Using ChatGPT for Quantitative Content Analysis: Opportunities and Challenges in Construction and Sustainability Research

Mona Foroozanfar, Frederic Bonneaud and Dominique Laffly
LRA - Laboratoire de Recherche en Architecture, Toulouse, France
mona.foroozanfar@gmail.com
frederic.bonneaud@toulouse.archi.fr
dominique.laffly@gmail.com

Abstract: Artificial Intelligence (AI), especially Large Language Models (LLMs) like ChatGPT, are changing the way researchers can process and analyse qualitative data. In this paper the use of ChatGPT is tested for Quantitative Content Analysis (QCA) by applying it to interview material about digital construction technology and sustainability. Two versions of the same data are compared: (1) complete transcripts of five interviews with professors, and (2) a shorter summarized version of the same interviews (the summaries were prepared by researcher). With the same workflow, ChatGPT did several steps: preprocessing (splitting the text into words, removing very common small words, and reducing words to their basic form), keyword extraction, thematic coding with five categories, and also a simple sentiment analysis. The aims were: (a) to see if ChatGPT can find the main themes in a reliable way, (b) to compare results from full transcripts versus summaries, and (c) to understand what practical advantages and problems appear when undertaking ChatGPT in a real research situation. The results were similar at the general level: Digital Technology and Sustainability were the strongest themes in both datasets, followed by Education/Training, Benefits, and Barriers. The sentiment analysis gave slightly positive values in both (+0.18 for transcripts, +0.16 for summaries). At a more detailed level, the transcripts included more technical words (for example “embodied carbon”, “Life cycle Analysis (LCA)” and standards), while the summaries included more general terms, which made the counts higher. Some practical issues also influenced the work: undertaking a free ChatGPT account caused interruptions, sometimes the tool changed its output style, and it was difficult to export charts or tables, these problems reduced reproducibility. In conclusion, ChatGPT can be useful for first steps in QCA and for saving time in early coding, but it is not enough for final or very detailed analysis. For better use, the following suggestions are provided: a combination of AI with human checking, making domain-specific dictionaries, undertaking clear and repeated prompts, and working with more stable or professional access. This study shows both the opportunities and the real problems when ChatGPT is used for content analysis in construction and sustainability research.

Keywords: ChatGPT, Quantitative content analysis, NLP, Construction, Sustainability, Digital technology

1. Introduction

Qualitative interviews are very important in research, because they show the views of experts. The interviews done in this research include practical experience, teaching ideas, and technical information that help to understand what makes digital technology adoption easier or more difficult. In the current research, the interviewees were asked to verify the results obtained from the survey, then Quantitative Content Analysis (QCA) was used for analysing interviews. QCA is a method that turns text into numbers, like codes, counts, and connections, and is often used to make summaries of qualitative data (Mayring, 2014; Krippendorff, 2018). Traditional QCA, however, needs trained people to code the text, repeat checks for reliability, and much manual work, which can take days even for small datasets.

Large Language Models (LLMs), such as the GPT family, offer a new way to accelerate quantitative content analysis by automating initial coding, extracting keywords and collocations, and producing structured counts that support faster exploratory work (Bijker et al., 2024; Tai et al., 2024). For researchers in construction as a field that now faces strong digital changes (like Building Information Modelling (BIM), digital twins, or AI in project management) and also needs to respond to environmental issues, LLMs may help to quickly analyse interviews and measure patterns (Azhar, 2011; Succar and Kassem, 2015; Sacks et al., 2018).

But LLMs also have problems: their answers can change in each run, results are not always stable (especially with free accounts), sometimes the system makes up details that are not real and they may lose detailed technical meaning (Bender et al., 2021; Ziems et al., 2024). For this reason, it is important to test how ChatGPT really performs in QCA tasks.

This study shows an experiment where the same interview data was used in two forms: full transcripts and shorter summaries. Both versions were analysed with the same ChatGPT process. The main research questions were:

- Can ChatGPT find the main themes in expert interviews about digital construction technology and sustainability?

- How different are the results when ChatGPT works with full transcripts compared to shorter summaries?
- What are the practical advantages and problems when ChatGPT is used for QCA in real research situations?

