

# Large Language Model-Assisted Qualitative Analysis: A Case Study of Medical Students with Academic Difficulties

Xiaohan Zhang, Shaohua Li\*

Renji College, Wenzhou Medical University, Wenzhou 325000, China

\*Corresponding author email: nncayenne@163.com

---

**Abstract:** Large language models (LLMs) are increasingly used to support qualitative research, yet practical accounts of how they can be incorporated without displacing human interpretation remain limited. This study presents a researcher-supervised workflow for processing and analysing semi-structured interviews with medical students experiencing academic difficulties. Thirty interviews were transcribed, checked against recordings, de-identified, and analysed through three stages: event/theme identification, element-level examination of focal events, and cross-event interpretive construction. The workflow identified recurring difficulties involving unclear learning goals, limited learning resources, professional-identity tension, and family expectations. LLMs supported the organisation of long transcripts and the comparison of related experiences; researchers verified outputs, retained contextual qualifications, and developed the final interpretation. The findings illustrate that LLMs can make qualitative analysis more traceable and efficient when their outputs are treated as provisional analytic material rather than authoritative conclusions. We propose practical safeguards for transcription, privacy, validation, and documentation. This case study offers a concise, reproducible pathway for using LLMs to assist—not replace—qualitative inquiry.

**Keywords:** Large language models; qualitative research; medical students; thematic analysis.

---

## 1. Introduction

Qualitative research is indispensable for examining how people understand and navigate complex experiences. Its strength lies in contextual interpretation, but its iterative practices—transcription, coding, comparison, memoing, and theory building—are time-intensive and difficult to document in full. These features can make it challenging for readers to see how an account was constructed, even when researchers follow established standards for credibility and reflexivity (Denzin & Lincoln, 2011; Levitt et al., 2018).

Recent LLMs can summarise text, extract candidate themes, compare passages, and produce structured outputs. Studies have therefore explored their use in deductive coding, inductive thematic analysis, and replication of human analyses (Chew et al., 2023; Dai et al., 2023; Jalali & Akhavan, 2024). Their value is not simply speed. Because a model can work across a long transcript or a structured batch of segments, it can offer a consistent first-pass view of recurring language and relationships. At the same time, fluent output can conceal unsupported inferences, cultural misunderstandings, and inherited biases. LLM-generated text cannot, by itself, establish analytic validity (Roberts et al., 2024).

The central methodological question is consequently not whether an LLM can “do” qualitative analysis. It is how researchers can organise collaboration with an LLM while retaining responsibility for meaning, ethics, and theoretical claims. Existing reports often focus on a single task, such as code generation, rather than on the linked pathway from audio-derived text to an interpretive account. A compact, transparent workflow is particularly useful for applied fields where researchers work with sensitive interview data and need to explain their decisions clearly.

This study addresses that need through a case study of

medical students with academic difficulties. It asks: (1) how can LLMs support transcription refinement, de-identification, and qualitative analysis; (2) what analytic insights can be produced through a staged event/theme-centred workflow; and (3) what human checks are necessary to preserve qualitative rigour? Our purpose is not to test whether a particular model is superior to human analysts. Instead, we demonstrate a practical procedure in which model outputs remain provisional and auditable.

## 2. Method

### 2.1. Design and participants

This qualitative case study used purposive sampling to recruit students from a medical university in Zhejiang Province, China. Eligible students were currently enrolled and had accumulated more than 10 failed credits. Thirty students completed the core interview sample (20 men and 10 women; 18–21 years old; failed credits 10–23.5). After preliminary analysis, two additional interviews produced no substantially new themes; the core analysis reported here therefore uses the original 30 interviews. All participants provided written informed consent. The study was approved by the Ethics Review Board of Kangning Hospital (YJ-2023-27-01).

Semi-structured face-to-face interviews lasted approximately 40–60 minutes. Questions covered learning motivation, self-understanding, learning strategies, physical and psychological states, social support, and adaptation to the educational environment. This broad guide was used to invite participants to describe both immediate academic problems and the contexts in which those problems developed.

