

Qualitative Coding Analysis through Open-Source Large Language Models: A User Study and Design Recommendations

Tung T. Ngo*
thantung.ngo@tudublin.ie
Technological University Dublin
Dublin, Ireland

Dai Nguyen Van
dainv@neu.edu.vn
National Economics University
Hanoi, Vietnam

Anh-Minh Nguyen
minh.a.nguyen@ucdconnect.ie
University College Dublin
Dublin, Ireland

Phuong-Anh Do
anhdp@ftu.edu.vn
Foreign Trade University
Hanoi, Vietnam

Anh Nguyen-Quoc
anh.nguyenquoc@ucd.ie
University College Dublin
Dublin, Ireland

Abstract

Qualitative data analysis is labor-intensive, yet the privacy risks associated with commercial Large Language Models (LLMs) often preclude their use in sensitive research. To address this, we introduce ChatQDA, an on-device framework powered by open-source LLMs designed for privacy-preserving open coding. Our mixed-methods user study reveals that while participants rated the system highly for usability and perceived efficiency, they exhibited "conditional trust", valuing the tool for surface-level extraction while questioning its interpretive nuance and consistency. Furthermore, despite the technical security of local deployment, participants reported epistemic uncertainty regarding data protection, suggesting that invisible security measures are insufficient to foster trust. We conclude with design recommendations for local-first analysis tools that prioritize verifiable privacy and methodological rigor.

CCS Concepts

• **Computing methodologies** → Natural language processing; • **Software and its engineering** → **Open source model**; • **Security and privacy** → **Privacy-preserving protocols**; • **Human-centered computing** → Natural language interfaces; **User centered design**.

Keywords

Large Language Models, Open-Source, Human-AI Collaboration, Qualitative Data Analysis

ACM Reference Format:

Tung T. Ngo, Dai Nguyen Van, Anh-Minh Nguyen, Phuong-Anh Do, and Anh Nguyen-Quoc. 2026. Qualitative Coding Analysis through Open-Source Large Language Models: A User Study and Design Recommendations. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Qualitative research is central to the social sciences because it enables scholars to explain social phenomena and build theory by surfacing meaning, process, and lived experience that are difficult to capture with predefined measures. Yet the craft of qualitative data analysis (QDA), particularly manual open coding, remains labour-intensive and cognitively demanding, constraining how much data researchers can analyse rigorously and transparently [2]. Recent advances in third-party large language models (LLMs) (e.g., ChatGPT, Gemini, and Perplexity), therefore appear promising and growing in popularity: by automating aspects of qualitative analysis, they can reduce the time burden of open coding and expand analytic capacity [25].

This promise, however, is tempered by several ethical and governance challenges. Using commercial LLMs often entails transferring confidential participant data to external servers, raising concerns about data sovereignty, re-identification risk, and compliance with institutional review board (IRB) requirements [7, 28]. These risks are compounded by the opacity of proprietary model architectures, including the possibility that sensitive inputs may be inadvertently memorised and later reproduced, a vulnerability documented in large-scale generative models [3]. In many qualitative research studies, especially those involving vulnerable populations or sensitive settings, these constraints might make it difficult, and sometimes impossible, to justify the use of third-party LLMs.

Open-source LLMs offer a theoretically compelling alternative for sensitive qualitative analysis. Because they can be deployed locally or within institution-controlled environments, they enable researchers to analyse qualitative data without transmitting it beyond secure boundaries, thereby strengthening privacy protection and regulatory compliance [28]. Their transparency and modifiability also allow adaptation to domain language and analytic approaches, including lightweight fine-tuning through Low-Rank Adaptation (LoRA) [13] and grounding outputs in a researcher's corpus via Retrieval-Augmented Generation (RAG), which can reduce hallucinations and improve context retention [15].

