ChatGPT-Assisted Codebook Design for Learning Analytics Datasets in Multiple Languages: A Case Study

Ayaz Karimov[1[0000-0001-8868-4408]], Mirka Saarela[1[0000-0002-1559-154X]], Xiner Liu[2[0009-0004-3796-2251]], Zhanlan Wei[2[0009-0002-3931-6398]], Andres Felipe Zambrano[2[0000-0003-0692-1209]], Amanda Barany[2[0000-0003-2239-2271]], Ryan S. Baker[2[0000-0002-3051-3232]], Jaclyn Ocumpaugh[2[0000-0002-9667-8523]], Sabina Mammadova[3[0009-0002-1071-0818]], and Tommi Kärkkäinen[1[0000-0003-0327-1167]]

[1] University of Jyväskylä, Jyväskylä, Finland
[2] University of Pennsylvania, Philadelphia, PA, United States
[3] One World International School Singapore, Singapore, Singapore
ayazkarimov3@gmail.com

Abstract. This study investigates the use of ChatGPT for the development of inductive (data-driven) codebooks from qualitative datasets in underrepresented languages (Azerbaijani and Finnish). Although prior work has leveraged GPT-4o as a "co-researcher" that can support more efficient and comprehensive inductive codebook development, further work is needed to understand how consistent results are across languages and for translated text. The study found GPT- 4o to be useful for identifying relevant codes, but also found limitations, particularly in terms of the quality of example sentences generated for less-resourced languages. Social moderation by humans and construct evaluations were applied to refine the generated codebooks to ensure clarity and reduce redundancy. The results demonstrated that, while GPT-4o significantly aids in multilingual qualitative analysis, human intervention remains essential to validate and enhance the accuracy of the outputs. This research is particularly significant for the learning analytics field as it demonstrates scalable methods for multilingual qualitative analysis, a critical step in expanding the inclusivity and applicability of learning analytics across educational contexts.

Keywords: AI-assisted research, AI in qualitative research, ChatGPT, codebook development, multilingual qualitative research, thematic analysis

1 Introduction

Qualitative coding has many applications in learning analytics research, but remains highly time-consuming, producing a need for scalable, efficient solutions that can produce consistent, high-quality, and comprehensive qualitative codebooks [27, 30]. This demand is particularly relevant in multilingual learning analytics research, where analyzing data across multiple languages introduces challenges in terms of comparability and feasibility. Addressing these challenges requires tools that can bridge linguistic and contextual gaps, allowing researchers to explore similar educational phenomena in multiple cultural settings.

Large Language Models (LLMs) such as GPT (Generative Pre-trained Transformer) offer promising solutions in this space. LLMs can process natural language in highly flexible ways, which can assist researchers by complementing human judgment in hybrid qualitative coding workflows, including suggesting categories [3, 12], proposing

category refinements [9], and explaining both human and machine decisions and disagreements [39, 40]. However, further research is needed to explore their strengths and weaknesses, particularly in conducting these forms of research for underrepresented languages and contexts. Initial findings suggest that LLMs may produce different outcomes when analyzing data in underrepresented languages, defined for this purpose as languages less frequently represented in their training datasets [e.g., 6]. Traditional qualitative coding methods often require deep linguistic and cultural expertise, and while LLMs like GPT may match this expertise to varying degrees, the results may differ significantly based on the language of the data. This challenge persists even when text is translated into more commonly used languages such as English, as effective coding often depends on context that may not be fully present in the translated text or recognized by English-speaking coders.

Therefore, research is needed to better understand when and how LLMs like GPT can accurately identify qualitative themes in data from underrepresented languages. This paper explores the use of ChatGPT (GPT-4o, selected as it was the most current version available at the time of analysis and for its conversational capability to engage in dialogue with human researchers) to generate thematic codebooks for qualitative data analysis across both translated and untranslated transcripts of underrepresented language texts from two international learning contexts. Specifically, we examine codebooks generated by applying a human-GPT hybrid approach to thematic analysis [3] to focus group transcripts in Finnish and Azerbaijani—both in their original forms and English translations—to explore how this approach may produce high-quality, consistent codebooks across languages and contexts. Our goal is to evaluate whether ChatGPT can offer an effective and efficient method for creating codebooks that enhance cross-language consistency maintaining the cultural and contextual nuances of each language. Thus far, most of the discourse around qualitative analysis using LLMs has involved English-language data. The current paper attempts to address our field's gaps in understanding how LLMs perform for this task across different languages and contexts.

