Issues in Information Systems
Volume 26, Issue 3, pp. 115-126, 2025

DOI: https://doi.org/10.48009/3_iis_2025_2025_109

Leveraging large language models and in-context learning for construct identification in computational social science: A case study on wearable devices

Omar El-Gayar, *Dakota State University, omar.el-gayar@dsu.edu*
Abdullah Wahbeh *Slippery Rock University, abdullah.wahbeh@sru.edu*
Mohammad Al-Ramahi, *Texas A&M University - San Antonio, mohammad.abdel@tamusa.edu*
Ahmed Elnoshokaty, *California State University, San Bernardino, ahmed.elnoshokaty@csusb.edu*
Tareq Nasralah, *Northeastern University, t.nasralah@northeastern.edu*

Abstract

Large Language Models (LLMs) have opened new frontiers in Computational Social Science (CSS) by enabling the extraction, classification, and analysis of large-scale unstructured text data. This study aims to leverage LLMs to systematically encode theoretical constructs from user-generated content. We propose an LLM-powered construct identification framework that employs LLMs for automated encoding, validated against human-coded benchmarks. The framework was evaluated as a case study in the domain of wearable devices. Two experiments for binary and ternary encoding were tested. For both experiments, the LLM demonstrated high accuracy, precision, and recall in encoding theoretical constructs of user-generated content. The findings emphasize that LLMs can complement traditional methods in CSS, enabling scalable, efficient, and effective analysis of social phenomena across diverse domains.

Keywords: computational social science, large language models, wearable devices**.**

Introduction

Traditional methods such as surveys and qualitative coding have long been used to study human behavior, but they suffer from scalability limitations, response bias, and resource-intensiveness (Chang et al., 2014; Farhadloo et al., 2016). As a result, researchers increasingly explore alternative approaches that leverage machine learning and artificial intelligence techniques to extract theoretical constructs from naturally occurring textual data, such as social media posts and online user reviews (Cao et al., 2011; Xu et al., 2024). The advent of Large Language Models (LLMs), such as GPT, has opened new frontiers in Computational Social Science (CSS) by enabling the extraction, classification, and analysis of large-scale unstructured text data. While LLMs have demonstrated effectiveness in content analysis (Bijker et al., 2024; Hitch, 2024; Li et al., 2024), dataset annotation (Zhang, 2023), thematic analysis (Christou, 2024; Joel-Edgar & Pan, 2024; Perkins & Roe, 2024), text annotation for sentiment analysis (Azad, 2024; Belal et al., 2023; Mathebula et al., 2024), and knowledge and data encoding (Saouabe et al., 2024; Singhal et al., 2023), their ability to encode well-defined theoretical constructs in user-generated content remains underexplored.

This study addresses this gap by leveraging LLMs to systematically encode theoretical constructs in large-scale user-generated content, offering a scalable alternative to traditional methods. Specifically, we propose

an LLM-powered construct identification framework that employs LLMs for automated encoding and is validated against human-coded benchmarks. The study aims to
:

1. Develop a methodology for encoding theoretical constructs using LLMs.

2. Validate the AI-generated construct encodings against human-coded data.

By bridging AI-driven text analytics with construct operationalization, this study contributes to CSS by offering a scalable alternative to traditional methods, enhancing automated construct identification in online user reviews, and demonstrating the applicability of LLMs in CSS research. The remainder of this paper is structured as follows: The Literature Review discusses prior work on AI-driven text analytics and construct identification. The Methodology section details the proposed approach. The Results and Discussion sections summarize key findings, and the Conclusion highlights contributions, limitations, and future research directions.

