SQuID: Investigating Human-AI Collaboration through an Interactive Affinity Diagramming Tool

Yifan Wu
ywu32880@isi.edu
University of Southern California
Los Angeles, California, USA

Matthew Maillet
mcmaillet@umass.edu
University of Massachusetts
Amherst, Massachusetts, USA

Sebastian Favela
sebastian@tiltx.com
Smart Carrier
Clinton, New York, USA

Luíza Leschziner
lulesc@amazon.com
Amazon Inc.
New York, New York, USA

Lydia B Chilton
chilton@cs.columbia.edu
Columbia University
New York, New York, USA

Sarah Morrison-Smith
smorriso@hamilton.edu
Hamilton College
Clinton, New York, USA

Abstract

Qualitative data analysis is increasingly performed on large datasets, which poses challenges for traditional methods of interpretation. Artificial Intelligence systems can assist with data analysis in the initial stages, but questions remain about control and accountability. We present the Systematic Qualitative Information Diagrammer (SQuID), a mixed-initiative system that uses large language models (LLMs) to propose provisional groupings and hierarchies for affinity diagramming, which researchers can review. We evaluated SQuID in a laboratory study with 13 human-computer-interaction affinity diagrammers completing short, individual analysis tasks. We examined how participants interacted with SQuID during early-stage affinity diagramming, particularly with regard to accepting or rejecting AI-suggestions. Participants described reduced effort during repetitive tasks common in early structuring, alongside recurring concerns about how AI assistance could be justified to others. From these observations, we provide design implications for qualitative analysis tools that support qualitative data analysis while maintaining transparency and opportunities for human judgment.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**.

Keywords

Interactive Systems, Qualitative Analysis, Affinity Diagramming, Human-AI Collaboration

ACM Reference Format:
Yifan Wu, Matthew Maillet, Sebastian Favela, Luíza Leschziner, Lydia B Chilton, and Sarah Morrison-Smith. 2026. SQuID: Investigating Human-AI Collaboration through an Interactive Affinity Diagramming Tool. In *Proceedings of Graphics Interface 2026 (GI '26).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

GI '26, Kitchener-Waterloo, Ontario, Canada

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

1 Introduction

Qualitative research methods are challenged by the growing size of datasets. Traditional approaches emphasize human immersion in the data to develop accountability and nuanced insights, but the scale of contemporary datasets can make this process prohibitively time-consuming [4]. A central challenge is therefore how automation can assist humans while preserving accountability. Recent advances in large language models (LLMs) have renewed interest in human–AI collaboration for qualitative analysis tasks. LLMs are often used for early-stage data analysis, e.g., grouping or summarizing textual data. However, there remain concerns about transparency and methodological validity about the use of AI in qualitative data analysis. These concerns are especially relevant in human–computer interaction (HCI) and applied UX research, where qualitative findings often inform design decisions [5, 12, 14]. As a result, qualitative analysis serves as a useful context for examining how AI systems can support existing practices.

Affinity diagramming [54] is a widely used inductive technique for organizing unstructured qualitative data such as interview notes or field observations [4, 24, 40]. Researchers iteratively group data points and create labels, allowing categories to emerge from the data organically rather than from pre-conceived notions [59]. This process is central to affinity diagramming; researchers develop insight through repeated engagement with the data, making affinity diagramming both labor-intensive and difficult to automate [3, 7].

While recent AI and natural language processing (NLP) systems can support automated coding and clustering, they are often poorly suited for affinity diagramming. Commercial tools such as Mural [45], LucidChart [41], and Creately [11] support brainstorming and synthesis, but lack transparency into how groupings are produced and have little support for tracking analytic decisions. As a result, improved efficiency can come at the cost of accountability.

We developed the *Systematic Qualitative Information Diagrammer* (SQuID), a system that uses LLMs to propose provisional groupings for affinity diagramming and presents them through a sequence of explicit interaction checkpoints. SQuID is designed for HCI and applied UX researchers conducting early-stage qualitative analysis, where analysts often seek an initial organization of large datasets before conducting deeper interpretive work. Rather than producing fixed outputs, the system is designed such that humans interactively accept or reject AI groupings, treating AI-generated suggestions

as provisional and contestable. We evaluated SQuID in a mixed-methods laboratory study with 13 participants (0.1–7.0 years of affinity diagramming experience), who analyzed both a dataset we provided and a dataset from their own research. Through think-aloud protocols, interviews, and surveys, we examine how affinity diagrammers accept and reject AI-generated groupings when first encountering SQuID. We also elucidate design choices that shaped perceptions of control and responsibility.

The contributions of this paper include: (1) a workflow for AI-assisted affinity diagramming structured around small batches of data analysis and human-in-the-loop checkpoints, (2) empirical observations about how researchers engage with and question AI suggestions during the early stages of affinity diagramming, and (3) implications for the design of AI assistants for qualitative analysis.

2 Related Work

2.1 Affinity Diagramming

Affinity diagramming has been a widely used approach for qualitative data analysis for more than 30 years [40]. In affinity diagramming, qualitative data are physically arranged using a bottom-up approach to reveal relationships [40]. The iterative grouping process is a creative activity in which the physical movement of data and collaborative discussion play central roles in analysis. The affinity diagramming process can be time-consuming, often taking days or weeks for large datasets [24]. Digital adaptations rarely address this fundamental scalability challenge [24], motivating our search for methods to accelerate affinity diagramming without undermining the interpretive depth that makes it valuable.

Although affinity diagramming is often associated with physical practices, such as arranging sticky notes, it is increasingly performed digitally. Practitioners commonly use spreadsheets when working remotely or when maintaining a persistent, shareable document is important [26]. These spreadsheets can be edited directly or imported into graphical affinity diagramming tools such as Mural [45] or LucidChart [41], allowing researchers to move between spreadsheet-based and visual representations as needed. While digital formats sacrifice some of the tactile and spatial affordances that support in-person analysis [40], they are widely adopted in practice and provide a pragmatic intermediate representation for large or distributed qualitative projects. SQuID outputs affinity diagrams as a CSV file, designed to align with existing workflows that use spreadsheet software while supporting direct imports into visualization software for subsequent refinement.

2.2 Automated Qualitative Coding

Qualitative coding traditionally involves manually analyzing data through deductive (matching with existing codes), inductive (deriving themes from raw data), or mixed approaches. Although affinity diagramming and inductive coding developed separately, both are bottom-up analysis methods. In fact, affinity diagramming is often viewed as a visualized inductive coding process.

Automation has largely targeted deductive coding, where NLP, machine learning, and LLMs are used to apply existing codebooks to new data [9, 16, 21, 29, 43, 48, 49, 51, 58, 60, 61, 65]. While LLMs are not fully reliable, they often achieve reasonable accuracy with predefined codes [29, 34, 61]. In contrast, attempts to automate inductive coding remain limited. Early embedding-and-clustering methods, such as QuAD [24] could only group subsets of data and lacked labeling mechanisms. Later systems like QualiGPT [64] and ThemeViz [31] rely on LLMs for coding and theme development, offering very limited user interaction. Researchers have also explored LLMs for qualitative coding in fields like legal studies [15], clinical science [44, 50, 52, 62], hotel management [57], and psychology [55, 63]. Methods in this body of work were either too naive or too specialized for general use outside of their specific domains. However, they signaled the need for automated qualitative analysis across disciplines. Other recent work explored LLM-based inductive coding pipelines and hierarchical approaches (e.g., Dai et al. [13], Katz et al. [32], Lam et al. [34]), but these works primarily emphasize algorithmic design rather than user interaction. As a result, questions about how analysts engage with inductive automation remain under-explored.

2.3 Human-AI Collaboration in Data Analysis

Prior work on automated data analysis systems emphasizes the importance of user interpretation and trust [10]. When users cannot interpret AI decisions, they may develop miscalibrated trust in the model, either overtrusting or undertrusting its output. [56]. Recent research frames explanation as an interactive dialogue rather than a static output, emphasizing users' desire to request and contextualize explanations over time [17, 18, 37]. This perspective motivates SQuID's rationale-on-demand mechanism, which encourages researchers to question AI-generated groupings.

HCI research on mixed-initiative interaction and interactive machine learning similarly emphasizes collaboration over full automation [1, 19, 28]. Prior work frames iteration, transparency, and user control as essential for effective human-AI partnerships [1, 19, 28]. Subsequent work shows that analysts benefit from tools that provide checkpoints for inspection and correction, rather than relying solely on predictive accuracy [2, 33]. Within qualitative analysis, researchers are often willing to adopt AI for specific stages, such as accelerating early organization, but resist automation that undermines interpretive judgment or agency [20, 30]. Studies consistently find that LLMs better support deductive than inductive coding, reinforcing calls for systems that scaffold researcher analysis instead of replacing it [46, 53].

