Automating Qualitative Data Analysis with Large Language Models

Angelina Parfenova
Lucerne University of Applied Sciences and Arts
Technical University of Munich
angelina.parfenova@hslu.ch

Alexander Denzler
Lucerne University of Applied Sciences and Arts
alexander.denzler@hslu.ch

Juergen Pfeffer
Technical University of Munich
juergen.pfeffer@tum.de

Abstract

This PhD proposal aims to investigate ways of automating qualitative data analysis, specifically the thematic coding of texts. Despite existing methods vastly covered in literature, they mainly use Topic Modeling and other quantitative approaches which are far from resembling a human's analysis outcome. This proposal examines the limitations of current research in the field. It proposes a novel methodology based on Large Language Models to tackle automated coding and make it as close as possible to the results of human researchers. This paper covers studies already done in this field and their limitations, existing software, the problem of duplicating the researcher bias, and the proposed methodology.

1 Introduction

Qualitative research is an important asset in various fields such as marketing, media studies, social science, psychology, and medical research (Avjyan, 2005; Brennen, 2021; Mohajan et al., 2018; Leeson et al., 2019). It stands out from quantitative methods in its ability to go deeper into research questions and capture individual experiences. However, it doesn't have the straightforward statistics or clear answers often found in quantitative research. This makes it harder to draw conclusions and prove hypotheses when dealing with a vast collection of unstructured text documents (Bumbuc, 2016).

The primary way to analyze data in qualitative research involves open coding, a process that requires meticulously reading through texts to pinpoint significant thoughts, ideas, attitudes, and topics (Glaser and Strauss, 2017). Following this, axial coding helps identify how these codes interrelate and groups them into broader categories (Saldana, 2016). This method is time-consuming, often stretching over weeks (Alshenqeeti, 2014), as it demands intensive manual effort and professional expertise to analyze a large number of documents. Given these challenges, there's a growing interest in automating or simplifying the text analysis process to make it less labor-intensive.

While there has been some progress in automating the analysis of interview data, using techniques like Topic Modeling (Parfenova, 2024; Leeson et al., 2019) and Wordnet hierarchies (Guetterman et al., 2018), these approaches mainly highlight keywords already present in the text. They don't generate the nuanced "ideas" and "thoughts" that come to mind upon reading it. Therefore the main goal of this research is to automate the coding procedure of qualitative data (mainly interviews) to make the result of analysis as close as possible to human researchers' results.

In this proposal, we explore existing approaches for analyzing interview data and suggest a new method for automating the full coding procedure. The aim is to develop a model that can analyze interviews minimizing the variability and biases that can be introduced by human researchers. Future work will involve producing software that can assist organizations and researchers in managing and interpreting large volumes of textual data efficiently.

2 Related Work

Before covering existing approaches and software dedicated to qualitative analysis, we need to explore how the coding is done by professional human coders.

2.1 Current coding practice

Each statement or significant segment of dialogue within an interview is assigned a "code" that summarizes its main idea. Depending on the researcher it can be represented as a word or even a phrase, the main goal is to encapsulate the key message of the citation (Miles and Huberman, 1994). Once coded, these segments are then organized into broader cat-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 83–91

egories that reflect the underlying patterns and relationships within the dataset. Categories are higher-order classifications that codes are grouped into. These categories emerge from the data and help in developing a theory that is grounded in the data itself (Glaser and Strauss, 2017). In practice, if codes are allowed to be either a word or a phrase, categories are mostly one or two words (collocation).

Describing the process in simpler terms, first, we summarize the main idea of each citation in the interview. Then, we start grouping them into bigger categories. This means looking at all the little ideas we've found and seeing how they fit together into larger themes. We ask questions like, "Do these codes share something in common?" or "Are they talking about the same bigger idea?" This helps us organize our findings better (Parfenova, 2024). We visualize it in the Figure 1.

These categories and codes themselves are then organized into a concept map, similar to a mind map, the example of which is portrayed in Figure 2 (note: it is only the part of the graph based on citations we wrote above). This graph helps in visualizing the whole narrative of the interviews conducted.