Related Work

Several studies have used LLMs in various aspects of data analysis, including annotation, qualitative research, grounded theory, and sentiment analysis. LLMs have demonstrated strong potential in data annotation and encoding. Several studies have compared the performance of LLMs with human annotators, showing that models like GPT and Claude can perform at or above human levels in certain tasks. For example, GPT-3.5-turbo achieved a 25% increase in annotation accuracy compared to human experts (Gilardi et al., 2023), while Claude-1.3 outperformed human experts in the task of labeling textual responses in the absence of prior training data (Mellon et al., 2024). On the other hand, LLMs still face challenges and difficulties in understanding context, particularly in multilingual settings, as noted by Nasution & Onan, (2024). Finally, prompt engineering has been shown to significantly improve LLMs' annotation accuracy, as demonstrated by Vujinović et al., (2024).

Several studies employed LLMs in both deductive and inductive qualitative research for coding purposes. Research shows that LLMs can assist researchers by identifying codes and structuring qualitative data. For deductive coding, Tai et al., (2024) and Xiao et al., (2023) leveraged LLMs alongside human experts' codebooks and demonstrated substantial agreement. Similarly, using LLMs to support qualitative analysis has been found to complement human experts rather than replace them (Perkins & Roe, 2024). Furthermore, LLMs improved inductive coding, which involves extracting patterns and themes from unstructured data. Bryda & Sadowski, (2024) introduced generative semantic coding and lexical pattern coding, showing that these methods can automate qualitative analysis while maintaining analytical rigor. LLM-based coding methods, such as ARGUMENT2CODE (Zhao et al., 2024), improved inductive coding by generating analytical prompts and thematic codebooks. Similarly, LLM-based tools such as CoAIcoder (J. Gao et al., 2023) helped facilitate collaborative qualitative analysis, improving coding efficiency while reducing associated costs. Despite their advantages, LLMs do not always fully capture the complexity of human thematic analysis. Hamilton et al., (2023) found that LLMs generated themes that partially overlapped with human expert-generated ones with concerns over broader contextual elements such as faith and family. Similarly, Li et al. (2024) showed that the GPT-4 model performed well in identifying common themes in interview data. However, it was less reliable in capturing less frequent topics than human experts.

LLMs have also been used in grounded theory research to understand and explore their roles in qualitative code generation and iterative refinement. Sinha et al. (2024) showed that GPT-4 could identify overlooked segments and produce higher-level codes with an initial set of lower-level codes compared to an initial set of lower-level codes generated by human researchers. This suggests that while LLMs can assist in developing grounded theory frameworks, they still require human experts to help refine theoretical concepts. LLMs have also been applied to sentiment analysis, where they have outperformed traditional lexicon-based algorithms. Belal et al., (2023) showed that ChatGPT improved sentiment classification accuracy by 20% on tweet datasets and by 25% on Amazon review datasets. These findings highlight LLMs' ability to extract sentiment from unstructured data with greater precision than traditional and advanced analytical approaches.

According to the literature, many studies have demonstrated the potential of LLMs in different tasks, such as content analysis, sentiment analysis, thematic coding, and knowledge and data annotation. However, none have attempted to utilize these models to encode theoretical constructs from online user reviews. Overall, previous studies have primarily focused on more general or descriptive tasks, such as detecting broad themes or improving coding efficiency, rather than mapping data to established theoretical frameworks. Therefore, we believe there is a significant gap in the literature regarding how LLM could be leveraged for theoretically grounded coding. Addressing this gap is crucial, as using LLMs to systematically identify and encode theoretical concepts within user-generated content could open new avenues for rigorous, large-scale analysis of online behavior, especially when compared to traditional methods like surveys, which are limited by sample size and require extensive time and resources for data collection.

Methodology

Figure 1 shows the research method. The method starts with the selection of the problem domain, which consists of clearly defining the domain of the topic of interest, such as healthcare, education, and organizational behavior. Selecting the relevant problem domain is significant since it can help address the key gap in understanding how theoretical constructs are evident in real-world data. Once the domain has been identified, we need to specify where the data for analysis comes from. In the context of the current study as well as the objectives and research gap, data could be obtained from different sources such as surveys, interviews, online posts, and transcripts.

