Understanding the Experiences of Smokers and Vapers with Preparatory Activities Suggested in a Digital Smoking Cessation Intervention

Antonio Florin Lupu
Supervisor: Willem-Paul Brinkman
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Antonio Florin Lupu
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Inald Lagendijk

Abstract

Digital smoking cessation tools increasingly use chatbots to recommend preparatory activities intended to support behavior change. However, little is known about how users actually experience these activities and whether large language models (LLMs) can support the qualitative analysis of such experiences. This study explores how smokers and vapers described their experiences with chatbot-suggested activities in an online intervention and investigates whether an LLM can assist with analyzing their responses.

We conducted a reflexive thematic analysis on 650 filtered user responses and examined whether a local version of LLaMA 3.3 (8B) could support the generation of common patterns and themes from qualitative data, and the categorization of such data into predefined themes. Our findings show that participants had mixed experiences. Some felt encouraged and supported by the activities, while others found them confusing, irrelevant to smoking cessation, or difficult to complete due to contextual or environmental barriers.

While the model was able to generate reasonable themes that overlapped with human interpretations, it was unable to reliably label data points using a predefined coding schema, achieving a Cohen's Kappa of just 0.003. These findings suggest that LLMs may be useful in the early, exploratory stages of qualitative analysis, but currently lack the accuracy needed for theme application in complex, nuanced data.

1 Introduction

Digital health tools have become increasingly popular in supporting behavior change [1], especially in areas like smoking and vaping cessation [2]. One approach that has gained popularity recently is the use of conversational agents to aid in the quitting process. These chatbots can suggest preparatory activities meant to help users get ready to quit, such as reflecting on personal motivations, setting a quit date, or identifying personal triggers.

Understanding how smokers and vapers experience chatbot-suggested activities gives us some insight into how these digital interventions can influence whether someone stays motivated or not. If an activity feels helpful or encouraging, it might make a participant more confident about quitting. On the other hand, if it feels confusing or hard to do, it could make them less likely to engage with the intervention. Looking at how participants describe their experience with these activities in their own words gives us a better picture of what works and what does not. This kind of insight can help improve how digital health tools are designed and how support is given throughout the quitting process.

The aim of this paper was to provide a secondary data analysis based on the experiences of participants from the study by Albers et al. [3] regarding their preparatory activities. Examples of such activities are finding motivational quotes for quitting smoking, focusing on past achievements and successes, or increasing physical activity. Additionally, we acknowledged that qualitative analysis is often time-consuming. We were therefore also interested in exploring whether large language models (LLMs) could support qualitative work in practice, as a secondary question.

The main research question that this paper aims to answer is: ***How do smokers and vapers experience preparatory activities suggested by a chatbot during an online smoking cessation intervention?***

Due to the complexity and broadness of this question, it has been divided into 2 sub-questions:

1. *What do smokers' and vapers' free-text responses reveal about their experiences with chatbot-recommended preparatory activities?*
2. *To what extent can LLMs support qualitative analysis?*

By answering these questions, the project contributes to two main areas: first, it provides a more detailed understanding of how people experience digital interventions, allowing for the improvement of digital smoking cessation programs [4]; second, it investigates the use of artificial intelligence (AI) to support qualitative research [5].

The structure of the paper is as follows. Chapter 2 presents related work, including previous work on digital interventions for smoking cessation and the use of large language models in qualitative work. Chapter 3 explains the methods that were followed to answer the research question with all of its sub-questions, and Chapter 4 presents the results of our research. Chapter 5 discusses the results in the context of the literature and possible future improvements to the study, as well as limitations, while Chapter 6 underlines the ethics and reproducibility of the employed methods.

2 Related Work

2.1 Digital Interventions for Smoking Cessation

Many digital interventions for smoking and vaping cessation have focused on their effectiveness in improving quit rates and reducing nicotine use. For example, meta-analyses by Civljak *et al.* [6] and Taylor *et al.* [2] highlight how web-based or app-based tools can support users through structured programs, reminders, and goal setting. These studies often evaluate their success through outcomes that can be measured, such as abstinence rates after a certain period. While these results are valuable for understanding what works at a broader level, they often miss the more nuanced insights into how users feel during the intervention or how they respond to individual components like chatbot messages or preparatory exercises.

2.2 Prior Studies on the Same or a Similar Intervention

Several previous studies have researched different aspects of the same chatbot-based smoking cessation program explored in this paper. Iftimescu [7] analyzed what users expect and need from their daily interaction with the chatbot, showing that personalization and emotional support affected user engagement. Ekinci [8] focused on how users respond to the chatbot's communication style, looking at both moments of appreciation and frustration, especially when interactions felt repetitive or impersonal. Naydenov [9] and Oudheusden [10] studied how users experience working with human coaches in the same smoking cessation context, focusing on feedback loops and the perceived value of human involvement. These projects offer insights into different parts of the intervention, but none have specifically analyzed how users experience preparatory activities themselves.

Albers *et al.* [11] conducted a mixed-methods study to identify user needs by combining feedback from interaction sessions with reflections on future usage scenarios. Their results emphasized the importance of perceived usefulness, timing, motivation to change, social support, and external barriers in shaping how people interact with digital tools. While their dataset included user feedback on preparatory activities, the focus of their analysis was on synthesizing general behavioral needs rather than examining how participants responded to specific activities in practice. In contrast, our study looks more narrowly at participants' lived experiences with each task. This offers a more granular perspective of how people felt about each activity they were asked to do during the intervention.

