DeTAILS: Deep Thematic Analysis with Iterative LLM Support

Ansh Sharma
a646shar@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Karen Cochrane
karen.cochrane@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

James R. Wallace
james.wallace@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Thematic analysis is widely used in qualitative research but can be difficult to scale because of its iterative, interpretive demands. We introduce DeTAILS, a toolkit that integrates large language model (LLM) assistance into a workflow inspired by Braun and Clarke's thematic analysis framework. DeTAILS supports researchers in generating and refining codes, reviewing clusters, and synthesizing themes through interactive feedback loops designed to preserve analytic agency. We evaluated the system with 18 qualitative researchers analyzing Reddit data. Quantitative results showed strong alignment between LLM-supported outputs and participants' refinements, alongside reduced workload and high perceived usefulness. Qualitatively, participants reported that DeTAILS accelerated analysis, fostered trust through transparency and control, but primarily supported verification-oriented reflexivity, falling short of deeper positional reflexivity. We contribute: (1) an interactive human–LLM workflow for large-scale qualitative analysis, (2) empirical evidence of its feasibility and researcher experience, and (3) design implications for trustworthy AI-assisted qualitative research.

ACM Reference Format:
Ansh Sharma, Karen Cochrane, and James R. Wallace. 2018. DeTAILS: Deep Thematic Analysis with Iterative LLM Support. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The growing scale of online datasets, such as social media posts, has created a pressing need for methods that can support efficient yet rigorous analysis [18]. Qualitative research methods such as thematic analysis (TA) [12] enable scholars to understand meanings, experiences, and perspectives, producing rich insights into complex social phenomena (e.g., [30, 76]) and situating findings within broader cultural or situational contexts [25]. These approaches are particularly valuable for describing behaviour, exploring lived experience, and identifying concepts or variables that might otherwise be overlooked [69, 73]. However, despite these strengths, TA is time- and labour-intensive, and scaling it to large corpora presents practical challenges for consistency and depth [14, 42, 47].

Importantly, TA is not a single technique but a family of approaches [25]. Reflexive Thematic Analysis (RTA), developed by Braun and Clarke, emphasizes the researcher's interpretive role, the evolving and flexible nature of codes, and reflexivity as a way to acknowledge how analytic decisions shape findings [3, 13]. In contrast, codebook TA stabilizes codes into more structured frameworks, supporting consistency, rigour, and team-based analysis [10, 74]. Both approaches are widely used, but they carry different commitments to reflexivity, flexibility, and replicability, which become particularly salient when scaling analysis to large datasets.

Large language models (LLMs) have recently been proposed as tools to help meet these scaling challenges. Their ability to generate, classify, and summarize text makes them promising assistants for coding and theme development [29, 86]. Yet existing AI-assisted qualitative tools often produce outputs that are generic or decontextualized, require extensive refinement, and operate through opaque processes that undermine trust [16, 33, 47]. Analysts frequently prefer manual methods when reflexivity and interpretive authority are central [13]. As a result, despite the promise of LLMs, their adoption in qualitative research remains limited, with open questions around how to balance efficiency with transparency, trust, and interpretive depth.

To address these challenges, we developed *DeTAILS: Deep Thematic Analysis with Iterative LLM Support*, a researcher-centred toolkit inspired by Braun and Clarke's six phases of analysis. Although originally motivated by RTA, the scale of Reddit datasets required design choices such as editable but stabilized codebooks, which ultimately aligned the workflow more closely with codebook TA [10, 74]. Our evaluation with 18 qualitative researchers demonstrated both the potential and the tensions of this approach. DeTAILS substantially accelerated coding and theme development and was valued for its structured workflow and support for researcher control. At the same time, participants noted areas for improvement: reflexivity often centred on verifying AI outputs rather than interrogating analytic assumptions, novices at times over-trusted fluent outputs, and the workflow risked constraining interpretive flexibility. Through our own analysis of these interviews, using a codebook-style approach similar to what DeTAILS ultimately supported, we developed themes that informed a set of design recommendations for future iterations. These findings showcase the system's promise while also highlighting what needs to change in order to better align with reflexive thematic analysis.

This paper makes three contributions to HCI and qualitative research. First, it introduces DeTAILS, an interactive toolkit that integrates LLM support across thematic analysis phases while balancing acceleration with researcher control. Second, it provides a nuanced empirical account of how qualitative researchers of varying expertise engaged with LLM assistance, highlighting both efficiencies and tensions around agency, reflexivity, and trust. Third, it offers

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

design implications for the next generation of AI-assisted qualitative tools, including opt-in AI involvement, iterative workflows that support both forward and backward reflection, and safeguards against over-trust. In doing so, this work addresses current gaps in how AI systems are applied to qualitative research, where existing tools often privilege efficiency over reflexivity, provide opaque or decontextualized outputs, and leave unanswered questions about trust and adoption.

While inspired by RTA, our findings show that current LLM-supported workflows systematically fall short of supporting reflexivity, instead producing codebook-style practices. This paper therefore contributes not a system that achieves RTA, but an empirical account of why this gap persists and how future systems might address it.

2 Related Work

Thematic Analysis (TA) is a foundational qualitative method for identifying patterns of meaning ("themes") within textual data [2, 10, 12]. Reflexive TA (RTA), developed by Braun and Clarke [13], has become especially influential. In RTA, the researcher plays an active, interpretive role in theme development rather than applying a fixed codebook or pursuing inter-coder reliability [88]. Codes and themes are not pre-defined but emerge through close engagement with the data and ongoing critical reflection [12, 66, 78]. Reflexivity is central: researchers acknowledge their subjectivity and positionality, recognizing how analytic decisions shape knowledge production [56, 62]. This interpretive freedom allows for the generation of rich, situated insights but also demands rigour in documenting decisions and engaging critically with one's own role in the analysis [34, 56, 66].

These qualities made RTA an attractive starting point for our work. We sought to design a system that could preserve interpretive depth and reflexivity while helping researchers handle larger, more complex corpora such as social media data. Yet RTA is inherently time- and labour-intensive, particularly at scale [42, 64, 66]. Researchers must develop deep familiarity with their material and iteratively revisit analytic decisions, which becomes impractical with thousands of posts or transcripts [42, 66, 78]. Collaborative projects further add overhead as teams negotiate divergent perspectives and reconcile competing theme definitions [17, 42, 59]. Without large teams or extended timelines, analysts risk producing superficial themes or overlooking patterns in the data [14, 42, 89]. For example, platforms such as Reddit can generate corpora so vast that manual reflexive engagement is nearly impossible [7, 38].

Although computational tools exist for tasks such as text clustering or keyword extraction [8, 48, 65], few align with the interpretive and reflexive commitments of RTA [5, 51, 63]. Prior work on computational grounded theory and mixed-methods text analysis demonstrates how algorithmic techniques can surface large-scale patterns, but cautions that such tools often lack transparency and can encourage superficial engagement with data [5, 63, 65]. Very few systems combine algorithmic support with features that foreground researcher oversight, iterative refinement, and reflexivity. Our work was motivated by this gap: we aimed to explore how LLMs could be embedded within an interactive workflow that supports the interpretive and reflexive principles of RTA while making large-scale qualitative analysis more tractable.

2.1 Large Language Models as Qualitative Coding Assistants

Large language models (LLMs) are transformer-based neural networks trained on extensive corpora to generate, classify, and summarize text [32, 68, 99]. Because they produce fluent, contextually relevant prose and can respond conversationally, they appear well suited to tasks such as suggesting candidate codes or drafting theme summaries. These capabilities have prompted interest in computational tools that assist human analysts [11, 38, 41], and HCI researchers have begun to explore LLMs as qualitative coding assistants [26, 29, 31, 37, 43, 44, 85, 96].

Despite promising results, studies report notable limitations in both deductive and inductive coding. In deductive workflows, few-shot prompting helps align outputs with a predefined codebook but often yields overgeneralized or incorrect labels [23, 35, 52, 53, 60, 72, 81, 92]. For inductive coding, LLMs can quickly produce plausible code lists and theme summaries, yet their suggestions depend heavily on prompt design [28, 97] and still require analysts to verify, refine, and interpret the outputs [50]. Widely documented issues—including hallucinations, overconfidence, shortcut reasoning, and generic explanations—also risk undermining analytic rigour [4, 9, 20, 21, 36, 40, 54, 57, 67, 79, 83, 84, 87, 94]. These challenges underscore the need for careful human oversight [4].

To address these concerns, recent HCI work has proposed human-in-the-loop workflows that foreground researcher judgment [38, 39, 93]. Long et al. [57] advocate AI workflow design patterns that interleave automated suggestions with deliberate checkpoints where users vet and approve outputs. For example, CollabCoder embeds LLM suggestions within a collaborative coding interface designed for multi-analyst teams: researchers independently conduct open coding on the data, optionally accept AI-generated suggestions, and then reconcile their codes through system-supported discussion and consensus-building to create and group a shared codebook [36]. This design improves efficiency while preserving reliability, as the AI handles pattern detection and summarization while human analysts retain interpretive authority and negotiate meaning. Together, these studies show that LLMs can augment qualitative coding when embedded in workflows that deliberately preserve reflexivity and analytic rigour.

However, while systems like CollabCoder demonstrate the value of LLMs for collaborative, multi-analyst reconciliation, there remains a gap in supporting individual researchers conducting large-scale analysis independently. Existing tools often focus narrowly on specific stages like code generation or inter-coder agreement, leaving the full spectrum of concept mapping, iterative theme development, and rigorous evidence verification to the analyst alone. In particular, ensuring the traceability of AI-generated themes back to verbatim data—without the risk of hallucinated quotes—remains a critical challenge for individual workflows. Our work contributes to this space by designing an interactive system tailored for the

single researcher that scaffolds these activities from initial concept definition through code refinement and final theme synthesis. By using an editable codebook as a boundary object for recursive human–machine dialogue, and integrating explicit quote-verification safeguards, we aim to make large-scale qualitative analysis more tractable while sustaining the reflexive rigour emphasized in prior work [36, 57, 93].

2.2 Trust and Adoption of Automated QDA Tools

Trust in automated qualitative data analysis depends on researchers' confidence that computational tools operate transparently, fairly, and reliably. Large language models are trained on vast internet corpora, and their internal mechanisms are opaque; consequently, they may reproduce dominant-culture perspectives and misinterpret underrepresented voices [16, 20, 79]. When the analytic lens involves subtle cultural differences or sensitive topics, analysts worry that an AI might reinforce biases or overlook context, undermining confidence in machine-generated themes. This unease is compounded by the opaqueness of model predictions—researchers often cannot see why an LLM produced a specific code, making it difficult to judge whether the result reflects the data or the model's priors.

