Position: Thematic Analysis of Unstructured Clinical Transcripts with Large Language Models

Seungjun Yi[1*], Joakim Nguyen[1], Terence Lim[1], Andrew Well[2], Joseph Skrovan[1], Mehak Beri[1], YongGeon Lee[1], Kavita Radhakrishnan[1], Liu Leqi[1], Mia Markey[1], Ying Ding[1]
[1]The University of Texas at Austin, Austin, TX, USA
[2]Department of Cardiac Surgery, Vanderbilt University School of Medicine, Nashville, TN, USA
{charlie.yi,jhn001,terence.lim}@utexas.edu,
andrew.well@vumc.org, jskrovan@fastmail.fm, mb72384@eid.utexas.edu,
,{yg910524}@utexas.edu, kradhakrishnan@mail.nur.utexas.edu,
{leqiliu,mia.markey}@utexas.edu, ying.ding@ischool.utexas.edu

Abstract

This position paper examines how large language models (LLMs) can support thematic analysis of unstructured clinical transcripts, a widely used but resource-intensive method for uncovering patterns in patient and provider narratives. We conducted a systematic review of recent studies applying LLMs to thematic analysis, complemented by an interview with a practicing clinician. Our findings reveal that current approaches remain fragmented across multiple dimensions including types of thematic analysis, datasets, prompting strategies and models used, most notably in evaluation. Existing evaluation methods vary widely (from qualitative expert review to automatic similarity metrics), hindering progress and preventing meaningful benchmarking across studies. We argue that establishing standardized evaluation practices is critical for advancing the field. To this end, we propose an evaluation framework centered on three dimensions: *validity*, *reliability*, and *interpretability*.

1 Introduction

Thematic analysis (TA) is one of the most widely used methods for qualitative data analysis, providing a structured yet flexible framework for identifying, analyzing, and reporting patterns (*themes*[2]) within textual data [1, 2]. In clinical contexts, TA is frequently applied to unstructured patient interview transcripts to reveal underlying meanings and generate actionable insights for improving care [3], with the process commonly guided by Braun and Clarke's six-phase framework [1]. Researchers begin by familiarizing themselves with the data through repeated readings and the identification of important keywords (Step 1–2). These keywords are then used to generate initial *codes*[3] (Step 3), which serve as the foundation for identifying preliminary *themes* (Step 4). The *themes* are iteratively reviewed, refined, and clearly defined and named through interpretive analysis (Step 5). In the final stage, a structured conceptual model is developed to represent the thematic relationships and underlying constructs (Step 6). To ensure credibility, at least two trained researchers with domain expertise must independently engage with the data and participate in multiple rounds of discussion to achieve consensus. Each year, over 900,000 U.S. healthcare interviews undergo thematic analysis across academic, clinical, and research settings [4, 5]. Most involve unstructured data, requiring

*Corresponding author

[2]*Themes*: Broader patterns or categories that emerge from grouping related codes to capture significant aspects of participants' experiences or perspectives in relation to the research objectives

[3]*Codes:* concise labels that identify features of the textual data relevant to the research question

The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance

inductive thematic analysis (ITA), which extracts *themes* directly from transcripts without predefined *codes*[6, 7, 8]. Manual ITA demands over 6.1 million hours annually equivalent to 3,000 full-time jobs and $305.4 million in cost which limits scalability and delaying insights in fast-moving healthcare environments [9, 10, 11]. Automating ITA is therefore critical for timely, scalable, and reproducible *theme* generation.

Recent advances in large language models (LLMs) offer a promising alternative for automating key components of the six-step manual inductive thematic analysis process [12, 13]. LLMs can rapidly generate initial *codes* (Step 2) and candidate *themes* (Steps 3–5) from interview transcripts in under 10 minutes, a substantial reduction compared to the 5–8 hours required by human analysts [14, 15].