The aim of this study is to help other researchers considering the use of LLMs for QCA and to show the real challenges of working with common AI tools.

2. Literature Review

2.1 Quantitative Content Analysis and its Challenges

Content analysis is a well established method for converting textual material into structured data. Traditional QCA procedures emphasize clear category definitions, coder training, and inter coder reliability testing (Mayring, 2014). Content analysis can be approached qualitatively or quantitatively. Qualitative content analysis focuses on understanding meanings and patterns within communication through interpretation and categorization, offering depth but often being time-consuming and subjective. In contrast, quantitative content analysis measures textual features such as word frequencies, providing objectivity and replicability but sometimes oversimplifying complex meanings. In practice, both methods are complementary, as qualitative insights guide coding, and quantitative measures enhance reliability and comparability (Devi, 2009). Krippendorff (2018), also differentiates between qualitative and quantitative content analysis, both have strengths and limitations. QCA's main advantages are replicability and the ability to compare frequencies across groups or time, however, QCA is laborious, and human coders introduce subjectivity and fatigue, which can affect reliability (Neuendorf, 2017). However, despite its structured rigor, quantitative content analysis may overlook contextual nuances and interpretive depth. This limitation highlights the need to integrate human interpretation to complement statistical findings and ensure richer understanding.

2.2 Automated Text Analysis and LLMs

Automated text analysis undertaking Natural Language Processing (NLP) has advanced rapidly in recent years, driven by larger datasets, greater computing power, and the development of deep learning methods (Hirschberg and Manning, 2015; Young et al., 2018). Common techniques include splitting text into words (tokenization), finding frequent word groups (n-grams), undertaking dictionaries for positive or negative words (sentiment lexicons), and applying topic models to discover themes (Jurafsky and Martin, 2021). More recently, Large Language Models (LLMs) such as GPT-3 (Brown, 2020) and ChatGPT provide conversational and instruction-following capacities that make them attractive for non-expert users. Early studies show that LLMs can do some annotation and classification tasks with results close to human performance in certain areas (Gilardi, Alizadeh and Kubli, 2023). However, there are still problems: the outputs are not always consistent, results can change depending on how the prompt is written, and the process inside the model is not transparent (Bender et al., 2021; Ziems et al., 2024). Yet, while automation improves scalability and efficiency, the "black-box" (the internal decision-making process is not fully transparent or interpretable) nature of LLMs raises concerns about interpretability and validity. Future studies should therefore focus on balancing computational power with methodological transparency, especially in applied research contexts.

2.3 Digital Technologies and Sustainability in Construction

Digital technologies, are central to contemporary construction research. For example, BIM, as a process that relies on digital technology, enables integrated modelling, data exchange, and lifecycle analysis, is considered an enabler for sustainability practices like LCA (Azhar, 2011; Sacks et al., 2018). One of the methods for undertaking research in this domain is interview to understand adoption barriers (cost, skills, culture), benefits (efficiency, reduced rework) and to study effective factors on stakeholders' opinion to adopt them. Thus, accurately coding of interview content is critical for deriving actionable conclusions. Nevertheless, much of the existing literature emphasizes the potential of digital tools rather than their practical implementation challenges. More empirical studies are needed to assess how these technologies actually contribute to sustainability outcomes in real-world construction projects.

2.4 Hybrid Human–AI Workflows

In the literature, a practical view suggests undertaking a hybrid process: AI tools can be used for fast first coding, and human researchers can check and interpret the results. This method can save time and still keep the deeper analysis (Li, Dohan and Abramson, 2021; Mortelmans, 2025; Williamson, Van Rooyen and Dry,

2025). But more real examples are necessary to show how this approach can be applied in special areas such as construction management. However, while the hybrid model promises efficiency and balance, empirical validation in domain-specific contexts such as construction management remains limited. Further studies should explore whether AI-assisted coding genuinely enhances accuracy or merely accelerates surface-level interpretation.

3. Methodology

3.1 Data Collection and Preparation

The empirical data for this study came from five semi-structured interviews with professors and senior researchers in construction related fields. Each interview lasted between 45 and 60 minutes and focused on digital technologies adoption (BIM, digital twins, AI) and sustainability (LCA, embodied carbon, energy efficiency).

