# CSE3000 Research Project

## Human Insight vs. Artificial Intelligence: A Thematic Analysis Comparing Manual and LLM Approaches to Understanding How Smokers Experience Preparatory Activities in a Digital Cessation Intervention

**Keshav Nair**
**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 21, 2025

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

**Abstract**

Smoking remains a leading cause of preventable death, making effective cessation support a global health priority. While conversational agents (chatbots) offer a scalable solution, their success depends on understanding the user's experience. This study addresses two interconnected challenges: first, understanding the subjective experience of smokers with preparatory activities proposed by a chatbot, and second, evaluating the efficacy of Large Language Models (LLMs) in analyzing this qualitative feedback.

This research employs a comparative design. A manual thematic analysis of smokers' written reflections first established a baseline coding scheme. This scheme was then compared against the outputs of three LLMs, which were tasked with both generating themes independently and applying the predefined manual scheme. The accuracy of the LLMs' application was measured against the human baseline using Cohen's Kappa.

The manual analysis revealed that smokers' experiences were predominantly positive, showing strong motivation and a sense that the activities helped reinforce their quitting goals. This was concurrently challenged by expressions of skepticism about the activities' effectiveness and mentions of personal barriers to quitting. The comparative analysis demonstrated that while LLMs could identify these broad positive and negative topics, they failed to capture more subtle, attitude-based concepts, such as a user's willingness to engage with an activity despite their personal doubts. Furthermore, the models' accuracy in applying a predefined coding scheme was substantially lower than the human baseline.

This work makes two primary contributions. For digital health, the findings show that cessation aids must be designed to personalize activities to address specific user barriers and skepticism. Methodologically, the study provides a clear verdict on the current role of LLMs in this context: while LLMs show potential as an exploratory aid in theme generation, they are not yet a viable tool for applying a predefined coding scheme, making human analytical oversight essential for ensuring the depth and validity of qualitative research.

# 1 Introduction

Smoking remains a leading cause of preventable death, causing over 8 million deaths worldwide per year [14]. The issue also affects non-smokers, with an estimated 1.2 million deaths annually being caused by second-hand smoke [5]. The scale of this global health crisis necessitates the development of accessible and effective support systems for individuals seeking to quit. In response, digital health tools like conversational agents (chatbots) have emerged as a scalable means to support smoking cessation programs [10].

While these digital interventions show promise, two interconnected challenges remain. First, to design truly effective interventions, we must understand how preparatory activities are experienced from the user's subjective perspective, a topic not fully explored in the existing literature. Second, as these interventions generate vast amounts of qualitative data from user reflections, a methodological challenge emerges. Manual thematic analysis, while effective for interpreting human experiences, is labor-intensive and challenging to apply to large datasets.

This paper addresses both of these areas. At the same time that it explores the user experience, the advent of Large Language Models (LLMs) offers a promising way to automate and enrich the analysis of user feedback, potentially overcoming the limitations of manual methods. Therefore, this study has two aims: first, to explore smokers' experiences with preparatory activities, and second, to conduct a comparative analysis of manual versus LLM-driven thematic analysis.

Accordingly, this paper is guided by the primary research question: How do smokers experience proposed preparatory activities by a conversational agent as part of an online smoking cessation intervention? Complementing this, a key secondary research question explores how LLMs can support thematic analysis in this context. To answer these questions, this study employs a thematic analysis of an existing dataset of smokers' written responses to a survey regarding the preparatory activities provided by the chatbot [1], while also evaluating the effectiveness and accuracy of different LLM models against a manual analysis baseline.

This paper will first outline the thematic analysis methodology, then present the results from both the manual and LLM-driven analyses, and conclude by discussing the implications of these findings for digital health and qualitative research.

# 2 Related Works

This section reviews the key bodies of literature relevant to this study, focusing on the use of conversational agents in smoking cessation programs and the strategies employed to improve their effectiveness.

## 2.1 Digital Interventions for Smoking Cessation

Conversational agents are increasingly being researched as a means to deliver accessible and scalable smoking cessation support. Existing research has approached the design and evaluation of these digital tools from several angles. A primary focus has been on gauging their overall effectiveness, with systematic reviews and meta-analyses working to establish how well these agents help users quit [9]. Beyond simply measuring outcomes, studies

[1] The pool of reflective questions can be accessed at: https://github.com/PerfectFit-project/virtual_coach_rl_persuasion_algorithm/blob/main/reflective_questions.csv.

have also analyzed the specific communication types used by chatbots, to understand how different conversational styles impact user outcomes [3].

In addition to the agent's conversational style, research has also delved into the specific content of the interventions. The efficacy of preparatory activities and the role of virtual assistants in enhancing cessation counseling have been examined as crucial components for success [20]. To ensure users benefit from these activities, another key area of research has been identifying and implementing strategies to enhance user engagement with the platform [1].

However, while these studies confirm the potential of digital interventions and have explored their effectiveness, communication styles, and engagement tactics, a gap remains. Much of the existing literature focuses on quantitative metrics of success rather than on the user's rich, subjective experience with the features they offer. Understanding how users perceive, interpret, and feel about the preparatory activities proposed by a chatbot is essential for designing the next generation of truly user-centered interventions. This study aims to fill this gap by focusing specifically on this subjective, qualitative dimension of the user journey.

# 3 Methodology

This study employs a thematic analysis approach to systematically interpret a comprehensive dataset comprising smokers' experiences with preparatory activities within an online smoking cessation intervention. The overarching aim of this research is to understand how smokers perceive these proposed activities. To achieve this, the analysis is guided by the following set of specific sub-questions. Firstly, what key themes emerge from smokers' descriptions of their experiences with preparatory activities in the intervention? Secondly, how do the themes generated by LLMs in thematic analysis compare to those identified through manual thematic analysis? Finally, how do LLMs perform in applying a predefined coding scheme?

