3	3	4	3.33	3	3	2	2.67	3.00
	3	3	3	3.00	3	4	4	3.67	3	3	3	3.00	3.22
Images	5	4	5	4.67	4	4	3	3.67	3	3	3	3.00	3.78
	4	5	4	4.33	4	3	4	3.67	4	4	2	3.33	3.78
	4	4	5	4.33	4	5	4	4.33	3	5	4	4.00	4.22
	4	4	5	4.33	4	4	5	4.33	3	4	5	4.00	4.22
	3	2	1	2.00	4	2	4	3.33	4	3	4	3.67	3.00
Institutions	4	4	3	3.67	4	5	4	4.33	4	3	4	3.67	3.89
	4	4	2	3.33	4	5	3	4.00	4	3	1	2.67	3.33
	4	5	5	4.67	3	5	4	4.00	4	3	3	3.33	4.00
	4	3	5	4.00	4	5	4	4.33	4	3	3	3.33	3.89
	3	4	4	3.67	4	5	4	4.33	4	4	3	3.67	3.89
Machine Learning, ANN & DL	3	4	5	4.00	3	4	3	3.33	3	3	2	2.67	3.33
	4	4	3	3.67	4	4	4	4.00	4	3	3	3.33	3.67
	4	3	3	3.33	4	3	4	3.67	4	3	4	3.67	3.56
	4	4	3	3.67	4	4	3	3.67	3	4	4	3.67	3.67
	4	3	3	3.33	4	4	4	4.00	4	3	4	3.67	3.67
Media Studies and Visual Culture Studies	4	5	5	4.67	4	5	4	4.33	4	5	3	4.00	4.33
	4	3	5	4.00	4	5	4	4.33	3	4	4	3.67	4.00
	4	3	3	3.33	5	5	4	4.67	4	4	3	3.67	3.89
	4	4	4	4.00	4	3	3	3.33	4	3	3	3.33	3.55
	4	3	3	3.33	4	3	3	3.33	4	4	4	4.00	3.55
Pattern Recognition	3	3	4	3.33	4	2	3	3.00	4	2	2	2.67	3.00
	4	3	3	3.33	4	2	3	3.00	4	2	3	3.00	3.11
	4	2	3	3.00	4	2	4	3.33	4	2	4	3.33	3.22
	3	2	3	2.67	3	3	3	3.00	3	2	2	2.33	2.67
	4	4	3	3.67	4	4	4	4.00	4	3	2	3.00	3.56
Political & Economic Contexts of (Applied) AI	4	5	4	4.33	3	4	4	3.67	3	4	4	3.67	3.89
	4	4	4	4.00	4	3	5	4.00	4	4	3	3.67	3.89
	3	3	3	3.00	3	4	4	3.67	3	3	3	3.00	3.22
	3	4	3	3.33	4	3	4	3.67	3	3	3	3.00	3.33
	4	3	3	3.33	3	3	4	3.33	3	3	4	3.33	3.33
Project Description	4	3	4	3.67	4	4	4	4.00	3	3	2	2.67	3.45
	4	2	3	3.00	4	3	4	3.67	4	4	3	3.67	3.45
	4	2	3	3.00	4	3	4	3.67	3	2	2	2.33	3.00
	2	2	3	2.33	4	4	4	4.00	3	3	4	3.33	3.22
Publications	4	4	4	4.00	3	3	4	3.33	2	3	3	2.67	3.33
	4	3	3	3.33	4	4	5	4.33	4	4	3	3.67	3.78
	4	3	3	3.33	3	4	4	3.67	2	4	5	3.67	3.56
	3	4	4	3.67	3	4	4	3.67	3	3	3	3.00	3.45
Research Interest Challenges Limitations	3	5	5	4.33	3	4	4	3.67	3	5	3	3.67	3.89
	3	4	3	3.33	3	4	5	4.00	2	3	4	3.00	3.44
	4	3	3	3.33	4	4	5	4.33	3	4	4	3.67	3.78
	3	3	4	3.33	3	4	4	3.67	3	3	3	3.00	3.33
	3	2	2	2.33	3	4	5	4.00	3	3	5	3.67	3.33
Sensors & Infrastructures & Platforms	4	4	4	4.00	4	4	5	4.33	4	4	3	3.67	4.00
	3	3	4	3.33	4	5	4	4.33	3	3	3	3.00	3.55
	3	3	4	3.33	3	5	3	3.67	3	3	3	3.00	3.33
	4	3	3	3.33	4	4	5	4.33	4	3	4	3.67	3.78
	4	3	3	3.33	3	4	4	3.67	3	3	4	3.33	3.44
Speculations Ideologies Imaginations of AI	3	4	4	3.67	4	5	3	4.00	4	3	2	3.00	3.56
	4	3	4	3.67	4	4	4	4.00	4	2	3	3.00	3.56
	4	3	4	3.67	4	4	4	4.00	4	3	4	3.67	3.78
	4	1	3	2.67	4	3	3	3.33	3	2	3	2.67	2.89
	4	3	4	3.67	4	4	4	4.00	4	3	3	3.33	3.67
Terms & Definitions	5	3	4	4.00	4	2	4	3.33	4	4	4	4.00	3.78
	4	3	2	3.00	4	4	4	4.00	4	3	3	3.33	3.44
	4	2	2	2.67	4	3	4	3.67	3	3	3	3.00	3.11
	3	2	2	2.33	3	4	3	3.33	3	2	2	2.33	2.66
	4	1	2	2.33	3	3	4	3.33	3	3	4	3.33	3.00
Tools & Methods	4	5	4	4.33	4	4	4	4.00	4	4	4	4.00	4.11
	3	3	3	3.00	3	4	4	3.67	3	3	3	3.00	3.22
	2	4	4	3.33	2	4	4	3.33	2	2	3	2.33	3.00
	3	3	3	3.00	3	4	4	3.67	3	3	4	3.33	3.33
	4	3	3	3.33	4	4	4	4.00	4	4	3	3.67	3.67
Trust	3	2	4	3.00	3	3	3	3.00	3	3	2	2.67	2.89
	3	2	3	2.67	4	2	3	3.00	3	2	4	3.00	2.89
	4	3	3	3.33	4	3	4	3.67	4	4	4	4.00	3.67
	4	2	3	3.00	4	3	4	3.67	3	3	4	3.33	3.33
	3	2	2	2.33	4	3	3	3.33	4	3	4	3.67	3.11
Uses of AI for...	2	5	4	3.67	3	4	4	3.67	3	4	3	3.33	3.56
	4	3	3	3.33	3	4	3	3.33	3	3	4	3.33	3.33
	4	4	4	4.00	3	3	4	3.33	3	2	3	2.67	3.33
	3	2	1	2.00	3	4	3	3.33	3	3	2	2.67	2.67
	4	3	3	3.33	4	5	4	4.33	4	3	4	3.67	3.78
Average	3.64	3.22	3.41	3.42	3.66	3.75	3.86	3.76	3.41	3.19	3.27	3.29	3.49