To create the full transcripts, the interviews, which were in different languages, were recorded in text or audio format. The software Descript was used to convert the audio into text. This process produced verbatim transcripts that captured the exact words of each participant. After that, a summarized version of each interview was prepared by the researcher. In these summaries, filler sentences were removed, long explanations were paraphrased, and the main argumentative points were preserved. The summarized texts were approximately 30–40% of the length of the original transcripts. This approach allowed for a more concise dataset while keeping the essential content of the interviews for further analysis.

3.2 Predefined Categories and Lexicon

Although the main structures of the study were originally based on personal, educational, technical, and project specifications, which served as the framework for designing the questionnaire and presenting it to the interviewees, the analysis of the interview themes led to the identification of five key categories. These categories are:

- Digital technology: BIM, digital twins, IoT, AI, data platforms.
- Sustainability: carbon, LCA, energy efficiency, green building.
- Education/Training: curricula, workshops, student skills, conferences.
- Benefits: efficiency, reduced errors, improved decision-making, lifecycle optimization.
- Barriers: cost, licensing, skills gap, organizational resistance, interoperability issues.

3.3 ChatGPT-Based Analytical procedure

To analyse the interviews, ChatGPT-4o was used with a consistent prompt and workflow. The researcher provided instructions, and ChatGPT itself performed the main steps. First, for preprocessing, the model split the text into words (tokens), converted them to lowercase, and removed common English stop words (common words such as “the”, “and”, “of” that are usually removed before analysis because they do not carry meaningful information). Basic lemmatization (a text-processing step that reduces words to their base or dictionary form for instance, “running”, “runs”, “ran” → “run”) was handled automatically by ChatGPT’s language understanding. Then, for keyword and bigram (a pair of consecutive words analysed together to detect frequent word combinations, for instance, “digital technology”, “sustainable development”) extraction, ChatGPT returned the 30 most frequent single words and the 20 most frequent word pairs. During thematic mapping, the model assigned each word or word pair to one or more predefined categories, and the counts were aggregated per category. For sentiment scoring, a simple rule-based method was applied by ChatGPT: each sentence was evaluated for positive and negative words, and an overall score was calculated. Finally, all model responses and summary tables were recorded manually by the researcher. Each dataset, both full transcripts and summaries, was processed in separate ChatGPT sessions.

3.4 Validation and Cross-Checks

To check the plausibility of ChatGPT results, some manual validations were performed. First, ten sentences were selected randomly from the interviews, and the researcher manually decided what categories they should belong to. These sentences were then given again to ChatGPT with a prompt asking which category it assigned them to. Based on comparing the manual judgment and the model’s answer, the agreement was high for broad categories, such as Digital Technology or Sustainability, but lower for technical topics that required specific domain knowledge.

Second, some prompts were repeated multiple times to see if ChatGPT produces the same counts and assignments. Small differences appeared, showing that results are mostly stable but not perfectly identical in repeated runs. These steps allowed the researcher to get confidence that ChatGPT generally assigns sentences and words correctly for broad categories, while noting that careful human review is still needed for technical, complex or ambiguous content.

4. Results

4.1 Theme Frequency

The results of the word and phrase categorization performed by the ChatGPT analytical process are summarized in Table 1.

Table 1: Theme Frequencies

Theme	Full Transcript	Summary Text
Digital Technology	122	155
Sustainability	109	131
Education/Training	74	93
Benefits	68	79
Barriers	52	68

Key observations:

The order of themes is the same in both datasets: Digital Technology > Sustainability > Education/Training > Benefits > Barriers (Figure 1). This shows that ChatGPT was able to capture the main priorities in a stable way. At the same time, the number of counts is higher in the summary dataset. This difference seems to happen because the summaries use simpler and more repeated wording (for example, “technology,” “sustainability”), while the longer transcripts often included extra details or qualifiers.

