Automating Thematic Analysis with Multi-Agent LLM Systems*

Sreecharan Sankaranarayanan[1,*], Conrad Borchers[1,*], Sebastian Simon[2], Elham Tajik[3], Amine Hatun Atas[4], Berkan Celik[5], Francesco Balzan[6], Bahar Shahrokhian[7]

[1] *Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA*
[2] *Copenhagen University, Nørregade 10, 1172 København, Denmark*
[3] *Florida State University, 222 S Copeland St, Tallahassee, FL 32306, USA*
[4] *Galatasaray University, Ortaköy, Çırağan Cd. No:36, 34349 Beşiktaş/İstanbul, Türkiye*
[5] *Van Yuzuncu Yil University, Bardakçı, Yüzüncü Yıl Üniversitesi Kampüsü, 65090 Tuşba/Van, Türkiye*
[6] *University of Bologna, Via Zamboni, 33, 40126 Bologna BO, Italy*
[7] *Arizona State University, 1151 S Forest Ave, Tempe, AZ 85281*

Abstract

Thematic analysis (TA) is a method used to identify, examine, and present themes within data. TA is often a manual, multistep, and time-intensive process requiring collaboration among multiple researchers. TA's iterative subtasks, including coding data, identifying themes, and resolving inter-coder disagreements, are especially laborious for large data sets. Given recent advances in natural language processing, Large Language Models (LLMs) offer the potential for automation at scale. Recent literature has explored the automation of isolated steps of the TA process, tightly coupled with researcher involvement at each step. Research using such hybrid approaches has reported issues in LLM generations, such as hallucination, inconsistent output, and technical limitations (e.g., token limits). This paper proposes a multi-agent system, differing from previous systems using an orchestrator LLM agent that spins off multiple LLM sub-agents for each step of the TA process, mirroring all the steps previously done manually. In addition to more accurate analysis results, this iterative coding process based on agents is also expected to result in increased transparency of the process, as analytical stages are documented step-by-step. We study the extent to which such a system can perform a full TA without human supervision. Preliminary results indicate human-quality codes and themes based on alignment with human-derived codes. Nevertheless, we still observe differences in coding complexity and thematic depth. Despite these differences, the system provides critical insights on the path to TA automation while maintaining consistency, efficiency, and transparency in future qualitative data analysis, which our open-source datasets, coding results, and analysis enable.

Keywords

multi-agent systems, thematic analysis, large language models, LLMs, qualitative analysis, qualitative coding

1. Introduction

Since GPT-3 emerged in 2020 [1], generative AI systems have driven innovative usage for different tasks and in various contexts. Lately, such systems have also been identified as potential support for qualitative analysis [2]. While classic large language models (LLMs) were able to assist in complex processes such as thematic analysis (TA) in isolated steps (e.g., code creation), initial studies showed that a tightly coupled hybrid configuration with human researchers may lead to the most desirable outcomes [3].

Joint Proceedings of LAK 2025 Workshops, co-located with the 15th International Conference on Learning Analytics and Knowledge (LAK 2025), Dublin, Ireland, March 03–07, 2025.
* Corresponding author.
sreecharan93@gmail.com (S. Sankaranarayanan); cborcher@sc.cmu.edu (C. Borchers); sas@psy.ku.dk (S. Simon); et22e@fsu.edu (E. Tajik); ahatas@gsu.edu.tr (A. H. Atas); berkancelik@yyu.edu.tr (B. Celik); francesco.balzan3@uinbo.it (F. Balzan); bshahrok@asu.edu (B. Shahrokhian)
0000-0001-9122-6870 (S. Sankaranarayanan); 0000-0003-3437-8979 (C. Borchers); 0000-0003-3218-2032 (S. Simon); 0000-0001-6325-353X (A. H. Atas); 0000-0002-7068-8918 (B. Celik); 0000-0001-5962-4254 (F. Balzan); 0000-0002-3737-4714 (B. Shahrokhian)

TA is a qualitative research method for detecting, analyzing, and reporting patterns in qualitative data [4]. Inductive TA, a variant of TA, helps to describe the essence of the underlying data [5]. Typically, TA is a manual, multi-stage process that is both time-intensive and reliant on the collaboration of multiple researchers to ensure consistency. It involves iterative tasks—such as coding data, identifying categories, themes, and resolving possible discrepancies among coders, which are particularly laborious when handling large datasets.