Table 7: Evaluation results of all findings for all three evaluators and criteria

C Examplary Outputs

C.1 Codebooks

Codebook Comparison

Coder 1

- 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
- Format
- Generative AI
- Historical Perspectives on AI, ML, ANN
- Images
- Institutions
- Machine Learning, ANN, DL
- Media Studies - Bildwissenschaft - Visual Culture Studies
- Pattern Recognition
- Political & Economic Contexts of (Applied) AI
- Project Description
- Publications
- Research Interest, Challenges, Limitations
- Sensors, Infrastructures & Platforms
- Speculations, Ideologies, Imaginations of AI
- Terms & Definitions
- Tools & Methods
- Trust
- Uses of AI for...

Coder 2

- AI Critique
- AI History
- Automation of Work
- Commercialization
- Continuities In Research
- Data Availability
- Data Practices
- Definition of Discipline
- Depiction of AI
- Expertise
- Future Areas of Research
- History of Climate Science
- History of Discipline
- History of Facial Recognition
- History of Photography
- History of Physics
- Interview Technicalities
- Large Language Models
- Limitations of AI
- New Questions Through AI
- Pattern Recognition
- Personal Approach To AI
- Philosophical Implications of AI
- Politics of Infrastructure
- Possible AI Applications
- Practices In Climate Science
- Prediction
- Programming Practices
- Recent Developments In Research
- Recent Personal Work
- Recent Publications
- Research Practice
- Rule-Based AI
- Ruptures Through AI

