CentaurTA: A Self-Improving Human-Agents Collaboration Framework for Thematic Analysis

Lei Wang
Temple University
tom.lei.wang@temple.edu

Min Huang
Independent Researcher
minerstudy@gmail.com

Eduard Dragut
Temple University
edragut@temple.com

Abstract

Qualitative analysis is essential for studying complex social and behavioral phenomena, yet existing large language model (LLM) approaches face key limitations. Fully automated pipelines often compromise methodological rigor, while manual coding remains costly and labor-intensive. Although recent work emphasizes human–AI collaboration, many multi-agent systems focus primarily on theme-level outputs, provide limited human oversight, and overlook grained, data-level coding quality. We introduce **CentaurTA**, an iterative, self-improving human–agent framework for scalable thematic analysis. CentaurTA places humans in the loop to oversee and guide analysis, using expert feedback as a persistent learning signal to drive prompt-level refinement. By combining structured human feedback with rubric-based evaluation, CentaurTA provides fine-grained supervision for both open coding and theme construction while preserving methodological rigor. Experiments across multiple datasets, baselines, and LLM families show that CentaurTA improves coding alignment and transparency, highlighting the central role of human feedback in reliable qualitative analysis. Our code and data are available at https://github.com/Tom-Owl/CentaurTA.

1 Introduction

Thematic analysis (TA) is widely regarded as a flexible and foundational approach for making sense of qualitative data (Braun and Clarke, 2006). It typically unfolds through iterative cycles of coding and theme development, which require both analytic adaptability and interpretive judgment. Despite its methodological accessibility, however, conducting rigorous TA remains labor-intensive, cognitively demanding, and challenging to replicate (Rietz et al., 2020; Nowell et al., 2017). Recent advances in large language models (LLMs) have created new avenues for augmenting qualitative inquiry. These models are increasingly incorporated into qualitative research workflows, with the most substantial uptake occurring in TA (de Morais Leça et al., 2025). Prior work demonstrates that LLMs can accelerate early-stage analytic tasks such as open coding, generating candidate themes, and scaffolding codebook. For example, CoAIcoder (Gao et al., 2023) integrates a small language model to recommend codes during initial coding, while CollabCoder (Gao et al., 2024) embeds GPT-based assistance into a three-stage pipeline involving independent coding, code merging, and theme generation. In contrast, Qiao et al., 2025 propose a fully automated, multi-agent framework that performs TA without human involvement, raising concerns about bias, loss of reflexivity, and misalignment with core qualitative epistemologies.

Empirical studies caution against fully automated TA. Lee et al., 2024 find that the effective use of ChatGPT in qualitative research requires sustained human-AI collaboration (Wang et al., 2026; Zhu and Callison-Burch, 2025), with human expertise playing a central role in contextual interpretation (Lan et al., 2025), theoretical grounding, and validation. As a result, there is growing consensus that LLMs should function not as autonomous coders but as analytic partners embedded within human-centered workflows that preserve human interpretive authority (Morgan, 2025). However, despite this consensus, existing LLM-assisted QDA systems provide limited support for meaningful human-in-the-loop interaction. Earlier interactive approaches to semi-automated coding emphasize visualization or lightweight intervention, but do not enable systematic learning from human feedback (Drouhard et al., 2017; Marathe and Toyama, 2018; Malaviya et al., 2024). More recent LLM-based frameworks often incorporate human feedback only as a one-off corrective step rather than as a signal for iterative system improvement (Gao et al., 2025). For example, (Xu et al., 2025) introduce a multi-agent, human-in-the-loop framework

Findings of the Association for Computational Linguistics: ACL 2026, pages 15871–15884

for TA of clinical interviews. While the system incorporates expert intervention, it lacks an iterative self-improvement mechanism (Deng et al., 2025) that allows agents to adapt based on expert feedback. As a result, human experts must provide redundant corrections to a static framework. Moreover, the study does not include systematic human evaluation of code-level quality, limiting insight into how AI assistance affects analytic rigor.

Taken together, existing work reveals a critical gap: *the absence of a fine-grained, self-improving human–AI collaboration loop for TA*. Recent position research emphasizes that many human-in-the-loop systems (Chen et al., 2024c; Dragut et al., 2021; Zhang et al., 2019) conceptualize expert input as episodic supervision rather than as a learning signal (Zhang et al., 2025a) that enables continuous system adaptation (Shen et al., 2025). Hence, existing frameworks either emphasize automation at the expense of qualitative rigor or incorporate human feedback in ways that fail to support iterative model or agent improvement. This gap constrains both efficiency and scalability, particularly for large qualitative datasets where repeated manual correction becomes prohibitive.

To bridge this gap, we proposed **CentaurTA**, a self-improving human-agent collaboration framework for TA, which explicitly models human feedback as a persistent learning signal within an iterative analytic loop. Our approach enables LLM-based agents to refine their open coding and theme construction behavior over time based on expert intervention, reducing redundant errors while preserving human interpretive control. Our design is informed by recent advances in self-improving LLM agents and alignment research, which emphasize that reliable improvement requires treating human judgment as repeatable supervision rather than a one-off fix (Mondal et al., 2024; Chen et al., 2024b; Wang et al., 2024; Nayak et al., 2024). Our contributions are:

- **CentaurTA:** a self-improving, human-centered agent framework for thematic analysis with an Actor–Critic design and prompt-level optimization driven by expert feedback.
- **Rubric evaluation:** a constraint-based evaluation protocol for open coding and theme construction that yields fine-grained, actionable signals beyond coarse LLM-as-Judge metrics.
- **Empirical findings:** evidence that iterative human feedback improves alignment, that rubric-based early stopping helps prevent overfitting in self-improvement, and that learned principles can transfer to other platforms.