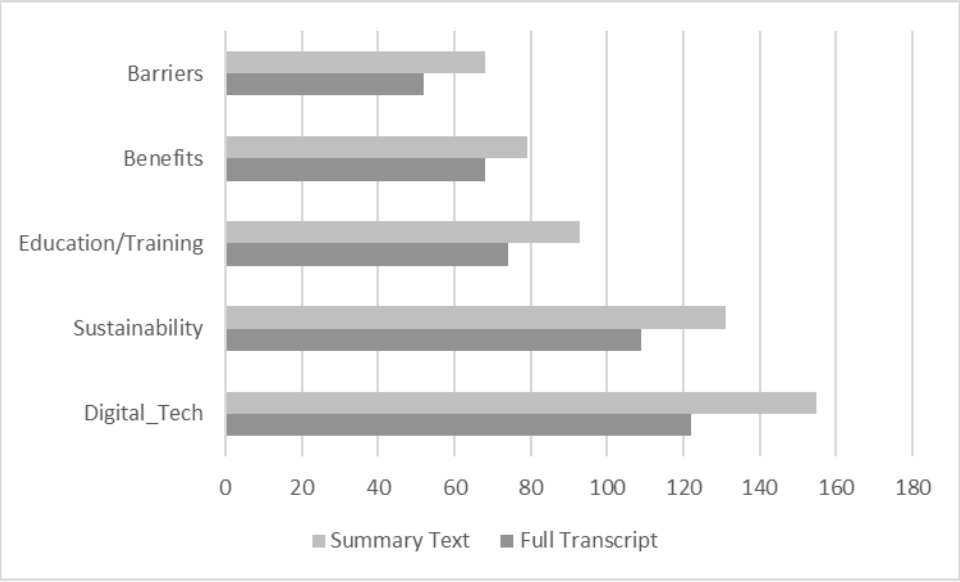

Figure 1: Comparison chart (full vs summary)

4.2 Keyword Patterns

For the full transcripts, the most frequent single words (unigrams) included: BIM, carbon, sustainability, energy, LCA, data, student, curriculum, cost, interoperability. In the summary dataset, the top unigrams were more general: technology, digital, sustainability, education, adoption, training.

The most frequent two-word combinations (bigrams) in the full transcripts highlighted technical concepts, such as “embodied carbon”, “life cycle”, “carbon footprint”, “BIM model”. In contrast, the summaries emphasized broader phrases, including “digital technology”, “sustainable development”, “student training”.

4.3 Sentiment

The compound score provides a single measure of the overall sentiment in the text. It combines positive, negative, and neutral elements into a value ranging from -1 to +1. Scores close to +1 indicate strongly positive sentiment, scores near -1 indicate strongly negative sentiment, and values around 0 suggest either neutral sentiment (no strong emotional tone) or a balance of positive and negative signals (mixed sentiment) (Hutto and Gilbert, 2014) . In this study, the full transcripts had a compound score of +0.18, and the summaries had +0.16, indicating mildly positive overall sentiment (Figure 2)."

Interviewees generally expressed optimism about the potential of digital technologies to support sustainability goals, while also noting practical concerns such as cost, skill gaps, and policy constraints.

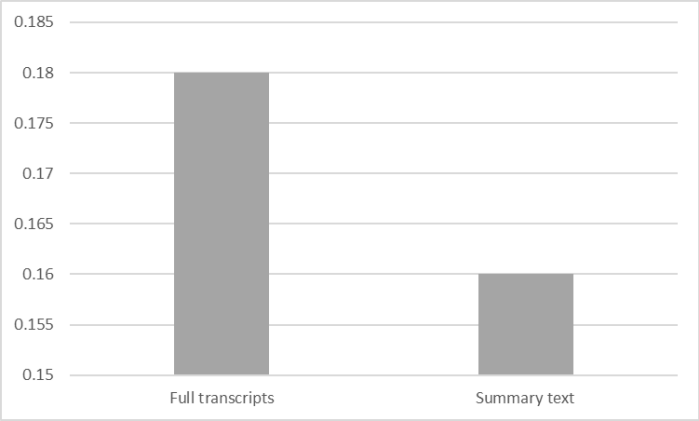

Figure 2: Sentiment comparison

4.4 Manual Spot-checks and Agreement

To evaluate the reliability of ChatGPT's coding, ten random excerpts from the interviews were checked manually. The level of agreement between the human coding and ChatGPT was relatively high, especially for the broad categories. The disagreements mostly appeared in sentences that were more complex in meaning, for example when nuance, sarcasm, or mixed opinions were present. In addition, technical sentences sometimes could belong to more than one category. For instance, "IFC mapping for LCA" could reasonably fit under both Digital Technology and Sustainability.