## 3.1 Process

The research process unfolds in three main stages. Initially, a manual thematic analysis was conducted to establish a baseline understanding of smokers' experiences and to develop a human-validated coding scheme. Next, the thematic analysis is conducted once again, this time using LLMs, which perform two distinct tasks: independently generating their own coding schemes from the data, and applying the pre-defined manual coding scheme to the same data.

## 3.2 Data sanitation

Before analysis, the dataset was sanitized to ensure data quality and consistency. The process involved several steps. First, all text responses were converted to lowercase and cleared of any punctuation to ensure uniformity. The multiple reflection answer columns for each participant were then concatenated into a single text field, creating one combined response per participant. Finally, any rows that were empty, contained non-English or unintelligible text, or otherwise provided no useful information for the analysis were removed from the dataset.

### 3.3 Manual Thematic Analysis

Following data sanitation, a manual thematic analysis was performed on the dataset. This manual approach establishes a baseline for comparison with less subjective methods, specifically Large Language Models (LLMs). The rising popularity and efficiency of LLMs in automating tasks necessitate their evaluation in complex analytical endeavors such as thematic analysis. Given the inherent nuance of performing a thematic analysis, a manual approach was prioritized, serving as a foundation.

The analysis was conducted on a data corpus comprising approximately 2,000 responses from smokers. Due to the substantial size of this corpus, the analysis was managed by processing the data in batches of 100 responses. The thematic analysis followed the six-stage process articulated by Braun and Clarke [4].

The first stage involved familiarizing oneself with the data. As advised by Braun and Clarke, it is ideal to read through the entire data set at least once before you begin your coding, as your ideas and identification of possible patterns will be shaped as you read through it. Adhering to this stage, the dataset was read before any coding occurred.

Subsequently, another pass-through of the dataset was required to label each item. However, these labels must be specific, distinguishing them from themes. This ensured that discrete segments of raw data were meaningfully identified and tagged.

Thereafter, the process focused on identifying themes. Thus, initial codes were to be grouped into overarching themes. This grouping was facilitated visually, where codes were documented and then arranged to identify conceptual similarities and relationships.

The fourth crucial stage involved reviewing themes. During this phase, it was confirmed that themes were well-supported, internally coherent, and distinct from one another.

Penultimately, the clear definition of each theme was required. This stage focused on precisely articulating each selected theme's essence and contribution to the overall narrative. According to Braun and Clarke, a detailed analysis must be conducted and written for each theme. As well as identifying the 'story' that each theme tells, it is important to consider how it fits into the broader overall 'story' that you are speaking about your data, concerning your research question.

Finally, the sixth stage was producing the report. In this concluding phase, the complex story derived from the data was meticulously articulated. The primary objective was to present a compelling and well-supported analysis that directly addressed the research question, convincing the reader of its utility and validity.

However, before the analysis was performed on the entire data corpus, some cross-referencing was necessary to reduce the bias of the analysis. This is crucial as thematic analysis is based on subjective opinions and perceptions. Thus, an initial batch of 100 data items was analyzed, the first 100 items of the dataset, which were then sent to a peer for analysis. This allowed for a comparison of two views on how the same data items should be coded (detailed in 5.2).

After this crucial stage, the thematic analysis was continued with the finalized coding scheme. Upon applying this scheme to a 40% subset of the data, each processed item was assigned to its respective final theme(s). To visualize the process of the thematic structure for these processed items, a mind map was developed (see Figure 1), which illustrates how their initial labels correspond to the final themes they were categorized under.

To analyze the reproducibility and robustness of the developed thematic coding scheme, an inter-rater reliability analysis was conducted. After the development of the new coding scheme, the same peer was trained on its application. Both the primary researcher and the