Empirical studies of social scientists illustrate a persistent tension: qualitative researchers appreciate the potential of LLMs to handle scale, yet they question the interpretability and trustworthiness of machine outputs [33, 47]. Many fear that subtle but important content will be missed or flattened, and that algorithmic bias could distort their findings. Practical barriers also play a role: numerous qualitative researchers lack programming expertise, leading them to default to manual methods —spreadsheets, memos or post-it notes — because these are familiar and under their control [51]. In other words, researchers often prefer intensive but trustworthy manual coding to an opaque tool whose outputs they cannot fully trust [13, 33, 47].

A consistent finding across recent evaluations is that LLMs are most effective as assistants rather than autonomous analysts [26, 29, 86]. Human oversight remains essential to ensure coding quality, contextual sensitivity and trustworthiness [20, 47, 49, 79, 80]. For example, Dai et al. [26] report that while an LLM could suggest reasonable open codes from raw text, human reviewers had to merge redundancies, correct misinterpretations and supply missing context. Similarly, Tai et al. [86] found that GPT-4 could classify text into high-level categories with high recall, yet it sometimes confidently produced incorrect justifications that only a human could catch. Deiner et al. [29] judged LLM-generated themes as often plausible but noted that their depth and specificity seldom matched human interpretations, with the model tending toward broad or generic themes. These outcomes underscore that LLMs can accelerate coding and pattern detection, but they do not replace human interpretive judgment.

Beyond interpretability, privacy and data-security concerns hamper adoption. Many commercial LLM services require uploading data to cloud servers, which creates ethical and legal risks when handling confidential interviews or personal histories. Perron et al. [70] highlight that proprietary cloud-based models "would almost certainly violate confidentiality" because whatever data are submitted can be stored, accessed or used for additional training without explicit consent, and providers rarely guarantee secure deletion. In response, researchers have begun exploring local or self-hosted LLMs that process data entirely on users' machines. For instance, a recent study in social work found that *local* LLMs enabled analysts to classify and extract information from sensitive child-welfare reports without transmitting any information to external servers. Such local models address confidentiality concerns while delivering performance comparable to proprietary systems, suggesting a promising path forward.

Despite growing interest, adoption of AI-assisted qualitative analysis remains limited. Existing tools seldom support the transparency, reflexivity and data privacy that qualitative researchers expect, leaving scholars to favour manual coding over opaque automated systems [33, 47]. Our work addresses this gap by designing a locally deployable system that embeds LLM suggestions within a transparent, stepwise workflow. Analysts can inspect, edit and contextualize model outputs at each stage, maintaining control over their data and mitigating bias, misinterpretation and confidentiality risks. By foregrounding researcher agency and reflexive rigour, we hope to increase confidence in—and adoption of—automated QDA support.

3 DeTAILS: Deep Thematic Analysis with Iterative LLM Support

DeTAILS is a self-contained, cross-platform desktop application that supports thematic analysis of Reddit data. Its design is inspired by Braun and Clarke's six-phase reflexive TA framework [14], but intentionally adapts it into a semi-structured workflow mediated by AI support. In particular, DeTAILS uses a *codebook as a boundary object* between the researcher and the LLM. While reflexive TA resists fixed codebooks, in practice a provisional, editable codebook provided a shared structure for surfacing, editing, and contesting LLM suggestions. This design choice reflects strategies often associated with codebook approaches [10, 74], while aiming to retain the flexibility and reflexivity central to RTA.

Researchers begin by loading data and generating an initial codebook from a subset of the dataset through human–LLM collaboration. This mirrors established TA practices where an initial subset guides broader coding [75], while making the boundary between human and AI contributions explicit. DeTAILS then applies the evolving codebook to the remaining dataset and clusters codes into suggested themes, but enables researchers to iteratively reflect on, refine, and build themes using those codes.

Importantly, researchers may revisit and change any previous phase of their work, and DeTAILS automatically propagates those changes forward. For instance, a researcher may add or remove codes, redefine them, or change themes and associated codes at any point—and DeTAILS will automatically update the corresponding applications to the corpus and downstream themes. This workflow is illustrated in Figure 1.

DeTAILS comprises a graphical front end and an analysis back end. The front end, implemented in Electron with a React UI, provides interactive interfaces for codebook creation, clustering, and

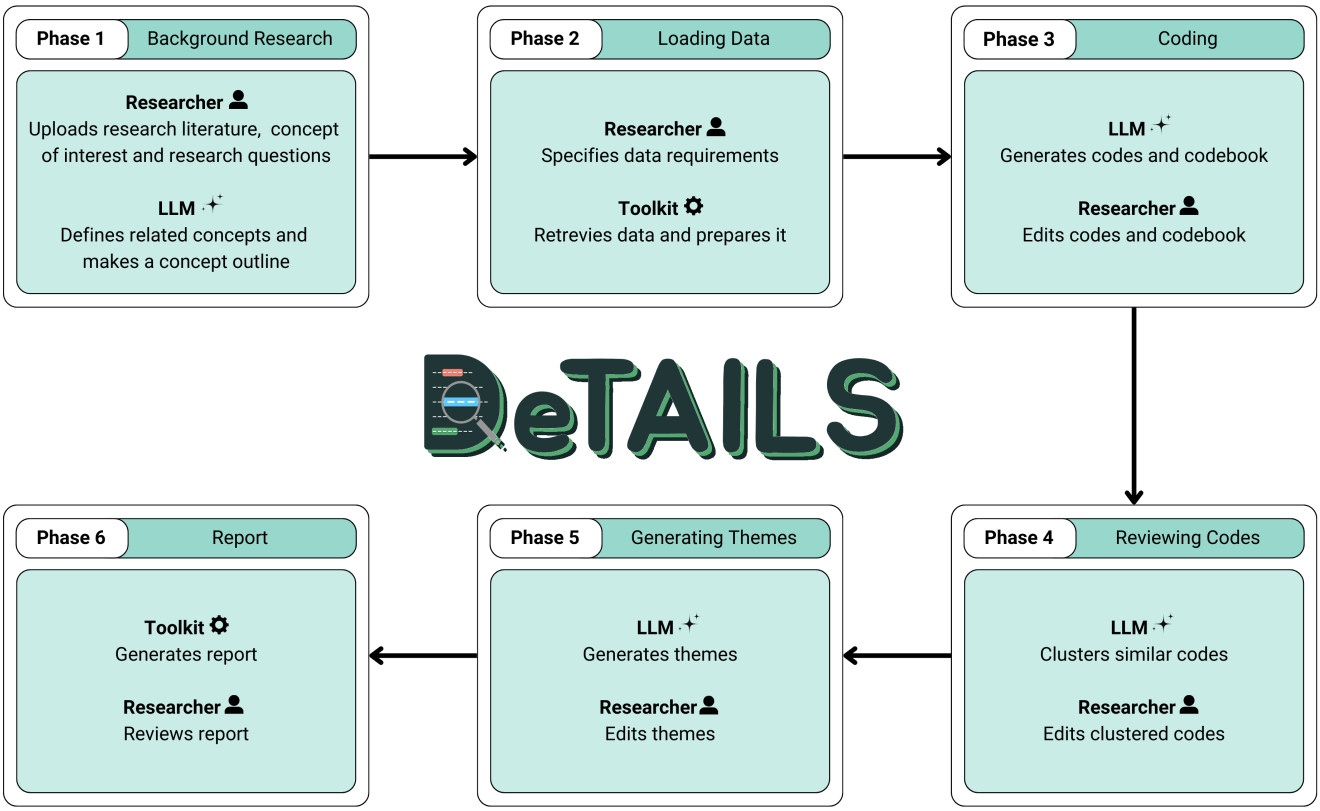


Figure 1: The DeTAILS workflow comprises six phases inspired by Braun and Clarke's Reflexive Thematic Analysis. Each phase requires collaboration between the researcher and LLM.

theme refinement. The back end integrates a Python analysis engine with a local vector database (ChromaDB) and an LLM runtime (Ollama) to support both remote APIs and local models. A modular orchestration layer built on LangChain enables interchangeable use of four providers (OpenAI, Google Vertex AI, Google AI Studio, and local Ollama models). Each phase of the workflow is guided by structured prompts that define LLM behaviour and output format. This modular design allowed us to balance transparency, replicability, and flexibility across different model providers. Full source code is available on GitHub.

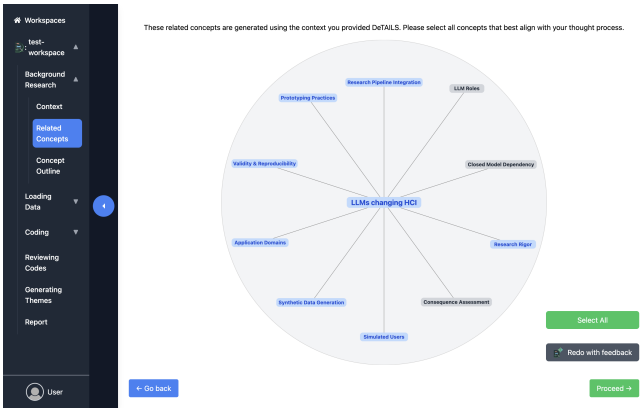
To illustrate the workflow, we now demonstrate how DeTAILS can be used to analyze posts from *r/UXResearch*. This example mirrors one conducted by study participants and highlights how the toolkit scaffolds key phases of thematic analysis at scale.

3.1 Phase 1: Background Research

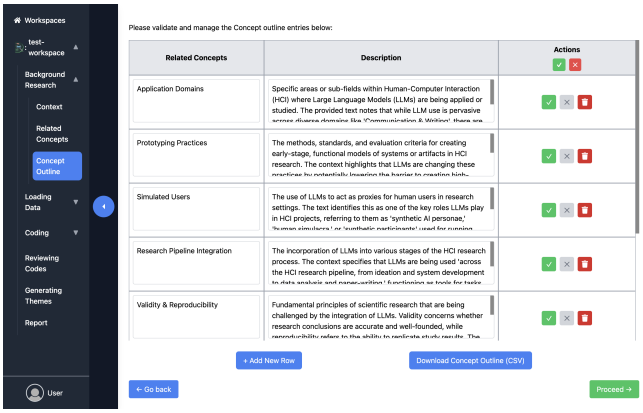
Qualitative researchers begin a thematic analysis by thoroughly understanding existing research and the broader context of their research topic [42, 66]. DeTAILS embeds this process in its workflow through stored contextual information — such as research questions and relevant literature — provided by the researcher. Researchers provide this information by uploading related documents, like relevant research papers, and through structured and free-form text entry. The information is then processed and indexed into a ChromaDB vector database, facilitating Retrieval-Augmented Generation (RAG) so that later LLM responses are grounded in theories from the researcher's sources.

