Prompting strategies for coding in AI-assisted qualitative data analysis

Prompting-Strategien für das Codieren in der KI-gestützten qualitativen Datenanalyse

Mireya Diaz[1]

1 Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, United States

Abstract

Introduction: Coding is a central step in many qualitative data analyses (QDA). Artificial intelligence (AI) has made inroads into the field with performance dependent on the prompting strategy used. This work examines the effectiveness of different prompting strategies on AI performance.
Methods: Coding and theme analyses were performed on abstracts from a 16-study meta-analysis examining the association between air pollution and preeclampsia incidence. These were also performed in a second related corpus containing fewer formal sources. Coding was performed with R, NVivo 14, and ChatGPT 4.1. Several prompting strategies in increasing degree of complexity (zero-shot, few-shots, and chain-of-thought – CoT) were evaluated regarding their accuracy against manual coding. Direct thematic coding with AI was also examined.
Results: Zero-shot prompting (with or without CoT) provided the best results for structured text. Requesting a full word-by-abstract matrix increases sensitivity. Supplementing CoT with general topic information and role playing made it the best strategy for more colloquial text. Directly inquiring thematic coding within AI showed good performance. Some prompting strategies are more prone to hallucinations. The latter are associated with errors in the frequency counting process while coding.
Discussion: This work offers performance assessment of an AI tool for coding and theme extraction through a variety of prompting strategies combined with head-to-head comparison with popular QDA tools using readily available data. During user interaction, AI chatbots provide suggestions regarding potential next steps which may enhance results during coding. AI tools may require replication to ensure consistent findings and mitigate their probabilistic nature. Integrating results from at least two tools (AI and traditional) provides more comprehensive and accessible results for QDA.

Keywords: qualitative data analysis, thematic coding, artificial intelligence, ChatGPT, prompting engineering, zero-shot, few-shots, chain-of-thought

Zusammenfassung

Einleitung: Die Kodierung ist ein zentraler Schritt bei vielen qualitativen Datenanalysen (QDA). Künstliche Intelligenz (KI) hat in diesen Bereich Einzug gehalten, wobei die Leistungsfähigkeit stark von der verwendeten Prompting-Strategie abhängt. Diese Arbeit untersucht die Wirksamkeit verschiedener Prompting-Strategien auf die Leistung der KI.
Methoden: Die Kodierung und Themenentwicklung erfolgten anhand von Abstracts aus einer Metaanalyse von 16 Studien, die den Zusammenhang zwischen Luftverschmutzung und der Inzidenz von Präeklampsie untersuchten. Zudem wurde ein zweiter, thematisch verwandter Datensatz analysiert, der weniger formelle Quellen enthielt. Die Kodie-

rung wurde mit R, NVivo 14 und ChatGPT 4.1 durchgeführt. Verschiedene Prompting-Strategien mit zunehmendem Komplexitätsgrad (Zero-Shot, Few-Shot und Chain-of-Thought – CoT) wurden hinsichtlich ihrer Genauigkeit im Vergleich zur manuellen Kodierung bewertet. Auch die direkte thematische Kodierung mittels KI wurde untersucht.
Ergebnisse: Zero-Shot-Prompting (mit oder ohne CoT) lieferte bei strukturierten Texten die besten Ergebnisse. Die Anforderung einer vollständigen Matrix (Wort-Abstract-Zuordnung) erhöhte die Sensitivität. Die Ergänzung von CoT durch allgemeine Themeninformationen und Rollenspiele machte diese Methode zur besten Strategie für eher umgangssprachliche Texte. Auch die direkte Abfrage thematischer Kodierungen durch die KI zeigte gute Ergebnisse. Einige Prompting-Strategien sind anfälliger für Halluzinationen; letztere stehen im Zusammenhang mit Fehlern bei der Häufigkeitszählung während des Kodierprozesses.
Diskussion: Diese Arbeit bietet eine Leistungsbewertung eines KI-Tools für Kodierung und Themenentwicklung unter Verwendung verschiedener Prompting-Strategien sowie einen direkten Vergleich mit gängigen QDA-Tools auf Basis leicht verfügbarer Daten. Bei der Interaktion mit Nutzern liefern KI-Chatbots Vorschläge für mögliche nächste Schritte, was die Ergebnisse der Kodierung verbessern kann. KI-Tools erfordern möglicherweise Replikationen, um konsistente Ergebnisse zu gewährleisten und ihre probabilistische Natur auszugleichen. Die Zusammenführung der Ergebnisse aus mindestens zwei Tools (KI und traditionelle Methoden) liefert umfassendere und besser zugängliche Ergebnisse für die QDA.

Schlüsselwörter: qualitative Datenanalyse, thematische Kodierung, Künstliche Intelligenz, ChatGPT, Prompt Engineering, Zero-Shot, Few-Shot, Chain-of-Thought

Introduction

Artificial intelligence (AI) has paved its way into almost every human activity. The text mining capabilities inherent to AI make certain processes within qualitative data analysis (QDA), such as coding, a perfect scenario where the transition in analysis assisted by AI can take place. In this setting, AI can simplify considerably the work leaving more time for the analyst to focus on interpretation and thinking.
Recent work [1], [2], [3], [4], [5] has assessed the integration of generative AI in QDA, particularly in the role of automating the coding process. These papers show that AI performs in alignment with the human analyst; that such performance is highly dependent on the instructions or prompts given to the tool, it is susceptible to errors in quotations, code naming, and hallucination; and that we should see it as a collaborative agent rather than a substitute for the human analyst. Still, there is need for guidance on which prompts will lead to the desired results [6] and how to best control its errors.
This paper addresses the performance of prompting in the key step of coding starting with keyword identification via word-frequency analysis, and the contribution of AI to this step compared to other traditional tools. Also, it assesses the feasibility of AI conducting thematic coding directly. These were examined in both highly structured (scientific abstracts) and less structured (interviews, opinions) text. Ultimately it seeks to fill the gap in guidance about prompting in these activities as noted by previous literature.