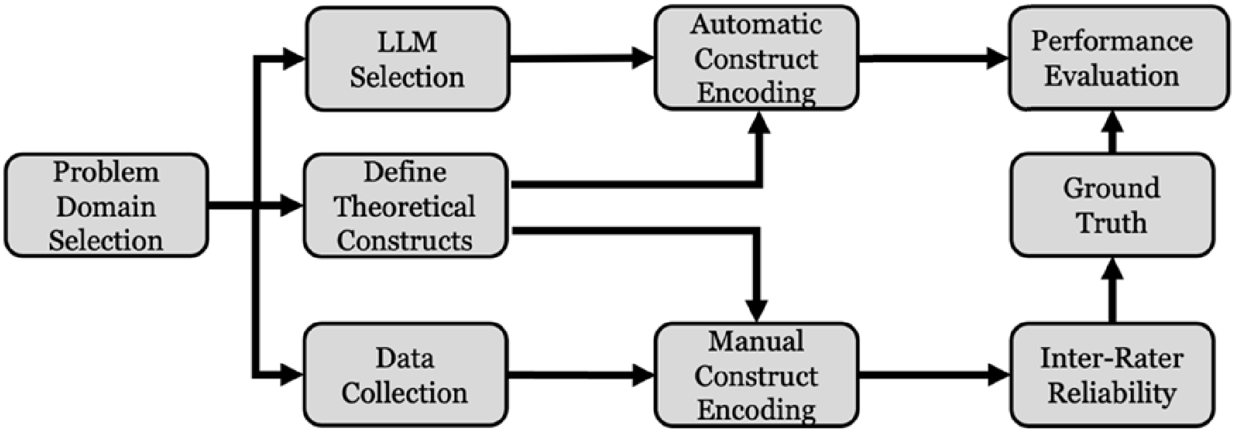

Figure 1. Research Method

Once the relevant data for a specific domain is collected, there is a need to define relevant theoretical constructs. For each construct, we need to define it, provide sample measurement items based on relevant literature to help the LLM model understand how the construct is understood and measured, and manually identify relevant data that capture the constructs to be used with the LLM model. Such details are stored in

a YAML file that allows users to encode domain-specific background knowledge (Alviano & Grillo, 2024). Appendix A shows a sample from the YAML file for the ease of use construct used in the prompt.
The experiment consists of a ternary encoding process, that utilizes the YAML file, in which a construct is marked as positive, negative, or absent in the review. For ternary encoding, for each review, we determine whether a construct is reflected or not as well as the sentiment.

$$x_{r,x}{}^{(ter)} \begin{cases} 1 \;\; if\; construct\; c\; is\; present\; (mentioned\; or\; reflected)\; positively\; in\; review\; r \\ -1 \;\; if\; construct\; c\; is\; present\; (mentioned\; or\; reflected)\; negatively\; in\; review\; r \\ 0 \;\; otherwise \end{cases}$$

The ternary encoding process consists of two complementary tasks. First, two independent domain experts will manually encode the constructs based on the domain knowledge in the YAML file. A third domain expert will then resolve any disagreements between the two. Afterward, we will measure inter-rater reliability (Landis & Koch, 1977) based on the results of the manual labeling and the resolution of disagreements. Once the manual process is completed, we obtain what we refer to as the ground truth.

Next, we automatically encoded the construct using the selected LLM based on the domain knowledge in the YAML file. Prompt engineering (Bijker et al., 2024; Lee et al., 2024) and in-context learning (Agarwal et al., 2025; Dong et al., 2024a; X. Gao & Das, 2024) are used to complete the encoding process. Prompt engineering is a technique for optimizing the interaction with LLMs to achieve desired outcomes. It involves the design, testing, and refinement of prompts to guide LLMs' behavior to achieve useful and desired responses (Chen et al., 2025). In-context learning is the capability of LLMs that allows them to adapt to new tasks without requiring parameter updates. Instead, in-context learning leverages task-relevant information provided in the input context, such as examples or instructions, to guide the model's behavior (Dong et al., 2024b). The process consists of iteratively prompting the LLM to encode the reviews across each construct. Using the manually encoded results as well as the LLM-based results, we generated a confusion matrix for the experiment and calculated the accuracy, precision, recall, and f-measure to evaluate the performance of the LLM model.

