ChatGPT for Automated Qualitative Content Analysis: Intercoder Reliability

Rimke Bijker, Stephanie S Merkouris, Nicki A Dowling, Simone N Rodda

Submitted to: Journal of Medical Internet Research
on: March 31, 2024

Table of Contents

ChatGPT for Automated Qualitative Content Analysis: Intercoder Reliability

Rimke Bijker[1] MSc, PhD; Stephanie S Merkouris[2] PhD; Nicki A Dowling[2] PhD; Simone N Rodda[1, 2] PhD

[1]Department of Psychology and Neuroscience Auckland University of Technology Auckland NZ
[2]School of Psychology Deakin University Burwood AU

Corresponding Author:
Simone N Rodda PhD
Department of Psychology and Neuroscience
Auckland University of Technology
90 Akoranga Drive
Auckland
NZ

Abstract

Background: Analysis of web-based data can provide insights into any number of conditions including the mechanisms of behavior change through to attitudes towards treatment. However, data analysis approaches like qualitative content analysis are notoriously time and labor intensive because of the time to detect, assess and code a large amount of data. Tools such as ChatGPT may have tremendous potential in automating at least some of the analysis.

Objective: The aim of this study was to explore the utility of ChatGPT in conducting qualitative content analysis through the analysis of forum posts from people sharing their experiences on reducing their sugar consumption.

Methods: Inductive and deductive content analysis were performed on 537 forum posts to detect mechanisms of behavior change. Thorough prompt engineering provided appropriate instructions for ChatGPT to execute data analysis tasks. Data identification involved extracting change mechanisms from a subset of forum posts. Precision of the extracted data was assessed by comparison with human coding. Based on the identified change mechanisms, coding schemes were developed with ChatGPT using data-driven (inductive) and theory-driven (deductive) content analysis approaches. The deductive approach was informed by the Theoretical Domains Framework using both unconstrained coding scheme and structured coding matrix. Ten coding schemes were created from a subset of data and then applied to the full dataset in 10 new conversations resulting in 100 conversations each for inductive and unconstrained deductive analysis. Ten further conversations coded the full dataset into the structured coding matrix. Inter-coder agreement was evaluated across and within coding schemes. ChatGPT output was also evaluated by the researchers to assess whether it reflected prompt instructions.

Results: The precision of detecting change mechanisms in the data subset ranged from 66% to 88%. Overall kappa-scores for inter-coder agreement ranged from 0.72-0.82 across inductive coding schemes and from 0.58-0.73 across unconstrained coding schemes and structured coding matrix. Coding into the best performing coding scheme resulted in category-specific kappa scores ranging from 0.67-0.95 for the inductive approach and 0.13-0.87 for the deductive approaches. ChatGPT largely followed prompt instructions in producing a description of each coding scheme although wording for the inductively developed coding schemes were lengthier than specified.

Conclusions: ChatGPT appears fairly reliable in assisting with qualitative content analysis. ChatGPT performed better in developing an inductive coding scheme which emerged from the data over adapting an existing framework into an unconstrained coding scheme or coding directly into a structured matrix. Potential for ChatGPT to act as a second coder also appears promising with almost perfect agreement in at least one coding scheme. The findings suggest ChatGPT could prove useful as a tool to assist in each phase of qualitative content analysis, but multiple iterations are required to determine the reliability of each stage of analysis.

(JMIR Preprints 31/03/2024:59050)
DOI: https://doi.org/10.2196/preprints.59050

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✔ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

ChatGPT for Automated Qualitative Content Analysis: Intercoder Reliability

Abstract

Background: Analysis of web-based data can provide insights into any number of conditions including the mechanisms of behavior change through to attitudes towards treatment. However, data analysis approaches like qualitative content analysis are notoriously time and labor intensive because of the time to detect, assess and code a large amount of data. Tools such as ChatGPT may have tremendous potential in automating at least some of the analysis.

Objective: The aim of this study was to explore the utility of ChatGPT in conducting qualitative content analysis through the analysis of forum posts from people sharing their experiences on reducing their sugar consumption.

Methods: Inductive and deductive content analysis were performed on 537 forum posts to detect mechanisms of behavior change. Thorough prompt engineering provided appropriate instructions for ChatGPT to execute data analysis tasks. Data identification involved extracting change mechanisms from a subset of forum posts. Precision of the extracted data was assessed by comparison with human coding. Based on the identified change mechanisms, coding schemes were developed with ChatGPT using data-driven (inductive) and theory-driven (deductive) content analysis approaches. The deductive approach was informed by the Theoretical Domains Framework using both unconstrained coding scheme and structured coding matrix. Ten coding schemes were created from a subset of data and then applied to the full dataset in 10 new conversations resulting in 100 conversations each for inductive and unconstrained deductive analysis. Ten further conversations coded the full dataset into the structured coding matrix. Inter-coder agreement was evaluated across and within coding schemes. ChatGPT output was also evaluated by the researchers to assess whether it reflected prompt instructions.

Results: The precision of detecting change mechanisms in the data subset ranged from 66% to 88%. Overall kappa-scores for inter-coder agreement ranged from 0.72-0.82 across inductive coding schemes and from 0.58-0.73 across unconstrained coding schemes and structured coding matrix. Coding into the best performing coding scheme resulted in category-specific kappa scores ranging from 0.67-0.95 for the inductive approach and 0.13-0.87 for the deductive approaches. ChatGPT largely followed prompt instructions in producing a description of each coding scheme although wording for the inductively developed coding schemes were lengthier than specified.

Conclusions: ChatGPT appears fairly reliable in assisting with qualitative content analysis. ChatGPT performed better in developing an inductive coding scheme which emerged from the data over adapting an existing framework into an unconstrained coding scheme or coding directly into a structured matrix. Potential for ChatGPT to act as a second coder also appears promising with almost perfect agreement in at least one coding scheme. The findings suggest ChatGPT could prove useful as a tool to assist in each phase of qualitative content analysis, but multiple iterations are required to determine the reliability of each stage of analysis.

Keywords: ChatGPT; natural language processing; qualitative content analysis; Theoretical Domains Framework

Introduction

Background

Emerging from a variety of fields such as psychology, nursing, media communication, and market research, content analysis has become one of the main methods for analyzing large qualitative datasets [1, 2]. Content analysis employs systematic and replicable methods to synthesize and describe patterns in textual, visual, or audio data and facilitate understanding of content and meaning of the data [1-4]. It can be used to analyze qualitative data obtained from interviews, focus groups, and patient records, as well as naturally occurring data, such as user-generated website content, discussion forums, social media, and customer databases [4, 5]. Furthermore, content analysis can employ quantitative methods, which focus on quantification of the data to enable calculation of prevalence data or use in statistical analyses, or qualitatively methods, which focus on distilling concepts or constructs from the data to facilitate understanding [6]. Qualitative content analysis also allows integration of quantitative components, for instance by summarizing the findings in frequency distributions, which can be particularly useful when working with large datasets [7].

Content analysis can be performed using an inductive, data-driven approach or a deductive, theory-driven approach. An inductive approach is suitable when the topic of interest is still emerging or lacking a firm scientific knowledge database [4, 8, 9]. In contrast, a deductive approach is more appropriate when there is an existing body of knowledge and the aim is to confirm hypotheses and code data into an existing framework or theory [3, 10]. While the practice and philosophy of qualitative content analytic approaches may vary, they have similar tasks to be executed as part of the data analysis process, including identification of relevant data, organisation of data, and data classification [6]. The systematic nature of this process provides opportunities for computer assistance. In fact, computer-assisted execution of qualitative analysis tasks has been around for some time with the aim of improving the speed and accuracy of the analysis process, especially when large datasets are involved [11].

Whether inductive or deductive, each approach of content analysis requires a preparation phase, which involves reading the text for familiarization and identifying the meaning units (i.e., the words or sentences in the text that are relevant to the research questions and contained within the unit of analysis) by allocating them initial labels that reflect the key content of the meaning units. [12]. It is recommended that this is repeated multiple times, whereby there is a cycle of reading and re-reading the data and adjusting the labels so that they represent the data that are being coded. [1]. The data are then subject to condensation where the meaning units are summarized into short descriptions which maintain the key content of the data and are allocated codes that are grouped according to content similarity [13]. Traditional qualitative analysis software offer automated coding but multiple studies have reported inaccuracies, suggesting that manual coding should remain the dominant approach [13, 14].