However, the integration of LLMs to fully automate the TA process remains in its early stages. Existing studies primarily focus on specific stages of thematic analysis, such as *coding* [16], and *theme* generation [17, 18]. They also vary widely in methodological choices, including the type of thematic analysis employed (i.e., inductive vs. deductive), the models used (e.g., `gpt-4o`, `Llama`), and the prompting strategies adopted (e.g., few-shot prompting, chain-of-thought), which makes it difficult to establish a coherent picture of progress. This challenge is further compounded by the rapid pace of LLM development, which continuously reshapes the methodological landscape. More importantly, evaluating the results of thematic analysis remains a major challenge. Evaluation practices remain inconsistent, relying primarily on expert qualitative assessment and comparisons between LLM-generated and human-generated themes, which limits both reproducibility and comparability across studies.

In this work, we aim to:

1. Provide an overview of recent developments and research progress in thematic analysis supported by large language models (LLMs).
2. **Highlight the need for standardized evaluation in LLM-based thematic analysis, and propose a evaluation framework based on validity, reliability, and interpretability.**
3. Emphasize the need for end-to-end frameworks to ensure the practical utility of thematic analysis in clinical settings.

2 Methods

Literature Search and Selection Our data collection targeted publications addressing thematic analysis (TA) in relation to large language models (LLMs). Using arXiv metadata (cutoff: 2025–08–15), we identified relevant papers through keyword-based searches of titles and abstracts. Summary statistics are reported in Table1, with full search details in the Appendix C. To address the limitation that some peer-reviewed papers may be missing from arXiv, we supplemented our search with Elicit [19] and manual queries on major indexing platforms (see Appendix E for details). Full procedures, verification steps, survey prompts, and keyword lists are provided in Appendix D and Table 3. From the collected literature, we reviewed 56 recent (past three years from the cutoff date) studies applying LLMs to thematic analysis, focusing on five dimensions: analysis type, models, datasets, prompting strategies, and evaluation methods. We included peer-reviewed publications, preprints, while excluding theses, dissertations, and incomplete work.

Expert Interview In addition to the systematic literature review, we conducted an in-depth interview with a practicing cardiac surgeon with eight years of clinical experience, who has previously performed manual TA of patient transcripts [20]. The interview was held over a two-hour Zoom session, during which the surgeon was provided with a summary of the literature review findings and invited to discuss both their implications and the broader role of LLMs in augmenting thematic analysis workflows. Detailed insights from this discussion are presented in Section 4.

3 Results

3.1 Trends and Dataset Characteristics of LLM-Supported Thematic Analysis (TA)

From Table 1, we observe that TA leveraging LLMs is a relatively recent phenomenon, with the first work appearing after the release of GPT-3.5 in Nov 2022 [21]. The number has been growing

exponentially since then (Figure 1). As of Aug 2025, total of 46 papers were identified in the arXiv metadata, which only three were related to healthcare.

Table 1: Summary of retrieved papers containing keywords related to thematic analysis and LLMs. The additional "(+11)" indicates papers identified outside of arXiv. Keyword list definitions are provided in Appendix Table 3.

Keyword Category	**Count**
Thematic Analysis (`core_ta`)	381
Inductive Thematic Analysis (`inductive_ta`)	14
Deductive Thematic Analysis (`deductive_ta`)	3
Thematic Analysis (`core_ta`) + LLMs (`llm_generic`)	45 (+11)
Thematic Analysis (`core_ta`) + LLMs (`llm_generic`) + Healthcare (`healthcare`)	3
Total TA (any category above)	428