While previous research demonstrates that LLMs can support isolated steps in TA, fully automated systems have sometimes been criticized as struggling to capture the nuanced, context-driven insights essential for robust qualitative research, due to their limitations in clarity, mutual exclusivity, and reliability [2]. Relatedly, LLMs are prone to hallucination and over longer outputs can deviate from their initial prompt subject. Hybrid models integrating human expertise with AI yield more balanced outcomes [3], yet they still rely on manual intervention for critical tasks, and it remains unclear which phases of inductive TA can be reliably outsourced to AI without compromising interpretative depth and validity. Addressing this gap is crucial for optimizing the balance between efficiency and analytical rigor, ultimately paving the way for fully or largely autonomous systems that deliver consistency, efficiency, and transparency while preserving the richness of human insight.

In this paper, we present a novel multi-agent system that leverages LLMs to automate larger task sequences autonomously while remaining robust, transparent and valid throughout its qualitative data analysis, building on and extending previously isolated LLM calls for intermediate steps. We first present a short overview of existing work on the topic, before outlining the architecture of our system.

2. Related Work

The many potential applications of LLMs have inspired researchers to explore their application for data analysis. Multiple studies have explored the use of LLMs, such as GPT, for tasks related to TA [2,6]. For instance, Barany et al. tested GPT-4 during inductive codebook development, assessing its potential to address key challenges in manual coding, including time constraints, inconsistencies, and human error. Their findings highlighted that a hybrid approach—where humans and AI collaborate—balances efficiency and reliability. In contrast, fully automated methods relying solely on GPT, while highly time-efficient, exhibited limitations in clarity, mutual exclusivity, and reliability [2]. Similarly, Paoli employed GPT-3.5 Turbo to conduct inductive TA on two datasets of semi-structured interviews: one with 13 video game players and another with 10 lecturers teaching data science. The study also identified the specific phases of TA where LLMs struggle. Paoli emphasized that Phase 1 (data familiarization) and Phase 6 (report generation) require human intervention, whereas other stages showed greater automation potential [6]. In addition to inductive approaches, several studies have investigated the use of LLMs for deductive qualitative analysis [7,8]. For example, Xiao et al. applied GPT-3 for deductive coding tasks, reporting substantial agreement with human coders on question complexity (Cohen's Kappa = 0.61) and fair agreement on syntactic structure (Cohen's Kappa = 0.38) [8].

In summary, most related studies investigate different versions of GPT and their potential to aid in developing codes and themes for both inductive and deductive TA. These studies typically assess reliability, validity, and interpretative depth by comparing GPT-generated outputs with those produced by humans. However, several limitations remain in applying GPT to TA. Yan et al. highlight key shortcomings, including trustworthiness (interviewed researchers stated the need to manually verify the results of LLM output), consistency (the same prompt does not generate the same results), data capacity, contextual understanding (interviewed researchers mentioned that interpretation of data relied on provided context exclusively), and interface constraints (impossibility to upload larger datasets in GPT-3.5) [3]. Similarly, Tai et al. examined GPT-3.5 for deductive coding and found that, while its performance, as measured in [9], was comparable to traditional human coding, notable limitations persisted, such as algorithmic constraints ("LLMs rely on patterns and

structures present in the training data, and if specific linguistic nuances or subtleties are absent, the model's understanding may be limited.", p. 11) and token limits (GPT-3.5 can process a maximum of 2048 characters per input) [9]. Overall, these studies indicate that existing methods are not fully automated yet. In other words, for existing systems, only having a researcher describe an analysis method (e.g. "conduct an inductive thematic analysis") and provide the dataset of interest to the system to obtain valid and reproducible results as output is not feasible yet. The present study makes progress towards that research goal by introducing a multi-agent system taking as input the process description and dataset, and producing as output a report with themes and codes directly linked to the underlying dataset, resulting in a robust, transparent and valid output.