2.2 Coder qualification and expertise

The coders responsible for this task are typically trained researchers or analysts with a background in qualitative methods. They possess an understanding of the research aims and are skilled in identifying the nuanced meanings within the text. It is their expertise that allows them to discern the subtleties in dialogue and assign appropriate codes that reflect the core message of the segment (Miles and Huberman, 1994).

Inter-coder reliability is essential to guarantee the credibility of qualitative data coding—it creates consensus among various coders in their application of codes. Usually, it involves pilot sessions where several researchers initially code a subset of data, and then an agreement on codes is achieved through discussion and comparison of coded segments. Thus, researchers try to avoid individual bias by voting system, basically agreeing which code is better for this particular segment. The degree of coder agreement is quantified through statistical measures like Cohen's Kappa or Krippendorff's Alpha (Krippendorff, 2018).

Although there are extensive descriptions of the methods used in analyzing texts, it is crucial to review prior studies that focused on qualitative data analysis using computational methods. Several papers have addressed this topic; let us provide a brief overview of these works.

2.3 Existing approaches

The first approach covered vastly in literature is topic modeling and word-to-vector conversion followed by a comparison of this NLP technique with an open coding procedure. If the revealed topic/code was similar in meaning to one revealed by the researcher, it was considered to be extracted properly (Leeson et al., 2019). In this research Topic modeling, specifically LDA, was conducted for each question and resulted in ten keywords with weights that represented the highlighted topics. This technique was good in covering topics discovered in the transcripts, however, the keywords extracted were not close enough semantically to the results of expert coders.

Another recent approach was to create a Topic Modeling alike model that combines BERT embeddings with HDBSCAN clustering to create clusters of keywords and then visualize them in the form of a graph as social scientists do with a concept map (Parfenova, 2024). The example of keywords extracted from the same set of interviews used as examples above is illustrated in Figure 3. The advantage of this method is drawing the concept map that consists of keywords and links between them based on co-occurrence in the topic, however, it doesn't generate ideas/thoughts based on the context of a citation but extracts words that already exist in a text. That way it is a completely different procedure rather than 'coding'.

Other works (Guetterman et al., 2018; Wei et al., 2015) have used WordNet to find the closeness between words and compose their semantic hierarchy. For example, "based on edge distance between appropriate synsets in this tree-like structure, one could consider that exercise and workout are very similar (an edge distance of 0), exercise and yoga are quite similar (an edge distance of 1), whereas exercise and straining are even less similar (an edge distance of 2)". Other similarity metrics were also used, such as Leacock and Chodorow similarity (Leacock, 1998) and Wu-Palmer similarity (Wu and Palmer, 1994). This method was also successful in the identification of codes. However, a significant limitation of this approach is its reliance on WordNet, which is not actively maintained, offers limited lexical coverage, and does not scale

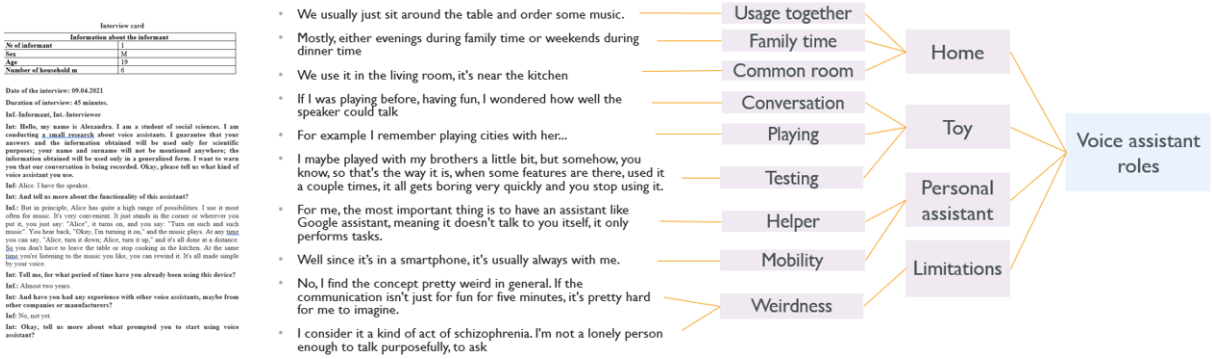

Figure 1: Coding process (Davydova, 2024). The text (in this case interview transcript) is being split into paragraphs. The main idea/thought of a paragraph is extracted and becomes a "code" (open coding). Then, this list of codes is grouped into higher-level topics (axial coding).

