Introducing ChatGPT to a researcher's toolkit: An empirical comparison between a rule-based and a large language model approach in the context of qualitative content analysis of political texts in Finnish

Ilona Kousa
University of Helsinki
ilona.kousa@helsinki.fi

Abstract

Large Language Models, such as ChatGPT, offer numerous possibilities and prospects for academic research. However, there has been a gap in empirical research regarding their utilisation as keyword extraction and classification tools in qualitative research; perspectives from the social sciences and humanities have been notably limited. Moreover, Finnish-language data have not been used in previous studies. In this article, I aim to address these gaps by providing insights into the utilisation of ChatGPT and drawing comparisons with a rule-based Natural Language Processing method called Etuma. I will focus on assessing the effectiveness of classification and the methods' adherence to scientific principles. The findings of the study indicate that the classic recall and precision trade-off applies to the methods: ChatGPT's precision is high, but its recall is comparatively low, while the results are the opposite for Etuma. I also discuss the implications of the results and outline ideas for leveraging the strengths of both methods in future studies.

1 Introduction

The field of Natural Language Processing (NLP) has recently undergone a significant transformation, largely driven by the widespread adoption and popularity of large language models (LLMs). LLMs, such as ChatGPT, offer numerous possibilities and prospects for academic research as well. Many researchers who have previously relied on traditional NLP methods are now considering the future trajectory of the field. The question arises: will other NLP methods become obsolete, with LLM applications replacing them in research?

Since the launch of ChatGPT in November 2022, there has been extensive discussion within the scientific community, and research articles have been published at an accelerated pace. Many of these studies have demonstrated that ChatGPT's performance in various tasks is comparable to that of humans in terms of quality.

In a study by Huang et al., (2023), ChatGPT was able to identify implicit hate speech well compared to humans. Guo et al., (2023) found that ChatGPT's capabilities to answer questions from several domains including finance, medicine, law, and psychology, were on par with those of human experts. Gilardi et al., (2023) reported that ChatGPT even outperformed humans in annotation tasks including relevance, stance, topics, and frames detection. On the other hand, ChatGPT's ability to produce consistent results has been questioned and caution has been advised regarding its application to text classification (Reiss, 2023). Some studies have found ChatGPT's zero-shot performance to be lacking, although prompt engineering and additional training have been shown to improve results (Shi et al., 2023; Yuan et al., 2023).

While ChatGPT has been extensively examined for a diverse range of tasks, there remains a gap in empirical research regarding its utilisation as a classification tool in qualitative research. Furthermore, perspectives from the social sciences and humanities have been notably limited thus far. In addition, Finnish-language data has not been used as research material.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 102–113
December 1-3, 2023

In this article, I will also introduce Etuma, an NLP tool that represents a traditional rule-based approach based on supervised learning methods, dictionaries, and grammar rules. The aim is to highlight the distinctive features of these two different approaches in one of the most common NLP tasks: text classification. I will focus on the qualitative content analysis of extensive datasets in the field of digital humanities, with a particular emphasis on topic classification, a central aspect of qualitative content analysis.

1.1 The scope of the study

The study aims to address the following research questions: 1) What distinguishes rule-based and LLM-produced classification in their effectiveness as qualitative content analysis tools? and 2) How viable are these methods in terms of scientific rigour, considering compliance with scientific principles such as reproducibility and transparency?

The motivation behind this research stems from a project that involves the categorisation of a large volume of Finnish-language texts. In the scope of this article, I will not discuss the underlying project and its results in detail, but rather focus on the description and validation of the methods. The main objective of this study is to describe the characteristics of different approaches and provide information to fellow researchers, who are considering using either method in qualitative content analysis.

In addition, beyond the features outlined in this article, there are several noteworthy concerns regarding the use of LLM techniques in research. These concerns include, for example, plagiarism and other unethical use, as well as challenges related to training data, including bias, misinformation, and vulnerability to adversarial attacks (Ray, 2023).

Firstly, I will describe the research setting, then report on the research materials, methods, and process. Then I will present and discuss the results and their limitations. I will conclude with insights and suggestions for future research.