AICoE

- AI Applications
- AI Critique
- Automation
- Bildwissenschaft
- Black Box Problem
- Climate Science
- Critical Theory
- Data Quality
- Digital Literacy
- Epistemological Questions
- Epistemology
- Ethics
- Extractivism
- Facial Recognition
- Future Directions
- Fuzziness
- Generative AI
- Human-AI Interaction
- Image Manipulation
- Infrastructures
- Interdisciplinary
- Machine Learning
- Media Influence
- Model Limitations
- Neocolonialism
- Neural Networks
- Pattern Recognition
- Prediction Challenges
- Style Transfer
- Surveillance Capitalism
- Uncertainty
- Visual Culture

Figure 2: Side-by-side comparison of the codebooks developed by two human coders and the AICoE system for analyzing the HiAICS data. The comparison highlights overlapping themes, distinct coding approaches, and varying emphases in categories such as technical concepts, historical perspectives, ethical considerations, and individual interviewee experiences.

C.2 Codebook Relations

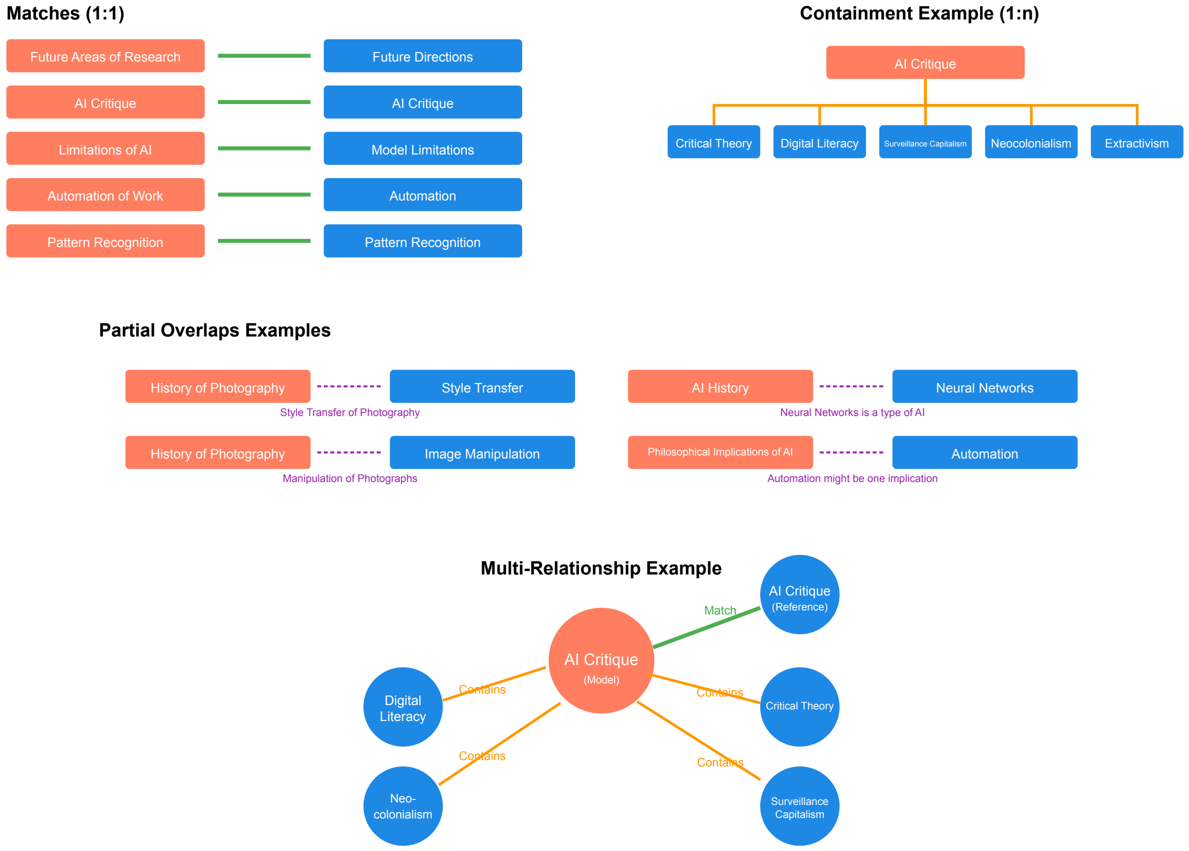


Figure 3: Examplary visualization of select relationships between codes between a human-developed codebook and the codebook of AICoE, as annotated by one of our expert annotators

C.3 Findings

Finding 1 Quality Score: 4.33

Pervasiveness of Algorithmic Biases

The coded segments illustrate that algorithmic biases are not limited to a specific domain but are a widespread issue affecting various applications of AI and machine learning. For instance, Speaker 0 in Interview_[redacted]_20240905 discusses how biases can lead to incorrect predictions in climate modeling when the system encounters new, unseen data. Similarly, Speaker 0 in Interview_[redacted]_20241016 highlights the persistent problem of bias in facial recognition technology. This pervasiveness underscores the need for a comprehensive approach to addressing biases, one that considers the unique challenges and implications of each domain.