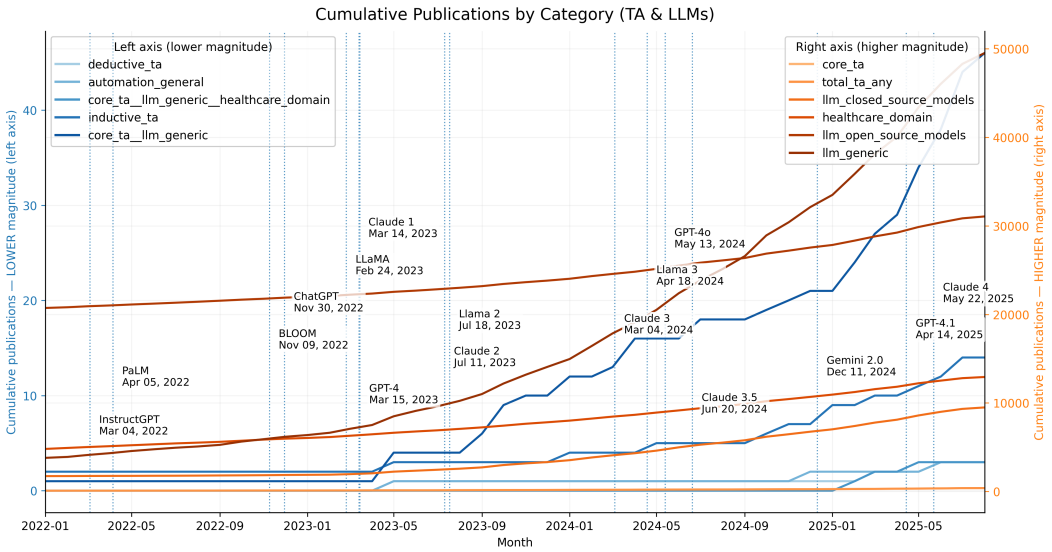

Figure 1: Cumulative number of publications mentioning thematic analysis (TA) and LLMs by category since Jan 2022. Research explicitly combining TA with LLMs began to emerge following the release of GPT-4. Vertical lines indicate major milestones in LLM development (see Appendix B). The categories shown correspond to those summarized in Table 1.

3.2 Survey of LLM-Assisted Thematic Analysis (TA)

We explored five key dimensions of LLM-assisted TA, including types of TA adopted, models leveraged, datasets analyzed, prompting strategies employed, and evaluation methods applied (Table 2).

Type of Thematic Analysis: Inductive vs. Deductive. *Deductive* TA, grounded in a priori codebooks, provides clearer definitions and comparability, whereas *inductive* TA generates themes directly from transcripts, enabling the derivation of novel insights. Among the 56 studies, inductive TA was the most prevalent, accounting for 36 papers (64%). Hybrid approaches that combined inductive and deductive elements were the second most common (22%), while purely deductive analyses were relatively rare (9%). A small fraction of work (5%) employed other or less clearly specified strategies. *Deductive* and *inductive* TA require different evaluation approaches. Consequently, evaluation methods developed for one are difficult to transfer to the other, which contributes to persistent inconsistency in evaluation practices.

Models Employed. Across the surveyed literature, the majority of studies relied on the OpenAI GPT family, with GPT-3.5 [22] and GPT-4 [23] variants serving as the primary analytic engines.

Table 2: Frequency summary across dimensions for 56 papers (Aug 2022–Aug 2025). Since individual papers may span multiple categories, counts can exceed 56. Categories deemed not applicable were excluded from the table.

Dimension	Attributes (# out of 56 / %)
Types of TA	Inductive (36/64%), Few-shot (9/16%), Hybrid (12/22%), Other (3/5%)
Models (by family)	GPT (32/58%), Claude (7/13%), LLaMA (6/11%), Gemini (6/11%), Mistral/Mixtral (3/5%), DeepSeek (4/7%), Qwen (2/4%), Other/Custom (7/13%)
Datasets by Domain	Social Media/Online Communities (14/25%), Education (12/21%), Software Engineering/Programming (11/20%), Arts/Humanities/Other (10/18%), Healthcare/Clinical (9/16%)
Prompting	Zero-shot (19/35%), Few-shot (9/16%), Chain-of-thought (7/13%), Self-consistency/Reflexion (8/15%), Tool/Agent use (7/11%)
Evaluation	Human qualitative review (22/40%), Automatic text metrics (16/27%), Task-based/utility eval (7/13%), Hybrid (11/20%)

