Auto-TA: Towards Scalable Automated Thematic Analysis (TA) via Multi-Agent Large Language Models with Reinforcement Learning

Seungjun Yi[1*], Joakim Nguyen[2], Huimin Xu[2], Terence Lim[3], Andrew Well[4], Mia Markey[1], Ying Ding[2,5]

[1] Department of Biomedical Engineering, University of Texas at Austin
[2] School of Information, University of Texas at Austin
[3]College of Natural Sciences, University of Texas at Austin
[4]Vanderbilt University Medical Center
[5]Dell Medical School, University of Texas at Austin
Corresponding Email: charlie.yi@utexas.edu, ying.ding@ischool.utexas.edu

Abstract

Congenital heart disease (CHD) presents complex, lifelong challenges often underrepresented in traditional clinical metrics. While unstructured narratives offer rich insights into patient and caregiver experiences, manual thematic analysis (TA) remains labor-intensive and unscalable. We propose a fully automated large language model (LLM) pipeline that performs end-to-end TA on clinical narratives which eliminates the need for manual coding or full transcript review. Our system employs a novel multi-agent framework, where specialized LLM agents assume roles to enhance theme quality and alignment with human analysis. To further improve thematic relevance, we optionally integrate reinforcement learning from human feedback (RLHF). This supports scalable, patient-centered analysis of large qualitative datasets and allows LLMs to be fine-tuned for specific clinical contexts.

1 Introduction

Congenital heart disease (CHD) affects approximately 1% of live births, with around 40,000 cases annually in the U.S., and over 12 million people living with CHD worldwide (Centers for Disease Control and Prevention, 2022; Liu et al., 2019). The lifelong journey of individuals and families affected by complex CHD brings emotional, logistical, and structural challenges that are often overlooked in traditional clinical metrics. Recent qualitative studies highlight the importance of mapping lived experiences and identifying outcomes that patients and caregivers themselves consider meaningful, particularly those related to capability, comfort, and calm (Mery et al., 2023).

Despite growing recognition of the importance of lived experience, post-discharge care still relies heavily on two primary forms of data: structured patient-reported outcomes (PROs) and unstructured narratives. Structured PROs, typically gathered through standardized processes, are essential for monitoring health status but often fail to capture the complexity of patient and caregiver needs (Valderas et al., 2008). Unstructured data, such as interviews and open-text survey responses, offer richer insights into emotional, social, and logistical experiences (Greenhalgh et al., 2019). However, these narratives remain underutilized due to the substantial time, cost, and expertise required for manual thematic analysis (TA) (Mery et al., 2023). Traditional TA is foundational to qualitative research but remains labor-intensive, time-consuming, and prone to inconsistency (Braun and Clarke, 2006; Nowell et al., 2017). Coding[1] just 10–15 interviews can take 40–60 hours, often requiring expert reviewers and therefore, limiting scalability in clinical settings (Watkins, 2017; Namey et al., 2008).

These limitations have driven recent efforts to automate TA using machine learning and large language models (LLMs). Hybrid frameworks now combine human judgment with automated components, enabling faster analysis while retaining interpretability (Dai et al., 2023b; Xu et al., 2025c). However, most still rely on human-in-the-loop workflows that require full transcript review limiting scalability. This raises a fundamental question: *what's the point of introducing LLMs if humans still need to go through the entire transcript?*

To address these limitations, we propose Auto-TA (Figure 1), a fully automated LLM pipeline that performs end-to-end thematic analysis on unstructured clinical narratives without requiring manual coding or full transcript review, with the ultimate goal of identifying meaningful outcomes[1] and gaps in care[1]. Unlike prior hybrid approaches that rely on human-in-the-loop workflows, our pipeline is designed to autonomously generate codes[1], extract themes[1], and evaluate alignment. To further enhance theme generation quality, we incorporate a multi-agent system where specialized LLM agents collaborate to improve

alignment with human-generated themes. Each agent takes on a distinct role—such as coder (generation agent), reviewer (feedback agent) — facilitating iterative evaluation and refinement that captures diverse perspectives and overcomes challenges related to semantic nuance. As an extension, we explore the integration of optional human feedback to refine theme generation. By leveraging reinforcement learning from human feedback (RLHF), we aim to improve thematic relevance, consistency, and alignment with real-world needs in clinical contexts, while preserving scalability. Drawing on expert ratings or predefined metrics such as coherence and distinctiveness, RLHF enables our system to iteratively optimize outputs beyond static prompting. This approach supports patient-centered research and informs clinical decision-making by identifying meaningful outcomes[1] and gaps in care[1] within large-scale unstructured qualitative data.

In summary, our contribution is as following:

- Automated Thematic Analysis: Proposed `Auto-TA` (Figure 1), an end-to-end LLM pipeline that performs TA on unstructured clinical narratives without requiring manual coding or transcript review.
- Multi-Agent Theme Refinement: Introduced a multi-agent LLM system with specialized roles to improve theme quality and alignment with human analysis.
- Scalable RLHF Integration: Optionally incorporates reinforcement learning from human feedback (RLHF) to improve thematic relevance and alignment with patient-centered outcomes, while preserving scalability.

2 Related Work

Automated Thematic Analysis (TA) with LLMs Efforts to automate TA with LLMs have progressed from hybrid, human-in-the-loop methods to autonomous multi-agent systems. *LLM-in-the-loop* approach (Dai et al., 2023a) accelerated coding by 60%, while single-agent models replicated high-level themes but lacked nuance (Paoli, 2024). Bias and hallucination risks remain a concern, especially in health and policy contexts (Lee et al., 2024; Khan et al., 2024). *LLM-TA* (Raza et al., 2025) improved lexical alignment with expert CHD themes but required manual transcript review. *TAMA* (Xu et al., 2025a) introduced a coder–reviewer–refiner agent framework, boosting HIT rate and reducing analysis time by 99%. Domain-specific pipelines validate efficiency gains but highlight ongoing challenges in bias and privacy (Qiao et al., 2025).

Thematic-LM performed multi-agent TA on Reddit threads related to climate change, highlighting opportunities for automation while raising open questions around semantic evaluation (Qiao et al., 2025).

Multi-Agent LLMs Multi-agent large language model (LLM) systems were first formalized by *CAMEL*, which introduced two role-playing GPT-3.5 agents that interact to decompose and solve tasks (Li et al., 2023). Such systems consist of two or more LLM-powered agents that communicate to collaborate or compete on a given problem. Building production pipelines has been simplified by *AutoGen*, a lightweight framework for spawning role-conditioned agents that converse and invoke external APIs (Wu et al., 2023). Systematic benchmarks such as *AgentBench* and *MultiAgentBench* confirm that coordinated or competitive teams consistently outperform single-model baselines across web navigation, coding, negotiation, and other tasks (Liu et al., 2023; Zhu et al., 2025). In evaluation settings, *ChatEval* shows that a debating panel of agents yields markedly more reliable text judgments than a solitary judge (Chan et al., 2024), while *COPPER* augments collaboration with counterfactual self-reflection fine-tuned by PPO to tackle credit-assignment issues (Bo et al., 2024). Overall, these results highlight role specialization, debate, and reflection as key inductive biases for scaling LLM capabilities, while domain-specific applications (e.g., qualitative health analytics) and group-level reward modeling remain open research challenges.

