Word Clouds as Common Voices: LLM-Assisted Visualization of Participant-Weighted Themes in Qualitative Interviews

Joseph T. Colonel
Icahn School of Medicine at Mount Sinai
joseph.colonel@mssm.edu

Baihan Lin*
Icahn School of Medicine at Mount Sinai
baihan.lin@mssm.edu

Abstract

Word clouds are a common way to summarize qualitative interviews, yet traditional frequency-based methods often fail in conversational contexts: they surface filler words, ignore paraphrase, and fragment semantically related ideas. This limits their usefulness in early-stage analysis, when researchers need fast, interpretable overviews of what participant actually said. We introduce **ThemeClouds**, an open-source visualization tool that uses large language models (LLMs) to generate *thematic, participant-weighted* word clouds from dialogue transcripts. The system prompts an LLM to identify concept-level themes across a corpus and then counts how many unique participants mention each topic, yielding a visualization grounded in *breadth of mention* rather than raw term frequency. Researchers can customize prompts and visualization parameters, providing transparency and control. Using interviews from a user study comparing five recording-device configurations (31 participants; 155 transcripts, Whisper ASR), our approach surfaces more actionable device concerns than frequency clouds and topic-modeling baselines (e.g., LDA, BERTopic). We discuss design trade-offs for integrating LLM assistance into qualitative workflows, implications for interpretability and researcher agency, and opportunities for interactive analyses such as per-condition contrasts ("diff clouds").

1 Introduction

Qualitative interviews are a cornerstone of HCI practice: they capture lived experience, tacit knowledge, and situated rationales that are difficult to elicit through logs or lab tasks alone (Hopf, 2004). But precisely because conversational data are rich, early-stage sensemaking can be slow and brittle. Time-constrained teams often rely on word clouds to orient themselves and to communicate initial patterns. Word clouds help researchers surface recurring terms and communicate high-level themes to stakeholders (Khusro et al., 2021). In principle, a quick visualization that "shows what people talked about" is invaluable. In practice, however, frequency-based word clouds tend to reflect *how* people talk rather than *what* they mean.

This misalignment is acute for spoken transcripts. Even with stop-word removal, the statistical surface of talk often dominates frequency ranks, such as disfluencies ("uh"), discourse markers ("like", "you know"), and coordination ("and"). Moreover, participants rarely reuse identical strings when describing similar concerns. One person may say "it felt in the way," another "kind of distracting," another "I kept noticing the device," and a fourth "it made me self-conscious." Traditional clouds fragment these into separate tokens, spreading salience thinly across synonyms and paraphrases. The resulting picture understates a theme's breadth and overstates lexical quirks, leaving analysts to manually reconcile meaning after the fact.

In our motivating study, clinicians and participants evaluated different recording-device configurations intended for psychiatric assessment. When we generated standard frequency clouds per device, familiar problems reappeared: conversational scaffolding rose to the top; multi-word concerns broke into stems; semantically aligned reactions (e.g., "distracting," "in the way," "felt watched") were scattered. The clouds neither matched researcher notes nor helped communicate trade-offs to stakeholders. A different aggregation principle was needed.

Recent advancements in large language models (LLMs) present new opportunities for enhancing qualitative analysis (Xu et al., 2025). Models such as Llama 3.3 can process long passages of unstructured text, identify latent topics, and recognize semantically important terms even when they are phrased differently across transcripts (Touvron et al., 2023). These capabilities make LLMs well-suited for tasks like summarization and topic

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 169–177

extraction, which are core components of qualitative synthesis.

In our use case, LLMs make a new design space feasible. Rather than counting words, we can ask a model to reason about concepts, recognize paraphrases, and collapse near-synonyms—capabilities that have matured as models improved long-context understanding. But naively inserting LLMs can reduce transparency. Our design goal, therefore, is to preserve the *immediacy and communicability* of word clouds while shifting the unit of analysis from tokens to *concepts*, and the weighting from raw counts to the *breadth of mention* across participants. In effect, we want a cloud that answers the question analysts and stakeholders actually ask: "How many people brought this up?"