Despite their advantages, open-source LLMs can pose substantial usability and maintenance challenges for qualitative researchers, particularly those without technical expertise. Effective use requires competence in machine learning workflows, command-line tools, and model configuration (e.g., quantization and inference parameters), creating barriers for researchers trained in interpretive or

human-centered methodologies rather than computational methods [2, 14]. Furthermore, the open-source ecosystem lacks integrated, well-designed workflows that support core qualitative research practices, forcing researchers to assemble ad hoc solutions that can undermine rigour and analytic quality [25]. This tension, high ethical suitability but limited usability and evidence-based performance, motivates the present study.

The contributions of our research:

- We developed ChatQDA, an on-device platform powered by open-source LLMs, designed to preserve data privacy while supporting qualitative researchers through an open-coding workflow.
- We reported findings from a user study examining how users experience and evaluate collaboration with ChatQDA during open coding, focusing on perceived usefulness, security and privacy, ease of use, trustworthiness, and behavioural intention.
- We distilled empirically grounded design recommendations for future open-source, LLM-based qualitative analysis tools.

2 Related Work

The integration of AI into QDA has evolved from clerical automation to interactive partnership. [20] illustrated this with Cody, a system prioritizing interactive rule refinement and reflection over processing speed. Addressing collaboration, [10] developed CoAICoder to mediate team interpretation, though they identify a trade-off between improved consensus and reduced code diversity. Contextualizing this shift, [9] argues that AI should assist rather than automate, enabling researchers to manage scale and abstraction without sacrificing methodological nuance.

The emergence of Large Language Models (LLMs) has further transformed AI into an interpretive partner capable of thematic synthesis and reflexive assistance. [12] and [16] observe that while AI excels at identifying concrete, descriptive themes, it often misses the subtle nuances detected by human contextual knowledge; notably, Morgan found that LLMs reverse the traditional inductive cycle by presenting broad concepts first for researchers to query rather than moving from raw data to themes. Modern frameworks such as GAITA [17] and GATA [1] formalize this as a guided partnership where the human remains the intellectual leader while delegating structured tasks like template refinement and data familiarization to the AI. Most recently, [22] and [6] have advocated for AI as a reflexive co-researcher or epistemic agent. While [22] integrated GPT-4 to generate micro-analytic memos that surface missed data segments in Grounded Theory analysis, [6]’s AbductivAI model utilizes Chain-of-Thought (CoT) prompting to involve AI agents and human coders in iterative validation loops, fundamentally converting AI from a passive tool into an active intellectual contributor in distributed cognition networks.

Recent open-source advancements have significantly narrowed the performance gap with proprietary systems. Llama 3.1 [21] rivals GPT-4 in general reasoning, while Kimi K2 [24] advances agentic intelligence through autonomous tool orchestration. Innovations in reasoning-focused architectures, such as DeepSeek-R1’s [11] pure reinforcement learning and the hybrid “thinking modes” of Qwen3

[27] and GLM-4.5 [23], enable self-reflection and adaptive computational allocation. These shifts are complemented by models like gpt-oss-120b [18], which provide customizable chain-of-thought capabilities essential for tailored research workflows.

3 Methods

3.1 ChatQDA - Chatbot Powered by Open-Source LLMs

A chat-based interface was selected for its intuitive alignment with participants’ familiarity with tools like ChatGPT and Gemini, as well as its flexibility for integrating advanced LLM capabilities. We developed ChatQDA based on Hugging Face’s Chat UI¹ framework, and the system design adheres to recommendations outlined in [26]. To ensure transparent reasoning, only LLMs capable of generating interpretable explanations were incorporated. Users can upload document files and query the system about their content, with contextual understanding enhanced by a large context window (e.g., extended token limits). The chat history for each session is dynamically embedded in the prompt to contextualize responses, enabling the system to refine answers iteratively based on user feedback. We selected the gpt-oss-20b [18] model for its efficient balance of high-tier reasoning and a compact hardware footprint optimized for on-device use. This 20.9B-parameter MoE model utilizes MXFP4 quantization to achieve a 12.8 GiB checkpoint, allowing deployment on hardware with only 16GB of RAM. While activating just 3.61B parameters per pass for low latency, it remains competitive on benchmarks, scoring 98.7% on AIME 2025 (with tools) and 85.3% on MMLU.