In this study, we investigate ChatGPT-4o's capabilities to support thematic analysis and codebook development process for multilingual datasets including transcripts in the underrepresented languages Azerbaijani and Finnish. In particular, we focus on analyzing how GPT performs when applied to both the datasets in the original native language and in English translations. By using GPT to help develop qualitative codebooks for each dataset, we aimed to assess its consistency across various linguistic and cultural contexts, as well as its potential to reduce manual coding efforts while maintaining accurate interpretation of the data. For this purpose, we follow one of the hybrid human-AI approaches proposed by [3], where GPT is used as a tool that inductively suggests a preliminary list of themes that are then refined by human researchers to ensure accuracy in the definitions and examples. We conducted this research in the context of transcripts from scenario-based focus groups with Azerbaijani and Finnish K-12 teachers which gathered data on their perspectives on learning analytics. ChatGPT was prompted to generate codes from the dataset in both languages and their translations. Afterward, social moderation was conducted by humans to review and refine these codes, ensuring relevance and eliminating redundancy. Additionally, a codebook

similarity evaluation was performed to check consistency with established coding frameworks, followed by a construct quality evaluation to assess the depth and clarity of the constructs generated. Through this process, we can better understand the potential and limitations of GPT-4o for cross-language/underrepresented language hybrid human-AI qualitative codebook development.

2 Background

Qualitative analysis has been widely adopted in learning analytics research and in educational research in general, as it allows researchers to delve into the subtle and often complex phenomena present in educational environments and student behaviors with deeper insights [2]. One of the key components of qualitative analysis is qualitative coding, where researchers identify themes within the data and then label the data to assess the prevalence of these themes and guide subsequent analysis [31]. Qualitative coding can follow either an inductive bottom-up approach—allowing patterns to emerge from the data—or a top-down deductive approach, applying codes from existing codebooks or frameworks [4].

There is nowadays substantial use of qualitative coding in English and other majority research languages, within the learning analytics community and related communities. However, this becomes more challenging for researchers attempting to study phenomenon across countries, where multiple languages may be present in data sets. Despite the increasingly global nature of learning analytics research [1], multilingual qualitative analysis [17] remains uncommon. One major reason is the time and resource-intensive nature of the qualitative process. Tasks such as data collection, transcription, and interpretation are by themselves labor-intensive, and conducting them across multiple languages is even more labor-intensive and requires a nuanced understanding of linguistic and cultural contexts [29].

Moreover, when conducting qualitative coding with data from underrepresented languages, it can sometimes be difficult to find sufficient numbers of individuals who both have the necessary linguistic fluency and qualitative coding expertise [32]. This problem is exacerbated when researchers are attempting to study or compare phenomena across languages [13] – few researchers are fluent in all languages relevant to a multi-country study. Furthermore, when native speakers code data on behalf of researchers, effective communication between researchers and linguistic experts can be challenging, particularly if the research team and the native speaker come from different cultural or academic backgrounds [15]. Coordinating iterative rounds of review, social moderation, and validation between multilingual team members often requires additional layers of communication and negotiation [25]. Even when the researchers themselves are highly fluent speakers of the relevant languages, there is still considerable complexity beyond what is seen for monolingual research. As a result, many researchers default to analyzing monolingual data, typically in English, even when multilingual data is available [33, 36]. This reliance on English-centric analyses marginalizes voices from underrepresented linguistic communities, who are often already marginalized, and can

diminish the richness of data by excluding culturally embedded meanings that are not easily translatable [34, 35].

LLMs such as GPT offer a potential solution to many of these challenges by providing scalable support for multilingual text processing, based on being trained on extensive multilingual (though still majority-language dominated) corpora, giving them the capacity to understand and generate coherent responses across a range of languages [26, 37]. LLMs have been utilized in various qualitative research capacities, such as assisting with automatic translation [11], detecting sentiment variations [28], identifying outliers or anomalous patterns [19], suggesting preliminary themes for codebook development [3], facilitating the identification of patterns and themes [20], and suggesting refinements to coding categories generated by humans [39]. However, their application for these tasks has largely been mostly limited to more well-resourced languages (in particular, English).

Even when applied to multilingual contexts, many such studies use LLMs primarily for initial pre-processing tasks, such as language translation [e.g., 22], rather than more complex tasks like thematic analysis or semantic exploration. Translating text into English before conducting qualitative coding [e.g. 23] or coding data in its original language and then translating the codes to English for future analysis [e.g. 10] are established processes in qualitative analysis. However, these approaches have a core limitation; language often conveys complex meanings based on the cultural contexts in which it is situated [7], and key ideas are often difficult to translate [14, 38]. Therefore, developing better methods to support researchers in understanding and interpreting meaning across languages could increase the feasibility and quality of multi-cultural, multi-national learning analytics research.

3 Methodology

3.1 Data Context

The data used in this research was collected during two scenario-based focus groups conducted online with 10 elementary school STEM teachers from Azerbaijan (4) and Finland (6) [18]. In each focus group, teachers were presented with hypothetical real-life scenarios where they faced specific challenges related to the employment of technology and learning analytics in the classroom. Teachers responded to questions designed to explore their reactions to these challenges, aiming to understand teachers' perspectives on the use of learning analytics and educational technology platforms. Each focus group was held via Zoom in the participants' native language and lasted approximately 90 minutes. The sessions were fully recorded, and participants were informed of the data collection and processing procedures both through pre-focus group emails and at the beginning of the sessions. Participants were also briefed on the research goals and consent procedures via email prior to the focus groups. To maintain confidentiality, teachers' identifiable information was redacted throughout the reporting process. In cases where teachers referred to each other by name during the session, we applied post-hoc hand redaction by replacing any mentioned names with their corresponding de-identified codes to ensure consistent confidentiality.