2 Data and methods

The context of this study is an ongoing research project focusing on political energy discourse in Finland. In the project, my goal is to analyse the public political debate, specifically examining the comments made by citizens and politicians. The objective is to gain a comprehensive understanding of the issues underpinning the energy debate and to explore the various themes that emerge from the collected material. Given the large volume of the research material, I will employ a combination of automatic text analysis and qualitative methods (Guetterman et al. 2018; Jänicke et al. 2015; Grimmer & Stewart, 2013).

In research work, several important criteria should be considered when selecting a method or a research tool. The tool should be suitable for scientific research in general, adhering to rigorous standards of reliability, validity, and ethical considerations. In this case, an important feature was also the tool's proficiency in the Finnish language, to ensure its ability to process and analyse Finnish texts. Additionally, the tool needed to provide a comprehensive overview of large volumes of research material by effectively categorising it into relevant topics.

With these criteria in mind, I sought to explore whether the widely popular ChatGPT could be a suitable method for conducting the analysis required for the project. In my previous work, I have utilised the Etuma tool for keyword extraction and topic classification. Therefore, I decided to compare the two approaches to delineate the strengths and weaknesses of each method.

2.1 Data

The original corpus was collected as a part of the broader research project. It consists of 110,295 social media comments from August 2022 to August 2023 and 25,872 parliament speeches from February 2022 to March 2023. The social media comments were collected from a web scraping tool called Mohawk Analytics (Legentic 2023) and the transcribed parliament speeches were downloaded from a database known as Parliament Sampo (Hyvönen et al, 2022).

For the purposes of this article, I limited the material to a smaller subset so that it would be easier to qualitatively assess the analysis results produced by each method. I employed a keyword search (“electric car” AND “subsidy”; “sähköauto” AND “tuki” in Finnish) to filter texts discussing a specific topic of interest in the project: electric car subsidies offered by the Finnish government. The subset corpus comprised 40 social media comments and 33 parliamentary speeches.

The social media data included 21 tweets from Twitter (currently X), 19 online news comments, and 4 discussion forum posts. The parliament speech corpus consisted of 13 speeches from the Finns party, 3 speeches from the Social Democratic party, 3 speeches from the Centre party, 3 speeches from the Green party, and one speech each from the National Coalition party and the Christian Democrats. In addition, the material included 5 responses from government ministers from the Social Democratic party, the Centre party, and the Green party. The original language of the texts was Finnish, but keywords, topics, and text quotes have been translated into English for this article.

The social media comments were typically short, but their length varied between 20 and 155 words per comment. The comments were critical towards the research topic, as exemplified by statements such as "*Electric car subsidies go to the wealthy and electricity subsidies also benefit the wealthy. Because of the current government, we are all impoverished."*. Several comments included misspelled words.

The parliament speeches were more extensive, with their length varying between 72 and 662 words per speech. The speeches contained a considerable amount of specialised technical and administrative vocabulary, for example "*subsidies for the purchase of electric and gas cars and distribution infrastructure are necessary actions as we move towards a fossil-free transport system*" and did not contain much informal language, typos or misspellings.

I copied the original texts into an Excel file and recorded the analysis results obtained with different methods in their respective columns. To clean the data, I removed mentions targeted to specific users (identified with the '@' character) in social media comments. Additionally, I randomly selected a sample of 10 social media comments and 10 parliamentary speeches for validating the results. I will describe the categorisation and validation processes in more detail in the following sections.

2.2 Methods

In this study, I utilised both Etuma and ChatGPT4 for the identical task: extracting keywords from texts and categorising them into topics. I focused only on keyword and topic classification and excluded sentiment analysis. However, it's important to note that ChatGPT does not extract keywords in the traditional sense. Instead, it generates language based on the patterns it has learned from its training data.

The initial phase of the study started in a zero-shot setting, where no training data or pre-defined categories were used. I analysed the corpus in September 2023 using ChatGPT version 4, accessed through the Poe.com platform, and with Etuma's browser-based NLP tool. I will detail this process further in the subsequent section.