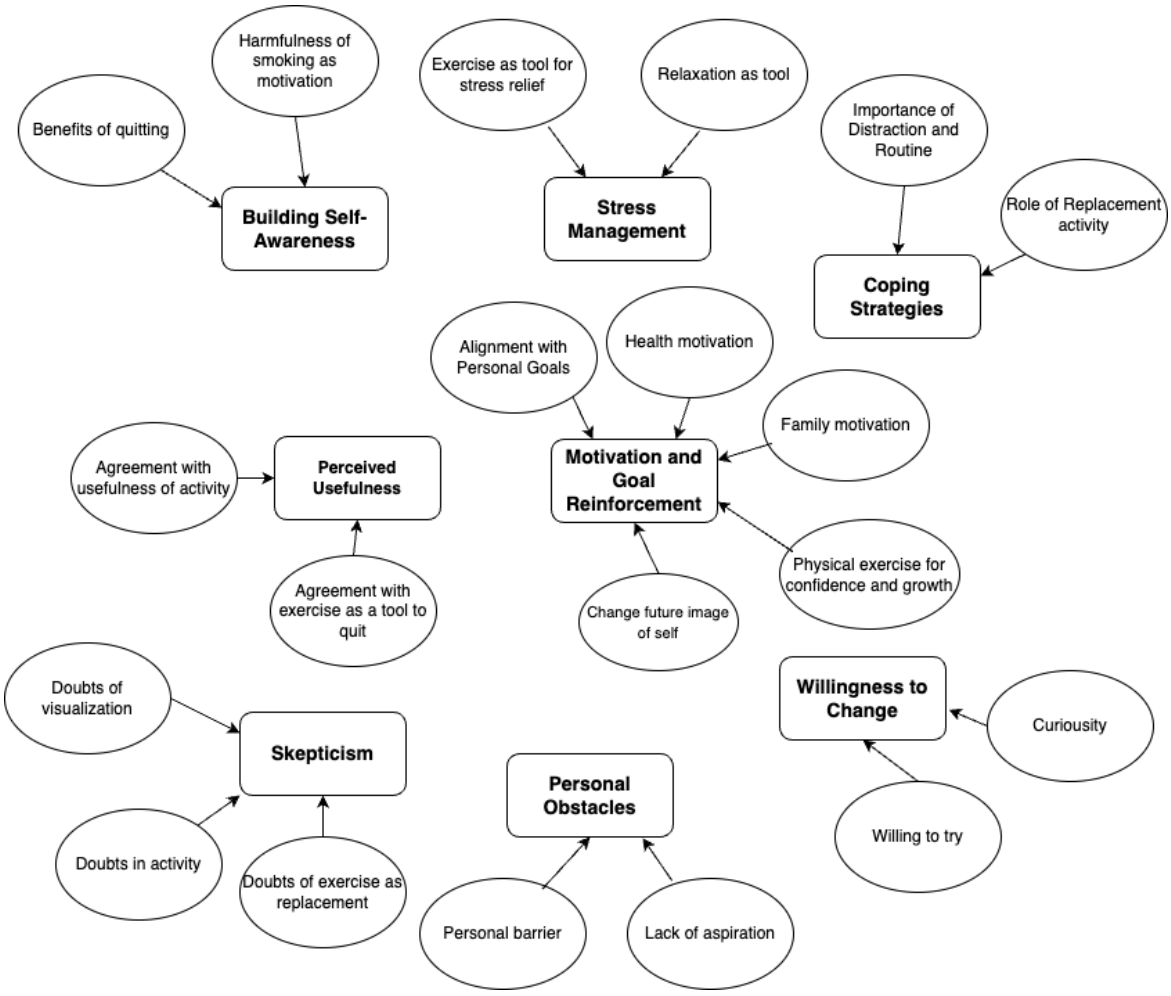


Figure 1: This figure visualizes the thematic analysis, showing the initial labels (circular nodes) and the themes (square nodes) they fall under.

peer independently applied the final coding scheme on a subset of 100 data items, distinct from the initial subset employed for the initial coding scheme development.

The level of agreement between the two coders was quantified using Cohen's Kappa coefficient ($\kappa$). This statistical measure was selected due to its ability to account for agreement occurring by chance. Hence, the tool extracted the agreement between coders on the application of the coding scheme, while providing a more rigorous assessment of reliability than a simple percentage agreement [7]. Given that the coding scheme allowed for the assignment of multiple themes to a single data item, Cohen's Kappa was calculated individually for each theme within the established scheme. This approach treats the application of each theme as a binary decision (present/absent) for each data item by each coder.

For each theme, a $2 \times 2$ contingency table was generated, cross-tabulating the coding decisions of the two raters (i.e., theme applied by Coder 1 and Coder 2; theme applied by Coder 1 but not Coder 2; theme not applied by Coder 1 but applied by Coder 2; theme not

applied by either coder). The observed proportion of agreement ($P_o$) and the proportion of agreement expected by chance ($P_e$) were derived from these tables. Cohen's Kappa was subsequently calculated using the formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

These computations were performed using Python with the `scikit-learn` library [16].

The resulting Kappa coefficients for each theme were interpreted according to the benchmarks proposed by Landis and Koch [12], where values $< 0.00$ indicate poor agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. This per-theme assessment of inter-rater reliability allowed for a detailed evaluation of the coding scheme's consistency and the clarity of individual theme definitions.

### 3.4 Thematic analysis using LLMs

Successfully applying LLMs on advanced tasks such as thematic analysis greatly relies on having a precise method of instruction, or prompt engineering. Merely copy-pasting the full dataset into the model and telling it to summarize will likely not produce the best outcome. Research by Wittmann [21] investigates whether guiding LLMs through a structured, multi-step process, as opposed to a single-prompt approach, enhances the depth, nuance, and accuracy of the generated insights. This highlights the necessity of a carefully chosen prompting strategy.

For this study, the Top-Down Structured Prompting (TDSP) technique, as detailed and evaluated by Wittmann, has been employed. This method was chosen because results strongly confirm that structured prompting techniques, particularly TDSP, significantly enhance the accuracy and depth of LLM-driven thematic analysis. Wittmann's study also found that, on average, 86% of findings identified by human researchers were at least partially identified in the outputs generated using structured prompting techniques like TDSP [21].

The TDSP process applied to our research involves four sequential stages. Firstly, the Generation of Overarching Themes and Sub-themes is performed by providing the LLM with the research question and the complete sanitized dataset to generate initial high-level themes and their corresponding sub-themes. Secondly, the Categorization of Data by Sub-themes utilizes the sub-themes generated in the first step to categorize the interview data, mapping sentences or segments of text to their relevant sub-theme. Accordingly, a specific prompt has been designed for execution within a script (See Appendix A for example prompts of each stage). Thirdly, the Extraction of Key Insights Within Sub-themes involves prompting the LLM with the associated data segments and the sub-theme's definition to extract specific, nuanced key insights for each sub-theme. Finally, the Synthesis of Comprehensive Evaluation sees the LLM synthesizing all the gathered information, themes, sub-themes, key insights, and supporting quotes into a comprehensive evaluation for each theme. The choice of TDSP is further justified by its advantages over other prompting techniques evaluated by Wittmann. TDSP capitalizes on the LLM's proficiency in pattern recognition and concise summarization of large datasets. Additionally, it provides clear, step-by-step guidance, preventing the LLM from being overwhelmed by the data. It also ensures context optimization by providing focused information at each stage, leading to the most consistent performance across all datasets in the study [21].

To address the research sub-questions comparing LLM and manual analyses of smokers' experiences, and to assess LLM accuracy, the LLM-driven thematic analysis using TDSP

has been conducted in two distinct ways. Firstly, the LLM has independently analyzed the dataset of smokers' reflections on preparatory activities to generate its coding scheme (themes and sub-themes). Secondly, the LLM has been provided with the coding scheme derived from the manual thematic analysis (detailed in Section 3.2) and tasked with applying these predefined themes to the dataset, allowing for a direct comparison of coding accuracy.

To elaborate on these methods, the LLM-driven analysis was conducted in two distinct ways using LM Studio, a local LLM platform. A local LLM was chosen to ensure the consistency and reproducibility of the results, as detailed further in Section 4.3. Three different models were loaded for this comparative analysis: DeepSeek R1 distill Qwen 7B, Qwen3 30B A3B, and Mistral Nemo Instruct 2407.