2.3 Using LLMs to Find Patterns and Generate Themes

Researchers have started exploring whether LLMs can help with qualitative analysis, particularly in identifying patterns in open-text data. Dai *et al.* [12] proposed an LLM-in-the-loop method, where GPT-3.5 was used to generate early codes from user interviews. The generated codes were found to be reasonable and sped up the early stages of analysis, though they still required review and editing by human researchers. Similarly, De Paoli [13] tested whether AI could handle the full process of inductive thematic analysis, from reading responses to summarizing key themes. They found that models like GPT could surface useful themes, especially when provided with clear instructions and structured prompts. However, one consistent limitation across these studies was the model's tendency to generate vague labels if the task was not clearly defined.

2.4 Using LLMs to Apply Themes

Chew *et al.* [14] tested GPT-3.5 on deductive coding tasks, where the model had to assign text responses to existing themes. They found that the model could perform well with examples that fit perfectly in a specific box, but struggled with ambiguity, especially when a response belonged to multiple themes. Xiao [15] also noted that LLMs often failed to distinguish between overlapping concepts unlesss the definitions were very specific. These findings suggest that while LLMs might reduce the manual workload of theme application, they cannot yet replace human coders with reasonable precision in their theme application, particularly in more subjective or emotionally complex data.

2.5 Prompt Structure for Qualitative LLM Analysis

De Paoli [13] explored how LLMs can be used for thematic analysis, which is a qualitative analysis method, and found that a prompt where the task is broken down into clear, step-by-step instructions leads to more accurate and useful outputs. In addition to breaking the task into stages, they also recommended using deterministic settings, such as setting the temperature of the model to zero, to make the results more reproducible. Khalid and Witmer [16] further suggest that prompts are more effective when they start by assigning a clear role to the model (e.g., "You are a qualitative researcher"), followed by explicit task instructions, relevant context, and finally a clear description of the expected output format.

3 Methodology

To answer the main research question and implicitly, all its subquestions, the dataset from Albers *et al.* [3] was analyzed using thematic analysis, a manual process outlined by Braun and Clarke [17]. They defined thematic analysis as a method for identifying, analyzing, and interpreting patterns within qualitative data. It allows researchers to make sense of complex, open-ended responses by organizing them into themes that reflect participants' perspectives. Smokers' and vapers' open-text responses have been coded to find recurring patterns or themes in the data. Later, the manual thematic analysis results were compared with those from AI-assisted thematic analysis methods. We will experiment with different LLMs to see if automatic thematic analysis yields accurate results regarding theme generation and application, respectively.

We used reflexive thematic analysis, following the approach by Braun and Clarke [17], to find patterns in how participants described their experiences. This method was a good fit because it is flexible and does not require us to follow a strict theory or framework. Compared to other qualitative analysis methods, such as content analysis, which mostly counts specific words or ideas, or grounded theory, which is used to build new theories, reflexive thematic analysis gave us the freedom to explore people's thoughts, feelings, and interpretations in more detail [18]. This was specifically useful in our case, since we were looking at open-text responses that included personal reflections, emotions, and different kinds of feedback.

3.1 Cleaning the dataset

The study of Albers et al. [3] gathered data from the interaction of smokers and vapers with a chatbot that was suggesting preparatory activities for smoking cessation. In the

questionnaire they deployed, users were asked: *How did you approach, do, or experience the preparatory activities suggested by the chatbot?*. Approximately 2300 answers were collected from this question. Due to the complexity of the question, namely asking users for their approach, the activity they specifically did, and their experience with it, all in one question, some filtering was needed. Most responders only answered one of the three parts of the above question; therefore, we manually filtered out all responses that did not contain information about the users' experience with the suggested preparatory activities. If a response only described what the user did, or how they approached the task, without mentioning how they felt about it or how it impacted them, we considered that response irrelevant for our study and filtered it out. After this filtering step, there were around 650 relevant responses that could be used during this thematic analysis.

3.2 Manual Thematic Analysis

Developing the Initial Coding Scheme
To begin the thematic analysis, we had to make an initial coding scheme. To avoid bias and to ensure some reproducibility, we used a peer approach to come up with an initial set of themes. To do this, we started with a random sample of 100 responses. These were analyzed to identify recurring ideas and early theme candidates. A peer researcher independently coded the same 100 responses and developed their own theme suggestions. We then compared and discussed our initial coding schemes, resolving differences and merging overlapping concepts. For example, we identified a theme called "Skepticism", but after this peer discussion, we concluded that there are two types of skepticism in the responses: participants either could not see how the suggested preparatory activity would help, which we decided to classify as "Perceived Usefulness", or they highlighted that they could not do the activity, which we decided fits best into the "Easy to do" spectrum. This process resulted in a final coding scheme, which aimed to improve consistency and trustworthiness in our analysis [17], [18].

Checking Reliability of the Coding Scheme
To check if the final coding scheme could be applied reliably and to introduce less bias, we conducted a second round of peer testing using a new random sample of 100 responses. The peer coder was trained using 30 responses different from this new random sample. The first 15 responses were done together with the peer coder, giving context and feedback along the way. The next 15 responses were done independently by the peer coder, after which we ended the training since the result was already satisfactory. After this training session, both coders assigned themes to the new random sample of datapoints (n=100), using the finalized schema. This allowed us to assess the inter-coder reliability and confirm that the themes were interpretable by others [19]. To measure inter-coder reliability, Cohen's kappa was calculated. The value of Cohen's kappa was 0.72, which shows a moderate agreement as per McHugh's interpretation [20]. We then continued labeling the dataset using this final coding scheme.