2.3 Research process

Figure 1 presents the key phases in the research process. During the study, I conducted both distant reading and traditional close reading in parallel (Jänicke et al., 2015). During the distant reading phase, I utilised computational methods to analyse the material based on topics and keywords, enabling a systematic examination of the data to identify patterns and trends. In the close reading phase, I engaged in an iterative process to uncover the topics present in the data, as well as their associated keywords and the context in which these keywords were discussed. This approach allowed for a deeper understanding of the data and its nuances. Finally, in the third step, the classification is refined to better align with the broader objectives of the research project, ensuring that it adequately captured the relevant information.

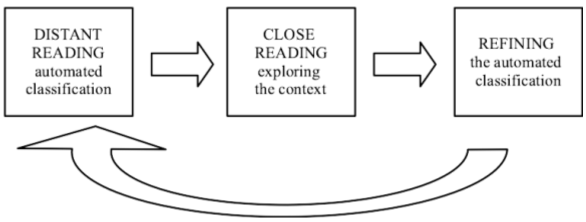

Figure 1: Research process.

The approach of combining distant and close reading has been previously employed successfully. For example, Guetterman et al. (2018) conducted a study where they compared the results of qualitative analysis using three different methods: 1) close reading, 2) automatic text analysis, and 3) a combination of the two, by analysing the same materials in separate research groups. Their findings indicated that the combination of traditional close reading and automated distant reading yielded the most comprehensive, high-quality, and detailed results.

In the following sections, I will describe in more detail the distinctive features of the process for both methods separately.

2.3.1 ChatGPT

The term language model (LM) refers to systems that have been trained to predict the probability of a given token (character, word, or string) (Bender et al., 2021). ChatGPT has been pre-trained on large datasets consisting of web-crawled text, including conversations, and fine-tuned by humans with the Reinforcement Learning from Human Feedback (RLHF) method (OpenAI 2023a).

I employed ChatGPT 4 through the Poe.com platform. Users on this platform can create their own chatbot and customise its settings according to their preferences. This includes configuring a default prompt, which serves as the initial message for the chatbot, as well as setting a temperature value. Increasing the temperature parameter allows the predictive model to take more risks, suggesting less likely alternatives and thereby reducing result consistency (OpenAI 2023b).

The prompt plays a central role in determining what kind of results a ChatGPT-powered bot generates. After some testing with different prompt wordings and temperature values, I created a chatbot with the following prompt: “You are an advanced artificial intelligence for text analysis, and you need to classify given texts based on topics. One sentence can contain more than one topic. Extract as many topics as possible. The temperature setting is 0. Format the output to be a simple list of keywords that appear in the text and what topic the keywords are classified into.”.

The research process is illustrated in Figure 1. During the initial analysis phase, I input the texts individually into the same chat conversation and recorded ChatGPT's responses in an Excel table. In the zero-shot setting, the system autonomously identified 47 topics and 935 keywords within the data. Concurrently, I validated the classification by conducting a close reading of the original texts.

In the second analysis phase I experimented with a few-shot approach, providing more detailed instructions within the prompt about the specific topics I was interested in. I noticed that the more precise my requirements were defined, the better results I obtained. For instance, prompts like "extract relevant keywords and topics related to commuting" or "how are coronavirus aids and electric car subsidies linked in the texts" produced desired results but demanded accurate information or hypotheses about the material's content. In addition, ChatGPT's memory did not extend very far in the conversation, so it could not answer questions about the entire corpus.

In the third phase of the research process, I employed prompt engineering to refine the results and minimize the potential impact of a poorly formatted prompt on the outcomes by following the instructions of White et al., (2023). Among the four prompt enhancement strategies they proposed, I found "Question refinement" to align best with my needs, although in this case it did not lead to an improvement in recall. A report detailing example chat interactions of the prompt engineering experiment can be found in Appendix 2.

2.3.2 Etuma

Etuma’s technological foundation is rooted in NLP research conducted at the University of Helsinki (Lahtinen 2000; Tapanainen 1999), which has since been continued commercially by Etuma (Etuma 2023). Etuma performs several NLP tasks on texts, such as morphological, syntactic, semantic, and sentiment analysis.

