FORUM: QUALITATIVE SOCIAL RESEARCH SOZIALFORSCHUNG

Volume 25, No. 3, Art. 5
September 2024

Further Explorations on the Use of Large Language Models for Thematic Analysis. Open-Ended Prompts, Better Terminologies and Thematic Maps

Stefano De Paoli

Key words: thematic analysis; large language models; semi-structured interviews; initial coding; thematic maps

Abstract: In this manuscript I build upon an initial body of research developing procedures for leveraging large language models (LLMs) in qualitative data analysis, by carrying out thematic analysis (TA) with LLMs. TA is used to identify patterns by means of initial labelling of qualitative data followed by the organisation of the labels/codes by themes.

First, I propose a new set of LLM prompts for initial coding and generation of themes. These new prompts are different from the typical prompts deployed for such analysis in that they are entirely open-ended and rely on TA language. Second, I investigate the process of removing duplicate initial codes through a comparative analysis of the codes of each interview against a cumulative codebook. Third, I explore the construction of thematic maps from the themes elicited by the LLM. Fourth, I evaluate the themes produced by the LLM against the themes produced manually by humans. For conducting this research, I employed a commercial LLM via an application program interface (API). Two datasets of open access semi-structured interviews were analysed to demonstrate the methodological possibilities of this approach. I conclude with practical reflections on performing TA with LLM, enhancing our knowledge of the field.

Table of Contents

Forum Qualitative Sozialforschung / Forum: Qualitative Social Research (ISSN 1438-5627)

1. Introduction

Recent advances in generative artificial intelligence (AI) and large language models (LLMs) allow social scientists to explore these models' potential for qualitative inquiry. One example of application of LLMs to qualitative research is the automated transcriptions of qualitative interviews (WOLLIN-GIERING, HOFFMANN, HÖFTING & VENTZKE, 2024). LLMs have risen to prominence recently with the world-wide diffusion of ChatGPT, an AI driven Chatbot for conversation with people, based on LLMs[1]. LLMs are AI systems that can be used by humans to manipulate language, and they excel at routine tasks such as summarisation or classification of text. This is possible because LLMs are AI systems trained on large corpuses of textual data. For example, the LLM Llama-3-70b from Meta was trained on 70 billion textual parameters. A comprehensive overview of LLMs was offered by SERRANO, BRUMBAUGH and SMITH (2023). [1]

There is a nascent area, where scholars are approaching thematic analysis (TA) using LLMs, following the six phases developed by BRAUN and CLARKE (2006). TA is a qualitative method of analysis where the researcher labels (codes) portions of data with relevant meaning and then organises these codes/labels into patterns (the themes). BRAUN and CLARKE stipulated that TA encompasses the following phases: 1. familiarisation with the data; 2. initial coding; 3. identification of themes; 4. revision of themes; 5. renaming and summarising of themes; and 6. write-up of the results. A review of the existing literature employing LLMs for TA is presented in Section 2. [2]

Scientists adopting LLMs for performing TA align themselves loosely with the adoption of machine learning in qualitative analysis (see e.g. HOXTELL, 2019; WALDHERR et al., 2019; WIEDEMANN, 2013), integrating automation into the process. However, LLMs constitute a potential paradigm shift compared to prior machine learning techniques, due to their significant capacity to manipulate language. In previous publications (DE PAOLI, 2023a, 2023b), I showed how Phases 2-6 of TA can be executed on semi-structured interviews with an LLM, albeit warning about the need to include appropriate methodological reflections and controls. Other authors proposed similar processes (DRÁPAL, WESTERMANN & SAVELKA, 2023) or recommended embryonal activities for TA with LLMs (LEE, VAN DER LUBBE, GOH & VALDERAS, 2023). As social scientists increasingly explore the capabilities of LLMs in qualitative analysis, earlier attempts to conduct TA may be superseded by more refined procedures and require fresh methodological thinking. [3]

In this text, I present a novel set of prompts for performing inductive TA with an LLM. Prompts are instructions that a user gives to the LLM to complete a defined task, while the output is called a response. Prompt and response together constitute the LLM context. It is not intuitively straightforward to obtain the desired response from the LLM with a prompt, and several attempts must be carried out

[1] GPT stands for generative pre-trained transformers. GPT3.5 and GPT4 are LLMs from OpenAI, and both underpin the widely known ChatGPT. However, the LLMs can also be used by users via the application program interface (API) directly.

before arriving at something satisfactory. This activity is called "prompt engineering" (e.g. CHEN, ZHANG, LANGRENÉ & ZHU, 2023, p.2). Strategies have been developed to engineer prompts, such as chain of thoughts (e.g. YU, HE, WU, DAI & CHEN, 2023), whereby the LLMs are instructed to explain the reasoning behind their response. However, it is also possible to instruct LLMs with zero-shot or few-shot prompting. Here the model is requested to approach a new problem in its entirety where the prompt either does not contain examples (zero-shot) or where the model is given a few examples of a task before producing a response on the new material (few-shot). In earlier work, I formulated a set of zero-shot prompts for TA, which I believe now can be re-evaluated and be superseded by prompts with better features. [4]

Some aspects of my prompts require initial scrutiny, with more details offered in Section 3. First, my former prompts did amount to a set of instruction requesting an LLM to produce a fixed number of codes or themes, for e.g. one interview, or the codebook. Already in previous work, I noticed this is not what human investigators would do. Analysts would see this instead as an open-ended process. I will demonstrate that open-ended prompting is possible for conducting initial coding and the generation of themes with LLMs, more closely approximating the workflow performed by human scholars. Second, some of the terms I relied upon in previous prompts did not align with the TA terminology, which may hinder a wider acceptance of the method by the social sciences community. I am convinced that by adopting the language of prompts that better aligns with the TA terminology will prove beneficial in this regard. It will be easier for researchers executing their analysis with LLMs to more adequately explain the methodology and procedures. The lack of open-ended prompts and the need for methodologically sound language in prompts are key knowledge gaps which I seek to address. Moreover, based on the results of the definition of themes, I will illustrate how it is possible to derive a thematic map, thus further enhancing the previous practice for doing a TA with LLMs. [5]

Lastly, I recently conducted work with a colleague (DE PAOLI & MATHIS, 2024) to assess how well the codes generated by an LLM can saturate the data (see SAUNDERS et al., 2018 for a definition of initial thematic saturation, or ITS). We promoted a new method for the removal of duplicated codes from the codebook, after each interview. The reduction of the duplicates is currently necessary when doing TA with an LLM, since each interview is analysed independently (one by one) with the LLM, therefore duplicate codes can recur across the analysis. To arrive at a consolidated codebook of unique codes (without repeated codes), it is necessary to identify these duplicates before creating themes in Phase 3. In this paper, I follow the reduction of duplicate codes procedure after each interview. [6]

Thus, I use an open access dataset composed of 15 semi-structured interviews related to investigating the practices of qualitative researchers. I rely on a second open access dataset for carrying out an evaluation, composed of nine interviews from the "Teaching Data Science" project at the University of California, Santa Barbara (CURTY, GREER & WHITE, 2022). [7]

The manuscript is organised as follows: In Section 2, I present a review of literature on the use of LLMs for qualitative analysis, initial coding and TA. In Section 3, I illustrate the methods and materials. In Section 4, I outline the results, including the initial coding, the definition of themes, the thematic maps and the evaluation. In Section 5, I discuss the findings and provide observations about the practical implications for TA with LLMs. [8]

2. Thematic Analysis and Initial Coding With LLMs

CLARKE and BRAUN (2017) defined TA as "a method for identifying, analyzing, and interpreting patterns of meaning ('themes') within qualitative data" (p.297). The authors also emphasised how TA is not connected to any specific theoretical position and is flexible and adaptable to context and data. The themes/patterns support social scientists "to address the research or say something about an issue" (MAGUIRE & DELAHUNT, 2017, p.3353). Themes are derived from the sorting and grouping of initial codes (what BRAUN & CLARKE [2006, p.89] called Phase 3). The process of initial coding (Phase 2) entails work for breaking down data/observations into small discrete parts (SALDAÑA, 2021). Initial codes are normally characterised by: 1. names/labels that synthetically capture the meaning of a portion of data; 2. a description of the code (a 2-3 lines narrative augmenting the meaning of the code). [9]

The flexibility of TA and its phases has led some experts to consider conducting TA with LLMs. In my prior contribution, I showed that Phases 2-6 of a TA can be emulated with a LLM, with a valid degree of quality (DE PAOLI, 2023a, 2023b). Using a LLM, I identified most of the themes that human researchers elicited in their own study of the same data. My procedure/workflow has since been tested and evaluated with multiple datasets of semi-structured interviews displaying consistency and reproducibility. DRÁPAL et al. (2023) performed TA with LLMs focusing on more fact-based data (court decisions), also following some of the phases proposed by BRAUN and CLARKE. DRÁPAL et al. used a human-led analysis as a metric for comparison. A comparative study of themes was also carried out by HAMILTON, ELLIOTT, QUICK, SMITH and CHOPLIN (2023), who argued for tools such as ChatGPT to supplement the work of human scientists. DAI, XIONG and KU (2023) presented a simplistic human LLM collaboration model using Cohen's K inter-reliability coding metric, without reporting the prompts adopted for the inquiry. Reporting prompts, however, is essential for enabling the reproducibility of results. In less developed contributions, authors also offered initial reflections on TA, but without attempting a full investigation (e.g. LEE et al., 2023), or undertook their analysis on a small dataset of user messages (SCHIAVONE, ROBERTS, DU, SAURO & LEWIS 2023). [10]

Other authors operated targeted activities on the initial coding phase (which overlaps with other qualitative methodologies). GAO et al. (2023) and XIAO, YUAN, LIAO, ABDELGHANI and OUDEYER (2023) have produced relevant investigations in this regard, indicating a significant level of agreement between LLM and human coders using the Cohen's K metric. However, the authors of these two publications adopted deductive coding, therefore using pre-defined initial codes to label the data. Other authors concentrated on discourse analysis, also performing a deductive procedure. Contributions in this direction include the paper by HUBER and CARENINI (2023) who used a pre-trained language model to operate, for example, inferences on discourse structure. CHEW, BOLLENBACHER, WENGER, SPEER and KIM (2023) also worked on a deductive approach to discourse analysis. [11]