Methods

Raw abstracts from a set of 16 studies used in a meta-analysis of air pollution and incidence of preeclampsia were used as the text for analysis [7]. Selection of a corpus with data from public sources avoid ethical issues (privacy, informed consent) present in QDA. Abstracts were ordered by publication date to follow the availability of information as well as to determine the abstract number at which saturation was reached if present. Words identified as highly prevalent (in nine abstracts or more, equivalent to more than half of them) across the abstracts by three tools (R, NVivo and ChatGPT 4.1) were calculated. A second set of less prevalent but still prevalent words appearing in five to eight abstracts (representing the 2nd quartile of the size distribution) were ascertained as means of identifying potential themes appearing in a smaller group of abstracts. Words appearing at the top of this group (i.e., seven or eight abstracts) were selected to estimate specificity. Words related via stemming to words in the highly prevalent group were excluded from the specificity assessment, such as pollutant. Accuracy (sensitivity and specificity against manual coding assisted by R) was calculated for ChatGPT 4.1 [8] prompting strategies.

Coding with R

The popular statistical language and software environment R [9] provides a wide range of text mining functions through tidy tools using the packages tidyverse, tidytext, tibble, and dplyr. Tidy data are formatted into a column for each variable, a row for each observation, and arranged into a table. Each sentence is a unit of observation. To process the abstracts: text was tokenized into words, stop words (i.e., common words such as articles, prepositions, conjunctions, etc. that do not carry significant meaning) were removed, and frequency of occurrence of remaining words was calculated. Codes were then generated manually using the most prevalent words across abstracts (labeled as Manual Coding). Words identified by R were used as reference standard. Sensitivity was estimated as the number of true positive (i.e. words appearing in more than half of the abstracts) divided by all the highly prevalent words and transformed to a percentage. Specificity was estimated by the number of true negatives (i.e. words less prevalent) divided by the moderately prevalent words (appearing in seven or eight abstracts) and transformed to a percentage.

Coding with NVivo

NVivo was used to validate the word extraction and manual coding performed with the assistance of R. Two alternative word frequency options for exploration in NVivo [10] were assessed. These are exact match with the 50 most frequent words excluding stop words and with a minimum length of two; and a second option which included the same conditions but used stemming rather than exact matching. The latter offers a comparison point for results obtained by traditional tools and AI which incorporates stemming in its processing.

Artificial intelligence coding through ChatGPT 4.1

ChatGPT 4.1 is the fourth generation in the large language model chatbot. Success of AI in the assigned task is highly dependent on the prompting strategy employed. This has been identified in previous work of ChatGPT within thematic analysis [11]. Prompts are textual inputs to the generative AI tool. These allow providing the tool with questions, instructions, and/or introducing the context so the tool can perform an appropriate job. The idea is to provide the tool with specificity, context, and detail to obtain better or desired outputs. A four-component framework [12] will set the tone of the conversation between the tool and the user to make it effective. These components are: persona (define the AI's identity, expertise); task (specify the function(s) to perform); context (provides tone and constraints); and format (defines the output's appearance). Different prompting strategies use these components in different degrees. The main characteristics of the three types of prompting strategies that we will examine in this work are described below. For all these three strategies, the persona defined sets the depth of the response.

1. **Zero-shot prompting:** it corresponds to the simplest form of prompting. This strategy lacks examples. It basically contains the task and the context.
2. **Few-shots prompting:** in addition to providing the task request (and context), the prompting includes examples of how the user would like the tool "to think" (including formatting cues).
3. **Chain-of-thought (CoT) prompting:** it was developed for complex tasks, such as numerical problems, symbolic reasoning, logic analysis [13]. The key is to decompose the complex task in a series of steps. These are provided to the tool as a series of instructions thus the tool can think step-by-step and complete the task effectively.

This paper focuses on the coding task key to many types of QDA, including thematic analysis. Coding corresponds to the second step of the six-step approach to thematic analysis [14]. ChatGPT has been previously used with this purpose. In fact, Naeem et al. [15] explored its efficiency compared to manual work. Two steps of this approach will be examined here: keyword selection and coding. For keyword selection AI will select the keywords from the data based on specific prompts provided to it. Within the six Rs framework to select keywords, we will emphasize repetition and richness.

In this work we evaluate the accuracy of word-frequency counting preceding coding by AI using different prompting strategies as summarized in Table 1. This general prompting strategy was performed sequentially in three stages with increasing complexity. The most successful of the prompts in stage one (zero-shot) was selected to be the skeleton for the prompts in stage two (few-shots and role playing), and the most successful prompt from stage two as the skeleton for prompts in stage three (CoT). Also, during the interaction with the chatbot it provided information worth considering within the strategy and thus the strategy was modified adding this information within iterations (see for example Strategy 1 in Table 1 under "added"). Each strategy is run in a new chat. For the different strategies we selected no internet access for the ChatGPT engine. This means that its web search capabilities were disabled, looking for eliciting its language synthesis functionality more than its language mimicking one. Two separate runs for each of the different prompting strategies were performed to assess consistency of findings.