3. System and Benchmark

3.1. Dataset

For this study, we use an open-source dataset consisting of 200 statements from four questions on Computer-Supported Collaborative Learning (CSCL), Collaborative Learning (CL), and the future of the CSCL field, contributed by researchers from diverse cultural and linguistic backgrounds [10]. This dataset provided an ideal testbed, as the definitions reflect variations in terminology, conceptual focus, and context while remaining concise and semantically close. The collaborative nature of the analysis task, requiring agreement among experts, added an additional layer of complexity, making it an excellent candidate for evaluating LLM performance in TA. Finally, our data were suited for this preliminary evaluation of our system since the knowledge in our data is likely new and not well-known to LLMs, as it represents the opinions of select experts in a research field. The criterion of novelty is important as genuinely novel data and analysis tasks test LLM's reasoning abilities as opposed to invoking aspects of its pre-existing knowledge base.

3.2. Multi-Agent System

Our system leverages the Claude Sonnet 3.5 model in a multi-agent system (MAS) architecture [21, 22] designed to mimic the collaborative coding process of humans (e.g., [20]). The model shows good performance on reasoning over text in the DROP benchmark (F1 Score = 0.87) [11], making it an adequate choice for text-processing tasks.

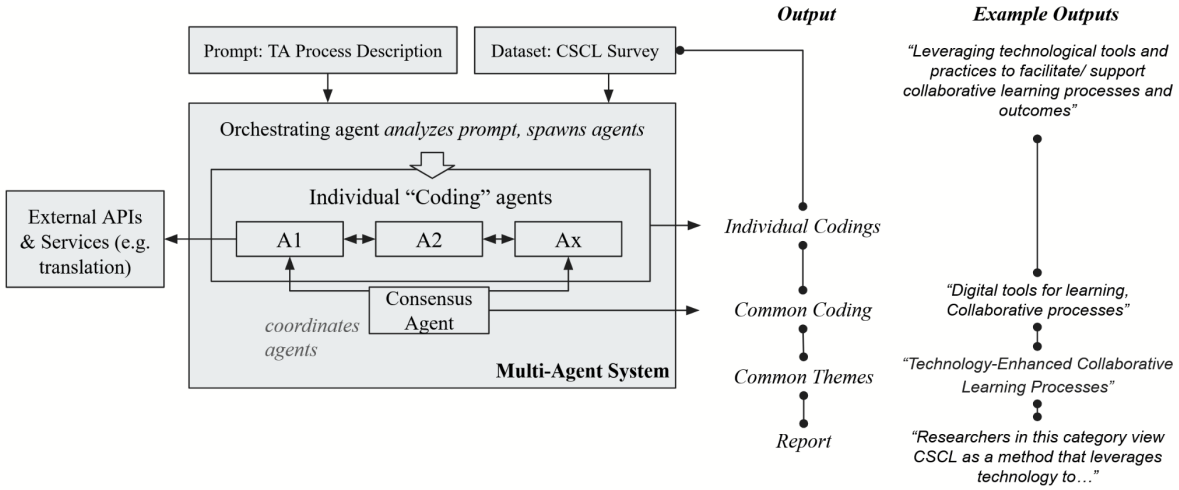

Figure 1: An illustration of the system design and its outputs

The system begins with a single orchestrator agent that is given an "agentic" prompt as input. This prompt provides the agent with the input data and instructs the agent about the task, in this case TA, along with a detailed description of the TA process. The prompt is considered agentic because the agent processes the prompt and spawns multiple sub-agents, as necessary, to complete sub-tasks. In our case, for example, agents are spawned to complete each of the individual steps in the TA

process, i.e., individual coding of data, consensus-building, theme identification, etc. Sub-agents may be spawned for other steps not part of the TA process as well, such as data pre-processing, so long as those instructions are included in the prompt. For example, some of the input data is in French and the agent is instructed to translate these to English before performing the thematic analysis. The orchestrator agent thus spawns a translator sub-agent to complete these translations prior to proceeding with the TA steps. The agents spawned for the TA process are shown in Figure 1. Individual coding agents produce codes which a consensus agent coordinates to produce common *codes*. This may involve multiple model calls from the individual agents to reconcile differences in codes identified by the consensus agent. Once consensus is achieved, these codes are assigned by the consensus agent to the entire dataset. From these codes, independent agents, as before, aggregate them into code-sets representing a theme in the data. Once consensus is reached on these themes and their descriptions, the final report is assembled. The complete output, system and user prompt can be found here[1].