A smaller number incorporated Claude [24] or Gemini [25] models, most often as comparative baselines rather than central analytic tools. Open-source families such as LLaMA [26], Mistral [27], and DeepSeek [28] appeared in exploratory or ensemble configurations, reflecting growing interest in transparency and cost-efficient alternatives. Only a handful of papers examined fine-tuned or specialized deployments (e.g., InCoder [29], CodeGen [30], ProgGP [31]), typically for narrow use cases. A detailed breakdown by model family is provided in Appendix F.

Datasets by Domain. The surveyed studies drew on diverse domains, with social media and online communities [32, 33, 34] most common (25%), followed by education [35, 36] (21%), software engineering and programming [35, 37] (20%), and healthcare or clinical contexts [13, 15, 38] (16%). The remaining 18% were arts, humanities, and other fields such as court cases, disciplinary decisions, and design workshops.

Prompting Strategies. Studies most commonly relied on zero-shot prompting (i.e., with no specific examples given), often implemented through multi-stage pipelines, with smaller proportions using few-shot templates, chain-of-thought reasoning, or iterative refinement methods such as self-consistency and Reflexion. A minority explored multi-agent or tool-based prompting, while some did not employ prompting at all, focusing instead on fine-tuned or human-coded approaches. Overall, zero-shot prompting dominated for its accessibility and alignment with inductive thematic analysis, whereas more elaborate prompting designs reflected attempts to stabilize outputs or approximate collaborative coding practices. A fuller breakdown is provided in Appendix G. Also, most existing approaches still rely on human-in-the-loop workflows that require full transcript review, which severely limits scalability. Since familiarization (Step 1 of Braun & Clarke's six-step thematic analysis process [1]) with transcripts is the most time-consuming step, retaining this requirement is essential to the practical utility of LLM-assisted thematic analysis in clinical settings.

Evaluation. Evaluation of LLM-assisted thematic analysis remains fragmented, with no universally accepted gold standard. Human qualitative review was the most common approach ($\approx 40\%$), involving inter-coder agreement checks, reflexive discussions, and expert evaluations of plausibility and thematic coverage, although reproducibility was often limited. Inter-coder agreement was typically assessed using Krippendorff's alpha [39], a reliability coefficient that accounts for chance agreement, supports different levels of measurement, and handles incomplete data, or Cohen's kappa [40], which measures agreement between two raters under the assumption of complete data and equal category distribution. Other criteria beyond inter-coder agreement were measured through surveys in which experts rated on a 1–5 Likert scale [15, 16]. About one quarter of studies (27%) evaluated LLM-generated themes against human ground truth using automatic similarity metrics. Lexical overlap metrics such as Jaccard Index, fuzzy match, ROUGE [41], BLEU [42], GLUE [43], and METEOR [44] assess similarity based on shared words or n-grams[4]. Semantic similarity metrics, including cosine similarity, BERTScore [45], compare vector embeddings[5]. Classifier-based accuracy

[4]Overlapping sequences of one or more consecutive words.

[5]Numerical representations of text that capture semantic meaning, generated by models such as `all-mpnet-base-v2` and `text-embedding-3-large`.

and F1 scores were occasionally applied. A smaller subset (13%) evaluated outputs indirectly through task-based measures (e.g., learning outcomes), while about one fifth (20%) adopted hybrid designs combining human judgments with computational metrics described above [16].

However, major challenges in evaluation remain. Even when the same metric is reported, differences in underlying embeddings hinder comparability and synthesis across studies. Issues of transparency further compound these challenges: the surveyed work do not disclose the complete ground truth, often citing privacy or IRB restrictions, while others provide no explicit justification for withholding this information. For the usage of automatic metrics, a major limitation of directly assessing similarity between human- and LLM-generated themes is the lack of one-to-one mappings. Thematic analysis inherently produces sets of themes that only partially overlap: a single LLM-generated theme may correspond to multiple human themes, while some human themes may have no direct counterpart.