3.2 User Study

To understand how qualitative researchers experience open-source LLM assistance in open coding, we conducted a mixed-method user study. The study examines researchers’ experiences when conducting open coding manually and when collaborating with ChatQDA during the same analytic task.

3.2.1 Participants. We deliberately recruited two user groups as a comparative case study [29]. The first group comprised novice qualitative researchers (fewer than two years of experience with qualitative data analysis and limited prior use of AI for QDA). The second group comprised more experienced qualitative researchers (at least two years of qualitative research experience and some prior exposure to AI-supported QDA). In total, we conducted four sessions (two participants per group). Before participation, we explained the study purpose and procedures and obtained informed consent. To reduce the risk that participants’ feedback would be driven primarily by first impressions of the interface (as indicated in our pilot), we provided a detailed guide describing ChatQDA’s functions, usage scenarios and allowed participants to try the system several times prior to the study task.

3.2.2 User Task and Requirements. Each participant conducted two open-coding activities using the same dataset of a 5-page semi-structured interview transcript.

¹<https://github.com/huggingface/chat-ui>

- Task 1 - Manual open coding (baseline): Participants conducted open coding, following their usual procedures. This stage served both to familiarise participants with the dataset and to establish a baseline for comparison.
- Task 2 - ChatQDA-assisted open coding. Participants then collaborated with ChatQDA to open-code the same materials for 1 additional hour.

3.2.3 Data Collection. Following the two coding activities, participants completed a short post-task survey administered in English via Google Forms. The survey used multi-item, 5-point Likert scales (1 = strongly disagree; 5 = strongly agree) to measure perceived security and privacy [4], perceived ease of use, perceived usefulness [8], and trust [5]. We then conducted a semi-structured interview with each participant to probe the reasoning behind their survey responses and to elicit concrete examples from the session. Interviews were conducted in person, lasted approximately 40 minutes, audio-recorded, transcribed using OpenAI Whisper [19], and manually checked for accuracy.

3.2.4 Data Analysis. Survey responses were analysed using descriptive statistics to summarise patterns across constructs. Interview transcripts and ChatQDA interaction logs were analysed using deductive thematic analysis [3], guided by the survey constructs (perceived usefulness, ease of use, security and privacy, and trust-worthiness). We began with participant-level coding to develop initial codes and construct-level themes, then iteratively refined the coding scheme and theme definitions through discussion among the co-authors. The resulting themes characterise opportunities and challenges in LLM-supported open coding and inform design recommendations for future qualitative analysis tools.

4 Key Findings

From the study, we were able to gather two kinds of data: quantitative data from the surveys and qualitative data from the think-aloud sessions and interviews, which are described in detail below for each major theme.

4.1 Perceived Security

Perceived security ratings clustered around the neutral midpoint (*Median* = 3), reflecting cautious confidence with no strong endorsement. Items regarding preventing third-party modification and interception got slightly higher ($M = 3.0$) than those concerning general security measures and unauthorized access ($M = 2.75$). Overall, participants did not express strong security concerns, but neither did they strongly endorse the system's security, suggesting that perceived security may depend on clearer, more concrete assurance (e.g., transparent explanations of safeguards, evidence of local-only processing, and auditable protections).

The largely moderate (neutral) security ratings appear to reflect epistemic uncertainty rather than a negative judgment. One participant explicitly framed their midpoint choice because of limited technical expertise: *"Although I'm a qualitative researcher, I'm not a technology engineer... I don't really trust that this platform can prevent a third party from modifying the data I process. I'm not sure."*

4.2 Perceived Privacy

Across the four perceived privacy items, responses clustered in the upper-mid range (3–4), indicating moderate to elevated privacy concern when using the ChatQDA. Concern was highest for data collection and privacy during use (75% agree; $M = 3.75$), followed by potential unauthorised secondary use of research data ($M = 3.50$). Concern was lowest for unauthorised sharing with other entities ($M = 3.25$). Overall, participants' privacy concerns centred more on how much data the system collects and how data are handled during processing, rather than on explicit third-party sharing.