Thematic coding agreement between ChatGPT 4.1 and manual coding for Strategy 8 was assessed via consensus by identifying the overlap between both sets of codes. For thematic coding agreement between replicate runs it was estimated as the percent of abstracts showing an exact match in which the codes appeared divided by all identified codes; either both replicates identified the code or both replicates did not identify the code to be counted as an agreement.

Table 1: Prompts to perform coding with ChatGPT 4.1 without internet access

Strategy 1: Zero-shot prompt 1 *Iteration 1 (S1a)*: Read the following 16 abstracts. Identify words that appear in 9 or more of these abstracts. Indicate how many times the selected words appear in total. (Based on output from the chatbot the following iterations took place.) *Iteration 2 (S1b, added)*: Ignore stop words. *Iteration 3 (S1c, added)*: Perform stemming. *Iteration 4 (S1d, added)*: Provide full word-by-abstract matrix. *Iteration 5 (S1e, added)*: Provide untruncated full word-by-abstract matrix in a CSV file.
Strategy 2: Zero-shot prompt 2 providing general underlying topic Read the following 16 abstracts about the association between ambient air pollution and preeclampsia. Identify words that appear in 9 or more abstracts. Indicate how many times the selected words appear in total. Ignore stop words. Provide full word-by-abstract matrix.
Strategy 3: Few-shot prompt – role playing for expert qualitative data analyst Iteration 1 (S3a): Act as an expert qualitative data analyst. Read the following 16 abstracts. Identify words that appear in 9 or more abstracts. Indicate how many times the selected words appear in total. For example: the word air appears in 15 of the 16 abstracts. It appears in each of the abstracts in csv format as follows: 5, 6, 2, 3, 0, 7, 3, 1, 2, 1, 5, 3, 3, 3, 1, 4. Thus, the word air appears a total of 49 times. Ignore stop words. *Iteration 2 (S3b, added)*: Perform a breakdown of more words.
Strategy 4: Zero-shot Chain-of-thought prompt *Iteration 1 (S4a)*: Read the following 16 abstracts. Perform the following steps using these abstracts. Remove all stop words. Normalize word forms considering synonyms and variants. For each remaining word, count its frequency in each abstract. Identify all words that appear in 9 or more abstracts. For each such word, provide: the number of abstracts in which it appears. A CSV style list of its counts per abstract. The total number of appearances across all abstracts. Format the output as a table and as CSV lines. *Iteration 2: (S4b, added)*: Please provide the full word frequency matrix
Strategy 5: Zero-shot Chain-of-thought prompt with general topic and role playing Act as an expert qualitative data analyst. Read the following 16 abstracts about the association between ambient air pollution and preeclampsia. Perform the following steps using these abstracts. Remove all stop words. Normalize word forms considering synonyms and variants. For each remaining word, count its frequency in each abstract. Identify all words that appear in 9 or more abstracts. For each such word, provide: the number of abstracts in which it appears. A CSV style list of its counts per abstract. The total number of appearances across all abstracts. Format the output as a table and as CSV lines.
Strategy 6: Few-shots Chain-of-thought prompt *Iteration 1 (S6a)*: Act as an expert qualitative data analyst. Read the following 16 abstracts about the association between ambient air pollution and preeclampsia. Perform the following steps using these abstracts. Remove all stop words (common English words such as “the”, “and”, “of”). Normalize word forms considering synonyms and variants (e.g., lowercase, unify variants such as “pre-eclampsia” and “preeclampsia”, singular and plural as “association” and “associations”, “NO2” and “nitrogen dioxide”). For each remaining word, count its frequency in each abstract. Identify all words that appear in 9 or more abstracts. For each such word, provide: the number of abstracts in which it appears. A CSV style list of its counts per abstract. The total number of appearances across all abstracts. Format the output as a table and as CSV lines (example: air,5,6,2,3,0,7,3,1,2,1,5,3,3,3,1, 4,49). *Iteration 2 (S6b, added)*: Please provide the full word frequency matrix
Strategy 7: Few shots Chain-of-though prompt (each step one at a time) Steps 1 through 5 as in Strategy 6 but provided one at a time rather than all at once.
Strategy 8: Request thematic coding directly *Iteration 1 (S8a)*: Act as an expert qualitative data analyst. Read the following sixteen abstracts. Perform thematic coding with them. (All 16 abstracts were submitted together) *Iteration 2 (S8b, added)*: 16 abstracts adding one at a time and indicating that in the instruction “Read the following xxx abstracts.”

Assessing generalizability

A set of 12 narratives was collected from the internet covering the words air pollution and pre-eclampsia after a Google search using these terms, to assess generalizability of the findings when dealing with less formal and structured QDA data sources such as interviews, opinions, etc. Further search within the identified sites including newspapers, news channels, and YouTube was performed with the terms. The main selection criterion was that the narrative dealt with both terms. If it only dealt with one, it was discarded. The narratives identified are a mix of online newsletters commenting on journal articles [16], [17], [18], [19], [20], [21], video posts and TV news seg-

ments uploaded into YouTube [22], [23], [24], [25], [26]. Videos were transcribed with the assistance of closed captions. Narratives were anonymized prior to analysis. After this pre-processing, the narratives were subjected to the same steps as the abstracts, that is ordering by publication or airing day, manual coding assisted with R validated with NVivo and then, ran through different prompts in ChatGPT 4.1. For the latter only the subset with best results observed for the abstracts were assessed. Accuracy (sensitivity and specificity) in selection of prevalent words (i.e., in more than half of narratives, and just close to half of the narratives respectively) were also used as the performance measure for coding prompts. Coverage of codes with respect to manual coding was used to assess direct request of thematic coding as done with abstracts. Two independent runs of the strategies were executed.