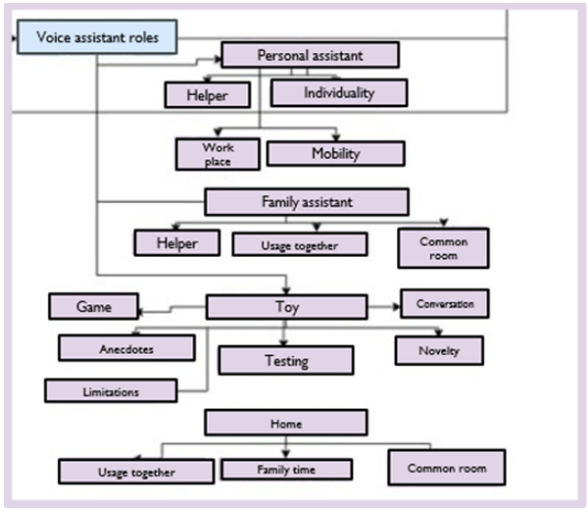

Figure 2: Example of graph (Davydova, 2024)

without considerable human effort for updates. Furthermore, this method has been primarily applied to structured interviews with specific, directed questions, which are not typically the interviews that present the most analytical challenge or demand the most time.

2.3.1 Approaches Based on Large Language Models (LLMs)

With the emergence of Large Language Models (LLMs) such as GPT variants, there has been a paradigm shift in how we approach text analysis and labeling. The ability of LLMs to understand and generate human-like text has opened new avenues in various fields, including computational social science (CSS) and content moderation. This section explores using LLMs in the context of document annotation, relation extraction, and concept linking, providing insight into the challenges presented by current research.

Labeling with LLM In the field of computational social science (CSS), the annotation of documents is a foundational step in analyzing social phenomena. Traditionally, this process has been both time-consuming and labor-intensive, often requiring manual labeling of large corpora. LLMs have made this task easier by enabling researchers to annotate documents at scale. However, despite the efficiency of LLMs, their annotations are not without flaws and often exhibit biases and imperfections. To overcome these issues, a novel algorithm has been introduced, emphasizing design-based supervised learning (DSL) (Egami et al., 2024). The DSL estimator combines imperfect LLM-generated labels with a limited set of high-quality, gold-standard labels, which are created by experts in social science thoroughly annotating a representative sample of documents. The DSL algorithm then combines these accurate labels with the larger set of imperfect LLMs. It does this by adjusting the LLM labels based on discrepancies with the gold standard, resulting in improved 'pseudo-outcomes'. These are then used in statistical analyses, ensuring results that are both robust, due to the expert input, and scalable, thanks to the automation provided by LLMs.

Another study shifts the focus into trying to experiment with variations in prompts and batch sizes to improve the quality of hate-speech labeling (Matter et al., 2024). Utilizing manual annotation as a benchmark, GPT-3.5 and GPT-4 models were fine-tuned to detect nuances in violent speech. The results showed that the best GPT-4 model achieved Cohen's Kappa scores of 0.54 and 0.62 when compared to two human coders, respectively, indicating a moderate to substantial agreement. Weighted and macro F1 scores further supported the model's re-