Beyond limited technical expertise, a plausible explanation is the gap between the theoretical privacy benefits and the practical realities of open-source LLM use. While participants recognised that local deployment should, in principle, strengthen privacy protections, they remained cautious because they could not independently verify how models are built, distributed, and maintained, or how data might be handled in practice. As one participant noted: *"In theory, open-source LLMs may protect user data better than closed ones, but in practice we can't know how developers use the information, or whether an open-source model contains malicious code to take user data."*

4.3 Perceived Ease of Use

Across the perceived ease-of-use items, responses were generally positive, indicating that ChatQDA was easy to learn and understand. All participants rated 4 (agree) for "Learning to use this technology for qualitative analysis is easy for me" and for "My interaction with this technology for qualitative analysis is clear and understandable". Perceived ease of becoming skillful was more mixed (25% = 2, 50% = 3, 25% = 4; $M = 3.00$), suggesting strong initial learnability but some uncertainty about developing higher proficiency.

The qualitative comments help explain why participants rated ChatQDA as easy to learn and clear to interact with. A novice researcher emphasised prompting as an accessible skill: they found it straightforward to generate and iterate multiple prompts, experimenting with keywords and instructions to progressively improve the output.

4.4 Perceived Usefulness

Across perceived usefulness items, responses were generally positive (mostly 3–5), indicating that ChatQDA was viewed as potentially helpful for qualitative analysis, particularly for efficiency. Ratings for overall usefulness clustered around neutral to agree ($M = 3.00$), and were slightly higher for making qualitative analysis easier ($M = 3.25$). The strongest endorsement was accomplishing tasks more quickly (50% rated 5; $M = 3.75$, $Md = 4$), while perceived performance improvement showed greater variability (evenly distributed across 2–5; $M = 3.50$).

Overall, data from all sources show that participants anticipated clear time-saving benefits, while views on broader performance gains were more varied, suggesting that usefulness may be most immediate for efficiency and may depend on longer-term use or stronger evidence of improvements in analytic quality.

4.5 Trustworthiness

Across trustworthiness items, ratings clustered around the mid-point, indicating moderate trust in ChatQDA. Overall trust was neutral to slightly positive ($M = 3.00$), while perceived reputation was lowest ($M = 2.75$), suggesting ChatQDA was not yet seen as an established research tool. Perceived competence/effectiveness was more favourable ($M = 3.25$, $Md = 3.5$). Reliability ratings were mixed, with half of the participants agreeing but one participant selected 1 ($M = 3.00$, $Md = 3.5$).

Qualitative data help explain this pattern of conditional trust. Participants emphasised that trustworthiness varied by sub-task: excerpt extraction was seen as acceptable, whereas labelling and first-order coding were viewed as less dependable. As one participant explained: *"It depends on the task. For excerpt extraction, the quantitative accuracy is about 50–60%, but in terms of quality, the researcher still needs to review the data, because open-source models often only identify surface-level, direct statements that are close together in the conversation. If the interviewee speaks indirectly, stating a claim in the first sentence but giving an illustrative example at the end, the AI often captures only the first sentence and ignores the example."* Trust was further weakened by perceived hallucination and inconsistencies across analytic steps, illustrated by one participant's observation: *"For label extraction, the AI tends to perform much worse and shows signs of hallucination. For example, it identifies only 8 excerpts, but when assigning first-order codes it becomes 10. Usually it assigns one code per excerpt; the AI cannot identify meaning-based relationships across excerpts to group them under the same code."* Participants also highlighted sensitivity to prompts as a threat to consistency: *"Consistency is only slightly above average. When I change the instructions, the results change immediately, and that can push the outputs further apart. For a less experienced qualitative researcher, this could be confusing when looking at the results produced."* Finally, while the system was perceived as effective for surface-level meaning *"When using this platform, I feel that at the direct, surface level of meaning it processes quite quickly and is relatively effective"* participants remained sceptical about its ability to support deeper interpretive work: *"I feel that for deeper layers of meaning, implicit meanings and meanings that require a broader context, reading across more parts of the data, and higher inference, the platform is not very good yet."* Together, these insights suggest that participants' trust in ChatQDA was bounded: the tool was valued as a fast assistant for explicit content, but viewed as less reliable for nuanced interpretation and consistent coding without careful researcher oversight.