Reinforcement Learning (RL) is a computational framework in which an agent learns, by trial and error, to choose actions that maximize long-run cumulative reward in an environment (Sutton and Barto, 1998). Within the LLM domain, the three-stage RL-from-human-feedback (RLHF) recipe of supervised fine-tuning, reward-model training, and policy optimization with Proximal Policy Optimization (PPO) has become the standard since *InstructGPT* demonstrated large gains in helpfulness and compliance (Ouyang et al., 2022; Schulman et al., 2017). Follow-up work such as *Helpful and Harmless* (Bai et al., 2022) scales the pipeline and explores reward–KL trade-offs, while RLHF for specific text tasks, such as summary generation

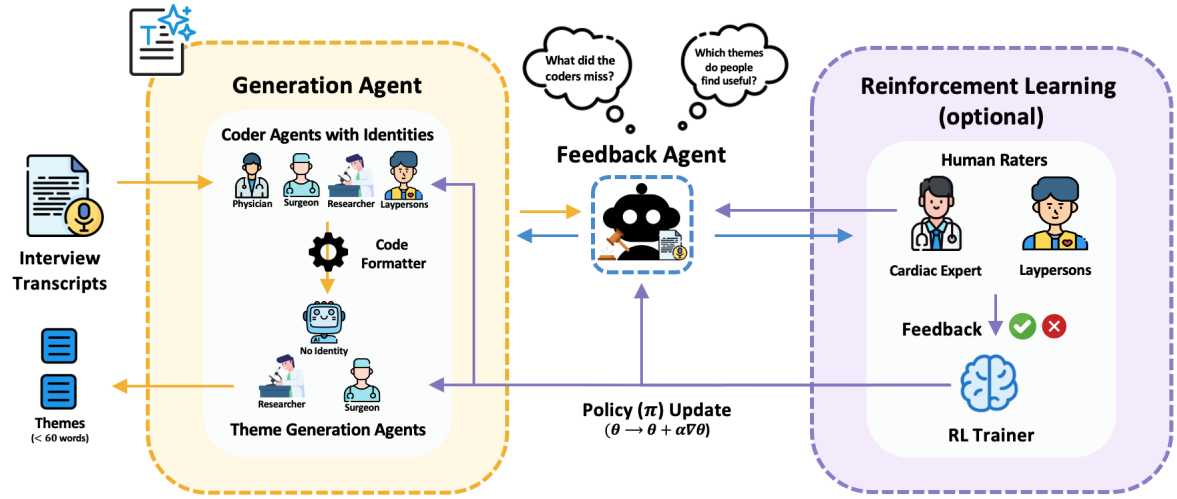


Figure 1: **The Auto-TA Framework.** An end-to-end multi-agent LLM pipeline for thematic analysis of unstructured clinical narratives. Coder agents with diverse identities generate preliminary codes from interview transcripts, which are processed into themes by downstream agents. A feedback agent evaluates outputs and provides iterative refinement. Optional integration of reinforcement learning from human feedback (RLHF) allows the system to optimize thematic relevance and alignment with patient-centered outcomes. The implementation of RLHF is currently in progress.

(Stiennon et al., 2020), confirms that preference-based rewards outperform maximum-likelihood fine-tuning. Recent variants introduce hierarchical or self-refinement rewards, but consensus is still emerging around stable optimization and domain-specific signal design, leaving room for customized reward modeling in qualitative health analytics.

Definitions and Abbreviations Table 1 defines key terms and medical abbreviations used throughout the paper.

3 Method

3.1 The *Auto-TA* Framework

Figure 1 presents an overview of the Auto-TA framework. It comprises three interacting group of agents: **Generation Agents**, **Feedback Agents**, and **Reinforcement Learning**, which operate in coordination to perform automated TA. The full process is summarized below:

Generation Agents We instantiate two categories of generation agents:

a) **Coder Agents with Identities.** A pool of k=4 role-conditioned GPT-4o[1] agents[2] are each given the full interview transcript as input. If the transcript exceeds the model's input batch limit, it is divided into contiguous chunks $\mathbf{x} \in \mathbb{R}^{\leq 1500}$, and each agent processes the same chunk independently. Each agent then emits a set of initial codes $\mathcal{C}_r = \{c_{r,1}, \ldots, c_{r,n_r}\}$, where $r \in R$ denotes the role-specific identity of the agent and n_r is the number of codes produced by role r. Example role prompts are provided in Appendix F, where all identities follow a similar structure.

b) **Theme-Generation Agents.** The codes from all roles are merged and passed to a secondary set of agents that cluster semantically similar codes and generate preliminary themes $\Theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_m^{(0)}\}$. Each theme is concise, often within 60 words. Some agents operate without identity conditioning, while others may retain role-specific perspectives to support diverse theme formulation. In subsequent steps, feedback from the feedback agent is used to iteratively refine and improve the generated themes.

Together these agents reduce a transcript to a concise thematic representation without human review, completing Steps 1–3 of reflexive TA (Braun

[1] All prompts use temperature = 0 for determinism.

[2] *Physician, Surgeon, Researcher, Layperson* were assigned per Cardiac Expert guidance to simulate prior TA. Future work should consider alternative identities.

Term	Definition
Code	A discrete analytical unit that captures a key pattern within the data, generated directly from the dataset and retaining its interpretive significance without being reducible to smaller meaningful components. *Coding* refers to the process of generating these codes from the original text. A *coder* is an individual who performs *coding* by identifying and labeling meaningful segments of data. Examples are in Appendix G.2.
Themes	Inductively derived patterns or topics of meaning that emerged from participants' narratives about their lived experiences, identified through iterative coding and consensus across multiple coders. These themes were then categorized into meaningful outcomes and gaps in care, with outcomes further grouped using the capability, comfort, and calm framework.
Meaningful Outcomes	Patient- and family-prioritized goals that reflect capability (*doing the things in life you want to*), comfort (*experience of physical/emotional pain/distress*), and calm (*experiencing health care with the least impact on daily life*).
Gaps in Care	Gaps in care refer to unmet needs or breakdowns in healthcare delivery that hinder patients and families from receiving consistent, compassionate, and coordinated support across the care journey.

Table 1: **Key Definitions and Abbreviations** used in this paper, consistent with Mery et al. (2023). Additional abbreviations are listed in Appendix B.

and Clarke, 2006).

Feedback Agents The feedback agent $\mathcal{F}$ acts as an autonomous critic. Given $\Theta^{(t)}$ at refinement round t, it produces:

- **Evaluation Scores**: Obtain $\mathbf{s}^{(t)} = \langle \mathcal{C}, \mathcal{D}, \mathcal{T} \rangle$ on the trustworthiness dimensions defined in Section 4.
- **Edit Themes**: Suggests specific edits to themes, such as ADD (adding), SPLIT (splitting), COMBINE (combining), or DELETE (deleting), each targeted to a particular theme.