Recent systems such as CollabCoder and PaTAT demonstrate that interactive refinement can improve efficiency while preserving analytical rigor [21, 22]. Across this literature, a common theme is that the value of AI in qualitative analysis lies not in one-shot outputs, but in mechanisms for refinement and oversight. Importantly, prior work distinguishes between faithful explanations of a model's internal reasoning and post-hoc rationalizations that provide plausible, human-interpretable accounts of system behavior [18]. SQuID's rationale-on-demand mechanism is explicitly designed as a form of post hoc rationalization intended to prompt researchers' reflection rather than to reveal the underlying model's true internal decision-making. Because such rationales can appear coherent without being fully faithful, SQuID pairs them with traceability to underlying data, allowing researchers to verify and reinterpret AI-generated

groupings. In this way, explanations function as provisional analytic artifacts that support interactive transparency rather than authoritative accounts to be accepted uncritically.

Prior work shows that trust and user control matter in human–AI collaboration, but it says little about how these ideas specifically play out in inductive analysis, where structure emerges over time rather than being defined in advance. Our work fills a gap by providing a close look at how researchers actually use an AI-based system to make affinity diagrams and conduct inductive analysis, showing how questions about authority, control, and accountability can arise even during short, individual use.

3 System

The Systematic Qualitative Information Diagrammer (SQuID) is a Python-based prototype designed to support affinity diagramming with large language models (LLMs). Rather than treating clustering as a one-shot automated task, SQuID structures the diagramming process as an interactive workflow with explicit checkpoints that allow users to review, accept, or reject groupings as they emerge. The system prioritizes interaction and transparency over fully automated optimization, positioning the researcher as the decision-maker rather than the system. SQuID is open source at https://github.com/MoCHI-Research/SQuID and model-agnostic. During our studies, we used OpenAI's `GPT-4-0613` for grouping and `text-embedding-ada-002` for semantic similarity, although SQuID can be configured to use any models.

Table 1 compares SQuID with tools available on the market as well as workflows and interactive systems in the literature. The design choices of SQuID, such as identifying themes bottom-up and supporting multi-level data hierarchies, have been motivated and informed by affinity diagramming. Among the systems we surveyed, some apply highly automated workflows that need minimum or zero user input [31, 32, 34, 64], while others require extensive input to ensure user control and agency [13, 22]. SQuID explores the balance between user control and autonomy, offering a highly automated system that still allows users to review and decide codes.

3.1 Interaction-Centered Workflow

Figure 1 illustrates SQuID's workflow for a single pass. Data are processed incrementally, with users reviewing intermediate groupings after each batch. At each checkpoint, users can accept assignments they agree with or reject specific data-label pairings for reconsideration in later batches. This design foregrounds analyst judgment, allowing structure to emerge through iterative refinement rather than being deferred until the end of the pipeline.

Figure 2 shows SQuID's interface alongside a traditional affinity diagram for comparison. In conventional affinity diagramming (Figure 2, left), researchers physically or digitally arrange data points into labeled groups through direct manipulation. SQuID's interface (Figure 2, right) presents data points in a list on the left and AI-generated labels on the right, with toggle buttons allowing users to mark pairings for re-analysis. Data points marked for re-analysis are returned to the pool and reassigned in the next batch, preserving the iterative character of manual affinity diagramming while reducing the mechanical labor of initial organization. The interface is intentionally simple, prioritizing legibility of the data.

The workflow is designed to support gradual abstraction through visible intermediate structure. SQuID makes partial results available after each batch rather than presenting a completed diagram at the end, allowing analysts to develop familiarity with the emerging organization as they proceed. Exportable artifacts preserve the provenance of decisions across passes, supporting both downstream analysis and the kind of process reconstruction that qualitative researchers need when accounting for their analytic choices.

3.2 Batching and Prompting

Although modern LLMs support large context windows, we found that grouping performance degrades as the number of data points increases, even when token limits are not exceeded. Prior work [36, 38, 39] indicates this weakness of LLMs when responding to long inputs. When processing large datasets, LLMs tend to produce overgeneralized groups and to lose data points entirely. To address these limitations, SQuID processes data in batches to maintain grouping quality and responsiveness. We empirically determined a batch size of k=75 data points based on internal piloting to balance output coherence and latency for GPT-4-0613.

Data points are short text segments, each representing a single self-contained idea, observation, or relevant detail from the source material. For example, a data point could be a brief quote from an interview transcript or a condensed paraphrase of a field observation. Within each batch, data points are numbered and submitted to the LLM using one-shot prompting [8] to constrain output format:

```
system: "As an expert in affinity diagrams, group the
following data using as many groups as you find necessary.
Ensure each group has a group label that describes them.
Take care not to include one same data point in multiple
groups. Provide the groups in the format:
'Group 1: Label 1. 1, 2, 3
Group 2: Label 2. 4, 5, 6' "

user: "1. I like watermelons.
2. Do you like blueberries? I hate them!
3. My favorite food is cod.
4. I love apples.
5. Shrimp is gross."

assistant: "Group 1: Preferences for fruits. 1, 2, 4
Group 2: Preferences for seafood. 3, 5"

user: [Actual data to group]
```

Prompts are issued independently per batch without context from previous batches, meaning the LLM reasons over a local window rather than the full dataset. Data points the LLM omits from its grouping are caught by post-processing code and flagged as "Not Grouped"; like any other grouped item, they are then presented to the researcher for review, who may reject them back into the unprocessed pool. All rejected data and data that have yet to be processed by SQuID are combined using custom Python code and incrementally presented to the user as remaining batches. The merge step described in section 3.3 addresses label inconsistency that arises across independently processed batches.

Table 1: Comparison of SQuID with other systems.

System	Control & Input	Interaction Timing	Large Data Support	Data Hierarchy Support	Output Details	Output & Interpretability
Tools on the market						
LLMs (e.g., [47])	Prompt	Before & in the loop	Limited by model	Manual	Prompt-based	Prompt-based
Mural AI [45]	None	None	Limited by model	Manual	Visualized groups & themes	None
Workflows						
LLM-in-the-loop [13]	Labeled exemplar; human-AI discussion	In the loop	Limited by model	Two level limit	Two-level themes; higher-level coding	Human-AI discussion
LLooM [34]	None	None	Distill-cluster-synth workflow	Built-in	Themes; prevalence; description, score & example per theme	None
GATOS [32]	Optional hybrid	Before the loop	Embedding-based workflow	Two level limit	Two-level themes	Optional codebook
Interactive Systems						
PaTAT [22]	Manual annotation; accept/reject; label editing	In the loop	One entry at a time	None	AI theme recommendations	Rule synthesis; label prediction model
QualiGPT [64]	Theme #; prompt; format	Before the loop	Batching	None	Themes; description, entry count, sample quote	None
ThemeViz [31]	Theme #; prompt	Before the loop	Limited by model	None	Visualized groups & themes	None
This Work						
SQuID	Accept/reject pairings; label hierarchy; merging threshold	Before & in the loop	Batching-merging workflow	Built-in	Hierarchical themes for every data entry	Post hoc reason generation

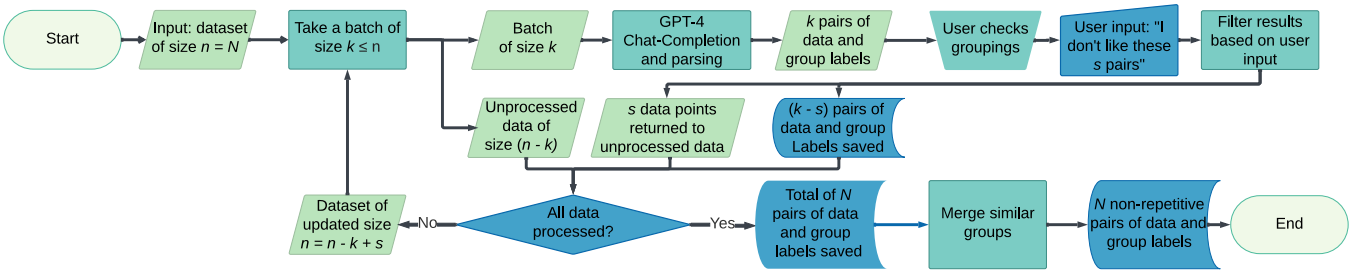

Figure 1: SQuID processes qualitative data in batches, facilitates user review and regrouping at intermediate checkpoints, merges semantically equivalent labels, and outputs structured groupings suitable for further passes or export.

3.3 Hybrid Merge and Hierarchical Structure

Because SQuID processes batches independently, the LLM sometimes generates semantically equivalent labels with different wordings across batches. SQuID includes a merge step to consolidate these at the end of each pass. It uses an embedding model (text-embedding-ada-002) to generate vector representations of all group labels, then applies the Girvan-Newman algorithm [23] to identify communities of semantically similar labels by treating cosine similarity between vectors as edge weights. The threshold edge weight used for merging is configurable by the researcher, with a default value of 0.91 for GPT-4, determined empirically through internal piloting using multiple sample datasets. Lower values produce fewer, broader groups, and higher values preserve more distinctions between labels.