In the first approach, these models were used to independently generate a set of overarching themes and sub-themes from the dataset, which were then compared against the themes from the manual analysis. In the second approach, the models were tasked with applying the predefined coding scheme derived from the manual thematic analysis to a subset of 100 items. The performance in this task was then evaluated against the manual application on the same subset using Cohen's Kappa to rigorously analyze the degree of similarity.

# 4 Responsible Research

Applying artificial intelligence to sensitive health topics demands a rigorous commitment to ethical and methodological integrity. This study was therefore guided by core principles of data privacy and analytical validity, ensuring a responsible approach was taken throughout the research process.

## 4.1 Data Integrity

The research was based on a dataset collected by Albers et al. [1], which received ethical approval from the TU Delft University Human Research Ethics Committee. To ensure consistency for analysis, the data underwent a sanitation process that did not alter the substance or meaning of the responses. All data processing steps and criteria for data exclusion are described and justified in the methodology section (Section 3.2).

## 4.2 Methodological Integrity

The study employed a thematic analysis to analyze the open-text responses. To enhance the credibility and trustworthiness of the findings and to limit the biases of any one analytical process, the research incorporated a manual peer analysis. This two-coder process, in which a peer reviewed and applied the coding scheme, served two critical purposes: it acted as a validation step for the primary researcher's interpretations and helped identify and correct for idiosyncratic biases. This process establishes a verifiable benchmark for the thematic analysis, contributing to the overall reproducibility and credibility of the outcomes.

## 4.3 Responsible AI

While local LLMs are often chosen to ensure the privacy of sensitive data, this justification does not apply to the present study as the dataset is publicly available [1]. Instead, the decision to employ local LLMs was grounded in the principles of methodological consistency and reproducibility. Cloud-based commercial LLMs can be updated by their providers

without notice, meaning the analytical instrument could change during the research process. Using a static, local model ensures that the analysis is reproducible and that the LLM does not train on the input data, which would otherwise lead to inconsistent outcomes.

Despite this methodological control, the use of LLMs still presents significant ethical challenges, most notably the inherent biases within the models themselves [2]. While a local LLM provides a consistent analytical tool, it can still reproduce biases from its vast original training data. In the context of this study, this bias could manifest in misinterpreting culturally specific expressions of stress, failing to recognize slang related to smoking habits, or misclassifying the tone of a user who is expressing skepticism through indirect language, thereby skewing the thematic analysis towards more literal interpretations.

# 5 Results

## 5.1 Manual Thematic Analysis

The manual thematic analysis of smokers' reflections revealed eight key themes that capture their experiences with the preparatory activities. These themes are detailed below with illustrative quotes from the dataset.

### 5.1.1 Motivation and Goal Reinforcement

The most predominant theme identified in the analysis was Motivation and Goal Reinforcement, which was central to the user experience. This theme captured how well the preparatory activities aligned with participants' personal reasons for quitting. For the vast majority of participants, this connection was strong and positive. One user, for example, articulated this clearly by stating an activity "match my decision" because "i want to became *[sic]* more physically active and i want to quit smoking and have a healthier life".

While there were rare instances where users felt an activity did not align with their personal motivations, this was distinct from general skepticism. Such cases were uncommon, as expressions of doubt were more frequently related to the perceived usefulness of an activity itself (or general skepticism) rather than its connection to a user's underlying goals.

### 5.1.2 Perceived Usefulness

Closely related was the theme of Perceived Usefulness, encompassing users' direct judgments on an activity's effectiveness, while also being very prevalent. This theme contained a range of feedback, from positive to negative perspectives. An example of this being "i think this activity will help me to not think about smoking and will get my mind off of the fact that i want to smoke". While a negative example of this would be, "it matches my decision but i dont feel like itll help because smoking is a physical addiction".

### 5.1.3 Coping Strategies

On a more practical level, Coping Strategies emerged as a theme, reflecting the in-the-moment utility that users found in the activities. It focused on how the chatbots suggestions helped them manage immediate challenges, particularly by serving as a necessary distraction from the urge to smoke. For example, a user stated, "it helps me to distract the urge to smoke".

#### 5.1.4 Willingness to Change

The analysis also captured a key user mindset in the theme of Willingness to Change. This reflected a strong, underlying commitment to the quitting process that often overrode specific doubts about an activity, with some participants expressing they were "intrigued to see if it works" even if they were not fully convinced. Another user stated, "im willing to give anything a go that may help me quit", expressing zero belief in the activity itself.

#### 5.1.5 Skepticism

Providing a counterbalance to the positive engagement, Skepticism collected the natural doubts and mixed attitudes users expressed. This feedback ranged from specific concerns, such as one user who was "not sure increasing exercise alone will help me to quit", to more general disbelief in certain methods, "i dont belive *[sic]* in that kind of things so i dont think it helps".

#### 5.1.6 Building Self-Awareness

Beyond practical strategies, users also valued gaining new insights into their behaviors, a concept captured in the theme of Building Self-Awareness. The activities were often seen as tools for self-monitoring, with one user noting, "i think understanding when and how much i smoke will really help me understand and reduce the amount which will lead to quit", as well as becoming more aware of the consequences of their actions.

#### 5.1.7 Stress Management

Stress Management further highlighted the practical focus of users, capturing the crucial connection they made between stress and their smoking habits. This theme is distinct from Coping Strategies, as it focuses primarily on stress as a trigger, which was prevalent among smokers. Participants frequently discussed activities in terms of their potential to manage or avoid stress, which they saw as a primary trigger for smoking. This is highlighted in the following response, "i totally agree because i often smoke cogarettes when im tense so i guess that learning how to relax would be really helpful".

#### 5.1.8 Personal Obstacles

Finally, the theme of Personal Obstacles gave voice to the internal, self-identified barriers users faced. Distinct from general skepticism, these were specific challenges that users felt hindered their progress, such as one participant who admitted they were "trying to become more physically active but lack aspiration", or another participant stating "im not sure im motivated enough".