Following the preparation phase, machine learning may assist in the development of a coding scheme (also called a coding frame or data dictionary) that can be used to systematically organize the data by allocating each code into that scheme. When the analysis involves large datasets, coding schemes are typically constructed on a subset of data, after which the coding scheme can be applied to the full dataset [15]. For the inductive content analysis approach, the development of a coding scheme involves organizing codes and initial groupings into categories according to the research question [3, 4]. Categories are continuously adjusted and are often split in sub-groups to ensure they are discrete and mutually exclusive, while still adequately reflecting the data [12]. Meaning units might have to be adapted as well to ensure that each code belongs to only one category [2]. Multiple

researchers can also develop lists of categories that are combined into one coding scheme via negotiation [16]. For each category, labels are developed to define the category and, commonly, a separate coding manual is composed to specify coding rules [2].

For the deductive content analysis approach, codes are similarly organized into categories, however, the categories are pre-defined in a coding matrix which is informed by a theory or framework [3, 4]. The matrix can then be made into a coding scheme which reflects the data being coded, by labeling and relabeling the categories in line with the current dataset [3]. This unconstrained deductive approach results in a coding scheme where different categories may emerge that still reflect the data and are consistent with the theory or framework. Alternatively, a structured deductive approach can test categories, concepts, and hypotheses by coding data directly into predefined categories of the coding matrix [3]. When using a structured coding matrix the categories are not redefined to reflect the data being coded, thereby giving the researcher the opportunity to focus on the degree of alignment between the matrix and current dataset [3].

Whether following an inductive or unconstrained deductive approach, development of a coding scheme is an iterative process which involves continuously updating the coding scheme until a version is achieved which can be used to reliably code the entire data set [17]. Development and refinement of coding schemes is generally conducted by the research team and may either involve others with expertise in the framework or theory or a literature review [3, 4]. Recent evidence suggests that machine learning and natural language processing techniques are promising for relatively straight-forward tasks such as data extraction and classification of unstructured qualitative data [18, 19]. Employing such techniques, however, typically requires programming skills that are unavailable to most researchers.

After the development phase, the quality of the coding schemes can be evaluated through assessing reliability of application. This phase involves two or more researchers classifying the data into the defined categories and comparing the resulting annotated datasets [2]. Manual coding may be aided by technology in terms of sorting and renaming categories (e.g., using excel or qualitative analysis software) but this is also prone to fatigue and error, especially where coding schemes have multiple categories and sub-categories or where interpretation of meaning is part of the selected approach [3, 10, 13, 20, 21]. For example, reliability is dependent on the number of correctly coded items which may be at the higher order or sub-category level. While percent agreement is often reported [21], a more robust approach is through kappa inter-rater agreement (henceforth referred to as inter-coder agreement, a more appropriate term when classification relates to nominal categories) [28]. This inter-coder agreement determines the degree to which the coders agree beyond what would be expected by chance alone [22]. The calculation of kappa score relies on each category in the coding scheme being mutually exclusive and the unit of analysis being independent in that it contains only one statement or code [22]. Where the kappa score indicates low inter-coder agreement, the coding scheme might require further refinement until application of the coding scheme achieves high inter-coder agreement [2]. The final coding scheme can be used to classify the content of the full dataset and produce frequency data for further analysis. Where previous evaluation of the coding scheme was based on a data subset, reliability might be assessed again once the full dataset is annotated by multiple researchers as coding of large datasets may be prone to errors due to researcher fatigue and subtle changes in interpretation over time [2, 6, 13]. While reliability is commonly evaluated by inter-coding agreement, an alternative option is intra-coding agreement in which the same person codes the same data on two separate occasions [23].

In addition to evaluating reliability, validity of the findings from content analysis can be enhanced via triangulation, which can be achieved by combining methodologies, involving various

investigators, or incorporating multiple theories [4, 8, 13, 20, 24]. In content analysis, increased validity might be achieved by transparency in reporting and potential for replication, confirmation that categories relate to the research questions, member checks, evaluation of the appropriateness of the coding scheme by content experts, and the use of different data sources [3, 4, 13]. Other considerations relate to credibility (i.e., how well data analysis procedures ensure inclusion of all relevant data, which may be improved through consensus amongst the research team or external experts) and dependability (i.e., how the analytic process may alter data over time – such as recoding data -, thereby implying a need for careful logging of changes and consistent decision making) [13].

ChatGPT

Advances in artificial intelligence (AI) and machine learning technologies have greatly impacted working life by automating manual processes and assisting in laborious tasks. One of the more prominent developments by OpenAI is the Chat Generative Pre-trained Transformer (ChatGPT), a large language model (LLM) that has been trained using a wide range of data, which enables it to synthesize human inputs and generate human-like responses [25]. Inputs are provided in the form of 'prompts' which contain questions or instructions that are processed to provide responses in line with the training data. According to an investigation of Altmetrics, a platform that tracks online attention to research, the interest in ChatGPT appears to be particularly high among scientists [26]. Since its launch in November 2022 [27], scientists have enlisted the assistance of ChatGPT for a range of tasks, including conceptualizing ideas, summarizing scientific literature, generating analytic code, and writing up manuscripts [28, 29]. In this field, ChatGPT's potential has primarily been discussed in relation to scientific writing and editing. For example, ChatGPT has been used to conduct rapid literature reviews, translate and edit texts, structure manuscripts to increase readability and fulfill journal guidelines, write logically sound abstracts, and provide suggestions on how to address reviewer comments [25, 30-32].

Research suggests that ChatGPT may be a promising tool for qualitative analysis methods [33-36]. As a natural language processing tool that applies computation techniques and semantic similarities to analyze and synthesize natural language and speech [37], ChatGPT may enhance the streamlining of tasks and increase the efficiency of qualitative research projects. Nevertheless, interviews with those involved in qualitative analysis suggest the time needed for prompt engineering, understanding complex responses, and organizing unstructured output, combined with concerns about accuracy and validity of the output may not outweigh the benefits of incorporating ChatGPT as a tool in qualitative research [38].

Study aims

The aim of this study was to explore the utility of ChatGPT in conducting qualitative content analysis through the analysis of forum posts from people sharing their experiences on reducing their sugar consumption. Specifically, given the various analytic approaches that can be employed, the aim was to explore the utility of ChatGPT in both inductive (i.e., data-driven) and deductive (i.e., using an existing framework) approaches. For the deductive approach we used the Theoretical Domains Framework (TDF) as this is a commonly used framework to guide evaluation of the implementation of evidence-based practice in health care settings [39]. A secondary aim was to provide insight into the mechanisms of sugar reduction most frequently discussed in forum data by providing frequency data generated with the inductive and deductive approaches.

Methods

Overview

An overview of the study methods is presented in Figure 1. The qualitative content analysis process was broken down into several tasks - referred to as queries - which we identified as potential for automation using ChatGPT. Prompt engineering was used to create appropriate prompts with instructions in line with the respective tasks (i.e., the queries). In approximation of a manual coding process, ChatGPT was first used to generate a dataset of condensed meaning units which were extracted from forum posts on sugar reduction. Based on a subset of the generated dataset, ChatGPT was then used to create 10 versions of an inductive, data-driven coding scheme and 10 versions of an unconstrained deductive, theory-driven coding scheme. The unconstrained deductive coding scheme was based on the TDF. Thus, different versions of the coding schemes were developed in parallel, in contrast to a manual process where coding schemes are developed iteratively and reflect more refined updates from previous versions. ChatGPT was then used to annotate the full dataset by applying the coding schemes. ChatGPT was also used to code the full dataset into a structured coding matrix based on the TDF where data was coded directly into the predefined categories guided only by the training data of ChatGPT's underlying LLM. All coding scheme versions and the coding matrix were applied to the full dataset 10 times in 10 new conversations, whereby each conversation was considered an independent coder. By doing so, the resulting annotated datasets could be compared to assess the inter-rater reliability of the coding schemes.

Figure Study flow

All ChatGPT queries were conducted from 12-18 June 2023 when ChatGPT operated under the GPT-3.5-turbo model [40]. We opted to use the ChatGPT web application with unpaid subscriptions to ensure that our approach can be applied by anyone interested in using qualitative content analysis, regardless of the programming skills and financial resources they have available to conduct their research.

Data source

The data source of this study was from a study on mechanisms of sugar reduction which has been previously described [41]. Briefly, Google searches were conducted to identify online content in which internet users mentioned mechanisms related to changing sugar consumption. In this search, potentially relevant posts were sourced from consumers platforms (e.g., forums, message boards), popular media platforms (e.g., blogs, online magazines, news articles), and professional platforms (governmental and treatment provider websites). User-generated posts (including context, where applicable) were manually transferred to a Microsoft Excel spreadsheet and allocated labels and codes reflecting the change mechanisms identified in the forum posts. Mechanisms were then analyzed to create an overview of the range of change mechanisms applied to reduce sugar consumption. For the current study, we retained all consumer posts which resulted in a dataset with 539 unique forum posts.