A key function of Etuma is ontological classification, based on which it groups keywords referring to the same theme into more general classes called topics. For example, the keywords “*electric car*”, “*e-car*” and “*battery vehicle*” would be categorised into the same topic called Electric cars. It is important to note that although Etuma refers to the classification with the term topic, the method should not be confused with topic modelling methods, which are based on unsupervised machine learning, whereas Etuma employs dictionaries and supervised learning.

Using Etuma, I followed the same research process depicted in Figure 1. The initial analysis step involved uploading the original dataset in CSV format into the Etuma analysis system. Within the Etuma interface, I then applied filters as described above, to extract the specific sub-dataset relevant to this research. In the distant reading phase, the system identified 415 topics and 1621 keywords within the data. During the close reading phase, I conducted a review of the most frequently occurring topics and their corresponding keywords. Then I reviewed less frequent topics at a broader topic-name level.

In the subsequent phase, the emphasis is on refining the classification to improve the relevance and precision of the analysis by merging and splitting topics and transferring keywords between them. Etuma has a built-in user interface for these tasks, as refining the classification is an integral part of the research process. The extent of this phase depends on the goals of the research, the amount of material and precision of the classification. After the classification-validation process is completed, new classification rules are updated to the Etuma system, with the option that the customised rules can be reused. The purpose of the process is to improve the relevance of the classification to adapt to the specific requirements of the study. However, in this article I will focus on the zero-shot situation where no fine-tuning has been implemented.

3 Empirical analysis

In this section, I will present the key findings obtained from the analysis conducted by Etuma and ChatGPT on the corpora. Firstly, I will describe the characteristics of keyword extraction and topic classification for both methods, along with relevant examples. It is important to note that the purpose of these key figures is to compare the classifications, without taking a stance on what constitutes the ideal classification. Secondly, I will present a comparison of the methods using a smaller sample, employing traditional metrics such as recall, precision, and F1 score. This analysis will provide a quantitative evaluation of the performance of each method. Additionally, Appendix 1 contains a list of the most frequently occurring topics and keywords identified during the analysis.

3.1 Classification characteristics

Table 1 illustrates the differences in the number of unique keywords and topics identified by each method. As a general observation, ratio between the number of topics and keywords was similar in both corpora, indicating that the text type had no significant effect on the results.

		ChatGPT 4	Etuma
Social media	Keywords	246	435
	Topics	15	144
Parliament speeches	Keywords	722	1311
	Topics	40	378

Table 1: Unique keywords and topics in corpora

Keywords Both methods successfully analysed the Finnish-language material without significant deficiencies or shortcomings. However, there were differences in the keywords produced by the methods. The most noticeable difference was in the number of keywords: Etuma extracted more than one and a half times the number of unique keywords compared to ChatGPT. Additionally, Etuma tended to have more one-word keywords and ChatGPT generated more multi-word keywords.

The parliamentary speeches contained many acronyms. Both methods correctly classified common abbreviations such as EU (the European Union) and Yle (the Finnish Broadcasting Company). Etuma also extracted some acronyms from the parliamentary speeches (e.g., MAL, KAISU) but did not classify them to an exact topic. Initially, ChatGPT did not recognize these acronyms as keywords. When prompted separately, ChatGPT correctly classified MAL as "Maankäyttö, asuminen ja liikenne" (Land use, housing and transport) but did not identify "*KAISU*" as "Keskipitkän aikavälin ilmastopolitiikan suunnitelma" (Medium-term climate change policy plan).

ChatGPT correctly classified more names of Members of Parliament (e.g., Li, Tynkkynen) compared to Etuma. Typos and slang are common in social media materials. Etuma provides a list of keywords it does not recognize, and among them, there were 31 unique keywords that were misspelt and thus left uncategorized. Based on my observations, ChatGPT analysed typos correctly more frequently. However, a detailed analysis of the feature was not conducted in this study.

Topics In terms of unique topics, the difference between Etuma and ChatGPT was even more pronounced, almost tenfold. As can be deduced from the results, ChatGPT tended to employ broader topics (Economy, Politics), while Etuma's classification was more granular (Subsidies, Social security). Furthermore, it is worth noting that some of ChatGPT's unique topics overlapped (e.g., "Economics", "Economics and Finance", "Economy", "Economy and Finance"), leading to even fewer distinct classification themes than the count of unique topic names identified.