If RL is disabled, the proposals are applied heuristically: themes with $\mathcal{C} < 0.7$ trigger ADD; $D_L < 0.20$ triggers COMBINE, etc. The resulting $\Theta^{(t+1)}$ is re-scored until $\|\mathbf{s}^{(t+1)} - \mathbf{s}^{(t)}\|_1 < 0.05$ or a maximum of three iterations[3].

[3]The specific thresholds and number of iterations are subject to change. In practice, these values are tuned based on empirical validation and expert feedback to meet the goals of inductive TA. For example, in the AAOCA dataset, cardiac experts provide feedback to ensure that the thresholds and parameters align with clinically meaningful interpretations.

Reinforcement Learning When expert raters are available, `Auto-TA` can optionally switch to a RLHF loop:

a) Human raters assign binary rewards ($r \in 0, 1$) to each theme set, where 1 indicates meeting quality standards and 0 indicates otherwise. While we use *Coverage, Actionability, Distinctiveness, and Relevance* as evaluation criteria adopted from Xu et al. (2025a), users may define their own criteria based on specific needs.

b) A reward model R_ϕ is updated via MSE loss on the rated batches.

c) The policy parameters θ of the Theme-Generation agent are optimized with Proximal Policy Optimisation (PPO):

$$\theta \leftarrow \theta + \alpha \nabla_\theta \, \mathbb{E}_{\pi_\theta}\big[\, R_\phi(\Theta^{(t)}) - \beta \mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{SFT}}\big)\big]$$

Here, π_θ denotes the current policy of the theme-generation agent, and $R_\phi(\Theta^{(t)})$ is the reward assigned by a model evaluating the quality of the generated theme set $\Theta^{(t)}$. The KL divergence term $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{SFT}})$ measures deviation from the base supervised model π_{SFT}, which acts as a regularizer to prevent the policy from drifting too far. The coefficient β controls the strength of this regularization, balancing reward maximization and policy alignment. The update step uses a learning rate α to scale the gradient ∇_θ, which indicates the direction to adjust parameters θ to increase the expected objective. The expectation $\mathbb{E}_{\pi_\theta}$ is taken over actions sampled from the current policy, reflecting average performance under the agent's behavior.

The feedback agent is retained as a critic during training; its scores are concatenated to the reward model input, enabling reward shaping without additional human cost. If no human feedback is supplied, `Auto-TA` degrades to the heuristic loop above.

In summary, the architecture automates TA in under 10 minutes per 10k-word transcript. The optional RLHF path enables the system to adapt to researcher preferences over time without requiring full transcript re-reads. The resulting themes are categorized into outcomes and care gaps, with outcomes further organized using the capability, comfort, and calm framework (Mery et al., 2023).

End-to-End Workflow of Auto-TA

Step 1: Transcript Processing Each transcript is fed to k=4 role-conditioned coder agents. If the transcript exceeds the input limit, it is divided into chunks and broadcast to all agents.

Step 2: Code Aggregation and Theme Generation Codes from all roles are merged and clustered by theme-generation agents to produce preliminary themes $\Theta^{(0)}$.

Step 3: Feedback Evaluation A feedback agent critiques the initial themes and produces quality scores $\mathbf{s}^{(0)}$.

Step 4: Theme Refinement Themes are iteratively improved via heuristic edits or PPO updates, producing $\Theta^{(1)}$.

Step 5: Repeat Steps 3–4 until convergence or a maximum of $t_{\max} = 5$ iterations.

Step 6: Output Final theme set $\Theta^{\star} = \Theta^{(t_{\text{conv}})}$ and its associated audit trail.

3.2 Dataset

This study analyzes a targeted subset of transcripts from a broader qualitative project on the lifelong experiences of individuals and families affected by single-ventricle congenital heart disease (SV-CHD) (Mery et al., 2023). We use de-identified transcripts from nine moderated focus groups involving 42 parents of children diagnosed with Anomalous Aortic Origin of a Coronary Artery (AAOCA). Each 90-minute session captures narrative-driven discussions on diagnosis, care pathways, emotional burdens, and decision-making. The transcripts average 10,987 words (SD: 1,537; median: 11,457).

We also incorporate human-generated themes from Mery et al. (2023) as references to compute traditional metrics against LLM-generated themes. For consistent referencing, each quote was manually assigned Quote IDs, a unique identifier. Further corpus details are in Appendix A.1.

4 Evaluation

4.1 Evaluation Criteria

We adopt the four criteria of *credibility*, *confirmability*, *dependability*, and *transferability* from the trustworthiness framework in qualitative research introduced by Lincoln and Guba (1985) and later applied to TA by Nowell et al. (2017) and Korstjens and Moser (2018). These criteria were also recently employed in TA using LLMs (Qiao et al., 2025).

Credibility and Confirmability ($\mathcal{C}$) We evaluate credibility and confirmability jointly by assessing the degree to which the generated themes are grounded in the original data. Specifically, we retrieve the associated segments using Quote IDs and task an evaluator agent with determining whether each theme is consistent with its supporting quotes. This consistency check identifies cases where the theme reflects the quoted content accurately versus instances of hallucination or bias. Let Q denote the set of all coded quotes, and $Q_{\text{ref}} \subseteq Q$ be the subset of quotes that are used to generate at least one theme. We define $\mathcal{C}$ as:

$$\mathcal{C} = \frac{|Q_{\text{ref}}|}{|Q|} \times 100$$

A higher value of $\mathcal{C}$ indicates stronger alignment between themes and the underlying data, reflecting both accurate representation (*credibility*) and traceable justification (*confirmability*) in the analysis. $\mathcal{C}$ can be enhanced through *prolonged engagement*, *triangulation* (e.g., multiple coders or data sources), and *member checking*, thereby maximizing the overlap between participant intent and the researcher's interpretations (Lincoln and Guba, 1985; Nowell et al., 2017; Korstjens and Moser, 2018).

Dependability ($\mathcal{D}$) Dependability reflects the stability of theme generation across independent runs. We compute lexical overlap using bidirectional ROUGE scores (Lin, 2004). Specifically, $\mathcal{R}_1$ and $\mathcal{R}_2$ represent the ROUGE-1 (unigram) and ROUGE-2 (bigram) overlap scores, respectively. Given two sets of themes A and B generated from independent runs, we define:

$$\mathcal{R}_1^{A\to B} = \frac{|\text{unigrams}(A) \cap \text{unigrams}(B)|}{|\text{unigrams}(A)|} \quad (1)$$

$$\mathcal{R}_2^{A\to B} = \frac{|\text{bigrams}(A) \cap \text{bigrams}(B)|}{|\text{bigrams}(A)|} \quad (2)$$

$$\mathcal{R}_1 = \frac{1}{2}\left(\mathcal{R}_1^{A\to B} + \mathcal{R}_1^{B\to A}\right) \quad (3)$$

$$\mathcal{R}_2 = \frac{1}{2}\left(\mathcal{R}_2^{A\to B} + \mathcal{R}_2^{B\to A}\right) \quad (4)$$

$$\mathcal{D} = \frac{1}{2}\left(\mathcal{R}_1 + \mathcal{R}_2\right) \quad (5)$$

We evaluate $\mathcal{D}$ across 10 independent generations per transcript and report results aggregated across all 9 transcripts. Higher values of $\mathcal{D}$ indicate stronger inter-run consistency. To further support dependability, we maintain a methodological log and employing coding strategies (Korstjens and Moser, 2018).