All data were from open-access sources and freely available without sign-in requirement or agreement with a specified set of terms and conditions. The retained posts did not contain any user-identifiable data. As such, the study was exempted from ethical approval in accordance with Auckland University of Technology Ethics Committee guidelines [42].

Prompt engineering

Research has highlighted the importance of using high-quality prompts for LLMs, given that

different prompts that appear to reflect similar instructions may lead to highly variable responses [43-45]. Two researchers (RB, SM), therefore, engaged in a systematic process of prompt engineering to generate and compare prompts. To enhance the quality of our prompts, we used a combination of various techniques that have been suggested to improve responses in conversations with LLM such as ChatGPT [46-48]. For example, we started conversations with prompts that included information on the context of our study and then applied techniques to engage ChatGPT in an iterative prompt engineering process in which instructions referred to writing relevant prompts, providing alternative prompts, refining prompts based on additional information, listing pros and cons of prompts, and recommending prompts given the relevant pros and cons. Through this process we refined and selected at least three prompts per analysis task that were all run several times to assess which prompts contained the most optimal instructions for use in the qualitative content analysis query process. Assessment at this stage was based on the extent to which responses were in line with the prompt instructions such as correct response format, no apparent issues in response content and comprehensiveness of response.

The final set of prompts for the query process all had a similar construction in that each prompt started with a) detailed instructions on the task to be executed, including notes on what to watch out for, if applicable, followed by b) specifications on the required response format and/or answer template, and ending with c) information on which the response had to be based (i.e., list of forum posts or change mechanisms). In some prompts the instructions followed a step-by-step structure [49] as prompt engineering showed that this resulted in a more comprehensive response. See Textbox 1 for a prompt example (with step-by-step structure) and Supplementary materials for a complete overview of the final prompts (Multimedia Appendix 1: Final prompts). At the time of this research, ChatGPT had a limit on the number of tokens (i.e., units that are meaningful to an LLM, such as words, word fragments, or punctuation signs) processed in each combined prompt and response. Consequently, for each query we re-prompted ChatGPT with the same instructions but different subsets of the data on which responses had to be based until all data was processed.

Textbox 1. Example of a prompt used in this study.

Below are forum posts reflecting real-life experiences related to changing sugar consumption. For each of these posts, please extract all potential mechanisms/strategies/techniques for reducing sugar consumption. Each mechanism should be described in a brief and concise manner, using up to 10 words. If multiple mechanisms can be extracted from a single post, each should be mentioned separately. Please note that mechanisms may not be explicitly presented as such, so be sure to interpret and extract them from the forum posts accordingly.

First format the response as:

“

Post 1:
- mechanism 1: {max 10 word description}
- mechanism 2: {max 10 word description}
- etc (if applicable)

Post 2:
- mechanism 1: {max 10 word description}

- mechanism 2: {max 10 word description}
- etc (if applicable)
"

Then please reformat the response as an excel spreadsheet with in the first column the post number, the second the mechanism number, and the third the mechanism description. Please format the table as csv and put the table in code block.

Forum posts:
"""
[insert forum posts]
"""

After prompt engineering, four new OpenAI accounts were created and used on four different computers to run the large number of queries. As such, the accounts were used to conducted tasks on a different computer simultaneously or repeat tasks on the same computer within an interval of 48 hours. New conversations were started for each task and ended when all data pertaining to the task in the query were processed (i.e., one run of the task). Conversations were logically labelled to enable ease of look-up at a later stage (e.g., Task1_account1_run3, Task5_account4_run1).

Query Process

Data Preparation

Data preparation involved creating a dataset of condensed meaning units based on the change mechanisms identified in the forum posts. Posts from the original published sugar study were randomized using a random number generator function in Excel. The forum posts were included alongside step-by-step instructions to identify potential change mechanisms from the provided posts (see Textbox 1). The first step included directions to extract the mechanisms for reducing sugar consumption from the posts and to summarize the change mechanisms in a brief description (up to 10 words). The resulting brief descriptions functioned as the condensed meaning units to be used for coding scheme development. In the next step, ChatGPT was instructed to reformat the response into a table with post number, mechanism number (within post) and the brief description of each identified change mechanism. All output tables in the conversation were transferred to an Excel spreadsheet which received a label consistent with the conversation name. The data preparation query was repeated to emulate a manual data preparation phase involving multiple coders and enabling comparison of brief descriptions across coded datasets. The query was repeated 10 times in 10 different conversations, thereby resulting in ten datasets reflecting 10 different coders.

Comparison of the ten datasets generated by ChatGPT with human coding was based on the change mechanisms identified in 108 forum posts (i.e., a 20% data subset). Three researchers (RB, SM, SR) compared the condensed meaning units with the forum posts and indicated which condensed meaning units correctly reflected a change mechanism described in the forum post and which condensed meaning units incorrectly identified a change mechanism from the post. Datasets were double-coded, and disagreements were discussed until consensus was achieved. The dataset with the highest percentage of correctly identified change mechanisms across all identified change mechanisms (i.e., the best precision) was selected for further use in the analysis. Specifically, the

20% data subset was used for the development of the inductive and unconstrained deductive coding schemes after which the developed coding schemes were applied to the full dataset of change mechanisms. For the structured deductive approach, the full dataset was coded directly into a TDF coding matrix. Change mechanisms in the selected dataset were given a unique identifier from 1 to *n* to enable data linkage in subsequent tasks.

Inductive Approach: Development and Application of Data-Driven Coding Schemes

The inductive approach started with a task that reflected the development of a coding scheme where categories are derived from the data [3]. The prompt developed for this task instructed ChatGPT to organize meaning units from the subset of data (20%) into categories based on underlying patterns in the data. The prompt included explicit instructions to ensure that there was no overlap between the categories. Furthermore, the instructions and answer template indicated that ChatGPT had to provide a 20-word label to define each category. The task was repeated in 10 new conversations to create 10 versions of an inductively developed coding scheme.

The next task in the inductive approach reflected the classification of the full dataset into the inductively developed coding scheme. The output from the previous task was inserted into a prompt where instructions specified that change mechanisms were to be classified under the best matching category in accordance with the 20-word category definitions. ChatGPT was also instructed to format the output as a table which listed the mechanism unique identifier, the brief description of the change mechanism and the best matching category from the coding scheme. As there were ten inductively developed coding schemes, the standard prompt was also adapted ten times to reflect the various versions of the coding scheme. To reflect a process with ten independent coders, each coding scheme was applied ten times by starting ten new conversations per coding scheme.

Unconstrained Deductive Approach: Development and Application of Theory-Driven Coding Schemes

The deductive approach started with the development of an unconstrained coding scheme that was informed by the TDF. The prompt instructed ChatGPT to identify domains (akin to categories in the inductive approach) from the TDF and redefine these domains to reflect the current dataset subset. For the first step, the instructions specified grouping all mechanisms under the best aligning domain with explicit statements that a group could only reflect one domain and that each change mechanism could only be listed once. Instructions also specified listing each change mechanism under the domain in which they were grouped to ensure each mechanism was listed once. Prompt engineering illustrated that although ChatGPT training data includes sources on the TDF, prompt responses tend to incorporate fabricated or adapted domain names. Therefore, the prompt included a list of all 14 domains to remove any ambiguity in instructions. For the second step, ChatGPT was instructed to provide a 20-word definition for each domain based on the group of change mechanisms it listed under the domain in the first step. Instructions further detailed that domains not identified in step one should be labeled as “N/A”. The task was repeated in 10 new conversations to create 10 versions of a deductively developed coding scheme.

The next task in the deductive approach reflected classification of the full dataset into the domains of the deductively developed unconstrained coding scheme. The prompt instructed ChatGPT that change mechanisms were to be classified under the best matching category in accordance with the 20-word category definitions. There was no reference to the TDF in the instructions to ensure that

ChatGPT based its response on each provided coding scheme and to minimize the risk of unintentionally infusing the instructions with additional information based on ChatGPT's pre-existing knowledge of the TDF. The prompt was adapted ten times (i.e., once for each of the unconstrained coding schemes) and ten new conversations were started per coding scheme (to mimic a process with ten coders).

Structured deductive approach: Coding directly into TDF coding matrix

The full dataset was classified directly into the domains of the TDF using a coding matrix which listed all 14 TDF domains [39] but was empty of definitions for the domains. Thus, this approach did not consider the underlying data and solely relied on ChatGPT's pre- existing knowledge of the TDF and how it would describe the mechanisms of sugar reduction. The prompt included instructions to classify each change mechanism under only the TDF domain that best matched each mechanism and specified the response format to be in the form of a table with columns for the mechanism unique identifier, short description, and the best matching TDF domain. Direct coding into the TDF was conducted over ten new conversations, reflective of a process with ten coders.