Results

R text mining (word-frequency counting and manual coding) results

R identified 27 words appearing in nine or more abstracts. Two additional terms representing numbers 95 and 10 are present in nine or more abstracts but are excluded from these 27 words given their vagueness. In our case they are part of "95% confidence interval" and "PM10" (a pollutant). However, they need other terms to have meaning.

Table 2 illustrates that information in these abstracts can be categorized into nine codes. All these codes can be identified (using words appearing in more than half of the abstracts, and seven of them are well represented (Methods and Features are poorly represented). All the codes can be well identified using words appearing in more than a quarter of the abstracts (i.e., five or more abstracts).

NVivo 14 results

Only two words were different in frequency between R text mining code and NVivo 14 using exact matching of words. Five words detected as prevalent in nine or more abstracts by R did not make the cut among the 50 more prevalent words within NVivo. Using the stemming feature in NVivo 14, identifies 12 words in which more instances of the word are encountered by NVivo 14 than by R, which continues doing exact matching. In these 12 instances an average of 8.8 more words in overall is identified by NVivo 14. In general, NVivo 14 registers where in the text the words appear. However, these should be searched one by one rather than appearing in a summary table as it is possible with both R and ChatGPT. Another matching strategy available in NVivo 14 is matching with generalizations. Although not used in this work, this strategy applies relationships among words that can help identifying potential codes.

ChatGPT 4.1 results

Figure 1 shows a receiving operating characteristic (ROC) curve summarizing the accuracy results obtained with the different prompting strategies to ChatGPT 4.1 for the first run (gray circles), the second run (white squares), and their average (dark gray diamonds). Depending on the strategy, ChatGPT 4.1 identifies between one quarter to three quarters of the words in the set of 27 highly prevalent words for the first run, but only 44% in the second run. Strategies S1e (zero shot with full word by abstract matrix) and S4b (zero shot chain-of-thought with full word by abstract matrix) are the ones with the greatest sensitivity for the first run. When both runs were averaged, sensitivity was below 70% for all strategies.

There are words counted in more abstracts than in which they effectively appear, an average of two abstracts more. This corresponds to the phenomenon known as hallucination, and the different prompting strategies aimed at

Table 2: Coding based on complete abstracts and exemplary words found in at least nine abstracts and additional prevalent words appearing in fewer abstracts

Manual coding	Words appearing in at least nine abstracts	Words appearing in at least five abstracts†
Exposure	Nitrogen, particulate, matter, exposure, air, pollution	Pollutants (8), carbon (7), dioxide (7), PM2.5 (7), oxides (5), ozone (5)
Setting	Residential*, ambient	Hospitals (5)
Data source	Data, study	Records (7), medical (6)
Health condition	Preeclampsia	Hypertension (8)
Methods	Models*, estimated	Regression (8), logistic (6), land (5), use (5), estimate* (8), cohort (7)
Measures	CI, risk, adjusted, association, effect*	Odds (7), confidence (6), interval (6), ratios (6), levels (8)
Features	Trimester	Singleton (7), age (5), preterm (5)
Host	Women, pregnancy, maternal, delivery, gestational, birth	
Findings	Increased, increases, association	Outcomes (8), elevated (5), significant (5), observed (8)

* Words identified in at least nine abstracts via stemming. † (Total number of abstracts).

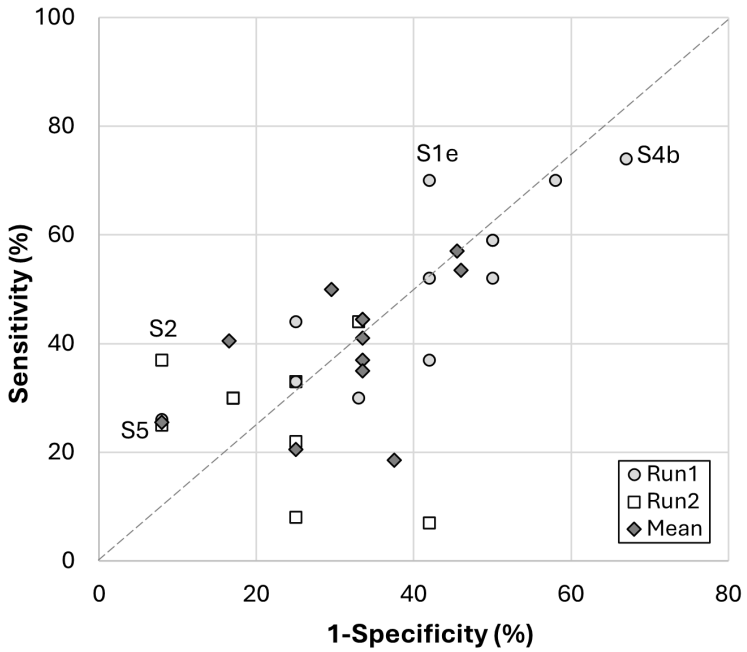

Figure 1: ROC curve for prompting strategies applied to structured abstracts

improving the recognition of the words by the engine trying to avoid these hallucinations, not necessarily successfully in all the strategies though. Hallucinations also manifest in identifying additional words that are not as prevalent – not at least at the level of nine or more abstracts. These correspond to false positives under the condition of highly prevalent words. Twelve words were under this condition and were selected to assess the specificity of the tool with CoT (S5) showing better performance in this measure.