Transferability ($\mathcal{T}$) Transferability concerns the extent to which themes Θ, generated from a dataset D, can be meaningfully applied to a new but contextually similar corpus D'. In qualitative research, high $\mathcal{T}$ is traditionally supported through *thick description*—that is, detailed accounts of participants, settings, and analytical choices that enable readers to assess contextual relevance (Lincoln and Guba, 1985; Nowell et al., 2017).

In our framework, we use transferability by dividing the dataset into a training set and a validation set. Specifically, we use 7 transcripts to generate Θ_{train} and 2 transcripts to generate Θ_{val}, then compute bidirectional ROUGE to assess overlap:

$$\mathcal{R}'_1 = \frac{1}{2}\left(\mathcal{R}_1^{\Theta_{\text{train}}\to\Theta_{\text{val}}} + \mathcal{R}_1^{\Theta_{\text{val}}\to\Theta_{\text{train}}}\right) \quad (6)$$

$$\mathcal{R}'_2 = \frac{1}{2}\left(\mathcal{R}_2^{\Theta_{\text{train}}\to\Theta_{\text{val}}} + \mathcal{R}_2^{\Theta_{\text{val}}\to\Theta_{\text{train}}}\right) \quad (7)$$

$$\mathcal{T} = \frac{1}{2}\left(\mathcal{R}'_1 + \mathcal{R}'_2\right) \quad (8)$$

To robustly estimate transferability, we compute $\mathcal{T}$ across all $\binom{9}{2} = 36$ possible 7-train / 2-validation transcript splits. For each split, we perform independent thematic analysis on both subsets and evaluate bidirectional ROUGE overlap between the resulting theme sets. We report the mean and standard deviation of $\mathcal{T}$ across these 36 combinations to capture overall generalizability and variation across different training-validation configurations. A higher $\mathcal{T}$ indicates that themes generated from one subset of the corpus generalize well to others, suggesting good conceptual transfer across the dataset.

4.2 Other Metrics

To evaluate the alignment between LLM-generated and human-generated themes, we explore alternative metrics beyond those used in prior work (Raza et al., 2025; Xu et al., 2025b). Jaccard Similarity (J) and HIT Rate (R), commonly used in earlier studies, rely on surface-level word overlap and binary thresholding, which makes them less effective at capturing paraphrastic variation and deeper semantic relationships. This motivates the use of more comprehensive and semantically informed measures of thematic alignment.

Cosine Similarity (C_{bi}) To quantify semantic alignment at the individual theme level, we calculate the cosine similarity between each pair (t_i, ℓ_j) using embeddings from a sentence-level transformer model (`all-mpnet-base-v2`). Bidirectional Cosine similarity(C_{bi}) is defined as

$$C(t_i, \ell_j) = \frac{\mathbf{v}_{t_i} \cdot \mathbf{v}_{\ell_j}}{\|\mathbf{v}_{t_i}\|\|\mathbf{v}_{\ell_j}\|},$$

where $\mathbf{v}_{t_i}$ and $\mathbf{v}_{\ell_j}$ are the embedding vectors of t_i and ℓ_j, respectively. To summarize alignment, we first report the *unidirectional* mean maximum similarity from human to LLM themes:

$$C_{T\to L} = \frac{1}{n}\sum_{i=1}^{n} \max_j \text{cosine}(t_i, \ell_j),$$

which reflects how well each human theme is captured by its closest LLM-generated counterpart.

To account for potential asymmetry, we also compute the reverse direction, measuring how well each LLM theme aligns with the closest human theme:

$$C_{L\to T} = \frac{1}{m}\sum_{j=1}^{m} \max_i \text{cosine}(\ell_j, t_i).$$

The *bidirectional cosine alignment score* is then defined as the average of both directions:

$$C_{\text{bi}} = \frac{1}{2}\left(C_{T\to L} + C_{L\to T}\right),$$

providing a balanced measure of mutual semantic alignment between theme sets.

Levenshtein Distance (D_L) We compute the average maximum normalized Levenshtein similarity between each human theme and LLM theme. Full formulation is provided in Appendix C.

BLEU Score (B) We use BLEU (Papineni et al., 2002) to measure n-gram overlap (up to 4-grams with brevity penalty) between human and LLM themes, and report the maximum score per human theme and averaging over all.

5 Results

Impact of Agent Identities Our results suggest that assigning domain-specific identities to agents lead to substantial improvements in credibility ($\mathcal{C}$), with the Cardiac Surgeon and Qualitative Researcher identities achieving the highest scores (Table 2). All identity-augmented agents outperformed the baseline in $\mathcal{C}$, with gains ranging from +11.54 to +16.28. Dependability ($\mathcal{D}$) is largely unaffected or slightly reduced, likely due to variability from non-uniform agent behavior. Transferability ($\mathcal{T}$) improves with identity augmentation, with the Medical Doctor and Psychologist agents showing the highest gains ($+0.026$, $+0.027$), suggesting expert-informed themes generalize better. $\mathcal{T}$ shows minimal change across all 36 combinations with low standard deviation, indicating consistent performance and supporting the robustness of `Auto-TA`.

Theme Alignment with Human Ground Truth We evaluated the alignment between LLM-generated and human-generated themes using C_{bi}, D_L, and B (Table 3). Higher alignment for the Cardiac Surgeon agent likely reflects the presence of a cardiac expert in the human analyst group. Alignment appears influenced by annotator expertise; different expert compositions (e.g., more qualitative researchers) could shift alignment patterns. Even when surface-level overlap is low, identity-augmented agents often capture subtle or overlooked aspects of the data not fully represented in the human-generated themes. This suggests that identity conditioning can expand the thematic space in meaningful and complementary ways.

This observation also points to the limitations of traditional evaluation metrics that rely on surface-level comparison between human and LLM-generated themes. As shown in Table 4, low scores on metrics such as B do not necessarily imply incorrect themes. Rather, they underscore the need for evaluation criteria better suited to the interpretive nature of qualitative research, where multiple valid versions of TA can coexist.

Our preliminary results showed no significant changes using the four criteria from previous studies, including distinctiveness and actionability (Figure 5). This may be due to the use of a coarse 5-point integer scoring system within a form-filling paradigm using LLM evaluators, which may lack the resolution to capture micro-scale improvements. To address this, we propose adopting a scalar scoring system (e.g., continuous values between 1.00 and 5.00) to better reflect incremental changes in theme quality during iterative refinement.