We demonstrated the proposed methodology as a case study in the field of wearable devices. Wearable devices are considered an important domain in the field of healthcare and one of the most widely used personal devices after smartphones. Wearable devices can serve as tools for understanding the factors that drive behavioral change (Sjöklint et al., 2015). They also provide opportunities to study psychological and social phenomena, such as motivation, empowerment, and social interactions (Karapanos et al., 2016; Ryan et al., 2019).

Results

Online users' reviews about different brands/models of wearable devices were collected from Amazon and Best Buy websites using a search query on Brandwatch, a social media data collection and analytics tool (*Crimson Hexagon and Brandwatch*, 2020). A total of 50 user reviews were used to evaluate the LLM's final performance. These reviews were selected to represent a range of sentiments and device types, ensuring that the evaluation captures diverse feedback.

Definition of Theoretical Constructs

In social science, constructs refer to abstract psychological traits or social phenomena that researchers aim to measure or understand (Sethi & King, 1991). In the field of wearable devices, we choose relevant constructs based on relevant literature. These constructs include hedonic motivation (HM), perceived ease of use (PEU), perceived usefulness (PU), device appeal (DA), customizability (C), device quality (DQ), device connectivity (DC), perceived values (PV), and credibility support (CS) (Elnoshokaty et al., 2022). For each construct, a short definition is provided based on the literature, a set of items used to measure the constructs were identified, and sample user reviews were identified as shown in Table 1 for demonstration.

Table 1. Perceived Ease of Use Definition, Measurement Items, and Sample Review

Definition:
Perceived ease of use is defined as the degree to which a person believes that using a particular system would be free of effort. In the context of wearable devices, it refers to users' perception of how using a wearable device is free of effort. In general, users are more satisfied with a wearable device when the device requires little effort. According to the literature, several studies have examined users' satisfaction and acceptance of wearable devices and showed that perceived ease of use positively impacts users' satisfaction.

Measurement Items:

1. "I find the wearable device easy to use"
2. "Learning to operate the wearable device is straightforward for me"
3. "Interacting with the wearable device does not require a lot of mental effort"
4. "I find it easy to get the wearable device to do what I want it to do"
5. "The functions of the wearable device are clear and understandable"
6. "I can use the wearable device without written instructions"
7. "I find it easy to navigate through the features of the wearable device"

Sample Users' Review:

- R1: "It was very easy to set up and the app is very easy to use"
- R2: "simple set up, easy-to-understand functions"
- R3: "Setup was simple, the instructions in the box were easy to understand"
- R4: "User friendly. Easy to use. Very basic and simple, but that was perfect for what I wanted"
- R5: "All of the apple products I own are easy and simple to use to the best abilities"
- R6: "Great sound, easy controls, simple to use"
- R7: "I have an Apple Watch. I set it up for her. It was easy. The instructions were pretty clear"
- R8: "We bought this for my niece. The set up was easy and user friendly"
- R9: "The simplicity of the apple watch is amazing. It makes it simple to use and understand"
- R10: "Replaced my Fitbit charge 2. Easy to read. Simple to use apps like timer or alarm"

Manual Construct Encoding

Two independent researchers who are experts in the wearable devices field manually encoded the 50 reviews for each construct. A third researcher resolved any disagreements among the two independent researchers. We achieved an average kappa statistic of 0.876 for the ternary encoding, indicating an almost perfect agreement between the researchers.