Table 3 illustrates the coding generated via manual coding with information from the keyword extracted in R and in NVivo. It also shows the coding generated by ChatGPT 4.1 when asked directly to perform thematic coding without the process of keyword extraction (Strategy 8). Both processes generate nine codes which explain the information contained in the 16 abstracts. Although they contain the same number of codes, they do not necessarily match on a one-to-one fashion. For example, "Data source" and "Measures," and to a lesser extent "Host," are codes identified from manual coding while these are not selected by ChatGPT 4.1. What manual coding identified as "Methods," ChatGPT was more specific as to whether these methods corresponded to measure the exposure or to perform the analysis, a valid distinction. Likewise, the two manual codes "Features" and "Findings," ChatGPT 4.1 differentiates them into "Effect modification" and "Exposure timing," and into "Association strength" and "Confounding adjustment" respectively.

Table 4 summarizes the themes identified by ChatGPT 4.1 using strategies 8a and 8b for the first run (Total and S8a), for the second run (Total-2 and S8a-2), and the percent agreement in terms of abstracts in which themes are identified or not by both runs (%Agree). When we alter Strategy 8a by adding one abstract at a time (Strategy 8b) rather than providing them all at once, we can identify how the different codes emerge sequentially. We can observe in Table 4 that codes identified in Strategy 8a correspond to those which appear in seven or more abstracts except for the notion of "air pollution" which although it appears in the sixteen abstracts it is not indicated in a general context, but it is represented by the specific pollutants studied. Table 4 also indicates that saturation has not been reached yet because two codes emerge from the four last abstracts. These are "Joint effects and multiple exposures," and "Built and natural environment." We could embed them within a more general code "Environmental exposure" or simply "Exposure" as suggested by manual coding. Another issue that Strategy 8b reveals is a certain degree of variability in the results obtained via ChatGPT 4.1. For example, "Confounding adjustment" is identified as a code in 12 abstracts by Strategy 8a but only in seven by Strategy 8b. Likewise "Association strength" is identified in 16 abstracts by Strategy 8a but only in 12 in Strategy 8b under "Magnitude & relevance/consistency of effects".

Agreement between the two runs for strategy S8b varied between 44% and 100%. "Pre-eclampsia as main outcomes" appears as a main theme separately from other adverse outcomes in the second run. This was much less marked in the first run, corresponding to the lowest level of agreement at 44%. Another theme that differed somewhat between the two sets of runs is that of "Quantification of risk". In the second run, it gave priority to a "Dose-response/exposure gradient" while it was "Quantification of risk" per se in the first run.

For the narratives, the subset with best results observed for the abstracts (S1e, S4a, S4b, S5, S7b) were assessed. Thirteen words appeared in seven or more abstracts and formed the set to estimate sensitivity while 10 words appeared in five or six abstracts and formed the set to

Table 3: Codes generated by direct thematic coding from AI (Strategy 8a) compared to those generated by manual coding of prevalent words

Manual coding	Abstracts	ChatGPT 4.1 coding	Abstracts
Exposure	16	Pollutant type	16
Setting	14	Geography context	15
Data source	10		0
Health condition	16	Outcome type	16
Methods	13	Exposure assessment methods Methodological issues	13 2
Measures	14		0
Features	12	Effect modification Exposure timing	2 10
Host	15	Effect modification	2
Findings	16	Association strength Confounding adjustment	16 12

Table 4: Summary of thematic coding using Strategy 8b

ChatGPT 4.1 code	First abstract	Total	S8a	Total-2	S8a-2	% agree	Manual coding
Pre-eclampsia as main outcome	1	3		12	X	44	Health condition
(Other) maternal (adverse pregnancy) outcomes	1	16	X	14	X	88	Health condition
Environmental exposure/ air pollution	1	16		16	X	100	Exposure
Exposure assessment methods	1	16	X	16	X	100	Methods
Quantification of risk	1	3		7	X	63	Findings
Geographic & demographic context	1	9	X	6		69	Setting
Contributions to existing evidence	1	1		0		94	Findings
Temporal considerations/ trends	2	14	X	15	X	94	Features
Methodological considerations & limitations	2	7	X	3		63	Methods
Public health/ policy implications	2	3		5		50	Findings
Specific pollutants	3	14	X	14	X	75	Exposure
CV effects during pregnancy	3	4		11		56	Health condition
Magnitude & relevance/ consistency of effects	3	12	X	13	X	81	Findings
Adjustment for confounding	4	7	X	9	X	63	Findings
Effect modification & population subgroups	6	11	X	11	X	100	Host
Research gaps, further studies	9	2		0		88	
Joint effects & multiple exposures	13	1		2		94	Exposure
Built & natural environment	15	2		3		94	Setting
Dose-response/ exposure gradient	2 (run 2)	0		5		69	Findings

First abstract indicates the abstract in which the code emerged for the first time. Total column summarizes the number of abstracts identifying the corresponding code for strategy S8b (run 1), and Total-2 for run 2. Column S8a indicates whether the code was also identified with S8a in the first run, or in the second run (S8a-2). %agree indicates the percent of abstracts in agreement between the two runs in which the theme was identified within strategy S8b.