DeTAILs then suggests a set of related concepts based on the provided contextual information. These concepts are used as inputs to develop an initial codebook, but also provide the researcher an opportunity to reflect on the direction their analysis is taking, and to fine-tune. They are developed through a two-phase process: first researchers select concept labels (Figure 2a) from a radial map, second DeTAILS generates an editable 'Concept Outline' list (Figure 2b) that includes their definitions. Once the researcher approves of these concepts, DeTAILS uses them to generate an initial codebook for anlaysis.

For instance, in the context of an analysis of "LLMs Changing HCI" in discussions on *r/UXResearch*, the initial set of related concepts based on related literature and research questions might include concepts like 'Simulated Users', 'Synthetic Data Generation', or 'Validity & Rigour'. In this initial pass, the researcher might choose to exclude concepts that fall outside of their research interests like 'Closed Model Dependency'. DeTAILS then generates a concept outline with more in-depth descriptions for each concept of interest.



(a) Background Research: Researchers provide context (e.g., research questions and related documents), and DeTAILS suggests related concepts via an interactive radial map to guide analysis.

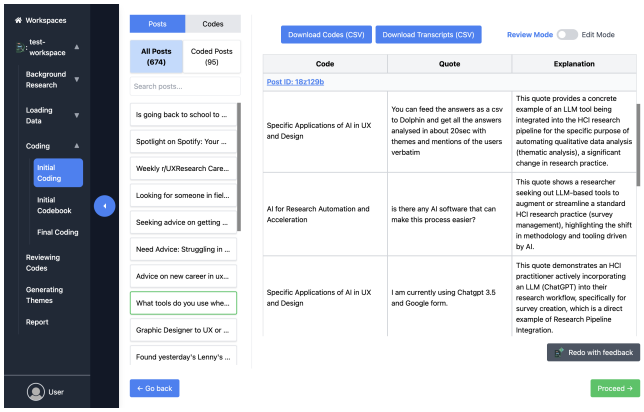


(b) Concept Outline: Researchers refine and finalize a table of key concepts and definitions, which form the foundation for generating an initial codebook in later phases.

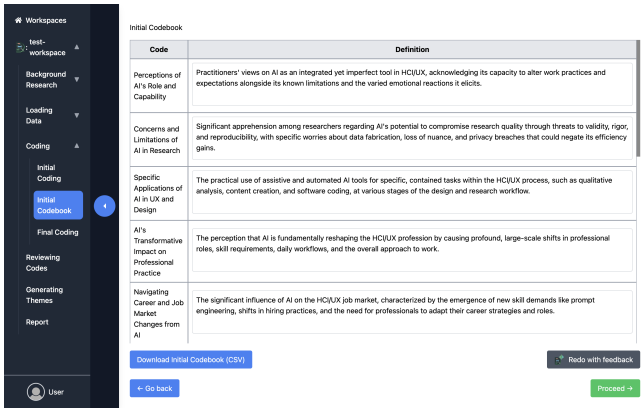
Figure 2: Background Research: DeTAILS supports early-stage analysis by surfacing related concepts and building a structured concept outline, providing a grounded starting point for subsequent coding.

3.2 Phase 2: Loading Data

DeTAILS can load subreddit data via academic torrent dumps [82, 90] as well as text files from local disk. It has built-in support to load torrent files via a user's native Transmission client. It can download the files, decompress them, and filter based on the subreddit of interest. Since monthly Reddit dumps are mostly made up of data that is not of interest to the researcher, DeTAILS also provides a means to work with local, pre-filtered NDJSON files. Once loaded, the dataset can be filtered by adjusting date ranges, removing empty posts, or by keyword search. DeTAILS can similarly load data from locally provided plain text files, for instance those collected during interviews. While not supported at this time, we suggest that DeTAILS can be easily extended to work with other data sources. For instance, Reddit's upcoming API, or other social media platforms like Threads.



(a) Initial Coding: DeTAILS applies LLM-generated codes to a subset of the data, providing a first-pass categorization.



(b) Initial Codebook: Researchers review and refine the LLM-suggested codebook before applying it to the full dataset.

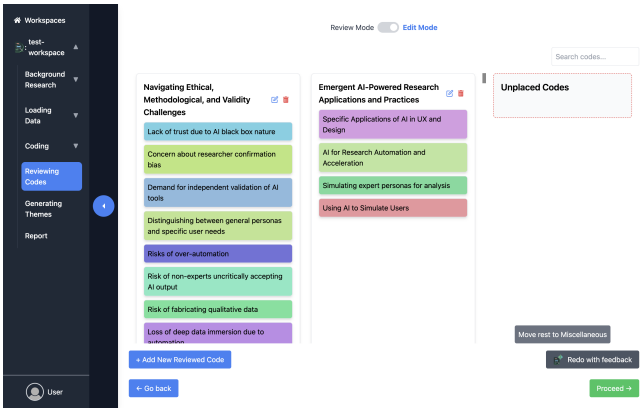
Figure 3: Coding Phase: LLMs generate preliminary codes, which are then refined by researchers into an initial codebook to guide analysis across the corpus.

3.3 Phase 3: Coding

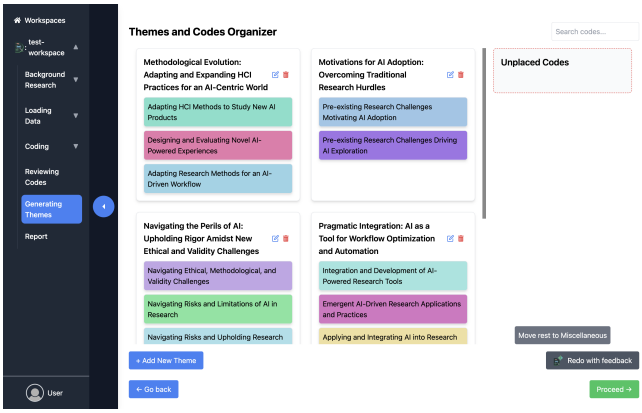
In this Phase the researcher and DeTAILS work together to collaboratively define a codebook through three activities:

DeTAILS first uses the provided background information and research questions to create codes for a split of the data to create an initial codebook (Figure 3a). The researcher can then view these codes and their application to the data through example posts and quotes, and revise them as they wish. The use of LLMs to generate this initial codebook is a key difference from previous work that relied on existing codes and examplar data for automated coding. Instead, we use these LLM-generated codes to bootstrap the rest of the coding process.

Second, the researcher can then manually edit this codebook directly (Figure 3b) by adjusting the code labels and definitions. DeTAILS provides the codebook in a plain text format, and the researcher can directly edit the codes by modifying their descriptive text.



(a) Reviewing Codes: DeTAILS clusters semantically similar codes into higher-level categories. Researchers can refine these clusters by editing definitions, merging, or reorganizing codes, with changes applied across the corpus.



(b) Generating Themes: DeTAILS proposes overarching themes from code clusters, presented in an interactive bucket interface. Researchers can adjust, merge, or create new themes to reflect their analytic lens.

Figure 4: Phases 4 and 5: Reviewing Codes and Generating Themes. These phases support higher-order interpretation, allowing researchers to refine code clusters and develop coherent, researcher-driven themes.

Third, the researcher can prompt DeTAILS to apply the revised codebook to the remainder of the corpus. We refer to this work as *global coding*. As in the initial coding stage, DeTAILS provides a tabular interface through which the researcher can again view how the codes have been applied through a list of codes and their associated text. Researchers can also view each post individually and edit existing codes or add new codes to it.

Importantly, DeTAILS ensures the correctness of LLM-generated outputs by verifying each quote and code during each of these steps. After each LLM call DeTAILS cross-checks each quoted text segment against the original Reddit post to confirm that it appears verbatim. Any quote that cannot be found in the source text is discarded as a likely hallucination.

3.4 Phase 4: Reviewing Codes

In this stage, all codes and their corresponding explanations are batched and sent to the LLM for clustering. The model groups semantically similar codes and assigns each cluster a consolidated 'reviewed code'. These clusters are then visually presented via buckets, illustrating their relationships (Figure 4a). Researchers can edit the visualized clusters by dragging them between blocks or create, rename, and delete buckets. They can also fine-tune the clusters through a 'Redo with feedback' function that uses a text-based prompt for the LLM. Once the reviewed code groupings are finalized, an alternate view displays these codes alongside their associated quotes and explanations, maintaining a clear link to the source data.

3.5 Phase 5: Generating Themes

The final analytical step involves generating overarching themes from the reviewed codes. DeTAILS clusters all codes into a set of potential themes for the corpus. Similar to the Reviewing Codes phase (Figure 4a), these suggested themes are presented via an interactive 'bucket' interface, where each theme contains a mutually exclusive set of affiliated codes (Figure 4b). The researcher can drag codes between themes, or create new themes, merge or rename them. DeTAILS also provides an LLM-driven refinement option, 'Redo with feedback' function, where researchers can explain through a text-based prompt how they would like the themes to change.

3.6 Phase 6: Report

Once themes are finalized, DeTAILS assists in compiling and exporting the findings. The report interface provides options to organize the report, for instance, by 'Theme & Code' or 'Post by Post'. Unlike most current tools that limit data downloads to their proprietary formats, DeTAILS enables exports of the final, structured report in the widely compatible CSV format, suitable for sharing, further analysis in other tools, or direct incorporation into publications. Although the final write-up of the analysis still remains a manual process, DeTAILS generates a structured report — segmenting each result into Code-Quote-Explanation sections — that researchers can quickly review.

4 Methodology

We conducted a mixed-methods study to examine how qualitative researchers engaged with DeTAILS when applying thematic analysis to their own data. Our goal was to understand not only the tool's usability and functionality, but also how it supported (or constrained) researcher agency, reflexivity, and analytic practice. Participants provided informed consent, and the study received ethics clearance from our institution's Research Ethics Board.

4.1 Participants and Recruitment

We recruited 18 qualitative researchers with varying levels of experience and disciplinary backgrounds in thematic analysis (Table 1). To capture a range of perspectives, we used purposive sampling with inclusion criteria of prior involvement in thematic analysis and qualitative coding. Participants were categorized by TA experience as novice (less than 1 year experience, n=6), proficient (1–4

years experience, n=6), or expert (more than 4 years experience, n=6). This stratification ensured feedback from both newcomers and seasoned practitioners.

Participants were recruited via direct email invitations using academic and professional networks. Prior to their session, each participant was asked to send their research questions, 2–5 representative research papers related to the questions, and the name of a subreddit they were familiar with and interested in analyzing.