Automatic Construct Encoding

GPT-4.1-mini was selected to perform the encoding process for the constructs. GPT-4.1-mini is highly effective for text and content analysis, with the ability to handle long-context comprehension and thousands of tokens. It performs very well with insights extraction and patterns identifications, making it ideal for analyzing large volumes of text data and generating detailed content insights (OpenAI, 2025). The GPT-4.1-mini ability to handle thousands of tokens as well as generate insights from text makes it an ideal model for encoding constructs based on the analysis of opinions expressed by users across a broad range of wearable devices. By leveraging GPT-4.1-mini advanced instruction-following capabilities, it can accurately identify constructs and encode them within the reviews. Using the same set of 50 reviews, GPT-4.1-mini was used to encode all the constructions. For ternary encoding, we provided GPT-4.1-mini the YAML file.

Comparing the resulting encoding from GPT-4.1-mini with the manual ground truth, we achieved an average kappa statistic of 0.741 for ternary encoding, respectively, indicating a substantial agreement between the manual encoding and LLM encoding. To compare the performance of the LLM encoding based on the manual encoded reviews, accuracy, precision, and recall were calculated for each construct. We also calculated overall accuracy, precision, and recall. Table 2 shows the confusion matrix for the ternary encoding process, comparing the automated GPT-4.1-mini encoding with manual encoding. The model correctly identified 75 instances where a construct is present or reflected in a user review with a positive sentiment (a positive class, 1). Additionally, GPT-4.1-mini accurately detected 302 instances where a construct is not explicitly reflected, an absent class (0), and 29 instances where a construct is present and reflected in the user review with a negative sentiment, a negative class (-1). However, the ternary encoder misclassified 14 positive cases as absent, 2 positive cases as negative, 6 absent cases as positive, 15 absent cases as negative, 3 negative cases as positive, and 4 negative cases as absent. Despite these misclassifications, the overall high accuracy of correctly encoded instances suggests that the ternary encoding approach effectively captures the intended constructs from user reviews.

Table 2. Confusion Matrix for Ternary Encoding

		GPT-4.1-mini Encoding		
		-1	0	1
Manual Encoding	-1	29	4	3
	0	15	302	6
	1	2	14	75

Table 3 shows the evaluation metrics for the ternary encoding processes using all constructs and reviews, assessing the performance of GPT-4.1-mini in terms of accuracy, precision, recall, and F-measure. The ternary encoding approach achieved an accuracy of 90.2%, with a precision of 90.8%, recall of 90.2%, and an F-measure of 90.4%, indicating strong classification performance.

Table 3. Overall Evaluation Metrics for Ternary Encoding

Accuracy	Precision	Recall	F-measure
0.902	0.908	0.902	0.904

Table 4 shows the evaluation metrics for the ternary encoding processes for each construct using the 50 reviews, assessing the performance of GPT-4.1-mini in terms of accuracy, precision, recall, and F-measure. The results show varying levels of performance across different constructs, with accuracy, precision, recall, and F-measure scores ranging from 0.800 to 1.000. Overall, the findings suggest that some constructs, such as customizability, exhibit near-perfect performance, while others, like perceived usefulness, demonstrate relatively lower performance. Most constructs, however, display strong performance, with accuracy scores above 0.900, indicating a high degree of reliability and effectiveness in their measurement.

Table 4. Evaluation Metrics for each Construct using Ternary Encoding

Construct	Accuracy	Precision	Recall	F-measure
Perceived Ease of Use	0.900	0.919	0.900	0.906
Perceived Usefulness	0.800	0.840	0.800	0.809
Hedonic Motivation	0.940	0.961	0.940	0.945
Device Appeal	0.940	0.904	0.940	0.919
Customizability	1.000	1.000	1.000	1.000
Device Quality	0.880	0.897	0.880	0.883
Perceived Values	0.820	0.866	0.820	0.831
Device Connectivity	0.920	0.928	0.920	0.912
Credibility Support	0.920	0.945	0.920	0.927

Discussion

Results from the case study showed the effectiveness of LLMs, specifically GPT-4.1-mini, in encoding theoretical constructs from user-generated content in the domain of wearable devices. The results demonstrate that LLMs could be an important tool in CSS through the identification of social constructs based on textual data. The results obtained from the ground truth and the GPT-4.1-mini in terms of inter-rater reliability, indicating substantial agreement among the raters, suggest that prompting GPT-4.1-mini with an in-context learning approach could achieve performance comparable to the human experts in terms of detecting constructs from users' reviews. The ternary encoding approach achieved an accuracy of 90.2% and provided insights into how constructs were reflected in user reviews.