Post processing

The output of conversations applying the coding schemes to the full dataset were manually transferred to Microsoft Excel spreadsheets and labelled in accordance with the conversation names. As described previously [50], ChatGPT-generated data commonly requires various post-processing steps to clean the data and to check whether responses are in line with instructions. Checks were performed on all output to see whether the change mechanisms descriptions in the output were identical to those included in the prompt. Doing so revealed instances of hallucination where ChatGPT altered brief descriptions in the output generated after lengthier prompts. Where hallucinations were identified, ChatGPT was re-prompted with the instructions followed by only those change mechanisms that were incorrectly processed. Furthermore, post processing involved minor adjustments to clean category and domain names where the full label was not reported. In some conversations where coding was directly into the TDF, output showed instances where more than one domain was allocated per change mechanism. In these cases, the first of the annotated domains was selected for use in data analysis. Finally, the clean datasets were aggregated by coding scheme version. This entailed creating one spreadsheet per coding scheme where the first column reflected the mechanism unique identifier, the second column the brief descriptions, and the remaining ten columns the category or domain name allocated to the change mechanism by conversations in which the respective coding scheme was applied.

Quality evaluation

Results from the preparation phase were presented as number of change mechanisms identified in the subset and number that ChatGPT correctly identified against human coding. We also present the total number of change mechanisms identified in the full dataset across each of the conversations. Results from the development of the inductive and unconstrained deductive coding schemes were presented as number of categories or domains identified and median and range of label word count per coding scheme version. Furthermore, the content of the labels was inspected to check whether coding schemes were in line with the instructions (e.g., no overlap in categories or labels). The structured deductive approach was not evaluated on any of these metrics as the coding matrix contained all 14 TDF domains and lacked labels with domain definitions.

Reliability of the coding schemes was evaluated by calculating Fleiss' kappa [51] for inter-coder

agreement with more than two coders. Specifically, inter-coder agreement calculation was performed for each version of the coding scheme developed using ChatGPT and the unstructured coding matrix by comparing allocated codes in the aggregated datasets. Kappa-statistics were calculated per coding scheme (i.e., overall kappa score) and per category (or domain) within the coding scheme (i.e., category-specific kappa score) to compare reliability across and within coding schemes. Kappa scores below 0 were interpreted as poor agreement, scores from 0.00-0.20 as slight agreement, scores from 0.21-0.40 as fair, scores from 0.41-0.60 as moderate, scores from 0.61-0.80 as substantial, and scores from 0.81-1.00 as almost perfect agreement [52].

Furthermore, frequency data was generated for each of the coding schemes and the coding matrix. To do so, the mode across the ten ChatGPT coders per change mechanism was retrieved for each of the aggregated datasets (i.e., the datasets which combined the annotated data from ten conversations). This meant frequency data reflected a majority agreement among the ChatGPT coders rather than a consensus decision-making approach. Thus, frequency tables were created per coding scheme after a decision-making approach where final classification of the mechanism was based on the mode of the annotated data per applied coding scheme. Frequency data and kappa scores were obtained using Stata software version 18.0 (StataCorp, College Station, TX).

Results

Evaluation of preparation phase

As displayed in Table 1, the number of change mechanisms identified from forum posts varied across datasets (i.e., the ChatGPT conversations from the data preparation), ranging from 571 to 623 condensed meaning units. Precision rates between ChatGPT and human coding ranged between 0.66 to 0.88. Two datasets yielded a precision of 0.88, which meant a decision needed to be made on which one to bring forward to the next phase. We selected the dataset which identified the larger number of change mechanisms based on the data subset (n=127), as this would enable a greater number of change mechanisms to be incorporated in the development of the coding schemes.

Table 1. Precision in identifying change mechanisms during preparation phase

Dataset generated in data preparation	Change mechanisms identified in full dataset (n)	Change mechanisms identified in 20% data subset (n)	Mechanisms correctly identified from subset (n)	Precision
Dataset 1	584	128	111	0.87
Dataset 2	587	119	105	0.87
Dataset 3*	585	127	112	0.88
Dataset 4	602	118	104	0.88
Dataset 5	571	131	110	0.84
Dataset 6	616	128	109	0.85
Dataset 7	619	132	114	0.86
Dataset 8	584	124	107	0.86
Dataset 9	623	135	115	0.85
Dataset 10	591	121	101	0.66

* Dataset selected for development of coding schemes

Evaluation of the inductive approach

Results from the inductive approach are presented in Table 2. The inductively developed coding schemes contained a variable number of categories, ranging from 5-13 categories. The category labels were overall lengthier than instructed, with a median label word count of 15-51 across coding schemes. Overall kappa scores indicated substantial or almost perfect inter-coder agreement for all coding schemes. The majority of definitions included specific examples of change mechanisms. Categories with labels reflecting a mixed or residual group of change mechanisms had category-specific kappa scores indicating moderate or less than moderate inter-coder agreement. Frequency tables based on a majority agreement decision-making approach showed that categories reflecting substituting sugary products (16-29%), gradually reducing sugar consumption (12-20%), and a joint category reflecting both substituting and reducing sugar (41%) were the topmost frequently mentioned categories in the final annotated datasets. See supplementary material for a detailed overview of all inductively developed coding schemes, related metrics, and frequency tables (Multimedia Appendix 2: Inductively developed coding schemes and meta data)

Table 2. Inter-coder agreement for application of coding schemes from the inductive approach

Aggregated datasets by version of data-driven coding scheme	Categories in coding scheme	Label word count		Inter-coder agreement
	n	median	range	Overall kappa [a]
Version 1	10	21	15-25	0.76
Version 2	9	26	13-34	0.69
Version 3	5	28	24-29	0.79
Version 4	10	30	21-36	0.84
Version 5	9	15	12-20	0.77
Version 6	6	28	23-36	0.83
Version 7	10	25	19-31	0.77
Version 8	10	16	11-22	0.72
Version 9	13	51	10-69	0.77
Version 10	5	17	15-19	0.72

[a] $P<.001$ for intercoder agreement for all versions of the coding scheme

Coding scheme 4 had the best inter-coder agreement overall ($\kappa=0.84$; $P<.001$) and was selected for further evaluation. As indicated in Table 3, inter-coder agreement on the categories within this coding scheme were almost perfect (i.e., $\kappa>0.80$; $P<.001$) for all but two categories. Particularly high category-specific inter-coder agreement ($\kappa=0.91$; $P<.001$) was observed for a category which included specification of what was not included in the category (i.e., "*[…] This approach contrasts with gradual reduction or moderation strategies*"). There was overlap in the names of two categories (i.e., "*Sugar alternatives and substitutes*" and "*Substitution and replacement approaches*"), however, the category definitions clearly delineated these categories and both categories had almost perfect inter-coder agreement ($\kappa= 0.84$; $P<.001$).

We also examined the content of coding scheme version 2 which had the poorest overall inter-coder agreement ($\kappa= 0.69$; $P<.001$). The poorest category-specific agreement ($\kappa= 0.46$; $P<.001$) within this coding scheme was for a miscellaneous category. Furthermore, overlap was observed in the definitions with similar examples provided for multiple categories. For example, "*seeking advice and support*", "*seeking support from others*", and "*seeking parental guidance*" appeared in three different categories, and "*avoiding purchasing sugary foods*" and "*changing shopping habits*"

appeared in two different categories.

Table 3. Inductively developed coding scheme with best overall inter-coder agreement.