We contribute a method and artifact that operationalize this shift in a way that fits qualitative workflows. Our open-source tool, ThemeClouds, leverages Llama 3.3 to assist in generating semantic word clouds from qualitative interview transcripts. Rather than relying solely on term frequency, the tool uses LLM reasoning to extract salient terms and conceptually related groupings, producing visualizations that better reflect the themes embedded in natural dialogue. By incorporating lightweight user control, the system balances LLM assistance with researcher agency, supporting interpretation while preserving transparency and flexibility.

Our work builds on prior literature in textual visualization and qualitative coding tools (Bateman et al., 2008; Lennon et al., 2021). While previous approaches have highlighted the risks of misleading word clouds or opaque model outputs, we aim to demonstrate how thoughtful design centered around customization and interpretability can help researchers co-construct word clouds with LLMs in qualitative workflows. The remainder of this paper describes the architecture and design decisions behind the system, demonstrates its application to interview data, and reflects on broader implications for LLM-assisted tools in qualitative analysis.

Our contribution is methodological and pragmatic. We do not claim a new theory of qualitative analysis; instead, we provide a lightweight, defensible, and *participant-weighted* alternative to frequency clouds that better aligns early-stage summaries with how analysts reason and report. We show how to integrate LLM assistance without obscuring the analytic process, emphasizing controls, artifacts, and audit trails that allow researchers to trust, contest, and adapt outputs.

2 Related Work

2.1 Word clouds as communicative summaries

Word (or tag) clouds have enduring appeal because they compress large corpora into a glanceable visual summary, where word frequency maps to font size. Early tools like Wordle made word clouds ubiquitous on the web (Steele and Iliinsky, 2010). Kaser and Lemire formalized the layout problem, showing how to use 2D packing and typesetting techniques to draw tag clouds efficiently (Barth et al., 2014). Subsequent work evaluated how visual features affect readability and selection (Rivadeneira et al., 2007; Bateman et al., 2008). As a result, classic word clouds can be "aesthetically pleasing" and easy to create but have well-documented limitations for analytic tasks.

These efforts improved the communicative surface, yet the core statistic – token frequency – remains brittle in conversational settings, where disfluency and paraphrase are the norm. Our approach retains the familiar word-cloud form while changing the underlying weighting to reflect population-level salience.

2.2 Speech-derived clouds and semantic grouping

Spoken language transcripts differ markedly from traditional text sources like news articles or reviews as they are spontaneous, unedited, and often noisy. Disfluencies such as filler words ("um", "like"), false starts, and repetition are commonplace. The transcript format introduces both unique structure (turn-taking, repair, backchannels) and noise (ASR errors, fillers). These properties challenge the direct application of word cloud techniques developed for clean, edited corpora. Prior work in visualization, natural language processing (NLP), and accessibility has begun addressing these issues, especially in the context of spoken interactions.

Several systems have explored real-time word cloud generation from speech. Iijima et al. designed an interface for deaf and hard-of-hearing users that visualizes each speaker's utterances as personalized word clouds, enabling better topic tracking in meetings (Iijima et al., 2021). Importantly, their system filters out non-content words, addressing the prevalence of noise in speech. Chandrasegaran et al. similarly integrate ASR with word clouds in TalkTraces (Chandrasegaran et al., 2019), emphasizing that when enhanced with topic modeling and embedding-based filtering, word clouds

can help users follow evolving spoken discussions. These works highlight the value of preprocessing speech transcripts to improve word cloud clarity.

The semantic structure of speech also requires more than frequency-based layouts. Wang et al. proposed ReCloud (Wang et al., 2020), which clusters semantically similar terms using NLP techniques, allowing users to grasp themes rather than isolated keywords. Skeppstedt et al. extended this idea with Word Rain (Skeppstedt et al., 2024), embedding word semantics along a visual axis and combining font size with TF-IDF bar charts. Though both methods were tested on written corpora (reviews, climate texts), they underscore how semantic grouping and de-biasing frequency are crucial for domains where redundancy and ambiguity are common.

Together, these studies suggest that effective word cloud generation from speech transcripts must account for semantic ambiguity and high noise levels. This motivates approaches that combine filtering for content-bearing terms and semantically aware tags to produce meaningful visualizations of conversational speech. Our method builds on this trajectory by externalizing grouping decisions to an LLM while preserving analyst control over prompts, topic cardinality, and the final mapping.