The system mirrors the coding and theme-finding process of humans insofar as it adapts to problems in consensus-finding phases by re-iterating over codes, reflecting the dynamic and creative nature of inductive thematic analysis. Moreover, the system detects and labels unrelated responses and translates any responses not in the prompted language by spawning dedicated translator agents (see Figure 1 for an illustration of the system).

4. Performance

In the following paragraphs, we describe three criteria–robustness, transparency, and validity–which TA tools like ours need to fulfill to perform highly-automated TA with sufficient quality. For each criterion, we describe preliminary evaluations of our system.

4.1. Robustness

Robustness in automated TA is defined as reasonable magnitudes of change in system output over multiple independent runs of the system under varying conditions or contexts [9]. Reproducibility is a major challenge in qualitative data analysis. While some authors argue TA is an inherently creative task and should vary when done repeatedly [13], automated systems should produce the same output over multiple runs and exhibit coherent results under comparable conditions to ensure reproducibility. However, it is important to note that LLMs are designed to produce variation in output as they are fundamentally probabilistic [25] and depending on parameters such as temperature. Temperature changes the probability distribution of the predicted next word. Our system makes use of a temperature greater than 0 to produce some variation in the individual coding phase. Nevertheless, the overall system's output was semantically consistent over 20 runs, a notable result given the variation of classic LLM systems [9].

4.2. Transparency

Transparency is the possibility of tracing and documenting the process and intermediary outputs of the system. Traditional LLM systems are blackboxes, i.e., opaque: There is no possibility to capture how the result was created from the text input, including for open-source LLMs which improve model parameter access but still require complex audits [23, 24]. Only recently have models like GPT acquired the ability to "reason," essentially auto-prompting themselves to produce answers to more complex problems or questions [25]. Such intermediate prompts can be analyzed and are oftentimes part of the output itself. In our case, the system produces not only a final report but also themes and codes. Not only are codes and themes reported, but codes are linked to data points, and themes linked to the codes. The system thus provides a high level of transparency. For instance, the system

[1] https://github.com/se6astians1mon/TA-MAS/blob/main/TA%20MAS%20Output%20For%20LAK%20Paper.pdf

identified the theme “Collaborative Processes”. It provided a description highlighting that “This theme focuses on the process and outcomes of people engaging in shared activities, emphasizing the collaborative nature of learning.” and that “Researchers in this category view Collaborative Learning as a process where learners work together on shared activities. The emphasis is on the collaborative nature of the learning experience, which can occur both synchronously and asynchronously”. The description reflects the occurrence in underlying data points of synchronous and asynchronous settings of collaboration. The associated codes were “Shared Activity, Group Collaboration, Mutual Assistance, Reflective Activities”. The code for “Shared activity” was used to describe data points like “sequence d'apprentissage lors de laquelle les élèves sont impliquées dans une action conjointe et partagent une vision commune des tâches et de problèmes soulevés[2]” (a data item that a translation agent had previously translated for the other agents) or “Collaborative learning is an approach in which participants work together on tasks, often in small groups.” Given the chain of themes, codes and data points, the system output exhibits similar transparency to human-conducted TA (human TA reports and codebooks typically include concrete examples of data they are based on).

4.3. Validity

In general, a test or tool is deemed valid for a specific purpose if it accurately measures what it is intended to measure [19]. What TA intends to measure is derived from human consensus about the essence of the data and its themes. Past work has measured automated TA validity via alignment with human-derived themes [2, 9]. Inductive TA also has the purpose of informing the reader of the essence of the data in a final report [12]. Hence, It is thus important that the theme descriptions are clear and comprehensible to the intended audience.

We evaluated the system by comparing its performance to a manually conducted inductive TA by a research team. Three independent coders generated 553 codes from 200 open-ended responses, and a fourth researcher assisted in resolving discrepancies to achieve consensus on the final themes. This process demonstrated excellent reliability, through a value of 0.934 for Krippendorff’s Alpha [10].