Interviews were conducted in a private setting by trained members of the research team. Participants were encouraged to describe concrete episodes rather than provide only general

evaluations of their academic performance. Interviewers used follow-up questions to clarify sequence, perceived causes, significant others, and consequences. This approach generated narratives that could be examined as connected experiences while also allowing participants to introduce concerns not anticipated in the interview guide. Field notes recorded interview conditions and immediate methodological reflections; they were used during later discussions to avoid treating the transcript as a complete representation of the encounter.

## 2.2. Data preparation and privacy protection

Recordings were first converted to text using speech-recognition software. An LLM was then asked to flag likely recognition errors, punctuation problems, repetitions, and ambiguous words while preserving the wording and sequence of the speaker. It did not replace the audio record. Researchers checked suggested revisions against the recording, especially where accents, colloquial expressions, emotional speech, or specialised terminology could alter meaning. The verified transcript, rather than the model output, was treated as the analytic source.

Before analysis, identifying information was replaced with consistent placeholders. Names, precise locations, institutional units, dates, and distinctive biographical details were generalised where necessary. An LLM assisted in flagging possible identifiers, but a researcher reviewed each change because de-identification can itself remove context or overlook indirect identifiers. Only de-identified excerpts and process materials are suitable for sharing; raw interview recordings and identifiable transcripts are not treated as open data.

## 2.3. Three-stage analytical workflow

Analysis followed a three-stage workflow adapted from iterative thematic and grounded-theory-informed practices (Braun & Clarke, 2006; Charmaz, 2014). The unit of early analysis was a relatively coherent event or theme, rather than an isolated word or sentence. This choice allowed the team to preserve actors, setting, sequence, and consequences while using an LLM to organise extensive text. For every stage, researchers compared outputs with source passages, revised prompts when necessary, and recorded the rationale for accepting, rejecting, or modifying suggestions.

**Table 1.** Three-Stage Workflow for LLM-Assisted Qualitative Analysis

| Stage                        | LLM-supported activity                                                                | Researcher responsibility                                                                       |
|------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| 1. Identify events/themes    | Produce a candidate list of coherent experiences; group related passages.             | Check boundaries and wording against transcripts; select analytically meaningful themes.        |
| 2. Examine focal events      | Organise material by context, actors, actions, emotions/cognitions, and consequences. | Verify nuance; preserve contradictions and negative cases; choose representative excerpts.      |
| 3. Construct interpretations | Suggest cross-event links and a preliminary relationship map.                         | Compare cases, assess plausibility, relate claims to evidence, and formulate the final account. |

Prompts specified the task, the expected output format, and the instruction not to invent details. The model was given segmented text when needed, followed by cross-segment comparison. The team retained prompt versions, outputs, and decision notes as an audit trail. This documentation does not eliminate subjectivity; rather, it makes human and machine contributions more visible (Aguinis & Solarino, 2019; Tracy, 2019).

Credibility was approached through iterative comparison rather than through a claim of mechanical objectivity. Two researchers independently reviewed the candidate event list and discussed differences in event boundaries, labels, and supporting passages. Where the model suggested a theme that was not well supported, the suggestion was removed or reframed as a question for re-reading. Where researchers disagreed, they returned to the full interview rather than deciding on the basis of a summary. This procedure was especially important for distinguishing a recurring pattern

from a striking but isolated statement. The resulting analysis therefore combines structured model assistance with familiar qualitative practices of constant comparison, reflexive discussion, and attention to deviant cases.

## 3. Results

### 3.1. Event/theme identification

The first-stage review generated eight candidate events/themes. Four were most central to the account of sustained academic difficulty: lack of clear learning goals, limited learning resources and strategies, tension around professional identity, and family expectations combined with financial or emotional pressure. These themes were not independent variables. Participants often described them as overlapping experiences that changed the amount of time and energy they could devote to study.