2.3 LLM-assisted thematic analysis

LLMs have been used to accelerate theme discovery, propose candidate codes, and reduce analytic burden, sometimes reaching near-human agreement in semi-structured settings. They enable scalable and semi-automated approaches to thematic analysis of qualitative interviews, especially in domains where manual coding is labor-intensive. In the biomedical context, Xu et al. introduced TAMA (Xu et al., 2025), a multi-agent LLM framework designed to assist clinicians in analyzing interviews related to congenital heart disease. By integrating human-in-the-loop feedback with AI-generated theme suggestions, TAMA enhances the accuracy and distinctiveness of identified themes, while significantly reducing the burden on expert coders. Similarly, Singh et al. developed RACER (Singh et al., 2024), an LLM-powered methodology applied to semi-structured interviews conducted during the COVID-19 pandemic. RACER achieved near-human agreement in theme extraction, demonstrating that LLMs can reliably support mental health research involving large volumes of qualitative data.

These successes suggest that concept-level reasoning over long documents is feasible. Our contribution is to harness these capabilities for a narrow but ubiquitous task (first-pass summarization via word clouds) while foregrounding human-centered properties (agency, transparency, workflow fit) that determine whether such tools are practically useful in HCI contexts.

3 Methods

ThemeClouds is designed to assist researchers in generating word clouds from qualitative interview transcripts by using LLMs to surface salient, semantically meaningful concepts, rather than relying on surface-level word frequency. The pipeline consists of three key stages: (1) identifying candidate concepts across a corpus, (2) mapping those concepts to individual transcripts, and (3) aggregating the results to produce a word cloud visualization. Our system prioritizes topic relevance, clarity, and interpretability over lexical frequency or length. Figure 1 outlines the proposed workflow.

We formalize the shift from tokens to concepts and from frequency to breadth. Let $\mathcal{T} = \{t_1, \ldots, t_M\}$ be transcripts (one per participant for a given condition) and let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be short concept-phrases proposed by an LLM for the corpus. For each transcript t and concept c, the mapping step produces a binary assignment $y(t, c) \in \{0, 1\}$ indicating whether the concept is clearly present in the transcript (the artifact optionally supports a soft score $\hat{p}(t, c) \in [0, 1]$ with threshold τ for binarization). The *breadth* of concept c is:

$$b(c) = \sum_{t \in \mathcal{T}} y(t, c),$$

the number of unique participants whose transcripts include the concept. The visual weight for c is $w(c) = g(b(c))$, where $g(\cdot)$ is a monotone scaling (linear by default; logarithmic and square-root options aid mid-rank legibility). We also support condition-wise contrasts by rendering $\Delta b(c) = b_{\mathrm{A}}(c) - b_{\mathrm{B}}(c)$ to make differences across device configurations glanceable.

3.1 Input and preprocessing

The system takes as input a collection of textual transcripts from qualitative interviews. These transcripts may come from usability studies, field interviews, focus groups, or other open-ended sources. Transcripts are assumed to be minimally cleaned

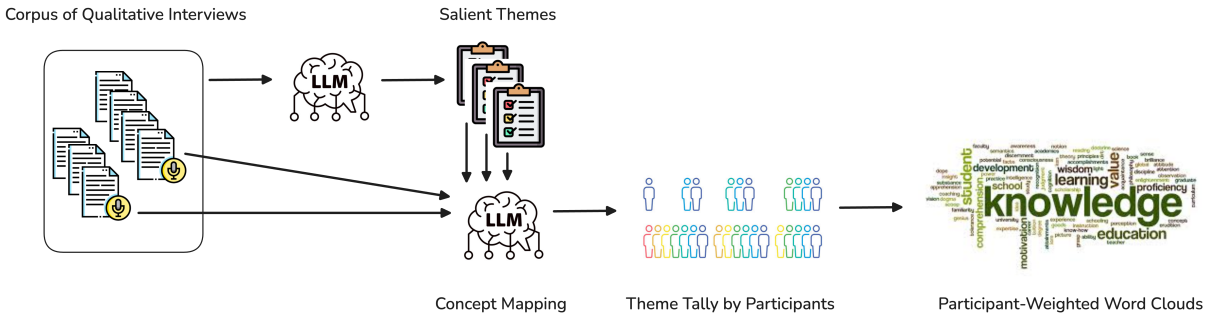


Figure 1: System overview for **ThemeClouds**: LLM-assisted *participant-weighted* thematic word clouds. An LLM first proposes a compact set of concept-level themes for the corpus. Each transcript is then mapped to this fixed theme list via binary presence judgments, yielding a per-theme count of *unique participants* (breadth). The final cloud sizes each theme by its participant prevalence (not token frequency). Prompts, per-transcript assignments, and counts form an audit trail that supports iteration and reproducibility.