Finding 2 Quality Score: 4.33

Interdisciplinary Approach to Visual Culture

The coded segments suggest that combining Bildwissenschaft (focusing on the singular, autonomous image) with media studies (considering image economies and infrastructures) provides a more comprehensive understanding of AI's impact on visual culture. This is evident in Speaker 0's remark from Interview_[redacted]_20241016, where they mention the need to combine approaches from Bildwissenschaft with media studies to deal with both the historical, singular image and the broader image ecologies. This integration is crucial for navigating the changing landscape of visual content production and analysis, especially with the advent of AI-generated images.

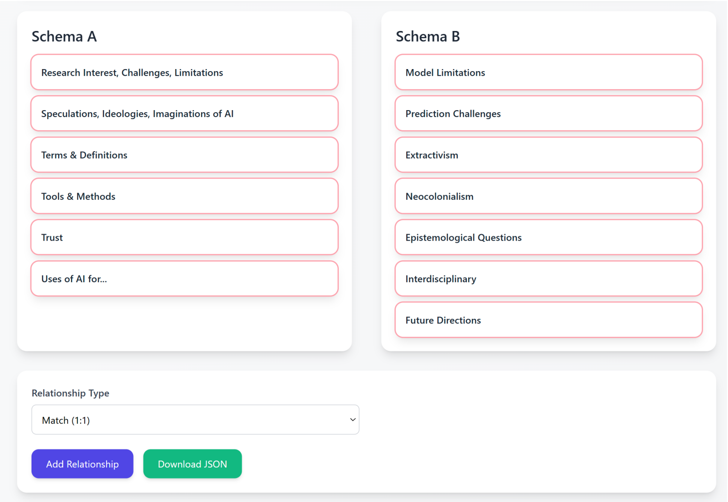
Finding 3 Quality Score: 4.00

Evolution of Expert Systems

The concept of expert systems has undergone significant evolution, from being heavily reliant on rule-based systems and knowledge engineering to embracing more data-driven approaches. This shift is evident in Speaker 1's discussion from Interview_[redacted]_20141016, where they mention, "Today, if you want to build a similar concept, an expert system, instead of interviewing the experts, medical doctors asking them about, tell me about these symptoms and this illness and this, et cetera, you would take data, raw data." This evolution suggests a move towards leveraging machine learning and potentially generative AI models, as hinted at with the mention of "generative pre-trained transformer" in the same interview.

Figure 4: High-quality findings generated by AI Co-Ethnographer from the HiAICS dataset, as rated by three evaluators. The Quality Score (1.00–5.00) represents the average across all evaluators and criteria.

D Human Evaluation Interfaces



Schema A

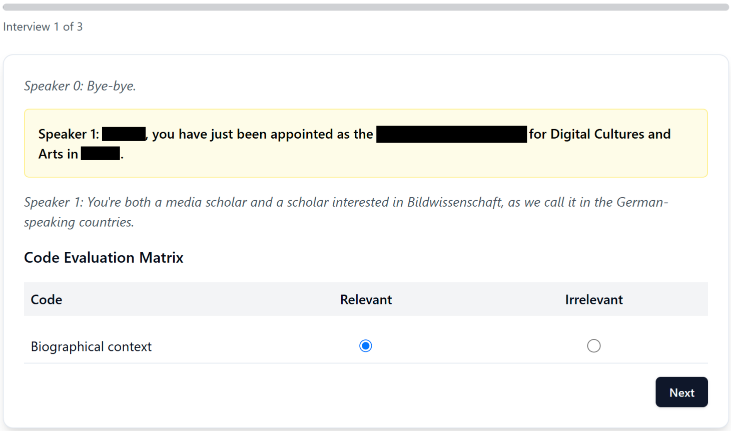
- Research Interest, Challenges, Limitations
- Speculations, Ideologies, Imaginations of AI
- Terms & Definitions
- Tools & Methods
- Trust
- Uses of AI for...