3.3 AI-Assisted Theme Generation

As part of our secondary research goal, we explored whether large language models (LLMs) could support qualitative analysis. We used LLaMA 3.3 (8B) as our LLM. The reason for this choice of model is that, in a recent systematic comparison of 12 bio-NLP tasks, the LLaMA-based models performed on par with the GPT series models [21]. These NLP tasks involved classification and summarization, which gave a strong indication that LLaMA 3.3 (8B) would be suitable for our qualitative analysis task. Furthermore, we used a local version of this LLM to ensure consistent behavior over time and to avoid the unpredictability of online models, which can be updated or retrained without notice, potentially affecting reproducibility.

We chose a two-step approach to better understand what LLMs can and cannot assist with in qualitative analysis. As a first focus, we tested whether the LLM could develop its own themes from the data, which shows how well it can identify patterns on its own. Then, we assessed how accurately the same LLM could apply a set of existing themes to the same dataset we labeled in the manual thematic analysis, which is a more precise and controlled task. This helped us to see whether LLaMA 3.3 (8B) was better at open exploration or at following specific instructions in a coding task.

Initial Prompt Design
The first attempts involved coming up with a relatively simple prompt. We provided the model with the research question and the entire dataset containing the participants' responses from the questionnaire and asked it to generate themes. The temperature settings were left at default (non-zero).

Initial Prompt - Theme Generation

You are a qualitative researcher. Analyze the following user responses using thematic analysis.

These responses are from smokers and vapers who interacted with a chatbot recommending preparatory activities for quitting. Your task is to extract themes based on what you read.

Research question: "How do smokers and vapers experience proposed preparatory activities by a chatbot during an online smoking cessation intervention?"

Please return:
1. A list of main themes.
2. A short explanation for each theme.

This prompt focused on being as general as possible, so it could also be used on different research questions or datasets while still being efficient and simple. However, with this prompt, the model often focused on *what activity was done* (e.g., walking, journaling, yoga) rather than *how the participant experienced it*, which was the focus of our research question. We suspect this occurred due to the complexity of the responses, where participants often first explained what

they did and how they approached the activity, and only later commented on how they felt about it.

Improved Prompt Design and Tuning

To address this, we refined the prompt to explicitly instruct the model to ignore the content of the activities and focus only on participants' evaluations of the experience. We also set the temperature of the LLM to zero to ensure deterministic and reproducible outputs.

With this prompt, and the temperature set to zero, LLaMA 3.3 (8B) provided relevant and consistent results. By clearly telling the model to focus only on how people felt about the activities, and not what they did, we got results that were more relevant to our research question. The output of the model based on this prompt will be shown and discussed more thoroughly in the Results section.

Improved Prompt - Theme Generation

You are a qualitative researcher. Analyze the following user responses using thematic analysis.

These responses are from smokers and vapers who interacted with a chatbot recommending preparatory activities for quitting.

IMPORTANT: Your task is to extract themes ONLY about how people experienced the activities (what they thought or felt about them). Do NOT make themes about what kind of activity was done (e.g., walking, yoga, journaling).

The research question is: "How do smokers and vapers experience proposed preparatory activities by a chatbot during an online smoking cessation intervention?"

Please return:
1. A list of main themes
2. A short explanation for each theme.

Each theme should reflect the participant's perception or experience of the activity, not what activity they actually did.

3.4 AI-Assisted Theme Application

Finding a Suitable Prompt

After testing whether LLaMA 3.3 could generate a meaningful coding scheme, the next step was to assess whether it could apply our manually developed coding scheme to unseen data. The goal was to evaluate the model's potential in automatically classifying participant responses into the correct themes.

To do this, we prompted LLaMA 3.3 (8B) with a full list of our finalized themes, including short explanations for each. The prompt used, which is shown below, clearly instructed the model to focus only on how users experienced the activities, not what activity they did. It also emphasized that the model should assign one or more applicable themes per response, and be strict in its assignments so as not to potentially assign all themes to the same response.

Prompt - Theme Application

You are a qualitative researcher helping with a thematic analysis project.

Your task is to read user responses and assign the most appropriate theme(s) from the list below. These responses come from smokers and vapers who completed chatbot-recommended activities to prepare for quitting.

IMPORTANT:
- Only focus on how the users experienced or evaluated the activities.
- Assign ONE or MORE themes to each response.
- Be strict. Only assign a theme if it clearly applies.

To this prompt, we added the full coding scheme from our manual analysis, which you can find in the Results section, along with explanations for each theme. The model was also given the full dataset of 650 participant responses to classify. These responses were provided unlabeled, so that the models' classifications could be later compared to the human classifications for evaluation.

4 Results

4.1 Manual Analysis Results

The following subsection presents the results of the manual thematic analysis. The analysis revealed nine distinct themes that describe how participants perceived the preparatory activities suggested by the chatbot. We grouped these themes into five broader categories, each representing a shared dimension of user experience.