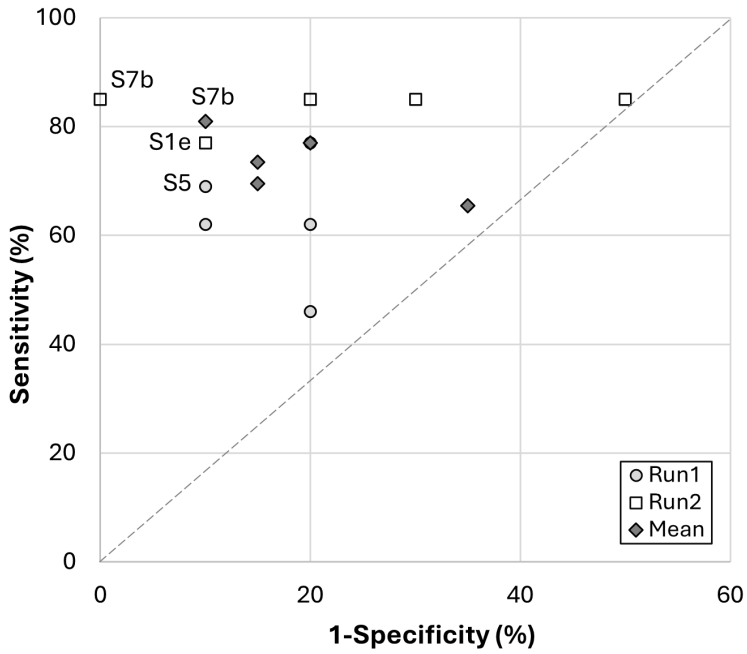

Figure 2: ROC for prompting strategies applied to narratives

estimate specificity. Sensitivity ranged between 46% to 77% with S7b showing the greatest sensitivity, and S5 the second best for the first run (Figure 2). It ranged between 77% and 85% for the second run. Specificity ranged between 80% and 90% with S7b exhibiting the smallest specificity and S5 and S4b the highest one for the first run. It ranged between 50% and 100% for the second run. These results contrast with those from the abstracts. For the abstracts S4b offered the best compromise between sensitivity (though poor) and specificity between the two runs, while for the narratives S7b was consistently the best performing prompting strategy followed by S5, and then S4b.

Five themes (Health condition, Exposure, Findings, Host, and Pollutant separate from Exposure) appeared using the most prevalent words approach in manual coding. A sixth theme (Methods) appeared using less prevalent words. All these themes are shared with the abstracts. When prompted directly for thematic coding, ChatGPT 4.1 also identified these six themes appearing in a more elaborate fashion and with sub-themes. In terms of agreement between the two runs, 15 codes emerged, with 8 codes overlapping (53%), one specific to the first run (Research Gaps and Limitations), and six codes specific to the second run, including “Noise Pollution”, “Epidemiologic Evidence”, and “Causality Caution”.

Discussion

Data coding, a key step in many QDA, is a perfect activity to experiment and become familiar with AI tools within the analytic classroom. We examined its performance and compared it with more traditional tools such as text analysis through the popular R and NVivo. Selecting abstracts from meta-analyses with a moderate number of studies allows examining these tools and learning the process of data coding without confronting issues of data privacy and informed consent.

Data coding with R and NVivo 14 was done through a process of identifying prevalent words via frequency-counting across the abstracts followed by building codes based on the meaning and relationship of these prevalent words. The process of frequency counting is very accurate in both tools. The same process was followed using ChatGPT 4.1. In this case the accuracy and thus identification of prevalent words depends on the prompting strategy used. Three different sets of strategies were assessed with increasing level of complexity from zero-shot prompting, to few-shots prompting, to the more elaborate CoT prompting. Zero shot (conventional or through a CoT) accompanied by a display of a full word by abstract matrix provided the best results for structured texts. Although these strategies presented a better performance, still frequency-counting and identification of words complying with a specific prevalence threshold by ChatGPT 4.1 are deficient. This process of finding highly prevalent words for coding and thematic analysis provides a practical way of finding shared themes contained in a corpus. However, if we want to find all themes, we need to be more flexible and consider less prevalent words too. Highly prevalent words are consonant with the concept of repetition for identifying keywords. However, it does not equal thematic relevance. Important themes may be discussed infrequently, as indeed it was highlighted by less prevalent words and how they defined codes not detected by the highly prevalent words. Another study [27] noticed that ChatGPT (version 4.0) failed to detect low-frequency codes, or that it suggests themes without subthemes. In our case, neither situation arose; rare codes were identified, as well as themes and subthemes. The limitation was in fact introduced by design, that is by specifying a

minimum number of abstracts in which these words should appear.

To assess the generalizability of results in terms of prompting strategies performance, less structured text was sought whose abstracts addressed the same overarching topic. Indeed, narratives included a large subset of themes contained in the abstracts in an informal tone. In the more informal sources, we ascertain that some of the previous findings hold, i.e., zero shot still providing good performance with or without CoT. However, adding general topic information and role playing to CoT provided the best result with both good sensitivity and specificity. The latter allows us to obtain most of the words that will assist in identifying codes.

A somewhat surprising finding is that CoT does not necessarily work as the best strategy in all settings. It did not for structured text, however it resulted in the best strategy when assisted with overall topic information and role playing for less structured text. Likely structured text requires less sophisticated prompting to uncover its content, in contrast to text including a more colloquial tone. However, the overall goal of identifying highly prevalent words includes multi-step reasoning, a scenario in which CoT excels and which is the reason why CoT was chosen. Others have identified situations in which CoT does not perform as well as expected. It can happen when

1. too many steps are involved causing problems in one or more intermediate steps (known as reasoning inflation which includes self-doubt and verification, back-tracking and retries, or repetition of previous steps [28]);
2. certain prompts are ignored in favor of memorized facts (known as prior knowledge bias produced by prior information that prevails over new information even when introduced in the prompt [29]); or
3. novel reasoning models already incorporate this stepping process internally and thus resist receiving it through the prompts.