Hallucination On a few occasions, ChatGPT demonstrated a behaviour known as hallucination, where it generated information that was not accurate or factual. For instance, it asserted that

"*Sulo Vileen*" (referring to a character from a Finnish TV series) is a colloquial term for ordinary Finns, akin to Joe Public in English. This occurrence also points to a limitation related to the training data in Finnish.

Prompting I tested various prompts with ChatGPT and repeated identical prompts in new chat interactions, which revealed that classification results for the same piece of text could change even though the content and prompt remained identical. As an example, during the initial analysis round, ChatGPT classified various keywords such as "*travelling to Spain*" "*musicians*" and "*price range*" under the same topic Social issues. However, in a new chat interaction, these same keywords were classified as International travel, Arts/Culture and Economy. This suggests that ChatGPT may have tried to simplify the classification by grouping less precise keywords into a smaller set of topics, indicating an internal learning mechanism guiding the classification.

3.2 Validation

To gain a more detailed understanding of the recall and precision levels of the methods, I conducted a comparative analysis with human classification. This involved calculating the traditional metrics of recall, precision, and the F1 score. During the validation phase, I randomly selected a sample of twenty texts from the material, consisting of ten social media posts and ten parliamentary speeches. Then I manually classified the texts by extracting the relevant keywords from them. At this stage, I tagged all potentially interesting keywords in the texts through which it would be possible to examine the material from various perspectives. Similarly, I did not provide specific instructions to Etuma and ChatGPT regarding the types of keywords to extract. As a result, I tagged a total of 151 keywords from the social media sample and 311 keywords from the parliament speech sample.

For each method, I compared the classification results with the human classification and calculated the recall using the following formula:

$$\frac{relevant\ extracted\ keywords}{all\ relevant\ keywords}$$

In addition, I calculated precision by reviewing the classification results and determining the number of keywords that were either left unclassified or classified incorrectly. The formula I used to calculate precision is as follows:

$$\frac{correctly\ classified\ keywords}{all\ extracted\ keywords}$$

The F1 score, a balanced measure that considers both precision and recall, is calculated as the harmonic mean of the two. I calculated the F1 score using the following formula:

$$2 * \frac{recall * precision}{recall + precision}$$

Table 2 presents the recall and precision levels, along with the F1 score that combines both metrics.

		ChatGPT 4	Etuma
Social media	**Precision**	0.96	0.70
	Recall	0.61	0.85
	F1 score	0.75	0.77
Parliament speeches	**Precision**	0.96	0.70
	Recall	0.58	0.81
	F1 score	0.72	0.75

Table 2 Recall, precision, and F1 score

Recall The recall level of Etuma's classification was higher in the social media sample (0.85) than in the parliamentary speech sample (0.81). For a single text, the recall ranged from 0.58 to 1.00, with an average of over 0.80 for both text samples. For ChatGPT the recall varied from 0.42 to 1.00 for individual texts, with an overall recall of 0.61 for the social media sample and 0.58 for the parliamentary speech sample.

A possible explanation for the difference between the two text types is that Etuma's tool is optimized for the analysis of relatively short customer feedback and survey responses, and not for the analysis of longer texts (Etuma 2023).

However, also ChatGPT's recall was higher for the social media sample. In the scope of this study, it is difficult to determine whether the difference is only due to the length of the texts, or whether the vocabulary and training materials used in the development of the methods also play a role. However, I observed that social media posts use more common language terms, while parliamentary speeches have more specialised terms that the tools did not always identify as keywords.

Precision Etuma's precision rate was 0.70 for both parliamentary speech texts and social media posts. However, different things affected the precision rate in the two samples. There were more misspelled words in the social media posts while there was more specialised vocabulary in parliamentary speeches. For example, from the sentence "*supplementary budget proposals allocate not only procurement support towards electric cars, but also support for ethanol and gas conversion opportunities*" Etuma did not recognize that the keyword "*gas conversion opportunities*" (kaasukonversiomahdollisuus in Finnish) referred to gas cars in this context.