3. Methods

The procedure I advance here requires conducting some key phases of TA with an LLM, in line with the six phases proposed by BRAUN and CLARKE (2006). In the initial coding (Phase 2) the researcher breaks down the data by assigning a code to each relevant discrete component/portion. From the list of codes, the analyst creates themes (second order categories), by comparing, ordering, and grouping initial codes (Phase 3). Further phases would require revising (Phase 4) and renaming the themes (Phase 5). Specifically, in this paper, I focus on Phases 2 and 3. [12]

The workflow I used is presented in Figure 1. For the work, I leveraged GPT3.5-Turbo-16k, an LLM with 16 thousand tokens for context. A token amounts to a short word (a group of characters). Therefore, the model can process circa 16 thousand short words, which include the input (the prompt) and the output (the response). I accessed the model via the OpenAI API programmatically through custom written python scripts, requesting the LLM to operate certain actions on data (e.g. identify initial codes on interview text) and to generate a response (e.g. a list of codes). An API is a set of protocols that a user can employ to exchange information with a software or service (see for a review on the subject OFOEDA, BOATENG & EFFAH, 2019). [13]

Through trial-and-error tests, I engineered a set of prompts instructing the LLM to perform Phase 2 and 3 of TA, considering the LLM as a black-box. I would give the LLM an input prompt, then assess the output, and then modify the prompt again until the desired output was achieved and could be reproduced. For much of the activities, I relied on the python completion function in Figure 2, which sets the model to be used (GPT3-5-Turbo-16k) and requests the model to respond to a prompt. In this function, I also kept the temperature parameter at zero. With temperature at zero the model has "limited creativity" and provides responses almost in a determinist manner.

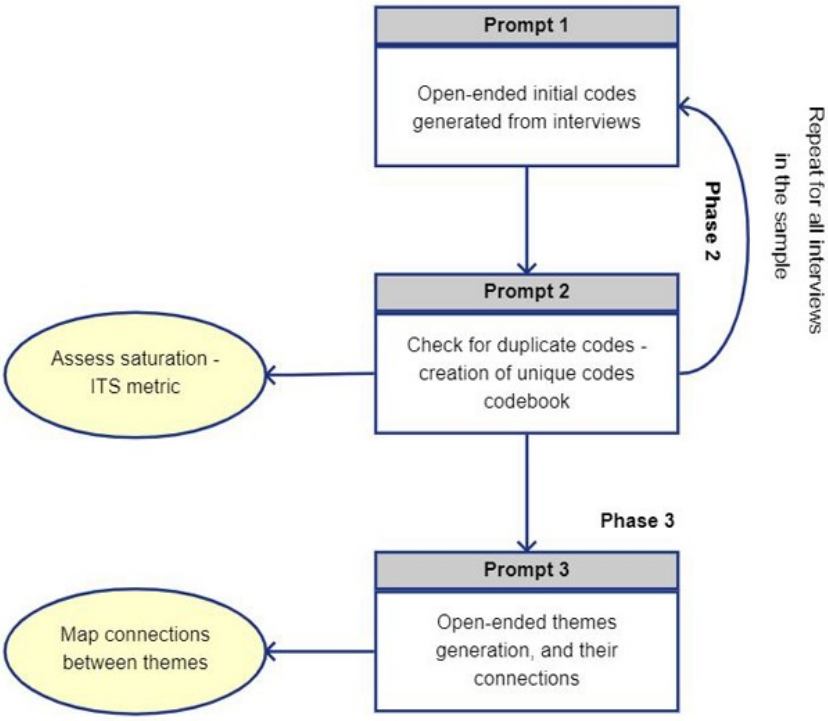
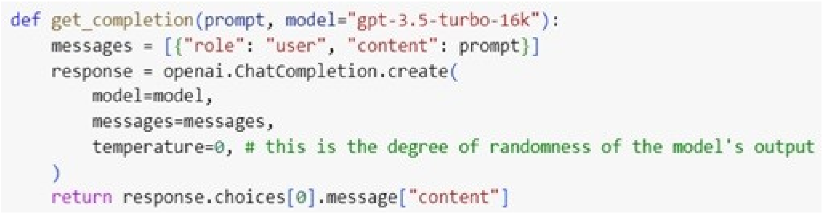

Prompt 1
Open-ended initial codes generated from interviews
Phase 2
Repeat for all interviews in the sample
Prompt 2
Check for duplicate codes - creation of unique codes codebook
Assess saturation - ITS metric
Phase 3
Prompt 3
Open-ended themes generation, and their connections
Map connections between themes

Figure 1: Workflow covering Phases 2 and 3 of BRAUN and CLARKE (2006)



```
def get_completion(prompt, model="gpt-3.5-turbo-16k"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0, # this is the degree of randomness of the model's output
    )
    return response.choices[0].message["content"]
```

Figure 2: Completion function for the LLM [14]

3.1 Phase 2 of thematic analysis

In the first step of the workflow, I address Phase 2, the initial coding. I applied Prompt_1 for this purpose, as presented in Box 1 and will comment on this prompt and compare it to the prompt I formulated in previous work. Comparison is important to appreciate the innovations represented by the new prompts.

~~~
prompt = f"""
Can you assist in the generation of a very broad range of initial codes (generate as many initial codes as needed to capture all the explicit or latent meanings or events), aiming to encompass a wide spectrum of themes and ideas present in the text below, to assist me with my thematic analysis. \
Provide a name for each code in no more than four words, an up to 40 words meaningful and dense description of the code and a quote from the respondent for each topic no longer than 80 words.\
Format the response as a json file keeping names, descriptions and quotes together in the json, and keeping them together in "Codes". \
```{text}```
"""
~~~

Box 1: Prompt_1 initial coding of an interview [15]

The main point of interest in Prompt_1 is in italics. In the text in italics, I present the language of the new prompting which should be compared with the prompt I devised in previous initial coding, shown in Box 2. In the earlier/old prompt below, I asked the model for a fixed number of codes to be generated (e.g. 15). As I noted in the introduction, this deviates from the approach likely taken by a human analyst, who does not normally have a fixed number of codes to produce. Moreover, the language I used in the previous/old prompting is ambiguous and not methodologically sound, as it does not align with TA terminologies.

> Identify the *15* most relevant themes in the text, provide a meaningful name for each theme in no more than six words, 12 words simple description of the theme, and a max 30 words quote from the participant.

Box 2: Extract from previous/old prompting for initial coding (DE PAOLI & MATHIS, 2024, p.9) [16]

Conversely, in Prompt_1 there are two important innovations. First, the instruction that I give to the LLM does not fix in advance the number of initial codes to be generated, and with the prompt, I also instruct the LLM that there is no upper limit to the codes that can be produced. The new Prompt_1 is an open-ended process for inductive coding, which resembles the procedure that would be followed by a human researcher. Second, the text of Prompt_1 aligns closely to the TA language, compared to the prompt in Box 2. In Prompt_1, I use the words "codes”, "themes" and "thematic analysis" explicitly. In previous prompts, I adopted the word "themes" in place of "codes", which seemed to correspond to the identification of initial codes. When creating themes for Phase 3, in previous prompts, I used the word "topics", which also seemed to produce valid responses from the LLM. The formulation I propose in the new Prompt_1 therefore better aligns with established TA terminology. To achieve this, I first asked the model if it recognised what TA was and how this method is performed as a sequence of phases. Having established that the LLM was familiar with the method, my strategy has been to test a variety of prompt formulations using TA terms until I obtain the desired response and to also reproduce the response. This is reflected in the prompt text (e.g. "assist me with my thematic analysis", "generate initial codes"). I suggest that the new language of Prompt_1 should facilitate the acceptance of the methodology more widely, alongside aligning the LLM-driven analysis to the established language of TA. [17]

To carry out the initial coding of a dataset of interviews, I passed the text of each interview as a variable in Prompt_1 (i.e. variable *{text}*). The prompt is contained in a python script that I used to interact with the LLM API via the completion function (Figure 2). With this function and the prompt, I was able to operate an inductive definition of the codes from the interview text. The LLM would then return a response which I captured with the python script. The response is a list of the codes for the interview under consideration. Each code is composed of: 1. a name; 2. a description; and 3. a quote retrieved from the interview. Indeed, with Prompt_1, in addition to the code/label, I requested from the LLM a description of each code of up to 40 words and to retrieve a quote from the interview. As I request in the prompt, the list of codes is formatted by the LLM as a json, a convenient format for data manipulation. In a json, the data is represented as "key: value" pairs, e.g. "quote": "*text of the quote*". I then parsed each code in the json in a python dataframe, a powerful data format. The structure of the dataframe is like a table containing the following columns: numerical index of the code (from 0 to n), code name, code description and quote. Each row of the dataframe is a code. [18]

In Section 4, I will reflect upon one additional dynamic of Prompt_1. In a further iteration, I requested the model to produce a shorter code description (15 words, rather than 40), leading to the creation of a larger number of initial codes. Therefore, whilst my open-ended prompt does not fix in advance the number of codes, it may be the case that the model considers the limits of the context (16k tokens in this case, which includes prompt+response) for the generation of codes. While the initial coding is based on requesting a shorter description, the LLM has therefore more tokens for producing a larger number of codes. [19]

The second prompt I engineered for Phase 2 (Prompt_2, in Box 3) delivers the progressive reduction of the codebook. Since my method is based on using the LLM to analyse each interview separately, the initial codes are elicited independently from interview to interview. The result is some duplication of initial codes across interviews, which would not normally happen when the coding is carried out by humans. Analysts using computer software for qualitative analysis can reuse existing codes from previous interviews to code a new portion of data representing very similar meanings. On the other hand, LLMs code each interview with no knowledge of previous codes, resulting in duplicate codes. Consequently, this requires the researcher using an LLM to identify these duplicates and expunge them to arrive at a final codebook containing only unique codes. In this contribution, I performed the identification of duplicates also with the LLM. [20]

In a recent contribution for testing analytical saturation when using LLMs, a colleague and I (DE PAOLI & MATHIS, 2024) suggested a method for constructing the codebook of unique codes incrementally. This is achieved by comparing each set of codes, generated by the LLM for each interview, with a cumulative codebook of unique codes, as shown in Figure 3. If a code is already in the unique codebook (i.e. a duplicate) then this code is discarded by the LLM. Conversely, if the code is not already present in the unique codebook (i.e. it captures something in the data that no other code does) then this is added to the codebook. At the first stage of this process the unique codebook is equal to the list of codes for the first interview. In further stages (until all the interviews' codes are checked), the unique codes are added to the unique codebook for subsequent comparison.

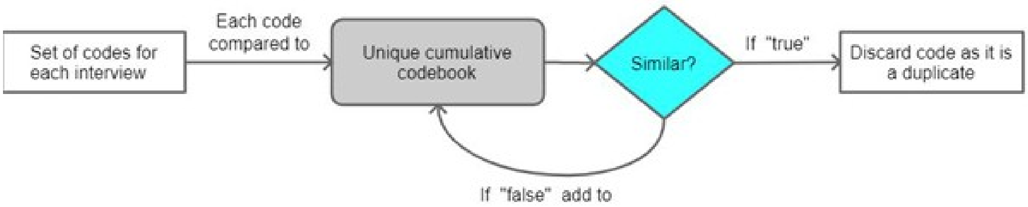

Set of codes for each interview
Each code compared to
Unique cumulative codebook
Similar?
If "true"
Discard code as it is a duplicate
If "false" add to

Figure 3: Process for the reduction of duplicate codes [21]

With Prompt_2, I asked the LLM to determine if each code conveys a "similar" idea or meaning to any of the unique codes in the cumulative codebook. The model yields the response "true" if the code is already in the unique codebook list (i.e. there is another code that conveys a similar meaning). Instead, if the LLM deems that a code is "not similar" to any of the unique codes it will respond "false", and add the code to the unique codebook. The process is then repeated until the last set of codes is checked for duplicates.