To address these limitations, we propose evaluating automated inductive thematic analysis along three key dimensions: **validity**, **reliability**, and **interpretability**. (1) **Validity.** Any set of generated themes may not respect a 1:1 mapping between humans and LLMs, so we compute **maximum-weight bipartite matching** on a similarity matrix $\mathbf{S}$, where S_{ij} denotes the similarity between human theme i and model theme j. Use both: (a) Lexical overlap: Jaccard; optionally ROUGE for n-gram overlap. (b) Semantic similarity: cosine over sentence embeddings or BERTScore / MoverScore / BLEURT [45], and report: Precision/Recall/F1@match, Coverage@τ (fraction of human themes matched above a threshold), *Redundancy* (mean intra-LLM theme similarity), and *Novelty rate* (LLM themes with no matched human counterpart). (2) **Reliability.** Krippendorff's α or Cohen/Fleiss κ can be used as diagnostics for vague definitions or unclear code boundaries. To test *stability*, re-run parts of the pipeline with different random seeds or bootstrap samples and compute Adjusted Rand Index (ARI) or Variation of Information (VI) for code assignments. At the theme level, system-generated themes can be aligned with human themes by comparing overlap in supporting excerpts. When applying LLMs to the pipeline, *confirmability* assesses whether themes are data-driven or driven by the internal biases inherent across different LLMs. (3) **Interpretability.** Analytical depth and nuance remain challenging [46]. Mitigations include rationale generation [47], expert adjudication [15], and adaptive codebooks [33]. Excerpts within a theme should be consistent in meaning, and different themes should be distinguishable. Embedding similarity can check *coherence* defined as the average similarity among quotes within each theme and *distinctiveness* as the distance between theme centroids [48]. Report *coverage* of (i) passages assigned to any theme and (ii) participants represented by each theme. The *credibility* of themes assesses whether the generated themes faithfully represent the data. For domain-specific contexts, combine automated metrics with human-in-the-loop validation [49, 34].

Other considerations: Cost of Thematic Analysis. The cost of using LLMs has become significantly more feasible. Recent analyses estimate inference prices around \$0.15–\$2.50 USD per million input tokens and \$0.60–\$10 USD per million output tokens, compared to the \$200,000+ typically required for manual thematic analysis [50, 51].

4 Discussion and Conclusion

Our results indicate that current LLM-based approaches to thematic analysis remain fragmented, especially in the evaluation methods used across studies. As the clinician in our study interview observed, evaluating inductive thematic analysis can feel like "*shooting in the dark*," given the absence of clear ground truths or universally accepted metrics. The current disparate evaluation practices prevents direct comparison across studies. Without such standardization, research efforts remain fragmented and the development of an end-to-end automated thematic analysis framework with LLMs is hindered.

In summary, while LLMs hold considerable potential to transform thematic analysis, progress will remain fragmented without common standards for evaluation. By advancing validity, reliability, and interpretability as shared dimensions, the field can move toward more rigorous, scalable, and clinically meaningful applications of automated qualitative research.

References

[1] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[2] Virginia Clarke and Virginia Braun. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, 26(2):120–123, 2013.

[3] Sirwan Khalid Ahmed, Ribwar Arsalan Mohammed, Abdulqadir J. Nashwan, Radhwan Hussein Ibrahim, Araz Qadir Abdalla, Barzan Mohammed M. Ameen, and Renas Mohammed Khdhir. Using thematic analysis in qualitative research. *Journal of Medicine, Surgery, and Public Health*, 6:100198, 2025.

[4] Melissa DeJonckheere and Lisa M Vaughn. Semistructured interviewing in primary care research: a balance of relationship and rigour. *Family Medicine and Community Health*, 7:e000057, 2019.