**Table 2.** Focal Events/Themes Identified in Interviews With Medical Students Experiencing Academic Difficulties

| Focal event/theme                | Illustrative analytic content                                                                                      | Contribution to interpretation                                             |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Unclear learning goals           | Uncertainty about what to learn, why the major mattered, and how present courses related to future work.           | Reduced sustained study investment.                                        |
| Limited resources and strategies | Difficulty planning, monitoring progress, seeking help, or recovering after failed assessments.                    | Made setbacks accumulate rather than resolve.                              |
| Professional-identity tension    | Mismatch between personal interests, perceived abilities, and expectations attached to medical training.           | Weakened commitment and confidence.                                        |
| Family expectations and pressure | Concern about family sacrifice, expectations for achievement, and the emotional burden of disappointing relatives. | Created pressure that could motivate briefly but also intensify avoidance. |

The LLM was useful at this stage because it could bring together passages that appeared at different points in the same interview. For example, a participant might initially discuss skipping study sessions, later mention uncertainty about career prospects, and finally describe distress after family conversations. The model could propose a connection among these passages. Researchers accepted such connections only after reading the surrounding narrative and checking that the participant, rather than the model, supported the relationship.

The theme list was deliberately kept at the level of experience rather than converted immediately into abstract psychological labels. This helped avoid a common risk of automated analysis: replacing participants' situated accounts with smooth but overly general categories. "Limited learning resources and strategies," for instance, included practical problems such as uneven study routines, difficulty locating help, and uncertainty about how to respond after a failed examination. It was not interpreted simply as lack of effort. Similarly, family pressure contained both support and strain. Some students described family concern as an impetus to persist, whereas others described it as intensifying self-blame. Keeping these contrasts visible prevented the analysis from treating a theme as uniformly harmful or beneficial.

The event/theme-centred format also allowed the team to compare the temporal ordering of experiences. For some students, uncertainty about professional fit appeared before academic decline; for others, it emerged after repeated poor performance. The LLM could identify passages with temporal markers and place them in a provisional sequence, but the sequence was checked manually because interview narratives often move back and forth in time. This comparison did not establish causal order. It did, however, show why a simple checklist of risk factors would be insufficient: similar factors could acquire different meanings depending on when and how they entered an individual student's story.

### 3.2. Element-level examination of a focal event

"Lack of learning motivation and clear goals" was examined in greater detail. The second-stage template separated the event into context, trigger, actions, emotional and cognitive responses, social influences, and consequences. This decomposition showed that motivation was rarely described as a stable personal trait. It was often linked to repeated unsuccessful efforts, uncertainty about whether the chosen major fitted the participant, and difficulty imagining a credible future role.

The model helped organise these elements in a consistent format across participants. However, human review was essential for two reasons. First, an expression of temporary exhaustion could be incorrectly elevated into a general claim of "low motivation." Second, the same reported action—for example, avoiding a class—could reflect boredom, shame, anxiety, illness, or competing family responsibilities. The team therefore used the structured output as a comparison aid, not as a substitute for interpretation. Disconfirming passages and within-case variation were recorded alongside recurrent patterns.

### 3.3. Interpretive construction

Across events, the analysis suggested a low-investment cycle. Uncertainty about goals and professional fit weakened engagement with difficult coursework. Limited study strategies and repeated failures then reduced confidence and

made future-oriented planning less credible. Family expectations could provide short-term encouragement, but when experienced as pressure they also increased guilt and avoidance. The resulting reduction in study investment exposed students to further setbacks, reinforcing the original uncertainty.

This cycle is an interpretive model, not a causal prediction. It captures one plausible account of how multiple experiences became mutually reinforcing in this case. The LLM contributed by listing cross-event relationships and by helping the team compare them across interviews; the final model was produced through iterative discussion, return to source material, and attention to alternative explanations. Several participants did not fit the pattern neatly, reminding the team that the model should guide inquiry rather than classify individuals.