(e.g., anonymized and transcribed verbatim) but do not require pre-coding or structuring. Because the method abstracts above tokens, we found that aggressive lexical normalization is unnecessary; we keep punctuation and stop-words intact for the LLM stage, using standard tooling like NLTK only for baseline clouds (Bird, 2006). Interviews are transcribed with Whisper (Radford et al., 2023).

3.2 Concept elicitation (corpus-level)

The goal is a compact, human-interpretable vocabulary that captures salient ideas without collapsing distinct concerns. We prompt a long-context LLM with the corpus (or stratified subsets) to propose N short concept-phrases, encouraging specificity (e.g. "in the way," "felt watched," "image quality"), discouraging generic terms ("user," "good," "bad"), and avoiding fillers or study-task scaffolding. Rather than returning frequent unigrams or bigrams, the model is guided via prompt engineering to prioritize short phrases, semantically specific topics, and coverage diversity across the corpus. We favor a diverse set that covers the thematic space rather than a large list that risks redundancy. The artifact includes our exact prompts and a small set of variations. Analysts can re-run this step to explore granularity. We also explicitly discourage the model from selecting filler words, generic terms like "user" or "system," or concepts that appear frequently but lack thematic depth.

In our evaluation, we prompt a poplar open-source LLM model, LLaMa-3.3-70B-Instruct (Touvron et al., 2023), to identify a set of N salient topics that best represent key concepts across the entire corpus with the following prompt.

```
You are analyzing interview transcripts where participants were asked to share their experiences using five webcam setups: [insta], [single iphone], [dual iphones], [logitech], and [obsbot].
The transcripts are organized in the following format: Each section begins with the webcam label (e.g., "### insta") followed by participant comments about that device.
Ignore filler words, repeated question prompts, or interviewer language. Focus only on participant speech that offers insight, reaction, or description.
Your task is to identify **exactly 20 meaningful and distinctive words or short phrases** that summarize participants' real experiences for **each webcam setup**.
Guidelines:
- Do NOT just pick the most frequent words.
- Select words or short phrases that are **emotionally descriptive**, **technically relevant**, or **highlight distinctive qualities** (positive or negative).
- Avoid: generic words (e.g., "thing", "camera"), filler words, or phrases repeated from the question.
For each setup, return a bullet list of 20 high-quality descriptors.
Output format:
### [setup] - ...
```

The result is a curated list of N topics that act as candidate entries for the word cloud. These phrases serve as a proxy for the major themes in the interviews, as judged by the LLM in context.

3.3 Concept mapping (per transcript)

In the second stage, the LLM is prompted to evaluate each transcript individually in relation to the N identified concepts or insights. For each transcript, the model receives: (1) the full content of that single transcript and (2) the fixed list of N topics produced in the prior step.

The model is then tasked with identifying which topics are clearly present in the given transcript. Importantly, the prompt encourages the model to make binary or categorical judgments rather than assigning soft weights or scores. This helps mitigate overfitting and keeps results interpretable for the end user. We use the following prompt:

```
You are analyzing a participant's response about the **device_name** webcam setup.
Below is a list of key descriptive terms and phrases that were identified across interviews for this webcam.
Your task is to determine **which (if any)** of these words or phrases are meaningfully reflected in the participant's comments — even if the exact wording is not used.
Focus on semantic alignment: if a participant implies or clearly expresses a concept that corresponds to one of the key terms, include it.
### Key Descriptive Terms for **device_name**:**keyword_list**
### Output Instructions:
Return ONLY a list of matching terms (one per line).
Do not include explanations, numbering, bullet points, or extra commentary.
A maximum of 20 key descriptive terms and phrases are allowed.
It is imperative to avoid false positives: if a keyword isn't reasonably supported, do not include it.
```

Through the above approach, given $\mathcal{C}$, we map each transcript independently by asking the LLM to judge concept presence using the fixed vocabulary. We default to binary assignments to keep outputs interpretable and to avoid length confounds: loquacious speakers should not inflate weights. Binary judgments also simplify spot-checks: analysts can audit questionable assignments by reading short excerpts of the transcript. The artifact includes an optional soft scoring mode ($\hat{p}(t, c)$) and guidance for threshold selection if analysts prefer graded presence.