We grouped the themes into larger categories to simplify the interpretation of the results and to reflect recurring patterns across related responses. While the individual themes highlight specific reactions of the participants, such as perceiving the activity as helpful, confusing, or personally mismatched, the groups help organize these into broader concepts. Some responses fell into more than one theme, because many participants expressed complex, sometimes contradictory, opinions about the activities.

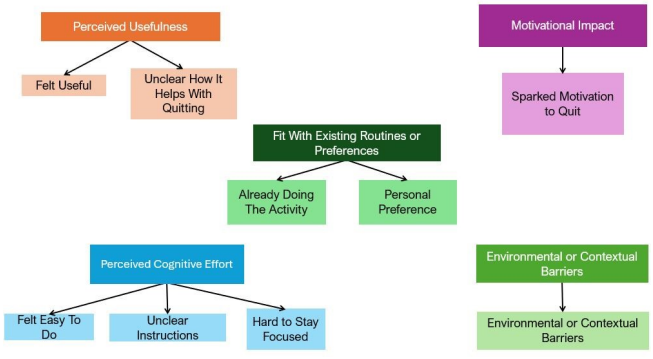

Figure 1: Thematic Map - Manual Thematic Analysis Results

The five groups, along with their corresponding themes, can be observed in Figure 1, and are described below with illustrative quotes taken directly from the dataset.

- **Perceived Usefulness** - This group reflects how helpful or relevant the activity felt for participants in the context of quitting smoking or vaping. Some participants felt that the activity had a clear and meaningful role and it was helpful in supporting their quitting journey. Others expressed confusion or doubt about how the activity was supposed to help. ***Felt Useful*** - Participants described the activity as something that was or not helpful in their quitting journey: *"I've made a rule 'I won't smoke, I'm doing it for my fiancé'. Honestly works pretty well." (P623).* ***Unclear How it Helps With Quitting*** - Participants sometimes engaged with the activity but struggled to see its relevance to smoking cessation: *"The thing is I am not sure that being physically active is going to help me quit smoking." (P137)*
- **Perceived Cognitive Effort** - This group captures how mentally demanding or clear participants found the activity. For some, the task felt easy and straightforward, while others found it confusing or mentally taxing. ***Felt Easy to Do*** - Participants described the activity as simple, intuitive, or low-effort, which encouraged engagement: *"I thought the task was easy and good to do, I got a new notebook [...]" (P131)* ***Unclear Instructions*** - Some participants did not understand what they were supposed to do or how to approach the activity: *"I didn't, I was a little confused on how to do this? [...]" (P33)* ***Hard to Stay Focused*** - Some participants reported getting distracted or having trouble concentrating on the task: "I was really excited to do this activity and then I got distracted by cleaning and forgot all about it." (P231)
- **Fit with Existing Routines or Preferences** - This group reflects whether the activity aligned with the participant's current habits or personal preferences. Activities that did not match with how people already structured their lives or what they typically enjoy were often dismissed. ***Already Doing the Activity*** - Some participants reported that they have already been doing the suggested task regularly, therefore diminishing its benefits or relevance: *"A lot of what the video talked about I already do. [...]" (P99)* ***Personal Preference*** - Others found that the activity did not suit them personally, either because it did not match their coping style or felt unappealing: *"I don't really like breathing exercises, I feel like another type of approach works better for me [...]" (P324)*
- **Motivational Impact** - This group describes whether the activity helped users feel more determined or inspired to quit smoking or vaping. While not all users reported feeling more motivated, those who did tended to speak positively about their overall experience. ***Sparked Motivation to Quit*** - These participants explained how the activity helped them feel more motivated, encouraged, or ready to make changes: *"I had written down my goals and began to exercise more which I found motivating." (P288)*
- **Environmental or Contextual Barriers** - This group captures how external factors, such as a person's environment or living conditions made the activity difficult or impossible to do. These barriers were not about motivation or interest, but rather situational factors like physical limitations, accessibility, personal context in life. Some participants reported that things like weather, anxiety, financial constraints, or other personal circumstances made it hard to follow through: *"I really wanted to start walking more and getting a bit fitter instead of taking taxis everywhere but I have slight agoraphobia so this didn't work for me." (P33)*

4.2 AI-Assisted Theme Generation Results

In our exploration of AI-assisted thematic analysis, we used LLaMA 3.3 (8B) to generate themes from participant responses. The model identified several relevant themes for answering our main research question. To evaluate the reproducibility of our manual results by the LLM, we decided to do a mapping between our manual themes and the AI themes. This mapping was done manually by comparing theme definitions and then pairing the ones that touched on the same topics. Table 1 presents our mapping. A more detailed view, including theme descriptions, can be observed in the Appendix A.

Table 1: Mapping of Our Themes to AI Themes

Human theme	AI theme
Perceived Usefulness	Activity Helpfulness
Perceived Cognitive Effort	Task Complexity
Fit with Existing Routines or Preferences	Redundancy in Recommendations
Motivational Impact	Motivational Impact
Environmental or Contextual Barriers	External Challenges

Overall, the AI-generated themes broadly aligned with the themes we developed through manual coding. While the model did not reproduce our categories with the same level of detail, the conceptual overlap suggests that LLaMA 3.3 (8B) can identify general patterns in the data that are consistent with human interpretations. This indicates some potential for using LLMs in the early stages of qualitative analysis, particularly for exploratory theme generation.