ChatGPT's precision was high, 0.96 for both samples. Errors typically related to the interpretation of the correct topic, rather than to keyword extraction. For example, from the sentence in a social media post "*With this populist fake news, you can get a few votes in the elections, and nothing else*", ChatGPT classified the keyword "*elections*" (vaalit in Finnish) into a topic called Politics and the keyword "*votes*" (äänet in Finnish) into topic Social issues. In Etuma's analysis the precision rate was predominantly affected by uncategorised keywords. The results indicate that the precision of the results obtained is not significantly influenced by the type of text being analysed.

F1 score The F1 score, which takes into account both recall and precision, was slightly higher for Etuma in both the social media and parliament speech samples.

4 Discussion

In this section, I revisit the research questions I presented in the introduction. Firstly, I discuss the effectiveness of the classification in qualitative content analysis from the perspectives of key aspects such as recall, granularity, precision, and refinement. Secondly, I assess the alignment of the methods with scientific principles, specifically focusing on repeatability, transparency, and research integrity.

4.1 Effectiveness of classification

Recall In this study, Etuma's recall was higher compared to that of ChatGPT. The result reveals a fundamental difference between the approaches. While ChatGPT concentrates on summarising the content, Etuma aims to provide a comprehensive description of the content.

Granularity ChatGPT focused on the main points and tended to overlook rhetorical expressions and topics mentioned less frequently and more indirectly. A lack of detail was also observed in a previous study when comparing ChatGPT's responses to those of human experts (Guo et al., 2023). In situations where the corpus contains a significant amount of noise or irrelevant data, ChatGPT's ability to emphasise essential information can be beneficial. However, there are scenarios where researchers specifically seek nuanced details and rhetorical language, which may not align with ChatGPT's primary focus.

Precision As anticipated from prior research (Ortega-Martín et al., 2023), ChatGPT' performed well in semantic disambiguation and integrating cultural context into its classification. The adaptability of information related to cultural context stands out as notable strength of LLMs. Spelling mistakes and specialised vocabulary are more challenging for a dictionary-based approach because it is not feasible to add all possible spelling variants and special vocabulary to the ontology. Even though both methods are susceptible to the exclusion of specific terms, abbreviations, and misspelled words based on the vocabularies and training data utilised, this study revealed that ChatGPT outperformed Etuma in these regards.

In this study, there were no noticeable deficiencies in the knowledge of the Finnish language for either method. While I did not experiment with other languages, it is important to note that analysing a less common language like Finnish might not be as accurate or comprehensive due to the limited training material available.

Refinement A high recall or precision score does not automatically imply the relevance of results to the researcher. As I have described in this article, an important part of the research process is the validation and fine-tuning of the results in an iterative process. The workload involved in this step depends on the recall and precision of the initial analysis performed by the automated method. If the recall rate is high, it might be possible to enhance precision by refining the analysis. Conversely, if the precision rate is extremely low, the researcher faces a substantial workload in validating and fine-tuning the classification.

In this study, my attempts to fine-tune ChatGPT's results were not successful, as demonstrated in Appendix 2. However, if employed differently, it may be feasible to fine-tune ChatGPT's classification as well. On the other hand, Etuma has built-in tools designed to improve recall and precision as it is part of the method's standard process.

4.2 Compliance with scientific principles

Repeatability The methods differ in terms of reproducibility due to their distinct approaches. With the Etuma tool, the outcome of the analysis remains consistent, unless the researcher alters the classification rules. In contrast, a characteristic of ChatGPT is that identical input can yield different outputs. Moreover, during this study, I noticed that ChatGPT produced different results from the same text using the same prompt, a phenomenon that is in line with findings from earlier research (Ortega-Martín et al., 2023; Reiss, 2023).

In this regard, the method resembles qualitative analysis conducted by human analysts, as the classification performed by two different individuals may not be identical. A potential way to address this challenge could involve using similar approaches used to enhance the validity and reliability of human classification, like independently annotating the same material several times and then comparing the results.

Transparency With ChatGPT, transparency was impacted by the challenge of generating a manageable classification structure that could be easily documented and refined. ChatGPT operates as more of a black box, while Etuma offers greater transparency due to its classification being built upon predefined dictionaries.