2 Datasets

In this study, we use 3 domain-specific datasets (Chen et al., 2022; Aljebreen et al., 2024; Yang et al., 2025a; Dugan et al., 2024). **USRS**: this dataset focus on self-regulated learning of Chinese undergraduates, which consists of 12 personal Reflection texts written in Chinese by National Scholarship recipients over the past two years from University H (Huang, 2024). **ASP**: this dataset explores the challenges and assistive technology for autistic job seekers (West et al., 2025). It contains first-person narrative (Subbiah et al., 2025; Wu et al., 2025a) in English from 15 participants representing three employment pathways: university-based coordinators, state vocational rehabilitation job coaches, and autistic individuals with prior experience in these programs (Garrison et al., 2025). **Dreaddit**: a public English-language corpus of short, multi-domain social media texts for stress detection. We use the half of test set for analysis, which can provide presentative test set at the same time align with other dataset make them have the same amount of text (Turcan and McKeown, 2019). USRS and ASP have received IRB approval. All datasets were anonymized prior to analysis, with identifiable information (e.g., names, addresses, and other sensitive personal details) removed or replaced by anonymous identifiers. Table 1 summarizes the dataset statistics used in our experiments.

3 Method

We present an overview of the proposed framework, **CentaurTA**, in Figure 1. Formally, let $\mathcal{D}$ denote a set of long-context documents (Zhang et al., 2024; Paul et al., 2022; Han et al., 2026) with research background as b. Each document $d \in \mathcal{D}$ is decomposed into an ordered list of sentences, denoted as $sent(d) = \{s_1, s_2, \ldots, s_n\}$. The objective of CentaurTA is to perform thematic analysis, which consists of two sub-tasks.

Task 1: Open Coding. The framework first produces a structured set of analytical units, each represented as $(code, quote, ref)$, where $code$ is an open-ended inductive code inferred from d, $quote \subseteq d$ is a concise textual fragment extracted

Dataset	Domain	Context	# Doc	Language	# w	# sent	Avg # w
USRS (Huang, 2024)	Education	Long	12	Chinese	25,718	471	2,143
ASP (Garrison et al., 2025)	Autistic Job Seekers	Middle	15	English	12,678	651	845
Dreaddit (Turcan and McKeown, 2019)	Stress Analysis	Short	214	English	19,358	822	98

Table 1: Data statistics

from the original document, and $ref \subseteq \{1, \ldots, n\}$ denotes the indices of the sentences in *sent* that support the code (Rao et al., 2026). Let $\mathcal{C} = \{c_i\}_{i=1}^{m}$ denote the global set of open codes generated in Task 1, obtained by applying open coding to each document $d \in \mathcal{D}$ and aggregating the results: $\mathcal{C} = \bigcup_{d \in \mathcal{D}} \mathrm{OpenCoding}(d|b)$.

Task 2: Theme Construction. Building on the outputs $\mathcal{C}$ of Task 1, the framework then derives higher-level thematic representations in the form (*theme*, *def*, *codes*, *rationale*). Here, *theme* denotes an abstract concept synthesized from multiple open codes, *def* provides a detailed definition of the theme, *codes* $\subseteq \mathcal{C}$ refers to a subset of the open codes generated in Task 1, and *rationale* explains how these codes are integrated to form the theme.

3.1 Document Batching and Batch Accuracy

To support long-context documents while keeping human feedback cognitively manageable, we process each document d in contiguous *sentence batches*. Specifically, we partition document into batches of size B sentences, and each interaction round operates on one batch. In this study, we set B=10 as a practical trade-off: shorter batches often lack sufficient local context for reliable thematic interpretation, while longer batches impose excessive review burden during human feedback and revision. We use *batch accuracy* to quantify alignment during iterative human–agent collaboration. For each generated item in a batch, the Critic outputs a binary decision $r_c \in \{Y, N\}$ and the human annotator provides a corresponding label r_h. For each code or theme, both the human annotator and the Critic Agent assign a binary label (Y/N). Batch accuracy is defined as the proportion of items labeled as Y within a batch.

3.2 Overview

We propose a self-improving Actor–Critic agent framework with human-in-the-loop to provide high-quality feedback for prompt optimization, illustrated in Figure 1. Unlike conventional reinforcement learning with parameter updates, our approach operates entirely in natural language space and performs optimization at the prompt level for rubric learning, enabling model-agnostic and lightweight adaptation.

3.3 Actor-Critic Module

The Actor-Critic module consists of two complementary agents: an Actor responsible for generation and a Critic responsible for evaluation. In contrast to single-model generation or direct reward modeling, the actor-critic paradigm allows the evaluation strategy to evolve independently from the generation process, which is particularly important when human feedback is sparse or expensive. Given an input x with research background as b and an Actor prompt p_a, the **Actor** generates a candidate output

$$y = \pi_\theta(x \mid p_a, b),$$

where π_θ denotes a LLM parameterized by θ. The Actor prompt p_a encodes task instructions, constraints, and optional reasoning guidance.

Conditioned on the input-output pair (x, y) and a Critic prompt p_c, the LLM-based Critic Q_ϕ evaluates the quality of the Actor's output and produces a binary reward signal:

$$r_c = Q_\phi(x, y \mid p_c, b), r_c \in \{Y, N\}.$$

3.4 Human-Agent Collaboration

Human expert feedback is incorporated to ensure that the Critic's evaluations are aligned with human preference. Given the Actor's output y and Critic's output r_c, a human annotator provides feedback in the form of binary labels and natural language decision notes as rationale, denoted as

$$r_h = \mathcal{H}(x, y, r_c)$$

Specifically, we design a two-stage human feedback protocol that cleanly separates efficiency from authority: *Stage 1 - Simulated human value feedback* A simulated expert agent (GPT-5.2 with domain knowledge) generates provisional feedback (draft labels and short rationales) for candidate outputs. These drafts are used solely to pre-screen and prioritize items for review, reducing the volume

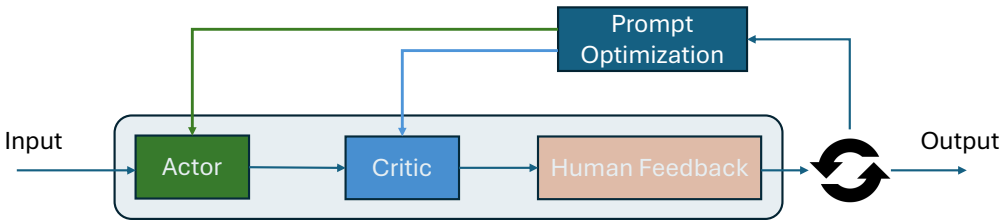


Figure 1: Overview of **CentaurTA**. The framework decomposes long documents into sentence batches, performs open coding and theme construction with an Actor agent, evaluates outputs with a Critic agent, and uses human feedback to refine role-specific prompts via a prompt optimizer.

of cases requiring expert attention. *Stage 2 - Domain expert feedback as final supervision.* Domain experts review, revise, and finalize the simulator's drafts in real time. Only the expert-validated labels and rationales are used to supervise prompt optimization for principles distillation (Jones et al., 2025). Details about the annotators are provided in Appendix A.