```
for t in range(l):
value=codes[t]
prompt = f"""
Then, determine if code: ```{code}``` conveys a very similar idea or meaning to any element in the list combined_unique: {", ".join(combined_unique)}. Your response should be either 'true' (similar idea or meaning) or 'false' (no similarity).
Format the response as a json file using the key code_in_combined_u
"""
```

Box 3: Prompt_2 identification of unique codes [22]

In DE PAOLI and MATHIS (2024), we observed that the unique cumulative codebook and the total cumulative codebook (all the codes from Prompt_1, including duplicates) both progress in the form of linear relations. The unique cumulative codebook grows at a slower rate than the total cumulative codebook (because duplicates are removed from each interview set). We postulated that this slower growth amounts to a form of saturation, and that the ratio of the unique codes over the total codes can offer a synthetic measure of inductive thematic saturation (ITS). Because of the linear growth of both codebooks this ratio corresponds to the ratio of the raise and run formula between the two linear functions. This ratio is a number between 0 (the ideal case where all the interviews in a dataset are the same and an LLM would generate the same initial codes always) and 1 (the ideal case when all the interviews are completely heterogeneous, with the LLM always generating different codes for each). The ITS ratio should not be too close to 1 as a way of measuring the presence of analytical saturation. In Figure 4, I present a simplified and fictious interpretation of the concept where the unique codebook has a slower growth than the total codebook. The ratio between the cumulative number of unique and total codes provides the ITS metric. In the example of Figure 4, this is the ratio of 103 unique codes over 370 total codes (at the last interview in the dataset e.g. 25), or 103/370=*0.28*.

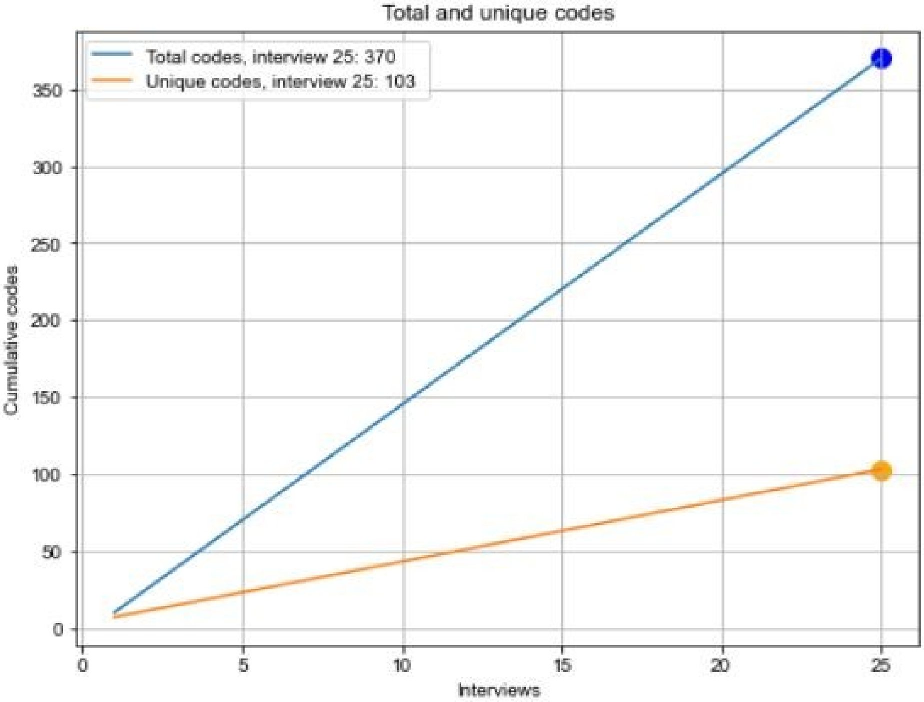


Figure 4: Simplified interpretation of saturation of initial codes for LLMs [23]

3.2 Phase 3 of thematic analysis

A third new prompt is Prompt_3 (Box 4) which I leveraged for Phase 3 of the TA. In the prompt, I followed a similar logic to that of Prompt_1. First, the language I used keeps the process open-ended without fixing beforehand the number of themes the model should generate (differently from previous proposed prompting). Second, I designed Prompt_3 again to align with the lexicon of TA, asking the model to determine themes by sorting and grouping initial codes. Third, in the opening of the prompt I asked the model to review the codebook of unique codes beforehand, ahead of defining the themes. The themes are then obtained by passing the unique codebook to the model (as a list variable called *codes_list*). In the prompt, I asked the model to read the codes in advance because during the engineering of the prompt I noticed that the model was encountering some attention problems. The first codes in the list did seem to skew the elicitation of themes, whereas later codes were then used less to identify themes. By asking the model to read the codes prior to processing, this errant behaviour comes under better control and is minimised. With this prompt, I also instructed the model about the structure of each code in the list, which includes the code numerical index, its name and description. The response I obtained from the LLM is a set of themes with a name, a description and the list of codes composing the theme (a numerical index which allows the codes to be readily retraced in the dataframe in preparation for the write up).