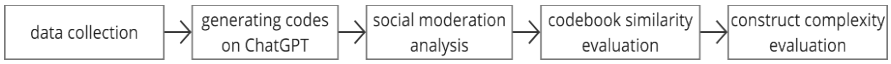


Fig. 1. Process flow for developing codebooks in different languages using ChatGPT.

The selection and development of the research scenarios emerged from a collaborative process among the authors (authors 1-2 and author 8) that aimed at ensuring that each scenario was both relevant to educators and reflective of real-world teaching challenges. The process began with an initial brainstorming session, where those researchers, drawing from their diverse backgrounds in education, learning analytics, and educational technology, identified key themes and issues commonly faced by teachers when integrating technology into their classrooms. Once the central themes—such as personalized learning, data accessibility, and ethical considerations—were established, six scenarios were initially designed to reflect real-life situations that would be familiar for teachers and prompt meaningful discussion about the use of learning analytics in the classroom. Each scenario was reviewed and refined by the group. To accommodate the language needs of the focus groups, the scenarios were initially developed in English and then translated into both Finnish and Azerbaijani by researchers fluent in those languages, as the focus groups were conducted in these languages. To ensure the scenarios elicited valuable insights, follow-up questions were also crafted alongside each scenario to further guide participants in exploring the practical and ethical implications of integrating learning analytics into classroom, and everyday teaching practices.

Appendix A (see https://zenodo.org/records/15558223) details the six scenarios, which included (1) Learning Analytics for a Virtual DNA Simulation, (2) Personalized Approach to Student Struggles, (3) Limited Access to Student Performance Data, (4) Presenting Student Performance through De- tailed Data, (5) Learning Analytics as a Tool to Communicate with Parents, and (6) Ethical Considerations in the Adoption of Learning Analytics.

3.2 Codebook Generation

Codebook Development. We used the web-based version of ChatGPT (GPT-4o) to generate initial draft codebooks of qualitative themes from the transcripts. To prevent any potential influence from prior interactions between the authors and the chatbot, we disabled the memory feature and ensured that all previous conversations were deleted after each prompt execution. For the English translations of the Finnish and Azerbaijani datasets, we used prompts in English, while for the Azerbaijani and Finnish datasets, we used prompts in their native languages (Table 1).

Table 1: GPT-4o prompts across languages

Language	Prompt
English	Hi ChatGPT, I want to do a qualitative data analysis of the attached document, please provide a detailed codebook with constructs, definitions and examples for analyzing it.

Language	Prompt
Azerbaijani	Salam ChatGPT, mən əlavədə olan sənədin keyfiyyət təhlilini (qualitative analysis) aparmaq istəyirəm, zəhmət olmasa, onu təhlil etmək üçün konstruksiyalar, təriflər və nümunələr ilə ətraflı kod kitabçası təqdim et.
Finnish	Hei ChatGPT, haluaisin tehdä laadullisen temaattinen analyysi liitteenä olevasta asiakirjasta. Anna yksityiskohtainen koodikirja, jossa on rakenteita, määritelmiä ja esimerkkejä sen analysoimiseksi.

Since GPT performs better with smaller segments of data, as noted by [39], we divided the dataset into five pieces, corresponding to the six scenarios. Each segment (75-85 lines) was organized manually by the first author to ensure that one discussion topic was completed before the next topic began. Given that GPT may produce varying responses even with the same prompt and input data, we repeated the same prompt for each data segment three times and collected all ChatGPT responses to verify consistency of the outputs. We conducted analysis both on original versions of each data set and their English translations, which were generated by ChatGPT and verified by a native speaker. In total, we prompted GPT 60 times (3 iterations/runs× 5 data segments × 2 Languages x 2 Translations). Moreover, every time the model generated definitions, the first author reviewed the content to ensure that the outputs remained grounded in the actual data and the original transcripts. Throughout this process, no hallucinations were observed.

Addressing Redundancy and Relevance through Social Moderation. During the sessions, researchers systematically reviewed all codes and eliminated 1) redundant coding categories and 2) coding categories that only applied to one instance in the data and therefore lacked broader relevance. We used social moderation [16] to evaluate and consolidate the coding categories generated by GPT. During social moderation, multiple researchers collectively discuss and review the data to reach an agreement on codes' definitions and interpretations [5]. Each dataset was discussed in a separate meeting between two researchers in the relevant languages (author 1 and author 9 for the Azerbaijani dataset and author 1 and a native speaker for the Finnish dataset), for a total of four 45-minute sessions. Any disagreements between researchers were resolved through discussion, and if necessary, a third researcher (author 2) was consulted to reach a final consensus. This refinement process created consistent final codebooks for each dataset and addressed conceptual redundancy between codes, which was important for the broader comparisons across languages outlined in Section 3.2.4.