5 Discussions

Overall, participants found ChatQDA easy to use and potentially useful, especially for speeding up open-coding tasks. However, they expressed cautious confidence about security, moderate privacy concern, and conditional trust in the system's outputs, particularly for deeper interpretation and consistent coding. These tensions help explain why behavioural intention was generally positive but polarised (most willing to adopt, one clear outlier unwilling to use/recommend).

5.1 Make Privacy and Security Visible

Even with on-device deployment, participants did not strongly endorse security or privacy. Their ratings reflect uncertainty and limited ability to verify what the system does with data. Improve ChatQDA by: (1) adding clear in-app statements and controls on what is stored, logged, and how to delete it; (2) showing a "local-only/no network calls" indicator; and (3) providing model provenance (version/source/checksum) to reduce concerns about open-source supply-chain risks.

5.2 Build Trust through Traceability and Consistency

Trustworthiness varied by sub-task: excerpt extraction was acceptable, but labelling/first-order coding raised concerns about hallucination, inconsistency across steps, and sensitivity to prompt changes. Participants also noted the system handled surface meaning better than latent meaning requiring broader context. Improve ChatQDA by: (1) forcing evidence-based outputs (highlighted spans/line references for every excerpt/code); (2) adding guardrails to prevent "new" excerpts/codes appearing across stages; (3) enabling reproducible runs (saved prompts/settings) and a lockable codebook; and (4) supporting grouping related excerpts under shared codes.

5.3 Move from "Fast" to "Methodologically Defensible"

Usefulness was strongest for efficiency, while perceived performance gains were more varied, suggesting ChatQDA's value is clearest as a time-saving assistant, but broader adoption depends on supporting analytic rigour. Improve ChatQDA by: integrating memoing and audit trails, linking codes to evidence, and providing lightweight prompt templates for common qualitative tasks to help users become skilful (not just competent).

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

- [1] Isil Anakot, Andrew Katz, Kai Jun Chew, and Holly Matusovich. 2025. Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis. *International Journal of Qualitative Methods* 24 (April 2025), 16094069251338898. doi:10.1177/16094069251338898
- [2] Cauã Ferreira Barros, Bruna Borges Azevedo, Valdemar Vicente Graciano Neto, Mohamad Kassab, Marcos Kalinowski, Hugo Alexandre D. Do Nascimento, and Michelle C.G.S.P. Bandeira. 2025. Large Language Model for Qualitative Research: A Systematic Mapping Study. In *2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE)*. IEEE, Ottawa, ON, Canada, 48–55. doi:10.1109/WSESE66602.2025.00015
- [3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2633–2650. <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>
- [4] Juan Carlos Roca, Juan José García, and Juan José de la Vega. 2009. The importance of perceived trust, security and privacy in online trading systems. *Information Management & Computer Security* 17, 2 (June 2009), 96–113. doi:10.1108/09685220910963983 _eprint: <https://www.emerald.com/ics/article-pdf/17/2/96/1208198/09685220910963983.pdf>
- [5] Kwek Choon Ling, Dazmin Bin Daud, Tan Hoi Piew, Kay Hooi Keoy, and Padzil Hassan. 2011. Perceived Risk, Perceived Technology, Online Trust for the Online Purchase Intention in Malaysia. *International Journal of Business and Management* 6, 6 (June 2011), p167. doi:10.5539/ijbm.v6n6p167