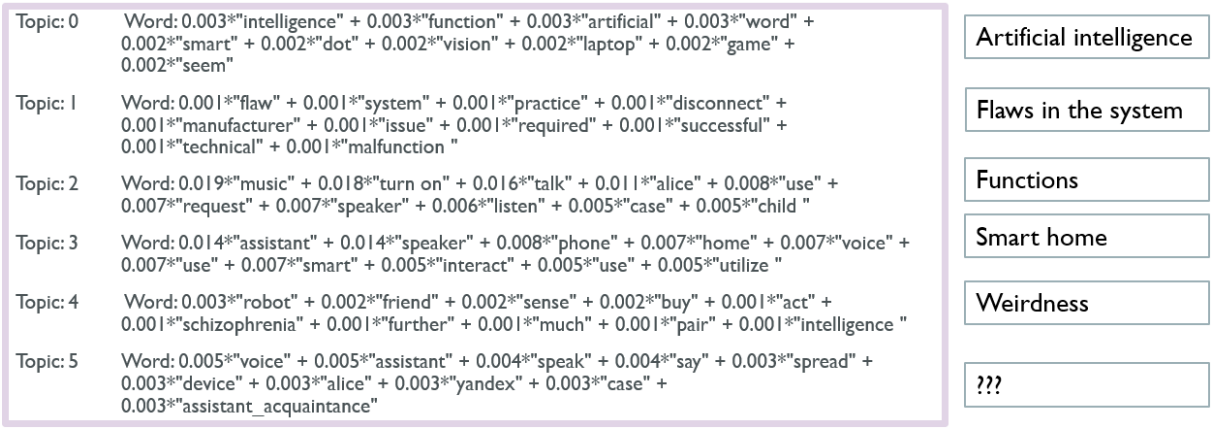

Figure 3: Keywords extracted using Topic Modeling and their possible interpretation

liability. These findings suggest that while LLMs like GPT-4 show promise in automating content moderation, their annotations when benchmarked against manual coding, reveal some discrepancies that require further adjustment.

2.3.2 Deductive and Inductive Coding in Qualitative Research

Coding in qualitative research can be categorized into deductive and inductive methods. Deductive coding uses a pre-established codebook applied to the data, while inductive coding generates codes directly from the data itself.

Some studies, such as (Xiao et al., 2023) and (Spinoso-Di Piano et al., 2023), have explored automatic code generation using NLP methods, primarily focusing on deductive coding where labels are predefined. In contrast, our approach employs an inductive coding process based on "grounded theory" (Glaser and Strauss, 2017), allowing knowledge to emerge from the data. Preliminary knowledge is utilized only during categorization, with initial coding being entirely inductive.

Our proposed method is ideal for inductive coding, aiming to identify patterns and themes organically. Acknowledging the computational methods supporting deductive coding, future research could investigate hybrid methods that combine both inductive and deductive elements, combining the strengths of each approach.

2.4 Relation extraction and concept linking

In the domain of natural language processing, the task of relation extraction and concept linking is crucial for transforming unstructured text into a structured form that highlights the relationships between entities. In social science, the practice of labeling these relationships within concept maps varies; some researchers annotate the connections between codes explicitly, while others do not, due to the lack of a standardized approach. The pros of labeling are that it can clarify the nature of relationships and facilitate a deeper understanding of complex interactions within the data. On the other hand, the cons include the potential for subjectivity and the added layer of analysis that could complicate the interpretation of data.

Currently, the detection of relationships often relies on Large Language Models (LLMs) (Loureiro et al., 2023; Trajanoska et al., 2023; Bratanic, 2022; Yao et al., 2023; Pan et al., 2023), which, despite their growing sophistication, still face challenges such as the accurate identification of relations in documents covering diverse topics (Friedman et al., 2022; Feder et al., 2022). The lack of universally accepted link types further complicates the task, as this can lead to ambiguity and inconsistent findings across different studies (Picco et al., 2023; Cabot and Navigli, 2021). Additionally, extracting an exhaustive list of entities can create overly complex networks that make analysis even more complicated rather than reveal significant patterns.

Given these considerations, it is worth discussing whether labeling relationships in knowledge extraction is necessary or if an unlabeled graph is enough for sociological research. The answer may not be absolute; the decision to label relationships should be guided by the specific research objectives and the nature of the data being analyzed. Further research is required to develop more standardized methods for relation extraction that could benefit the social sciences and other disciplines.