4.3 AI-Assisted Theme Application Results

To evaluate the model's performance in applying subthemes, we compared its output to our coding using Cohen's Kappa [20], a statistical measure of inter-rater reliability that we also used previously during the manual analysis. We found the agreement between LLaMA's assignments and our manual labels to be 0.003, which falls in the *none* category of agreement according to McHugh's interpretation [20]. This result indicates that the model could not reliably apply

predefined qualitative codes to new data, even with a detailed prompt and a well-defined schema.

5 Discussion

5.1 Discussing the Findings in the Context of the Literature

The study was guided by the main research question: *How do smokers and vapers experience preparatory activities suggested by a chatbot during an online smoking cessation intervention?* To address this, we analyzed open-ended user responses using both manual thematic analysis and AI-assisted methods. Two sub-questions were used to explore this broader question from multiple angles. Below, we revisit each one and discuss the main findings in the context of the literature.

1. What do smokers' and vapers' free-text responses reveal about their experiences with chatbot-recommended preparatory activities?
Through manual thematic analysis, we identified nine themes grouped under five overarching categories: perceived usefulness, perceived cognitive effort, motivational impact, fit with existing routines or preferences, and environmental or contextual barriers. These results show that users' experiences were diverse and often complex, as multiple themes could be assigned to the same response. Some participants described the activities as motivating or helpful, while others found them irrelevant, confusing, or mismatched with their daily lives or preferences. Several also reported environmental or personal limitations that prevented them from engaging fully with the activity: *"I really wanted to start walking more and getting a bit fitter instead of taking taxis everywhere but I have slight agoraphobia so this didn't work for me." (P33)*

Our manual analysis produced categories describing user experiences, such as *Perceived Usefulness*, *Motivational Impact*, and *Environmental Barriers*. These categories align with previous qualitative research on digital health interventions. For example, Ekinci [8] found that users responded strongly to the communication style and the emotional tone of a cessation robot, which influenced both engagement and perceived relevance, similar to our theme of Perceived Usefulness. Likewise, Iftimescu [7] reported a need for personalized activities, which matches our theme of *Personal Preference*.

The structure and breadth of our themes also align with recommendations by Braun and Clarke [17], who underlined the importance of thematic depth over frequency. Our decision not to quantify how often specific themes occurred was intentional, matching their view that qualitative research should put emphasis on meaning and nuance, rather than numerical quantification. Moreover, our labels overlap with the LLM-generated themes, which serves as an indication that our manual analysis captured reasonable patterns in the text. However, the manual analysis included subtler distinctions (e.g., *Unclear How It Helps With Quitting* and *Already Doing The Activity*) that were not captured as cleanly by the model, which reinforces the idea that human interpretation is valuable in subjective and complex contexts.

In summary, participants' experiences with preparatory activities were mixed and shaped by factors such as motivation, task relevance, emotional response, and external barriers. While some found the activities helpful or inspiring, others felt they were too generic, hard to relate to, or impractical in their personal context.

2. To what extent can LLMs support qualitative analysis?
We explored two types of tasks in which LLMs could support qualitative research: (1) generating themes from qualitative data, and (2) applying a predefined coding scheme to new data.

When using LLaMA 3.3 (8B) to generate themes from raw data, the results resembled our manually developed themes. Categories like *Activity Helpfulness*, *Task Complexity*, and *Motivational Impact* were directly comparable to those in our coding scheme. This supports findings by De Paoli [13] and Dai *et al.* [12], who reported that LLMs can reasonably generate meaningful themes when guided by a clear, focused prompt. However, our AI-generated themes were typically broader and more generic than those from our manual analysis. The model did not always separate overlapping concepts or detect nuanced distinctions that emerged through iterative peer discussion in our coding process. This matches findings by Xiao [15], who noted that LLMs tend to produce more abstract or vague theme labels unless provided with highly specific task framing and feedback.

The prompt design also impacted the quality of the results. Our initial attempts yielded activity-based themes rather than ones focused on experiences. Only after refining the prompt, so that it specified that the model should focus on how participants felt about activities, did the results become more relevant. This aligns with the prompt design principles outlined by Khalid and Witmer [16], who recommended clear role-setting, step-by-step task descriptions, and deterministic temperature settings for reproducibility in qualitative tasks.

In contrast, when LLaMA 3.3 (8B) was used to apply our predefined coding scheme to participant responses, it performed poorly. The model achieved a Cohen's Kappa [20] of only 0.003 when compared to our human-coded data, indicating almost no agreement. Given that the human coders reached moderate agreement using the same schema, we concluded that the LLM failed to apply the themes reliably. This was likely due to the complexity of the input responses and the limitations of prompt-only methods without fine-tuning or contextual support.

While our results showed that the LLaMA 3.3 (8B) model was not able to reliably apply a predefined coding scheme to qualitative responses, other studies have reported more promising outcomes when using LLMs for deductive thematic analysis. For instance, Chew *et al.* [14] and Dai *et*

al. [12] found that LLMs like GPT-3.5 could assist in the application of predefined codes, especially when the data was relatively unambiguous or when responses clearly matched a single theme. In those studies, the models achieved reasonable alignment with human coders, suggesting that LLMs may be useful in low-complexity settings or when coding schemes are rigid and well-defined.