Importantly, human feedback is not used to directly optimize the Actor or the Critic agents. Instead, it serves as a supervisory signal for the prompt optimization module, which refines the CoT prompts of both Actor and Critic agents. Moreover, the learning signal is not derived from model–model agreement: the simulator's outputs are proposals, while experts determine the final supervision. This design lowers repetitive annotation cost while preserving expert control over what the system learns.

3.5 Prompt Optimization

The Prompt Optimizer is an LLM-based agent parameterized by a prompt p_o that coordinates the refinement of the chain-of-thought (CoT) prompts for both the Actor and the Critic, enabling prompt level self improvement. The Actor and Critic prompts are refined using role specific supervision signals. The Actor prompt is updated based on human feedback over generated outputs, whereas the Critic prompt is optimized to better match human assessments of its binary evaluations. The prompt optimization process is defined as:

$$p_a^* = \text{PromptOpt}(x, y, r_h, p_a \mid p_o, b), \quad (1)$$

$$p_c^* = \text{PromptOpt}(x, r_c, r_h, p_c \mid p_o, b). \quad (2)$$

We refer to the evolving CoT prompts p_a^* and p_c^* as *learned principles*: they are explicit, natural language guidelines distilled from human rationales that steer generation (Actor) and evaluation (Critic). Learned principles can be snapshot at different iterations and transferred (Li et al., 2019) to other workflows (e.g., by replacing the prompts used in other systems) to test generalization.

This coordinated, parallel prompt refinement drives a self-improving system evolution. As the Actor produces increasingly higher quality outputs and the Critic becomes more aligned with human judgment, the system's dependence on human expert feedback naturally diminishes. Consequently, human supervision becomes progressively sparser over time, substantially improving the efficiency and scalability of the overall framework.

3.6 Evaluation Methods

To assess the performance of our CentaurTA framework, we employed three complementary evaluation methods: (1) LLM-as-Judge for credibility and conformability and (2) rubric based evaluation. These evaluation methods offer comprehensive insight into CentaurTA's performance for open coding and theme construction.

Credibility and Conformability 1) Credibility evaluates whether the generated codes, sub-themes, or themes can represent the data. 2) Conformability assesses whether the generated codes, sub-themes, or themes are data-driven and consistent with the original input context (Qiao et al., 2025). For both dimensions, the LLM-as-Judge (Zheng et al., 2023) makes binary decision for each output, and the **success rate** (percentage of correct cases) is reported as metric.

Rubric-based evaluation. To obtain fine-grained, verifiable quality signals for both open coding and theme construction, we instruct domain experts to build a rubric library (Asai et al., 2026; Yifei et al., 2025; Rao and Callison-Burch, 2026a; Chen et al., 2024a) that decomposes qualitative-method requirements into binary, checkable constraints, which are applied via an LLM-as-Judge prompt. For rubric based evaluation of open coding, we

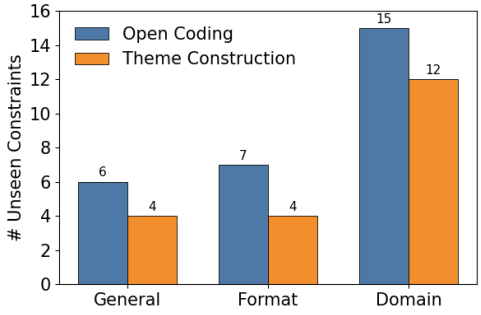


Figure 2: Domain Expert design unseen constraints for rubric-based evaluation

Domain	Method	Model	Credibility (%)	Conformability (%)
USRS	Atlas	-	50.00	50.00
	MC	GPT-5	100.00	100.00
	TA	GPT-5	100.00	100.00
	TA	GPT-5.2	90.00	90.00
ASP	Atlas	-	80.00	80.00
	MC	GPT-5	91.67	91.67
	TA	GPT-5	92.31	92.31
	TA	GPT-5.2	94.12	94.12
Dreaddit	Atlas	-	83.33	83.33
	MC	GPT-5	61.11	61.11
	TA	GPT-5	93.33	93.33
	TA	GPT-5.2	100.00	100.00

Table 2: Performance comparison of theme construction

define the following constraints for code format:

- **R1 (Context alignment):** Codes must be closely aligned with the stated research background and analytic focus.
- **R2 (Avoid over-abstraction):** Codes should remain descriptive and avoid excessive conceptualization or theory-level inference.
- **R3 (Semantic clarity):** Each code must be semantically self contained and clearly interpretable in isolation.
- **R4 (Evidence grounding):** The quoted text must accurately and sufficiently support the meaning of the code.
- **R5 (Semantic completeness):** Each code should explicitly specify its core semantic elements (e.g., actor, action, and object or information source) to avoid underspecification.
- **R6 (Faithful interpretation):** Codes must not introduce unsupported causal claims or interpretations beyond what is warranted by the quoted evidence.
- **R7 (Conciseness and precision):** Codes should be concise and focused, avoiding non-essential modifiers while preserving essential meaning.

We ensure that the resulting constraints are new and diverse (Lin et al., 2024; Pyatkin et al., 2025; Li et al., 2026; Feng et al., 2025b; Rao and Callison-Burch, 2026b). As shown in Figure 2, for *open coding*, each {code, quote, sentence_id} item is evaluated using 18 constraints: 6 *general* rules (e.g., avoiding unsupported causal inference and enforcing clear conceptual boundaries), 7 *format/linguistic* constraints (e.g., noun/gerund phrasing, length limits, semantic self containment, and referential clarity), and 5 *domain specific* constraints tailored to each dataset. For *theme construction*, each theme is evaluated with 12 constraints: 4 *general* theme-construction rules (e.g., level consistency and inter theme distinctiveness), 4 *format* constraints (e.g., concise naming and complete inclusion of supporting codes), and 4 *domain-specific* constraints enforcing domain-appropriate thematic organization.