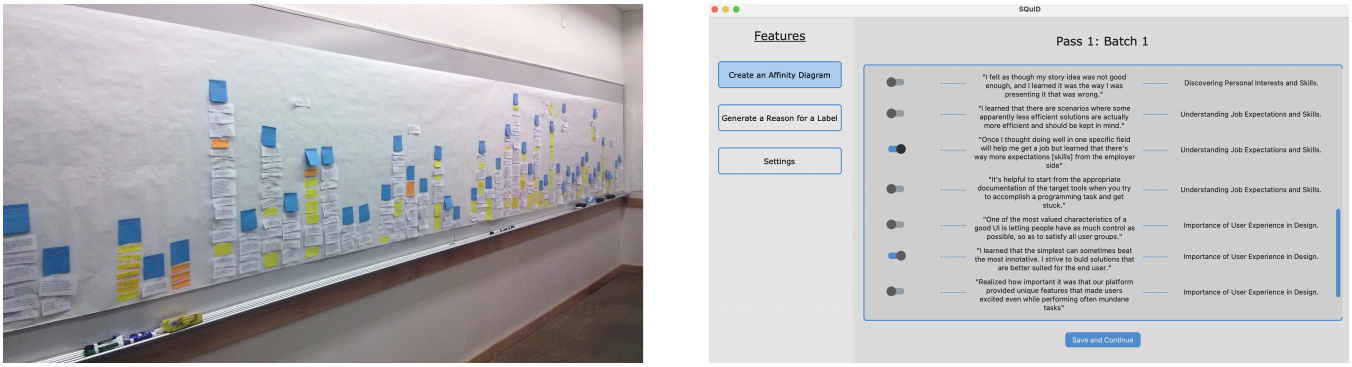

Figure 2: (Left) A traditional affinity diagram with data points organized into labeled groups. (Right) SQuID's diagramming interface showing AI-generated labels on the right and corresponding data points on the left. Toggle buttons allow users to remove data-label pairings for reassignment in subsequent batches.

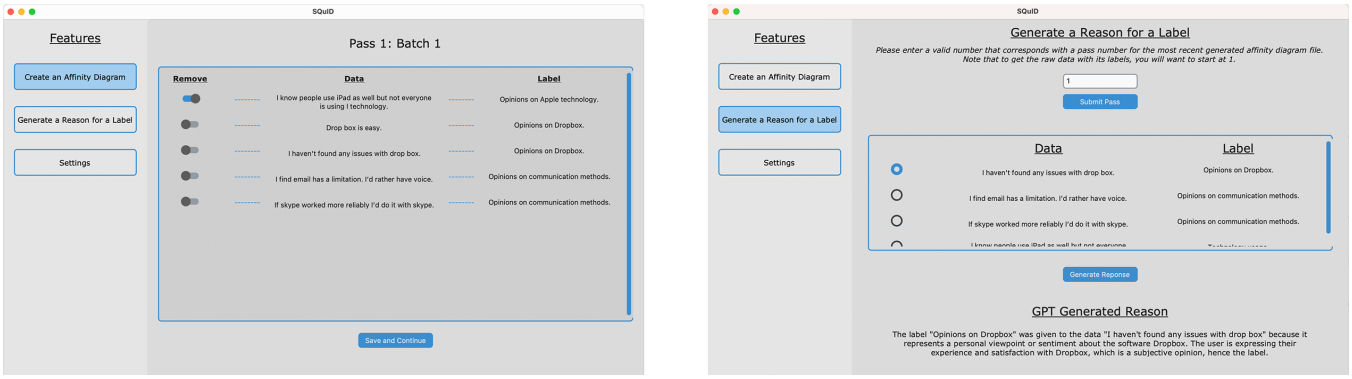

Figure 3: (Left) SQuID's interface following a completed pass, showing data-label pairings. (Right) The Generate a Reason for a Label tab displaying a post-hoc rationale generated for a selected data-label pairing. Rationales are designed to support researcher reflection rather than faithfully expose the model's original reasoning.

Labels within each community are merged by prompting the LLM to select the single most representative label from among the semantically equivalent candidates. For example, the labels *Importance of Self-Reflection and Personal Growth*, *Lessons from Self-Reflection and Personal Growth*, and *Personal Growth and Self-Realization* are semantically equivalent, with pairwise cosine similarities ranging from 0.926 to 0.968 (well above the default threshold of 0.91), and would be merged by SQuID into a single representative label: *Personal Growth and Self-Reflection*.

Girvan-Newman was selected over alternatives [6, 42] because it does not require a predetermined number of clusters and performs well at the small scales typical of group label sets, which rarely exceed a few dozen items per pass. This approach is more transparent than directly prompting an LLM to identify similar labels, since the similarity computation is deterministic.

To demonstrate the effectiveness of SQuID's batching and merging algorithm, we compared its performance with GPT-5.2 [47], OpenAI's flagship model, as a baseline. For direct comparison, we used the consumer interface of ChatGPT and the same one-shot prompt presented in Section 3.2. We used 0.91 as the merging threshold for SQuID with GPT-4 and 0.87 for SQuID with GPT-5.2, based on manual inspection into the semantic similarity of the model outputs. Our dataset [24] contained 431 data points describing undergraduate students' experiences with mistakes. We compared SQuID and the baseline on both the full 431-item dataset and the 200-item subset used in our user study (Section 4.2), measuring group count and ungrouped data points. As Table 2 shows, GPT-5.2 missed approximately 30% of data points with the entire dataset, while SQuID categorized over 95% of data in both conditions. Users of SQuID can further reduce missed data points to zero by rejecting ungrouped items back into the pool via the review feature.

3.4 Multi-Pass Analysis and Export

SQuID supports multi-pass analysis by treating the labels produced in one pass as data points for the next, enabling iterative abstraction toward higher-order themes. This reflects common qualitative practice, where analysts move from granular codes toward broader categories through repeated passes over the data. Users determine when sufficient abstraction has been reached and can terminate the process at any pass. Results are exported as a CSV file in which each original data point is associated with its assigned labels across all passes. This format is compatible with spreadsheet software

Table 2: Comparison with LLM baseline.

	200 DATA POINTS		431 DATA POINTS	
	Group Count	Data Not Grouped	Group Count	Data Not Grouped
GPT-5.2	12	10 (5.00%)	30	129 (29.93%)
SQuID with GPT-4	25 → 20*	0	50 → 34*	14 (3.25%)
SQuID with GPT-5.2	37 → 30*	5 (2.50%)	80 → 40*	16 (3.71%)

*Before merging → after merging

and graphical affinity diagramming tools such as Mural [45] and LucidChart [41], supporting import into existing workflows for further refinement and collaborative discussion.

3.5 Human-in-the-Loop Components

SQuID's human-in-the-loop design operates through two mechanisms: batch-level review and rationale-on-demand. After each batch is processed, the interface presents data points alongside their AI-generated labels. Users review each pairing and toggle any they disagree with for removal. Toggled items are returned to the unprocessed pool and reassigned in subsequent batches. Figure 3 (left) shows the interface following a completed pass, with accepted pairings retained and toggled pairings marked for reassignment. This continuous review process means that analysts are not simply confirming a completed output but shaping the diagram as it develops, maintaining a form of analytic agency throughout.

The Generate a Reason for a Label feature, shown in Figure 3 (right), allows users to request a post-hoc rationale for any data-label pairing after a pass is complete. The rationale is generated by the LLM and explains why a data point was assigned its label in terms of semantic content. These rationales are not faithful reconstructions of the model's internal reasoning; they are post-hoc accounts designed to support reflection and sensemaking. These rationales give analysts language to evaluate and, where necessary, contest a grouping, particularly when preparing to discuss or present results to others.

```
system: "Respond to the user by explaining why the data
presented received the mentioned label. Be specific about
the data and make clear connections for the user."

user: "Provide a reason for giving the label [selected
label] to the following data: [selected data point]."
```

These human-in-the-loop components position SQuID as an interactive partner in the diagramming process rather than a one-shot clustering tool. Users continuously guide the emerging structure, and the system provides resources for understanding and defending the groupings it produces.