Code Quality Evaluation. After the social moderation analysis, we conducted a separate process in which the codes were independently reviewed to decide on their inclusion in the final codebook. For the Azerbaijani dataset, two researchers (author 1 and author 9) fluent in Azerbaijani performed the review. For the Finnish dataset, author 1 collaborated with two Finnish native speakers, who independently reviewed the codes and supported us in these decisions. This step, distinct from the construct quality evaluation, focused on ensuring the practical relevance and validity of each code. The researchers' assessments added an additional layer of refinement for use in the subsequent analysis. To evaluate the complexity and effectiveness of the constructs generated by GPT for qualitative coding, we con- ducted a second round of assessment following the social moderation analysis. This second evaluation focused on the codes that had passed through the social moderation process. We sought expert evaluations from two coders for each language (author 1 and

author 9 for the Azerbaijani dataset & author 1 and a native speaker for the Finnish dataset) who were already familiar with the dataset, using a set of dimensions previously used to analyze construct definitions in qualitative coding, including clarity, concreteness, objectivity, granularity, and specificity, as well as the usefulness of examples for each code ([21], see Table 2). Ratings were given using a 1 to 5 Likert scale, where 1 indicated low performance and 5 indicated high performance.

Table 2: Code quality assessment items

Assessment	Definition
1. Clarity	(as opposed to ambiguity) measures how well-defined and easily comprehensible a concept is, without ambiguity or confusion.
2. Concreteness	(as opposed to abstractness) refers to the quality of being specific, tangible, and perceptible by the senses. In other words, concrete concepts are those that can be observed, experienced, or measured in a direct manner, where abstractness refers to the quality of being theoretical, conceptual, or non-physical.
3. Objectivity	(as opposed to subjectivity) means verifiable information based on facts and evidence. Subjective means information or perspectives based on feelings, opinions, or emotions.
4. Granularity	(as opposed to coarseness) concerns the ability to represent and operate on different levels of detail in data, information, and knowledge that are located at their appropriate level.
5. Specificity	(as opposed to generality) measures the extent to which a construct is precise and distinct, rather than conflating it with other related concepts. Generality, on the other hand, refers to the quality of being broad, general, or inclusive of a wide range of subjects, ideas, or situations.
6. Quality of example sentence	Evaluates the usefulness, clarity, and alignment of the example sentences generated for each code in helping coders understand and apply the construct.

In addition to rating these concepts across the 6 dimensions, the coders of the dataset—author 1 and author 8 for the Azerbaijani dataset, and author 1 and author 2 for the Finnish dataset — also assigned a simple "yes" or "no" rating to indicate whether each construct was useful. This rating was applied because even though a construct could receive a low rating in some of these categories, human coders could find it valuable to be included. On the other hand, a construct with high ratings could still have overly high overlap with other constructs or not appear frequently in the dataset, and therefore, human coders would prefer to exclude it from the final codebook.

Conceptual Overlap Evaluation. To evaluate the conceptual overlap identified across datasets, author 1 (a highly fluent speaker of all three languages) identified semantically similar codes that appeared across the Azerbaijani, Finnish, and their English translation datasets by grouping them under broader thematic categories. For instance, overlapping codes like *Parental Involvement* and *Teacher-Parent Communication* were considered as the same thematic category (*Parental Involvement*). The final labels were selected based on their ability to clearly define the underlying concepts, ensuring that each category was distinct and non-overlapping. Definitions were standardized across datasets to establish a common frame for comparison and evaluation.

4 Results

4.1 Preliminary Code Generation

We prompted ChatGPT 60 times (3 runs/iterations × 5 data segments × 4 datasets) obtaining 477 codes in total across all datasets. The number of codes generated in each query differed depending on the language of the dataset. For the Azerbaijani dataset, 72 codes were generated when the dataset was presented in its native Azerbaijani language, but that number nearly doubled (N=130) when the same dataset was translated into English. The Finnish dataset followed a slightly different pattern. When processed in Finnish, it yielded 143 codes, slightly more than the 132 codes generated when translated into English.

4.2 Preliminary Code Generation

The social moderation process considerably reduced the initial number of codes across all datasets. After moderation, a total of 50 codes (10.5%) remained from the original 477 (see Table 3). 15 codes (11.5%) remained for the Azerbaijani dataset in English, 11 codes (15.3%) for the Azerbaijani dataset in Azerbaijani, 11 codes (9.1%) for the Finnish dataset in English, and 13 codes (8.3%) for the Finnish dataset in Finnish.