The confusion matrices validate the robustness of the encoding process, showing a low number of misclassifications. While some positive constructs were misclassified as absent and vice versa, the overall precision and recall scores suggest that GPT-4.1-mini is highly effective in encoding constructs. The ternary encoding approach is particularly valuable in scenarios where sentiment is important, as it differentiates whether a construct was present with a positive sentiment or a negative one reflecting users' experiences with wearable devices in relation to the selected constructs. Results showed the potential of LLMs, in this case GPT-4.1-mini, to enhance traditional methods of studying social phenomena. Unlike conventional surveys and focus groups, which often suffer from low participation rates and high costs, automated construct encoding allows for scalable, real-time analysis of user perceptions.

While the study achieved high agreement between manual and AI-generated encodings, cases of misclassification suggest that LLMs may still struggle with context-dependent meanings and linguistic variations. However, results based on these datasets serve as proof of concept and the ability to encode theoretical constructs using LLM. Future research could further analyze the sensitivity of the output to in-context learning. Moreover, it is worth noting that in many experiments, human experts are prone to errors, such as subjective bias, inconsistency, and lack of attention. According to the literature, LLMs can also exhibit bias and be inconsistency (Bail, 2024). Overall, this study demonstrates that LLMs, such as GPT-4.1-mini, offer a powerful and scalable tool for analyzing large-scale social data. The findings pave the way for further integration of generative AI in CSS, particularly in areas where traditional methods face limitations in scale and efficiency.

Conclusion

This study demonstrated the potential of LLMs, specifically GPT-4.1-mini, in encoding theoretical constructs from user reviews in the domain of wearable devices. The results highlighted the effectiveness of GPT-4.1-mini in CSS research, especially for encoding constructs. Ternary encoding provided valuable sentiment insights, making it useful for different analytical needs. Despite minor misclassifications, the results suggest that GPT-4.1-mini can complement traditional research methods by enabling scalable, real-time analysis of social phenomena. Future research could further refine these models to enhance contextual understanding and adaptability across various domains.

References

Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2025). Many-Shot In-Context Learning. *Advances in Neural Information Processing Systems*, *37*, 76930–76966.

Alviano, M., & Grillo, L. (2024). *Answer Set Programming and Large Language Models interaction with YAML: Preliminary Report*. CEUR-WS. https://iris.unical.it/handle/20.500.11770/376947

Azad, S. (2024). The Effectiveness of GPT-4 as Financial News Annotator Versus Human Annotator in Improving the Accuracy and Performance of Sentiment Analysis. In O. P. Verma, L. Wang, R. Kumar, & A. Yadav (Eds.), *Machine Intelligence for Research and Innovations* (pp. 105–119). Springer Nature. https://doi.org/10.1007/978-981-99-8129-8_10

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Belal, M., She, J., & Wong, S. (2023). *Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis* (arXiv:2306.17177). arXiv. https://doi.org/10.48550/arXiv.2306.17177

Bijker, R., Merkouris, S. S., Dowling, N. A., & Rodda, S. N. (2024). ChatGPT for Automated Qualitative Research: Content Analysis. *Journal of Medical Internet Research*, *26*(1), e59050. https://doi.org/10.2196/59050

Bryda, G., & Sadowski, D. (2024). From Words to Themes: AI-Powered Qualitative Data Coding and Analysis. In J. Ribeiro, C. Brandão, M. Ntsobi, J. Kasperiuniene, & A. P. Costa (Eds.), *Computer Supported Qualitative Research* (pp. 309–345). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65735-1_19