Table 3: Participant demographics

PID	Age	Exp.	Research Area
026	31	7	UX Design
043	23	0.2	UX/UI Design
118	26	4	HCI
132	40	5	HCI
212	22	0.3	HCI
350	34	7	Industrial Design
356	24	1.5	Educational Technology
368	25	2	Human-AI Collaboration
576	22	2	Cognitive Science & HCI
593	26	3	HCI
604	24	2	HCI
734	23	1	Human-AI Interaction
796	27	2	Integrated Innovation/ Prod. Management

4 Method

4.1 Participants

Thirteen participants (ages 22–40, $M = 26.69$, $SD = 5.31$; six female, seven male) were recruited via email from six North and Central American institutions. All had experience with affinity diagramming, ranging from approximately two months to seven years ($M = 2.84$ years, $SD = 2.28$). Each participant was assigned a random three-digit identifier. Sessions were conducted remotely over Zoom, lasting approximately 80 minutes on average ($SD = 9.06$), and participants were compensated $40. The study was approved by our institutional review board (IRB #S24-035). Participant demographics are summarized in Table 3. Recruitment prioritized participants with direct affinity diagramming experience; this sample size is consistent with qualitative research norms, where theoretical saturation is the benchmark for sufficiency [25].

4.2 Procedure and Tasks

We conducted a laboratory study to examine how researchers interact with SQuID during early-stage affinity diagramming. Each session began with a brief demonstration of SQuID, after which participants completed two timed tasks while thinking aloud.

Task 1 (Provided dataset). Participants used SQuID to analyze a dataset of 200 short survey responses describing undergraduate students' experiences with making mistakes. This task familiarized participants with the system and elicited initial feedback on interaction design. Participants were given approximately 20 minutes to work, with a reminder at 5 minutes remaining.

Task 2 (Own dataset). Participants then used SQuID on a dataset from their own research that they had previously analyzed using manual methods. Dataset sizes ranged from 12 to 319 items ($M = 126.15$, $SD = 113.17$). This task increased ecological relevance by using participants' previously analyzed datasets, enabling comparison to their prior interpretations under time constraints. As in Task 1, participants worked for approximately 20 minutes.

We did not include a baseline condition where participants were asked to create an affinity diagram without SQuID. Full affinity diagramming typically unfolds over days or weeks and often involves

collaborative negotiation, making a controlled, within-session baseline infeasible. However, during semi-structured interviews, we asked participants to compare the analysis of their data with SQuID to their own prior analysis using manual methods. Time limits and individual tasks were chosen to observe how participants engage with AI-generated structure when first encountering such support without causing fatigue, rather than to replicate full qualitative practices that typically unfold over extended periods.

4.3 Measures

After each task, participants completed a short survey (10 items, 0–10 scale) assessing perceived utility, usability, clarity, and sense of control shown in Table 5. Perceived utility here refers to how participants experienced the system's contribution to early structuring, not a comparison to manual methods. Although we did not include a manual baseline, we retained self-report items about perceived utility to capture how participants experienced the redistribution of analytic work when interacting with AI-generated structure. We also measured perceived workload using the Raw NASA-TLX. Participants rated six workload dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration) using visual analog scales ranging from 0 (very low) to 10 (very high). We used a 0-10 continuous visual analog scale for consistency with our custom survey instrument. We omitted the pairwise weighting step following current practice [27], and report dimensions descriptively without aggregation into a weighted composite. Following the session, participants took part in a semi-structured interview probing their experience, points of trust or concern, comparison to manual affinity diagramming, and how they might incorporate SQuID into their research workflow.

4.4 Analysis

Survey responses were collected on a 0–10 visual analog scale (VAS), with 5 as neutral. Because items were conceptually related, we analyzed them both individually and as composites (perceived utility, usability/clarity, control/agency). To examine task differences (Task 1 vs. Task 2), we used mixed-effects models with Task as a fixed effect and participant intercepts, emphasizing effect sizes and confidence intervals given the small sample ($n = 13$). Robustness checks revealed no consistent task effects. Thus, we report only the survey responses after Task 2, when both tasks had been completed.

Raw NASA-TLX dimensions were analyzed descriptively, reporting means and standard deviations for each of the six dimensions as well as an unweighted overall mean computed by averaging across dimensions. No inferential tests were applied to TLX scores, given the sample size and the descriptive purpose of the measure. TLX data were used to characterize the general workload profile of the task rather than to compare conditions or test hypotheses about workload. As with the survey questions, robustness checks revealed no consistent task effects between Task 1 and Task 2. Thus, we report only the RAW NASA-TLX responses after Task 2, when both tasks had been completed.

Acceptance rate was defined as the proportion of system- generated data-label pairings that participants did not reject. The number of data points analyzed and rejected was derived from manual video review of Task 2 interactions. Task 1 was excluded because participants were still learning the system. During Task 1, several participants explicitly indicated that they initially forgot they could reject data-label pairings until prompted by a researcher.

Table 4: Acceptance of AI suggestions during Task 2.

PID	Rejections	Data Points	Percent Accepted
026	8	308	97.40%
043	1	21	95.24%
118	17	110	84.55%
132	47	319	85.27%
212	2	40	95.00%
350	0	17	100.00%
356	1	20	95.00%
368	53	225	76.44%
576	2	208	99.04%
593	120	209	42.58%
604	64	75	14.67%
734	4	76	94.74%
796	0	12	100.00%
Mean	24.54	126.15	83.07%
SD	36.59	113.17	25.77%

To examine whether prior qualitative analysis experience influenced system evaluations, experience was converted to numeric years (e.g., "≈1-2 months" coded as 0.2; "1.5 years" as 1.50). We computed Pearson correlations between experience and Task 2 ratings (overall mean and theory-driven composites), emphasizing effect sizes and 95% confidence intervals rather than significance testing due to the small sample.

Think-aloud and interview data were analyzed using reflexive thematic analysis [7]. One researcher read through the complete dataset and developed an initial set of themes and subcategories inductively, and then applied codes to the full dataset. The resulting coded dataset and themes were reviewed by the rest of the research team, who suggested revisions until consensus was reached.

5 Results

Among the thirteen participants, twelve completed both tasks. Participant P593 encountered technical issues with the toy dataset in Task 1 and therefore only completed Task 2, but still experienced all system features, completed surveys and the think-aloud protocol, and participated in the interview. Acceptance rates varied widely (mean 83.07%, range 14.67–100%; Table 4). Throughout the Results and Discussion, we treat survey responses as complementary to observed interaction behavior and qualitative reflection, rather than as stand-alone evidence of system effectiveness.

5.1 Interactions with the System

Across all thirteen participants, the average acceptance rate was 83.07% ($SD = 25.77$); however, this aggregate statistic masks a strongly bimodal distribution. Eleven participants consistently accepted the vast majority of the system's suggestions, with acceptance rates between 76.44% and 100% ($M = 92.97\%, SD = 7.58$). In

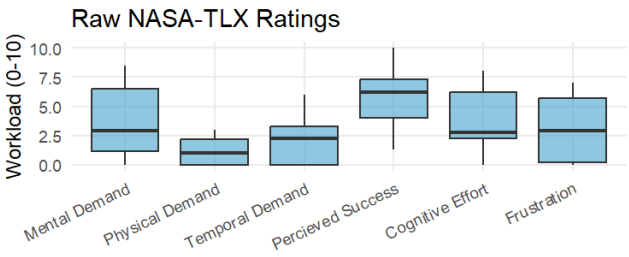

Figure 4: Raw NASA-TLX dimension scores (0–10 scale, n = 13). Overall workload averaged $M = 3.31$ ($SD = 1.29$).

contrast, two participants rejected most suggestions, with substantially lower acceptance rates of 42.58% and 14.67% ($M = 28.63\%$, $SD = 19.74$). We note that rejections by PID 604 appeared to stem from characteristics of their dataset, further underscoring that low acceptance does not necessarily indicate disengagement but may reflect data–system mismatch. When participants cluster into two distinct acceptance regimes, linear associations between experience and acceptance (Pearson's $r = 0.09, p = 0.76$; Spearman's $\rho = .04, p = 0.88$) or between dataset size and acceptance (Pearson's $r = -0.12, p = 0.70$; Spearman's $\rho = -0.48, p = 0.10$) are unlikely to emerge. These results suggest that acceptance behavior is better explained by dataset characteristics or interaction style than by individual experience.

5.2 Survey Ratings and Raw NASA-TLX

To test for differences between survey responses after each task, we used mixed-effects models with Task as a fixed effect and participant intercepts and found no consistent effects of Task on survey responses. Thus, we report only the results from the survey taken after both tasks. Participants' ratings showed moderate variability across items ($SDs = 1.49$–2.79, $CVs \approx 0.24$–0.54), with most SDs falling between 2 and 3. Survey results across participants are shown in Table 5.

Raw NASA-TLX scores (Figure 4) indicated relatively low perceived workload overall ($M = 3.31$, $SD = 1.29$). Physical demand was minimal ($M = 1.17$, $SD = 1.16$), consistent with a primarily cognitive task, and temporal demand was modest ($M = 2.18$, $SD = 1.95$), suggesting participants did not feel rushed. Mental demand ($M = 3.73$, $SD = 2.86$) and effort ($M = 3.82$, $SD = 2.52$) were somewhat higher, reflecting the cognitive work of evaluating and responding to system-generated groupings. Frustration remained low ($M = 3.01$, $SD = 2.60$). Participants reported moderate perceived success ($M = 5.97$, $SD = 2.43$), indicating that most felt able to accomplish the task.