A comparison of the AI-generated themes with the manual (human) analysis revealed that the proposed multi-agent system successfully automated much of the TA workflow. The system’s results nearly matched three of the four manually identified themes (e.g. “Role of technology” describing in what ways technology supports collaborative learning) of each of the four questions in the dataset, with partial alignment on the fourth (the MAS producing themes like “Common goals and objectives” whereas the coding team found the theme “goal specification” with a larger variety of codes). These findings underscore the system’s capacity to deliver high-quality TAs while substantially reducing the required time and effort.

Feedback from the human coding team confronted with the output of the MAS indicated a large overlap between automated coding and theme identification. Coders stated that the output would have been an interesting contribution to the initial, manual TAs at the theme creation stage, where researchers had to return repeatedly to the code and data level to produce meaningful themes, highlighting the coherence between themes and code sets by the multi-agent system.

5. Discussion

Extending the use of LLMs for TA proposed in recent research, the presented multi-agent system constitutes the next step in assisting researchers in the complex task of TA in largely automated ways [26]. Its successful application to a domain-specific dataset of CSCL definitions is initial evidence for the possibility to automate TA through the use of generative AI. While our empirical findings are not the first to report partial alignment between human-coded and AI-coded themes [2],

[2] *(french)* learning sequences in which students are engaged in a common action and share a common vision on tasks and identified problems

our approach critically differs in introducing an autonomous, multi-agent system with subtask division and transparent documentation of the analytics process between agents. The system's preliminary evaluation of robustness, transparency, and validity makes it a promising application for researchers and their qualitative datasets.

However, several limitations should be addressed to further refine the analysis process: For example, we observed that the system occasionally struggled with ambiguous or highly context-dependent text (for instance, one question in the survey asked participants to distinguish both cooperative and collaborative learning, with a clear tendency on task division revealed by the manual analysis but not by the MAS), underscoring the need for human oversight. This limitation could be overcome using techniques like Retrieval Augmented Generation (RAG) [13], but such evaluation remains subject to future research. Indeed, it is possible that some form of human supervision, as argued in past research, is indispensable in the analytical process [2].

5.1. Technical Limitations

The system has been tested on a dataset of 200 open-ended survey responses, which were limited to a single sentence. Many artifacts, such as interview data in human-centered design research in our field [28], as well as others [27] feature data with considerably longer text. LLMs have limitations in the context window, i.e., the size of the number of tokens in the input prompt. While these limits have increased considerably for state-of-the-art models (from GPT-3 with 4,000 tokens to Claude Sonnet 3.5 with 200,000 tokens [14]), there remains a limit, especially if agents in a MAS have to handle the context of other agents during collaborative tasks. The optimal splitting of longer text for agentic collaboration in our system remains subject to future research. In our current system, we expect financial and compute costs to scale non-linearly with the complexity of input data.

5.2. Performance Assessment

We have considered robustness, transparency and validity in comparison with a human-performed, manual thematic analysis. In this study, we only assessed robustness through repeated use with the same prompt. However, it is unlikely that researchers produce an exact same prompt even for the same task - robustness should therefore include some degree of robustness towards variation in the input prompt. While we judged transparency in the automated analysis as "sufficient", we highlight the possibility to further investigate the output capturing the model reasoning process behind each code and theme, which is subject to future research. Not only would this allow for better transparency, it could also allow human coders to critically assess their own reasoning in TA. We hypothesize that such detailed information could inform the design of adversarial configurations of automated TA-systems challenging each other's reasoning to further improve analysis quality. It will be interesting to see whether such competing systems produce a stabilizing result over time, potentially confirming the perspective of "one truth" in data, or not [12].

Relatedly, we ask how **validity** could be better defined - what does it mean if topics "nearly" align? How many topics should align to consider an automated TA valid? Should human-derived themes continue to be considered the ground truth for analysis? Recent research has begun to establish evaluation frameworks and metrics for assessing AI-driven thematic analysis. For example, Dunivin proposed a framework combining quantitative measures (e.g., intercoder reliability via Cohen's Kappa) with qualitative assessments of interpretative depth and content fidelity [15]. In parallel, Zhang et al. demonstrated that integrating these metrics enables a robust comparison between AI-generated outputs and those produced by human analysts, thereby clarifying which phases of inductive thematic analysis can be reliably automated without sacrificing interpretative nuance [16]. In future work we intend to build on and extend these assessment metrics to provide benchmark datasets and methods to assess fully or highly automated systems like ours. Furthermore, our initial assessment through the team having performed the manual TA analysis, should be confirmed and extended on by external experts presented with clear guidance on both codes and

themes. Future work will establish a standardized process and explore to what the process of assessing automated systems can itself be assisted by systems like ours.