2.4.1 Existing softwares

Qualitative researchers often turn to specialized software like Atlas.ti[1], Dedoose[2], and MAXQDA[3] for manual coding, which facilitates text analysis by allowing for efficient tagging and categorization within a user-friendly environment. However, these tools do not fundamentally alter the nature of the analysis process but rather provide a digital convenience for traditional workflows.

The challenge thus remains to create an innovative, accessible tool that not only simplifies the coding process but also enhances the analytical capabilities of researchers, enabling them to extract deeper insights from qualitative data without the need for advanced programming skills.

2.4.2 Discussion

Integrating AI into qualitative data analysis presents a promising approach to overcoming the limitations of human coding, such as subjective bias (Bumbuc, 2016), agreement challenges between individual coders (Krippendorff, 2018), and the time-consuming nature of manual coding (Saldana, 2016). Human coders, while better understanding the nuances of sentences they code, often struggle with consistency and objectivity, leading to variability in data interpretation (Saldana, 2016). AI, with its capacity for rapid data processing and application of standardized coding rules, offers a solution to these challenges by ensuring a more uniform and efficient analysis (Bengio et al., 2013), allowing to deal with larger amounts of data. Thus, the development of a model capable of imitating the human coding process, at the same time overcoming the abovementioned challenges, could serve not only as a supporting solution for human coders but as a standard itself.

3 Research Proposal

This PhD proposal seeks to explore the ways of automating the analysis of quantitative data, mainly interview transcripts. The main goal is to test the proposed approach on different sets of interviews and compare it with expert coding. Next, we outline the main research questions:

1. How do social scientists code interviews and ensure consistency of coding while collaborating? What preliminary knowledge are they using while deriving categories from codes?
2. How can we identify codes and group them into meaningful categories using LLMs?
3. If the researcher is biased while analyzing interviews, is there a way to replicate this bias to make the model work like a real human researcher?

It is important to note that the concept map described in previous sections is the last step after the ones mentioned above. As its nature lies in relationship extraction and information visualization, it is considered to be the next separate research topic. Thus, it will not be covered in the proposed methodology for this proposal, though it was important to describe it as the concluding part of qualitative data analysis.

The accomplishment of these research goals will help to systematically organize data, which can be massive and complex, into understandable and manageable themes. This enables researchers to identify patterns and insights that are not immediately apparent, thus adding depth to the research findings.

Exploring the domain knowledge of social scientists can significantly improve the accuracy of automated models. This knowledge can lead to the creation of algorithms that are more aligned with human cognition, which is especially important when analyzing nuanced human communication.

While bias is typically something to be minimized, understanding it can be useful, particularly in developing AI that can replicate human-like understanding. It's important to recognize that complete objectivity is unattainable, and acknowledging bias allows for a more reflexive approach to data analysis.

4 Proposed Approach

The proposed methodology follows the cognitive process of a social scientist who typically keeps in mind the study's framework during thematic analysis. As described in section 2.1 on Current coding practice, thematic analysis is a two-step process consisting of open and axial coding. The first step involves summarizing each citation's main idea/thought, and the second consists of categorizing all ideas into higher-order categories (Fig.1). According to existing manuals, the open coding phase doesn't involve preliminary knowledge of

[1] https://atlasti.com/ Accessed: 12.12.2023
[2] https://www.dedoose.com/ Accessed: 12.12.2023
[3] https://www.maxqda.com/ Accessed: 12.12.2023

the research topic (Glaser and Strauss, 2017), while axial coding heavily relies on it (Miles and Huberman, 1994). The next sections describe how each stage of the coding process can be automated.

4.1 Open Coding: summarization

The methodology's initial phase consists of summarizing a sentence's main idea, however with social scientist professional bias. That's why the first stage of automating the process involves finetuning several LLMs on the dataset with professionally extracted codes by human experts. Models will be finetuned using PEFT (Parameter Efficient Finetuning) Low-Rank Adaptation (LoRA) (Hu et al., 2021) to reduce the number of parameters that need to be tuned to around 1%. The fine-tuning quality is subsequently evaluated using the BERT score (Zhang et al., 2019), and ROUGE score (Lin, 2004).