6 Limitations

A key limitation of our approach lies in the assumption that alignment with human-generated themes is a sufficient indicator of thematic quality. In practice, there may be multiple valid sets of *right* themes for the same transcript, depending on the perspective and interpretive lens of the analyst. As such, high alignment with a human-coded reference set does not necessarily imply better or more meaningful TA. This observation aligns with findings from prior work. For instance, a cardiac expert cited in Mery et al. (2023) noted that approximately two-thirds of themes tend to be straightforward and expected, regardless of the analyst's background. However, the remaining one-third are more variable and dependent on individual expertise and interpretation. Beyond this conceptual limitation, several practical constraints remain. The framework has not yet been tested across different domains, leaving its ability to generalize to other clinical or non-clinical settings uncertain. In its current form, the architecture does not support interaction among agents, which limits opportunities for collaborative reasoning or negotiated theme development. Evaluation is also limited to a single comparison with human-coded themes, offering little insight into the stability of outputs across multiple runs. Moreover, the system shows notable sensitivity to prompt wording, where minor variations can lead to substantially different thematic results and may raise concerns about reproducibility in real-world applications.

7 Conclusion and Future Work

We presented `Auto-TA`, a fully automated, multi-agent LLM framework for end-to-end thematic analysis of unstructured clinical narratives. Role-

Agent Identity	$\mathcal{C}$	$\Delta\mathcal{C}$	$\mathcal{D}$	$\Delta\mathcal{D}$	$\mathcal{T}$	$\Delta\mathcal{T}$
NO IDENTITIES (BASELINE)	82.13 ± 18.96	–	0.400 ± 0.017	–	0.308 ± 0.018	–
CARDIAC SURGEON	**98.41 ± 4.76**	**+16.28**	0.395 ± 0.019	-0.005	0.318 ± 0.027	+0.010
QUALITATIVE RESEARCHER	97.56 ± 3.34	+15.43	0.397 ± 0.014	**-0.003**	0.324 ± 0.023	+0.016
MEDICAL DOCTOR	96.83 ± 8.98	+14.70	0.389 ± 0.025	-0.011	0.334 ± 0.007	+0.026
PSYCHOLOGIST	93.67 ± 2.35	+11.54	0.359 ± 0.018	-0.041	0.325 ± 0.015	**+0.027**

Table 2: **Performance of Identity-Augmented Agents** Evaluation results across core metrics: credibility ($\mathcal{C}$), dependability ($\mathcal{D}$), and transferability ($\mathcal{T}$). Values denote the mean and standard deviation computed over nine transcripts. Bolded values indicate the best performance, and underlined values denote the second-best. Δ columns indicate improvements over the baseline. Higher values are better for $\mathcal{C}$, $\mathcal{D}$, and $\mathcal{T}$. Definitions for each evaluation metric are in Section 4.

Agent Identity	Cosine (C_{bi})	ΔC	Levenshtein (D_L)	ΔD_L	BLEU (B)	ΔB
NO IDENTITIES (BASELINE)	0.132 ± 0.027	–	0.301 ± 0.027	–	0.019 ± 0.008	–
CARDIAC SURGEON	0.115 ± 0.053	-0.017	0.259 ± 0.089	-0.042	0.020 ± 0.009	$+0.001$
QUALITATIVE RESEARCHER	0.107 ± 0.046	-0.025	0.252 ± 0.082	-0.049	0.014 ± 0.007	-0.005
MEDICAL DOCTOR	0.112 ± 0.018	-0.020	0.287 ± 0.025	-0.014	0.018 ± 0.006	-0.001
PSYCHOLOGIST	0.121 ± 0.029	-0.011	0.282 ± 0.021	-0.019	0.020 ± 0.006	$+0.001$

Table 3: **Semantic and Lexical Alignment Metrics.** Mean ± standard deviation scores across three metrics used to evaluate alignment between LLM-generated and human-generated themes: bidirectional Cosine Similarity (C_{bi}), Levenshtein Distance (D_L), and BLEU Score (B). Δ columns represent absolute changes from the baseline (No Identities). Higher values indicate better alignment, except for D_L, where lower values indicate better alignment.

#	**Human Theme**	**Closest LLM Themes (`iteration=5`)**
1	Clarity of potential risks and outcomes	Desiring Comprehensive Data on Long-Term Outcomes Seeking Clarity and Understanding about My Child's Condition
2	Freedom from hypervigilance related to the condition	Feeling Overwhelmed by Emotional Uncertainty Living with Constant Anxiety about Child's Health
3	The diagnosis given in a compassionate and empathic way	Experiencing Relief from Diagnosis and Support Feeling Overwhelmed by Medical Decisions
4	A sense of control over the future	Balancing Relief and Anxiety about the Future Seeking Reassurance and Clear Communication from Medical Professionals
5	Being heard and taken seriously by clinicians	Seeking Reassurance and Clear Communication from Medical Professionals Feeling Isolated During the Medical Journey
6	Individualized support for management decision-making	Advocating for Child's Medical Needs Navigating Surgical Decision Anxiety
7	Receiving support from others	Finding Strength in Family and Community Support Finding Joy in Family Connections
8	Being appropriately informed	Seeking Reassurance and Clear Communication from Medical Professionals Desiring Clear Communication from Medical Professionals
9	Partnership with the care team	Seeking Reassurance and Clear Communication from Medical Professionals Advocating for Child's Medical Needs
10	Feeling that my child is safe	Coping with Health Crisis Trauma Feeling Overwhelmed by Medical Decisions
11	Not feeling responsible for the diagnosis and its timing	Struggling with Feelings of Guilt and Responsibility Desiring Proactive Healthcare Measures to Prevent Crises
12	Appropriately coping with stress, anxiety and depression	Feeling Overwhelmed by Emotional Uncertainty Managing Ongoing Anxiety about My Child's Health

Table 4: **Examples of semantic alignment between human and LLM-generated themes.** Each human-generated theme is matched with the two LLM-generated themes exhibiting the highest cosine similarity scores, based on the similarity matrix heatmap (Figure 4b) after five refinement iterations incorporating feedback from the feedback agent. These LLM-generated themes are drawn from agents with diverse identities in the `Auto-TA` pipeline. Although traditional alignment scores are modest, the most similar LLM themes often reflect comparable underlying meanings, and in some cases, elaborate or extend the human-coded interpretations.

conditioned agents enhanced alignment with human-coded themes, particularly when identities matched annotator expertise. Despite low traditional metric scores, identity-augmented agents captured valid insights and expanded the thematic space. Auto-TA demonstrates potential for scalable, expert-informed qualitative analysis, supporting future extensions through domain adaptation, reinforcement learning, and agent collaboration.