Hallucination, a concept related to validity and trustworthiness, in our case and likely in the task of coding from keyword extraction arises from errors in the frequency counting process, as evidenced by the full word by abstract matrix. It is not clear if this arithmetic failure is in the counting per se, or in keeping track which words are observed in which abstract since both issues can be appreciated in those matrices. Using ChatGPT 4.1 to generate the codes via the frequency count process would have missed a third of the codes while another third would have been more detailed. The zero-shot prompting strategies with full word matrix serve as guidance for practitioners as what can be used for coding, while for learners exploring all strategies should be part of the learning process of both coding in QDA and use of AI.

We noted that thematic saturation was not reached after 16 abstracts, a number close to empirical saturation thresholds in many QDA. However, in our example the new codes appearing from the later abstracts (more recent in time) reflect the lack of saturation in the field itself. Research related to joint effects of pollutants and multiple exposures, as well as in depth assessment of the built environment is more recent in time in comparison to research dealing with the other codes identified.

Asking ChatGPT 4.1 to generate the data codes directly without using a word frequency counting strategy led to the best results for coding itself. It identified most codes we extracted using a manual approach. In our specific example nine codes summarize well the content of these abstracts and the underlying themes developed. The ability of ChatGPT to perform thematic analysis has been evaluated by several authors [4], [5], [11], [30], [31] who also assessed single prompting strategies and how these could augment or automate coding and thematic analysis. Most of these studies evaluated either zero-shot or few-shot prompting in earlier versions of ChatGPT (3.5 and 4.0). These authors identified similar advantages (speed, reasonable coding, and theme identification) and limitations (hallucinations) as the present work did. Only one study [32] has assessed the same three prompting strategies evaluated here (still in a previous ChatGPT version) in a corpus of 14 texts of various sources (magazines, blog posts, news articles, transcript) and sizes (252 words to 21.6 thousand words). These authors are more critical of their findings in comparison to the current and previous works. They also had greater expectations of the role of AI, which is fully automation of the coding and thematic analysis processes. They needed eight cycles of conversations to start observing results that conformed to their expectations. They indicated that demand for a manual oversight would negate any ascribed efficiency gains. They also indicated that these previous studies which obtained comparable results drew opposite conclusions. So, it seems that the level of satisfaction with AI's performance for QDA will depend on the users' expectations of AI role, i.e., AI as an assistant vs. an independent analyst.

Combining the codes derived manually and those generated by ChatGPT 4.1 provides a more comprehensive and detailed structuring of the information contained in the corpus analyzed. This indicates that likely using a couple of these tools rather than just one tool can assist analysts to a more successful QDA. Likewise, when using AI tools, it will be advisable to run at least a couple of replicates of the same process to determine consistency in the findings given the random nature of its underlying learning process. We observed in both structured and less structured corpora, that although there is at least a 50% consensus between two runs, independent codes emerge in individual runs and thus running at least a couple of them increases the chance to tap into those non-overlapping codes and themes. This issue relates to reliability. This is necessary in case that the parameter temperature that controls the level of randomness in the chatbot is not accessible to the user as frequently occurs. Additionally, performing independent runs would mimic the process of different coders analyzing the data.

It is important to warn readers of two issues: one relates to the scope of the findings, the other to comparisons of

results across studies involving AI. The findings should be considered in the context of the two related corpora examined and be cautious when extrapolating to other corpora. Also, it is very important to consider the version of the AI tool – given that from one generation to the next, and even within the same generation with few modifications (e.g. from ChatGPT 4 to ChatGPT 4.1) – may lead to substantial enhancements to the tools that may affect positively QDA processes and thus issues raised with a previous version may be solved by those modifications of a newer version.

Conclusion

AI chatbots simplify the process of coding within QDA. Either, simple zero shot strategies for structured text, chain-of-thought with supplemental general topic information and role playing for less structured text, or a direct request of thematic coding combined with a classical QDA tool allow the analyst to derive a comprehensive set of codes. Users of both QDA and AI should be exposed to the whole process of prompting to gain a more comprehensive and rewarding learning experience.

Note

Competing interests

The author declares that she has no competing interests.