4.2 Materials and Apparatus

We focused on Reddit data given its heterogeneity and scale [71], which has led to its use in numerous HCI studies (e.g., [19, 46, 55, 75]). Working with Reddit data allowed us to evaluate DeTAILS on large-scale, noisy real-world text, while also enabling each participant to explore content relevant to their interests. We pre-loaded each participant's chosen subreddit data into DeTAILS, downloading all posts and comments from the last full month [90] and loading a new workspace for their session. Each participant analyzed 30 posts drawn from their own subreddit; posts varied in length, reflecting the heterogeneity of real Reddit content rather than a controlled stimulus set.

Sessions were conducted using a MacBook Air with an Apple M1 chip (8-core CPU, 7-core GPU) and 16 GB RAM. This machine ran DeTAILS with the `gemini-2.5-pro-preview-03-25` LLM via the Google Vertex AI API. We chose this configuration to evaluate the workflow with a state-of-the-art model, isolating questions about the human–LLM interaction design from those about model capability. We acknowledge that this choice is in tension with the privacy motivations described in Section 2.2, where we argued that local LLMs are important for handling sensitive qualitative data. DeTAILS' modular orchestration layer (Section 3) supports local models via Ollama precisely so that researchers working with confidential data can deploy the same workflow without transmitting data to external services; we return to this tension and its implications in Section 7. The full text of all system prompts used is provided in Appendix A.

Participants who attended in person used this laptop directly. For remote participants, we utilized Microsoft Teams' screen-sharing and remote control functionality: the author shared their screen running DeTAILS, and the participant was granted control to operate the interface virtually. This setup allowed online participants to fully interact with the toolkit (e.g., clicking buttons, highlighting text, editing code labels) as if they were seated at the machine. Audio and video were streamed via Teams for communication. Two participants opted for in-person sessions in our lab, while the remaining 16 participated remotely via Teams.

4.3 Study Procedure

Each study session lasted between 2 and 2.25 hours and followed a structured sequence: an initial semi-structured interview, a guided walkthrough of DeTAILS, an interactive analysis task using DeTAILS, a follow-up exploration of large-scale results, and post-study questionnaires.

The central activity was having participants use DeTAILS to conduct thematic analysis on 30 discussion post transcripts from a subreddit of their choice (each transcript consisting of an original post and its comments). We selected a subset of 30 posts to ensure that interactive analysis was feasible within the session time while still providing sufficient material to generate meaningful codes and themes. Participants were asked to approach this analysis as if it were part of their own research. The task was structured according to the toolkit's workflow, with participants in control and the researcher observing, supporting, and prompting reflection when needed.

After completing the 30-post analysis, participants were introduced to pre-computed results for the full month of subreddit data (December 2024) that had been prepared in advance. They were able to browse and inspect outputs from each phase at this larger scale. This stage allowed participants to reflect on how the workflow operated across two scales, and to provide additional feedback on features, usability, and desired improvements. Participants could ask clarifying questions about the tool throughout.

The first author designed the study protocol in consultation with the third author and conducted all participant interviews.

4.4 Data Collection and Analysis

We collected both quantitative and qualitative data. Quantitative data included system logs of participant interactions (e.g., time per phase, edits, suggestion usage) and three post-study surveys (NASA-TLX, AttrakDiff, Perceived Usefulness). Statistical analyses were conducted by the first author, with consultation from the third author. We computed weighted precision, recall, and F1 scores using cosine similarity across Related Concepts, Concept Outline, Initial Coding, and Global Coding; semantic similarity for Initial Codebook definitions; and macro F1 metrics for Reviewing Codes and Generating Themes. To compute semantic similarity reliably across these text artifacts, we utilized Google's text-embedding-005 model to generate embeddings. Drawing on soft-evaluation frameworks established in natural language processing for similarity measures (such as BERTScore [98]), our cosine similarity calculations employed a continuous weighting approach rather than rigid binary thresholds. Specifically, the weighted true positive value for each comparison was incremented continuously by the cosine similarity of the corresponding text embeddings. Time across study phases was analyzed using repeated-measures ANOVA with Mauchly's test for sphericity; Greenhouse-Geisser corrections or Friedman tests were used when assumptions were violated. Post-hoc Wilcoxon signed-rank tests (Bonferroni-corrected) followed significant effects. For group comparisons (Novices, Proficients, Experts), assumption checks (Shapiro-Wilk, Levene's) guided the use of one-way ANOVA, Welch's ANOVA, or Kruskal-Wallis tests. Pairwise comparisons employed t-tests, Welch's t-tests, or Mann-Whitney U tests with Bonferroni-adjusted $\alpha = 0.0167$.

Qualitative data came from the post-session semi-structured interviews and conversations while the participant was using the system, which were transcribed verbatim. The second author, an expert in qualitative analysis, led the thematic analysis with input from the other authors. This analysis was conducted manually rather than in DeTAILS, but emulated the codebook-style workflow that the system itself supports. Codes and themes were iteratively refined over three rounds, a process resembling the recursive engagement of reflexive TA but operationalised in a more structured,

ID	Experience	Expertise	Field	Occupation	AI Used for TA	Subreddit	Date	Description
P1	<1 year	Novice	XR Gaming	Master's Student	ChatGPT	r/artificial	06/07/2008 – 31/12/2024	A subreddit dedicated to news, research, and discussions on artificial intelligence and its applications.
P2	<1 year	Novice	Explainable AI	PhD Candidate	ChatGPT	r/ChatGPTCoding	06/12/2022 – 31/12/2024	Community for ChatGPT code samples, usage tips, and creative AI-driven demos.
P3	1–2 years	Proficient	NLP Classification	Jr. Consultant	–	r/WeAreTheMusicMakers	09/09/2008 – 31/12/2024	Music production, composition techniques, and audio engineering discussions.
P4	2–3 years	Proficient	Human–AI Collaboration	PhD Candidate	Atlas.TI (AI features)	r/humanresources	29/05/2008 - 31/12/2024	HR best practices, talent management insights, and collaboration strategies.
P5	3–4 years	Proficient	Secure HCI	Master's Student	GPT-4	r/Chatbots	18/01/2009 – 31/12/2024	Chatbot development, conversational AI frameworks, and real-world use cases.
P6	5–6 years	Expert	Social Stigma	PhD Candidate	NVivo (AI features)	r/beyondthebump	29/04/2012 – 31/12/2024	Support for new parents navigating postpartum and early parenting challenges.
P7	<1 year	Novice	NLP Classification	Post-doc Researcher	Multiple LLMs	r/ArtificialInteligence	06/05/2016 – 31/12/2024	AI research, ethics debates, and industry application discussions.
P8	<1 year	Novice	Human–Robot Interaction	Post-doc Researcher	Multiple LLMs	r/CaregiverSupport	30/03/2013 – 31/12/2024	Advice and emotional support for family and professional caregivers.
P9	<1 year	Novice	Youth Smoking	PhD Candidate	ChatGPT	r/Canadian_ecigarette	23/10/2013 – 31/12/2024	Vaping products, health impacts, and regulation updates in Canada.
P10	<1 year	Novice	Healthcare Trust	Undergraduate Student	–	r/AskBalkans	05/03/2019 – 31/12/2024	Q&A on Balkan culture, travel tips, and current affairs.
P11	4–5 years	Expert	Accessible Gaming	PhD Candidate	–	r/disability	12/03/2008 – 31/12/2024	News, resources, and perspectives for individuals with disabilities.
P12	4–5 years	Expert	Data Integration	PhD Candidate	–	r/bigdata	05/01/2011 – 31/12/2024	Big data storage, processing pipelines, and predictive analytics.
P13	3–4 years	Proficient	Health Systems	PhD Candidate	–	r/ImmigrationCanada	10/02/2013 – 31/12/2024	Immigration advice, visa procedures, and settlement experiences.
P14	1–2 years	Proficient	Haptic UX	Product Analyst	ChatGPT	r/userexperience	19/06/2018 – 30/12/2024	UX design methods, usability testing, and prototyping best practices.
P15	12–13 years	Expert	Public Health	PhD Candidate	–	r/blackladies	16/11/2012 – 31/12/2024	Safe space for Black women to share, support, and celebrate achievements.
P16	3–4 years	Proficient	VR Analytics	PhD Candidate	ChatGPT	r/virtualreality	28/08/2011 – 31/12/2024	VR hardware, software platforms, and immersive design discussions.
P17	10–11 years	Expert	Dementia Technology	PhD Candidate	ChatGPT	r/dementia	10/01/2010 – 31/12/2024	Resources and support for those affected by dementia and their caregivers.
P18	11–12 years	Expert	Human Factors	Assistant Professor	–	r/smartwatch	21/01/2013 – 31/12/2024	Smartwatch tech, app development, and wearable computing trends.

Table 1: Demographics of Study Participants: Years of Experience, TA Expertise Level, Field of Specialization, Current Occupations, AI used for TA, Subreddit details: Name, Start Date - End Date, Description

codebook-oriented manner. This approach enabled us to capture patterns in participant reasoning, experience, and feedback across the phases of the workflow.

5 Quantitative Results

We report results on (1) alignment between DeTAILS-generated outputs and participants' refinements (F1 scores), (2) time spent across workflow phases, and (3) user-reported workload (NASA-TLX), perceived usefulness, and pragmatic/hedonic qualities (AttrakDiff).

5.1 F1 Scores

F1 scores increased consistently across phases (Figure 5). In Background Research, Related Concepts suggestions achieved a mean weighted F1 of 0.86 (SD = 0.13), improving to a median of 1.00 (IQR = 0.03) at Concept Outline. In the Coding phase, Initial Coding had a median F1 of 0.96 (IQR = 0.09), rising to a median of 0.99 (IQR = 0.05) in Global Coding. Reviewing Codes produced a median macro F1 of 1.00 (IQR = 0.14), and Generating Themes reached complete agreement (F1 = 1.00).

No significant expertise differences emerged. Related Concepts satisfied parametric assumptions, yielding a nonsignificant ANOVA ($F_{2,15} = 0.22$, $p = 0.8032$). Other phases violated normality but not homoscedasticity; Kruskal–Wallis tests were nonsignificant ($H_2 = 1.95$–5.82, $p = 0.0544$–0.7664). Generating Themes was excluded due to zero variance.

5.2 Time Analysis

Time analysis showed Initial Coding took significantly longer than other phases (Figure 6). Mauchly's test indicated violation of sphericity ($W = 0.197$, $p = 0.003$); applying Greenhouse–Geisser ($\varepsilon = 0.623$), repeated-measures ANOVA revealed a strong effect of Phase ($F_{2.49,42.37} = 22.77$, $p < 0.001$, $\eta_p^2 = 0.57$), corroborated by a Friedman test ($\chi_4^2 = 44.62$, $p < 0.001$). Post-hoc Wilcoxon tests (Bonferroni-corrected) showed Initial Coding was significantly longer than Related Concepts, Concept Outline, Initial Codebook, and Global Coding ($p \leq 0.004$). Global Coding also exceeded Initial Codebook ($p = 0.004$).