Schema B

- Model Limitations
- Prediction Challenges
- Extractivism
- Neocolonialism
- Epistemological Questions
- Interdisciplinary
- Future Directions

Relationship Type: Match (1:1)

Figure 5: Evaluation interface that allows human annotators to specify the relationships between different codebooks



Interview 1 of 3

Speaker 0: Bye-bye.

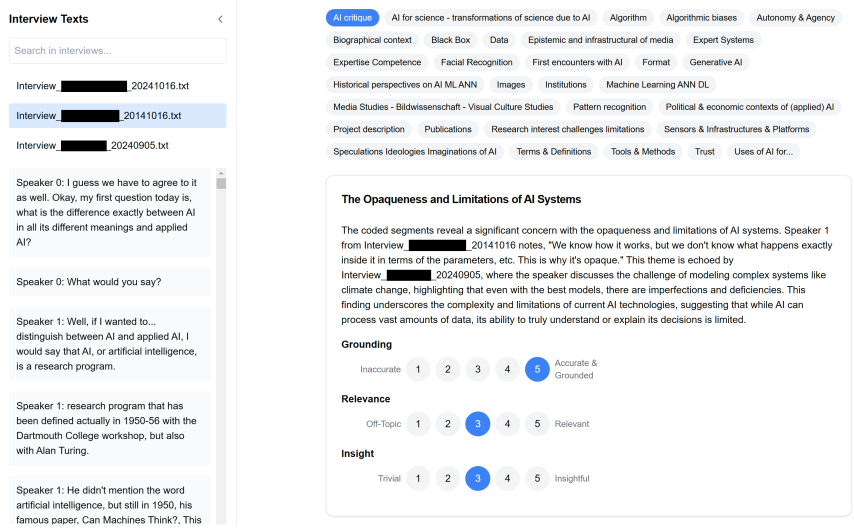
Speaker 1: [redacted], you have just been appointed as the [redacted] for Digital Cultures and Arts in [redacted].

Speaker 1: You're both a media scholar and a scholar interested in Bildwissenschaft, as we call it in the German-speaking countries.

Code Evaluation Matrix

Code	Relevant	Irrelevant
Biographical context	●	○

Figure 6: Evaluation interface used by human annotators to assess the relevance of code assignments



Interview Texts

Search in interviews...
Interview_[redacted]_20241016.txt
Interview_[redacted]_20141016.txt
Interview_[redacted]_20240905.txt

Speaker 0: I guess we have to agree to it as well. Okay, my first question today is, what is the difference exactly between AI in all its different meanings and applied AI?

Speaker 0: What would you say?

Speaker 1: Well, if I wanted to... distinguish between AI and applied AI, I would say that AI, or artificial intelligence, is a research program.

Speaker 1: research program that has been defined actually in 1950-56 with the Dartmouth College workshop, but also with Alan Turing.

Speaker 1: He didn't mention the word artificial intelligence, but still in 1950, his famous paper, Can Machines Think?, This

AI critique
AI for science - transformations of science due to AI
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
Format
Generative AI
Historical perspectives on AI ML ANN
Images
Institutions
Machine Learning ANN DL
Media Studies - Bildwissenschaft - Visual Culture Studies
Pattern recognition
Political & economic contexts of (applied) AI
Project description
Publications
Research interest challenges limitations
Sensors & Infrastructures & Platforms
Speculations Ideologies Imaginations of AI
Terms & Definitions
Tools & Methods
Trust
Uses of AI for...

The Opaqueness and Limitations of AI Systems

The coded segments reveal a significant concern with the opaqueness and limitations of AI systems. Speaker 1 from Interview_[redacted]_20141016 notes, "We know how it works, but we don't know what happens exactly inside it in terms of the parameters, etc. This is why it's opaque." This theme is echoed by Interview_[redacted]_20240905, where the speaker discusses the challenge of modeling complex systems like climate change, highlighting that even with the best models, there are imperfections and deficiencies. This finding underscores the complexity and limitations of current AI technologies, suggesting that while AI can process vast amounts of data, its ability to truly understand or explain its decisions is limited.

Grounding

Inaccurate 1 2 3 4 [✓ 5] Accurate & Grounded

Relevance

Off-Topic 1 2 [✓ 3] 4 5 Relevant

Insight

Trivial 1 2 [✓ 3] 4 5 Insightful

Figure 7: Evaluation interface used by human annotators to assess theoretical findings generated by AICoE