Future work should explore evaluation methods that go beyond surface-level alignment, focusing on the distinctiveness, utility, and interpretive depth of LLM-generated themes in specific clinical or research contexts. Bridging machine efficiency and human interpretability will likely require iterative collaboration with domain experts. Incorporating expert feedback into a reinforcement learning framework could enable adaptive refinement based on human judgment. Expanding the framework to new domains will help assess its generalizability, while structured dialogue or negotiation among agents may better mirror human qualitative analysis. An identity generation agent that selects optimal personas based on dataset context could further improve relevance. Finally, examining reproducibility and sensitivity to prompt variation is essential for real-world reliability.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*.

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

Centers for Disease Control and Prevention. 2022. Data and statistics on congenital heart defects. https://www.cdc.gov/heart-defects/data/index.html. Accessed: 2025-04-30.

Chi-Min Chan, Weize Chen, Yusheng Su, Yuchen Xiang, Xuancheng Ren, and Heyan Huang. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023a. Llm-in-the-loop: Leveraging large language model for thematic analysis. *Preprint*, arXiv:2310.15100.

Xinyao Dai and 1 others. 2023b. Hybrid human-ai thematic analysis with large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Trisha Greenhalgh, Joe Wherton, Chrysanthi Papoutsi, and 1 others. 2019. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, 21(11):e13170.

Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. 2024. Automating thematic analysis: How LLMs analyse controversial topics. *arXiv preprint arXiv:2405.06919*.

Irene Korstjens and Albine Moser. 2018. Series: Practical guidance to qualitative research. part 4: Trustworthiness and publishing. *European Journal of General Practice*, 24(1):120–124.

V. Vien Lee, Stephanie C. C. van der Lubbe, Lay Hoon Goh, and Jose Maria Valderas. 2024. Harnessing chatgpt for thematic analysis: Are we ready? *Journal of Medical Internet Research*, 26:e54974.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Yvonna S. Lincoln and Egon G. Guba. 1985. *Naturalistic Inquiry*. Sage Publications.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2023. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Yanyan Liu, Yunning Chen, Pan Yin, and 1 others. 2019. Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet Child & Adolescent Health*, 3(12):918–927.

Carlos M Mery, Andrew Well, Kate Taylor, Kathleen Carberry, José Colucci, Christopher Ulack, Adam Zeiner, Michelle Mizrahi, Eileen Stewart, Christine Dillingham, Taylor Cook, Arotin Hartounian, Elizabeth McCullum, Jeremy T Affolter, Heather Van Diest, Alexandra Lamari-Fisher, Stacey Chang, Scott Wallace, Elizabeth Teisberg, and Charles D Fraser Jr. 2023. Examining the real-life journey of individuals and families affected by single-ventricle congenital heart disease. *Journal of the American Heart Association*, 12(4):e027556.

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

Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1):1609406917733847.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Stefano De Paoli. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model. *Social Science Computer Review*, 42(4):997–1019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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

Muhammad Zain Raza, Jiawei Xu, Terence Lim, Lily Boddy, Carlos M. Mery, Andrew Well, and Ying Ding. 2025. LLM-TA: An LLM-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease. *arXiv preprint arXiv:2502.01620*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Jose M Valderas, Anna Kotzeva, Mireia Espallargues, and 1 others. 2008. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Quality of Life Research*, 17(9):1125–1135.

Daphne C Watkins. 2017. Qualitative research: The importance of conducting research that doesn't "count". *Health Promotion Practice*, 18(5):601–602.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, and Andrew Well et al. 2025a. TAMA: A human–ai collaborative thematic analysis framework using multi-agent LLMs for clinical interviews. *arXiv preprint arXiv:2503.20666*.

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. 2025b. Tama: A human-ai collaborative thematic analysis framework using multi-agent llms for clinical interviews. *Preprint*, arXiv:2503.20666.

Lina Xu and 1 others. 2025c. Agentta: Multi-agent thematic analysis with role-specialized llms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *Preprint*, arXiv:2503.01935.

Appendix

A Dataset Details

A.1 Original Corpus Composition and Participant Summary

Between February and September 2020, transcripts were collected from 19 moderated focus groups (Experience Groups, 90 minutes, 3–7 participants), 35 semi-structured 1:1 interviews (60 minutes), and 4 co-design workshops aimed at validating emerging journey maps. The full composition of the parent corpus is shown in Table 5.

Table 5: Parent corpus composition

	# Sessions	Participants	Approx. Words
EG Sessions	19	134	∼260 k
1:1 Interviews	35	56†	∼210 k
Workshops	4	58	∼50 k
Total	58	170	∼520 k

†35 family participants + 21 stakeholders

Of the 96 survey respondents, most were parents (n=52) or patients (n=29), with smaller numbers of siblings (n=9), partners (n=4), and fetal case participants (n=4). Females comprised the majority across groups, especially among parents (77%) and patients (55%). Most respondents identified as White (81% of parents, 78% of patients), with smaller proportions identifying as Hispanic/Latino or Black.

A.2 Quote Identification and Traceability

Each statement in the transcripts is tagged with a unique Quote ID to support reference and traceability during TA. The format `[P#_S###]` identifies both the speaker and the sequence: `P#` denotes the participant (e.g., `P1` = Participant 1), and `S###` marks the specific utterance from that participant. This system facilitates coding accuracy and enables cross-referencing between annotations and original context. A total of 85 participants contributed to the subset analyzed in this study.

A.3 Visualizing Transcript Embeddings via t-SNE

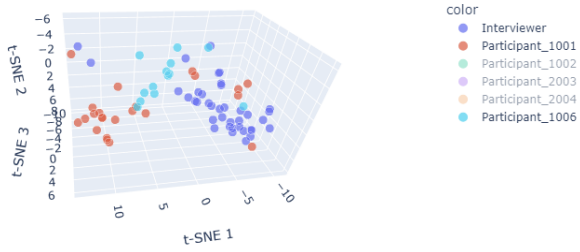


Figure 2: 3D t-SNE projection of sentence embeddings from Interview 1, showing clusters by speaker identity.

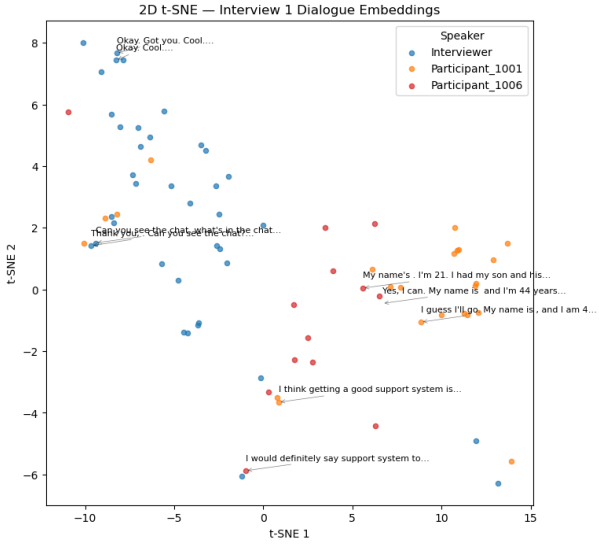


Figure 3: 2D t-SNE projection of sentence embeddings with representative labels and speaker roles.