By expertise, only Related Concepts differed: Welch's ANOVA ($F_{2,9.26} = 6.51$, $p = 0.017$) showed Proficients (M = 5.37 min) spent more time than Novices (M = 2.49 min, $p = 0.004$). No other group differences were significant.

5.3 NASA-TLX

Participants reported low to moderate workload. We utilized the Raw TLX (RTLX) scoring method, calculating overall workload as an unweighted average of the subscales. We observed a mean of 39.02/100 (SD = 26.45) for mental demand, 37.09/100 (SD = 22.97) for effort, 28.70/100 (SD = 23.81) for temporal demand, and notably low frustration (Mdn = 7.62, IQR = 25.24). Self-rated performance was high (M = 72.94, SD = 14.93). Overall workload averaged 26.3/100 (SD = 12.4). We found no significant differences between expertise groups. One-way ANOVA, Welch's ANOVA, or Kruskal–Wallis tests found no effects for any subscales ($p = 0.1993$–0.8919).

5.4 Perceived Usefulness Scale

Participants rated DeTAILS highly useful (Mdn = 4.00, IQR = 0.75 on a 5-point scale), with 86% of responses at 4 or 5. Strongest items were "accomplishes tasks more quickly" (Mdn = 4.50, IQR = 1.00) and "makes it easier to do my job" (Mdn = 4.00, IQR = 1.00). Fifteen of 18 participants rated at least four items as 4 or 5. Group comparisons revealed no significant differences. Only "Improves Performance" met parametric assumptions (M = 3.94, SD = 0.54), yielding a nonsignificant ANOVA ($F_{2,15} = 0.17$, $p = 0.8433$). Other items tested with Kruskal–Wallis were also nonsignificant ($p = 0.1137$–0.8679).

5.5 AttrakDiff

AttrakDiff ratings were overwhelmingly positive. Pragmatic qualities scored highly (e.g., structured: Mdn = 6.00, IQR = 1.00; straightforward: Mdn = 6.00, IQR = 0.00; simple: Mdn = 2.50, IQR = 1.00, where 1 = simple). Hedonic ratings were similarly strong (e.g., creative: Mdn = 6.00, IQR = 2.00; captivating: Mdn = 6.00, IQR = 0.75). No significant expertise effects were observed for most items ($p = 0.1196$–1.0000). A trend emerged for Simple/Complicated ($\chi_2^2 = 5.88$, $p = 0.0529$): novices (Mdn = 2.00, IQR = 0.75) perceived DeTAILS as simpler than experts (Mdn = 3.00, IQR = 0.75), with post-hoc tests suggesting a difference at corrected α.

6 Qualitative Results

Our analysis produced five overarching themes that capture how researchers navigated their interactions with DeTAILS: asserting agency, evaluating performance, negotiating trust, experiencing partnership, and reflecting on emotional and ethical implications. Within each theme, participants described both opportunities and challenges of integrating AI into qualitative workflows.

6.1 Asserting Researcher Agency & Control

Participants emphasized that DeTAILS did not replace their analytic authority but required active oversight. Two codes captured this theme. The first, *Agency Through Understanding AI's Limitations*, described how participants stressed the need for human interpretation. P7 explained, "It cannot be completely automated with an AI model. It needs some input… it is very difficult to completely automate," underscoring that analytic work requires subjectivity and nuance. The second code, *Verifying AI Output as an Act of Agency*, highlighted validation as a deliberate responsibility. P17 reflected, "I didn't agree with the explanation… the code was fine, but the quote I felt didn't align," showing how participants treated verification as an assertion of responsibility. Together, these codes show that agency was reasserted through recognition of AI's limits and critical oversight of its outputs.

6.2 Evaluating the Tool's Performance & Utility

Participants evaluated DeTAILS in terms of benefits and shortcomings, expressed across seven codes. The code *Accelerating the Workflow* captured how participants valued speed, with P1 observing, "This was definitely faster… what would take hours was done in less than 5 minutes." Yet participants also cautioned that speed did not guarantee better analysis, captured in the code *Critiquing Lack of Context or Nuance*. As P10 explained, "There's a big gap in

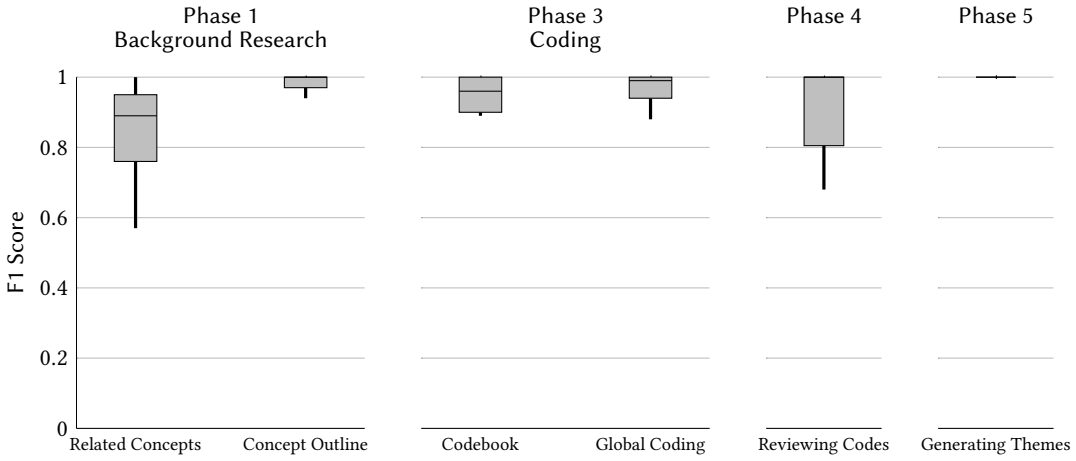


Figure 5: Box and whisker plot of F1 Scores across each phase of reflexive thematic analysis. Background Research and Coding are measured by a weighted F1, whereas Reviewing Codes and Generating Themes use a Macro F1.

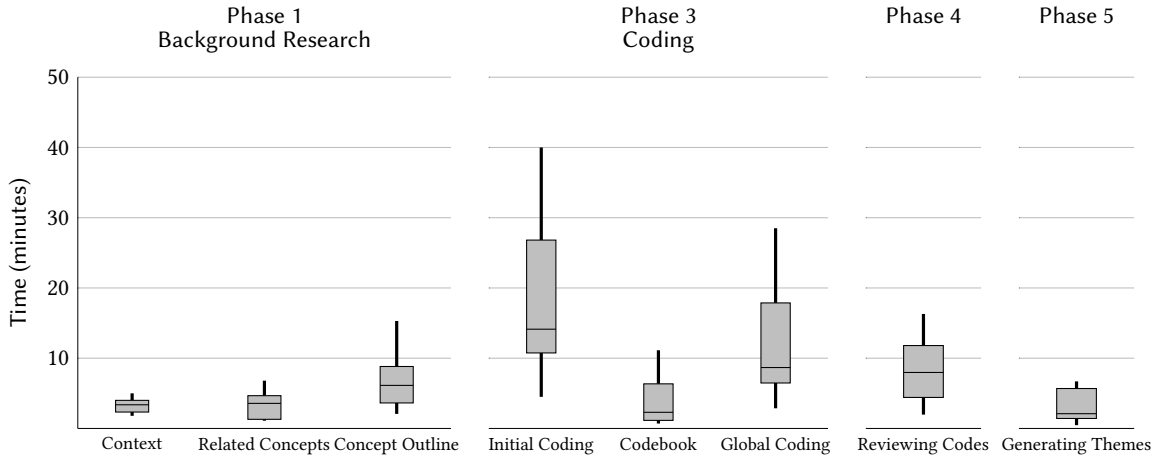


Figure 6: Box and whisker plot of time spent across each phase of reflexive thematic analysis.

what the LLM was able to do… the quotes were a bit less nuanced, a little less personal," illustrating how efficiency sometimes came at the cost of depth.

A second cluster of codes focused on comparisons and data issues. The code *Comparison with Existing Tools and Methods* highlighted how participants situated DeTAILS against their existing practices. P16 admitted, "I wasn't a huge fan of the initial coding… I would rather do it myself. Once I do some percentage… then maybe the AI can add." The code *Data-Specific Challenges and Observations* emphasized how limitations in the Reddit dataset shaped analytic outcomes. P12 noted, "When people are posting recent information, the number… is so low that it might not appear as a theme." These codes together show that the system's performance was tied not only to its design but also to the characteristics of the data.

The final set of codes pointed to usability and adoption. *Evaluating Tool Design and Usability* emphasized smooth interactions but also highlighted areas of friction; P18 appreciated visualizations but wished features were easier to find. *Future Use and Adoption* captured enthusiasm about integrating the tool into future projects, with P14 reflecting, "I enjoyed using it… I know it could have a very big impact, especially for researchers going through a lot of data." Finally, *Valuing High-Quality, Nuanced AI Contributions* described moments when participants felt the system added real analytic value. P15 noted, "The sophisticated language helped with the grouping in themes… it wasn't just like a 1+1=2." Together, these codes illustrate how participants weighed speed, usability, and nuance when assessing DeTAILS' overall utility.

6.3 Trust and Verification

Trust in DeTAILS was not automatic but had to be earned through explanation, alignment, and validation. Two codes captured this negotiation directly. *Building Trust Through Agreement and Disagreement* described how alignment with expectations could strengthen confidence, as P17 shared, "…this again verifies what I'm thinking

and also builds upon it." In contrast, *Distrust in Opaque Processes* pointed to skepticism when outputs felt shallow or overly accommodating. P4 remarked, "…sometimes I feel like the AI is trying to manage our expectations… so it's like false positives."

Three additional codes captured how participants sustained trust through ongoing oversight. *Researcher Validation as a Core Task* reflected how outputs were systematically checked, with P13 describing them as "probably accurate… to the few select number of posts we were asked to review." *Trust Through Transparency and Explainability* emphasized that explanations were a foundation for cautious trust, though they did not replace human judgment. As P11 explained, "If I was using this for a study, I would spend probably a week… really checking it." Finally, the code *Valuing AI Uncertainty to Support Judgment* described how participants sometimes treated ambiguous outputs as prompts for reflection rather than flaws. Together, these codes illustrate that trust in DeTAILS was relational and conditional, anchored in researcher validation and critical engagement.

6.4 The Human–AI Partnership

Participants also framed their work with DeTAILS as a kind of partnership, represented across five codes. *Framing the AI as a Teammate or Partner* captured how participants saw the system as collaborative, though never autonomous. P11 remarked, "If the LLM is my partner… I would still probably take the step of reading these things over." *Experiencing Dialogue and Co-Interpretation* described how outputs prompted reflection and refinement; P2 explained, "The requirement for prompting was really low… it really streamlines things," highlighting a conversational rhythm of feedback and adjustment.