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511–521. https://doi.org/10.1016/j.dss.2010.11.009

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, *63*, 67–80. https://doi.org/10.1016/j.dss.2013.08.008

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *0*(0). https://doi.org/10.1016/j.patter.2025.101260

Christou, P. (2024). Thematic Analysis through Artificial Intelligence (AI). *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2024.7046

Crimson Hexagon and Brandwatch. (2020). Brandwatch. https://www.brandwatch.com/p/crimson-hexagon/

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., & Sui, Z. (2024a). *A Survey on In-context Learning* (arXiv:2301.00234). arXiv. https://doi.org/10.48550/arXiv.2301.00234

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., & Sui, Z. (2024b). *A Survey on In-context Learning* (arXiv:2301.00234). arXiv. https://doi.org/10.48550/arXiv.2301.00234

Elnoshokaty, A., El-Gayar, O., Wahbeh, A., Al-Ramahi, M., & Nasralah, T. (2022). Drivers and Challenges of Wearable Devices Use: Content Analysis of Online Users Reviews. *AMCIS 2022 Proceedings*. https://aisel.aisnet.org/amcis2022/sig_health/sig_health/20

Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, *90*, 1–11. https://doi.org/10.1016/j.dss.2016.06.010

Gao, J., Choo, K. T. W., Cao, J., Lee, R. K.-W., & Perrault, S. (2023). CoAIcoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Trans. Comput.-Hum. Interact.*, *31*(1), 6:1-6:38. https://doi.org/10.1145/3617362

Gao, X., & Das, K. (2024). Customizing Language Model Responses with Contrastive In-Context Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), Article 16. https://doi.org/10.1609/aaai.v38i16.29760

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. *International Journal of Qualitative Methods*, *22*, 16094069231201504. https://doi.org/10.1177/16094069231201504

Hitch, D. (2024). Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research*, *34*(7), 595–606. https://doi.org/10.1177/10497323231217392

Joel-Edgar, S., & Pan, Y.-C. (2024). Generative AI as a Tool for Thematic Analysis: An Exploratory Study with ChatGPT. *UK Academy for Information Systems Conference Proceedings 2024*. https://aisel.aisnet.org/ukais2024/8

Karapanos, E., Gouveia, R., Hassenzahl, M., & Forlizzi, J. (2016). Wellbeing in the Making: Peoples' Experiences with Wearable Activity Trackers. *Psychology of Well-Being*, *6*(1), 4. https://doi.org/10.1186/s13612-016-0042-6

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, *29*(9), 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Li, K. D., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study. *Journal of Medical Internet Research*, *26*(1), e56500. https://doi.org/10.2196/56500

Mathebula, M., Modupe, A., & Marivate, V. (2024). ChatGPT as a Text Annotation Tool to Evaluate Sentiment Analysis on South African Financial Institutions. *IEEE Access*, *12*, 144017–144043. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3464374

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1), 20531680241231468. https://doi.org/10.1177/20531680241231468

Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, *12*, 71876–71900. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3402809

OpenAI. (2025). *Introducing GPT-4.1 in the API*. https://openai.com/index/gpt-4-1/

Park, E. (2020). User acceptance of smart wearable devices: An expectation-confirmation model approach. *Telematics and Informatics*, *47*, 101318. https://doi.org/10.1016/j.tele.2019.101318

Perkins, M., & Roe, J. (2024). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.22

Ryan, J., Edney, S., & Maher, C. (2019). Anxious or empowered? A cross-sectional study exploring how wearable activity trackers make their owners feel. *BMC Psychology*, *7*(1), 42. https://doi.org/10.1186/s40359-019-0315-y