Given the extremely low agreement score (0.003), we assumed that the discrepancy was due to the model's incorrect classifications rather than errors in our manual coding. This assumption is based on the fact that our coding scheme was developed iteratively, discussed with a peer, and tested through a reliability check, which showed moderate agreement between two human coders. It is likely that the model misinterpreted nuances in the responses or failed to apply the themes consistently, especially when users expressed multiple or contradictory sentiments. Our dataset consisted of nuanced, open-ended descriptions of experiences that often conveyed multiple or ambiguous feelings. Unlike brief snippets or structured survey responses, our inputs required a deep contextual interpretation, since the question itself from the questionnaire was very complex.

In summary, LLMs show promise for supporting qualitative analysis in early, exploratory stages, especially for generating broad themes. However, their ability to apply predefined codes to complex, subjective data remains limited. Without additional training, few-shot examples, or better contextual grounding, their performance cannot be comparable to human coders in tasks that require a fine-grained interpretation.

5.2 Limitations

This study has several limitations, including dataset scope, model constraints, and subjectivity in qualitative analysis. Firstly, the responses we analyzed came from a single dataset, collected as part of one specific chatbot-based intervention. Due to this, our findings might not reflect the experiences of people who did not complete the intervention or who do not usually engage with this type of digital support. The conclusions we draw are limited to this specific context.

Secondly, we relied exclusively on a single local LLM for both generating and applying themes. Conducting a thorough comparison of different models would have required extensive benchmarking and would have fallen outside the scope of this paper. Instead, we chose to focus on creating prompts that are efficient and robust. After all, new and more powerful LLMs continue to emerge. Therefore, by focusing on prompt design rather than model selection, our findings remain applicable even as better models appear.

Finally, while we took steps to increase the reliability and consistency of our manual thematic analysis, such as developing themes collaboratively and testing for coding reliability, qualitative research always involves a degree of subjectivity. Interpreting open-ended text involves choices about what patterns matter and how they should be described. Even with peer feedback and reliability checks, another researcher might have grouped or labeled responses differently. Our themes represent one interpretation of the data, but not the only possible one.

5.3 Future Work

To improve the accuracy of AI in qualitative research, future studies could explore several directions. One approach is to train or fine-tune language models on domain-specific qualitative data. Prior work by Gilardi and al. [22] and Gao and al. [23] has shown that fine-tuned models perform better on language processing tasks, which suggests that similar benefits could apply to thematic analysis.

Another strategy involves providing the model with a few correctly labeled examples (few-shot learning) to guide its outputs. This has been found to improve reliability in classification and summarization tasks in previous LLM studies [24][25], particularly when examples are well-aligned with the task and data format.

Lastly, future research could investigate how question phrasing and data structure affect AI performance. Open-ended questions that combine multiple tasks (e.g., "approach, do, or experience") may introduce ambiguity for both humans and machines. Simplifying or separating such prompts into focused items has been recommended by McCurdy and al. [26] for better thematic clarity and could be beneficial when working with LLMs.

6 Responsible Research

This study aimed to explore how smokers and vapers experienced preparatory activities suggested by a chatbot in the context of an online smoking cessation intervention. Because this topic touches on both health behavior and the use of artificial intelligence in qualitative research, we took a responsible and thoughtful approach in how the study was conducted and how findings were reported.

6.1 Research Data

This research was based on data collected by Albers *et al.* [3], which received ethical approval from the TU Delft University Human Research Ethics Committee. Subsequently, data trimming and falsification have been avoided by not making any modifications to the dataset. Any exclusion criteria for the data responses have been described and justified in the methodology. Furthermore, data fabrication has been prevented by also reporting negative results.

6.2 Ethical Use of Methodology

To analyze participants' open-text responses, we used reflexive thematic analysis [17], a method that allows for rich exploration of personal experiences. To increase the trustworthiness of the process, we included peer review and intercoder reliability checks. Two researchers created a shared coding scheme, and independently applied it. We calculated Cohen's kappa to measure agreement. This helped

reduce personal bias and ensured that our findings were consistent, understandable to others, and easily reproducible.

We also made a conscious decision not to quantify how often each theme occurred. Instead of counting responses, using frequencies or percentages, which could misrepresent the nuance of qualitative insights, we focused on understanding patterns of experience across diverse users. As Braun and Clarke [17] argue, qualitative work should prioritize meaning and depth over numeric summaries.

6.3 Handling Sensitive Topics Responsibly

Smoking cessation is a complex and often emotional process. Many responses reflected personal struggles, doubts, or strong feelings about the activities. We approached all responses with respect, recognizing that every participant brought their own background and perspective.

We would also like to clarify that our findings are not meant to speak for all smokers or vapers. The findings are grounded in a specific dataset [3], collected during a single intervention study, and should not be generalized without caution. Statements like *"I didn't find this helpful"* or *"This doesn't work for me"* were analyzed in context and interpreted as individual perspectives, not representative claims.

6.4 Responsible Use of AI

A secondary aim of the project was to explore whether large language models (LLMs) could support thematic analysis. These AI tools can save time by suggesting patterns or helping sort through large amounts of text. However, they can also change over time if used online, making it harder to reproduce the same results at a later time. To avoid this, we used a local version of LLaMA 3.3 (8B) through Ollama. This way, we could ensure the model remained consistent throughout the project, and our results could also be reproduced using the same model version.