The code *Positioning AI as a Second Opinion* captured how outputs were often treated as confirmatory checks, with P13 noting they were "probably accurate… to the few select number of posts we asked to review." *Recognizing the Boundaries of Partnership* emphasized that interpretive authority remained with humans, as P7 stressed, "It cannot be completely automated… it needs some input." Finally, *Negotiating Roles in Collaborative Analysis* captured how researchers actively shaped their inputs to direct the AI's contributions, with P12 reflecting, "…when you're looking for quality… that's what problem I need to change the questions a little bit." Together, these codes show that participants positioned DeTAILS as a supportive partner that could accelerate, provoke, or confirm analysis, but only within carefully defined boundaries.

6.5 The Reflexive & Emotional Experience

Participants reflected on the emotional and reflexive dimensions of working with DeTAILS, expressed across five codes. The code *Experiencing Surprise and Delight* captured moments when outputs exceeded expectations. P1 noted, "It brought out topics that I didn't even think it would… thinking about the future and understanding user demographics, that's a really good thing." In contrast, *Encountering Frustration or Skepticism* emphasized when outputs felt shallow or misleading, as P4 explained, "…sometimes I feel like the AI is trying to manage our expectations… so it's like false positives."

The code *Reflecting on Bias and Preconceptions* highlighted how participants became aware of problematic assumptions in AI systems. P11 recalled, "I was using ChatGPT… and it literally used the R slur in one of the suggested names," underscoring risks of bias in outputs. *Emotional Responses as Drivers of Reflexivity* described how emotions shaped reflexive engagement; P10 admitted, "I feel extremely lazy and I feel like I don't deserve to have this title of researcher, if I'm using an LLM to do my research for me," showing how discomfort prompted reflection on researcher identity. Finally, *Considering Ethical or Existential Implications* drew attention to risks around data integrity and representation. P17 warned, "If you forget to remove an identifier… I just worry about the safety of the data." Collectively, these codes highlight that working with AI was as much affective and ethical as it was technical, prompting participants to critically situate the tool within their research practice.

Across these five themes and their associated codes, participants portrayed their engagement with DeTAILS as a complex negotiation of efficiency, trust, control, partnership, and reflexivity. They valued the system's ability to accelerate tasks, provoke insights, and act as a supportive collaborator, while also emphasizing the need for human oversight, boundary-setting, and ethical responsibility. Their accounts suggest that integrating AI into qualitative research is not merely a technical matter but a relational and ethical practice in which analytic authority remains firmly with the researcher.

7 Discussion

DeTAILS was designed to explore whether LLMs could accelerate the labour of qualitative coding while maintaining researcher agency, reflexivity, and trust. Our evaluation showed substantial efficiency gains: participants completed datasets in about 33 minutes compared to the seven to eight hours typically required for manual analysis. Survey data confirmed perceptions of efficiency and usability, and log data indicated high alignment between model outputs and human refinements. At the same time, participants stressed the need for careful review and refinement, highlighting that analytic authority remained with the researcher.

Relative to prior systems such as NVivo auto-coding [58] or topic-modelling approaches [22, 61], DeTAILS offers both efficiency and interpretive alignment. High agreement scores and positive manageability ratings underscore how stepwise verification checkpoints scaffolded trust, leading participants to describe the AI as a "junior coder" or "second opinion." This reframes AI from a black-box generator to a collaborative partner. This also situates DeTAILS within broader debates about 'big qual' methods [18, 45], which seek to scale qualitative analysis without abandoning depth. Our results show that AI can accelerate labour while maintaining researcher engagement, extending these debates beyond descriptive approaches like topic modelling. Our contribution goes beyond scaling descriptively — it shows how interpretive engagement can also be scaffolded at scale.

7.1 Implications for Theory and Reflexivity

Although DeTAILS scaffolded reflexivity, it did not fully achieve the iterative, self-critical reflexivity central to RTA [13]. Log data showed that most participant time was spent reworking initial codes

rather than revisiting themes, and the system's forward-only propagation meant refinements at later stages could not cascade backwards. This design, coupled with participants' survey responses that they would "still read everything over," produced a workflow resembling codebook thematic analysis [10, 42]: reflexivity occurred through checking and correcting model outputs, rather than interrogating one's own analytic assumptions. That participants reported high efficiency and usability while still expressing caution — P11 estimated she would spend "a week checking" before trusting the outputs — illustrates how quant and qual results converge on this tension.

Quantitative indicators also highlight the limits of automation. Despite high F1 agreement, participants noted that quotes were "less nuanced" (P10) and that the outputs often lacked personal tone. This mismatch shows that statistical alignment does not equate to interpretive adequacy, echoing concerns that LLMs flatten complexity [24, 79]. AttrakDiff trends hinted at a "canny mountain" effect: novices were more likely to describe the system as "human" while experts were more sceptical, suggesting risks of over-trust in polished outputs. These findings underscore that some "grunt work" of qualitative analysis—iterative reading, grappling with ambiguity—remains irreducibly human and methodologically valuable [34, 66].

Taken together, our results suggest a division of labour. Quantitatively, DeTAILS provides strong evidence of efficiency, usability, and alignment; qualitatively, participants confirmed that the AI accelerated but did not replace analytic reasoning. This shows that LLMs may be well suited for scaffolding codebook-style workflows, while RTA's deeper reflexivity demands design features not yet realized. Rather than seeing this as a failure, we argue it clarifies the epistemological stakes: AI can support, but not substitute, the slow and interpretive work that gives qualitative research its depth [15, 91]. By explicitly delineating where automation adds value (speed, reliability, consistency) and where human labour must remain central (reflexivity, meaning-making, ethics), future tools can extend thematic analysis to larger datasets without eroding its methodological integrity.

Our findings also speak to theoretical debates about whether LLMs can meaningfully participate in qualitative inquiry. Reflexive thematic analysis (RTA) emphasizes researcher subjectivity, positionality, and iterative meaning-making as central to the analytic process [13, 34]. DeTAILS supported some of this reflexive work, but in ways that shifted its orientation: participants' reflexivity often focused on questioning the AI rather than questioning their own assumptions. This suggests that AI tools can scaffold certain forms of reflexivity, but risk narrowing it to verification unless deliberately designed to foreground positional reflection.

We therefore conclude that the role of AI in qualitative research should not be imagined as replacing human interpretation, but as scaffolding it [6]. By recognizing the stages where AI support is productive, the stages where human labour remains indispensable, and the theoretical commitments that must guide design, we can move toward tools that respect the epistemological richness of qualitative inquiry while extending its reach. Ultimately, our study shows that AI can extend — not diminish — the epistemic value of qualitative inquiry, but only when systems are designed to centre researcher reflexivity, ethics, and interpretive authority. In doing so, DeTAILS contributes to ongoing debates about how qualitative research can adapt to scale while preserving its interpretivist commitments.

7.2 Designing for Reflexive and Trustworthy AI Collaboration

Triangulating across interviews, logs, and surveys, it is clear that participants' trust grew not from automation alone but from being able to scrutinize, verify, and correct outputs. Quantitative manageability scores and agreement metrics bolster these accounts, showing that trust was scaffolded through transparency and editability. Yet to move closer to reflexive thematic analysis, systems must go beyond facilitating correction: they need to deliberately create space for reflexivity.

Classic criteria of trustworthiness—credibility, dependability, confirmability, and transferability [56]—help explain why this matters. High F1 scores and efficiency metrics provide evidence of dependability, but participants' concerns about nuance and voice remind us that credibility and confirmability still rest on human reflexivity, not machine alignment. Similarly, AttrakDiff scores showing strong manageability do not ensure that interpretive depth has been preserved. These findings underscore Morrow's [62] and Finlay's [34] view that reflexivity is not optional but foundational: researchers must continuously interrogate their own positionality and the power relations shaping analysis.

Our study showed that DeTAILS prompted reflexivity, but primarily about the AI's suggestions rather than researchers' own interpretive assumptions. This is not reflexivity in the RTA sense [13], but a more pragmatic checking of outputs. To foster genuine reflexive practice, tools must actively slow analysts down and foreground alternative interpretations. Backward propagation, reflexive memoing, structured pauses, positionality statements, and analytic audit trails [1] could all help shift reflexivity from verifying AI outputs toward interrogating the researcher's own assumptions.

These design imperatives also speak to broader quality debates. Tracy's [89] "big tent" criteria call for rich rigour, sincerity, and resonance. Efficiency gains and high agreement may contribute to rigour, but they risk overshadowing sincerity and resonance if outputs are too generic or sanitized. Rather than abandoning LLMs for these tensions, we argue for designing explicitly against them: embedding mechanisms that slow users, surface uncertainty, and preserve the interpretive labour—iterative reading, grappling with ambiguity, holding space for participants' voices — that cannot be automated without undermining qualitative inquiry.

7.3 Implications for Design

Our findings highlight key tensions between efficiency and reflexivity, trust and skepticism, and researcher agency and automation. While DeTAILS succeeded in accelerating thematic analysis and supporting transparency, participants' experiences also revealed limits: reflexivity often centred on verifying AI outputs rather than interrogating analytic assumptions, forward-only propagation constrained iteration, and polished outputs risked over-trust. Building on these insights, we translate our results into five design considerations for next-generation qualitative analysis systems.

Support Iterative and Bidirectional Workflows. DeTAILS reduced repetitive coding work but reinforced a codebook orientation by

only allowing edits to propagate forward. Reflexive thematic analysis, by contrast, requires recursive movement between codes, themes, and raw data [12, 13]. Future systems should support both forward and backward propagation so that refinements at later stages cascade to earlier ones. This would preserve the recursive character of reflexive analysis and scaffold more deliberate opportunities for reflection.

Make Uncertainty Visible. Quantitative results showed strong agreement scores (F1 up to 1.0) and low workload ratings, but participant interviews revealed skepticism and extensive time spent correcting outputs. This tension highlights the "canny mountain" problem, where AI outputs appear polished and thus invite over-trust, particularly for novices. Designers should resist over-automation and instead surface model uncertainty, flag low-confidence outputs, and provide scaffolds that encourage critical engagement [47, 80].

Give Researchers Control Over AI Involvement. Participants frequently described wanting to disable or bypass the AI when outputs felt intrusive or generic. Mandatory AI assistance risks displacing reflexivity onto model suggestions rather than the researcher's own analytic lens. Future tools should include opt-in/opt-out controls at each stage, clear provenance markers distinguishing human from AI contributions, and options to advance without invoking the model [24, 27].