This process is repeated for every transcript in the corpus. For each topic, we then compute a relative count of the number of transcripts in which the topic was marked as present. This produces a simple but robust measure of topic salience across the corpus.

3.4 Visualization and contrasts

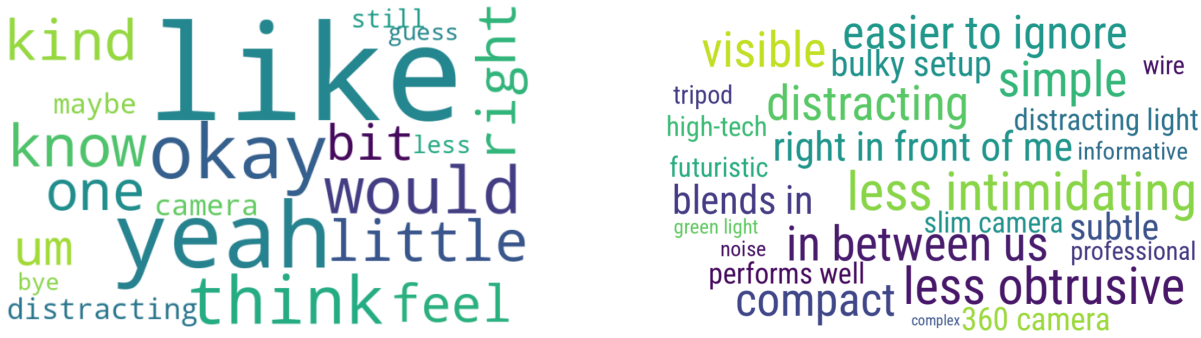
The final step uses these tallied topic counts to construct a word cloud. We render a conventional word cloud where size encodes $w(c)$. Each of the N topics or concepts to be highlighted is included. Because the units are now people, font size directly communicates population-level salience: often the most defensible signal when communicating with product teams or clinical stakeholders. In another word, the font size of each phrase is scaled based on how frequently it was mentioned across the subjects recruited for the qualitative interviews. Terms that were mentioned in most or all transcripts are rendered largest, while rare or marginal topics appear smaller.

For comparative analysis, we can also produce condition-wise "diff clouds" by coloring or separating concepts whose $\Delta b(c)$ exceeds a small margin. This reveals what a device configuration uniquely amplifies or suppresses.

3.5 Analyst-in-the-loop workflow

A central design goal is *researcher agency*. The system includes controls for adjusting the number of topics or concepts to note, the word cloud layout, font scaling, and prompt variants. Analysts can (1) edit the prompt, (2) adjust N, (3) seed or pin concepts they care about, (4) re-run elicitation to split overly broad concepts, and (5) audit and correct per-transcript assignments. This allows researchers to explore different perspectives on their data while retaining interpretability and structure. While the LLM outputs are fixed per run, users can rerun the topic generation with new prompts or adjusted constraints to suit different analytic goals.

We persist an *assignment table* with rows as



(a) Frequency-based word cloud

(b) LLM-assisted word cloud

Figure 2: Side-by-side comparison for one device condition (top 20 items shown). **(a)** A traditional frequency cloud—even after stop-word filtering—elevates conversational surface tokens and fragments paraphrases. **(b)** Our LLM-assisted, participant-weighted ThemeClouds collapses paraphrases into themes and sizes each by the number of unique participants who mentioned it, foregrounding actionable concerns from the interviews.

transcripts and columns as concepts so that any cloud can be reconstructed, inspected, or exported to downstream thematic coding. This audit trail helps teams defend qualitative findings in mixed-methods reports.

4 Human-Centered Design Considerations

A tool succeeds in HCI not only by being accurate, but by fitting how people actually work. We therefore prioritized five properties.