Another concern is that the LLM we used might have been trained on thematic analysis results from similar research projects. Studies have shown that LLMs can memorize parts of their training data, including content from books, websites, datasets used in research, or research results [27]. This can make it unclear whether the model we used was reasoning about the actual text it received through the prompt or repeating something it had seen before as a result of other thematic analysis projects. We mitigated this by comparing the themes generated by AI with the themes from our manual analysis, evaluating whether the LLM generated reasonable themes based on the given data or not. However, we do acknowledge that the model may have been exposed to similar qualitative work in training, and that this is a potential source of bias.

6.5 Societal Implications

This research underlines how smokers and vapers experience digital health tools, offering possible insights for designers and health professionals. Understanding what users find helpful, confusing, or demotivating can lead to better-designed interventions that support real behavioral change.

It also shows that while AI can help with research, it should be used with caution. As more researchers and developers begin using AI for qualitative work, it is essential to keep humans in the loop to ensure quality, fairness, and ethical use.

7 Conclusions

This study set out to understand how smokers and vapers experienced preparatory activities recommended by a chatbot in a digital smoking cessation intervention, and whether large language models (LLMs) can support the analysis of those experiences. Two subquestions guided this work:
(1) *What do smokers' and vapers' free-text responses reveal about their experiences with chatbot-recommended preparatory activities?*
(2) *To what extent can LLMs support qualitative analysis?*

In response to the first question, the open-ended feedback showed that users had a wide range of reactions. Some felt encouraged, motivated, or better prepared to quit smoking, while others found the activities confusing, irrelevant, or difficult to carry out due to personal or situational barriers, such as phobias or preferences for specific activities over others. These varied responses suggest that digital interventions may need to account more carefully for individual context, preferences, and constraints to be effective.

To address the second question, we explored whether a local LLM (LLaMA 3.3 8B) could support qualitative analysis by generating themes and applying a coding scheme. The model was able to produce general themes that aligned with human-coded patterns, indicating some promise in exploratory analysis. However, when asked to apply a predefined coding scheme to new data, the model performed poorly, showing almost no agreement with human coders. This suggests that while LLMs may be helpful in the early stages of qualitative research, they are not yet reliable for more structured or interpretive tasks without additional training or human oversight.

In conclusion, this study highlights the varied ways users experience preparatory activities in a digital smoking cessation intervention, shaped by emotional, practical, and contextual factors. While LLMs showed potential for exploring broad patterns in open-text data, they remain limited in tasks requiring nuanced interpretation. Together, these findings suggest that both intervention design and qualitative analysis benefit from a balance of automation and human insight.

Acknowledgement

I would like to thank Willem-Paul Brinkman, my supervisor, for their guidance, feedback, and support throughout this project. I also appreciate the insights and foundational work of previous students whose research helped shape the

direction of this study. Lastly, I would like to thank my fellow research group members for their ongoing feedback and collaboration.

During the writing process, I used OpenAI's ChatGPT to improve the wording and flow of the text. I also used ChatGPT at the start of the project to help generate an initial list of relevant literature to explore, which informed the early direction of my research. The content, analysis, and interpretation presented in this paper are my own.

References

[1] L. Laranjo, A. G. Dunn, H. Tong, *et al.*, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018. DOI: 10.1093/jamia/ocy072.

[2] G. M. J. Taylor, M. N. Dalili, M. Civljak, A. Sheikh, and J. Car, "Internet-based interventions for smoking cessation," *Cochrane Database of Systematic Reviews*, vol. 9, 2017. DOI: 10.1002/14651858.CD007078.pub5.

[3] N. Albers, F. Melo, M. Neerincx, O. Kudina, and W.-P. Brinkman, *The impact of human feedback in a chatbot-based smoking cessation intervention: An empirical study into psychological, economic, and ethical factors - data and analysis code for the phd thesis chapter*, version 1, Dataset, 2025. DOI: 10.4121/1d9aa8eb-9e63-4bf5-98a3-f359dbc932a4.v1.

[4] L. Yardley, B. J. Spring, H. Riper, *et al.*, "Understanding and promoting effective engagement with digital behavior change interventions," *American Journal of Preventive Medicine*, vol. 51, no. 5, pp. 833–842, 2016. DOI: 10.1016/j.amepre.2016.06.015.

[5] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[6] M. Civljak, L. F. Stead, J. Hartmann-Boyce, A. Sheikh, and J. Car, "Internet-based interventions for smoking cessation," *Cochrane Database of Systematic Reviews*, vol. 7, 2013. DOI: 10.1002/14651858.CD007078.pub3.

[7] V. G. Iftimescu, *Towards effective smoking cessation: Understanding the needs of daily smokers from ehealth chatbot interactions*, Bachelor's thesis, Accessed: 2025-06-02, Delft, The Netherlands, 2024. [Online]. Available: http://resolver.tudelft.nl/uuid:af3e4e41-27fc-4f4b-9d46-e46707e59d46.

[8] A. Ekinci, *Eases and difficulties of talking to a virtual coach for quitting smoking and becoming more physically active: A mixed-methods analysis*, Bachelor's thesis, Accessed: 2025-06-02, Delft, The Netherlands, 2022. [Online]. Available: http://resolver.tudelft.nl/uuid:74b565fb-899f-44ef-ab73-4de286093df6.

[9] Y. Naydenov, *Users' attitude towards adding human feedback when preparing for quitting smoking/vaping with a virtual coach: A mixed-methods analysis*, Bachelor's thesis, Accessed: 2025-06-02, Delft, The Netherlands, 2024. [Online]. Available: http://resolver.tudelft.nl/uuid:4d0e49fc-e645-4e9c-9ae8-e04f21445f77.