Scaffold Reflexive Practice Deliberately. Our study showed that participants' reflexivity often centred on verifying the LLM's suggestions rather than interrogating their own assumptions. While this still prompted reflective engagement, it diverges from the deeper positional reflexivity emphasized by Braun and Clarke [13, 34]. Future systems should embed deliberate reflexive prompts—for example, space to record positionality statements, moments to reflect on analytic choices, or checkpoints asking why particular codes or themes matter. Such scaffolds could help ensure reflexivity remains researcher-driven rather than model-driven.

Adapt to Analytic Styles with Transparency. We observed wide variation in preferences for code granularity, theme phrasing, and merging strategies. Some participants valued fine-grained codebooks, others preferred higher-level clusters. Current systems, including DeTAILS, treat such preferences as static. Future tools should adapt over time to an analyst's style, but with safeguards: adaptation must remain transparent, reversible, and always under explicit researcher control. This balance would allow tools to reduce repetitive work while respecting the diversity of qualitative practices [66, 78].

8 Limitations and Future Work

Our study was short-term and conducted in a controlled setting, with participants analyzing small Reddit datasets during single sessions. Although many participants expressed interest in applying DeTAILS to their own projects, longer-term deployments are needed to understand how the toolkit integrates into ongoing research and how it shapes analytic rigour over time. Such work would also clarify whether our findings hold for richer qualitative materials such as interviews, field notes, or survey responses [25, 30], which may raise new requirements for preprocessing, navigation, and code review. These forms of data also demand deeper reflexivity than short sessions allow, making longitudinal studies particularly important.

We also did not compare DeTAILS directly with widely used qualitative analysis software such as NVivo or Atlas.ti, nor with simpler LLM workflows such as direct prompting in ChatGPT. Without such baselines, our efficiency observations (e.g., participants completing analyses in roughly 33 minutes) should be read as indicative rather than as established time savings, since the tasks are not directly comparable to a full manual analysis in scope or depth. A head-to-head comparison would clarify where an LLM-supported workflow adds the most value, and which features from traditional tools might be incorporated. Relatedly, we observed no statistically significant differences between novice and expert participants. This null result likely reflects the small sample and short study design, and should not be read as evidence that the system equalizes expertise. Prior work cautions that novices are especially vulnerable to over-trust in fluent AI outputs [36, 83]. Future research should therefore investigate how expertise interacts with LLM-supported analysis in sustained use, and how systems can provide scaffolds that support novices' engagement without encouraging uncritical acceptance.

Our implementation also explored only a narrow range of prompting strategies and model types. We relied on a general-purpose LLM (`gemini-2.5-pro-preview-03-25`) with a chain-of-thought workflow, without testing alternatives such as ReAct prompting [95], reasoning-intensive prompting [77], or domain-specific fine-tuned models [28]. LLM stochasticity further means that repeated runs on the same data may produce different initial codes and themes, complicating direct reproducibility; the human-in-the-loop checkpoints partially mitigate this by anchoring the codebook in researcher review, but two analysts using DeTAILS on the same corpus may still arrive at different codebooks. Pinning model versions, model temperature, logging prompts and model outputs (as DeTAILS does) supports partial reproducibility, but developing guidelines for pairing analytic tasks with appropriate models, and for aligning prompt strategies with qualitative epistemologies, remains important future work.

This study was also conducted in mid-2025, during a period of rapid advances in LLM capabilities. Some of the technical limitations we observed may be addressed quickly as models evolve, while new challenges will undoubtedly emerge. Nevertheless, we believe the core findings of this paper—around reflexivity, trust, agency, and the role of human interpretive labour—remain significant. They point to enduring questions about how qualitative research can engage with AI systems, and how design can ensure that these systems scale analysis without eroding its interpretive commitments.

9 Conclusion

We presented DeTAILS, a toolkit for human–LLM collaboration in thematic analysis. DeTAILS integrates LLM support across analytic phases while keeping the researcher in control through features such as persistent memory snapshots, redo-with-feedback loops, and fully inspectable codes and themes. These design choices foreground researcher judgment, promote transparency, and enable direct verification of outputs against raw data.

Our study with 18 qualitative researchers showed that DeTAILS substantially accelerated coding and theming without displacing interpretive depth. Participants completed analyses quickly, achieved high agreement with generated themes, and reported feeling "in the driver's seat" with the AI positioned as a supportive assistant. Trust was not assumed but earned through iterative checkpoints and editable outputs, highlighting how workflow design can foster agency and reflexivity.

This work makes three contributions: a toolkit that demonstrates how LLMs can scaffold large-scale analysis; an empirical study that reveals both benefits and tensions in human–AI collaboration; and design guidelines for future systems that balance acceleration with interpretive authority. More broadly, our findings show that LLMs can extend—but not replace—qualitative inquiry. By coupling computational scale with researcher reflexivity, DeTAILS illustrates how AI tools can amplify interpretive practice while preserving the epistemological commitments that make qualitative research valuable.

Acknowledgments

This work was supported by NSERC Discovery Grant RGPIN-2022-03268.

References

[1] Jos Akkermans and Sophie Kubasch. 2021. The Research Audit Trail: Methodological Guidance for Application in Practice. *Electronic Journal of Business Research Methods* 19, 1 (2021), 1–12. https://academic-publishing.org/index.php/ejbrm/article/view/2033

[2] Jennifer Attride-Stirling. 2001. Thematic networks: an analytic tool for qualitative research. *Qualitative Research* 1, 3 (2001), 385–405. arXiv:https://doi.org/10.1177/146879410100100307 doi:10.1177/146879410100100307

[3] Darshini Ayton. 2023. Thematic analysis. *Qualitative Research–a practical guide for health and social care researchers and practitioners* (2023).

[4] B. Baker, J. Hoffmann, et al. 2025. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. *arXiv preprint arXiv:2503.11926* (2025). https://arxiv.org/abs/2503.11926

[5] Eric PS Baumer, David Mimno, Shion Guha, Eirik Quan, and Geri Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. ACM, 1343–1357.

[6] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. arXiv:2001.08435 [cs.SI] https://arxiv.org/abs/2001.08435

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[10] Richard Boyatzis. 1998. Transforming Qualitative Information: Thematic Analysis and Code Development. (01 1998).

[11] Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of Topic Models. *Found. Trends Inf. Retr.* 11, 2–3 (July 2017), 143–296. doi:10.1561/1500000030

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (01 2006), 77–101. doi:10.1191/1478088706qp063oa

[13] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. arXiv:https://doi.org/10.1080/2159676X.2019.1628806 doi:10.1080/2159676X.2019.1628806

[14] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. doi:10.1080/14780887.2020.1769238

[15] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. arXiv:https://doi.org/10.1080/14780887.2020.1769238 doi:10.1080/14780887.2020.1769238

[16] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society* 3, 1 (2016), 2053951715622512. arXiv:https://doi.org/10.1177/2053951715622512 doi:10.1177/2053951715622512

[17] Ashley Castleberry and Amanda Nolen. 2018. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning* 10, 6 (2018), 807–815. doi:10.1016/j.cptl.2018.03.019

[18] Abinaya Chandrasekar, Sigrún Eyrúnardóttir Clark, Sam Martin, Samantha Vanderslott, Elaine C. Flores, David Aceituno, Phoebe Barnett, Cecilia Vindrola-Padros, and Norha Vera San Juan. 2024. Making the most of big qualitative datasets: a living systematic review of analysis methods. *Frontiers in Big Data* Volume 7 - 2024 (2024). doi:10.3389/fdata.2024.1455399

[19] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 32 (2018), 25 pages. doi:10.1145/3274301

[20] B. Chen, R. Patel, A. Lee, et al. 2024. Large Language Models in Qualitative Research: Can We Do the Data Justice? *arXiv preprint arXiv:2410.07362* (2024). https://arxiv.org/abs/2410.07362

[21] Y. Chen, X. Li, J. Kim, et al. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting. *arXiv preprint arXiv:2310.11324* (2023). https://arxiv.org/abs/2310.11324

[22] Yingying Chen, Zhao Peng, Sei-Hill Kim, and Chang Choi. 2023. What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures* 17 (01 2023), 1–20. doi:10.1080/19312458.2023.2167965

[23] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv preprint arXiv:2306.14924* (2023). https://arxiv.org/abs/2306.14924

[24] R. Chew, P. Singh, et al. 2024. Artificial Intelligence and Qualitative Research: The Promise and Perils of Large Language Model (LLM) 'Assistance'. *Poetics* 103 (2024), 101892.

[25] John W. Creswell and Cheryl N. Poth. 2016. *Qualitative Inquiry and Research Design: Choosing among Five Approaches* (4 ed.). Sage publications.

[26] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. (2023). arXiv:2310.15100 [cs.CL] https://arxiv.org/abs/2310.15100

[27] Danielle Hitch. 2024. Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research* 34, 7 (2024), 595–606. doi:10.1177/10497323231217392

[28] S. De Paoli. 2024. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of AI in Qualitative Research. *Social Science Computer Review* 42, 3 (2024), 678–695.

[29] M. S. Deiner, Y. Zhang, J. Williams, et al. 2024. Large Language Models Can Enable Inductive Thematic Analysis of a Social Media Corpus in a Single Prompt: Human Validation Study. *JMIR Infodemiology* 4, 1 (2024), e59641.

[30] Norman K Denzin and Yvonna S Lincoln. 2011. *The Sage handbook of qualitative research*. sage.

[31] J. Drápal, P. Novák, and P. Švec. 2023. Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies. SSRN Electronic Journal.

[32] T. Eloundou, P. Mohtashami, J. Wu, A. Askell, J. Clark, et al. 2024. Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196* (2024). https://arxiv.org/abs/2402.06196

[33] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (Oct. 2021), 25 pages. doi:10.1145/3479856

[34] Linda Finlay. 2002. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. *Qualitative Research* 2, 2 (2002), 209–230. arXiv:https://doi.org/10.1177/146879410200200205 doi:10.1177/146879410200200205

[35] Yasir Gamieldien, Jennifer M. Case, and Andrew Katz. 2023. Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4487768

[36] 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. arXiv:2304.07366 [cs.HC] https://arxiv.org/abs/2304.07366

[37] Jie Gao, Zhiyao Shu, and Shun Yi Yeo. 2025. MindCoder: Automated and Controllable Reasoning Chain in Qualitative Analysis. arXiv:2501.00775 [cs.HC] https://arxiv.org/abs/2501.00775

[38] Robert Gauthier and James Wallace. 2022. The Computational Thematic Analysis Toolkit. *Proceedings of the ACM on Human-Computer Interaction* 6 (01 2022), 1–15. doi:10.1145/3492844

[39] Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 17 pages. doi:10.1145/3491102.3502076

[40] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 362, 19 pages. doi:10.1145/3544548.3581352

[41] M. F. P. Gillies, Dhiraj Murthy, Harry Brenton, and Rapheal Olaniyan. 2022. Theme and Topic: How Qualitative Research and Topic Modeling Can Be Brought Together. *ArXiv* abs/2210.00707 (2022). https://api.semanticscholar.org/CorpusID:252683229

[42] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. 2012. *Applied Thematic Analysis.* SAGE Publications, Inc., Thousand Oaks, California. doi:10.4135/9781483384436

[43] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1040, 25 pages. doi:10.1145/3613904.3642834

[44] Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker Lee. 2024. Prompt-based and Fine-tuned GPT Models for Context- Dependent and -Independent Deductive Coding in Social Annotation. In *LAK '24*. 518–528.