A.4 Ethics Statement

The original study was approved by the UT Austin Institutional Review Board [Protocol #2019080031] and registered on ClinicalTrials.gov [NCT04613934]. The dataset was generated in the study by Mery et al. (2023).

B Additional Abbreviations

Table 6 lists abbreviations that appear frequently in the paper for ease of reference.

C Evaluation Metrics Details

Levenshtein Distance (D_L) To capture lexical similarity between themes, we compute the normalized Levenshtein distance between each pair (t_i, ℓ_j), defined as the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other.

Term	Definition
Agent	An autonomous computational entity (LLM) that interacts with other agents or the environment to perform specific tasks.
CHD	Congenital Heart Disease
SV-CHD	Single-Ventricle Congenital Heart Disease
AAOCA	A congenital heart condition where a coronary artery arises from the aorta in an atypical location.
TA	Thematic analysis, a method for identifying and reporting patterns, as proposed by Braun and Clarke (2006).
LLM	Large Language Model
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
PPO	Proximal Policy Optimization (Schulman et al., 2017)
IRB	Institutional Review Board

Table 6: Glossary of key terms and abbreviations.

The normalized form is given by:

$$\text{Levenshtein}(t_i, \ell_j) = \frac{\text{edit\_distance}(t_i, \ell_j)}{\max(|t_i|, |\ell_j|)},$$

where $|t_i|$ and $|\ell_j|$ denote the lengths of the respective strings. For interpretability, we convert distance into a similarity score:

$$\text{sim}_{\text{lev}}(t_i, \ell_j) = 1 - \text{Levenshtein}(t_i, \ell_j),$$

so that values closer to 1 indicate higher surface-level string similarity. We report the average maximum similarity per human theme:

$$D_L = \frac{1}{n} \sum_{i=1}^{n} \max_j \text{sim}_{\text{lev}}(t_i, \ell_j),$$

which assesses how well each human theme is lexically approximated by its most similar LLM-generated theme.

D Dependability ($\mathcal{D}$) and Transferability ($\mathcal{T}$) Score Details

Table 7 presents detailed ROUGE-1 and ROUGE-2 scores used to assess theme dependability ($\mathcal{D}$) and transferability ($\mathcal{T}$) across agent identities.

E Supplementary Results on Theme Alignment

This section provides supplementary visualizations related to theme alignment with human ground

Agent Identity	$\mathcal{R}_1$	$\Delta\mathcal{R}_1$	$\mathcal{R}_2$	$\Delta\mathcal{R}_2$	$\mathcal{R}'_1$	$\Delta\mathcal{R}'_1$	$\mathcal{R}'_2$	$\Delta\mathcal{R}'_2$
No Identities (Baseline)	0.564 ± 0.020	–	0.236 ± 0.018	–	0.437 ± 0.020	–	0.180 ± 0.018	–
Cardiac Surgeon	0.563 ± 0.022	-0.001	0.227 ± 0.021	-0.009	0.449 ± 0.028	$+0.012$	0.188 ± 0.026	$+0.008$
Qualitative Researcher	0.538 ± 0.020	-0.026	0.255 ± 0.016	$+0.019$	0.456 ± 0.030	$+0.019$	0.192 ± 0.022	$+0.012$
Medical Doctor	0.546 ± 0.028	-0.018	0.232 ± 0.026	-0.004	0.464 ± 0.009	$+0.027$	0.204 ± 0.006	$+0.024$
Psychologist	0.488 ± 0.026	-0.076	0.229 ± 0.018	-0.007	0.452 ± 0.020	$+0.015$	0.198 ± 0.016	$+0.018$

Table 7: Lexical overlap scores used in evaluating dependability and transferability. $\mathcal{R}_1$ and $\mathcal{R}_2$ denote bidirectional ROUGE-1 and ROUGE-2 scores, respectively, computed across two independent theme generation runs from the same transcript (see Eq. 3–4); their average defines dependability $\mathcal{D}$ (Eq. 5). $\mathcal{R}'_1$ and $\mathcal{R}'_2$ denote analogous ROUGE scores computed between themes generated from different transcript subsets (train vs. validation) and define transferability $\mathcal{T}$ (Eq. 8). Higher values indicate stronger consistency or generalizability.

truth. Figure 4 shows improved cosine similarity between LLM and human themes from iteration 0 to 5. Figure 5 indicates that reviewer-assigned scores remained static, highlighting the need for more sensitive metrics. Table 8 shows semantically aligned theme pairs from the Cardiac Surgeon agent, highlighting shared or complementary meanings despite low similarity scores.

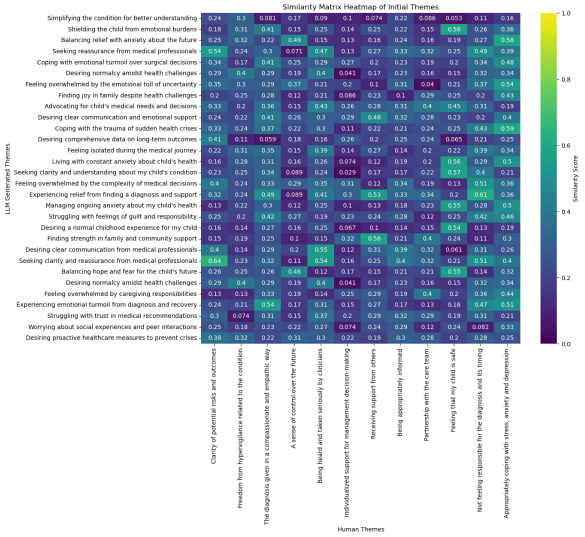
F Example Prompt with Identities

Prompt for Surgical Coder Identity

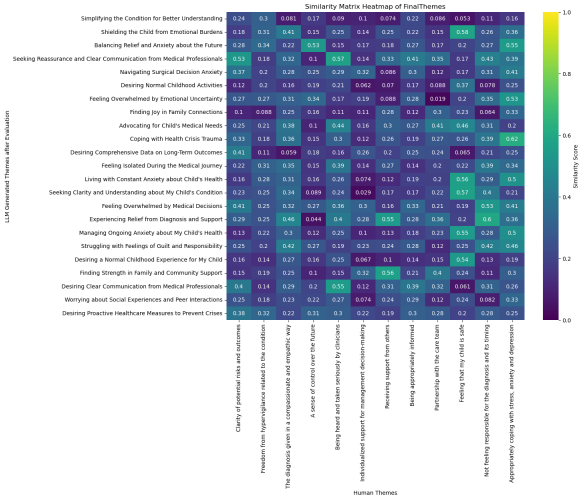
You are an experienced cardiac surgeon specializing in Anomalous Aortic Origins of Coronary Artery (AAOCA). Your role is to code patient narratives from a surgical perspective.
Our end goal is to perform TA on the provided transcript (Braun and Clarke, 2006).
The steps that should be performed are as follows:

1. **Familiarization**
   Read and re-read the data to become deeply familiar with it.
2. **Generating Initial Codes**
   Systematically code interesting features across the dataset.
3. **Searching for Themes**
   Group codes into potential themes, collating relevant data.
4. **Reviewing Themes**
   Check if themes work in relation to coded extracts and the full dataset.
5. **Defining and Naming Themes**
   Refine each theme and define its essence and scope.
6. **Producing the Report**
   Final analysis and write-up with evidence-rich examples.