5.3. Parameter Choice

LLMs attempt to produce human-like responses depending on their parameters, context and training dataset and are trained on large amounts of data produced by humans. Huber and Carenini highlight that biases embedded in training data and opaque decision-making processes can lead to distorted thematic outputs—especially in contexts that demand nuanced, context-driven interpretations such as educational research [17]. Incorporating explainable AI techniques, such as adversarial audits and chain-of-thought prompting [18], to enhance transparency and accountability seems a promising trajectory to improve the reliability of automated iTA systems and their alignment with robust research standards.

5.4. Ethical Considerations

Despite the promise of techniques like chain-of-thought prompting and adversarial audits to enhance transparency, Khan et al. note that "the interplay between model training, reinforcement learning, prompt wording and the dataset used in thematic analysis is complex and can lead to biased results whose causes are difficult or impossible to isolate" (p. 12) [28]. This complexity goes beyond technical opacity to encompass socio-technical opacity - the difficulty in tracing the influences and interests embedded in LLM design [29] - which raises concerns about deploying LLMs as autonomous agents in interpretative tasks such as thematic analysis due to the risk of a subtle transformation or reiteration of social, political and epistemic norms [30]. Consequently, many scholars advocate using LLMs as supportive tools, keeping the "human in the loop" (or even reinstating the "machine in the loop" to underscore the supporting role of LLMs in the process [31]) to preserve the essential interpretative and sense-making role of human analysts.

In summary, this study demonstrates the feasibility of using LLMs to support and automate TA, particularly in coding and theme identification. While substantial overlap with human TA was found for a domain specific dataset, challenges remain to achieve the same in-depth analysis that domain experts with experience in TA can achieve. Future work will focus on enhancing the system's ability to handle ambiguous data and expand its applicability to other datasets and languages. Our open-source datasets, coding results, and analysis will enable this future work[3].

Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 3.5 within a multi-agent LLM system to perform automated thematic analysis on qualitative data. After using this tool, the authors reviewed and interpreted the outputs as needed and take full responsibility for the publication's content.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems

[3] https://github.com/se6astians1mon/TA-MAS/blob/main/TA%20MAS%20Output%20For%20LAK%20Paper.pdf

(NeurIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

[2] Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., & Baker, R. S. (2024). ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), Artificial intelligence in education: AIED 2024 (Vol. 14830, pp. 99–107). Springer. https://doi.org/10.1007/978-3-031-64299-9_10

[3] 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 (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 191, 1–7. https://doi.org/10.1145/3613905.3650732

[4] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

[5] Borchers, C., Ooge, J., Peng, C., & Aleven, V. (in-press). How Learner Control and Explainable Learning Analytics on Skill Mastery Shape Student Desires to Finish and Avoid Loss in Tutored Practice. In LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland. ACM, New York, NY, USA.

[6] De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limit of the approach. Social Science Computer Review, 42(4), 997–1019. https://doi.org/10.1177/08944393231220483

[7] Ramanathan, S., Lim, L.-A., Rezazadeh Mottaghi, N., Buckingham Shum, S. (2025). When the prompt becomes the codebook: Grounded Prompt Engineering (GROPROE) and its application to belonging analytics. In LAK25: The 15th International Learning Analytics and Knowledge Conference (pp. 1–12). Dublin, Ireland. ACM. https://doi.org/10.1145/3706468.3706564

[8] Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., Oudeyer, P. Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 75–78). https://doi.org/10.1145/3581754.3584136

[9] Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M.,Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. International Journal of Qualitative Methods, 23, 16094069241231168. https://doi.org/10.1177/16094069241231168

[10] Simon, S., Borchers,C., Ataş, A. H., Tajik,E.,Celik, B., Čarapina, Mia, Liu, Y-D., Shahrokhian, B., Sankaranarayanan, S., Balzan, F., Molinari, G., Jagušt, T. & Liang, L. (2024). Exploring shared conceptual ground in Computer-Supported Collaborative Learning: A survey. (Manuscript submitted for review)