LLMs are tested with a variety of open coding prompts that range from explicit instructions to more nuanced requests that mimic the considerations of a social scientist:

- Explicit Instruction: Summarize the main idea/thought of a sentence.
- Informal Request: Can you tell me what the main idea of this sentence is in just a few words?
- Expert Angle: From the perspective of a social scientist, summarize the following sentence as you would in thematic coding.
- Impersonalization: If you were a social scientist, what code would you give to this citation?
- Detailed Explanation: Explain in a couple of words the primary thought expressed in the following text.
- Simplified Task: What's the gist of this sentence?

LoRA will be subsequently compared to prompt-engineering giving several examples of coded sentences and asked to code the same way the rest of the dataset. One of our hypotheses is that PEFT will result in higher BERT scores than prompting because some codes from social scientists are highly domain-specific (e.g. citation from the training data: " If a woman comes in, you can see from her that she doesn't drink alcohol and she doesn't have bad habits, both when interviewed and on further follow-up, and the pregnancy is going well, there may be complications, but some minor ones.", code: Habitus)

4.2 Axial Coding: Categorization

Following the open coding phase, the LLM categorizes the generated codes into thematic groups. This process is driven by a set of contextual prompts derived from the research itself, designed to generate meaningful categories by the model. Examples of such contexts involve mentioning the goal of the research, hypotheses, interview guide, theoretical framework, etc. Everything that might help with giving categorization more context.

The LLM processes the input codes $\mathcal{C}$ and the research context $\mathcal{Q}$ to produce a set of thematic categories $\mathcal{K}$. The evaluation includes assessing how well the codes from diverse prompts such as "Do these codes share something in common?" and "Group these codes into meaningful categories." converge into coherent groups that reflect the underlying themes of the dataset. The number of categories extracted is not predefined and can vary as well as it varies among human researchers.

If we take a look at Fig.1 we see several extracted categories from the set of open codes. However, the choice of categories usually depends on the individual researcher and might vary. That's why there is no certain way to internally evaluate the quality of categorization. At this stage, it will be necessary to perform an expert evaluation which is described in detail in the Evaluation section.

5 Dataset

For the fine-tuning of the Large Language Model (LLM), we propose utilizing a curated dataset comprising coded interview citations collected from social scientists, both academic researchers and students doing qualitative research in social science. An illustrative example, as demonstrated in Figure 4, presents data in a citation-label format. For instance, a citation s_i such as "Well since it's a smartphone, it's usually always with me," would be associated with a label l_i denoted as "mobility".

A potential challenge is the limited size of the dataset. However, recent studies, such as those by (Zhou et al., 2023), have shown that LLMs can still perform exceptionally well even when fine-tuned with a minimal dataset.

Code	Citation
Mobility	Well since it's in a smartphone, it's usually always with me.
Helper	For me, the most important thing is to have an assistant like Google assistant, meaning it doesn't talk to you itself, it only performs tasks.
Playing	For example I remember playing cities with her...
...	...