Your job is to perform **ONLY Step 2**.
You **MUST** provide the unique Quote IDs and descriptions about the codes according to the context of the original text. The Quote IDs in the transcript are marked as `[P1_S002]` for example.
Here is the Original transcript:



(a) Cosine Similarity at Iteration 0



(b) Cosine Similarity at Iteration 5

Figure 4: Comparison of LLM-human theme similarity across iterations. Higher similarity scores in later iterations suggest better thematic alignment after feedback-driven refinement.

G Example Prompt and Generated Codes

G.1 Example Prompt for Code Generation

#	Human Theme	Closest LLM Themes (Cardiac Surgeon, `iteration=0`)
1	Clarity of potential risks and outcomes	Postoperative Outcomes and Recovery Concerns About Long-Term Outcomes and Data Scarcity
2	Freedom from hypervigilance related to the condition	majority of the mental health affect from PTSD Initial Underestimation of Condition Severity
3	The diagnosis given in a compassionate and empathic way	Relief in Diagnosis; Initial Diagnosis and Confusion
4	A sense of control over the future	Post-Surgery Clearance and Future Concerns
5	Being heard and taken seriously by clinicians	Privacy and Protection of Child's Emotional Well-being
6	Individualized support for management decision-making	Decision for Surgery; Decision-Making and Surgical Considerations
7	Receiving support from others	Role of Support Systems
8	Being appropriately informed	Privacy and Protection of Child's Emotional Well-being
9	Partnership with the care team	the kids on his team."; Trust in Medical Team
10	Feeling that my child is safe	Navigating Parent-Child Communication
11	Not feeling responsible for the diagnosis and its timing	Diagnosis and Initial Reactions
12	Appropriately coping with stress, anxiety and depression	Depression; Anxiety

Table 8: **Examples of semantic alignment between human and LLM-generated themes.** Each human-generated theme is paired with two closely aligned LLM-generated themes from the CARDIAC SURGEON identity. Although traditional alignment scores between human- and LLM-generated themes are low, many theme pairs share similar underlying meanings, and some even build upon or encompass each other.

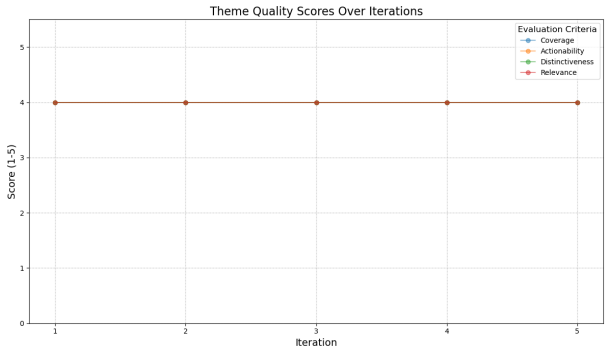


Figure 5: Theme quality scores (Coverage, Actionability, Distinctiveness, and Relevance) across refinement iterations. Flat trends suggest that integer-based evaluation lacks granularity to detect micro-level improvements.

Prompt: Surgical Perspective on Coding Patient Narratives

You are an experienced cardiac surgeon specializing in Anomalous Aortic Origins of Coronary Artery (AAOCA). Your role is to code patient narratives from a surgical perspective.
Our end goal is to perform TA on the provided transcript (Braun and Clarke, 2006).
The steps that should be performed are as follows:

1. **Familiarization:** Read and re-read the data to become deeply familiar with it.
2. **Generating Initial Codes:** Systematically code interesting features across the dataset.
3. **Searching for Themes:** Group codes into potential themes, collating relevant data.
4. **Reviewing Themes:** Check if themes work in relation to coded extracts and the full dataset.
5. **Defining and Naming Themes:** Refine each theme and define its essence and scope.
6. **Producing the Report:** Final analysis and write-up with evidence-rich examples.

Your task is to perform ONLY Step 2. You **MUST** include unique Quote IDs (e.g., `[P1_S002]`) with descriptions grounded in the transcript.

G.2 Example of Generated Codes

Note that Vanilla refers to agents with no designated identities.

Initial Codes from Surgeon Perspective

1. **Understanding AAOCA** *Quote ID: [P1_S001]*
   Use of acronyms and terminology (AAOCA, ALCA) reflects familiarity with medical vocabulary.
2. **Variation in Medical Conditions** *Quote ID: [P2_S001]*
   Illustrates how the anomaly presents differently depending on which coronary artery is affected.
3. **Simplifying Medical Terminology** *Quote ID: [P4_S001]*
   Parents often simplify terminology for ease of communication with others.
4. **Lack of Understanding Among Non-Medical Individuals** *Quote ID: [P4_S003]*
   Indicates that detailed terminology is typically only understood by healthcare professionals.
5. **Emotional Impact of Medical Complexity** *Quote ID: [P5_S001]*
   Parents express emotional burden due to difficulty explaining the condition.

6. **Privacy and Child's Emotional Well-being** *Quote ID: [P2_S003]*
   Parents shield children from information to avoid psychological stress.

...

Initial Codes from Vanilla Identity

1. **Use of Acronyms and Simplification** *Quote ID: [P1_S001], [P2_S001], [P4_S001], [P5_S001], [P3_S001]*
   Participants refer to the condition using acronyms like "AAOCA" and simplified terms to enhance comprehension.

2. **Emotional Impact and Coping** *Quote ID: [P5_S008], [P2_S007], [P4_S007], [P1_S002]*
   Expressed emotions include devastation, stress, and relief; a recurring theme across interviews.

3. **Lack of Information and Data** *Quote ID: [P4_S010], [P5_S017], [P1_S013]*
   Participants highlight frustrations stemming from the rarity of AAOCA and limited clinical data.

4. **Decision-Making Challenges** *Quote ID: [P4_S008], [P2_S011], [P5_S021]*
   Ambiguity in choosing surgery vs. watchful waiting due to mixed medical recommendations.

5. **Protective Parenting and Monitoring** *Quote ID: [P2_S014], [P3_S012], [P5_S024]*
   Parents remain highly vigilant, often suppressing their concerns to avoid alarming children.

...

6. **Sharing Experiences and Offering Help** *Quote ID: [P5_S027], [P1_S021]*
   Participants value peer mentorship and express willingness to support newly diagnosed families.

7. **Impact on Daily Life and Activities** *Quote ID: [P3_S004], [P4_S012], [P5_S015]*
   The condition disrupts routines and limits children's participation in physical activities.