5. Discussion

This study shows that ChatGPT can effectively capture high level thematic structures in qualitative interview data, offering fast and generally consistent results. The model is particularly useful for detecting dominant themes and reducing the time required for exploratory coding, making it accessible even for researchers without advanced NLP expertise. However, important differences were observed between full transcripts and summaries: summaries emphasized broad categories and sometimes overstated consensus, while full transcripts preserved technical and less frequent but meaningful terms that are vital for nuanced interpretation and policy relevant insights. Practical challenges included instability of free accounts, occasional variability across repeated runs, and export limitations, while methodological challenges included the model's difficulty with technical specificity and subtle meanings without human oversight. These findings suggest that LLM assisted analysis is best used in a hybrid workflow, where ChatGPT accelerates first pass coding and theme detection, but careful human validation remains essential to ensure accuracy, resolve ambiguities, and capture domain specific detail.

6. Limitations and Future Work

6.1 Limitations

- Dataset size and scope: The experiment included five interviews. This is enough for a methodological pilot, but the results cannot be fully generalized. Undertaking a larger dataset might reveal different patterns.
- Prompt and session dependence: The outputs were sensitive to how tasks and instructions were phrased when given to ChatGPT. Even though prompt templates were applied, small changes in wording could affect results.

- Single tool focus: This study relied on ChatGPT as an example of LLM workflows. Other models could produce different outcomes.
- Export reproducibility: Practical issues on the live platform, such as download failures and session timeouts, limited operation.
- Platform limitations and free plan use: This study used ChatGPT with free plan access, which imposed session interruptions, timeouts, and limited functionality. Paid or newer versions of ChatGPT (e.g., GPT-5, professional subscriptions) could provide more stable sessions, smoother interaction, better graphing and visualization options, and potentially more accurate analyses. Therefore, results observed here might differ if professional access were used.
- Summary preparation by the researcher: Since the researcher manually prepared the summarized versions of the interviews, some terms and concepts may have been replaced with more general equivalents. This could have influenced the results, for example by slightly inflating counts for high level categories or reducing the visibility of specific technical terms.

6.2 Future Work

- Scale up: Apply the procedure to a larger set of interviews to test its reliability and to explore new or emerging topics.
- Hybrid validation study: Systematically compare ChatGPT outputs with human coding done by multiple coders.
- Domain adaptation: Guide or train LLMs undertaking construction and sustainability specific lexicons to improve the accuracy of categorization.
- Software integration: Combine ChatGPT outputs with visualization tools for better reporting and reproducibility.
- Method standards: Develop standard prompt templates and reporting checklists for AI-assisted QCA to ensure reproducibility across studies.

7. Conclusion

This study explored the use of ChatGPT for quantitative content analysis of interviews in construction and sustainability. Results show that ChatGPT can effectively detect dominant themes such as Digital Technology, Sustainability, and Education/Training, providing consistent macro level insights across full transcripts and summaries. However, limitations remain: summaries may inflate general category counts, free platform instability and time to time variability can affect reproducibility, and subtle technical details or domain specific lexicons, risk misclassification without human review.

Based on these findings, researchers and practitioners are advised to approach ChatGPT as a supportive tool rather than a replacement for human judgment. Careful design of prompts, thorough documentation of sessions, and the use of hybrid workflows where humans validate and refine AI generated coding are key to capturing nuanced or technical content accurately. Predefining domain specific lexicons, such as related tags to BIM or LCA, improves categorization accuracy. For larger projects, professional or paid access can provide more stable sessions, smoother interaction, and better visualization and analytical capabilities. By combining these practices, ChatGPT can accelerate the early phases of content analysis, helping researchers efficiently process qualitative data and uncover meaningful patterns. As LLMs evolve, these methods should be refined and standardized to ensure that AI assisted QCA is both efficient and scientifically defensible.