### 5.2 Reliability of Manual Coding Scheme

This section presents the results of the inter-rater reliability for the application of the aforementioned coding scheme. This was assessed using Cohen's Kappa, with the detailed scores presented in Table 1.

Quantified by Cohen's Kappa, an average Kappa value of 0.730 across themes reflects overall substantial agreement between the coders. Most individual themes achieved substantial or almost perfect agreement, with 'Coping Strategies' ($\kappa$=0.896), 'Stress Management'

Table 1: Inter-Rater Reliability Scores (Cohen's Kappa) for Thematic Coding

| Theme Name | Cohen's Kappa ($\kappa$) |
|---|---|
| Building Self-Awareness | 0.446 |
| Coping Strategies | 0.896 |
| Motivation and Goal Reinforcement | 0.675 |
| Perceived Usefulness | 0.819 |
| Personal Obstacles | 0.661 |
| Skepticism | 0.736 |
| Stress Management | 0.852 |
| Willingness to Change | 0.755 |
| **Average Kappa Score** | **0.730** |

($\kappa$=0.852), and 'Perceived Usefulness' ($\kappa$=0.819) reflecting almost perfect agreement. Although the "Building Self-Awareness" theme generated a modest agreement of $\kappa$=0.446, which could reflect that its definition or application may have greater inherent subjectivity. Overall, high levels of inter-rater reliability across most themes increase confidence in the trustworthiness and reproducibility of the thematic patterns uncovered in the dataset.

## 5.3 Thematic Analysis using LLMs

### 5.3.1 Coding Scheme Generation

The coding scheme generation has been performed following the initial stage of the TDSP process. This led to coding schemes being generated for each LLM used: the DeepSeek R1 Distill Qwen 7B, Qwen3 30B A3B, and Mistral Nemo Instruct 2047 models. The themes generated have been compared to the baseline coding scheme (derived from the manual thematic analysis), as seen in Table 2. The DeepSeek R1 model did not produce meaningful themes; the themes were merely keywords from the prompts. Therefore, the model was excluded from further analysis.

### 5.3.2 Coding Scheme Application by LLMs

Following the generation of coding schemes by Large Language Models (LLMs), their ability to apply the predefined manual coding scheme was assessed. The goal of this evaluation was to determine how effectively the models could replicate human coding of qualitative data. The effectiveness of the three models was measured using Cohen's Kappa measurement, which was previously employed for inter-coder reliability (see Section 3.1). The average Kappa scores are presented in Table 3, with the baseline being the score from Section 5.2.

A direct comparison of the generated themes against the manual coding scheme, as presented in Table 2, reveals several key observations. Both Qwen3 and Mistral Nemo generated themes that corresponded to the manual themes of Motivation, Skepticism, and Building Self-Awareness. However, differences in specificity and grouping were also observed. For example, both models grouped the manual theme of 'Personal Obstacles' within their themes related to skepticism, and Qwen3 mapped 'Stress Management' to the broader category of 'Behavioral Strategies and Self-Discipline'. Furthermore, some manual themes were not identified by the LLMs at all; neither model generated a theme equivalent to 'Willingness

Table 2: LLM Coding Scheme Comparison

| **Manual Analysis** | **Qwen3** | **Mistral Nemo** |
|---|---|---|
| Motivation and Goal Reinforcement | Motivational Techniques and Future Identity | Motivation and Commitment |
| Coping Strategies | Distraction Through Activities | — |
| Willingness to Change | — | — |
| Skepticism | Skepticism and Individual Differences | Concerns and Skepticism |
| Building Self-Awareness | Awareness and Monitoring | Monitoring and Awareness |
| Stress Management | Behavioral Strategies and Self-Discipline | — |
| Personal Obstacles | Skepticism and Individual Differences | Concerns and Skepticism |
| Perceived Usefulness | — | Activity Engagement |

*Empty cells indicate that the model did not generate an equivalent or similar theme.*

to Change', and Mistral Nemo did not produce distinct themes for 'Coping Strategies' or 'Stress Management'.

Table 3: Average Cohen's $\kappa$ Scores for LLM Application of the Manual Coding Scheme

| **Model Name** | **Average Cohen's $\kappa$** |
|---|---|
| Qwen3 30B A3B | 0.135 |
| DeepSeek R1 distill Qwen 7B | 0.054 |
| Mistral Nemo Instruct 2407 | 0.326 |
| *Manual Inter-Rater (Human Baseline from Table 1)* | *0.730* |

# 6 Discussion

## 6.1 Insights from the Manual Analysis of Smokers' Experiences

The user's engagement appears to be primarily driven by a combination of goal alignment and perceived value. The most predominant theme, Motivation and Goal Reinforcement, shows that users responded positively when they could connect a preparatory activity to their health objectives. This finding empirically supports the principle identified by Strecher, who noted that message personalization increases perceived relevance, which in turn boosts user engagement [19]. This motivation was closely linked to Perceived Usefulness, a prevalent theme where users assessed whether an activity would be effective. The evaluation was pragmatic: some users believed an activity could serve as a mental diversion from smoking, while others concluded it would be insufficient to combat a physical addiction. Together, these themes suggest users were not just passive participants but were actively calculating an activity's potential return on investment for their quitting effort.

The analysis also indicates a user base seeking actionable tools for both immediate and long-term challenges. The theme of Coping Strategies reflects a demand for in-the-moment

solutions to manage the direct urge to smoke by providing a distraction. Stress Management is a more specific manifestation of this, where users identified stress as a key trigger and sought activities, such as relaxation, to counteract it directly. Beyond immediate solutions, the theme of Building Self-Awareness indicates that users also valued activities as tools for self-monitoring, which could help them understand their smoking patterns and form a long-term strategy for reduction.