Given the research background, source text, and system outputs, an LLM based judge returns a JSON decision matrix with Y/N labels for each rubric; we compute **rubric accuracy** as the average constraint satisfaction rate across items (optionally stratified by rubric category).

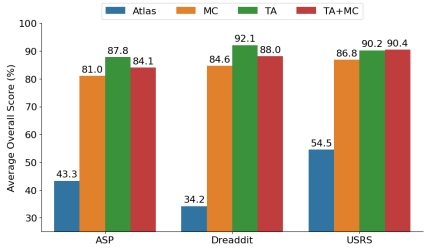
4 Results

4.1 Experimental Setting and Details

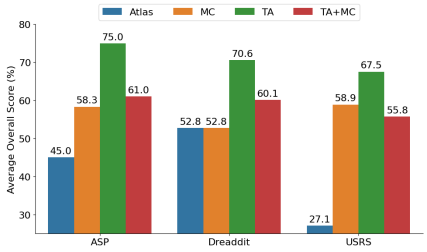
We evaluate the performance of our CentaurTA framework against 2 baselines: 1) MindCoder (Gao et al., 2025), an LLM-powered workflow for qualitative analysis, while enabling humans to conduct meaningful interpretation. 2) Atlas.ti (ATLAS.ti, 2025), a commercial platform for AI Coding.

4.2 The Necessity of Rubric-based Evaluation

Previous work has employed an LLM-as-Judge paradigm to evaluate qualitative codebooks in terms of credibility and conformability. As shown in Table 2, we conduct thematic analysis using three platforms, Atlas, MindCoder, and CentaurTA, across three domains: USRS, ASP, and Dreaddit. While credibility and conformability provide quantitative assessments of codebook quality, this evaluation paradigm has several limitations. 1) It lacks mechanisms for evaluating open-code phrases; 2) it does not provide fine-grained feedback signals to support iterative codebook improvement; and, 3) it fails to access the alignment between AI-assisted tools with human expert. To address these issues, we introduce a rubric-based evaluation framework for multifaceted assessment of thematic analysis across both the open-coding and theme construc-



(a) Baseline comparison for the open coding task evaluated with a rubric-based LLM-as-Judge.



(b) Baseline comparison for the theme construction task evaluated with a rubric-based LLM-as-Judge.

Figure 3: Baseline comparisons for open coding and theme construction tasks under rubric-based LLM-as-Judge evaluation. MC denotes MindCoder, TA denotes CentaurTA, and TA+MC denotes applying the learned principles of CentaurTA to MindCoder.

tion stages. As shown in Figure 3, the proposed CentaurTA framework consistently achieves better performance, particularly in the theme construction stage. More detailed results are provided in Appendix B, Table 4.

Moreover, the principles learned through the human–agent collaboration process can be transferred to other AI platforms to enhance their performance on thematic sub-tasks, demonstrating the transferability of the proposed method. Figure 4 illustrates an example of such learned principles. We observe that as human–agent feedback interactions increase, the actor agent's guiding principles continue to evolve and become more specifically focused on open coding. For example, the agent increasingly emphasizes aligning codes with the original context, ensuring semantic completeness, and avoiding fragmented coding. These observations indicate that the proposed human–agent collaboration framework successfully transfers human expert knowledge to the actor agent through iterative self-improvement, in which human feedback serves as a supervisory signal guiding the prompt optimizer to refine the actor agent's coding principles, thereby improving alignment over successive iterations.

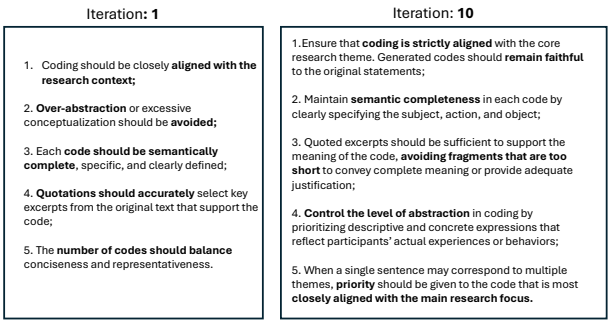


Figure 4: Learned principles of the actor agent for open coding at the 1st and 10th iterations.

4.3 Effectiveness of Human–Agent Collaboration

Iterative human–agent collaboration substantially enhances open coding performance. Following the batching protocol in Section 3.1, we measure *batch accuracy* as the agreement between Critic decisions and human labels within each 10-sentence batch. As illustrated in Figure 6a, batch accuracy increases monotonically with the number of human feedback iterations across two expert interaction trajectories. After approximately 20 interaction rounds, batch accuracy consistently exceeds 90%, indicating that the agentic system achieves near-complete alignment with human experts for a given document piece. These results demonstrate that the proposed CentaurTA can reliably align agent behavior with expert-level thematic analysis through iterative self-improving. In our framework, human experts are only required to supervise and revise the critic agent's decisions. All labor-intensive components, including manual annotation and prompt engineering, are handled automatically by the critic agent and prompt optimization agent, enabling high-quality open coding with substantially reduced human effort.

Figure 5 presents the detailed rubric-based scores for coding format during the open coding process. We observe that different rubrics exhibit distinct evolutionary trajectories as the number of iterations increases. This pattern reflects the inherently complex nature of thematic analysis, in which multiple objectives are optimized simultaneously and improvements along one constraint may occur at the expense of another. In particular, we observe occasional tension between R5 and R7. R5 requires that codes explicitly specify the actor, information source, and purpose, whereas R7 emphasizes that coding expressions should be concise and precise,

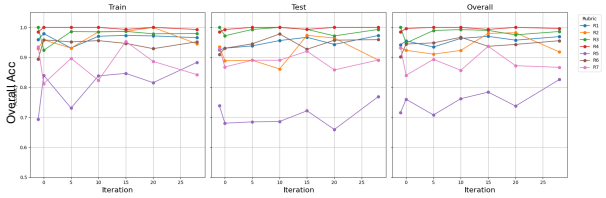
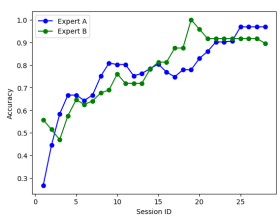
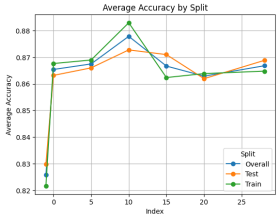


Figure 5: Evolution of rubric-based coding format scores across iterations in the open coding process, highlighting trade-offs between conciseness and semantic completeness.