The interpretive construction was strengthened by repeatedly asking what evidence would challenge the proposed cycle. In some interviews, students maintained high effort despite uncertainty about their major; in others, clear career goals did not protect against practical obstacles or emotional distress. These cases prevented the team from presenting the model as a single pathway. Instead, they suggested that academic difficulty develops through configurations of conditions. The analytic task was to describe the configuration evident in the data, including its variations, rather than to infer an invariant personal deficit. The LLM was helpful in retrieving passages relevant to these comparisons, but it did not determine which differences were theoretically important. That decision required researchers to consider the narrative context, the study purpose, and the risk of oversimplifying participants' experiences.

## 4. Discussion

This case study shows how LLMs can be incorporated into qualitative work without treating them as autonomous analysts. The practical contribution is the use of event/theme-centred units across three linked stages. Compared with an exclusively line-by-line approach, these units retain a recognisable episode of experience and make it easier to move between close reading and cross-case comparison. The approach is compatible with established qualitative practices while taking advantage of an LLM's ability to structure large amounts of text.

The findings also clarify the division of labour. LLMs were most helpful for repetitive, organisational work: surfacing candidate passages, grouping related material, formatting element-level summaries, and making preliminary links explicit. Researchers performed the work that carries interpretive and ethical responsibility: deciding whether a proposed theme was warranted, preserving ambiguity, recognising culturally situated meanings, evaluating exceptions, and deciding what could be reported. This division is consistent with the view of LLMs as research assistants rather than sources of independent qualitative knowledge (Bano et al., 2023; Roberts et al., 2024).

A further implication concerns transparency. Qualitative reproducibility should not mean that every team must reach identical conclusions from a contextual dataset. A more defensible goal is analytic traceability: readers should be able to see the dataset boundaries, prompts, outputs, revisions, and decisions that produced an interpretation. Retaining these materials enables reasoned comparison across analysts while

preserving the possibility of alternative readings. It also makes model-specific influence visible, which is important because outputs can change with the model, settings, and prompt wording.

This perspective reframes an LLM-generated analysis as a benchmark for dialogue rather than a benchmark for correctness. A shared first-pass output can give multiple researchers a common object to critique: they may confirm a suggested grouping, recover an omitted context, or defend a competing interpretation with evidence. Such use can be valuable in team-based projects, where differences between analysts are often informative but difficult to record. It should not be confused with validation by agreement with the model. The relevant question remains whether a proposed interpretation is grounded in the material, coherent with the research question, and appropriately limited by the available evidence (Morse et al., 2002; Gioia et al., 2013).

The workflow has practical safeguards. Audio verification is necessary when models revise transcripts; automated “corrections” may silently change a speaker’s meaning. De-identification requires manual review, because a model may miss contextual identifiers or erase analytically relevant detail. During analysis, prompts should request evidence-linked output and prohibit invented quotations. Researchers should maintain an explicit log of model version, prompt changes, and decisions, and should report how disagreements or implausible outputs were handled. These procedures are modest, but they convert an opaque interaction into a reviewable research process.

The workflow is also intended to be feasible for researchers without advanced programming expertise. Its basic components are familiar: a secure transcript-management procedure, clearly delimited prompts, a structured worksheet for reviewing outputs, and regular team discussion. The important requirement is not technical complexity but disciplined boundary-setting. Each task should be small enough that a researcher can inspect the source evidence behind the result. For example, an event list can be reviewed against linked transcript passages before it is used to generate a more detailed template; a relationship map can be treated as a draft for discussion instead of as a finished finding. Such staged use reduces the temptation to accept a polished answer merely because it appears coherent. It also allows the protocol to be adjusted to local ethical rules, language practices, and the analytic tradition of a particular study.