[10] J. C. van Oudheusden, *Analyzing users' introductions to human coaches: Insights from ehealth applications introductions*, Bachelor's thesis, Accessed: 2025-06-02, Delft, The Netherlands, 2024. [Online]. Available: http://resolver.tudelft.nl/uuid:dac0c377-6e2b-4d0c-8273-7404568e58c3.

[11] N. Albers, M. A. Neerincx, K. M. Penfornis, and W.-P. Brinkman, "Users' needs for a digital smoking cessation application and how to address them: A mixed-methods study," *PeerJ*, vol. 10, e13824, 2022. DOI: 10.7717/peerj.13824. [Online]. Available: https://doi.org/10.7717/peerj.13824.

[12] S.-C. Dai, A. Xiong, and L.-W. Ku, "LLM-in-the-loop: Leveraging large language model for thematic analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, Dec. 2023, pp. 9993–10 001. DOI: 10.18653/v1/2023.findings-emnlp.669. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.669/.

[13] S. De Paoli, "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*, 2023. DOI: 10.1177/08944393231220483.

[14] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, and A. Kim, "Llm-assisted content analysis: Using large language models to support deductive coding," *arXiv preprint*, 2023, arXiv:2306.14924.

[15] e. a. Xiao, "Using large language models to support deductive coding," *arXiv preprint*, 2023, in DACOP; cited in related literature.

[16] M. T. Khalid and A.-P. Witmer, "Prompt engineering for large language model–assisted inductive thematic analysis," *arXiv preprint*, 2025. arXiv: 2503.22978 [cs.HC].

[17] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. DOI: 10.1191/1478088706qp063oa.

[18] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, "Thematic analysis: Striving to meet the trustworthiness criteria," *International Journal of Qualitative Methods*, vol. 16, no. 1, p. 1 609 406 917 733 847, 2017.

[19] J. L. Campbell, C. Quincy, J. Osserman, and O. K. Pedersen, "Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement," *Sociological Methods & Research*, vol. 42, no. 3, pp. 294–320, 2013. DOI: 10.1177/0049124113500475.

[20] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012. DOI: 10.11613/BM.2012.031. [Online]. Available: https://doi.org/10.11613/BM.2012.031.

[21] A. N. Omitted, “Benchmarking large language models for biomedical natural language processing,” *Nature Communications*, 2025, Systematic evaluation of GPT and LLaMA on 12 BioNLP benchmarks.

[22] F. Gilardi and et al., “Chatgpt outperforms traditional methods in political text classification,” *Nature Human Behaviour*, 2023.

[23] L. Gao and et al., “Llms for health-related sentiment and topic detection: A comparative study,” *Journal of Biomedical Informatics*, 2023.

[24] Z. Zhao and et al., “Calibrate before use: Improving few-shot performance of language models,” *ICML*, 2021.

[25] T. Brown and et al., “Language models are few-shot learners,” *NeurIPS*, 2020.

[26] M. McCurdy and et al., “Designing qualitative surveys: A literature review on question framing,” in *CHI*, 2020.

[27] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, and K. Lee, “Extracting training data from large language models,” *Proceedings of the USENIX Security Symposium*, 2023.

A Mapping from Our Thematic Analysis Themes to AI's Themes

Table 2: Mapping from Our Thematic Analysis Themes to AI's Themes

Human theme	**Human theme's description**	**AI theme**	**AI theme's description**
Perceived Usefulness	This group reflects how helpful or relevant the activity felt for participants in the context of quitting smoking or vaping. Some participants felt that the activity had a clear and meaningful role and it was helpful in supporting their quitting journey. Others expressed confusion or doubt about how the activity was supposed to help. Examples include: Felt Useful (helpful for quitting) and Unclear How it Helps With Quitting (struggling to see relevance).	Activity Helpfulness	Participants expressed differing views on whether the activities felt meaningful. Some clearly saw their value, while others doubted their relevance or effectiveness.
Perceived Cognitive Effort	This group captures how mentally demanding or clear participants found the activity. For some, the task felt easy and straightforward, while others found it confusing or mentally taxing. Includes Felt Easy to Do, Unclear Instructions, and Hard to Stay Focused.	Task Complexity	The perceived difficulty of completing the activity was a common topic. Some participants found the tasks simple and manageable, while others were confused or thought they were hard to do.
Fit with Existing Routines or Preferences	Reflects whether the activity aligned with participants' current habits or preferences. Some were already doing the activity or found it personally unsuitable. Includes Already Doing the Activity and Personal Preference.	Redundancy in Recommendations	Some participants indicated they were already engaging in suggested activities, leading them to view the recommendations as unnecessary or uninspiring.
Motivational Impact	Describes whether the activity helped users feel more determined or inspired to quit. Those positively motivated spoke about increased commitment and readiness to change.	Motivational Impact	The activities sparked increased commitment in some users. Participants described feeling more prepared, reflective, or hopeful about quitting.
Environmental or Contextual Barriers	Captures external factors such as environment or personal context that made the activity difficult or impossible to do. Examples include physical limitations, anxiety, or financial constraints.	External Challenges	Participants described external factors like mental health issues, work overload, or lack of physical access that interfered with their ability to carry out the tasks.