(a) Human-Agent Collaboration for self-improving



(b) Rubric-based evaluation for early stopping

Figure 6: Human–Agent collaboration and self-improving workflow.

eliminating non-core modifiers. As a result, enforcing excessive conciseness may sometimes lead to the omission of critical semantic elements, such as the acting subject or information source, explaining the observed trade-off between these two rubrics.

4.4 Rubric-Based Early Stopping to Prevent Overfitting in Self-Improving Process

Although iterative human feedback consistently improves batch-level accuracy in Figure 6a, excessive self-improvement may lead to overfitting, where learned principles become overly specialized to a single document fragment. To explore this overfitting risk in the self-improvement process, we extract learned principles (i.e., the optimized Actor/Critic CoT prompts) at different stages on the USRS domain (iterations 0, 1, 5, and up to 25) and evaluate them on both training and held-out test sets using our proposed rubric-based evaluation method. Results are shown in Figure 6b. We observe rubric scores increase steadily up to approximately 10 iterations and decline thereafter on both splits, suggesting that continued feedback leads to overfitting. Unlike conventional overfitting, where training performance improves while test performance degrades, overfitting here occurs when batch-level accuracy continues to rise (Figure 6a) but overall rubric-based performance decreases on both training and test sets. This finding shows that rubric-based evaluation can serve as a signal for early stopping in human–agent collaboration. Our proposed CentaurTA can achieve its maximum performance gain with only 10% of the full document for self-improving human-agent collaboration for open coding, substantially reducing human effort compared to prior approaches require manual review of all open coding outputs and customized prompt engineering.

4.5 The Transition from Human-driven to Agent-driven Self-improvement

As shown in Figure 7, the Critic Agent supports self-improvement by comparing its feedback with human annotations and providing discrepancy signals to the prompt optimizer. The left region (highlighted in red) represents the human-driven phase, where continuous expert feedback aligns the agent with human preferences and leads to increasing batch accuracy. As training progresses, the system transitions to an agent-driven regime, in which the Critic Agent performs self-critique and autonomously generates optimization cues. The resulting accuracy fluctuations in the right region indicate effective self-refinement with reduced human intervention.

5 Analysis

Agreement Between LLM Judge and Human Annotators Expert alignment was established during rubric construction (Cohen's $\kappa = 0.78$ among three domain experts), reflecting agreement on the evaluation criteria. We further conducted an additional validation study: a domain expert manually labeled 300 open-coding results in the USRS domain and we compared these labels with rubric-based LLM-as-Judge decisions (Yang et al., 2025b). Inter-annotator Agreement (IAA) was evaluated using Cohen's κ and constraint-level accuracy. The LLM judge achieves 90% average rubric accuracy, comparable to the human expert's 89%. The overall IAA score is $\kappa = 0.68$, indicating substantial agreement. Agreement is higher for general rules ($\kappa = 0.71$) and slightly lower for domain-specific constraints ($\kappa = 0.64$), which are inherently more nuanced. These results suggest that the rubric-based protocol provides reliable, expert-aligned constraint verification rather than coarse end-to-end scoring.

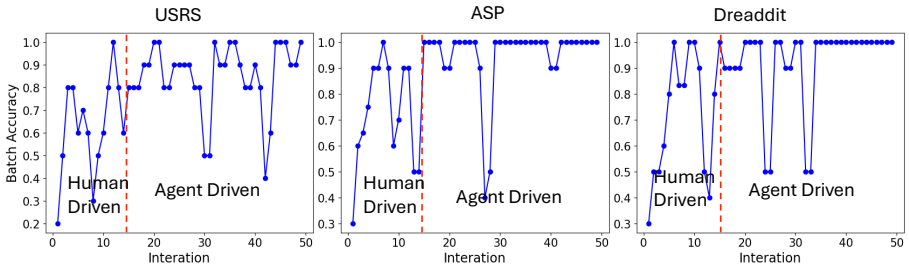


Figure 7: The transition from human-driven to agent-driven self-improvement.

Ablation Study To clarify the contributions of different feedback sources and system components, we conducted controlled experiments on the *Open Coding* task in the USRS domain (batch size = 10 sentences, maximum 25 refinement iterations). Figure 8 shows the impact of removing different feedback sources. The full CENTAURTA system achieves 90% rubric accuracy with an average runtime of 25 minutes across 10 refinement rounds. Removing iterative feedback reduces accuracy to 81%, indicating that iterative refinement is necessary for reliable performance. Using only domain-expert feedback maintains comparable accuracy (90%) but increases runtime substantially to 42 minutes, while using only simulated-human feedback improves efficiency (7 minutes, 5 rounds) but reduces accuracy to 85%, suggesting that performance gains arise from the complementary two-stage feedback mechanism rather than the simulator alone.

6 Related Work

Multi-agent Systems for Thematic Analysis Multi-agent frameworks extend LLM-based Thematic Analysis by assigning specialized roles (e.g., coder, aggregator, reviewer) to improve scalability and diversity of interpretation (Lee et al., 2024; Gao et al., 2023, 2024; Feng et al., 2025a). Thematic-LM coordinates multiple agents to maintain codebooks and aggregate codes across corpora (Qiao et al., 2025). However, existing agentic systems typically lack mechanisms to learn from human feedback over time and are often evaluated only at the theme level, without assessing code-level quality. Moreover, reliance on automatic metrics or LLM-based evaluators risks misalignment with human qualitative judgment (Wang and Dragut, 2024). Other recent agentic approaches share similar limitations, including limited human evaluation and static agent behavior (Xu et al., 2025; Yi et al., 2025). While reinforcement learning has been proposed to improve agent performance, its applicability to qualitative analysis is unclear given the small number of codes and themes typically involved. MindCoder integrates LLM outputs, self-critiques, and human revision within an interactive framework (Gao et al., 2025), but relies heavily on manual inspection and prompt engineering, creating scalability bottlenecks.

Feedback-Driven Self-Improvement in LLM Agents Recent advances in self-improving large language model (LLM) agents have moved beyond prompt-level self-correction toward explicit feedback-driven learning loops. Prior work has shown that intrinsic self-correction without a reliable external signal can be brittle and may amplify bias or indecision (Zhang et al., 2025b). To address this limitation, subsequent approaches introduce dedicated critic agents, distilled critic models (Mondal et al., 2024), and decomposition-based critique–refine pipelines (Ferraz et al., 2024) to strengthen feedback quality.