Table 3: Code quality assessment items

Dataset	Translation	Generated (N)	Selected (N)	Selected (%)
Azerbaijani	Original	72	11	15.3%
	English	130	15	11.5%
Finnish	Original	132	13	8.3%
	English	143	11	9.1%
Total		477	50	10.5%

4.3 Code Quality

The results from the code quality assessment, shown in Appendices B and C (see https://zenodo.org/records/15558223), reflect the evaluations of GPT-generated constructs based on clarity, concreteness, objectivity, granularity, specificity, code utility (binary), and quality of example sentences. Both codebooks created using the Azerbaijani dataset (in Azerbaijani and in English) received consistently high ratings across multiple dimensions of our codebook complexity assessment. Objectivity and clarity were the two categories that obtained the highest average ratings across all the constructs for both codebooks (4.86 and 4.77 in the original Azerbaijani, and 4.97 and 4.70 in the English translation). On the other hand, the perceived quality of examples was the category of the evaluation with the lowest aver- age rating across constructs (3.77 for both codebooks). For instance, constructs such as *Benefits of Learning Analytics (Öyrənmə Analitikasının Faydaları)* and *Challenges Limitations (Problemlər və Limitlər)* received high scores in objectivity (5.0 for all), clarity (4.5 to 5.0), and concreteness (4.0 to 5.0). These constructs also obtained a high rating for granularity and specificity, with most scores ranging between 4.0 and 4.5.

However, the example sentences provided for these constructs received lower scores (between 3.0 and 4.5), which indicates room for improvement in their selection. This suggests that GPT may have difficulty identifying high-quality examples of the constructs it suggests from the dataset. Constructs such as *Cheating & Academic Integrity* and *Real-Life Applications* stood out by receiving a perfect 5.0 in several categories, including objectivity and specificity. Although the constructs were clear and relevant to the human coders, the example sentences received ratings of 5.0 and 3.5, respectively. This variation reflects a discrepancy between the coders' and GPT's interpretation of what constitutes useful examples. For instance*, Improvement Recommendations* received a high average score of 4.5 for all five criteria, but the example sentence ("I suggest using the simulations. I think that it would be better to have key stages in simulations so students may pass to the next step because they also face distraction problems during a lesson.") received a lower score of 2.5 because it focused on platform recommendations rather than student-focused improvements. From the perspective of construct evaluation, there were no substantial differences between the native Azerbaijani and the translated English version, which means that both versions were comparable for GPT in suggesting relevant qualitive codes.

Compared to the codebooks for the Azerbaijani dataset, the codebook for the original Finnish received lower average ratings across several categories, in terms of clarity (3.4), concreteness (3.4), and granularity (3.4), objectivity (3.2), and specificity (3.6) indicating that GPT struggled more with interpreting the dataset in its native Finnish language. In particular, constructs like *Ethical Principles* and *Parental Involvement* showed reduced clarity, concreteness, objectivity, and granularity, with some ratings dropping as low as 1.5. Additionally, the example sentences were rated much lower in the Finnish version, averaging around 3.4. For instance, the example sentence provided for the *Gamified Platforms* code ("The school implements a learning analytics work tool which is specially designed for the study of natural sciences") did not mention any gamification elements such as points, rewards, competition, or game mechanics, which are key features of the gamification. This misalignment between the example sentence and the code exemplifies GPT's difficulty in accurately identifying and providing relevant examples when working with the original Finnish language text.

By contrast, the codebook for the Finnish dataset developed using the English translations showed similar results to those obtained with the Azerbaijani dataset, with the highest ratings achieved for clarity (4.36), concreteness (4.18) and objectivity (4.14) the categories with the highest average ratings across the constructs of this codebook. For example, some constructs like *Teacher's Role* and *Challenges and Limitations* scored highly across most criteria, particularly in clarity (4.5), objectivity (4.5 and 5.0), and granularity (4.5). The *Teacher's Role* construct, defined as "teachers' personal experiences and views regarding the use of educational technologies and learning analytics," reflected a well-understood concept in both datasets. Similarly, the *Challenges and Limitations* construct, defined as "difficulties faced by teachers in implementing and using learning technologies effectively," also received high marks for clarity and objectivity. However, quality of example sentences provided by GPT received the lowest ratings across all the dimensions of the evaluation (i.e., 3.36 in the original Finish and 3.86 in the English translation). For instance, the construct *Parent Roles &*

Communication with Guardians also received a high evaluation, with clarity rated at 5.0 and objectivity at 4.5, but the example sentence quality was rated slightly lower at 4.0. In contrast, constructs like *Gamified Platforms* showed mixed results, with clarity, concreteness, and granularity rated between 3.5 and 4.0, but a much lower rating of 1.5 for the example sentences. One of the example sentences provided by GPT for *Gamified Platforms* was: "The school implements a learning analytics work tool which is specially designed for the study of natural sciences." This sentence received a low rating because it lacks direct relevance to the concept of *Gamified Platforms* and fails to reflect the core idea of gamification within educational tools. Likewise, the *Teacher's Role* construct was better defined in the English translation, receiving a score of 4.0 for the example sentence, while in the original Finnish version, it scored much lower at 2.5. This suggests that the English translation seemed to work better for capturing the nuances of the dataset by allowing for clearer and more relevant examples, which contrasts with the Azerbaijani case where translation into English sometimes led to higher conceptual ambiguity.