Category	Definition	Category-specific inter-coder agreement	Frequency distribution	
		kappa [a]	Number of change mechanisms coded into category[b]	%
Psychological and behavioral strategies	Employing various psychological and behavioral techniques to change sugar consumption, such as seeking support, addressing addiction, recognizing the problem, planning ahead, and finding alternative distractions.	0.83	116	19.8
Substitution and replacement approaches	Replacing sugary foods and drinks with healthier alternatives, including options without added sugars, high fiber carbs, fruits, nuts, or drinks like cinnamon tea, black coffee, or soda water.	0.84	111	19.0
Gradual reduction and moderation methods	Gradually reducing sugar consumption over time, either by reducing portion sizes, gradually decreasing sugar in tea/coffee, or incorporating a gradual process for adapting to reduced sugar. Moderation and small daily changes are emphasized.	0.85	85	14.5
Knowledge and awareness-based approaches	Gaining knowledge about the harmful effects of excessive sugar consumption, checking sugar content in products, reading articles for information, and being aware of hidden sugars in various food products. Seeking resources and advice is also encouraged.	0.87	65	11.1
Environmental and practical strategies	Implementing changes in the environment to support reduced sugar consumption, such as changing grocery shopping habits, keeping cabinets stocked with healthy snacks, discarding carb-filled foods, and not buying sugary items in the first place.	0.88	51	8.7
Health and well-being focus	Emphasizing the benefits of reduced sugar consumption on energy and overall health, incorporating nutrient-rich foods, ensuring adequate sleep, exercising to reduce stress, and considering health consequences and diabetes complications as motivation.	0.72	40	6.8
Elimination and cold turkey approaches	Completely quitting sugar consumption abruptly or going “cold turkey” for better health. This approach contrasts with gradual reduction or moderation strategies.	0.91	35	6.0
Support and community	Seeking parental guidance, useful advice, moral support, weight loss buddies, or	0.95	30	5.1

engagement	support from others. Sharing success stories and helping others break addiction are also emphasized.			
Sugar alternatives and substitutions	Exploring and utilizing various sugar alternatives and substitutes, including natural sweeteners like stevia, honey, or Swerve, as well as incorporating naturally sweet options like fruit.	0.84	28	4.8
Personal determination and accountability	Making a firm decision to quit sugar consumption, acknowledging weak moments and impulsive behavior, taking small steps, acknowledging the time and trial-and-error process, building willpower, and using tools like MyFitnessPal to track sugar intake.	0.67	24	4.1

[a]*P*<.001 for intercoder agreement for all categories of the coding scheme; [b]Based on majority agreement decision-making approach across ChatGPT coders (i.e., by selecting the mode of the codes per change mechanism).

Evaluation of the unconstrained deductive approach

Results from the unconstrained deductive approach are presented in Table 4. None of the developed TDF coding schemes contained all 14 domains. The number of domains identified from the data subset was variable, ranging from 6-10 domains. Definition word counts were in line with the prompt instructions for all coding schemes (median between 12-17), with maximum definition word counts mostly being close to 20 words. The extent to which there was inter-coder agreement overall was moderate or substantial across coding schemes.

Table 4. Inter-coder agreement for application of the coding schemes from the unconstrained deductive approach

Aggregated datasets by version of theory-driven coding scheme	Domains	Label word count		Overall inter-coder agreement
	n	median	range	kappa [a]
Version 1	10	16	12-19	0.58
Version 2	8	13	12-18	0.62
Version 3	7	14	13-17	0.52
Version 4	7	16	14-24	0.73
Version 5	10	12	9-13	0.73
Version 6	10	15	14-22	0.53
Version 7	7	12	11-19	0.53
Version 8	9	14	13-17	0.62
Version 9	6	17	13-19	0.60
Version 10	10	16	12-22	0.58

[a]*P*<.001 for intercoder agreement for all versions of the coding scheme

Across all TDF coding schemes, domain-specific kappa scores ranged from 0.06 to 0.89 (Multimedia Appendix 3: Domain-specific scores per TDF coding scheme). The domains "*Beliefs about consequences"*, "*Environmental context and resources"*, and "*Social influences"* were identified in all versions of the coding scheme, while the domains "*Optimism"*, "*Reinforcement"*, and

"*Social/professional role and identity"* were in none. The domain "*Social influence"* had the highest domain-specific inter-coder agreement (κ0.79-0.89; *P*<.001) whereas the domain "*Memory, attention, and decision processes"* had the lowest domain-specific inter-coder agreement (κ0.06-0.63; *P*<.001). There was no overlap in domain labels within coding schemes. See supplementary material for a detailed overview of all TDF coding schemes, related metrics, and frequency tables (Multimedia Appendix 4: Unconstrained TDF coding schemes and meta data). Coding scheme version 5 had the greatest number of TDF domains and equal highest inter-coder agreement (κ=0.73; *P*<.001) and was therefore selected for further evaluation. As indicated in Table 5, domain-specific inter-coder agreement on six of 10 domains within this coding scheme was substantial or near perfect.

We compared domain labels across the coding schemes with the best and poorest inter-coder agreement. This revealed that domains with better domain-specific inter-coder agreement included examples of change mechanisms in their labels. An example of this was seen for the domain "*Behavioral regulation"* labeled as "*The self-directed process of monitoring, controlling, and modifying behaviors related to sugar consumption*" in one version (κ= 0.43; *P*<.001) and more descriptively labeled as "*Techniques and strategies used to regulate and control sugar consumption, including portion control, substitution, gradual reduction, and self-monitoring*" in another version (κ= 0.77; *P*<.001).

Table 5. Unconstrained deductive coding scheme with best overall inter-coder agreement.

Domains of the Theoretical Domains Framework	Definition	Domain-specific inter-coder agreement	Frequency distribution	
		Kappa[a]	Number of mechanisms coded into domain[b]	%
Behavioral regulation	Implementing strategies, techniques, and habits to regulate and control sugar consumption.	0.77	369	63.1
Knowledge	Acquiring information and understanding about the effects of sugar on health and nutrition.	0.79	54	9.2
Beliefs about consequences	Understanding and acknowledging the positive outcomes of reducing sugar consumption motivate behavior change.	0.75	48	8.2
Social influences	External factors and support from others influence sugar consumption and dietary choices.	0.87	38	6.5
Environmental context and resources	The impact of the physical environment and available resources on sugar consumption habits.	0.61	34	5.8
Beliefs about capabilities	Confidence in one's ability to change sugar consumption habits and overcome addiction.	0.57	24	4.1
Goals	Setting specific targets and objectives related to reducing sugar	0.68	10	1.7

	intake.			
Emotion	Emotional factors that influence sugar cravings and behavior change.	0.35	5	0.9
Intention	Having a clear purpose and determination to change sugar consumption behavior.	0.59	3	0.5
Memory, attention, and decision processes	Cognitive processes that involve memory, attention, and decision-making in relation to sugar consumption.	0.13	0	0.0

[a]*P*<.001 for intercoder agreement for all domains of the coding scheme
[b]Based on majority agreement decision-making approach across ChatGPT coders (i.e., by selecting the mode of the codes per change mechanism).

Evaluation of structured deductive approach

This approach used a structured coding matrix where change mechanisms were coded directly into TDF domains without first specifying domain labels. This meant that instead of 10 different coding schemes each applied using 10 different ChatGPT conversations, the structured deductive approach applied the single published coding matrix using 10 different conversations. The overall kappa scores for the coding matrix indicated substantial inter-coder agreement (κ=0.66; *P*<.001). As indicated in Table 6, domain-specific inter-coder agreement on seven domains of the TDF coding matrix had substantial or near perfect agreement and four domains had moderate agreement. The highest domain-specific inter-coder agreement was observed for the domain "*Social influences*" (κ= 0.85; *P*<.001), which similarly yielded almost perfect inter-coder agreement in all but one of the unconstrained coding schemes (as per previous section). The lowest domain-specific inter-coder agreement was observed for the domain "*Optimism*" (κ= 0.33; *P*<.001), which was not featured in any of the unconstrained coding schemes and for the domain "*Emotion*" (κ= 0.35; *P*=.551), which yielded fair to substantial inter-coder agreement in the unconstrained deductively developed coding scheme.

Table 6. Inter-coder agreement on the structured deductive approach with TDF coding matrix

TDF domain	Domain-specific inter-coder agreement	Frequency distribution	
	Kappa[a]	Number of change mechanisms coded into domain [b]	%
Behavioral regulation	0.66	281	48.0
Beliefs about consequences	0.73	82	14.0
Environmental context and resources	0.56	82	14.0
Knowledge	0.79	46	7.9
Social influences	0.85	29	5.0
Memory, attention, and decision processes	0.50	19	3.2
Social/professional role and identity	0.73	18	3.1
Beliefs about capabilities	0.54	12	2.1

Goals	0.74	7	1.2
Emotion	0.36[c]	4	0.7
Intentions	0.85	2	0.3
Optimism	0.33	2	0.3
Skills	0.46	1	0.2
Reinforcement	0	0	0

[a]*P*<.001 for intercoder agreement for all domains of the coding scheme, unless otherwise stated.
[b] Based on majority agreement decision-making approach across ChatGPT coders (i.e., by selecting the mode of the codes per change mechanism).
[c]*P* .511

Discussion

Principal Results

The current study is among the first to use ChatGPT to automate a range of tasks related to qualitative content analysis of web-based data on behavior change. Preparation for the analysis process was done by identifying relevant change mechanisms from forum data on reducing sugar consumption, which we did through ChatGPT with an estimated 88% precision rate. Based on a subset of change mechanisms, ten coding schemes were developed using an inductive approach where categories and category labels were informed by the data. Another ten coding schemes were developed using an unconstrained deductive approach categories reflected relevant domains of the TDF and were relabeled in line with the current data on sugar reduction. Developed coding schemes largely followed prompt instructions but were highly variable in the number of categories across and within approaches. Using ChatGPT to code the full change mechanism dataset into each coding scheme showed moderate to almost perfect inter-coder agreement, where inter-coder agreement of the inductively developed coding schemes was generally superior to the deductively developed coding schemes. A structured deductive approach was also applied by coding directly into the original TDF coding matrix without specifying domains labels to reflect the current dataset. Overall inter-coder agreement for this approach exceeded that for the majority of coding schemes from the unconstrained deductive approach but it was lower than overall agreement observed for the inductively developed coding schemes.