[45] Michelle K. Jamieson, Gisela H. Govaart, and Madeleine Pownall. 2024. Making the most of big qualitative datasets: a living systematic review of analysis methods. *Frontiers in Big Data* 7 (2024), 1455399. doi:10.3389/fdata.2024.1455399

[46] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 192 (2019), 33 pages. doi:10.1145/3359294

[47] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (April 2021), 23 pages. doi:10.1145/3449168

[48] Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press.

[49] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2025. 'I'm Categorizing LLM as a Productivity Tool': Examining Ethics of LLM Use in HCI Research Practices. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW102 (May 2025), 26 pages. doi:10.1145/3711000

[50] Andrew Katz, Gabriella Coloyan Fleming, and Joyce Main. 2024. Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development. *arXiv preprint arXiv:2410.03721* (2024). https://arxiv.org/abs/2410.03721

[51] Harmanpreet Kaur, Xinyi Zhu, Shion Guha, Michael Muller, and Ben Shneiderman. 2024. Qualitative Researchers' Practices, Challenges, and Desires for AI Support in Thematic Analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–15.

[52] Elisabeth Kirsten, Annalina Buckmann, Abraham Mhaidli, and Steffen Becker. 2024. Decoding Complexity: Exploring Human–AI Concordance in Qualitative Coding. *arXiv preprint arXiv:2401.19275* (2024). https://arxiv.org/abs/2401.19275

[53] I. Klyukvin. 2024. Enhancing qualitative research in psychology with large language models: a methodological exploration and examples of simulations. *Qualitative Research in Psychology* 21, 2 (2024), 123–145.

[54] S. Koyejo and B. Li. 2023. *How Trustworthy Are Large Language Models Like GPT?* Technical Report. Stanford Human–Centered Artificial Intelligence. https://hai.stanford.edu/research/publications

[55] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3, Article 172 (2021), 28 pages. doi:10.1145/3432919

[56] Yvonna S. Lincoln, Egon G. Guba, and Joseph J. Pilotta. 1985. Naturalistic inquiry: Beverly Hills, CA: Sage Publications, 1985, 416 pp., $25.00 (Cloth). *International Journal of Intercultural Relations* 9, 4 (1985), 438–439. doi:10.1016/0147-1767(85)90062-8

[57] Tao Long, Katy Ilonka Gero, and Lydia B. Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow. In *DIS '24*. ACM, New York, NY, USA, 1–22. doi:10.1145/3609317.3625971

[58] Lumivero. 2025. Automatic coding techniques. https://help-nv.qsrinternational.com/15/win/Content/coding/automatic-coding-techniques.htm. Accessed: 2025-07-25.

[59] Catherine MacPhail, Nomhle Khoza, Laurie Abler, and Meghna Ranganathan. 2016. Process guidelines for establishing Intercoder Reliability in qualitative studies. *Qualitative Research* 16, 2 (2016), 198–212. arXiv:https://doi.org/10.1177/1468794115577012 doi:10.1177/1468794115577012

[60] Han Meng, Wuter Chefchanch, Chieh-Yang Huang, Ting-Hao Kenneth Huang, and Q. Vera Liao. 2024. Exploring the Potential of Human–LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma. *arXiv preprint arXiv:2405.05758* (2024). https://arxiv.org/abs/2405.05758

[61] Atsushi Miyaoka, Lauren Decker-Woodrow, Nancy Hartman, Barbara Booker, and Erin Ottmar. 2023. Emergent Coding and Topic Modeling: A Comparison of Two Qualitative Analysis Methods on Teacher Focus Group Data. *International Journal of Qualitative Methods* 22 (03 2023), 160940692311659. doi:10.1177/16094069231165950

[62] Susan L. Morrow. 2005. Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology* 52, 2 (2005), 250–260. doi:10.1037/0022-0167.52.2.250

[63] Michael Müller, Chris Mattson, and Ethan Kibler. 2016. Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing Companion (CSCW Companion)*. ACM, 541–545.

[64] James Nelson. 2017. Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qualitative Research* 17, 5 (2017), 554–570. doi:10.1177/1468794116679873

[65] Laura K. Nelson. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49, 1 (2020), 3–42.

[66] 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 (2017), 1609406917733847. doi:10.1177/1609406917733847

[67] Matthew Nyaaba, Min SungEun, Mary Abiswin Apam, Kwame Owoahene Acheampong, and Emmanuel Dwamena. 2025. Optimizing Generative AI's Accuracy and Transparency in Inductive Thematic Analysis: A Human-AI Comparison. arXiv:2503.16485 [cs.HC] https://arxiv.org/abs/2503.16485

[68] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023). https://arxiv.org/abs/2303.08774

[69] Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice.* Sage publications.

[70] Brian E. Perron, Hui Luan, Bryan G. Victor, Oliver Hiltz-Perron, and Joseph Ryan. 2024. Moving Beyond ChatGPT: Local Large Language Models (LLMs) and the Secure Analysis of Confidential Unstructured Text Data in Social Work Research. *Research on Social Work Practice* (2024). https://www.researchgate.net/publication/383174310 Preprint.

[71] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (2021). doi:10.1177/20563051211019004

[72] Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dan Calacci, and Andrés Monroy-Hernández. 2025. QuaLLM: An LLM-based Framework to Extract Quantitative Insights from Online Forums. *arXiv preprint arXiv:2405.05345* (2025). https://arxiv.org/abs/2405.05345

[73] Pranee Liamputtong Rice. 1996. In quality we trust! The role of qualitative data in health care. *Medical Principles and Practice* 5, 1 (1996), 51–57.

[74] Kate Roberts, Anthony Dowell, and Jing-Bao Nie. 2019. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC medical research methodology* 19, 1 (2019), 1–8.

[75] B Rotolo, G Bin Noon, H. H. Chen, and Z. A. Butt. 2023. Public Attitudes Towards Vaccine Passports in Alberta During the "Pandemic of the Unvaccinated": A Qualitative Analysis of Reddit Posts. *International Journal of Public Health* 68 (2023), 1606514. doi:10.3389/ijph.2023.1606514 Published December 22, 2023.

[76] Bobbi Rotolo, Natasha Janak Sanghadia, Ève Dubé, James R. Wallace, Janice E. Graham, Christine Huel, Shannon E. MacDonald, Terra Manca, and Samantha B. Meyer. 2026. Investigating attributes of stigma towards individuals who did not receive a COVID-19 vaccine: An analysis of user comment threads on Reddit conducted by the Canadian Immunization Research Network (CIRN). *Social Science & Medicine* 391 (2026), 118911. doi:10.1016/j.socscimed.2025.118911

[77] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* (2024).

[78] J. Saldana. 2021. *The Coding Manual for Qualitative Researchers.* SAGE Publications. https://books.google.ca/books?id=RwcVEAAAQBAJ

[79] J. Savelka, J. Mitrović, and S. Koleva. 2023. Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *arXiv preprint arXiv:2309.17147* (2023). https://arxiv.org/abs/2309.17147

[80] Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, and Sarita Schoenebeck. 2025. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 481, 17 pages. doi:10.1145/3706598.3713120

[81] Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and LuEttaMae Lawrence. 2024. The Role of Generative AI in Qualitative Research: GPT-4's Contributions to a Grounded Theory Analysis. In *Symposium on Learning, Design and Technology.* 17–25.

[82] RaiderBDev stuck_in_the_matrix, Watchful1. [n. d.]. Reddit comments/submissions 2005-06 to 2024-12. ([n. d.]).

[83] L. Sun, Y. Zhang, Y. Qian, et al. 2024. TrustLLM: Trustworthiness in Large Language Models. *arXiv preprint arXiv:2401.05561* (2024). https://arxiv.org/abs/2401.05561

[84] L. Sun, X. Zhao, Y. Qian, et al. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232* (2023). https://arxiv.org/abs/2311.05232

[85] Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T. P. Dobber, and Jos A. Bosch. 2024. Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study. In *LREC-COLING '24.* 5609–5621.

[86] R. H. Tai, S. Lin, X. Chen, et al. 2024. An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods* 23 (2024).

[87] K. Takagi, Y. Saito, and T. Yamada. 2023. Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks. *arXiv preprint arXiv:2305.08714* (2023). https://arxiv.org/abs/2305.08714

[88] Gareth Terry, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. Thematic analysis. 17-37 pages. https://uwe-repository.worktribe.com/output/888518 Have permission to upload AAM 24 months after publication..

[89] Sarah J. Tracy. 2010. Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research. *Qualitative Inquiry* 16, 10 (2010), 837–851. arXiv:https://doi.org/10.1177/1077800410383121 doi:10.1177/1077800410383121

[90] Watchful1. [n. d.]. Subreddit comments/submissions 2005-06 to 2024-12. ([n. d.]). https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/?

[91] Ryan Thomas Williams. 2024. Paradigm shifts: exploring AI's influence on qualitative inquiry and analysis. *Front. Res. Metr. Anal.* 9 (Dec. 2024), 1331589.

[92] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion).* Association for Computing Machinery, New York, NY, USA, 75–78. doi:10.1145/3581754.3584136

[93] W. Xu. 2023. AI in HCI Design and User Experience. In *Handbook of Human Factors in Web Design.* Springer, 1–20.

[94] Z. Xu, L. Chen, and J. Guo. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint arXiv:2401.11817* (2024). https://arxiv.org/abs/2401.11817

[95] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629* (2022).

[96] He Zhang, Chuhao Wu, Jingyi Xie, Jie Cai, John M. Carroll, and et al. 2025. Harnessing the Power of AI in Qualitative Research: Exploring, Using and Redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans* 100 (2025), 144. doi:10.1016/j.chb.2024.101144

[97] He Zhang, Chuhao Wu, Jingyi Xie, ChanMin Kim, and John M. Carroll. 2023. QualiGPT: GPT as an Easy-to-Use Tool for Qualitative Coding. *arXiv preprint arXiv:2310.07061* (2023). https://arxiv.org/abs/2310.07061

[98] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[99] Wayne Xin Zhao, Kun Zhou, Wenhan Huang, Meng Zhao, Zhen Li, Yuxuan Zhang, et al. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023). https://arxiv.org/abs/2303.18223