7 Conclusions

This work introduces **CentaurTA**, a self-improving human–agent collaboration framework for thematic analysis that treats expert feedback as a persistent learning signal rather than a one-off correction. CentaurTA operationalizes thematic analysis into open coding and theme construction, and couples an Actor–Critic module with a prompt-optimization loop so that both generation and evaluation criteria can be refined over successive human feedback rounds.

A central novelty of our approach is a *rubric-based evaluation* protocol that provides fine-grained, verifiable signals for both stages of thematic analysis. Beyond coarse end-to-end judgments (e.g., credibility and conformability), our rubrics evaluate code- and theme-level quality under explicit methodological and structural constraints, enabling interpretable diagnostics, transfer-

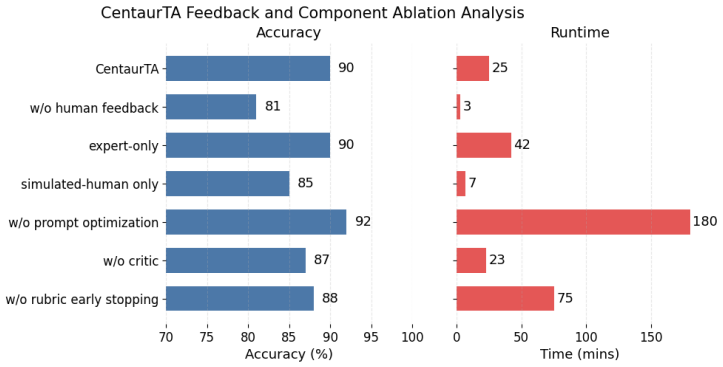


Figure 8: Impact of removing different feedback sources on accuracy and runtime

able *learned principles* (i.e., optimized Actor/Critic CoT prompts), and rubric-based early stopping to mitigate overfitting during self-improvement.

Across three domains (USRS, ASP, and Dreaddit) and multiple baselines, CentaurTA achieves consistently strong thematic outcomes, and iterative human–agent collaboration improves open-coding alignment with experts (exceeding 90% batch accuracy after $\sim$20 interaction rounds). Moreover, rubric trajectories reveal that peak generalizable performance can be reached with limited feedback (e.g., $\sim$10% of a document), substantially reducing redundant expert corrections; the learned principles can also be transferred to improve other AI-assisted workflows.

8 Acknowledgments

This work was supported in part by the U.S. NSF under awards 2107213 and 2026513. We thank Michael West for valuable suggestions on thematic analysis evaluation. We also thank Yifei Li for inspiring us to explore rubric-based evaluation methods. We thank Drs. Stephen MacNeil and Hongchang Gao for their constructive feedback.

9 Limitation

We acknowledge several limitations for future exploration. First, open-source LLMs could be employed as the underlying models for agents. Second, more sophisticated techniques for prompt optimization could be investigated, such as reinforcement learning or evolutionary algorithms. Third, while our rubric-based evaluation protocol provides fine-grained signals, it may not capture all aspects of thematic quality; Finally, human evaluations for interpretability and error analysis can gain deeper insights (Wen et al., 2025; Wen, 2025; Wen and Rezapour, 2025).

Ethical Considerations

This work studies LLM-assisted thematic analysis of qualitative text and proposes a human–agent collaboration framework that explicitly preserves human interpretive control. The study does not involve experiments on human subjects beyond expert-in-the-loop feedback for evaluating and guiding model outputs. Such feedback constitutes low-risk annotation of non-sensitive text and does not involve intervention, deception, or collection of personal identifiers.

We evaluate CentaurTA on three datasets spanning education, assistive employment, and social media. Two datasets (USRS and ASP) received prior IRB approval from their original studies, and all datasets were anonymized before analysis, with identifiable information removed or replaced by anonymous identifiers. We adhere to the licenses and usage terms of all datasets and do not release any new raw personal data.

Because qualitative analysis involves interpretive judgment, improper automation may introduce bias, over-abstraction, or epistemic distortion. To mitigate these risks, CentaurTA requires explicit evidence grounding for each code, incorporates rubric-based constraints derived from qualitative methodology, and places human experts in the loop to oversee, correct, and guide agent behavior.

Finally, we acknowledge the environmental and computational costs associated with LLMs. Our framework operates entirely at the prompt level without parameter updates, enabling lightweight adaptation and reducing unnecessary computation. We believe these design choices support responsible deployment of LLMs in qualitative research while maintaining methodological rigor and human accountability.

References

Abdullah Aljebreen, Weiyi Meng, and Eduard C. Dragut. 2024. Analysis and detection of "pink slime" websites in social media posts. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2572–2581, New York, NY, USA. Association for Computing Machinery.

Akari Asai, Jiashu He, Rui Shao, and 1 others. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650:857–863.

ATLAS.ti. 2025. Ai coding powered by OpenAI. `https://atlasti.com/ai-coding-powered-by-openai`. Accessed: 2025-12-24.

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

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024a. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Xiusi Chen, Hongzhi Wen, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, and Wei Wang. 2024b. IterAlign: Iterative constitutional alignment of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1423–1433, Mexico City, Mexico. Association for Computational Linguistics.

Zhijia Chen, Lihong He, Arjun Mukherjee, and Eduard Dragut. 2024c. Comquest: Large scale user comment crawling and integration. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD '24, page 432–435, New York, NY, USA. Association for Computing Machinery.

Zhijia Chen, Weiyi Meng, and Eduard Dragut. 2022. Web record extraction with invariants. *Proc. VLDB Endow.*, 16(4):959–972.

Matheus de Morais Leça, Lucas Valença, Reydne Santos, and Ronnie de Souza Santos. 2025. Applications and implications of large language models in qualitative analysis: A new frontier for empirical software engineering. In *Proceedings of the 2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering*, WSESE '25, page 36–43. IEEE Press.

Shijian Deng, Kai Wang, Tianyu Yang, Harsh Singh, and Yapeng Tian. 2025. Self-improvement in multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1987–2006, Suzhou, China. Association for Computational Linguistics.