Comparison with other studies

There have been a few exploratory studies using ChatGPT to analyze small datasets with analysis approaches that bear similarities to the inductive and deductive content analysis approaches used in our study. For example, one study used ChatGPT to conduct reflexive thematic analysis on a feature newspaper article about long Covid [35]. Analysis of the text was data driven as ChatGPT was prompted to generate a list of categories featured in the data, which, in contrast to the approach of our study, was then compared to a list of categories independently developed by a researcher. Findings revealed that the lists were largely similar although the ChatGPT-generated list included more focused categories than the broader human-developed categories. Both lists were combined to refine the categories and ChatGPT was queried to confirm fit of the refined categories with the data, reflective of a negotiation step when combining category lists. Other research used ChatGPT to assess problem solving through content analysis of 40 short internet chats in which university students aimed to solve a mathematical issue [33]. Using an inductive approach, the authors prompted ChatGPT to generate a list of categories with description (comparable to coding schemes

in our study) for each chat and provide an overall problem-solving score based on the extent to which categories were featured in the chat. Reliability of ChatGPT output over time was evaluated by repeating exact prompts at a different date in new conversations, which revealed only a moderately positive correlation between the output at different time points. Our study showed more promising results, as we observed substantial to almost perfect inter-coder agreement when evaluating reliability of the inductively developed coding schemes by repeating prompts in conversations across time points and OpenAI accounts.

Aforementioned data from problem-solving chats were also analyzed using a deductive approach [33]. To do so, the authors prompted ChatGPT with questions reflecting categories of a theory-driven coding scheme which had previously been applied by researchers to code the data. Overall problem-solving scores were calculated from the number of categories that were featured in a chat. Prompts were repeated at two time points to calculate intra-coder agreement. The findings showed 90% intra-coder agreement in output generated on separate occasions. Our study similarly evaluated the reliability of ChatGPT output by repeating prompts in conversations across time points and OpenAI accounts, albeit across ten conversations as opposed to two. Interestingly, our findings were more promising for the inductive approach while we observed only moderate to substantial overall agreement for the coding schemes from the unconstrained deductive approach.

The deductive approach with structured coding matrix in our study is somewhat similar to the approach taken in another study which evaluated feedback on a university course [34]. The study used ChatGPT to code 200 student comments directly into a pre-defined list of categories which were based on the literature but not accompanied by a label to define the categories. Researchers subsequently double checked the ChatGPT-annotated data and agreed with 85% of the allocated codes. An alternative double check was performed by prompting ChatGPT to rate how well an allocated category reflected the content of the comment. About 10% of the annotated data received low ratings, suggesting incorrectly coded data. These findings do not necessarily indicate an error in ChatGPT coding rather they highlight specific categories of the coding matrix that might need to be refined.

Our objective was to explore the utility of ChatGPT in conducting qualitative content analysis with an emphasis on diminishing the human workload and time spent on analysis tasks and error proneness related to human fatigue when analyzing large datasets. In line with this, human input in our study was focused on prompt engineering and data preparation, while the inductive and deductive analysis was conducted by prompting ChatGPT with a set of instructions related to those tasks. As such, it was beyond our scope to combine ChatGPT-generated coding schemes with a human-developed coding scheme or to compare ChatGPT-annotated data with human-annotated data. However, as we used secondary data, it is possible to compare findings related to our secondary aim with the change mechanisms identified in the original study which relied solely on manual coding [41]. The original study identified 25 different categories which were organized into four phases of readiness to change. The most frequent categories were similar to the inductive approach categories “Substance substitution”, “Knowledge and information” and “Avoidance”. However, the inductive and deductive approaches most frequently occurring category was related to behavioral regulation which the previous study coded into smaller categories that reflected the studies overarching theory and coding scheme.

Implications and future directions

Our study suggests that inter-coder agreement, particularly when employing the inductive approach, was superior to that achieved through deductive approaches. This finding aligns with the

fundamental principles underlying each approach, as the inductive approach, developed solely on the current dataset, aims to describe its content. This approach is particularly valuable in novel or emerging areas of study or when pre-existing data, concepts, or theories are scarce [3]. In fields such as behavior change, numerous explanatory and predictive models and theories offer insights not only into what is happening but also why. Our study indicates that coding within a model or theory yields higher inter-coder agreement when the coding scheme is tailored to reflect the current dataset. However, adjusting a published coding matrix poses the risk of missing opportunities to advance widely applicable models like the TDF, which has demonstrated universal utility [39, 53]. Moreover, as a framework, the TDF is more explanatory than predictive and some of its categories overlap (e.g., planning appears in two different categories), complicating data analysis methods such as content analysis [54]. Future research may benefit from applying the insights from our study to behavior change theories characterized by more clearly defined categories, such as the Theory of Planned Behavior to further explore ChatGPT utility.

The difference between inter-coder agreement for inductive and deductive approaches might be further explained by prompt engineering and category development employed by ChatGPT. For example, category labels from the inductive approach mostly exceeded the maximum word count specified in the instructions but also commonly included examples, which was not the case for the labels from the deductive unconstrained approach. LLMs process data based on semantic relations in text and as illustrated by literature [55], queries with richer semantic data (e.g., synonyms) increase LLM performance. Examples included in the category labels are likely semantically related to the short descriptions of change mechanisms which might in turn result in more consistent annotation of the data. Future research is warranted to confirm whether coding schemes which include examples increase inter-coder reliability across ChatGPT output.

Our findings underline the importance of including a step in the analysis process where coding schemes are refined based on data that are to be coded, as opposed to using a structured coding matrix. Depending on the topic investigated, certain elements of a theory or framework may be less applicable to the data but indicate new insight into the topic [3]. This may explain why domains that were not featured in the unconstrained coding schemes showed low domain-specific agreement when applying the structured coding matrix. Without the option to code into domains not featured in the data subset the overall inter-coder agreement using this approach may have been better. A noteworthy issue with the structured deductive approach was that despite instructions to only code mechanisms into the best matching domain, there was still considerable data across coding schemes allocated to more than one domain. The issue persisted despite adapting instructions for the task during the prompt engineering phase. This could be an artifact of the underlying TDF knowledge, which in some settings result in an overlap between domains [54]. It might also be that with the lack of domain labels, and thus limited semantic information, instructions are more ambiguous, thereby leaving more room for interpretation and resulting in less consistent output across conversations.

Through a thorough process of prompt engineering, we developed a set of structured prompts which can easily be adapted, expanded, and tailored by other researchers intending to use ChatGPT for qualitative content analysis. The extent to which prompts need adaptation and further prompt engineering depend on the research topic and approach. Replicating this study on a different topic would require an adequate description of that topic which might be limited to a substitution of the topic key words used in the prompts (i.e., “sugar consumption” and “change mechanisms”). Where other theories are selected for the inductive approach, it is advisable to check ChatGPT’s familarity with the theory. Depening on the topic and theory, we suggest specifying instructions with other notes on what to look out for during task execution. The use of appropriate synonyms is also recommended to improve ChatGPT performance [55] as terminology may vary across research

groups, disciplines, and communities. With other versions of ChatGPT, token limit may not be a restriction and we suggest researchers experiment with adding more context to the prompt. To further automate the preparation task, comparison with human coding may be replaced by having ChatGPT rate accuracy of previous output, as done in previous research [34].

The approaches used in this study can be tested on other data types, such as survey responses, interviews, or focus group discussions. However, the nature of the data (e.g., depth versus breadth, level of structuredness) might affect ChatGPT performance on analysis tasks and research is warranted to refine prompt engineering and assess reliability of the output based on other data types. Regardless of the research approach, when using tools like ChatGPT for qualitative analysis, researchers should be aware of ethical considerations to prevent harm, such as transparency, informed consent, data privacy, and potential biases in LLM training data or disseminated results [56, 57]. It is warranted that such issues are considered properly to assure research projects abide by ethical standards.