References

1. Gao J, Guo Y, Lim G, et al. CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In: Floyd Mueller F, Kyburz P, editors. CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; 2024 May 11-16; Honolulu, HI, USA. New York, NY: Association for Computing Machinery; 2024. Article No. 11. DOI: 10.1145/3613904.3642002
2. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. International Journal of Qualitative Methods. 2023 Oct 30;22:1-10. DOI: 10.1177/16094069231211248
3. Xiao Z, Yuan X, Liao Q, Abdelghani R, Ouedeyer PY. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 with deductive coding. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces; 2023 Mar 27. p. 75-78. DOI: 10.1145/3581754.3584136
4. De Paoli S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. Social Science Computer Review. 2024 Aug;42(4):997-1019. DOI: 10.1177/08944393231220483
5. Turobov A, Coyle D, Harding V. Using ChatGPT for thematic analysis. Working Paper. Bennet Institute for Public Policy, University of Cambridge; 2024.
6. Nguyen-Trung K. ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? Quality and Quantity. 2025 Jun 2;59:4945-4978. DOI: 10.1007/s11135-025-02165-z
7. Bai W, Li Y, Niu Y, Ding Y, Yu X, Zhu B, Duan R, Duan H, Kou C, Li Y, Sun Z. Association between ambient air pollution and pregnancy complications: A systematic review and meta-analysis of cohort studies. Environ Res. 2020 Jun;185:109471. DOI: 10.1016/j.envres.2020.109471
8. OpenAI. ChatGPT (version 4.1) [Large language model]. 2023. Available from: https://chat.openai.com/chat
9. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org
10. Lumivero. NVivo (Version 14) [Computer software]. 2023. Available from: https://lumivero.com/products/vivo
11. Lee VV, van der Lubbe SCC, Goh LH, Valderas JM. Harnessing ChatGPT for Thematic Analysis: Are We Ready? J Med Internet Res. 2024 May 31;26:e54974. DOI: 10.2196/54974
12. Case Western Reserve University. AI Prompt Engineering 101: Foundational Skills. Training Workshop. 2026 Mar 3. Available from: https://case.edu/hr/professional-development/live-training-demand-courses/ai-prompt-engineering
13. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Ed Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: NIPS'22: Proceedings of the 36th Conference on Neural Information Processing Systems; 2022 Nov 28 - Dec 9; New Orleans (LA), USA. (Advances in Neural Information Processing Systems; 35). p. 24824-24837. DOI: 10.52202/068431-1800
14. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101. DOI: 10.1191/1478088706qp0630a
15. Naeem M, Smith T, Thomas L. Thematic analysis and artificial intelligence: A step-by-step process for using ChatGPT in thematic analysis. International Journal of Qualitative Methods. 2025;24:1-18. DOI: 10.1177/16094069251333886
16. One in 20 cases of pre-eclampsia may be linked to air pollutant. Science Daily; 2013 Feb 7. Available from: https://www.sciencedaily.com/releases/2013/02/130206185852.htm
17. Sinpetry L. Air Pollution Linked to 1 in 20 Cases of Pre-Eclampsia. Softpedia; 2013 Feb 8. Available from: https://news.softpedia.com/news/Air-Pollution-Linked-to-1-in-20-Cases-of-Pre-Eclampsia-328058.shtml
18. Sjøgren K. Traffic noise and pollution increase risk of pre-eclampsia during pregnancy. Science Nordic; 2016 Oct 27. Available from: https://www.sciencenordic.com/denmark-environment-videnskabdk/traffic-noise-and-pollution-increase-risk-of-pre-eclampsia-during-pregnancy/1438950
19. Galvin G. Air pollution tied to hypertension in pregnant women. US News & World Report; 2019 Dec 18. Available from: https://www.usnews.com/news/healthiest-communities/articles/2019-12-18/air-pollution-tied-to-hypertension-in-pregnant-women-study
20. McHugh E. Clean the Air: Protecting Expectant Mothers from the Hazards of Pollution. MedpageToday; 2024 Jun 15. Available from: https://www.medpagetoday.com/opinion/second-opinions/110667
21. HealthDay staff. Pregnant Woman Exposed to 45 Common Chemicals. Study Finds. US News & World Reports, HealthDay News; 2026 Jun 17. Available from: https://www.usnews.com/news/health-news/articles/2026-06-17/pregnant-woman-exposed-to-45-common-chemicals-study-finds

22. Epidemiology. Impact of Road Traffic Pollution on Pre-eclampsia and Pregnancy-induced Hypertensive Disorders [Video]. YouTube; 2017. Available from: https://www.youtube.com/watch?v=C6moDXFI47U

23. FIU in DC. Air pollution and Pre-eclampsia [Video]. YouTube; 2022 Aug 28. Available from: https://www.youtube.com/watch?v=aN8NX3ZUiPM

24. Kumar A. How does pollution affect pregnancy [Video]. YouTube; 2025 Nov 28. Available from: https://www.youtube.com/shorts/WI85EZiI26Y

25. CBS New York. Max Minute: New Data Reveals Damaging Effects Of Climate Change On Pregnant Women [Video]. YouTube; 2020 Jun 18. Available from: https://www.youtube.com/watch?v=cQFGqLeaGug

26. News9Live. Invisible Danger: Air Pollution's Impact on Pregnancy [Video]. YouTube; 2025 Dec 18. Available from: https://www.youtube.com/watch?v=8k00wKh4Drg

27. Salazar M, Chaw M, Hellier Y, Hsia S, Gruenberg K. Comparison of Qualitative Analyses Conducted by Artificial Intelligence Versus Traditional Methods. Am J Pharm Educ. 2025 Dec;89(12):101882. DOI: 10.1016/j.ajpe.2025.101882

28. Lian X, Krichene W, Huang B, et al. Quantization inflates reasoning: token inflation as a hidden cost of low-bit reasoning models. arXiv. 2026. DOI: 10.48550/arXiv.2606.25519

29. Chochlakis G, Pandiyan NM, Lerman K, Narayanan S. Larger language models don't care how you think: why chain-of-thought prompting fails in subjective tasks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6-11; Hyderabad, India. arXiv; 2024. DOI: 10.48550/arXiv.2409.06173

30. Hitch D. Artificial intelligence augmented qualitative analysis: The way of the future? Qualitative Health Research 2024;34(7):595-606. DOI: 10.1177/10497323231217392

31. Sen M, Sen SN, Sahin TG. A new era for data analysis in qualitative research: ChatGPT! Shanlax International Journal of Education. 2023;11:1-15.

32. Nguyen DC, Welch C. Generative artificial intelligence in qualitative data analysis: analyzing -or just chatting? Organizational Research Methods. 2026;29(1):3-39. DOI: 10.1177/10944281251377154

Corresponding author:
Mireya Diaz, PhD
Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, United States
mcd8@case.edu

Please cite as
Diaz M. Prompting strategies for coding in AI-assisted qualitative data analysis. GMS Med Inform Biom Epidemiol. 2026;22:Doc14. DOI: 10.3205/mibe000312, URN: urn:nbn:de:0183-mibe0003124

This article is freely available from
https://doi.org/10.3205/mibe000312

Published: *2026-09-22*