```
prompt = f"""
Read first the list of initial codes of my thematic analysis: {", ".join(codes_list)}. Initial
codes are in the following format: [index]: code_name. code_description.\
Determine all the possible themes by sorting, comparing, and grouping initial codes.
Provide a suitable number of themes along with a name, a dense description (70 words),
and a list of codes (index) for each theme.
Ensure the themes capture the richness and diversity of the initial codes\
"""
```

Box 4: Prompt_3 for the identification of themes [24]

3.3 Thematic map

I suggest using the results of Prompt_3 further, to create a thematic map (see BRAUN & CLARKE, 2006, p.90, for an elaboration of the concept). Thematic maps are "mapping aids [...] that enhance the researcher's ability to identify and understand potential themes in relation to each other, and the overall dataset" (TERRY, HAYFIELD, CLARKE & BRAUN, 2017, p.28). A thematic map does not allow researchers to produce new themes, rather it is a way to map the potential relations across themes. A thematic map serves as a visual tool for exploring and revising the analysis (BRAUN, CLARKE & WEATE, 2016). A visual map of themes may also help investigators identify different levels or sub-themes. [25]

The analyst would usually draw the map by connecting themes and sub-themes according to their interpretation of emerging thematic connections. In the following my approach is instead to use the response from Prompt_3 to identify the number of codes that are shared across themes. Using this information, I draw a sort of network of nodes and edges, akin to a very basic social network analysis graph (WASSERMAN & FAUST, 1994), leveraging shared connections across themes and the robustness of their strength. The procedure I propose here may differ from how human analysts draw their thematic maps. Nonetheless, the result is still a thematic map: a visual aid for thinking about themes and their connections. [26]

3.4 Evaluation

I conducted an evaluation to assess if themes produced with Prompt_3 are "good enough" and potentially comparable with themes produced by humans. For the comparison, I operated at two levels: 1. performed a semantic similarity check using a language model (LM) called SBERT (REIMERS & GUREVYCH 2019) to compute a similarity score; 2. applied two additional LLMs to assess if human themes and LLM themes (from GPT3.5-Turbo-16k) convey a similar meaning. [27]

Semantic similarity is a measure of likeness between texts and is a common natural language processing (NLP) task (see for a review CHANDRASEKARAN & MAGO, 2021). For assessing semantic similarity, a user provides two texts to a pre-trained LM (e.g. two pairs of themes/descriptions) and the LM is asked to compute a similarity score. The LM computes a cosine similarity score (with values between -1 and 1) for the pair of texts. When the score is close to 1 there is high semantic similarity: the two texts are semantically alike. If the score is close to zero, there is no semantic similarity. Lastly if the score is close to -1, then the two texts are at the opposite ends of similarity. Therefore, comparing pairs of themes for cosine similarity produced by humans and by an LLM can support researchers to identify if the themes produced by an LLM have at least some degree of semantic likeness with human-generated themes. [28]

For the second evaluation, I used two LLMs (different from the LLM adopted for the analysis) and prompted them to tell me if two pairs of themes/descriptions (human-generated Vs LLM-generated) conveyed the same meaning or not. I employed the LLMs Mistral-7b-Instruct-V0.2[2] and Llama-8b-Instruct[3], two open-source models which are smaller than GPT3.5-Turbo-16k but that still have significant efficiency to operate the textual comparison. I deployed the evaluation LLMs in a google colab python notebook and I prompted them to deliver a score between 0 and 1 indicating how well a pair of themes convey the exact same meaning. Within the prompt I also asked the LLMs to supply a justification for the score. [29]

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2 [Date of Access: May 10, 2024]

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct [Date of Access: May 10, 2024]

3.5 Material

For performing this work, I used two open access datasets. The first is composed of 15 semi-structured interviews on the practices of qualitative researchers, with a focus on open research and open access. I adopted this dataset for producing codes and themes with the LLM and for producing thematic maps. The dataset comes from the project "Fostering Cultures of Open Qualitative Research" (HANCHARD & SAN ROMAN PINEDA, 2023) conducted by the University of Sheffield and was retrieved from its open access repository[4]. In this paper, my focus is to illustrate the process of doing TA with LLMs rather than delving into specific results. Consequently, describing certain aspects of the datasets is less pertinent than the exploration of the methodological approach. [30]

I utilised a second dataset for the evaluation. This is a dataset of nine interviews on "Teaching Data Science" (CURTY et al., 2022). The reasons why I am using a different dataset are as follows: First, for the dataset I mentioned above, no analysis is available beyond a simple report which does not appear to include any themes or second order categories that could be employed as the basis for a comparative evaluation. Second, I used the "Teaching Data Science" dataset in previous work and an evaluation was already carried out to check the validity of LLM themes produced with the old prompts. For this dataset, a set of themes is available in a scientific report (ibid.). These themes can be leveraged for comparative purposes. [31]

4. Results

4.1 Phase 2: initial codes

I will present now the generation of the initial codes (Phase 2 of TA) with Prompt_1. In Table 1, I show the total number of codes produced by the model with Prompt_1, and the total of unique codes when duplicates are removed (using Prompt_2). As I discussed in Section 3, with Prompt_1, I instructed the LLM to elicit the code/labels which are accompanied also by a description and a quote. I tested Prompt_1, requesting the model to respond with code descriptions of two different lengths (40 words and 15 words). As I display in Tables 1 and 2, this led the model to supply a different total (and unique) number of codes. With the 15 words description request, the model returns a larger number of initial codes. This behaviour of the LLM might be explained by the model adjusting its response (in terms of the number of codes) based on the available tokens. A shorter description for each code offers potentially more tokens that the LLM can use for supplying additional codes. On average for each interview, the model generated around 12 codes for the 40 words and 15 codes for the 15 words descriptions.

[4] https://orda.shef.ac.uk/articles/dataset/Fostering_cultures_of_open_qualitative_research_Dataset_2_Interview_Transcripts/23567223 [Date of Access: October 24, 2023]

Codebooks	**Nr. of Codes**
Total cumulative	183
Unique cumulative	93

Table 1: Codebooks with a 40 words description

Codebooks	**Nr. of Codes**
Total cumulative	211
Unique cumulative	144

Table 2: Codebooks with a 15 words description [32]

The numbers in Table 1 and Table 2 enable me to calculate the inductive thematic saturation (ITS) metric, as I discussed earlier by calculating the ratio between the unique and total codes:

$$\text{ITS (40 words)} = \frac{\text{Unique codes}}{\text{Total codes}} = \frac{83}{193} = 0.51$$

$$\text{ITS (15 words)} = \frac{\text{Unique codes}}{\text{Total codes}} = \frac{144}{211} = 0.68$$

There is some saturation as both measures move away from the value of 1 (which is the case where no duplicates are found). Whilst the prompt with the shorter description request allowed me to produce more initial codes, it also appears that the prompt requesting a longer description allows for better inductive thematic saturation or ITS. In essence, the reduction of the codebook I performed appears stronger for codes with longer descriptions. [33]

In Table 3 and Table 4, I show some examples of the codes provided by the LLM which include the code name, a description of the code and a quote from the dataset.

Code Name	**Description**[5]	**Quote**
Challenges of open data	"The respondent acknowledges the challenges of making qualitative data open, including issues of anonymity, confidentiality, and the time-consuming process of preparing the data for sharing. They also mention ethical concerns about the potential use of open data by other researchers."	"You know, so quantitative data, you know, that might just be a spreadsheet of numbers for example, to, you know, put it simplistically I suppose and that is not too difficult to make code, you know, the challenges with that as well. But That's not too difficult. Whereas, if you've got, you know, for example, photographs, that might have people in them, there's issues around anonymity, and Data confidentiality."
Data description	"The importance of well-described open datasets, including clear and accurate descriptions, recognized terminology, and interoperability across domains"	"I want to know that the description will be clear and will give me enough information that I know, this is the dataset that I need, and I'm going to be able to use it and the variables are going to be well described."

Table 3: Examples of codes generated with Prompt_1 and up to 40 words description.

Code Name	**Description**	**Quote**
Open-ended questions	"Using open-ended questions in semi-structured and unstructured interviews"	"I tend to use more open-ended questions in my interviews. They're semi-structured interviews, but they can quite often be open as well."
Open access data	"The respondent discusses the idea of publishing data sets and the ethics surrounding it."	"I think all data sets that would be made public, but I should have thought more about the ethics of it."

Table 4: Example of codes generated with Prompt_1 and up to 15 words description (GPT3.5-Turbo-16k, December 3, 2023) [34]

[5] These parts of Tables 3, 4 and 7 are the verbatim descriptions provided by the model.

4.2 Phase 3: themes

For researchers, Phase 3 of TA entails sorting and organising initial codes into themes. To perform Phase 3, I passed the entire list of unique initial codes (in the form of their numerical index, name and description) in Prompt_3, and in response from the model, I obtained a set of themes, their descriptions and the list of the underlying codes (as their numerical indexes). [35]

Since, with the prompt, I asked the model to identify an open-ended number of themes, the type and number of themes can vary when the prompt is run multiple times. This aligns with the observation that LLMs do not display an entirely deterministic behaviour and will potentially produce marginally different responses with the same prompt (MADAAN et al., 2023). However, as I will demonstrate later, an overlap occurs between the responses when Prompt_3 is run multiple times. [36]

First, I suggest looking at (part of) the response of the model from Prompt_3. In Box 5, I present what a theme looks like when formatted as a json object (with the pairs key: value(s)):