Correlational analyses between affinity diagramming experience and survey composites revealed a consistent negative direction: more experienced participants tended to rate the system lower across all three composites ($r = -0.30$ to $r = -0.45$), though none reached statistical significance and confidence intervals were wide. No consistent pattern emerged between experience and NASA-TLX subscales (overall composite $r = -0.04$).

Table 5: Survey items and responses (VAS 0—10).

Question	M	SD	95% CI
Utility			
I found the diagrammer was effective in creating an affinity diagram.	6.02	2.32	[4.76, 7.28]
I think the affinity diagram generated was helpful for categorizing data.	6.31	2.79	[4.79, 7.82]
The system managed a lot of the work I needed to do to create an affinity diagram.	7.01	2.25	[5.74, 8.28]
Usability/Clarity			
I felt the tool was very clear about what was going on through the entire diagramming process.	5.06	2.71	[3.59, 6.53]
I think anyone could use this system without getting lost.	6.83	2.59	[5.42, 8.23]
I thought the system-generated diagrams correlated well with the data it was given.	6.60	2.21	[5.39, 7.80]
I enjoyed using the system.	7.13	2.09	[5.99, 8.26]
I thought the system created labels that made sense with the data.	6.72	2.20	[5.52, 7.92]
I thought the system created diagrams that were easy to understand.	6.34	1.49	[5.53, 7.15]
Control/Agency			
I feel that the system grants me a wide range of control when creating affinity diagrams.	4.12	2.05	[3.01, 5.24]
Composite: Utility	6.39	2.20	[5.20, 7.59]
Composite: Usability/Clarity	6.45	1.49	[5.63, 7.26]
Composite: Control/Agency	4.12	2.05	[3.01, 5.24]

5.3 Qualitative Data

5.3.1 Automation as a Scaffold for Early Structuring. All participants described SQuID as useful for accelerating early-stage analysis. During task 2, P368 exclaimed: *"My gosh, this could have saved me like seven hours."* P043 grounded this observation in practical terms, noting that their methods course had estimated manual affinity diagramming at two to three hours for fifty to sixty data points, making larger datasets slow to analyze. P132 framed the value of automation in more structural terms, describing SQuID as a way to make their participation in the analytic process more meaningful rather than simply faster: *"I'm trying to make my participation in the value loop more significant. And more respectful of my own time."*

Participants described two primary roles for automation in their work. Nine used SQuID to generate provisional groupings for subsequent refinement. Four others emphasized its capacity to reveal patterns they might not have identified on their own. Twelve of thirteen participants expressed interest in using SQuID in their own research; the exception, P350, noted that they would need to discuss accountability concerns with collaborators before adopting it. This hesitation anticipated concerns that became more prominent as participants worked with their own data.

Automation was not equally useful across all participants. P576 found SQuID's groupings of the provided dataset easier to evaluate than SQuID's groupings of their own data, because they had already developed an analytic framework for the latter and could not assess the system output independently of that framework. P604 similarly found automation most useful as a first pass on unfamiliar data, and less suited to analyses with a specific interpretive goal from the outset. These responses suggest that SQuID is most productive when participants enter data without strong prior expectations.

5.3.2 From Constructing to Checking. Using SQuID changed not just how quickly participants worked, but also what they were doing. Rather than reading through data, formulating labels, grouping items, and verifying those decisions, they found themselves starting from a generated structure and working backward to assess it. P026 described this directly: *"Now you are cross-checking it rather than [...] reading through it, and then processing it, and then creating the label and then cross-checking it. So that's like three, four steps more. So here you are skipping the three steps and creating the label, right?"*

Not all participants experienced this shift in workflow as straightforwardly productive. P604 identified a tension at the center of the shift: *"So the system is taking the complexity from me of having to categorize [the data]. But it's giving me the complexity of having to individually go through all of them."* The effort of construction was replaced by the effort of verification. For participants working with large datasets or with data they knew well, that tradeoff was not always favorable. P593 noted that checking the system's work introduced overhead that would not have existed in traditional analysis, even if they spent less time overall.

This shift in work demands also had consequences for how participants engaged with the data at the level of individual items. Five noted that the checking mode encouraged a kind of shallow reading: scanning labels rather than dwelling on specific data points. P026 observed that they moved through the material faster than they would have manually, and that this speed meant they were deciding whether a label fit rather than developing their own understanding of what a data point meant. P734 extended this concern, suggesting the checking interaction style could produce a researcher who arrives at a completed affinity diagram without having internalized the data that generated it. P212 offered a partial counterpoint, noting that the work of preparing data for SQuID already required close reading that produced incidental familiarity: *"A certain amount of stuff gets processed [...], just by having to read it over. I'm already starting to build a picture."* This suggests the severity of the construction-to-checking tension depends partly on how much preparatory work a researcher has done before the system begins.

The distribution of acceptance rates partially reflects this dynamic. The eleven participants who accepted most of the system's suggestions were largely operating in checking mode, treating the AI's structure as a reasonable draft to confirm. The two participants with substantially lower acceptance rates (P593 and P604) were doing something closer to active correction, engaging more adversarially with the output. Both were also working with data on which they had established analytic frameworks. This pattern suggests that checking mode is most comfortable when data is unfamiliar, and most strained when the analyst's expectations diverge from what the system produces.

5.3.3 Analytic Intent and the Limits of Computational Grouping. All thirteen participants encountered labels that did not align with their analytic goals, but the nature of those misalignments varied. The most revealing cases were not individual labeling errors but differences between how the system organized data and what participants were actually trying to accomplish.

P604 provided the clearest example of this misalignment. Working with user review data, they had intended to group responses by sentiment, tracing emotional valence across product categories. SQuID instead organized by source type, producing labels structured around platforms and websites rather than affect. They described this as not wrong but beside the point: *"It is not grouping by sentiment necessarily, but it did give me more data. This could be [...] like one type of affinity diagramming that I would do to get to know the demographics of where the reviews are coming from."* Their response was to reframe the output as a different and potentially useful analysis, while recognizing it was not the one they had set out to do. Working through the provided dataset P576 observed that SQuID produced neutral, descriptive labels where they expected directional ones. Their own practice involved capturing not just what participants described but what those descriptions implied about their needs. That distinction was not something the system's output could resolve on its own. P026, when working with data about personal identity and belonging, found that the system collapsed distinct emotional dimensions into a single category where they would have preserved differences between types of feeling: *"I think all the feeling aspect, it is labeling as 'personal experiences and growth' instead of separating them: What are the different feelings?"* The label was not inaccurate, but it flattened distinctions that were analytically significant to them. P212 identified a different kind of misalignment: not granularity but naturalness. Encountering the label *"Realization of user-centric design and development,"* they noted it read as an incomplete sentence and that *"as a human, I would never write this."* Labels that are recognizably machine-generated compound the accountability concerns described in Section 5.3.5, since they are harder to present as one's own analytic work even when the underlying groupings are sound.

A related concern emerged around the sequence of labeling and grouping. P132 described a deliberate practice of withholding label names until after all items had been grouped, specifically to avoid anchoring their interpretation prematurely: *"If I start setting a label first, I'm probably gonna bias myself to what the label name or label length should be before I go into the next groups."*. P118 shared this sentiment and also noted potential bias due to SQuID's workflow, which inverts this order: labels are presented alongside groupings from the first batch, before the analyst has seen the full dataset. *"By the LLM already labeling this, it kind of biases you towards [the AI-generated results], you may have come up with a different thing if you didn't use it."*. For participants like P132 and 118 who treat naming as a reflective act that depends on surveying the whole, this inversion was not merely inconvenient—it introduced a source of analytic bias that their manual process was specifically designed to prevent.

These misalignments were not simply label quality issues. They reflected a more fundamental difference between computational grouping logic, which operates on semantic similarity, and researcher analytic intent, which is shaped by theoretical commitments, prior knowledge, and the specific questions driving a study. P350 put it plainly: *"GPT tends to go for the most obvious way."* For participants with developed frameworks for their data, the most obvious organization was frequently not the most useful one.

On the other hand, the system also produced groupings that revealed meaningful connections users did not see previously. P368 discovered a category they had not identified in their own published analysis: a cluster the system labeled *Distrust for AI* that cut across data points they had coded differently. They reflected: *"I think what we were really missing there was that idea of trust in AI. And that probably would have been a better category theme [than the original manual analysis]."* P026 similarly noted moments where the system's groupings prompted them to consider framings they would not have reached independently. In both cases, the value came from the system not sharing their assumptions.

Taken together, this data suggests that SQuID's relationship to analytic intent is asymmetric. When participants had well-developed and specific frameworks, SQuID's output tended to be too general to be useful without substantial revision. On the other hand, SQuID produced useful analysis when participants were approaching unfamiliar data, or when the system organized along dimensions they had not considered. These patterns reflect a fit problem between the analyst's stage of inquiry and the system's capabilities.