- [6] António Pedro Costa, Grzegorz Bryda, Prokopis A. Christou, and Judita Kasperuniene. 2025. AI as a Co-researcher in the Qualitative Research Workflow: Transforming Human-AI Collaboration. *International Journal of Qualitative Methods* 24 (Sept. 2025), 16094069251383739. doi:10.1177/16094069251383739
- [7] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *Comput. Surveys* 57, 6 (June 2025), 1–39. doi:10.1145/3712001
- [8] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Management Information Systems Quarterly* 13, 3 (Sept. 1989), 319–340. doi:10.2307/249008 _eprint: https://misq.umn.edu/misq/article-pdf/13/3/319/903/6_davis.pdf.
- [9] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–25. doi:10.1145/3479856
- [10] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2024. CoAlCoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Transactions on Computer-Human Interaction* 31, 1 (Feb. 2024), 1–38. doi:10.1145/3617362
- [11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shitong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jia Shi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shutong Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Li, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yuxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (Sept. 2025), 633–638. doi:10.1038/s41586-025-09422-z
- [12] Leah Hamilton, Desha Elliott, Aaron Quick, Simone Smith, and Victoria Choplin. 2023. Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. *International Journal of Qualitative Methods* 22 (Oct. 2023), 16094069231201504. doi:10.1177/16094069231201504
- [13] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LORA: LOW-RANK ADAPTATION OF LARGE LAN- GUAGE MODELS. (2022).
- [14] David M. J. Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetis, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. Computational social science: Obstacles and opportunities. *Science* 369, 6507 (Aug. 2020), 1060–1062. doi:10.1126/science.aaz8170
- [15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. [n. d.]. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ([n. d.]).
- [16] David L. Morgan. 2023. Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods* 22 (Oct. 2023), 16094069231211248. doi:10.1177/16094069231211248
- [17] Kien Nguyen-Trung. 2025. ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? *Quality & Quantity* 59, 6 (Dec. 2025), 4945–4978. doi:10.1007/s11135-025-02165-z
- [18] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastian Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpouras, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b & gpt-oss-20b Model Card. doi:10.48550/arXiv.2508.10925 arXiv:2508.10925 [cs].
- [19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) (ICML '23). JMLR.org, Article 1182, 27 pages.
- [20] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445591
- [21] Kira Sam and Raja Vavekanand. 2024. Llama 3.1: An In-Depth Analysis of the Next Generation Large Language Model. (2024). doi:10.13140/RG.2.2.10628.74882 Publisher: Unpublished.
- [22] Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and LuEttaMae Lawrence. 2024. The Role of Generative AI in Qualitative Research: GPT-4's Contributions to a Grounded Theory Analysis. In *Proceedings of the Symposium on Learning Design and Technology*. ACM, Delft Netherlands, 17–25. doi:10.1145/3663433.3663456
- [23] GLM-4 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huihong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghui Xue, Mingming Zhao, Mingshuo Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangting Yang, Shaoyou Lu, Shijie Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihang Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguan Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. doi:10.48550/arXiv.2508.06471 arXiv:2508.06471 [cs].
- [24] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiguo Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu

- Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaying Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. 2025. Kimi K2: Open Agentic Intelligence. doi:10.48550/arXiv.2507.20534 arXiv:2507.20534 [cs].
- [25] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 75–78. doi:10.1145/3581754.3584136
- [26] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. doi:10.1145/3613905.3650732
- [27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. doi:10.48550/arXiv.2505.09388 arXiv:2505.09388 [cs].
- [28] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (June 2024), 100211. doi:10.1016/j.hcc.2024.100211
- [29] Robert K. Yin. 2018. *Case study research and applications : design and methods* (6th ed ed.). Sage. <https://cir.nii.ac.jp/crid/1970586434844156305>