- *Interpretability:* counting unique participants aligns with how analysts argue salience ("many people brought this up").
- *Transparency:* we expose prompts, concept lists, assignment tables, and scaling choices, making it easy to reconstruct decisions or contest them.
- *Agency:* analysts can tune granularity and rerun steps to explore alternative framings.
- *Frugality:* default settings work on small, noisy corpora typical of interviews, without heavy parameter sweeps.
- *Workflow fit:* outputs are designed to triage and guide subsequent coding, not to replace careful qualitative analysis, echoing prior HCI work on semantic grouping and hybrid visual summaries (Iijima et al., 2021; Wang et al., 2020; Chandrasegaran et al., 2019; Skeppstedt et al., 2024).

5 Qualitative Evaluation

To assess the utility of our approach in a real-world setting, we applied it to a set of qualitative interviews conducted as part of a clinical psychology research study. 31 participants evaluated five webcam setups for psychiatric outpatient clinical assessments, producing 155 interviews with a clinical research coordinator. These conversations were conducted as one-hour in-person session, in a naturalistic dialogue format, as participants and the clinical research coordinator collaboratively evaluated different hardware configurations. Audio was transcribed with Whisper (Radford et al., 2023).

For a representative device condition (31 transcripts), Figure 2 compares a standard frequency-based word cloud with NLTK stop-word removal (Bird, 2006) (a) with our LLM-assisted word cloud (b). Despite identical source data, the frequency cloud elevates general discourse terms and fragments multi-word concerns, while the LLM cloud foregrounds concrete, device-specific ideas consistent with researcher notes.

What the numbers mean. Because our weights are counts of unique participants, the magnitude of a label directly translates to breadth. If "distracting" appears in 20 of 31 transcripts for a device, its visual prominence is immediately defensible—helpful for design reviews and IRB or clinical discussions where conservative, population-grounded claims are preferred.

To situate the approach among common baselines, we also trained topic models such as LDA and BERTopic (Rehurek et al., 2011; Grootendorst, 2022; Lin et al., 2023a) on the full 155-document corpus. As in Table 1, given the small per-condition

Table 1: Comparison of outputs from BERTopic, LDA, and our participant-weighted thematic method on the interview corpus. Lists are reproduced from model outputs (verbatim) and our curated themes (top items).

BERTopic (Top Topics)	**LDA (Top Topics)**	**ThemeClouds**
1. yeah, like, maybe, okay, whatever, aware, still, um, issue, course	1. 0.005*like + 0.004*okay + 0.004*think + 0.003*would + 0.003*little	1. Small and compact
2. like, part, um, things, process, always, treatment, cause, really, way	2. 0.018*like + 0.008*okay + 0.007*yeah + 0.007*think + 0.007*little	2. Not distracting
3. definitely, would, bit, oh, uncomfortable, good, odd, want, fact, especially	3. 0.004*like + 0.003*um + 0.003*part + 0.003*would + 0.003*things	3. Easy to ignore
4. yeah, like, maybe, okay, whatever, aware, still, um, issue, course	4. 0.004*like + 0.004*okay + 0.003*think + 0.003*little + 0.003*um	4. Less noticeable
5. okay, little, look, think, least, light, um, blends, bright, get	5. 0.099*like + 0.034*little + 0.027*think + 0.026*okay + 0.025*bit	5. Not too visible
6. yeah, even, white, side, either, light, much, like, slightly, around	6. 0.139*like + 0.045*yeah + 0.027*um + 0.027*okay + 0.022*think	6. Fades into the background
7. okay, little, look, think, least, light, um, blends, bright, get	7. 0.174*like + 0.037*um + 0.023*things + 0.020*would + 0.020*part	7. Simple and straightforward
8. definitely, would, bit, oh, uncomfortable, good, odd, want, fact, especially	8. 0.103*like + 0.057*would + 0.036*bit + 0.031*definitely + 0.026*think	8. Convenient
9. yeah, even, white, side, either, light, much, like, slightly, around	9. 0.130*like + 0.029*um + 0.024*yeah + 0.018*would + 0.018*okay	9. Reminds me of a Polaroid
10. like, part, um, things, process, always, treatment, cause, really, way	10. 0.133*like + 0.046*okay + 0.046*little + 0.040*um + 0.040*think	10. Compact and spacious

sample size and conversational style, neither produced immediately legible, per-device themes without additional manual massaging. Our participant-weighted list, on the other hand, aligns closely with analyst field notes and per-device concerns recorded during the study, foregrounding concept-level themes (e.g., "Not distracting," "Discreet," "Blends into the desk") that multiple participants independently raised.