[5] Konstantina Vasileiou, Julie Barnett, Susan Thorpe, and Terry Young. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18(1):148, Nov 21 2018.

[6] Kaitlyn A. Campbell, Elizabeth Orr, Pauline Durepos, Linh Nguyen, Lanting Li, Caitlin Whitmore, Paula Gehrke, Lorraine Graham, and Susan M. Jack. Reflexive thematic analysis for applied qualitative health research. *The Qualitative Report*, 26(6):2011–2028, 2021.

[7] Bethan Taylor, Catherine Henshall, Sara Kenyon, Ian Litchfield, and Sheila Greenfield. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? a mixed methods study comparing rapid and thematic analysis. *BMJ Open*, 8(10):e019993, 2018. Published 2018 Oct 8.

[8] Daphne C. Watkins. Qualitative research: The importance of conducting research that doesn't "count". *Health Promotion Practice*, 13(2):153–158, 2012.

[9] Emily Namey, Greg Guest, Lucy Thairu, and Lauri Johnson. Data reduction techniques for large qualitative data sets. *Handbook for Team-Based Qualitative Research*, pages 137–161, 2008.

[10] Hamza Alshenqeeti. Interviewing as a data collection method: A critical review. *English Linguistics Research*, 3, 01 2014.

[11] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research*, 27(4):591–608, March 2017. Epub 2016 Sep 26.

[12] Seungjun Yi, Jaeyoung Lim, and Juyong Yoon. Protomed-llm: An automatic evaluation framework for large language models in medical protocol formulation, 2025.

[13] Seungjun Yi, Joakim Nguyen, Huimin Xu, Terence Lim, Andrew Well, Mia Markey, and Ying Ding. Auto-TA: Towards scalable automated thematic analysis (TA) via multi-agent large language models with reinforcement learning. In *ACL 2025 Student Research Workshop*, 2025.

[14] Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, Yan Leng, and Ying Ding. Tama: A human-ai collaborative thematic analysis framework using multi-agent llms for clinical interviews, 2025.

[15] Muhammad Zain Raza, Jiawei Xu, Terence Lim, Lily Boddy, Carlos M. Mery, Andrew Well, and Ying Ding. Llm-ta: An llm-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease, 2025.

[16] Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6456–6469, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[17] Insa Mannstadt, Susan M. Goodman, Mangala Rajan, Sarah R. Young, Fei Wang, Iris Navarro-Millán, and Bella Mehta. A novel approach for mixed-methods research using large language models: A report using patients' perspectives on barriers to arthroplasty. *ACR Open Rheumatology*, 6(6):375–379, 2024.

[18] M Deiner, V Honcharov, J Li, T Mackey, T Porco, and U Sarkar. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study. *JMIR Infodemiology*, 4:e59641, 2024.

[19] Janice Y. Kung. Elicit. *Journal of the Canadian Health Libraries Association*, 44(1):15–18, April 2023.

[20] C.M. Mery, A. Well, K. Taylor, K. Carberry, J. Colucci, C. Ulack, A. Zeiner, M. Mizrahi, E. Stewart, C. Dillingham, T. Cook, A. Hartounian, E. McCullum, J.T. Affolter, H. Van Diest, A. Lamari-Fisher, S. Chang, S. Wallace, E. Teisberg, and C.D. Jr Fraser. Examining the real-life journey of individuals and families affected by single-ventricle congenital heart disease. *Journal of the American Heart Association*, 12(5):e027556, March 2023. Epub 2023 Feb 21.

[21] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis, 2023.

[22] 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. Language models are few-shot learners, 2020.

[23] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report, 2024.

[24] Anthropic. Introducing claude. `https://www.anthropic.com/index/introducing-claude`, 2023.

[25] Google DeepMind. Gemini: A family of highly capable multimodal models. `https://deepmind.google/technologies/gemini/`, 2023.

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[27] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[28] DeepSeek AI. Deepseek llms. `https://github.com/deepseek-ai`, 2025.