Despite this positive engagement, two distinct types of barriers were identified. Skepticism represents a challenge to the intervention's methodology, with users doubting if a specific activity was sufficient or questioning the validity of the psychological approach. This finding is consistent with research by Wakeman, who identified skepticism as a common hurdle in online cessation technologies and called for more evidence-based interventions to build user trust. Fundamentally different from this external doubt is the theme of Personal Obstacles, which captures internal, self-identified barriers. Here, users did not question the activity but rather their own ability to follow through. This directly reflects the challenges identified by Asbring, who found that stress and difficult life circumstances significantly influence smoking habits and that a perceived lack of willpower is a key barrier to quitting [22].

The theme of Willingness to Change appears to resolve the conflict between motivation and these barriers. The analysis shows that a strong underlying commitment to quitting could override specific doubts about an activity. Users expressed a readiness to try any suggested method, even those in which they had little or no belief, simply because it might help them achieve their ultimate goal. This indicates that for a significant portion of users, the overall desire to change was a more powerful determinant of action than their confidence in any single part of the process. The importance of this internal drive resonates with Asbring's research, which highlights that overcoming a lack of willpower is crucial for success in smoking cessation [22].

## 6.2 Coding Scheme Generation by LLMs

The LLM-generated coding schemes successfully identified a significant portion of the manual themes, with the models collectively covering 7/8 (87.5%) of the themes from the manual analysis. This aligns with Wittmann's finding that structured prompting can identify, on average, 86% of human-identified findings [21]. However, this high-level success masks critical gaps in nuance. For instance, neither the Qwen3 nor the Mistral Nemo model identified the 'Willingness to Change' theme (see Table 2), a concept requiring interpretation of a user's underlying attitude, not just the explicit content of their response.

Another key difference was the LLMs' tendency to conflate related but distinct concepts. Both models grouped the manual theme of 'Personal Obstacles' within their broader themes related to skepticism, while the Qwen3 model mapped the specific idea of 'Stress Management' to the very general theme of 'Behavioral Strategies and Self-Discipline'. This suggests that while the LLMs are capable of identifying high-level, explicit topics like 'Motivation' and 'Skepticism,' they are less adept at capturing the subtle, interpretive themes that are foundational to a deep qualitative understanding, underscoring the continued importance of human oversight in the analytical process.

Despite these shortcomings, the theme generation process offered an unexpected benefit: the discovery of a theme that the manual analysis had overlooked. The theme of 'Activity Engagement', identified by the Mistral Nemo model, captured a distinct aspect of the user experience that had been implicitly grouped within the theme of 'Perceived Usefulness'

during manual analysis. This suggests that LLMs can serve as a valuable tool for conceptual triangulation; even if they fail to replicate a human's exact coding scheme, their ability to rapidly propose alternative structures can help researchers notice different facets of the data and refine their analytical thinking. This reinforces the conclusions from Prescott et al. [17], who conclude that while human coders demonstrate a greater ability to identify nuanced and interpretative themes, the observed consistency in themes generated by AI indicates that a move towards hybrid approaches is necessary. Using these technologies in collaboration with human coders can improve the efficiency of qualitative research, a conclusion also supported by Ibrahim, who found that LLMs can effectively augment a researcher's perspective on themes within the data [11].

### 6.3 Coding Scheme Application by LLMs

The results shown in Table 3 indicate varying degrees of success among the LLMs in applying the coding scheme, with Mistral Nemo Instruct 2407 achieving the highest average Kappa score among the tested models. However, when compared to the human baseline ($\kappa = 0.730$), the LLMs generally demonstrate that they are severely lacking at applying nuanced coding schemes to qualitative data. Even when provided with examples and clear definitions for each theme, the models struggled to consistently replicate human judgment. This performance gap is consistent with recent findings [15] that show a distinct performance dichotomy: human coders excel at labeling complex, abstract, or nuanced sentences, whereas LLMs tend to be more reliable with simpler, more literal statements. This helps explain the LLMs' struggles in this study, which involved highly interpretive themes that require a level of abstraction the models have not yet mastered. As noted by Castellanos et al., human coders often recognize nuanced themes related to context, emotions, and cultural subtleties that LLMs may miss [6]. Furthermore, the interpretative scope of LLMs can be a factor, as LLMs may not inherently understand when to draw on implicit or explicit meanings within a text; they require specific instructions to control the scope of interpretation [13]. Thus, the limitations of AI have been highlighted in related literature, and these results follow that path.

### 6.4 Limitations

This study has several limitations that should be acknowledged when interpreting the findings. Firstly, the manual thematic analysis, which established the human baseline, was conducted on a representative subset of the data rather than the entire corpus of approximately 2,000 responses. While this approach was pragmatic, it means the generalizability of the findings relies on this subset, and the complete thematic saturation of the full dataset was not fully confirmed, a point addressed in the future works of this paper.

Secondly, the methodological decision to use locally hosted LLMs was based on the need for consistent results. However, this choice involves a trade-off, as local models may not match the performance and flexibility of larger, cloud-hosted counterparts, which could potentially impact the quality of the AI-generated themes and analysis. For instance, research by Dorca Josa et al. [8] indicates that many people felt that ChatGPT, or similar models, outperformed the local LLM in terms of language accuracy, flexibility, and overall capabilities.

Another limitation concerns the nature of the participant sample. The data was sourced from smokers who voluntarily participated in the digital cessation intervention. This introduces a potential bias, as these individuals may possess a higher intrinsic motivation to

quit than smokers not actively seeking help. Consequently, the strong presence of themes such as 'Motivation and Goal Reinforcement' and 'Willingness to Change' might be particularly characteristic of this group, and the findings may not fully extend to a broader, less-motivated population of smokers.

Additionally, the study's conclusions are also specific to the preparatory activities within this particular chatbot intervention. The identified themes may not be transferable to the user experience of other smoking cessation methods, such as nicotine replacement therapy, group counseling, or even digital interventions that utilize a different set of activities or a different conversational style.