4.4 Conceptual Overlap

Across all datasets, 21 unique code categories were identified, each reflecting distinct themes within learning analytics. Appendix D (see https://zenodo.org/records/15558223) highlights the conceptual overlap of these codes across the four datasets. Six codes demonstrated congruence across all four datasets, including *Benefits of Learning Analytics, Challenges of Learning Analytics, Ethical & Privacy Concerns, Student Engagement & Interaction*, and *Teacher Experi- ences & Perceptions*. One code, *Recommendations for Improvement*, appeared in all but the original Azerbaijani data set. Two additional codes, *Interest & Motivation* and *Gamified Platforms*, were found only in the two Finnish codebooks, while one, Acces- sibility, appeared solely in the Azerbaijani codebooks. Four other codes were found in two different codebooks but not in those related by language: they appeared in the Eng- lish translation of the Azerbaijani data and in the original Finnish data (i.e., *Assessment Methods, Platform Utilization, Insights from Data for Effective Teaching*) or in the Azerbaijani original and the English translation of the Finnish (i.e., *Academic Outcomes*). Finally, seven additional codes were present in only one of the Azerbaijani data sets.

5 Discussion

This study explored the effectiveness of using ChatGPT to generate qualitative codebooks for multilingual datasets, focusing on both translated and untranslated data in underrepresented languages such as Azerbaijani and Finnish. One of the most significant findings from this study is the impact of translation on thematic code generation. The translated datasets in English consistently produced more codes than the original language datasets, which suggests that translation may make it easier for ChatGPT to extract the key nuances from the data. This contrasts with findings from previous studies that only involved human coders, which suggested that qualitative coding of translated texts often missed important linguistic subtleties due to the complexities of the

original language and the challenges inherent in translation [7, 34]. It is possible that ChatGPT's limited processing ability in underrepresented languages enabled it to translate but not to interpret within the underrepresented language itself, unlike human speakers of these languages who possess richer comprehension in those languages. These complexities underscore the importance of considering both linguistic and cultural contexts in multilingual research, and the different affordances of human and AI qualitative coders, factors worthy of consideration as multilingual data sets become more prominent in learning analytics research.

The difference in code generation between the Azerbaijani and Finnish datasets also highlights the variability in GPT-4o's performance across languages. Azerbaijani, as an underrepresented language, could have presented a unique challenge for the model, given the limited training data for underrepresented languages (in GPT-3's training dataset, Azerbaijani was the 50th most represented language, consisting of just over 1/100,000th of the data set [24]). However, the model's ability to produce a noticeable number of relevant codes in Azerbaijani still reflects its potential to support qualitative research, even in languages that are often neglected in natural language processing. Despite these successes, the study also reveals the limitations of AI-generated codebooks, particularly when it comes to construct quality and usefulness of the example sentence. In both datasets, the clarity and granularity of codes were generally high, but the example sentences were almost always weaker, especially in the Finnish dataset, despite Finnish's much higher prevalence in the OpenAI data set – occurring with over 100x the frequency of Azerbaijani. This suggests that although GPT-4o can effectively identify relevant themes and constructs, its ability to illustrate these with specific examples may be limited by language-specific nuances.

An important step in our qualitative research process was the construct evaluations, which provided valuable insights into the clarity, specificity, and overall utility of the codes generated by GPT. Conducting construct evaluation, though still rare, is crucial in ensuring that the constructs developed through AI-assisted processes are not only theoretically sound but also practically useful for human coders. These evaluations allow researchers to assess the interpretability and relevance of each code, identify any ambiguities or misalignments, and refine the codebook to better suit the needs of qualitative analysis. As [32] highlighted, challenges in cross-language qualitative research often arise from a lack of reliable evaluation processes, making it vital to include code evaluation as part of the research methodology. Without this evaluative step, there is a risk that the generated codes may be overly abstract or not fully aligned with the dataset's objectives, particularly in multilingual settings where language nuances can affect meaning. This process is especially important in multilingual qualitative analysis, as effective communication between researchers and linguistic experts is often difficult when the research team comes from different cultural or academic backgrounds [15].

Despite the differences in code generation between translated and untranslated datasets, there was significant thematic overlap for both languages. Key themes such as *Teacher Experiences & Perceptions, Parental Involvement*, and *Ethical Concerns* were identified across codebooks. This thematic consistency supports claims from previous literature suggesting that translation does not necessarily compromise the core thematic content of multilingual data [29]. However, the increased variety of codes in the translated datasets also highlights the need for a rigorous social moderation process to refine and consolidate codes. Without this process, the risk of redundancy or thematic overlap is higher, particularly in multilingual research where translation can introduce additional complexity. Therefore, human oversight remains critical in refining and validating the GPT-generated outputs (as also discussed in [3]).