Figure 4: Dataset

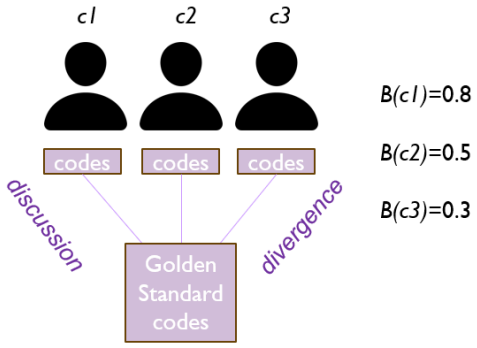

Figure 5: Coders consensus and individual biases

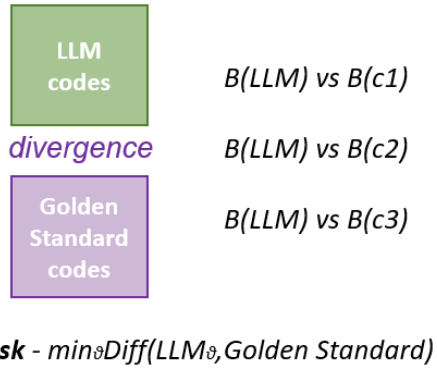

Figure 6: Comparing individual biases and LLM

6 Evaluation

The testing of the model will be based on comparison with experts. We will employ Krippendorff's Alpha to evaluate the reliability and consistency of the coding provided by our model compared to expert coders. This statistical measure is ideal for assessing the agreement among multiple coders on qualitative data.

Our approach involves constructing a contingency table where each row represents an interview citation and each column a coder (our AI model and human experts). The codes assigned to each citation by different coders are filled in this table. We will then calculate the observed disagreement among coders for each item and sum these to get the total observed disagreement. Expected disagreement, which is the disagreement expected by chance, will also be computed based on the distribution of each code.

The Alpha value is calculated using the formula $\alpha = 1 - \frac{ObservedDisagreement}{ExpectedDisagreement}$. A higher Alpha value (close to 1) indicates a higher agreement among the coders, suggesting that our model's coding aligns well with human experts. This method will provide a robust quantitative measure of the coding reliability of our AI model in qualitative data analysis.

We extend our testing framework by evaluating potential bias in the coding process. This involves comparing the coding consistency of our language model and a Golden Standard established by expert consensus.

The example of the evaluation framework is illustrated in Figures 5-6. The Golden Standard codes emerge from a discussion process among human coders ($c1$, $c2$, $c3$) to identify the most appropriate codes. To effectively incorporate a bias evaluation, we propose to compute bias scores for each human coder, $B(c1)$, $B(c2)$, $B(c3)$, to quantify their deviation from the Golden Standard. In parallel, we will calculate the bias score for the language model ($B(LLM)$), reflecting its divergence from the $GoldenStandard$. Our objective is articulated through the task of minimizing the difference between the model's bias and the Golden Standard, formalized as $minDiff(LLM_3, GoldenStandard)$, thereby striving to align the model's output with the unbiased consensus code.

7 Limitations

In this section, we will discuss the limitations of our current research as well as the challenges posed. One of the primary concerns is the issues associated with concept extraction and ontology building. Accurately selecting codes is inherently challenging, as the process relies on nuanced human knowledge. It raises a question: how can we develop a model sophisticated enough to replicate these complex human cognitive tasks?

Another issue is that language is intricate. People have their unique way of speaking, and they often communicate more than what they explicitly say. Our model strives to comprehend and code what people express, but it might not be able to do

so as accurately as humans. It may miss some of the implicit nuances in language that we naturally understand.

One major concern is that our dataset is relatively small. Since we are using a limited number of examples to fine-tune our model, there is a possibility that it may not perform as well as we expect. This could result in it being less effective in coding new interviews, as it hasn't had sufficient exposure during training.

Additionally, the extraction of relationships presents its own set of difficulties. Training an algorithm to navigate them poses a significant challenge, further complicated by the diverse strategies individual coders reach a consensus. The question of whether there exists a universal approach or if coders are utilizing various, possibly conflicting, techniques is yet to be answered.

Furthermore, the unclear nature of large language models (LLMs) introduces additional complexity. These models often act as "black boxes," making it challenging to discern the rationale behind their outputs. This obscurity necessitates the exploration of explainable AI, a significant area of research aimed at making AI's decision-making processes more transparent. Our project might encounter similar difficulties, interfering with our ability to fully understand and explain the model's behavior and decisions.

8 Ethics Statement

This research ensures data privacy by anonymizing all interview data and obtaining informed consent, in compliance with data protection regulations. While the goal is to enhance and assist human researchers, potential displacement effects are considered, striving to support rather than replace them. Efforts are made to mitigate biases in the LLMs, maintain fairness, and ensure transparency in the models' decision-making processes. Additionally, computational resources are optimized to minimize environmental impact.