Eduard Dragut, Yunyao Li, Lucian Popa, and Slobodan Vucetic. 2021. Data science with human in the loop. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 4123–4124, New York, NY, USA. Association for Computing Machinery.

Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Pena-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 220–229. IEEE.

Mingzhe Du, Luu Anh Tuan, Yue Liu, Yuhao Qing, Dong Huang, Xinyi He, Qian Liu, Zejun Ma, and See-kiong Ng. 2025. Afterburner: Reinforcement learning facilitates self-improving code efficiency optimization. *arXiv preprint arXiv:2505.23387*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. 2025a. Rethinking LLM uncertainty: A multi-agent approach to estimating black-box model uncertainty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12349–12375, Suzhou, China. Association for Computational Linguistics.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. 2025b. Vericot: Neuro-symbolic chain-of-thought validation via logical consistency checks. *Preprint*, arXiv:2511.04662.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. Coaicoder: Examining the effectiveness of ai-assisted human-to-human collaboration in qualitative analysis. *ACM Trans. Comput.-Hum. Interact.*, 31(1).

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: A lower-barrier, rigorous

workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jie Gao, Zhiyao Shu, and Shun Yi Yeo. 2025. Mindcoder: Automated and controllable reasoning chain in qualitative analysis. *Preprint*, arXiv:2501.00775.

Elizabeth Garrison, Stephen MacNeil, Donald A. Hantula, Michael West, Eduard Dragut, Matt Tincani, and Slobodan Vucetic. 2025. Exploring the challenges and assistive technology for autistic job seekers across employment pathways. *Research in Developmental Disabilities*, 167:105155.

Feijiang Han, Zelo ng Wang, Bowen Wang, Xinxin Liu, Skyler Cheung, Delip Rao, Chris Callison-Burch, and Lyle Ungar. 2026. Latex2layout: High-fidelity, scalable document layout annotation pipeline for layout detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):30907–30915.

Min Huang. 2024. *Beyond One-Dimension: A Study on Self-Regulated Learning and Its Typological Optimization of Chinese Undergraduates with National Scholarship*. Doctoral dissertation, Huazhong University of Science and Technology.

Haydn Thomas Jones, Natalie Maus, Josh Magnus Ludan, Maggie Ziyu Huan, Jiaming Liang, Marcelo Der Torossian Torres, Jiatao Liang, Zachary Ives, Yoseph Barash, Cesar de la Fuente-Nunez, Jacob R. Gardner, and Mark Yatskar. 2025. A dataset for distilling knowledge priors from literature for therapeutic design. *Preprint*, arXiv:2508.10899.

Fangping Lan, Abdullah Aljebreen, and Eduard Dragut. 2025. UniT: One document, many revisions, too many edit intention taxonomies. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23005–23024, Vienna, Austria. Association for Computational Linguistics.

V Vien Lee, Stephanie CC 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.

Sunzhu Li, Jiale Zhao, Miteto Wei, Huimin Ren, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Wei Chen. 2026. Rubrichub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. *Preprint*, arXiv:2601.08430.

Wenbo Li, Longyin Wen, Xiao Bian, and Siwei Lyu. 2019. Evolvement constrained adversarial learning for video style transfer. In *Computer Vision – ACCV 2018*, pages 232–248, Cham. Springer International Publishing.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *Preprint*, arXiv:2406.04770.

Chaitanya Malaviya, Subin Lee, Dan Roth, and Mark Yatskar. 2024. What if you said that differently?: How explanation formats affect human feedback efficacy and user perception. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3046–3065, Mexico City, Mexico. Association for Computational Linguistics.

Megh Marathe and Kentaro Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Sneha Mondal, Ashish Agrawal, Ritika, Preethi Jyothi, and Aravindan Raghuveer. 2024. DIMSIM: Distilled multilingual critics for Indic text simplification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16093–16109, Bangkok, Thailand. Association for Computational Linguistics.

David L Morgan. 2025. Query-based analysis: A strategy for analyzing qualitative data using chatgpt. *Qualitative Health Research*, page 10497323251321712.

Sid Nayak, Adelmo Morrison Orozco, Marina Have, Jackson Zhang, Vittal Thirumalai, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, James Harrison, and 1 others. 2024. Long-horizon planning for multi-agent robots in partially observable environments. *Advances in Neural Information Processing Systems*, 37:67929–67967.

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):1–13.

Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K. Roy-Chowdhury. 2022. Text-based temporal localization of novel events. In *Computer Vision – ECCV 2022*, pages 567–587, Cham. Springer Nature Switzerland.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. 2025. Generalizing verifiable instruction following. *Preprint*, arXiv:2507.02833.

Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. Thematic-lm: A llm-based multi-agent system for large-scale thematic analysis. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 649–658, New York, NY, USA. Association for Computing Machinery.

Delip Rao and Chris Callison-Burch. 2026a. Autorubric: Unifying rubric-based llm evaluation. *Preprint*, arXiv:2603.00077.

Delip Rao and Chris Callison-Burch. 2026b. Bibtex citation hallucinations in scientific publishing agents: Evaluation and mitigation. *Preprint*, arXiv:2604.03159.

Delip Rao, Eric Wong, and Chris Callison-Burch. 2026. Detecting and correcting reference hallucinations in commercial llms and deep research agents. *Preprint*, arXiv:2604.03173.

Tim Rietz, Peyman Toreini, and Alexander Maedche. 2020. Cody: An interactive machine learning system for qualitative coding. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 90–92.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, and 5 others. 2025. Towards bidirectional human–ai alignment. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*. Position Paper Track.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 185–203, Suzhou, China. Association for Computational Linguistics.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. 2024. E^2CL: Exploration-based error correction learning for embodied agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7626–7639, Miami, Florida, USA. Association for Computational Linguistics.

Lei Wang and Eduard Dragut. 2024. The overlooked repetitive lengthening form in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16225–16238, Miami, Florida, USA. Association for Computational Linguistics.

Lei Wang, Min Huang, and Eduard Dragut. 2026. Danceha: A multi-agent framework for document-level aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29714–29722.

Ximing Wen. 2025. Language model meets prototypes: Towards interpretable text classification models through prototypical networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29307–29308.

Ximing Wen and Rezvaneh Rezapour. 2025. A transformer and prototype-based interpretable model for contextual sarcasm detection. *arXiv e-prints*, pages arXiv–2503.