The findings of this study have practical implications for researchers, educators, and organizations involved in multilingual qualitative analysis. One of the most valuable contributions of ChatGPT-4o in this context is its ability to reduce the manual workload associated with generating qualitative codebooks across multiple languages. Traditionally, qualitative analysis, especially for multilingual datasets, has been time-consuming and resource-intensive, requiring human coders proficient in each language [8]. By leveraging LLMs like GPT-4o, researchers can automate initial stages of coding and thematic analysis, enabling faster and more efficient codebook development processes. This is particularly useful in large-scale international projects or in fields where multilingual datasets are common, such as education, healthcare, international development, and public policy. In educational research, for instance, GPT-4o's ability to assist in analyzing data from different linguistic backgrounds can help educators better understand how learning analytics tools are perceived and used by teachers and students across different cultures. This can lead to more contextually appropriate educational interventions tailored to the specific needs and cultural contexts of different regions.

One of the key limitations of this study is the relatively small and context-specific datasets, which comprised transcripts from learning analytics focus groups in Azerbaijani and Finnish. Although these languages are underrepresented in natural language processing research, the findings may not be generalizable to other languages, particularly those with different grammatical structures or cultural nuances. Additionally, the study was limited to the educational domain; GPT-4o's performance may differ when applied to datasets from other fields such as healthcare, social sciences, or legal studies. Another notable limitation is the model's struggle with generating high-quality example sentences, particularly in the native Finnish dataset, which suggests that GPT-4o may face challenges in capturing context-specific meaning and providing concrete illustrations in less-resourced languages. Finally, the study does not explore how GPT-4o com- pares with other language models or AI systems, leaving room for future research to investigate whether different models might yield more accurate or contextually appropriate results in multilingual qualitative analysis.

Future research should focus on expanding the dataset scope to explore how GPT-4o and similar LLMs handle a wider variety of languages, especially those that are structurally different from the ones studied here. For example, research could investigate how these models perform with languages that have very different grammatical structures or the thousands of languages much less present in OpenAI's training data than Azerbaijani. Such research could provide valuable insights into the generalizability of LLMs and identify areas where further training or fine-tuning might be necessary. Moreover, future research should further examine the ethical dimensions of using LLMs in qualitative research, particularly regarding potential biases in code generation and the impact of these biases on the validity of results in underrepresented languages.

Acknowledgments. The work was supported by the Swiss Cyber Institute and the Academy of Finland (grant no. 356314). We gratefully acknowledge Makreta Pyhäjärvi and Ville Heilala for their support as Finnish native speakers, including independent review of the Finnish codes and feedback that informed the final codebook.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Maria Bannert, Inge Molenaar, Roger Azevedo, Sanna Järvelä, and Dragan Gašević. 2017. Relevance of learning analytics to measure and support students' learning in adaptive educational technologies. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 568–569.

[2] Melissa Baralt. 2011. Coding qualitative data. Research Methods in Second Language Acquisition: A Practical Guide, 222–244.

[3] Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Felipe Zambrano, Alexandra L. Andres, Dara Bright, Mamta Shah, et al. 2024. ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In Proceedings of the International Conference on Artificial Intelligence in Education, 134–149.

[4] Andrea J. Bingham and Patricia Witkowsky. 2021. Deductive and inductive approaches to qualitative data analysis. Analyzing and Interpreting Qualitative Data: After the Interview, 133–146.

[5] Laila Burla, Birte Knierim, Jürgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel. 2008. From text to codings: Intercoder reliability assessment in qualitative content analysis. Nursing Research 57(2), 113–117.

[6] Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. arXiv preprint arXiv:2406.15625.

[7] Lucy G. Crane, Melanie B. Lombard, and Eric M. Tenz. 2009. More than just translation: challenges and opportunities in intercultural and multilingual research. Social Geography Discussions 5, 1, 51–70.

[8] Kevin Crowston, Eileen E. Allen, and Robert Heckman. 2012. Using natural language processing technology for qualitative data analysis. International Journal of Social Research Methodology 15(6), 523–543.

[9] Stefano De Paoli. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. Social Science Computer Review 42(4), 997–1019.

[10] Virginia Eatough, Jonathan A. Smith, and Rachel Shaw. 2008. Women, anger, and aggression: An interpretative phenomenological analysis. Journal of Interpersonal Violence 23(12), 1767–1799.

[11] Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving LLM-based machine translation with systematic self-correction. arXiv preprint arXiv:2402.16379.

[12] 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 CHI Conference on Human Factors in Computing Systems, 1–29.

[13] Maria Haak, Ines Himmelsbach, Marianne Granbom, and Charlotte Löfqvist. 2013. Cross-national and multi-language qualitative research: Challenges and recommendations. British Journal of Occupational Therapy 76(7), 333–336.