Saouabe, A., Oualla, H., & Mourtaji, I. (2024). Data Encoding with Generative AI: Towards Improved Machine Learning Performance. | EBSCOhost. *International Journal of Advanced Computer Science & Applications*, *15*(10), 53. https://doi.org/10.14569/ijacsa.2024.0151007

Sethi, V., & King, W. R. (1991). Construct Measurement in Information Systems Research: An Illustration in Strategic Systems. *Decision Sciences*, *22*(3), 455–472. https://doi.org/10.1111/j.1540-5915.1991.tb01274.x

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), Article 7972. https://doi.org/10.1038/s41586-023-06291-2

Sinha, R., Solola, I., Nguyen, H., Swanson, H., & Lawrence, L. (2024). The Role of Generative AI in Qualitative Research: GPT-4's Contributions to a Grounded Theory Analysis. *Proceedings of the 2024 Symposium on Learning, Design and Technology*, 17–25. https://doi.org/10.1145/3663433.3663456

Sjöklint, M., Constantiou, I., & Trier, M. (2015). The Complexities of Self-Tracking—An Inquiry into User Reactions and Goal Attainment. *ECIS 2015 Completed Research Papers*. https://doi.org/10.18151/7217479

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods*, *23*, 16094069241231168. https://doi.org/10.1177/16094069241231168

Vujinović, A., Luburić, N., Slivka, J., & Kovačević, A. (2024). Using ChatGPT to annotate a dataset: A case study in intelligent tutoring systems. *Machine Learning with Applications*, *16*, 100557. https://doi.org/10.1016/j.mlwa.2024.100557

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 75–78. https://doi.org/10.1145/3581754.3584136

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*(3), 103665. https://doi.org/10.1016/j.ipm.2024.103665

Zhang, Y. (2023). *Generative AI has lowered the barriers to computational social sciences* (arXiv:2311.10833). arXiv. https://doi.org/10.48550/arXiv.2311.10833

Zhao, F., Yu, F., & Shang, Y. (2024). A New Method Supporting Qualitative Data Analysis Through Prompt Generation for Inductive Coding. *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 164–169. https://doi.org/10.1109/IRI62200.2024.00043

Appendix A: Example Prompt for the Ease-of-Use Construct

```
def build_prompt(construct, review_text):
    prompt = f"Construct: {construct['name']}\n"
    prompt += f"Definition: {construct['definition']}\n"
    prompt += "Measurement Items:\n"
    for item in construct['measurement_items']:
        prompt += f"- {item}\n"
    prompt += "\nExamples:\n"
    for ex in construct['examples']:
        label = {1: "Positive", -1: "Negative", 0: "Not Reflected"}.get(ex['label'], "Unknown")
        prompt += f'Statement: "{ex["statement"]}"\n Reflects Construct? {label}\n'
    prompt += "\n---\n"
    prompt += f'Statement: "{review_text}"\n Reflects Construct? (Answer only one of: Positive, Negative, Not
Reflected)'
    return prompt
#---------------------------------------------------------------------------------------------------------------------#
```

Constructs

name: Perceived Ease of Use

definition: Perceived ease of use is defined as the degree to which a person believes that using a particular system would be free of effort. In the context of wearable devices, it refers to users' perception of how using a wearable device is free of effort. In general, users are more satisfied with a wearable device when the device requires little effort. According to the literature, several studies have examined users' satisfaction and acceptance of wearable devices and showed that perceived ease of use positively impacts users' satisfaction.

measurement_items:

- I find the wearable device easy to use.
- Learning to operate the wearable device is straightforward for me.
- Interacting with the wearable device does not require a lot of mental effort.
- I find it easy to get the wearable device to do what I want it to do.
- The functions of the wearable device are clear and understandable.
- I can use the wearable device without written instructions.
- I find it easy to navigate through the features of the wearable device.

examples:

- *statement: It was very easy to set up and the app is very easy to use.*
  label: 1
- *statement: The band was tight and uncomfortable at first.*
  label: 0
- *statement: It is not easy to use*.
  label: -1