While these observations are not a controlled user study, they illustrate a pattern we frequently saw during analysis: people-weighted concept clouds provide a more faithful "first glance" at what mattered to participants than token frequency or off-the-shelf topic models in this setting. It can effectively support researchers in identifying salient themes from conversational transcripts, even without structured codes or annotations.

6 Discussion and Limitations

Our tool demonstrates how large language models can be leveraged to assist in synthesizing qualitative feedback through semantic word clouds, offering an accessible, low-overhead entry point into exploratory analysis. While initial use cases show alignment with human interpretation, there are important limitations to consider.

6.1 Validity, bias, and controllability

LLM judgments depend on prompts and may overgeneralize. The system relies on static prompts and single-pass outputs, which may overlook nuances or misrepresent concepts without user intervention. We mitigate this by using a fixed vocabulary (reducing drift), binary mapping (reducing verbosity bias), and an assignment table that supports spot-checks and corrections. Analysts can also seed concepts to ensure coverage of domain-critical con-

cerns, an approach compatible with standard qualitative rigor practices.

6.2 Granularity and concept drift

The right granularity is contextual. Collapsing all camera-related concerns might hide distinctions between "felt watched" and "image quality." While prompt customization provides some control, more interactive or iterative workflows could better support researchers in refining outputs over time. Our workflow treats concept elicitation as an iterative process: split or merge concepts, re-run mapping, and compare clouds. We found small N (e.g., 12–25) balanced coverage and legibility, but analysts can tune N to their corpus.

6.3 Generalizability and small-data regimes

The method targets the small, noisy corpora typical of interviews and focus groups. Unlike topic models, which may prefer longer documents or larger datasets, our mapping step scales down: it asks a concrete question of each transcript with a fixed vocabulary. This makes the method robust when M is modest and concepts are grounded in context of the study and clinical application.

6.4 Ethics, privacy, and deployment

Interviews often contain sensitive information. Our artifact documents de-identification assumptions and supports local or compliant deployment. We view LLM assistance as a *scaffold* for human analysis, not a replacement: analysts should verify sensitive claims and avoid over-reliance on automated judgments in consequential settings.

We position people-weighted semantic clouds as a first-pass *orientation* tool. They help teams see what many participants noticed, seed codebooks, and communicate trade-offs. They do not obviate careful reading, synthesis, or theory-building. This stance aligns with prior HCI work that treats semantic grouping and hybrid visual encodings as aids to human reasoning rather than endpoints.

6.5 Interactivity and explanation

Static clouds are useful, but interactive affordances (such as hovering to see exemplar quotes, clicking to open transcripts, showing per-condition contrasts, toggling scaling) can turn the cloud into a navigational entry point for analysis. Because we persist per-transcript assignments, simple linkages suffice. We leave richer explanation (minimal rationales for concept presence) as future work consistent with analyst agency (Iijima et al., 2021; Wang et al., 2020; Chandrasegaran et al., 2019; Skeppstedt et al., 2024).

Future work will focus on improving model transparency, allowing users to inspect why certain phrases were chosen or how decisions were made at the transcript level, for instance in clinical decision support tools such as (Lin et al., 2023b,c, 2025). We are also exploring ways to incorporate multi-turn refinement and lightweight feedback mechanisms, enabling more dynamic human-LLM collaboration. In parallel, more formal evaluations across domains and user roles will be important to assess the tool's effectiveness, trustworthiness, and usability in varied qualitative research contexts.

7 Artifact

Our open-source ThemeClouds package [1] includes: (1) prompt templates for concept elicitation and per-transcript mapping; (2) scripts to reproduce Figure 2; and (3) anonymized assignment tables and per-concept participant counts suitable for auditing and alternative visualizations. The artifact also documents default parameters and prompt variants, so other researchers can reproduce and adapt the pipeline without brittle prompt hacking. We hope this work encourages further exploration into how LLMs can provide insight in qualitative workflows.