Finally, the study is bound by the inherent limitations of current artificial intelligence (AI) technology in qualitative research. LLMs excel at recognizing broad patterns but can struggle to grasp the complex, contextual, or implied meanings that a human researcher can interpret. Their performance is highly dependent on prompt specificity, and they can reproduce biases from their original training data, making critical human oversight essential for ensuring the validity of the results. These limitations, however, also open up several clear avenues for future research.

### 6.5 Future Work

Future research could enhance the methodology used in this study by exploring more sophisticated human-AI collaboration models, such as active learning. This iterative process, where a model queries a human researcher to annotate only the most informative or ambiguous data instances, is a well-established strategy for reducing annotation costs while maintaining model performance. In an active learning framework, the LLM could be used to identify and present the data points it finds most difficult to categorize, allowing the human analyst to focus their effort only where needed, potentially creating a more accurate and efficient thematic analysis process [18]. To further increase the robustness of the baseline findings, the complete dataset could be manually analyzed. While the current analysis of a 40% subset was substantial, analyzing the full corpus would definitively confirm whether thematic saturation had been reached and further validate the manually derived coding scheme.

The scope of this research could also be expanded in several ways. Including a wider array of local LLMs in the comparison could also help identify an optimal model or architecture for this specific task. Finally, future work could move beyond thematic analysis to explore alternative machine learning approaches for pattern recognition, such as topic modeling or sentiment analysis. These techniques could be applied to the dataset to reveal different types of insights into the smokers' experiences that a thematic structure may not capture.

## 7 Conclusions

This study sought to understand how smokers experience preparatory activities within a chatbot-guided smoking cessation intervention, while also evaluating the role of Large Language Models (LLMs) in conducting the thematic analysis. The findings for each research question are as follows.

In response to the main research question asking how smokers experience these activities, the analysis revealed a complex user journey. The key themes that emerged showed that the experience was predominantly positive, with smokers expressing strong motivation and connecting the activities to their personal quitting goals. However, this was balanced by

authentic expressions of skepticism and personal barriers. A central finding was that a user's underlying willingness to quit often persisted even when they harbored reservations about a specific activity.

Regarding the comparison of themes generated by LLMs versus manual analysis, the findings show a significant gap in nuance. While the LLMs could identify broad topics, they consistently failed to generate themes related to subtle, attitude-based concepts, most notably, a user's underlying willingness to change despite their doubts.

Finally, concerning the LLMs' performance in applying a predefined coding scheme, the results were definitive. The models' accuracy was substantially lower than the human baseline, demonstrating a clear inability to replicate nuanced human judgment even when provided with a clear framework and examples.

These findings lead to two primary contributions. For digital health, this research shows that to be effective, cessation aids must be designed to personalize activities to address specific user barriers and skepticism. Methodologically, the study offers a nuanced verdict on the role of AI in qualitative analysis. While LLMs show potential as an exploratory tool for theme generation, they are not yet a viable tool for applying a predefined coding scheme, making human oversight essential for ensuring the depth and validity of qualitative research.

# 8 Acknowledgments

I would like to thank my supervisor, Prof. Willem-Paul Brinkman, for their guidance, feedback, and support throughout this project. I also thank my fellow research group members for their helpful feedback and collaboration.

For transparency, it is noted that OpenAI's ChatGPT was used for assistance in this project. The tool was used to help refine the wording and flow of the text and to generate an initial list of literature at the beginning of the research. The core research, data analysis, and all interpretations presented in this paper are my own.

# References


[1] Nele Albers, Francisco S. Melo, Mark A. Neerincx, Olya Kudina, and Willem-Paul Brinkman. Psychological, economic, and ethical factors in human feedback for a chatbot-based smoking cessation intervention. *npj Digital Medicine*, 8(1):1–14, 2025.

[2] M. Bano, R. Hoda, D. Zowghi, and C. Treude. Large language models for qualitative research in software engineering: exploring opportunities and challenges. *Automated Software Engineering*, 31(1):8, 2024.

[3] Timothy W. Bickmore, Hoa Trinh, Xinxin Zhang, Sean O'Hara, Susan M. Sereika, Thomas K. Houston, Bianca A. Rosner, Lauren E. Markson, Gabriela Marcu, Heather E. Wilson, Tyler Fasolino, Michael Spataro, Amy Balzer, Sidney K. D'Mello, and Saeideh Kim. Chatbot-based coaching for post-hospitalization care: A multi-site randomized trial. *Journal of Medical Internet Research*, 26, July 2024.

[4] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[5] Giulia Carreras, Alessandra Lugo, Silvano Gallus, Barbara Cortini, Esteve Fernández, Maria José López, Joan B Soriano, Angel López-Nicolás, Sean Semple, Giuseppe Gorini, and TackSHS Project Investigators. Burden of disease attributable to second-hand smoke exposure: A systematic review. *Preventive Medicine*, 129:105833, 2019.

[6] Arturo Castellanos, Haoqiang Jiang, Paulo Gomes, Debra Vander Meer, and Alfred Castillo. Large language models for thematic summarization in qualitative health care research: Comparative analysis of model and human performance. *JMIR Medical Informatics*, 2023.

[7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[8] Aleix Dorca Josa and Marc Bleda-Bejar. Local llms: Safeguarding data privacy in the age of generative ai. a case study at the university of andorra. In *ICERI2024 Proceedings*, pages 7879–7888, Seville, Spain, November 2024. IATED Academy.

[9] Linwei He, Divyaa Balaji, Reinout W. Wiers, Marjolijn L. Antheunis, and Emiel Krahmer. Effectiveness and acceptability of conversational agents for smoking cessation: A systematic review and meta-analysis. *Nicotine & Tobacco Research*, 25(7):1241–1250, July 2023.

[10] Linwei He, Erkan Basar, Emiel Krahmer, Reinout Wiers, and Marjolijn Antheunis. Effectiveness and user experience of a smoking cessation chatbot: Mixed methods study comparing motivational interviewing and confrontational counseling. *Journal of Medical Internet Research*, 26:e53134, August 2024.