5.3.4 Rationale-on-Demand as a Reflection Aid. SQuID's "generate reason" feature produces a post-hoc explanation for why a data point was assigned to a particular group. It was one of the most frequently discussed features in the interviews. Participants did not primarily use it to understand the system's internal logic; rather, they used it to decide whether to accept a given grouping. P734 articulated this most clearly during the interview portion of their session. Reflecting on the "generate reason" feature, they said: *"So you're not basically using the AI to do the analysis, but actually using the AI to help you understand your own analysis, which you did by yourself better. So it could actually give you better points to help you better justify the labels you made."* The rationale, in their view, was not a window into the system but a resource for the researcher when accounting for an analytic decision.

P593 described connecting access to explanation directly to their willingness to trust the output: *"I also like the fact that I can get explanations from the AI, which again, kind of goes back to building trust with the system. So even if the system makes a mistake, I at least understand why it did."* For P593, understanding the system's reasoning, even flawed reasoning, gave them a basis for deciding how to respond.

Nine participants noted that they would have preferred access to this feature while SQuID was making groupings, not only after the pass. P576 described the constraint directly: *"I saw this label, and I want to see why it happened there. But I have to finish everything and go back to the reasoning section."* P593 encountered the limitation mid-task, navigating to the feature and being told they would need to complete the full process first. Unable to get a rationale, they resorted to bulk rejection as a workaround, deciding: *"I'm going to take out the one that was not grouped and all of the ones that say learning from mistakes and feelings"*—removing entire label categories rather than evaluating individual assignments. P026 discovered the constraint through a clarifying question during the session, then reflected afterward that they *"would have still wanted to see how it generated that label"* but *"didn't get to use that."* P368 did not encounter the constraint directly but anticipated it as a design problem, proposing a rationale button embedded in the diagramming table so they could consult it while deciding whether to accept or reject a grouping.

Four participants also noted limitations in what the explanations covered. Rationales addressed individual data point assignments without connecting to other items in the group or to the emerging structure of the diagram as a whole. P356 suggested that explanations would be more useful if they described why multiple data points belonged together rather than why a single item was assigned a label. This distinction matters: participants were not only evaluating individual assignments. They were trying to understand what a group meant and whether they could claim it as their own analytic conclusion. A rationale that operates at the level of one pairing at a time does not support that kind of holistic reasoning. P212 identified a further temporal gap, comparing SQuID unfavorably to AI systems that display a visible reasoning trace during processing: *"Squid doesn't have that, it spits something out, so there's no window into its thought process."* For P212, this absence was what made SQuID feel like a tool rather than a collaborator, suggesting that rationale access and real-time process visibility may serve different functions: one supports evaluation, the other supports the sense of working alongside the system.

5.3.5 Accountability as a Social and Methodological Concern. Participants who understood SQuID well, found its groupings plausible, and intended to use it in future work still described discomfort about a specific scenario: being asked to explain and defend an AI-assisted affinity diagram to someone else. The problem was not epistemic uncertainty about what the system had produced. It was social, rooted in the anticipated demands of advisors, collaborators, and reviewers rather than in doubts about the output itself.

P350 described this most fully. Having worked through the provided dataset and found the system's groupings broadly reasonable, they reflected on what would happen if they presented those results to their advisors: *"Imagine I present this to my advisors as the author of this classification. No, I would not be that confident. Especially because they would ask questions. So what was the reasoning for this cluster? Why this team? Why not that? And I would be, 'Oh, well, the AI decided like that.'"* They followed this with a direct statement about their threshold for sharing the work: *"I wouldn't pass this information to my colleague saying, 'Oh yeah, I did this.'"* Their reluctance was not about output quality, but rather about the authorship claim embedded in presenting the analysis as their own.

P734 extended this issue into a broader methodological concern. They described a scenario in which a researcher could collect transcripts, run them through SQuID, and produce a completed affinity diagram without having meaningfully engaged with the data at any stage. Their assessment was direct: *"In a worst-case scenario, I can imagine a very lazy researcher gives maybe his undergrads the task of collecting data and making a transcript. And without being*

in the session, they just take all the transcripts, post it into SQuID, get the labels, and not really even know anything that went on in the study." This framing positions accountability not as a personal concern about attribution, but as a structural problem for qualitative research. In their account, the value of affinity diagramming comes from the immersion it requires; a tool that bypasses that immersion also bypasses the epistemic work the method is meant to do.

P604 connected accountability directly to the experience of effort. They had rated their effort during the study as zero, then revised upward: *"I would like it to be two, though, which is what I mean by giving me control. So I also have accountability while going through the system."* Their comment reframes control not as a usability preference but as a methodological need. Having to work through the analysis was what gave them a basis for claiming the output. Frictionless automation was both less satisfying and less accountable. P212 framed this as a general property of delegated work rather than an AI-specific concern, comparing their confidence in SQuID to confidence in code written by others: *"I know that someone who is more likely to make mistakes than I am is doing this, so there's a base level of uncertainty there, and it's not specific to [ChatGPT]."* This positions the accountability concerns described by other participants within a broader pattern of delegation confidence, suggesting that design interventions may benefit from attending to general properties of delegated work (visibility, verifiability, traceability) rather than focusing narrowly on AI-specific explanations. P350 also described how human collaboration differs from AI-assisted analysis. In their own practice, affinity diagramming involves negotiation among people with different backgrounds, frameworks, and interpretations of the same data. That process produces not just a diagram, but a shared understanding that can be explained and defended. The system does not participate in that negotiation. Its outputs arrive fully formed, without the back-and-forth that would make them feel like collaborative affinity diagramming: *"When I'm collaborating with other humans, I can understand what's their background. And a lot of times, people bring their own agendas in."* The lack of social dimension was part of what made the AI-generated structure feel like it was not fully theirs.

5.3.6 Comparison to Current Practices in Affinity Diagramming. During Task 2, participants drew direct comparisons with SQuID's output. P368 found the system's groupings closely matched categories their team had derived collaboratively: *"This is like the exact categories—unless it has access to my paper somehow, which I kind of doubt."* P350 reached a similar conclusion: *"It's very close to the one I have. Some terminology is different, but quite close."* For initial grouping, SQuID approximated manual output substantially faster.

Participants were equally clear about what that efficiency traded away. P026 noted that manual analysis requires granular engagement, the system does not replicate: *"For a more granular level of understanding, I would still need to manually go through them."* P734 located the deeper risk: *"The devil with qualitative analysis is in the details. Yes, you could get the categories based off of reporting, but there's also benefit to doing the more difficult process."* The comparison participants found hardest to bridge was the collaborative dimension of manual diagramming. P734 described the negotiation that traditional affinity diagramming requires: *"There are always disagreements. You have to discuss with your collaborators 'Why did you interpret this data this way?' And in the process of arguing, you get to understand the data more."* P118 echoed this: *"This is the reason why we do it with multiple people. When we disagree, we explain our reasoning and reach consensus."* SQuID could produce groupings comparable in content to what a team might produce, but not the shared understanding that reaching those groupings generates.

6 Discussion

The themes that emerged from our study were consistent across both datasets and across domains, suggesting they reflect properties of SQuID's interaction design rather than artifacts of any particular dataset. Taken together, they share a common orientation: effective AI assistance for affinity diagramming is not primarily a matter of grouping accuracy, but of preserving the conditions under which researchers can claim, explain, and defend their analytic work. Table 6 summarizes our design implications for AI-assisted qualitative analysis tools.

6.1 Automation as Early-Stage Support

Participants consistently described the groupings suggested by SQuID as a starting point to react to rather than a conclusion to accept, and this held across very different levels of experience and very different datasets. When output is a draft, and rejection is possible, researchers can engage with AI suggestions analytically rather than deferring to them.

P132 and P118's framing of this dynamic was distinctive. Rather than describing SQuID as saving time, they described it as raising the quality of their own contribution: automation was valuable because it redirected their effort toward judgment rather than mechanical labor. This reframing suggests that the value of early-stage AI support may be less about reducing total work than about concentrating researcher attention where it matters most. A system that handles initial organization frees the analyst to do the work that only they can do.

Automated grouping also supported reflective analysis in ways participants did not always anticipate. P368's discovery of a "Distrust for AI" category in their own published data is the clearest example. Moments like this were grounded in interaction with specific outputs rather than abstract reflection; participants could inspect the data points assigned to a label, evaluate whether the grouping held, and decide whether to accept or contest it. The provisional character of AI suggestions, combined with the ability to reject them, preserved a form of analytic agency even when participants accepted the majority of what the system produced.