8 Conclusion

We introduced ThemeClouds, a participant-weighted, concept-level approach to word clouds using LLMs to count *who* raised *which* ideas, aligning early-stage summaries with the way HCI and UX analysts argue salience. In an audiovisual (AV) study for clinical assessment, the method surfaced actionable concerns that frequency clouds and topic-modeling baselines obscured. By emphasizing transparency, agency, and auditability, it bridges NLP advances and qualitative practice, offering a pragmatic step toward interactive, human-centered, LLM-assisted analysis.

Acknowledgments

We thank the participants and research staff who made this study possible, and colleagues who provided feedback during development. This work is supported by NIH grant 1U01MH136535.

[1] https://github.com/linlab/ThemeClouds

References

Lukas Barth, Sara Irina Fabrikant, Stephen G Kobourov, Anna Lubiw, Martin Nöllenburg, Yoshio Okamoto, Sergey Pupyrev, Claudio Squarcella, Torsten Ueckerdt, and Alexander Wolff. 2014. Semantic word cloud representations: Hardness and approximation algorithms. In *Latin American Symposium on Theoretical Informatics*, pages 514–525. Springer.

Scott Bateman, Carl Gutwin, and Miguel Nacenta. 2008. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tung-Yen Chuang, and Kwan-Liu Ma. 2019. Talktraces: Real-time capture and visualization of verbal content in meetings. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Christel Hopf. 2004. Qualitative interviews: An overview. *A companion to qualitative research*, 203(8):100093.

Ryo Iijima, Akihisa Shitara, Sayan Sarcar, and Yoichi Ochiai. 2021. Word cloud for meeting: A visualization system for dhh people in online meetings. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4.

Shah Khusro, Fouzia Jabeen, and Aisha Khan. 2021. Tag clouds: past, present and future. *Proceedings of the national academy of sciences, India section A: physical sciences*, 91(2):369–381.

Robert P Lennon, Robbie Fraleigh, Lauren J Van Scoy, Aparna Keshaviah, Xindi C Hu, Bethany L Snyder, Erin L Miller, William A Calo, Aleksandra E Zgierska, and Christopher Griffin. 2021. Developing and testing an automated qualitative assistant (aqua) to support qualitative analysis. *Family medicine and community health*, 9(Suppl 1):e001287.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Ravi Tejwani. 2023a. Neural topic modeling of psychotherapy sessions. In *International workshop on health intelligence*, pages 209–219. Springer.

Baihan Lin, Djallel Bouneffouf, Yulia Landa, Rachel Jespersen, Cheryl Corcoran, and Guillermo Cecchi. 2025. Compass: Computational mapping of patient-therapist alliance strategies with language modeling. *Translational Psychiatry*, 15(1):166.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023b. Psychotherapy ai companion with reinforcement learning recommendations and interpretable policy dynamics. In *Companion Proceedings of the ACM Web Conference 2023*, pages 932–939.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023c. Supervisorbot: Nlp-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 7149–7153.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Radim Rehurek, Petr Sojka, and 1 others. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Anna W Rivadeneira, Daniel M Gruen, Michael J Muller, and David R Millen. 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 995–998.

Satpreet Harcharan Singh, Kevin Jiang, Kanchan Bhasin, Ashutosh Sabharwal, Nidal Moukaddam, and Ankit B Patel. 2024. Racer: An llm-powered methodology for scalable analysis of semi-structured mental health interviews. *arXiv preprint arXiv:2402.02656*.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, and Matts Lindström. 2024. From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, 23(3):217–238.

Julie Steele and Noah Iliinsky. 2010. *Beautiful visualization: Looking at data through the eyes of experts*. " O'Reilly Media, Inc.".

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ji Wang, Jian Zhao, Sheng Guo, Chris North, and Naren Ramakrishnan. 2020. Recloud: semantics-based word cloud visualization of user reviews. In *Graphics Interface 2014*, pages 151–158. AK Peters/CRC Press.

Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, and 1 others. 2025. Tama: A human-ai collaborative thematic analysis framework using multi-agent llms for clinical interviews. *arXiv preprint arXiv:2503.20666*.