Ethics declaration: The interviews used in this study were conducted as part of the author's doctoral research. All participants were informed about the purpose of the study and gave their verbal consent to participate voluntarily. No sensitive or personally identifiable information was collected, and all data were anonymized prior to analysis.

AI declaration: In this study, ChatGPT was used to assist the qualitative analysis of interviews conducted for a doctoral thesis. Both full transcripts and summarized versions were provided to examine whether the model produces consistent results. ChatGPT performed the analysis and explained its methods, and the outputs were then reviewed and compared by the researcher. Based on these results, the manuscript was prepared.

References

Azhar, S. (2011) "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry," *Leadership and management in engineering*, 11(3), pp. 241–252.

Bender, E.M. *et al.* (2021) “On the dangers of stochastic parrots: Can language models be too big? 🦜 ,” in. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bijker, R. *et al.* (2024) “ChatGPT for Automated Qualitative Research: Content Analysis,” *J Med Internet Res*, 25. Available at: https://doi.org/10.2196/59050.

Brown, T. (2020) “Language models are few-shot learners. Advances in Neural Information,” in. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. Retrived from https://proceedings. neurips. cc/pa*.

Devi, N.B. (2009) “Understanding the qualitative and quantitative methods in the context of content analysis,” in. *Qualitative and Quantitative Methods in Libraries, International Conference organised by International Federation of Library Associations*, pp. 26–29.

Gilardi, F., Alizadeh, M. and Kubli, M. (2023) “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, 120(30), p. e2305016120.

Hirschberg, J. and Manning, C.D. (2015) “Advances in natural language processing,” *Science*, 349(6245), pp. 261–266.

Hutto, C. and Gilbert, E. (2014) “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in. *Proceedings of the international AAAI conference on web and social media*, pp. 216–225.

Jurafsky, D. and Martin, J.H. (2021) “Word senses and WordNet,” *Speech and Language Processing*, 10.

Krippendorff, K. (2018) *Content analysis: An introduction to its methodology*. Sage publications.

Li, Z., Dohan, D. and Abramson, C.M. (2021) “Qualitative Coding in the Computational Era: A Hybrid Approach to Improve Reliability and Reduce Effort for Coding Ethnographic Interviews,” *Socius: Sociological Research for a Dynamic World*, 7.

Mayring, P. (2014) “Qualitative content analysis: theoretical foundation, basic procedures and software solution.”

Mortelmans, D. (2025) “NVivo and AI: (Semi)-Automatic Coding,” in Mortelmans, D., *Doing Qualitative Data Analysis with NVivo*. Cham: Springer Nature Switzerland (Springer Texts in Social Sciences), pp. 229–250. Available at: https://doi.org/10.1007/978-3-031-66014-6_19.

Neuendorf, K.A. (2017) *The content analysis guidebook*. sage.

Sacks, R. *et al.* (2018) *BIM handbook: A guide to building information modeling for owners, designers, engineers, contractors, and facility managers*. John Wiley & Sons.

Succar, B. and Kassem, M. (2015) “Macro-BIM adoption: Conceptual structures,” *Automation in construction*, 57, pp. 64–79.

Tai, R.H. *et al.* (2024) “An Examination of the Use of Large Language Models to Aid Analysis of Textual Data,” *International Journal of Qualitative Methods*, 23, p. 16094069241231168. Available at: https://doi.org/10.1177/16094069241231168.

Williamson, C., Van Rooyen, A. and Dry, R. (2025) “In Tandem With Artificial Intelligence: A working Framework for Coding in ATLAS.ti™,” *International Journal of Qualitative Methods*, 24, p. 16094069241306551. Available at: https://doi.org/10.1177/16094069241306551.

Young, T. *et al.* (2018) “Recent trends in deep learning based natural language processing,” *ieee Computational intelligenCe magazine*, 13(3), pp. 55–75.

Ziems, C. *et al.* (2024) “Can large language models transform computational social science?,” *Computational Linguistics*, 50(1), pp. 237–291.