```
"Theme name": "Support and Guidance in Open Research"
"Theme description": "This theme highlights the need for support and guidance in conducting open research, particularly in the context of qualitative research. It includes codes related to the need for clear guidelines, funding, and education on open research practices."
"Underlying codes": "['[11]', '[29]', '[38]', '[61]', '[62]', '[69]', '[70]', '[71]', '[72]', '[78]', '[88]', '[89]', '[91]', '[92]', '[103]', '[110]', '[120]', '[130]']"
```

Box 5: Example of a theme as a json object from the model response (GPT3.5-Turbo-16k, December 3, 2023) [37]

The model returns a theme name, a theme description and then the index numbers of the codes composing the theme. The index corresponds to the row number of the code in the dataframe/table of unique codes. For a researcher, knowing the index is sufficient for identifying which codes compose the theme. The index value [11] for example corresponds to the 12th code in the dataframe table (since the index starts at 0). The index value [29] corresponds to the 30th code, etc. [38]

In Table 4, I expose the 9 themes generated using the 40-words unique codebook. The number of codes composing each theme varies from 18 to five. Themes cover aspects related to open science and open access, and to performing qualitative research. Looking closely at the themes, I suggest it is possible to notice that the theme "Future directions and implications" does not encompass many codes and might not be a relevant theme. Also, the theme "Challenges in humanities research" appeared to be composed of codes all adjacent to one another (by looking at the index numbers), which suggested to me that this theme was derived from codes coming from just one of the 15 interviews, and therefore may also not be an entirely relevant theme.

Theme Name	**Theme Description**[6]	**Codes Total**	**Examples of Codes**
Open science and data sharing	"This theme focuses on the respondent's commitment to open science and their efforts to make their research transparent and accessible. It includes codes related to open science practices, data sharing, ethics, consent, and challenges faced in practicing open science."	17	[77] Barriers to open access
Qualitative data management and analysis	"This theme explores the challenges and approaches to managing and analysing qualitative data. It includes codes related to data collection, data analysis, qualitative research tools, data management planning, and the limitations of quantifying qualitative data."	16	[43] Analytical coding
Challenges and barriers in open research	"This theme focuses on the challenges and barriers faced in practicing open research, particularly in the context of qualitative research. It includes codes related to funding, anonymization, consent processes, participant privacy, copyright restrictions, and the lack of guidance and support for open research."	18	[60] Challenges in making research open
Research methods and approaches	"This theme explores the various research methods and approaches used by the respondent. It includes codes related to data collection methods, data analysis approaches, research tools, interdisciplinary research, and the use of theoretical frameworks."	18	[16] Qualitative research tools
Participant recruitment and consent	"This theme focuses on the challenges and considerations related to participant recruitment and obtaining informed consent. It includes codes related to willingness to archive data, variations in participants' willingness to have their data archived, considerations of regional accents in transcriptions, and the importance of consent forms."	8	[53] Data management and consent

[6] Reported below are the verbatim descriptions provided by the model when I used Prompt_3. The same applies for Table 7.

Theme Name	Theme Description	Codes Total	Examples of Codes
Future directions and implications	"This theme explores the respondent's future plans for open qualitative research and the implications of open science practices. It includes codes related to future directions, guidelines, resources, funding, and the potential benefits of reusing and sharing qualitative data."	5	[70] Future plans
Field of research and expertise	"This theme focuses on the respondent's field of research and their expertise. It includes codes related to the respondent's broad area of research, their background and role, their experience in specific fields such as environmental psychology and criminal justice research, and their academic journey. "	13	[64] Workflow and policy questions
Open access and data sets	"This theme explores the respondent's thoughts on open access and the publication of data sets. It includes codes related to open access, finding and accessing data sets, ownership of data, concerns about potential misuse or unethical purposes, and the benefits and challenges of making data open access."	15	[21] Finding and accessing data sets
Challenges in humanities research	"This theme focuses on the specific challenges faced in humanities research, particularly in relation to open research. It includes codes related to working with textual data, interdisciplinary research, copyright restrictions, challenges in making research open, and the lack of understanding and guidelines for data sharing in humanities research."	8	[57] Copyright restrictions
Data analysis and interpretation	"This theme explores the respondent's approaches to data analysis and interpretation. It includes codes related to data analysis tools, qualitative methods, thematic analysis, interpretive research, case studies, and the importance of providing evidence to support arguments."	9	[73] Qualitative methods

Table 4: Themes generated by the model from the 40-words unique codebook ("base" iteration) (GPT3.5-Turbo-16k, December 3, 2023) [39]

In Table 5, I present the theme names from three additional runs/iterations of Prompt_3. The number of themes produced by the model may vary between iterations. However, I suggest that it is also possible to recognise themes which are repeated across all iterations (including the "base"), which I highlight in italics in Table 5. Some other themes also appear similar and have similarities with those in the "base" iteration (also in italics). This observation suggested to me that it is possible to identify a fundamental set of themes, despite potential variability in the model's responses in subsequent iterations.

Nr. of Theme	**Iteration_2**	**Iteration_3**	**Iteration_4**
1	*Open science and data sharing*	*Open science and data sharing*	*Open science and data sharing*
2	*Qualitative data management and analysis*	*Qualitative data management and analysis*	*Qualitative data management and analysis*
3	*Challenges and barriers in open research*	*Challenges and barriers in open research*	*Challenges and barriers in open research*
4	Interdisciplinary research and collaboration	Field of research and expertise	Interdisciplinary research and collaboration
5	*Research methods and approaches*	*Research methods and approaches*	*Research methods and approaches*
6	*Ethical considerations and participant consent*	*Participant recruitment and consent*	*Ethical considerations and participant consent*
7	*Future directions and challenges in research*	*Future directions and implications*	*Future directions and planning*
8		Access and usability of research data	Specific research fields and domains
9		Impact and benefits of open access	
10		Challenges in humanities research	

Table 5: Themes generated using the 40-words unique codebook, 3 additional iterations [40]

4.3 Proposition for a thematic map

The list of code indexes (from the "base" themes in Table 4) also provides a researcher the opportunity to cross-check connections between themes based on the codes they share. For the nine themes, there are 45 pairs of combinations. Seventeen of them had zero shared codes. In Table 6, I show the pair of themes sharing at least one code.

Theme 1	**Theme 2**	**Shared Codes**
Open science and data sharing	Challenges and barriers in open research	6
Open science and data sharing	Research methods and approaches	1
Open science and data sharing	Participant recruitment and consent	1
Open science and data sharing	Future directions and implications	2
Open science and data sharing	Open access and data sets	7
Open science and data sharing	Challenges in humanities research	2
Qualitative data management and analysis	Research methods and approaches	9
Qualitative data management and analysis	Field of research and expertise	2
Qualitative data management and analysis	Open access and data sets	1
Qualitative data management and analysis	Challenges in humanities research	3
Qualitative data management and analysis	Data analysis and interpretation	8
Challenges and barriers in open research	Research methods and approaches	1
Challenges and barriers in open research	Participant recruitment and consent	1
Challenges and barriers in open research	Future directions and implications	2
Challenges and barriers in open research	Field of research and expertise	1
Challenges and barriers in open research	Open access and data sets	7
Challenges and barriers in open research	Challenges in humanities research	3

Theme 1	**Theme 2**	**Shared Codes**
Challenges and barriers in open research	Data analysis and interpretation	1
Research methods and approaches	Field of research and expertise	3
Research methods and approaches	Challenges in humanities research	2
Research methods and approaches	Data analysis and interpretation	7
Participant recruitment and consent	Open access and data sets	2
Future directions and implications	Open access and data sets	1
Future directions and implications	Challenges in humanities research	1
Field of research and expertise	Challenges in humanities research	4
Field of research and expertise	Data analysis and interpretation	2
Open access and data sets	Challenges in humanities research	3
Challenges in humanities research	Data analysis and interpretation	2
	Average	*2.75*

Table 6: Shared codes between themes (40 words) [41]

In Table 6, some themes only share one code, indicating that their connection may be potentially weak. The average number of shared codes is 2.75. In Section 3, I proposed to use basic network analysis concepts to build a thematic map, serving as a visual representation to explore the connections across themes. I suggest focusing solely on pairs of themes where the shared codes exceed the average, to construct a thematic map (Figure 5). This choice impacts negatively on some themes, with, for example, the theme "Participant recruitment and consent" being expunged from the map. The number of shared connections used to build these maps is ultimately a decision left to the analyst. Another approach could be to include all themes in the map. In any case, in Figure 5, I advocate that all the themes appear connected, even though the relative strength of the connection based on the shared codes varies. Earlier I noted that the theme "Challenges in humanities research" was possibly not relevant since its underlying codes are all coming from the same interview. In Figure 6, I present the map when this code is removed and this helps to see that the themes may potentially compose two strong concepts: The first is related to open science/open access, and the second to the process of conducting qualitative research.

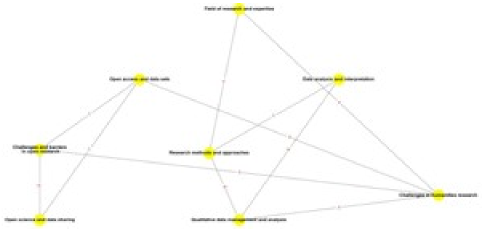

Field of research and expertise