6.2 The Labor Redistribution Problem

The shift from constructing to checking is more consequential than a simple efficiency gain. When researchers construct an affinity diagram manually, the labor of reading, grouping, and labeling is also the labor of developing a relationship with the data. Immersion and analysis are not separable stages, but SQuID decouples them. The system produces structure quickly, and the researcher's task becomes evaluating that structure rather than building it for themselves. This is faster, but also shallower in understanding, and is a design consequence of moving construction work to the AI. P734 located this risk not in SQuID specifically, but in the checking

Table 6: Summary of design implications for AI-assisted qualitative analysis tools, derived from participant observations.

Theme	Design Implication	Suggested Feature / Example
Automation as early-stage scaffold	AI outputs should be positioned as drafts, not final conclusions. Systems are most productive when researchers enter data without strong prior frameworks.	Present AI groupings as "suggested" with visible accept/reject affordances; prompt researchers to specify analytic intent (research questions, preferred label formats) before each pass begins to reduce mismatch between system output and analytic purpose.
Labor redistribution	Shifting from constructing to checking can reduce data familiarity. Systems should encourage deliberate engagement rather than passive confirmation. Exposing constituent data points during review also functions as a safeguard against hallucination: researchers are positioned to catch fabricated labels that a fully automated pipeline would not expose.	Require researchers to read a sample of items before accepting a grouping; present low-confidence assignments as prompts for closer reading rather than silently routing them to the next batch.
Granularity of control	Per-item rejection is necessary but not sufficient. Researchers need to edit labels, reassign items across groups, and intervene at the category level.	Provide inline label editing and cross-group reassignment; make these controls available during the diagramming process, not deferred to a post-hoc review stage.
Rationale access and scope	Explanations for AI groupings are needed at the moment of decision, not only after a full pass is complete, and should describe why multiple items belong together rather than why a single item was assigned a label.	Embed a "generate reason" button in the diagramming table, accessible per pairing during each batch review; provide rationales at the group level to support holistic reasoning about emerging categories.
Process provenance	Researchers need to account for their own analytic decisions, not just the system's behavior. Existing audit trails document outputs; they do not document the researcher's reasoning process.	Add structured reflection prompts at key decision points; generate session summaries capturing accept/reject patterns and flagging moments of uncertainty; export decision logs alongside groupings.
Accountability as a social concern	Concerns about justifying AI-assisted analyses to advisors and collaborators arise even when output quality is adequate. These are social and methodological concerns, not just epistemic ones.	Support documentation of AI involvement by analytic stage (e.g., AI-generated initial groupings vs. researcher-revised final themes); enable exportable audit artifacts for process reconstruction.
Conveying analytic intent	Researchers need mechanisms to convey study context and rejection reasoning to the system.	Brief study description input before processing; short annotations on rejected pairings that inform subsequent batches.

interaction style as a general pattern. This pattern made it possible to collect transcripts, run them through a system, and produce a completed affinity diagram without having read the data, formulated a single label, or made a decision that required understanding what a data point meant.

This does not mean SQuID's efficiency is without value. For early-stage orientation with unfamiliar data, for large datasets where full manual engagement is impractical, and for producing a provisional structure to be refined through subsequent immersion, the redistribution of labor can be productive. The problem arises when this checking mode is treated as a final product rather than a preliminary step, and when the speed of completion is taken as evidence of analytic depth. Design responses to this problem should support rather than bypass immersion; e.g., by making checking interactions slower and more deliberate, by requiring engagement with specific data points before acceptance, or by revealing moments of uncertainty as prompts for closer reading rather than routing rejected items to the next batch.

6.3 Control as a Prerequisite for Accountability

Our survey data showed that participants rated perceived control as the lowest of the three composite measures (Q10, $M = 4.12$), and qualitative feedback clarified what that limited control felt like in practice. Participants did not primarily describe control as a usability preference, but rather as a methodological need that determined whether they could claim ownership over the analysis. P604's comment is the clearest statement of this connection. Having initially rated their effort at zero and then revised upward, they were not complaining about difficulty. Their point was that having to make decisions, encounter resistance, and work through uncertainty was what gave them grounds to claim the analysis.

SQuID constrained control in two ways that mattered specifically for accountability. First, granularity: participants could reject individual pairings but could not edit labels or reassign items across groups, which meant there was no mechanism for correcting a label that was wrong in its framing rather than its assignment. Second, timing: the generate reason feature was unavailable until the final pass, so participants made acceptance decisions throughout the process without access to the tool that most supported those decisions. Both constraints had the same practical effect: they limited participants' available interventions to ones that were insufficient to ground a claim of analytic authorship.

6.4 Accountability Through Process Provenance

Participants' accountability concerns point to a problem that has received little attention in the design of AI analysis tools. Participants were not primarily asking for more transparency about the system's internal reasoning. They wanted to be able to account for their own analytic decisions by explaining what they did, why they grouped things as they did, and what they noticed the system did not do. We refer to this as a provenance problem; not in the sense of the system's audit trail, but in the sense of the researcher's account of how the analysis developed. A CSV export documents SQuID's output. It does not document why a researcher accepted one grouping and rejected another, what they noticed that the system missed, or how their understanding of the data shifted as they worked. P350's description of a future advisor meeting illustrates this directly. The questions they anticipate concern their own reasoning, not the system's. In qualitative research, how you arrived at your analysis is part of what reviewers evaluate, not just the result.

Prior work on explainable AI has focused primarily on making system behavior legible to users. Our findings suggest that for qualitative analysis tools, a different kind of explanation is also needed: one that helps researchers articulate their own reasoning, not just the system's. Features that would support this include structured reflection prompts at key decision points, session summaries that capture the researcher's accept-reject pattern and flag moments of uncertainty, and exportable decision logs that record what was considered and set aside, not just what was produced. These concerns also have practical implications for how AI-assisted affinity diagramming should be reported. Documenting AI involvement by analytic stage (e.g., distinguishing between AI-generated initial groupings and researcher-revised final themes) provides readers with a more accurate account of the interpretive process than treating the analysis as uniformly human-led or uniformly automated. Preserving audit artifacts, including pass-by-pass exports and records of rejected pairings, creates a basis for reconstructing how the analysis developed. System-generated rationales are best described in reporting as resources that supported the researcher's own judgment, not as justifications for specific analytic decisions.

7 Limitations and Future Work

This work has several limitations that suggest avenues for future research. First, SQuID does not retain information across batches or passes, limiting holistic reasoning about evolving group structures within the diagram. Its explanations were sometimes difficult to interpret because they were grounded in single labels or data points rather than relational contrasts across clusters. These issues point to opportunities to incorporate memory mechanisms [35]. To simplify output, SQuID also produces non-overlapping data groups; this differs from manual affinity diagramming, where themes may overlap, span across structures, or deliberately allow ambiguity between categories. This design decision shows the tension between algorithmic complexity and the open-ended structures that can characterize human-led affinity diagramming. Furthermore, while we used CSV and tabular representations as an auditable format common in remote qualitative work, future systems should explore spatial representations that better support flexible analysis. Our evaluation also relied on OpenAI models available at the time of data collection, which no longer represent the state of the art. Future work could compare alternative algorithms and workflows across models, especially open-source models that can be deployed locally to protect sensitive data.

Second, our work is limited by the experimental design. Our participant pool consisted entirely of HCI researchers, who may be more familiar with AI tools than researchers in other domains, limiting generalizability. While this sample size is consistent with qualitative research norms for theoretical saturation [25], participants were recruited from six North and Central American institutions and skewed toward moderate experience levels. Future work should replicate these findings with more diverse participant pools, including practitioners from industry and participants with less formal exposure to affinity diagramming.

The study design also imposed constraints; tasks were limited to 25 minutes and conducted individually, whereas affinity diagramming in practice often unfolds over days and involves collaboration. Time pressure and the absence of peer discussion may have increased reliance on automation and reduced the depth of label vetting, influencing acceptance rates and perceived control. We also did not study early-stage data familiarization practices, such as repeated reading prior to structuring. As a result, our findings do not speak to how AI assistance interacts with early immersion in the data, but rather to how analysts engage with AI once some familiarity is already present. Future work can examine how richer contextual information can be incorporated into AI support through longitudinal, in-situ studies.

8 Conclusion

This work examined how affinity diagrammers interact with AI assistance during initial encounters in a time-constrained laboratory setting. Participants reported reduced effort during repetitive affinity diagramming tasks and described shifting their attention toward inspecting and contesting AI suggestions, rather than manually organizing data. At the same time, participants expressed concerns about accountability when considering how AI-assisted diagrams might be presented to advisors or collaborators. Participants varied widely in how they relied on automation, underscoring that AI assistance does not produce uniform interaction strategies even among experienced qualitative researchers. Our findings suggest that AI-assisted qualitative analysis tools align more closely with affinity diagramming workflows when they emphasize interaction, transparency, and opportunities for intervention throughout the analytic process, rather than replacing interpretive work.