[14] Esther Helmich, Sayra Cristancho, Laura Diachun, and Lorelei Lingard. 2017. "How would you call this in English?" Being reflective about translations in international, cross-cultural qualitative research. Perspectives on Medical Education 6, 127–132.

[15] Monique M. Hennink. 2008. Language and communication in cross-cultural qualitative research. In Doing Cross-Cultural Research: Ethical and Methodological Perspectives, 21–33.

[16] Leslie Rupert Herrenkohl and Lindsay Cornelius. 2013. Investigating elementary students' scientific and historical argumentation. Journal of the Learning Sciences 22(3), 413–461.

[17] Prue Holmes, Richard Fay, Jane Andrews, and Mariam Attia. 2015. How to research multilingually: Possibilities and complexities. Research Methods in Intercultural Communication: A Practical Guide, 88–102.

[18] iTEC Scenarios, Design and Prototyping. n.d. Participatory design workshop guidelines. Retrieved September 18, 2024, from http://itec.aalto.fi/participatory-design/workshop-guidelines/

[19] Aodong Li, Yunhan Zhao, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. 2024. Anomaly detection of tabular data using LLMs. arXiv preprint arXiv:2406.16308.

[20] Xiner Liu, Andres Zambrano, Amanda Barany, Jaclyn Ocumpaugh, Jeff Ginger, Matt Gadbury, H. Chad Lane, and Ryan S. Baker. In press. Investigating learner interest and observation patterns in a Minecraft virtual astronomy environment. In Proceedings of the International Conference on Quantitative Ethnography.

[21] Xiner Liu, Jiayi Zhang, Amanda Barany, Maciej Pankiewicz, and Ryan S. Baker. In press. Assessing the potential and limits of large language models in qualitative coding. In Proceedings of the International Conference on Quantitative Ethnography.

[22] Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and M.F. Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports 14(1), 9603.

[23] Nurfaradilla Mohamad Nasri, Nurfarahin Nasri, and Mohamad Asyraf Abd Talib. 2021. Cross-language qualitative research studies dilemmas: A research review. Qualitative Research Journal 21(1), 15–28.

[24] OpenAI. 2023. GPT-3: Language models are few-shot learners: Supplemental materials. Retrieved September 21, 2024, from https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv

[25] Judith Oxley, Evra Günhan, Monica Kaniamattam, and Jack Damico. 2017. Multilingual issues in qualitative research. Clinical Linguistics & Phonetics 31(7–9), 612–630.

[26] Rajvardhan Patil and Venkat Gudivada. 2024. A review of current trends, techniques, and challenges in large language models (LLMs). Applied Sciences 14(5), 2074.

[27] Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Xiaofeng Wang, Anh Nguyen Duc, and Pekka Abrahamsson. 2024. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. arXiv preprint arXiv:2402.01386.

[28] Jay Ratican and James Hutson. 2024. Advancing sentiment analysis through emotionally-agnostic text mining in large language models (LLMs). Journal of Biosensors and Bioelectronics Research.

[29] Ueli Reber. 2019. Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. Communication Methods and Measures 13(2), 102–125.

[30] Kate Roberts, Anthony Dowell, and Jing-Bao Nie. 2019. Attempting rigour and replicability in thematic analysis of qualitative research data: A case study of codebook development. BMC Medical Research Methodology 19(1), 1–8.

[31] David Williamson Shaffer and Andrew R. Ruis. 2021. How we code. In Proceedings of the Second International Conference on Quantitative Ethnography, 62–77.

[32] Allison Squires. 2009. Methodological challenges in cross-language qualitative research: A research review. International Journal of Nursing Studies 46(2), 277–287.

[33] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. arXiv preprint arXiv:2006.06202.

[34] Bogusia Temple and Alys Young. 2004. Qualitative research and translation dilemmas. Qualitative Research 4(2), 161–178.

[35] Fenna Van Nes, Tineke Abma, Hans Jonsson, and Dorly Deeg. 2010. Language differences in qualitative research: Is meaning lost in translation? European Journal of Ageing 7, 313–316.

[36] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 363–372.

[37] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. arXiv preprint arXiv:2404.00929.

[38] Ahtisham Younas, Angela Cuoco, Ercole Vellone, Sergi Fàbregues, and E. Barrios. 2022. Contextual coding in qualitative research involving participants with diverse sociocultural backgrounds. The Qualitative Report 27(11), 2509–2527.

[39] Andres Felipe Zambrano, Xiner Liu, Amanda Barany, Ryan S. Baker, Juhan Kim, and Nidhi Nasiar. 2023. From nCoder to ChatGPT: From automated coding to refining human coding. In Proceedings of the International Conference on Quantitative Ethnography, 470–485. Springer Nature Switzerland, Cham.

[40] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M. Carroll. 2023. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv preprint arXiv:2309.10771.