[29] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen tau Yih, Luke Zettlemoyer, and Mike Lewis. Incoder: A generative model for code infilling and synthesis, 2023.

[30] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis, 2023.

[31] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.

[32] Mohammed-Khalil Ghali, Abdelrahman Farrag, Sarah Lam, and Daehan Won. Beyondwords is all you need: Agentic generative ai based social media themes extractor, 2025.

[33] Tingrui Qiao, Caroline Walker, Chris W Cunningham, and Yun Sing Koh. Thematic-LM: a LLM-based multi-agent system for large-scale thematic analysis. In *THE WEB CONFERENCE 2025*, 2025.

[34] JaMor Hairston, Ritvik Ranjan, Sahithi Lakamana, Anthony Spadaro, Selen Bozkurt, Jeanmarie Perrone, and Abeed Sarker. Automated thematic analyses using llms: Xylazine wound management social media chatter use case, 2025.

[35] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara J. Ericson, David Weintrop, and Tovi Grossman. How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment, 2023.

[36] Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. "don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education, 2025.

[37] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–20. ACM, May 2024.

[38] Walter S. Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255:108356, 2024.

[39] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 2004.

[40] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[41] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81. Barcelona, Spain, 2004.

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[43] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[44] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[46] Karisa Parkington, Bazen G. Teferra, Marianne Rouleau-Tang, Argyrios Perivolaris, Alice Rueda, Adam Dubrowski, Bill Kapralos, Reza Samavi, Andrew Greenshaw, Yanbo Zhang, Bo Cao, Yuqi Wu, Sirisha Rambhatla, Sridhar Krishnan, and Venkat Bhat. Human vs. llm-based thematic analysis for digital mental health research: Proof-of-concept comparative study, 2025.

[47] Zackary Okun Dunivin. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks, 2024.

[48] Tadeusz Calinski and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. Variance Ratio Criterion (VRC).

[49] Rewina Bedemariam, Natalie Perez, Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Elizabeth Conjar, Ikkei Itoku, David Theil, Aman Chadha, and Naumaan Nayyar. Potential and perils of large language models as judges of unstructured textual data, 2025.

[50] Mehmet Hamza Erol, Batu El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. Cost-of-pass: An economic framework for evaluating language models, 2025.

[51] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The economics of large language models: Token allocation, fine-tuning, and optimal pricing. Technical report, Cowles Foundation Discussion Paper No. 2425, Yale University; also available on arXiv, 2025. Online; February 11, 2025.

A Keyword Dictionary

Table 3: Keyword dictionary (`KEYWORDS_DICT`) used for thematic analysis literature search.

Dictionary Key	Keywords
`core_ta`	thematic analysis; qualitative thematic analysis; theme analysis; thematic coding; theme coding; identify themes; theme identification; Braun and Clarke; Braun & Clarke; six-phase framework; six phase framework
`inductive_ta`	inductive thematic analysis; data-driven thematic analysis; bottom-up thematic analysis; open coding (qualitative)
`deductive_ta`	deductive thematic analysis; theory-driven thematic analysis; top-down thematic analysis; a priori codes
`llm_generic`	LLM; large language model; transformer-based model; foundation model; prompting; chain-of-thought; self-reflection (LLM); RAG; few-shot; zero-shot
`healthcare`	clinical interview; patient interview; healthcare; medicine; clinical transcript; medical transcript; qualitative health research; nursing

B Model Release Milestones

We identify the following milestones in the development of LLMs; however, depending on which models were highlighted in the literature review, only a subset of these key milestones were selected for inclusion in our plot. InstructGPT was introduced on March 4, 2022, followed by PaLM on April 5, 2022, and BLOOM on November 9, 2022. ChatGPT launched on November 30, 2022. In 2023, LLaMA appeared on February 24, Claude 1 on March 14, GPT-4 on March 15, Claude 2 on July 11, and Llama 2 on July 18. The year 2024 brought Claude 3 on March 4, Llama 3 on April 18, GPT-4o on May 13, Claude 3.5 on June 20, and Gemini 2.0 on December 11. In 2025, GPT-4.1 was released on April 14, followed by Claude 4 on May 22, Claude 4.1 on August 5, and GPT-5 on August 7.