Open access and data sets
Data analysis and interpretation
Challenges and barriers in open research
Research methods and approaches
Challenges in humanities research
Open science and data sharing
Qualitative data management and analysis

Figure 5: Thematic map (40 words) built using basic social network analysis techniques. Please click here for an enlarged version of Figure 5.

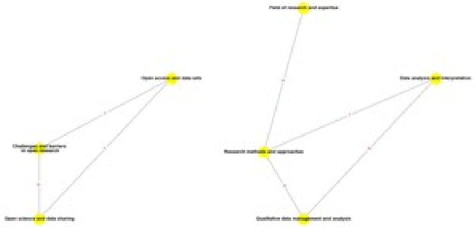

Field of research and expertise
Open access and data sets
Data analysis and interpretation
Challenges and barriers in open research
Research methods and approaches
Open science and data sharing
Qualitative data management and analysis

Figure 6: Thematic map (40 words) when one potentially weak theme is removed. Please click here for an enlarged version of Figure 6. [42]

This process can be repeated for the 15 words description set of themes. In Table 7, I display the "base" iteration and the themes identified by the model. The number of codes composing a theme is generally higher (because there are more codes in the unique codebook), while the number of themes is similar to the previous iterations. I argue that some themes also overlap with the themes from Table 4, indicating some consistency.

Theme Name	**Theme Description**	**Codes Total**	**Examples of Codes**
Data collection and analysis methods	"This theme encompasses the various methods and approaches used by the respondent in collecting and analysing qualitative data. It includes codes related to data collection techniques, such as interviews and workshops, as well as the use of qualitative analysis methods like mapping and diagrams."	45	[135] Case studies
Open science and open access	"This theme focuses on the respondent's commitment to open science and open access. It includes codes related to the sharing and reuse of data, efforts to make research transparent and accessible, and the challenges and benefits of open access in qualitative research."	35	[9] Reusing and sharing data

Theme Name	**Theme Description**	**Codes Total**	**Examples of Codes**
Challenges and considerations in qualitative research	"This theme explores the challenges and considerations specific to qualitative research. It includes codes related to the difficulties of working with qualitative data, the need for clear guidelines and funding, ethical considerations, and concerns about representation and interpretation of data."	56	[71] Workflow for open research
Research focus and field of study	"This theme relates to the respondent's specific research focus and field of study. It includes codes related to the respondent's area of expertise, such as museum gallery and heritage studies, cultural gerontology, environmental psychology, and management accounting."	10	[113] Research focus
Support and guidance in open research	"This theme highlights the need for support and guidance in conducting open research, particularly in the context of qualitative research. It includes codes related to the need for clear guidelines, funding, and education on open research practices."	18	[88] Data training
Use of technology and software	"This theme focuses on the use of technology and software in qualitative research. It includes codes related to the use of software for data analysis, such as Nvivo and Excel, as well as the use of visual methods and multimedia research outputs."	20	[31] Self-tracking practices
Ethical considerations and participant confidentiality	"This theme explores the ethical considerations and concerns related to participant confidentiality in qualitative research. It includes codes related to obtaining permissions, anonymizing data, and the potential impact of open access on participant confidentiality."	7	[136] Confiden-tiality
Relevance and implications of research	"This theme focuses on the relevance and implications of qualitative research. It includes codes related to the potential applications and benefits of research findings, as well as the need to consider the context and implications of research."	13	[111] Role of govern-ment and social research

Table 7: Themes generated by the model from the 15 words unique codebook ("base" iteration) (GPT3.5-Turbo-16k, December 3, 2023) [43]

Further iterations of Prompt_3, as I illustrate in Table 8, show some variation in the number of themes generated, but again there is consistency across the iterations (see themes in italics for exact matches, and similarity).

Nr. of Theme	**Iteration_2**	**Iteration_3**	**Iteration_4**
1	*Data collection and analysis methods*	*Data collection and analysis methods*	*Data collection and analysis methods*
2	*Open science and open access*	*Open science and open access*	*Open science and open access*
3	*Challenges and considerations in qualitative research*	*Challenges and considerations in qualitative research*	*Challenges and considerations in qualitative research*
4	Research focus and field of study	Interdisciplinary research and methods	Research focus and field of study
5	Support and guidance in open research	Open access and research culture	Open research and collaboration
6	Interdisciplinary research and methods	Research methods and tools	Qualitative vs quantitative research
7	*Ethical considerations and participant confidentiality*	*Ethics and privacy in qualitative research*	*Ethical considerations and participant confidentiality*
8	Open access and copyright issues	Career and professional identity	Challenges and opportunities in open access
9	Researcher's background and role		Research training and support
10			Relevance and implications of research

Table 8: Themes generated by the model from the 15 words unique codebook, three additional iterations [44]

In Table 9, I display pairs of themes that share at least one code. Using these pairs, I illustrate in Figure 7 the connections between themes when the shared codes exceed the average. In the map, I show that three themes are highly connected, potentially forming the kernel of a concept, and connecting more the relationship between the qualitative research themes and the open access themes, unlike Figure 6. However, I would suggest that the concepts from the map are more clearly identifiable in Figure 6. Moreover, when I plotted only the connections above the average, the theme on "Ethical considerations" was removed from the map. By lowering the shared connection requirement to six

(Figure 8), I suggest that the "Ethical consideration" theme clearly appears connected to the "Challenges" theme. In fact, the "Ethical considerations" and the "Data collection and analysis" themes may be sub-themes of the "Challenges and considerations in qualitative research" theme, which forms part of the conceptual kernel of this map. [45]

It is a matter of choice for the human analyst to decide how to use the connections based on shared codes to draw a map. A researcher may think, for instance, that themes which share a lower number of codes may nevertheless reveal more unexpected connections and new patterns. The examples and interpretations I have illustrated here are designed for illustrative purposes of the idea.

Theme 1	**Theme 2**	**Number of Shared Codes**
Data collection and analysis methods	Open science and open access	1
Data collection and analysis methods	Challenges and considerations in qualitative research	9
Data collection and analysis methods	Research focus and field of study	3
Data collection and analysis methods	Use of technology and software	10
Data collection and analysis methods	Relevance and implications of research	1
Open science and open access	Challenges and considerations in qualitative research	17
Open science and open access	Support and guidance in open research	14
Open science and open access	Use of technology and software	3
Open science and open access	Ethical considerations and participant confidentiality	5
Open science and open access	Relevance and implications of research	7
Challenges and considerations in qualitative research	Research focus and field of study	2
Challenges and considerations in qualitative research	Support and guidance in open research	12

Theme 1	**Theme 2**	**Number of Shared Codes**
Challenges and considerations in qualitative research	Use of technology and software	2
Challenges and considerations in qualitative research	Ethical considerations and participant confidentiality	6
Challenges and considerations in qualitative research	Relevance and implications of research	9
Support and guidance in open research	Ethical considerations and participant confidentiality	1
Support and guidance in open research	Relevance and implications of research	6
Use of technology and software	Relevance and implications of research	1
	Average	*6.5*

Table 9: Connections between themes (15 words).

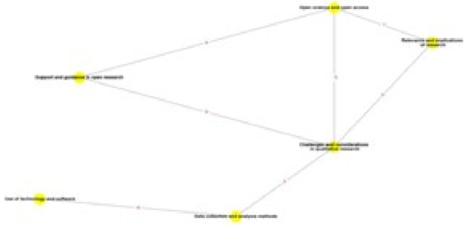

Open science and open access
Relevance and implications of research
Support and guidance in open research
Challenges and considerations in qualitative research
Use of technology and software
Data collection and analysis methods

Figure 7: Thematic map with at least seven shared codes. Please click here for an enlarged version of Figure 7.

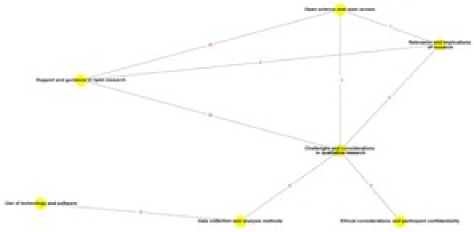

Open science and open access
Relevance and implications of research
Support and guidance in open research
Challenges and considerations in qualitative research
Use of technology and software
Data collection and analysis methods
Ethical considerations and participant confidentiality

Figure 8: Thematic map with at least six shared codes. Please click here for an enlarged version of Figure 8. [46]

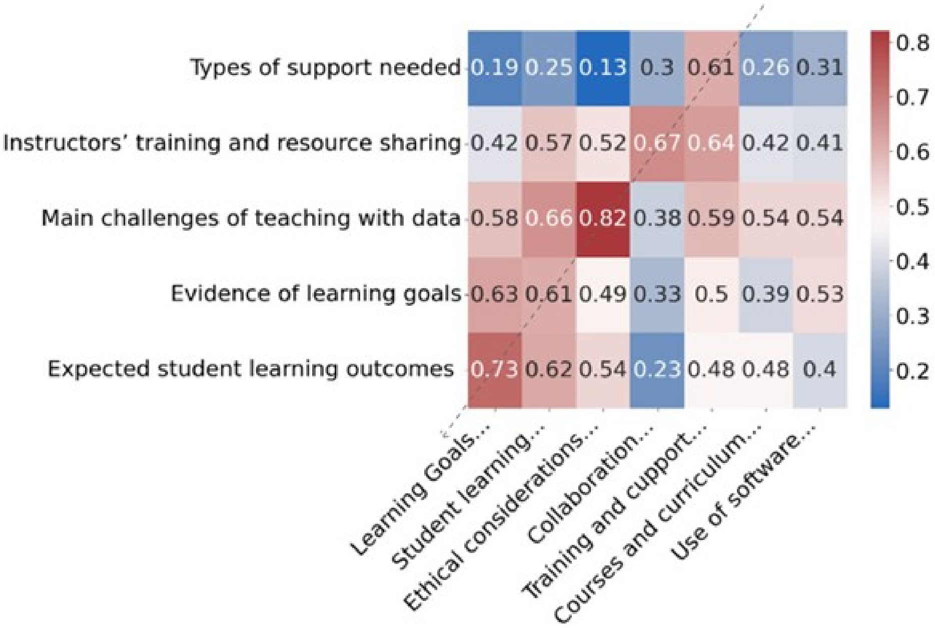
4.4 Evaluation

In this section, I offer an evaluation of the themes produced by the LLM with the new Prompt_3. My objective is to assess whether the LLM themes approximate the themes elicited by human analysts on the dataset on "teaching data science" (CURTY et al., 2022) that I briefly presented earlier. As I discussed in Section 3.5, the evaluation is based on comparing the LLM themes with both human-generated themes and the themes created with previous/old prompting (fixed themes, no TA language). In Table 10, I offer, in the second column, the original themes from CURTY, GREER and WHITE (2021) that I extracted from their report. Their analysis has five clear themes. In the third column, I show the themes I produced with Prompt_3 (one iteration) which is open-ended, and in this iteration the LLM generated seven themes. In the last column, I display a set of themes I generated using the old prompt with a fixed number of themes (which I set at seven, same as the number from Prompt_3). The themes I obtained from the LLM (with the two different prompts) have been re-ordered manually by me after performing the semantic similarity scores calculation to pair them for initial similarity.

ID	Original (Human Analysts)	Prompt_3 (Open-Ended)	"Old" Prompt (Response Fixed at Seven Themes)
1	Expected student learning outcomes and ways students engage with data	Learning goals and critical thinking in data analysis	Teaching with data and ethical considerations
2	Evidence of learning goals in instructional praxis	Student learning and manipulation of data	Student learning and support
3	Main challenges of teaching with data	Ethical considerations in teaching with data	Teaching methods and resources
4	Instructors' training and resource sharing	Collaboration and sharing of resources	Data collection and analysis
5	Types of support needed	Training and support for teaching with data	Training and support for instructors