It should be reiterated that our study focused primarily on reliability of ChatGPT-annotated data without addressing validity of the findings. Assessing the validity of the findings is an important area for future investigation. Validity checks could be incorporated by triangulation of investigators, for instance by having experienced qualitative researchers develop coding schemes and compare these with coding schemes developed using ChatGPT. Alternatively, the prompt with step-by-step instructions which we used to create a theory-driven coding scheme enables a less-time intensive method to check validity of the ChatGPT-generated coding scheme. In this type of prompt, the first part of output facilitates the possibility to assess whether the data underlying each category in the generated coding scheme accurately reflects the categories. Additionally, we encourage others using ChatGPT for qualitative content analysis to experiment with coding rules which may further increase the quality of the output and prevent overlap between categories. It should also be mentioned that we presented frequency data based on the coding schemes with the best inter-coder agreement. However, selection of the most optimal coding scheme may depend on other considerations, including expert reflection of categories and labels generated by ChatGPT, number of categories within a coding scheme, and the need to incorporate certain categories or mix inductive and deductive approaches. Moreover, appropriateness of the used analysis methods and techniques is dependent on the type of topic as well as the character and volume of the available data [19].

Limitations

The colloquial phrase “garbage in, garbage out” can be seen as a general rule in machine learning, meaning that any output retrieved is only as good as the data underlying the output [58]. As we followed a thorough process of prompt engineering and evaluation of the output did not reveal concerning deviations from the prompt instructions, we are relatively confident that the quality of the output was not substantially diminished by the instructions for the tasks. It should be noted though, that sourcing the data to be analyzed from online forums may have implications for the quality of the output. Forum data is user-generated and by nature less focused than data generated though targeted questions (e.g., as is the case with interview and survey data), potentially leading to convoluted data [59]. In our study this issue was partly circumvented by using secondary data which met certain eligibility criteria to ensure they related to lived experiences of people trying to change sugar consumption [41]. Still, using these data to generate a condensed set of change mechanisms did result in the inclusion of non-target data as evident from our estimate of 88% precision in identification of change mechanisms, which we considered acceptable for use in the inductive and deductive analysis approaches.

Whether precision levels should be considered acceptable is dependent on the goal of the study and is often a trade off with other metrics of model performance, including recall and accuracy [60]. When generating a dataset of change mechanisms identified from the forum posts, we focused on precision to optimize the amount of relevant data to be used in creating and applying the coding schemes. Imperfect precision could have led to distorted categories in a coding scheme as instructions specified to group all mechanisms, regardless of relevance, in discreet categories. It may also have caused the creation of a miscellaneous category, seen in various versions of the inductively developed coding schemes. Similarly, coding non-target change mechanisms into the coding schemes may have distorted the frequency data as these were also based on all data. We included instructions to code change mechanisms as not applicable where they did not fit into any category and this data could be further examined to build new knowledge of the TDF and how it applies to a range of different contexts.

Conclusions

AI assistance has the potential to make a massive impact on research involving qualitative content analysis. As demonstrated in the current study, ChatGPT can assist with each phase of the methodology, from condensing data and developing coding schemes to double coding and facilitating consensus meetings. While we have shown that ChatGPT can perform these tasks largely without human oversight, there are many risks associated with this approach, with the key risk being that the output may reflect the worst of ChatGPT's coding, and therefore, the output may not accurately reflect the data or research questions. We recommend that human involvement is necessary, but it could be reduced to just one or two researchers across each phase. Such human involvement would largely ensure that prompt engineering and interpretation remain aligned with the research goals and context. This study provides the foundations for qualitative content analysis with ChatGPT and can be tested and further developed as new AI emerges.

Acknowledgments

Funding for the study was provided by the Health Research Council of New Zealand (grant: 20/733).

The data sets generated and/or analyzed during this study are available from the corresponding author on reasonable request.

The authors have no perceived or actual competing interest in relation to this manuscript.

RB: Conceptualisation (supporting), data curation, formal analysis, investigation, methodology, validation, visualiation, writing original draft preparation and review and editing. SM: formal analysis, investigation, methodology, validation, and review and editing. ND: Review and editing. SR: Conceptualisation (lead), funding acquisition, investigation, project administration, supervision, validation and review and editing.

AI was used as part of the study design as reported in the methods section of this article.

References

1. Krippendorff, K., *Content analysis: An introduction to its methodology*. 2018: Sage publications.
2. Weber, R.P., *Basic content analysis*. Vol. 49. 1990: Sage.
3. Elo, S. and H. Kyngas, *The qualitative content analysis process.* J Adv Nurs, 2008. **62**(1): p. 107-15.
4. Hsieh, H.-F. and S.E. Shannon, *Three approaches to qualitative content analysis.* Qualitative health research, 2005. **15**(9): p. 1277-1288.
5. Fu, J., et al., *Methods for analyzing the contents of social media for health care: scoping review.* Journal of Medical Internet Research, 2023. **25**: p. e43349.
6. Kleinheksel, A., et al., *Demystifying content analysis.* American journal of pharmaceutical education, 2020. **84**(1): p. 7113.
7. Mayring, P. *Qualitative content analysis: Demarcation, varieties, developments*. in *Forum: Qualitative Social Research*. 2019. Freie Universität Berlin.
8. Mayring, P., *Qualitative Content Analysis.* Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 2000. **1**(2).
9. Pope, C., S. Ziebland, and N. Mays, *Qualitative Research in Health Care: Analysing Qualitative Data.* BMJ: British Medical Journal, 2000. **320**(7227): p. 114-116.
10. Lindgren, B.-M., B. Lundman, and U.H. Graneheim, *Abstraction and interpretation during the qualitative content analysis process.* International journal of nursing studies, 2020. **108**: p. 103632.
11. Hahn, C., *Doing qualitative research using your computer: A practical guide.* Doing Qualitative Research Using Your Computer, 2008: p. 1-232.
12. Graneheim, U.H. and B. Lundman, *Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness.* Nurse Educ Today, 2004. **24**(2): p. 105-12.
13. Bengtsson, M., *How to plan and perform a qualitative study using content analysis.* NursingPlus open, 2016. **2**: p. 8-14.
14. Zamawe, F.C., *The implication of using NVivo software in qualitative data analysis: Evidence-based reflections.* Malawi Medical Journal, 2015. **27**(1): p. 13-15.
15. Brooks, J., et al., *The Utility of Template Analysis in Qualitative Psychology Research.* Qual Res Psychol, 2015. **12**(2): p. 202-222.
16. Haney, W., et al., *Drawing on Education: Using Student Drawings To Promote Middle School Improvement.* Schools in the Middle, 1998. **7**(3): p. 38-43.
17. Gale, N.K., et al., *Using the framework method for the analysis of qualitative data in multi-disciplinary health research.* BMC medical research methodology, 2013. **13**: p. 1-8.
18. Young, I.J.B., S. Luz, and N. Lone, *A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis.* Int J Med Inform, 2019. **132**: p. 103971.
19. Karmegam, D., T. Ramamoorthy, and B. Mappillairajan, *A systematic review of techniques employed for determining mental health using social media in psychological surveillance during disasters.* Disaster medicine and public health preparedness, 2020. **14**(2): p. 265-272.
20. Neale, J., *Iterative categorization (IC): a systematic technique for analysing qualitative data.* Addiction, 2016. **111**(6): p. 1096-1106.
21. O'Connor, C. and H. Joffe, *Intercoder reliability in qualitative research: debates and practical*