C Data Collection and Keyword Search Procedure

We accessed the official arXiv metadata via the Kaggle API (cutoff date: 2025–08–15). Each record included fields such as title, abstract, and subject categories.

- **Keyword construction:** We compiled keyword lists for thematic analysis (core TA, inductive TA, deductive TA), large language models (LLMs), and the healthcare domain (Appendix Table 3).
- **Search method:** A case-insensitive string search was applied to the title and abstract fields. A paper was counted in a category if it contained at least one keyword from that list. Multiple matches within a category were counted once. For example, papers with several TA terms were recorded once under `core_ta`.

- **Category intersections:** Papers with both TA and LLM terms were classified under `core_ta` + `llm_generic`. Those additionally containing healthcare terms were counted in the three-way intersection (`core_ta` + `llm_generic` + `healthcare`).

The resulting counts are summarized in Table 1.

D Survey Prompts

The prompt used for the literature search in *Elicit* [19] was as following: *How can evaluation frameworks for automated inductive thematic analysis balance validity, reliability, and interpretability when comparing direct LLM coding and hybrid human–AI workflows?*

E Supplementary Literature Search Strategies

To complement the arXiv-based search and acknowledge the limitation that some peer-reviewed papers may be missing from arXiv (as it is a non-archival venue), we implemented two additional strategies:

- **Elicit-assisted search:** We used Elicit [19], an AI-powered research assistant designed to streamline the literature review process. All references identified through Elicit were manually verified and reviewed by the authors. Specific examples of the survey prompts are provided in Appendix D.
- **Manual database search:** We conducted manual searches on major indexing platforms[6] using the query *"thematic analysis"* in combination with domain-specific terms. The exact keyword lists are reported in Appendix Table 3.

F Models Employed

OpenAI GPT family (41 papers, ∼74%): GPT-3.5 and GPT-4 variants (Turbo, 16k/32k, GPT-4o, GPT-4V). Backbone of most studies. **Claude (8 papers, ∼15%)**: Claude-instant, Claude-2/3.5/3.7 Sonnet. Used mainly for comparison. **LLaMA & open-source (7 papers, ∼13%)**: LLaMA-2/3, CodeLlama, Mistral, Mixtral, Gemma, DeepSeek. Benchmarked against GPT; valued for transparency. **Google Gemini (6 papers, ∼11%)**: Gemini Pro/Flash/Ultra/2.5 Preview. Typically baseline; mixed performance on inductive tasks. **Other (5 papers, ∼9%)**: InCoder, CodeGen, ProgGP, genre-CTRL, Whisper, Sentence-T5. Niche or auxiliary roles.

G Prompting Strategies

Zero-shot (35%): Most common; single- or multi-stage pipelines for code extraction, theme consolidation, and naming. Efficient but sensitive to phrasing, raising reproducibility concerns. **Few-shot (16%)**: Exemplars, templates, or pseudo-code scaffolds embedded in prompts; improved controllability but introduced bias in exemplar selection. **Chain-of-thought (13%)**: Encouraged intermediate reasoning steps; supported transparency, though outputs were often verbose or inconsistent. **Self-consistency/Reflexion (15%)**: Iterative or recursive re-prompting; aimed at stability but benefits over single-pass outputs were mixed. **Tool/agent-based (11%)**: Multi-agent roles (coder, evaluator, judge) or debate frameworks; innovative but resource-intensive, with variable gains. **Not applicable (11%)**: No prompting used (e.g., fine-tuned models, human-coded datasets, conceptual papers).

[6] Google Scholar, PubMed, Scopus, Web of Science