9 Conclusion and future work

In conclusion, this proposal outlines a comprehensive framework for automating the extraction of information from qualitative research. By using the advanced capabilities of Large Language Models (LLMs) and integrating them with the expertise of social scientists, we aim to significantly reduce the time and effort required in the coding process.

In this proposal we have addressed the potential for replicating the bias inherent in human coding, recognizing that this aspect of qualitative analysis can be both a challenge and an opportunity. By understanding and potentially simulating these biases, we can approach the human-like analytical capabilities that are currently the domain of experienced researchers. The use of a curated dataset for fine-tuning the LLMs, along with the development of an algorithmic framework will be the first step in constructing an actual tool that facilitates qualitative analysis.

Future work will focus on the practical implementation of the proposed methodologies, including the fine-tuning of LLMs with the constructed dataset and the validation of the coding process against standard qualitative analysis. Additionally, we will explore the integration of multiple LLMs to simulate the collaborative nature of human coding teams. The end goal is the creation of user-friendly software that embodies the strengths of both manual and AI-assisted analysis, involving all stages of qualitative analysis from open coding to the construction of a concept map.

References

H. Alshenqeeti. 2014. Interviewing as a data collection method: A critical review. *English Linguistics Research*, 3:39–45.

E.G. Avjyan. 2005. Asynchronous on-line focus group: technology and procedures of conducting. *Southern Russian Journal of Social Sciences*, (1):116–129.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Tomaz Bratanic. 2022. From text to a knowledge graph: The information extraction pipeline. Neo4j Blog. Accessed: 2023-13-11.

Bonnie S Brennen. 2021. *Qualitative research methods for media studies*. routledge.

Ştefania Bumbuc. 2016. About subjectivity in qualitative data interpretation. In *International Conference Knowledge-Based Organization*, volume 22, pages 419–424.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Alexandra Davydova. 2024. Thesis: The integration of voice assistants in domestic interaction scenarios.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond.

Scott Friedman, Ian Magnusson, Vasanth Sarathy, and Sonja Schmer-Galunder. 2022. From unstructured text to causal knowledge graphs: A transformer-based approach.

Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

T.C. Guetterman, T. Chang, M. Dejonckheere, T. Basu, E. Scruggs, and V.G.V. Vydishwaran. 2018. Augmenting qualitative text analysis with natural language processing: Methodological study. *J. Med. Internet Res.*, 20.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Claudia Leacock. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: A Lexical Reference System and its Application*, pages 265–283.

W. Leeson, A. Resnick, D. Alexander, and J. Rovers. 2019. Natural language processing (nlp) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods*, 18.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Manuel V. Loureiro, Steven Derby, and Tri Kurniawan Wijaya. 2023. Topics as entity clusters: Entity-based topics from language models and graph neural networks.

Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. Close to human-level agreement: Tracing journeys of violent speech in incel posts with gpt-4-enhanced annotations.

Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

Haradhan Kumar Mohajan et al. 2018. Qualitative research methodology in social sciences and related subjects. *Journal of economic development, environment and people*, 7(1):23–48.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap.

Angelina Parfenova. 2024. Automating the information extraction from semi-structured interview transcripts. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM.

Gabriele Picco, Marcos Martínez Galindo, Alberto Purpura, Leopold Fuchs, Vanessa López, and Hoang Thanh Lam. 2023. Zshot: An open-source framework for zero-shot named entity recognition and relation extraction.

J. Saldana. 2016. *The Coding Manual for Qualitative Researchers*, 3rd edition. Sage, Los Angeles, CA.

Cesare Spinoso-Di Piano, Samira Rahimi, and Jackie Cheung. 2023. Qualitative code suggestion: A human-centric approach to qualitative coding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14887–14909, Singapore. Association for Computational Linguistics.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models.

Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. 2015. A semantic approach for text clustering using wordnet and lexical chains. *Expert Syst. Appl.*, 42(4):2264–2275.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *28th International Conference on Intelligent User Interfaces*, IUI '23. ACM.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2023. Exploring large language models for knowledge graph completion.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.