Ximing Wen, Wenjuan Tan, and Rosina Weber. 2025. Gaprotonet: A multi-head graph attention-based prototypical network for interpretable text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9891–9901.

M. West, M. Tincani, D. Hantula, and 1 others. 2025. Applying data science practices to identify characteristics of postsecondary autism support programs from their websites. *Journal of Autism and Developmental Disorders*.

Qiling Wu, Hyangeun Ji, Yuhyun Park, Sori Kim, Jiayu Yang, and Lei Wang. 2025a. East asian international doctoral students' role identity development in the united states: A collaborative autoethnography. *Journal of International Students*, 15(1):61–86.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. 2025b. Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11548–11565, Suzhou, China. Association for Computational Linguistics.

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

Tianyu Yang, Yuhan Liu, Sobin Alosious, Ethan A. Brown, Jason R. Rohr, Tengfei Luo, and Xiangliang Zhang. 2025a. Quest2DataAgent: Automating end-to-end scientific data collection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 500–514, Suzhou, China. Association for Computational Linguistics.

Tianyu Yang, Yuhan Liu, Sobin Alosious, Ethan A. Brown, Jason R. Rohr, Tengfei Luo, and Xiangliang Zhang. 2025b. Quest2DataAgent: Automating end-to-end scientific data collection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 500–514, Suzhou, China. Association for Computational Linguistics.

Seungjun Yi, Joakim Nguyen, Huimin Xu, Terence Lim, Andrew Well, Mia Markey, and Ying Ding. 2025.

Auto-ta: Towards scalable automated thematic analysis (ta) via multi-agent large language models with reinforcement learning. *Preprint*, arXiv:2506.23998.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Preprint*, arXiv:2509.00496.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Qi Zhang, Huitong Pan, Zhijia Chen, Longin Jan Latecki, Cornelia Caragea, and Eduard Dragut. 2025a. DynClean: Training dynamics-based label cleaning for distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2540–2556, Albuquerque, New Mexico. Association for Computational Linguistics.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, and Han Qiu. 2025b. Understanding the dark side of LLMs' intrinsic self-correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27066–27101, Vienna, Austria. Association for Computational Linguistics.

Shanshan Zhang, Lihong He, Eduard Dragut, and Slobodan Vucetic. 2019. How to invest my time: Lessons from human-in-the-loop entity extraction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2305–2313, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Andrew Zhu and Chris Callison-Burch. 2025. Overhearing llm agents: A survey, taxonomy, and roadmap. *Preprint*, arXiv:2509.16325.

A Human Annotator Details

We recruited 6 participants (Table 3) through university mailing lists and LinkedIn. Two self-identified as experts in thematic analysis and four as intermediate practitioners. Participants were divided into two groups: 1) Three annotators participated in the CentaurTA human-in-the-loop setting, providing iterative feedback during self-improvement. Each covered two domain datasets based on expertise. No discussion or cross-communication was allowed in this group to ensure independent supervision; 2) The remaining three formed an expert panel to curate the verifiable rubric library used for rubric-based evaluation. The panel conducted structured discussions to refine and validate the rubric constraints, reaching consensus on the validity and diversity of the final rubric library, with an agreement score (Cohen's Kappa) of 0.78. All participants were compensated at USD $20 per hour.

B Details of Rubric-based Evaluation

We provide more detailed rubric-based evaluation results for Open Coding and Theme Construction in Table 4.

C More Related Work

Related efforts in alignment research similarly emphasize maintaining human intent across iterative updates. For example, iterative constitutional alignment operationalizes rule-based guidance as a persistent optimization target (Chen et al., 2024b), while meta-judge and reward-modeling frameworks transform human evaluations into reusable training signals for continued improvement (Wu et al., 2025b). In interactive agent settings, feedback loops are further grounded in environmental interaction, where agents iteratively plan, act, verify, and revise based on observed outcomes (Wang et al., 2024; Nayak et al., 2024). Reinforcement learning approaches extend this paradigm by enabling sustained improvement through measurable objectives, such as optimizing code efficiency via execution feedback when supervised refinement reaches a plateau (Du et al., 2025).

PID	TA Exp.	Years of TA	Domain Expertise	Position	Human Feedback	Rubric curation
P1	Expert	10	USRS + ASP	Professor	Y	N
P2	Intermediate	4	USRS + Dreaddit	PhD student	Y	N
P3	Intermediate	3	ASP + Dreaddit	PhD student	Y	N
P4	Expert	7	USRS + Dreaddit	Professor	N	Y
P5	Intermediate	5	USRS + Dreaddit + ASP	PhD student	N	Y
P6	Intermediate	4	ASP	PhD student	N	Y

Table 3: Human annotators' expertise, roles, and task assignments.

Domain	**Method**	**domain**	**Code**				**Theme**			
			domain	**format**	**general**	**Avg**	**domain**	**format**	**general**	**Avg**
USRS	Atlas	-	43.39	60.5	59.56	54.48	25.0	50.0	6.25	27.08
	MC	GPT-5	80.1	89.18	91.15	86.81	44.83	55.17	76.72	58.91
	TA+MC	GPT-5	83.71	97.58	89.92	90.4	42.31	40.38	84.62	55.77
	TA	GPT-5	84.0	94.17	92.47	90.21	48.68	69.74	84.21	67.54
ASP	Atlas	-	26.09	50.97	52.86	43.31	60.0	60.0	15.0	45.0
	MC	GPT-5	65.87	83.92	93.36	81.05	72.92	48.96	53.12	58.33
	TA+MC	GPT-5	66.22	91.51	94.44	84.06	66.96	55.36	60.71	61.01
	TA	GPT-5	73.99	95.09	94.27	87.78	78.85	69.23	76.92	75.0
Dreaddit	Atlas	-	21.51	52.01	29.1	34.21	58.33	62.5	37.5	52.78
	MC	GPT-5	89.18	77.26	87.41	84.62	65.28	47.22	45.83	52.78
	TA+MC	GPT-5	88.35	85.02	90.78	88.05	64.29	49.11	66.96	60.12
	TA	GPT-5	95.24	93.93	87.18	92.12	66.67	70.0	75.0	70.56

Table 4: Rubric-based evaluation for Open Coding and Theme Construction.