[11] E. I. Ibrahim and A. Voyer. The augmented qualitative researcher: using generative ai in qualitative text analysis. SocArXiv, January 2024. Preprint posted online on Jan 22, 2024.

[12] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[13] Florian Mrozinski, Tobias Dargel, and David Krawinkel. Scaling hermeneutics: a guide to qualitative coding with llms for reflexive content analysis. *Research Square preprint*, 2024.

[14] World Health Organization. Tobacco, 2023. Accessed: May 8, 2025.

[15] Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6456–6469, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, Oct 2011.

[17] Maximo R Prescott, Samantha Yeager, Lillian Ham, Carlos D Rivera Saldana, Vanessa Serrano, Joey Narez, Dafna Paltin, Jorge Delgado, David J Moore, and Jessica Montoya. Comparing the efficacy and efficiency of human and generative ai: Qualitative thematic analyses. *JMIR AI*, 3(1):e54482, August 2024.

[18] Julia Romberg, Christopher Schröder, Julius Gonsior, Katrin Tomanek, and Fredrik Olsson. Have llms made active learning obsolete? surveying the nlp community. *arXiv preprint arXiv:2503.09701*, 2025.

[19] Victor J Strecher, Jennifer McClure, Gwen Alexander, Bibhas Chakraborty, Vijay Nair, Janine Konkel, Sarah Greene, Mick Couper, Carola Carlier, Cheryl Wiese, Roderick Little, Cynthia Pomerleau, and Ovide Pomerleau. The role of engagement in a tailored web-based smoking cessation program: Randomized controlled trial. *J Med Internet Res*, 10(5):e36, Nov 2008.

[20] Deepika V, Praveen S. Jodalli, and Avinash B R. The role of chatbots and virtual assistants in enhancing tobacco cessation counselling. *Frontiers in Digital Health*, 7:1503227, 2025.

[21] Ferdinand Wittmann. Enhancing thematic analysis with large language models: A comparative study of structured prompting techniques. *Preprint*, 2024. Leveraging Large Language Models to Automate and Enhance Inductive Thematic Analysis in Qualitative Research.

[22] Nina Åsbring, Samira Dini, Stephanie Madsen, and Joanna Stjernschantz Forsberg. Can a smoking cessation app benefit individuals in disadvantaged areas? a qualitative study on motivation, barriers, and perceptions of a digital app. *Preventive Medicine Reports*, 48:102925, Nov 2024.

# A Prompts used for Thematic analysis (TDSP process)

For each stage of the TDSP, we have example prompts below, which were used to perform the actual analysis:

1. You are an expert qualitative researcher specializing in thematic analysis. Your task is to analyze text data from smokers reflecting on their experiences with preparatory activities for smoking cessation, as guided by a conversational agent.

   Research Question: "How do smokers experience proposed preparatory activities by a conversational agent as part of an online smoking cessation intervention?"

   Dataset: I will provide you with a dataset of text responses from participants. Each response reflects their experience with preparatory activities. It will be from an excel sheet in a column, so different rows are different participants. Each row contains a combined text response so participants could have answered 1-4 questions in each row.

   Instructions: Based on the provided research question and dataset, please perform the following:

   Identify and list the main overarching themes that emerge from these smokers' descriptions of their experiences with the preparatory activities. For each overarching theme, provide a brief descriptive label. For each overarching theme, identify and list potential sub-themes that provide more specific details or facets of that broader theme. For each sub-theme, provide a brief descriptive label. Aim for clarity and conciseness in your theme and sub-theme labels. The themes should capture the essence of the smokers' experiences related to the preparatory activities. Output Format: Please structure your output as follows:

   - Overarching Theme 1:
     - Sub-theme 1.1:
     - Sub-theme 1.2:
   - Overarching Theme 2:
     - Sub-theme 2.1:
     - Sub-theme 2.2:

   ... and so on for all themes

   Data: [Paste responses here]

2. Context: We are conducting a thematic analysis on smokers' experiences. We have identified the following sub-themes from the overall dataset:

   - Overarching Theme 1:
     - Sub-theme 1.1:
     - Sub-theme 1.2:
   - Overarching Theme 2:
     - Sub-theme 2.1:
     - Sub-theme 2.2:

Task: Please analyze the following text response from a participant: "[Insert the concatenated text response from one row here]"

Instruction: Based on the definitions provided, which of the sub-themes (A, B, C, etc.) is/are most relevant to this participant's response? Please list the most relevant sub-theme(s). If multiple are relevant, list them in order of perceived importance. If none seem directly applicable, please indicate that.

3. You are an expert qualitative researcher. Your task is to extract key insights from participant responses that have been categorized under a specific sub-theme. These responses relate to smokers' experiences with CA-guided preparatory activities for smoking cessation.

   Main Research Question: "How do smokers experience proposed preparatory activities by a conversational agent as part of an online smoking cessation intervention?"

   Sub-theme for Analysis: - Sub-theme Label: - Sub-theme Definition/Description:

   Categorized Data for this Sub-theme: The following are text responses from participants that have been identified as relating to the sub-theme: []:

   Instructions: Based on the main research question, the sub-theme definition, and the categorized data provided: 1. Carefully read all the provided participant responses for this sub-theme. 2. Identify and articulate distinct key insights that elaborate on or explain the experiences related to this specific sub-theme. A key insight should be a specific pattern, idea, or observation emerging from these responses. 3. For each key insight you identify: a. Provide a concise descriptive statement of the insight. b. Select one or two direct quotes from the provided participant responses that best illustrate or support this insight. Ensure quotes are exact.

   Output Format: Please structure your output as follows:

   Sub-theme:

   Key Insight 1: - Supporting Quote(s): - "" - "[Another direct quote, if applicable]"

   Key Insight 2: - Supporting Quote(s): - "" - "[Another direct quote, if applicable]"

   (and so on for all identified key insights for this sub-theme)