6		Courses and curriculum in data science	Teaching and research integration
7		Use of software and tools for data analysis	Data literacy and skills development

Table 10: Themes from human analysts, LLM with new prompt and with old prompt [47]

In Figure 9 and Figure 10, I expose the results of the semantic similarity exercise. The values in the matrices are the semantic similarity scores between pairs of themes. In Figure 9, I report the comparison between the human themes (vertical) and the LLM themes from Prompt_3 (horizontal). I traced in the Figure 9 a diagonal (dotted line) corresponding to the five human themes and five most similar LLM themes. The scores on the traced diagonal vary between 0.82 to 0.61. For some pairs, there is quite high similarity (pairs 3 and 1) and for the others (2, 5 and 4) there is good similarity (>0.6). I argue, therefore, that the new prompt can produce themes which are "good enough" in terms of their semantic similarity with human themes. In Figure 10, I present the similarity scores comparing the human produced themes with the LLM themes from the old prompting style. The dotted line I trace on the diagonal illustrates that, in most cases, the results are comparable to those of Figure 9. One exception is perhaps the last of the human themes which does seem only marginally covered by the LLM themes. In these semantic similarity tests, I demonstrate that the LLM produces themes which, at least in part, are semantically alike to themes produced by human analysts.



	Learning Goals...	Student learning...	Ethical considerations...	Collaboration...	Training and support...	Courses and curriculum...	Use of software...
Types of support needed	0.19	0.25	0.13	0.3	0.61	0.26	0.31
Instructors' training and resource sharing	0.42	0.57	0.52	0.67	0.64	0.42	0.41
Main challenges of teaching with data	0.58	0.66	0.82	0.38	0.59	0.54	0.54
Evidence of learning goals	0.63	0.61	0.49	0.33	0.5	0.39	0.53
Expected student learning outcomes	0.73	0.62	0.54	0.23	0.48	0.48	0.4

Figure 9: Semantic similarity scores human themes vs. themes from Prompt_3

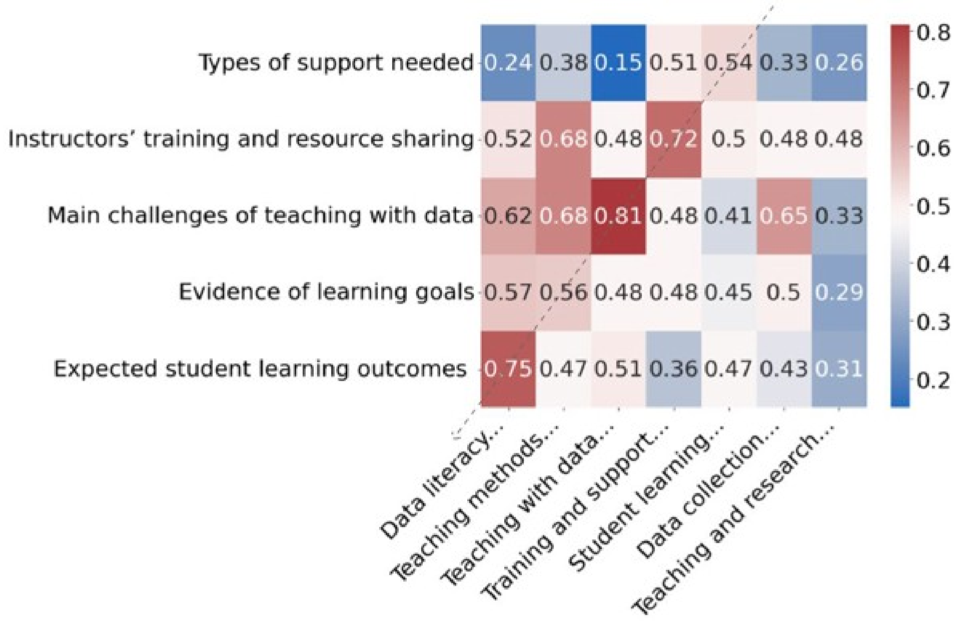


Figure 10: Semantic similarity scores human themes vs. themes from old prompting [48]

As a further evaluation, I employed two different LLMs to assess/compare if human themes and LLM themes from Prompt_3 convey a similar meaning. For this, I used a prompt asking two LLMs (Llama-3-8b and Mistral-2-7b) to respond with a score between 0 (completely different meaning) and 1 (exact same meaning) and a justification for the score. In Table 11, I present the outcomes of this comparison for the IDs 1 to 5 from Table 10. Moreover, in the last two rows of Table 11, I propose the results of the comparison of a theme with itself (i.e. LLMs were asked to compare the same text) and of a human theme with one of the LLM themes which received low semantic similarity scores. These last two rows validate my process and, as expected, in the first case the score is 1 and the models identify that the two themes are the same piece of text. In the case of the last row (where the human theme 1 is compared to the LLM theme 7), the models also identify that they are quite different. [49]

In the other rows of the table, I show that in four out of five comparisons the LLMs gave high scores (>0.8), identifying that the pair of themes convey very similar meanings. These high scores are supported by clear justifications of similarity in meaning. Only for one pair of themes (Row 4, "Instructors' training and resource sharing" vs. "Collaboration and sharing of resources") the score appears relatively low, with the models' justifications also clearly stating that the two convey potentially different meanings. In general, then, I suggest that the themes produced by the LLM with Prompt_3 can be deemed "good enough" as four out of five themes produced by the human analysts were identified by the LLM.

ID	Test_1: Llama[7]	Test_2: Mistral
1	Score: 0.83 "Justification: In both texts the importance of critical thinking and data literacy has been emphasised, with the second text providing more specific details on learning goals and careers."	Score: 0.8 "Justification: Both texts discuss the importance of critical thinking skills and data literacy, with the second text expanding on the learning goals and alignment to various careers."
2	Score: 0.800. "Justification: The texts share a common theme, with similar messages but varying levels of detail and scope."	Score: 0.85 "Justification: Both texts discuss the importance of students' abilities to manipulate and analyse data. The first text focuses more on the specific skills involved, while the second text provides a broader context."
3	Score: 0.83 "Justification: Both texts discuss the importance of ethics in teaching with data, but the first text is broader in scope, mentioning multiple challenges, whereas the second text focuses on ethical considerations, providing a more specific perspective."	Score: 0.95 "Justification: The two texts share a similar focus on the ethical challenges of teaching with data, with the first text expanding to include additional considerations such as data availability and re-use."
4	Score: 0.5 "Justification: The two texts do not convey the exact same message, but they share a common theme of collaboration and sharing among instructors."	Score: 0.3 "Justification: The texts discuss related but distinct aspects of instructors' professional development and resource sharing. The first text focuses on how instructors acquire the skills and resources to teach with data, while the second text discusses the benefits and methods of sharing resources among instructors."
5	Score: 0.8 "Justification: The two texts convey a similar message, focusing on the need for support and resources for teaching with data, but the second text provides more specific details on the types of support needed."	Score: 0.85 "Justification: Both texts emphasize the importance of ongoing workshops, technical support, and access to computing facilities. The first text also mentions the need for infrastructure improvements and resources, while the second text focuses on the training and support for instructors. Despite these differences, the core message remains the same."

[7] The justifications reported below are the verbatim text provided by the two LLMs.

ID	Test_1: Llama	Test_2: Mistral
1-1	Score: 1.0 "Justification: The two texts are identical, conveying the exact same message with no differences in wording or meaning."	Score: 1.0 "Justification: Both texts convey the exact same message regarding the importance of developing critical thinking skills and advancing students' data literacy."
1-7	Score: 0.25 "Justification: The two texts share a common theme of data analysis, but convey different ideas, with the first focusing on learning outcomes and data literacy, while the second focuses on software and tools used for data analysis."	Score: 0.2 "Justification: The first text focuses on student learning outcomes and engagement with data, while the second text discusses software and tools for data analysis. Although both texts relate to data analysis, they convey different aspects of the topic."

Table 11: Human and LLM themes comparison, performed with LLama-3 and Mistral-2 [50]

5. Discussion and Conclusion

In this paper, I proposed a set of new prompts for doing TA with an LLM. Filling the gap in the previous procedure I described in DE PAOLI (2023b), I introduced open-ended prompts for the generation of initial codes and themes from the analysis of a dataset of 15 interviews. The idea underpinning open-ended prompts is that these should potentially align better with the TA practice of human analysis. When executing an inductive TA, researchers do not approach the data with a pre-defined number of codes and themes to elicit. Instead, they remain open to exploring the data. With open-ended prompts, I took a step in this direction and toward the capacity for investigators to carry out a full inductive TA with LLMs. Moreover, the new prompts align with the terminology of TA. Ambiguous words I included in previous work such as "topics" or "items" (ibid.), have now been replaced with the keywords of TA. These include "initial codes", "themes", "sorting", "grouping" and "thematic analysis" and now form part of the text of prompts. This is significant as the new prompts I advance bring the TA with LLM closer to the common lexicon of social scientists and others who perform this type of analysis. I also conducted an evaluation of the themes created with the new prompts against themes created by human researchers on a second dataset, showing that the LLM themes are "good enough" and that they do seem to address most of the human results. Performing the evaluation on a different dataset also allowed me to confirm that the prompts can be used across different dataset domains. [51]

Further secondary observations I seek to highlight relate to the consideration that the prompt dependence on the results of the analysis cannot be eliminated entirely (at least not with the model used, which supports a 16K tokens context). I have illustrated how asking the LLM to produce different length descriptions in the initial coding (Prompt_1) results in the model producing a different number of codes (smaller set of codes when requesting a longer description). I also observed from the results that a higher number of initial codes (with a shorter description) does not necessarily lead to a better analysis. Using the ITS metric, derived by operating a codebook reduction after each interview, I have illustrated that the metric is smaller when using large descriptions for initial codes, and that this may be because the longer descriptions force the model to better identify codes which convey similar meanings. Moreover, when I performed the drawing of the thematic maps using the themes built from codes with longer descriptions, it appears easier to identify two distinct concepts, whereas this is less clear when using the themes created from initial codes with shorter descriptions. In any case, these observations will require further research performed either by myself or other scholars to assess their magnitude and relevance for the validity of the analysis. [52]

In this manuscript, I contribute to the embryonic body of research on doing TA with LLMs. I improved the procedure I described previously in DE PAOLI (2023a, 2023b) and DE PAOLI and MATHIS (2024). The method I propose is clearly distinct from deductive processes (thus far tried on initial coding only and not on themes generation) such as the one by XIAO et al. (2023) or from processes that only focus on the initial coding phase of a TA (see e.g. GAO et al., 2023). With the scientific community still at the beginning of exploring how to LLMs for TA, I am convinced that further research will be required to expand and develop the observations presented in this manuscript. The prompts I formulated can always be refined to increase the quality of the responses, and to further investigate the effect of the prompts (e.g. the length of description, variability in the response, etc.) in relation to the number of codes a model can produce. Moreover, I believe that the idea of drawing thematic maps based on the shared codes should be explored further. What I have advanced here serves as a demonstration of the concept, but further refined analysis may be possible, for example, measuring the relative frequency of codes to understand how they vary [53]