9 Acknowledgments

This work was conducted at a primarily undergraduate institution. The authors would like to give special thanks to Yu-Peng Chen and Violet Shi for proofreading this paper.

10 GenAI Usage Disclosure

ChatGPT, Claude, and Grammarly were utilized to assist in table formatting and proofreading.

References

[1] Saleema Amershi, Maya Cakmak, W Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[2] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 337–346. doi:10.1145/2702123.2702509

[3] Carl Auerbach and Louise B Silverstein. 2003. *Qualitative data: An introduction to coding and analysis*. Vol. 21. NYU press, USA.

[4] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.

[5] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers, USA.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008. doi:10.1088/1742-5468/2008/10/P10008

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[9] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 9 (June 2018), 20 pages. doi:10.1145/3185515

[10] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 443–452. doi:10.1145/2207676.2207738

[11] Creately. 2025. Affinity Diagram Tool | Affinity Diagram Online. https://creately.com/lp/affinity-diagram-tool/

[12] John W Creswell and J David Creswell. 2014. *Research desing: qualitative, quantitative and mixed methods approaches*. Vol. 54. United State of America: Sage Publications, USA.

[13] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9993–10001. doi:10.18653/v1/2023.findings-emnlp.669

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

[15] Jakub Drápal, Hannes Westermann, and Jaromir Savelka. 2023. *Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies*. IOS Press, USA, 197–206. doi:10.3233/FAIA230965

[16] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. arXiv:2401.15170 [cs.CL] https://arxiv.org/abs/2401.15170

[17] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 81–87. doi:10.1145/3278721.3278736

[18] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. doi:10.1145/3411764.3445188

[19] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (Miami, Florida, USA) *(IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. doi:10.1145/604045.604056

[20] 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

[21] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 29 pages. doi:10.1145/3613904.3642002

[22] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 19 pages. doi:10.1145/3544548.3581352

[23] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 7821–7826. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.122653799 doi:10.1073/pnas.122653799

[24] Ariel Goldman, Cindy Espinosa, Shivani Patel, Francesca Cavuoti, Jade Chen, Alexandra Cheng, Sabrina Meng, Aditi Patil, Lydia B Chilton, and Sarah Morrison-Smith. 2022. QuAD: Deep-Learning Assisted Qualitative Data Analysis with Affinity Diagrams. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 419, 7 pages. doi:10.1145/3491101.3519863

[25] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82. doi:10.1177/1525822X05279903

[26] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 95–104. doi:10.1145/2702123.2702561

[27] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting* 50, 9 (2006), 904–908. doi:10.1177/154193120605000909

[28] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[29] Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and Fine-tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 518–528. doi:10.1145/3636555.3636910

[30] 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

[31] Daye Kang, Zhuolun Han, Jiahe Tian, Muhan Zhang, and Jeffrey M Rzeszotarski. 2025. ThemeViz: Understanding the Effect of Human-AI Collaboration in Theme Development with an LLM-enhanced Interactive Visual System. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW494 (Oct. 2025), 29 pages. doi:10.1145/3757675

[32] Andrew Katz, Gabriella Coloyan Fleming, and Joyce B Main. 2026. Thematic analysis with open-source generative AI and machine learning: a new method for inductive qualitative codebook development. *Humanit. Soc. Sci. Commun.* 13, 1 (Jan. 2026). doi:10.1057/s41599-026-06508-5

[33] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5686–5697. doi:10.1145/2858036.2858529

[34] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 28 pages. doi:10.1145/3613904.3642830

[35] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[36] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can Long-Context Language Models Understand Long Contexts? arXiv:2311.04939 [cs.CL] https://arxiv.org/abs/2311.04939

[37] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[38] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 [cs.CL] https://arxiv.org/abs/2307.03172

[39] Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. Insights into LLM Long-Context Failures: When Transformers Know but Don't Tell. arXiv:2406.14673 [cs.CL] https://arxiv.org/abs/2406.14673

[40] Andrés Lucero. 2015. Using Affinity Diagrams to Evaluate Interactive Prototypes. 231–248 pages. doi:10.1007/978-3-319-22668-2_19

[41] Lucidchart. 2025. Affinity Diagram Software. https://www.lucidchart.com/pages/examples/affinity-diagram-software

[42] Claudia Malzer and Marcus Baum. 2020. A Hybrid Approach To Hierarchical Density-based Cluster Selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, USA, 223–228. doi:10.1109/MFI49285.2020.9235263

[43] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3173574.3173922

[44] Walter S Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. 2024. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine* 255 (2024), 108356. doi:10.1016/j.cmpb.2024.108356

[45] Mural. 2025. Affinity Clustering template | Mural. https://www.mural.co/templates/affinity-clustering

[46] Joel Oksanen, Andrés Lucero, and Perttu Hämäläinen. 2025. LLMCode: Evaluating and Enhancing Researcher-AI Alignment in Qualitative Analysis. arXiv:2504.16671 [cs.HC] https://arxiv.org/abs/2504.16671

[47] OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ (accessed on Mar. 10, 2025).

[48] Cassandra Overney, Belén Saldías, Dimitra Dimitrakopoulou, and Deb Roy. 2024. SenseMate: An Accessible and Beginner-Friendly Human-AI Platform for Qualitative Data Analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 922–939. doi:10.1145/3640543.3645194

[49] Andrew J. Perrin. 2001. The CodeRead System: Using Natural Language Processing to Automate Coding of Qualitative Data. *Social Science Computer Review* 19, 2 (2001), 213–220. doi:10.1177/089443930101900207

[50] Muhammad Zain Raza, Jiawei Xu, Terence Lim, Lily Boddy, Carlos M. Mery, Andrew Well, and Ying Ding. 2025. LLM-TA: An LLM-Enhanced Thematic Analysis Pipeline for Transcripts from Parents of Children with Congenital Heart Disease. arXiv:2502.01620 [cs.CL] https://arxiv.org/abs/2502.01620

[51] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 394, 14 pages. doi:10.1145/3411764.3445591

[52] Kota Sakaguchi, Reiko Sakama, and Takashi Watari. 2025. Evaluating ChatGPT in qualitative thematic analysis with human researchers in the Japanese clinical context and its cultural interpretation challenges: Comparative qualitative study. *J. Med. Internet Res.* 27 (April 2025), e71521. doi:10.2196/71521

[53] Alexandra Schofield, Siqi Wu, Theo Bayard de Volo, Tatsuki Kuze, Alfredo Gomez, and Sharifa Sultana. 2025. "My Very Subjective Human Interpretation": Domain Expert Perspectives on Navigating the Text Analysis Loop for Topic Models. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP22 (Jan. 2025), 30 pages. doi:10.1145/3701201

[54] Raymond Scupin. 1997. The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization* 56 (06 1997). doi:10.17730/humo.56.2.x335923511444655

[55] Evgeny Smirnov. 2025. Enhancing qualitative research in psychology with large language models: a methodological exploration and examples of simulations. *Qualitative Research in Psychology* 22, 2 (2025), 482–512. arXiv:https://doi.org/10.1080/14780887.2024.2428255 doi:10.1080/14780887.2024.2428255

[56] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 293–304. doi:10.1145/3172944.3172965

[57] Hala Sun, MiRan Kim, Soyeon Kim, and Laee Choi. 2025. A methodological exploration of generative artificial intelligence (AI) for efficient qualitative analysis on hotel guests' delightful experiences. *International Journal of Hospitality Management* 124 (2025), 103974. doi:10.1016/j.ijhm.2024.103974

[58] Robert H. Tai, Lillian R. Bentley, Xin Xia, Jason M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Barnas G. Monteith. 2024. An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods* 23 (2024), 14 pages. doi:10.1177/16094069241231168

[59] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. arXiv:https://doi.org/10.1177/1098214005283748 doi:10.1177/1098214005283748

[60] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 21 pages. doi:10.1145/3613904.3641960

[61] 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

[62] Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, Yan Leng, and Ying Ding. 2025. TAMA: A Human-AI Collaborative Thematic Analysis Framework Using Multi-Agent LLMs for Clinical Interviews. arXiv:2503.20666 [cs.HC] https://arxiv.org/abs/2503.20666

[63] Yang Yang and Liran Ma. 2025. Artificial intelligence in qualitative analysis: a practical guide and reflections based on results from using GPT to analyze interview data in a substance use program. *Quality & Quantity* 59, 3 (June 2025), 2511–2534. doi:10.1007/s11135-025-02066-1

[64] 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:2310.07061 [cs.HC] https://arxiv.org/abs/2310.07061

[65] Lishan Zhang, Han Wu, Xiaoshan Huang, Tengfei Duan, and Hanxiang Du. 2024. Automatic deductive coding in discourse analysis: an application of large language models in learning analytics. arXiv:2410.01240 [cs.CL] https://arxiv.org/abs/2410.01240

Received 30 January 2026; revised 3 April 2026; accepted 6 May 2026