guidelines. International journal of qualitative methods, 2020. **19**: p. 1609406919899220.
22. Cohen, J., *A coefficient of agreement for nominal scales*. Educational and psychological measurement, 1960. **20**(1): p. 37-46.
23. Joffe, H. and L. Yardley, *Chapter four: content and thematic analysis*. Research Methods for Clinical and Health Psychology. Marks D, Yardley L (ed): Sage Publications, London, 2003: p. 56-68.
24. Mayring, P., *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. 2014.
25. Lund, B.D. and T. Wang, *Chatting about ChatGPT: how may AI and GPT impact academia and libraries?* Library Hi Tech News, 2023. **40**(3): p. 26-29.
26. Raman, R., et al., *Early Research Trends on ChatGPT: Insights from Altmetrics and Science Mapping Analysis*. International Journal of Emerging Technologies in Learning (iJET), 2023. **18**(19): p. 13-31.
27. OpenAI, *Introducing ChatGPT*. 2022.
28. Hutson, M., *Could AI help you to write your next paper?* Nature, 2022. **611**(3 November 2022): p. 192-193.
29. Nature, *Tools such as ChatGPT threaten transparent science; here are our ground rules for their use*. Nature, 2023. **613** (26 January 2023): p. 612.
30. Gao, C.A., et al., *Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers*. NPJ Digit Med, 2023. **6**(1): p. 75.
31. Lund, B.D., et al., *ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing*. Journal of the Association for Information Science and Technology, 2023. **74**(5): p. 570-581.
32. Moskatel, L.S. and N. Zhang, *The utility of ChatGPT in the assessment of literature on the prevention of migraine: An observational, qualitative study*. Frontiers in Neurology, 2023. **14**: p. 1225223.
33. Siiman, L.A., et al. *Opportunities and Challenges for AI-Assisted Qualitative Data Analysis: An Example from Collaborative Problem-Solving Discourse Data*. in *International Conference on Innovative Technologies and Learning*. 2023. Springer.
34. Katz, A., et al., *Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy*. arXiv preprint arXiv:2305.11882, 2023.
35. Hitch, D., *Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future?* Qualitative Health Research, 2023: p. 10497323231217392.
36. Tabone, W. and J. de Winter, *Using ChatGPT for human–computer interaction research: a primer*. Royal Society Open Science, 2023. **10**(9): p. 231053.
37. Neelakantan, A., et al., *Introducing text and code embeddings*, in *OpenAI*. 2022.
38. Zhang, H., et al., *Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis*. arXiv preprint arXiv:2309.10771, 2023.
39. Cane, J., D. O'Connor, and S. Michie, *Validation of the theoretical domains framework for use in behaviour change and implementation research*. Implementation science, 2012. **7**: p. 1-17.
40. Brockman, G., et al., *Introducing ChatGPT and Whisper APIs*. 2023.
41. Rodda, S.N., et al., *I was truly addicted to sugar: A consumer-focused classification system of behaviour change strategies for sugar reduction*. Appetite, 2020. **144**: p. 104456.
42. Auckland University of Technology. *Exceptions to Activities requiring AUTEC approval (6). Applying for ethics approval: guidelines and procedures*. May 3, 2024]; Available from: https://www.aut.ac.nz/research/researchethics/guidelines-and-procedures#6.

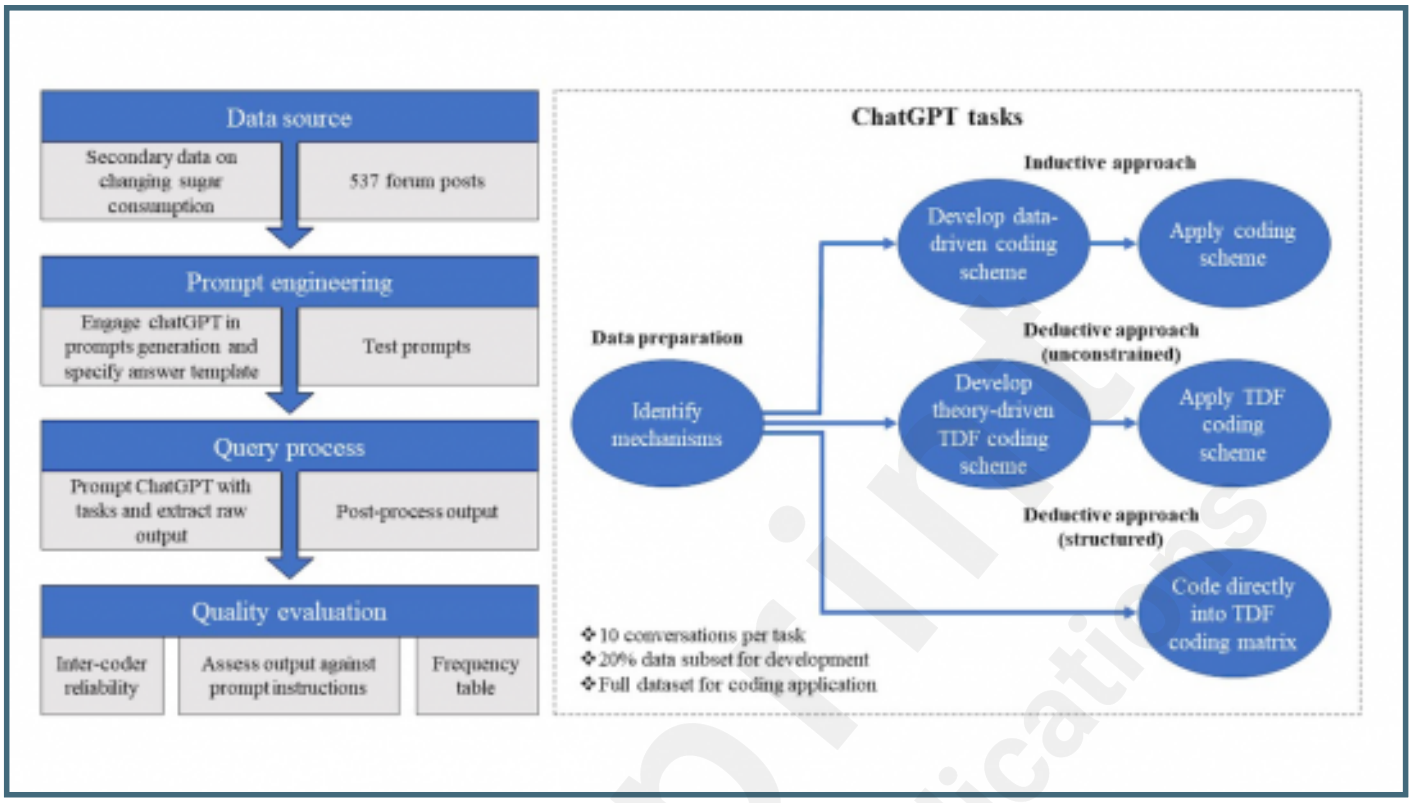
43. Lyu, Q., et al., *Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential*. Visual Computing for Industry, Biomedicine, and Art, 2023. **6**(1): p. 9.
44. Chen, S., et al., *Evaluation of ChatGPT family of models for biomedical reasoning and classification*. arXiv preprint arXiv:2304.02496, 2023.
45. Gao, C.A., et al., *Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers*. NPJ digital medicine, 2023. **6**(1): p. 75.
46. Giray, L., *Prompt Engineering with ChatGPT: A Guide for Academic Writers*. Annals of Biomedical Engineering, 2023.
47. White, J., et al., *A prompt pattern catalog to enhance prompt engineering with chatgpt*. arXiv preprint arXiv:2302.11382, 2023.
48. Ekin, S., *Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices*. Authorea Preprints, 2023.
49. Wei, J., et al., *Chain-of-thought prompting elicits reasoning in large language models*. Advances in Neural Information Processing Systems, 2022. **35**: p. 24824-24837.
50. Kocoń, J., et al., *ChatGPT: Jack of all trades, master of none*. Information Fusion, 2023. **99**.
51. Fleiss, J.L., B. Levin, and M.C. Paik, *Statistical methods for rates and proportions*. 2013: john wiley & sons.
52. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data*. biometrics, 1977: p. 159-174.
53. Francis, J.J., D. O'Connor, and J. Curran, *Theories of behaviour change synthesised into a set of theoretical groupings: introducing a thematic series on the theoretical domains framework*. Implementation Science, 2012. **7**: p. 1-9.
54. Phillips, C.J., et al., *Experiences of using the Theoretical Domains Framework across diverse clinical environments: a qualitative study*. Journal of multidisciplinary healthcare, 2015: p. 139-146.
55. Claveau, V. *Neural text generation for query expansion in information retrieval*. in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2021.
56. Ray, P.P., *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*. Internet of Things and Cyber-Physical Systems, 2023. **3**: p. 121-154.
57. Wang, C., et al., *Ethical Considerations of Using ChatGPT in Health Care*. J Med Internet Res, 2023. **25**: p. e48009.
58. Beam, A.L. and I.S. Kohane, *Big data and machine learning in health care*. Jama, 2018. **319**(13): p. 1317-1318.
59. McKenna, B., M.D. Myers, and M. Newman, *Social media in qualitative research: Challenges and recommendations*. Information and Organization, 2017. **27**(2): p. 87-99.
60. Kim, Y., J. Huang, and S. Emery, *Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection*. Journal of medical Internet research, 2016. **18**(2): p. e41.

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/f151d50c8dac6ae5f988669c6aa885a1.doc

Figures

Study Flow.



Multimedia Appendixes

Final prompts.
URL: http://asset.jmir.pub/assets/a04b2f02c77e971215b249e759da30ea.docx

Inductively developed coding schemes and meta data.
URL: http://asset.jmir.pub/assets/4730c7c7fb45ba6b6b4c8a21dc9e5a7dd9.docx

Domain-specific kappa scores per TDF coding scheme.
URL: http://asset.jmir.pub/assets/d2692d7d5d930289903652b5aeb56586.docx

Deductively developed coding schemes and meta data.
URL: http://asset.jmir.pub/assets/c9273e3f539f4cd524e6f6a8a264aeb1.docx