### 4.1. Limitations

The study is based on a single group of students in one educational context, and the interpretive model should not be generalised as a universal explanation of academic difficulty. The analysis was also conducted with a particular LLM workflow; output quality may vary by model, language, prompt design, and context-window limitations. We did not conduct a formal quantitative comparison of human and model coding, nor do we claim that the model’s suggestions were more accurate than those of human analysts. Finally, interview material concerning family expectations, identity, and distress is culturally and emotionally nuanced; some meanings may remain inaccessible to automated text processing. Future research could compare independent human teams, test the workflow in other populations, and examine how model versioning affects analytic stability.

## 5. Conclusion

Large language models can support qualitative research by making transcription refinement, de-identification checks, thematic organisation, and cross-case comparison more manageable. Their appropriate role is assistive and auditable. In this case study, a three-stage event/theme-centred workflow helped the research team develop an account of academic difficulty while keeping researchers responsible for verification and interpretation. The value of LLM-assisted qualitative analysis lies not in automated conclusions, but in a transparent collaboration that preserves contextual reading, ethical care, and human judgement.

For future studies, reporting should describe the model and version used, the data supplied to it, the sequence of analytic tasks, the main prompt templates, and the points at which researchers altered or rejected outputs. These details will help readers judge the scope of model involvement and will make constructive replication possible. In this way, the adoption of LLMs can strengthen—not dilute—the distinctive commitment of qualitative research to carefully evidenced, context-sensitive interpretation.

## Declarations

**Ethics approval and consent to participate.** The study was approved by the Ethics Review Board of Kangning Hospital (YJ-2023-27-01). Written informed consent was obtained from all participants.

**Data availability.** De-identified process materials may be made available on reasonable request, subject to ethical review and protection of participant confidentiality. Raw recordings and identifiable transcripts are not publicly shared.

**Competing interests.** The authors declare no competing interests.

**Funding.** This work was supported by the 2026 Wenzhou Philosophy and Social Sciences Planning Special Project for Ideological and Political Work (Grant No. 26SZH04YBM).

## References

- [1] Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, 40(8), 1291–1315. <https://doi.org/10.1002/smj.3015>
- [2] Arlinghaus, C. S., Wulff, C., & Maier, G. W. (2024). Inductive coding with ChatGPT: An evaluation of different GPT models clustering qualitative data into categories. *Open Science Framework*. <https://doi.org/10.31219/osf.io/gpny>
- [3] Bano, M., Zowghi, D., & Whittle, J. (2023). Exploring qualitative research using large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2306.13298>
- [4] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [5] Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Sage.
- [6] Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv*. <https://doi.org/10.48550/arXiv.2306.14924>
- [7] Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). LLM-in-the-loop: Leveraging large language models for thematic analysis. *arXiv*. <https://doi.org/10.48550/arXiv.2310.15100>

- [8] De Paoli, S., & Mathis, W. S. (2024). Reflections on inductive thematic saturation as a potential metric for measuring the validity of an inductive thematic analysis with LLMs. *Quality & Quantity*. <https://doi.org/10.1007/s11135-024-01950-6>
- [9] Denzin, N. K., & Lincoln, Y. S. (2011). *The SAGE handbook of qualitative research* (4th ed.). Sage.
- [10] Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research. *Organizational Research Methods*, 16(1), 15–31. <https://doi.org/10.1177/1094428112452151>
- [11] Jalali, M. S., & Akhavan, A. (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review*, 40(3), e1772. <https://doi.org/10.1002/sdr.1772>
- [12] Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology. *American Psychologist*, 73(1), 26–46. <https://doi.org/10.1037/amp0000151>
- [13] Li, K., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and human researchers in health care data analysis: Qualitative description study. *Journal of Medical Internet Research*, 26, e56500. <https://doi.org/10.2196/56500>
- [14] Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.
- [15] Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods*, 1(2), 13–22. <https://doi.org/10.1177/160940690200100202>
- [16] Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model assistance. *Critical Perspectives on Accounting*, 99, 102722. <https://doi.org/10.1016/j.cpa.2024.102722>
- [17] Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.
- [18] Tracy, S. J. (2019). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact* (2nd ed.). Wiley-Blackwell.