I also covered Phase 2 and 3 of TA (following BRAUN & CLARKE, 2006), namely the initial coding and the initial generation of themes. In previous work (see e.g. DE PAOLI, 2023b) I suggested that for performing Phase 4 an analyst could operate the model with a much higher temperature (i.e. the parameter which sets the degree of creativity for the model's response) for creating themes on three new iterations. Higher temperature should lead to more variation in the model response, and in this way, a researcher could observe the relative consistency of the themes and detect core themes. However, with the open-ended prompts it may just be enough for the analyst to carry out multiple iterations with the temperature at zero (no creativity), as I discussed earlier, to observe themes that are consistent across the iterations. For Phase 5, the renaming and summarising, the technique I offered in DE PAOLI (2023a, 2023b) can still apply. In those publications, my recommendation was to expose the model to each theme description and the underlying code names and to ask the model to put a name to this material, thus effectively renaming each theme. [54]

I believe that the direct implications of this endeavour for practice are manifold. In previous publications, I used a model accommodating 4097 tokens. The 16k model has a larger context and allows better processing of text, and at the time of writing we already have models with 128k tokens available, albeit at a much higher cost. In fact, the previous propositions I made (e.g. DE PAOLI, 2023b) are already surpassed in many ways. A key contribution for practice is the new prompts, which can be readily tested and deployed by researchers in their own analysis. Moreover, refining the language of prompts to better align with TA terminologies enables researchers to more effectively connect the method using LLMs with existing TA literature and practices. I suggested the possibility of deriving thematic maps from the properties of the themes (e.g. the shared codes), as a further step to align the approach with the traditional TA methodology. However, determining in what ways the shared codes should be used to create maps will require further work and further refining. Future testing could help to improve the prompts required for producing more refined analysis. Better alignment of the language of prompts with TA may also be explored further. Investigating the effects of the prompt (e.g. the length of descriptions) is another task which requires additional investigation, as is the problem of how large a codebook should be to deliver optimal generations of themes. [55]

Data Availability Statement

The data used are open access data available from FigShare and Zenodo. The authors of the datasets and the related repositories are cited and referenced in the manuscript. [56]

Acknowledgements

I would like to thank the two anonymous reviewers for giving valuable insights on how to make this a better paper. I also wish to thank my colleagues Alex LAW and Martin ZELLINGER for helping me revise the manuscript.

References

Braun, Virginia & Clarke, Victoria (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

Braun, Virginia; Clarke, Victoria & Weate, Paul (2016). Using thematic analysis in sport and exercise research. In Brett Smith & Andrew C. Sparkes (Eds.), *Routledge handbook of qualitative research in sport and exercise* (pp.191-205). New York, NY: Routledge.

Chandrasekaran, Dhivya & Mago, Vijay (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR), 54*(2), 1-37.

Chen, Banghao; Zhang, Zhaofeng; Langrené, Nicholas & Zhu, Shengxin (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint*, https://arxiv.org/abs/2310.14735 [Date of Access: June 29, 2024].

Chew, Robert; Bollenbacher, John; Wenger, Michael; Speer, Jessica & Kim, Annice (2023). LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint*, https://arxiv.org/abs/2306.14924 [Date of Access: December 11, 2023].

Clarke, Victoria & Braun, Virginia (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297-298.

Curty, Renata; Greer, Rebecca & White, Torin (2021). *Teaching undergraduates with quantitative data in the social sciences at University of California Santa Barbara: A local report*, https://doi.org/10.25436/E2101H [Date of Access: May 10, 2024].

Curty, Renata; Greer, Rebecca & White, Torin (2022). *Teaching undergraduates with quantitative data in the social sciences at University of California Santa Barbara* [Data set], https://doi.org/10.25349/D9402J [Date of Access: May 10, 2024].

Dai, Shih-Chieh; Xiong, Aiping & Ku, Lun-Wei (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint*, https://arxiv.org/abs/2310.15100 [Date of Access: January 6, 2024].

De Paoli, Stefano (2023a). Writing user personas with large language models: testing phase 6 of a thematic analysis of semi-structured interviews. *arXiv preprint*, https://arxiv.org/abs/2305.18099 [Date of Access: June 15, 2024].

De Paoli, Stefano (2023b). 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*, https://doi.org/10.1177/08944393231220483 [Date of Access: December 7, 2023].

De Paoli, Stefano & Mathis, Walther S. (2024). Reflections on inductive thematic saturation as a potential metric for measuring the validity of an inductive thematic analysis with LLMs. *arXiv preprint*, https://arxiv.org/abs/2401.03239 [Date of Access: June 15, 2024].

Drápal, Jakub; Westermann, Hannes & Savelka, Jaromir (2023). Using large language models to support thematic analysis in empirical legal studies. *arXiv preprint*, https://arxiv.org/abs/2310.18729 [Date of Access: January 16, 2023].

Gao, Jie; Guo, Yuchen, Lim; Gionnieve, Zhan; Tianqin, Zhang; Zheng, Li; Toby, Jia-Jun L. & Perrault, Simon T. (2023). CollabCoder: A GPT-powered workflow for collaborative qualitative analysis. *arXiv preprint*, https://arxiv.org/abs/2304.07366 [Date of Access: January 22, 2024].

Hamilton, Leah; Elliott, Desha; Quick, Aaron; Smith, Simone & Choplin, Victoria (2023). Exploring the use of ai in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods*, *22*, https://doi.org/10.1177/16094069231201504 [Date of Access: January 22, 2024].

Hanchard, Matthew & San Roman Pineda, Itzel (2023). *Fostering cultures of open qualitative research: Dataset 2—Interview transcripts*, https://doi.org/10.15131/shef.data.23567223.v2 [Date of Access: October 24, 2023].

Hoxtell, Annette (2019). Automation of qualitative content analysis: A proposal. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *20*(3), Art. 15, http://dx.doi.org/10.17169/fqs-20.3.3340 [Date of Access: January 16. 2024].

Huber, Patrick & Carenini, Giuseppe (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *arXiv preprint*, https://arxiv.org/abs/2204.04289 [Date of Access: December 12, 2023].

Lee, Vien V.; van der Lubbe, Stephanie C.; Goh, Leih H. & Valderas, Jose M. (2023). Harnessing ChatGPT for thematic analysis: Are we ready?. *arXiv preprint*, https://arxiv.org/abs/2310.14545 [Date of Access: December 12, 2023].

Madaan, Aman; Tandon, Niket; Gupta, Prakhar; Hallinan, Skyler; Gao, Luyu, Wiegreffe; Sarah, Alon, Uri; Dziri, Nouha; Prabhumoye, Shrimai; Yang, Yiming; Gupta, Shashank; Majumder, Bodhisattwa P.; Hermann, Katherine; Welleck, Sean; Yazdanbakhsh, Amir & Clark, Peter (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint*, https://arxiv.org/abs/2303.17651 [Date of Access: December 12, 2023].

Maguire, Moira & Delahunt, Brid (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education*, *9*(3), 3351-3359, https://ojs.aishe.org/index.php/aishe-j/article/view/335 [Date of Access: October 7, 2023].

Ofoeda, Joshua; Boateng, Richard & Effah, Jhon (2019). Application programming interface (API) research: A review of the past to inform the future. *International Journal of Enterprise Information Systems (IJEIS)*, *15*(3), 76-95.

Reimers, Niels & Gurevych, Iryna (2019). Sentence-bert: Sentence embeddings using Siamese Bert-networks. *arXiv preprint*, https://arxiv.org/abs/1908.10084 [Date of Access: May 16, 2024].

Saldaña, Johnny (2021). *The coding manual for qualitative researchers*. London: Sage.

Saunders, Benjamin; Sim, Julius; Kingstone, Tom; Baker, Shula; Waterfield, Jackie; Bartlam, Bernadette; Burroughs, Heather & Jinks, Clare (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, *52*, 1893-1907, https://doi.org/10.1007/s11135-017-0574-8 [Date of Access: October 10, 2023].

Schiavone, Will; Roberts, Chirstopher; Du, David; Sauro, Jeff & Lewis, Jim (2023). *Can ChatGPT replace UX researchers? An empirical analysis of comment classifications* [Online post], https://measuringu.com/classification-agreement-between-ux-researchers-and-chatgpt/ [Date of Access: June 12, 2023].

Serrano, Sofia; Brumbaugh, Zander & Smith, Noah A. (2023). Language models: A guide for the perplexed. *arXiv preprint*, https://arxiv.org/abs/2311.17301 [Date of Access: December 10, 2023].

Terry, Gareth; Hayfield, Nikki; Clarke, Victoria & Braun, Virginia (2017). Thematic analysis. In Willig Carla & Stainton Rogers Wendy (Eds.), *The Sage handbook of qualitative research in psychology* (2nd ed., pp.17-37). London: Sage.

Waldherr, Annie; Wehden, Lars O.; Stoltenberg, Daniela; Miltner, Peter; Ostner, Sophia & Pfetsch, Barbara (2019). Inductive codebook development for content analysis: Combining automated and manual methods. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *20*(1), Art. 19, https://doi.org/10.17169/fqs-20.1.3058 [Date of Access: January 16. 2024].

Wasserman, Stanley & Faust, Katherine (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Wiedemann, Gregor (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(2), Art. 23, https://doi.org/10.17169/fqs-14.2.1949 [Date of Access: January 16. 2024].

Wollin-Giering, Susanne; Hoffmann, Markus; Höfting, Jonasw & Ventzke, Carla (2024). Automatic transcription of English and German qualitative interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *25*(1), Art. 8, https://doi.org/10.17169/fqs-25.1.4129 [Date of Access: January 19, 2024].

Xiao, Ziang; Yuan, Xingdi; Liao, Vera Q.; Abdelghani, Rania & Oudeyer, Pierre Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In Association for Computing Machinery (Ed.), *Companion proceedings of the 28th International Conference on Intelligent User Interfaces* (pp.75-78). New York, NY: Association for Computing Machinery, https://dl.acm.org/doi/proceedings/10.1145/3581641 [Date of Access: June 30, 2023].

Yu, Zihan; He, Liang; Wu, Zhen; Dai, Xinyu & Chen, Jiajun (2023). Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint*, https://arxiv.org/abs/2310.04959 [Date of Access: December 10, 2023].

Author

Stefano DE PAOLI is professor of digital society at Abertay University in Dundee (Scotland). Stefano is interested in codesign, user research and qualitative methods. More recently he started working on large language models for data analysis.

Contact:

Professor Stefano De Paoli

Abertay University
Sociology Division
Bell Street, DD11HG, Dundee, UK

E-mail: s.depaoli@abertay.ac.uk
URL: https://www.abertay.ac.uk/staff-search/professor-stefano-de-paoli/

Citation

De Paoli, Stefano (2024). Further explorations on the use of large language models for thematic analysis. Open-ended prompts, better terminologies and thematic maps [55 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *25*(3), Art. 5, https://doi.org/10.17169/fqs-25.2.4196.