[11] Lam, Lina, (2025) GPT-4o Mini vs. Claude 3.5 Sonnet: A Detailed Comparison for Developers, https://www.helicone.ai/blog/gpt-4o-mini-vs-claude-3.5-sonnet

[12] Braun, Virginia, and Victoria Clarke. “Reflecting on Reflexive Thematic Analysis.” Qualitative Research in Sport, Exercise and Health 11, no. 4 (August 8, 2019): 589–97. https://doi.org/10.1080/2159676X.2019.1628806.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

[14] Zhu, Liang (2024) Claude 3.5 Sonnet vs GPT-4o: Context Window and Token Limithttps://prompt.16x.engineer/blog/claude-sonnet-gpt4-context-window-token-limit

[15] Dunivin, Z. O. (2024). Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. arXiv preprint, https://doi.org/10.48550/arXiv.2401.15170

[16] Zhang, He, Wu, Chuhao, Xie, Jingyi, Rubino, Fiona, Graver, Sydney, Kim, Chanmin, Carroll, John M., Cai, Jie (2024). When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding https://doi.org/10.48550/arXiv.2407.14925

[17] Patrick Huber and Giuseppe Carenini. 2022. Towards Understanding Large-Scale Discourse Structures in Pre-Trained and Fine-Tuned Language Models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2376–2394, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.170

[18] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., ... & Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155. https://doi.org/10.48550/arXiv.2308.08155

[19] Cohen, R. J., & Swerdlik, M. E. (2017). Psychological testing and assessment (9th ed.). McGraw-Hill Education.

[20] Qi, R., Li, W., & Lyu, H. (2024, November). Generation of Scientific Literature Surveys Based on Large Language Models (LLM) and Multi-Agent Systems (MAS). In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 169-180). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9443-0_14

[21] Ng, A. (2024, April 17). Agentic design patterns part 5: Multi-agent collaboration. The Batch. Retrieved February 19, 2025, from https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-5-multi-agent-collaboration/

[22] Tran, K. T., Dao, D., Nguyen, M. D., Pham, Q. V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-Agent Collaboration Mechanisms: A Survey of LLMs. arXiv preprint arXiv:2501.06322.

[23] Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., ... & Hadfield-Menell, D. (2024, June). Black-box access is insufficient for rigorous ai audits. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2254-2272). https://doi.org/10.1145/3630106.3659037

[24] Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. arXiv preprint arXiv:2402.01761. https://doi.org/10.48550/arXiv.2402.01761

[25] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.https://doi.org/10.48550/arXiv.2203.11171

[26] Asano, Y., Sankaranarayanan, S., Majd, S. A. K. R., & Bogart, C. (2021, November). A Thematic Summarization Dashboard for Navigating Student Reflections at Scale. In International Conference on Computers in Education.

[27] Rose, C., Sankaranarayanan, S., Shuang, B., & Bury, S. (2019, April). Opportunities for Text Mining in Service of Chemical Engineering. In 2019 Spring Meeting and 15th Global Congress on Process Safety. AIChE.

[28] Khan, A. H., Kegalle, H., D'Silva, R., Watt, N., Whelan-Shamy, D., Ghahremanlou, L., & Magee, L. (2024). Automating Thematic Analysis: How LLMs Analyse Controversial Topics. arXiv preprint arXiv:2405.06919. https://doi.org/10.48550/arXiv.2405.06919

[29] Balzan, F., Munarini, M., Angeli, L. (2024). Who Pilots the Copilots?. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science(), vol 14830. Springer, Cham. https://doi.org/10.1007/978-3-031-64299-9_42

[30] Amoore, L., "Machine learning political orders," Review of International Studies, vol. 49, no. 1, pp. 20–36, 2023. http://doi.org/10.1017/S0260210522000031

[31] Balzan, F., Zanellati, A., Zingaro, S.P., Gabbrielli, M. (2025). A 2-Step Methodology for XAI in Education. In: Meo, R., Silvestri, F. (eds) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023. Communications in Computer and Information Science, vol 2134. Springer, Cham. https://doi.org/10.1007/978-3-031-74627-7_6