The fine-tuning process of Etuma's classification is characterised by transparency and repeatability, as it is largely done manually, and every change leaves a trace in the system log. However, a challenge emerges from the extensive scope of classification, often requiring researchers to narrow their focus to, for example, a smaller subset of the corpus or the most prevalent topics.

Research integrity Despite the surrounding technological hype, researchers bear the responsibility to ensure that new technologies are not adopted too uncritically for scientific use. For example, various biases and information distortions due to training data and processes is an area that should be discussed. While this material appeared to be free from evident bias, it is important to acknowledge that in other types of content, biases may emerge. Additionally, ChatGPT's tendency to produce hallucinations, or inaccurate information, underscores the need for cautious evaluation of the data it generates. Furthermore, manually validating the analysis of a vast data set can be challenging, potentially allowing biases to go unnoticed.

In a broader perspective, it is important to consider the implications of tool development on the work of researchers. The findings of this study indicate that LLMs assume a significant portion of decision-making on behalf of researchers. While the idea of reducing workload is appealing, it is important to ensure that the autonomy of researchers is not compromised, potentially impacting the research process and even the results. As an example, an attempt to summarise complex information into broad topics may inadvertently overlook nuances or lead to potentially incorrect interpretations.

Moreover, relying solely on automated analysis tools can potentially direct researchers towards formulating research questions that align with the capabilities of the tools, rather than prioritising a comprehensive understanding of the phenomenon being studied. Additionally, it's important to note that although these tools are becoming more accessible, they do not always assure time savings or superior quality compared to manual methods. The utilisation of these tools can also be constrained by the fact that certain tools, such as public language model tools, may not be suitable for analysing sensitive data.

4.3 Strengths and limitations

The study has several strengths. Firstly, it addresses an existing knowledge gap by exploring the application of ChatGPT as a tool for qualitative analysis in Finnish. In addition, the perspective of the research is broadened using two distinct corpora. The study offers comparative insights for researchers who are considering employing either a large language model or a rule-based NLP approach for their analysis.

A limitation of the study is that the material is relatively narrow and focused on one specific research topic. Expanding the scope of the study would enhance the generalisability of the findings and provide a more comprehensive understanding of the methods' capabilities.

5 Conclusions

To summarise my findings, the utilisation of ChatGPT as a research tool poses challenges to the reliability of the research due to issues of repeatability and transparency. In the context of result usability, challenges arise from the occurrence of hallucinations and potentially from low recall.

A major limitation of a low recall rate is that it excessively restricts the researcher's autonomy in decision-making. In this project, my aim was to conduct a comprehensive classification that allows for a qualitative analysis of the material from various perspectives. Hence, I do not want the tool to determine what is important or interesting in the text on my behalf.

On the other hand, a drawback of the rule-based approach often lies in its lack of semantic meaning and context. Nevertheless, this deficiency can be addressed through refining, which, at least with the tools employed in this study, proves to be more straightforward with a rule-based method.

5.1 Implications for future research

In scientific research, repeatability and transparency are important features and the classification of qualitative content demands consistency, validity, and reliability. While ChatGPT may not yet substitute traditional NLP methods in these regards, it undeniably possesses strengths such as adept semantic analysis and information of cultural contexts.

In future research, it would be interesting to employ the methods in parallel and harness the strengths of both. Throughout the research process, I conceived numerous ideas on how to integrate the methods (indicated with a dashed line in Figure 2). The goal would be to utilise ChatGPT in a manner that ensures its shortcomings do not compromise the scientific principles.

Leveraging the LLM's capacity for semantic interpretations could enhance the semantic classification of another NLP method in the zero-shot phase, assisting in semantic filtering of research material based on the studied phenomenon.

In the close reading phase, LLM could aid researchers by generating automated summaries or in interpreting ambiguous or complex texts, suggesting alternative meanings and context to researchers in the validation process. The knowledge within the LLM based on the vast training data could extend beyond the corpus, aiding in the analysis of social discourse, for instance.

Furthermore, in the classification refinement phase, LLM's ability to identify semantic meanings and its creative capabilities could be used to formulate new topics or classification frameworks based on the feedback from the validation process.

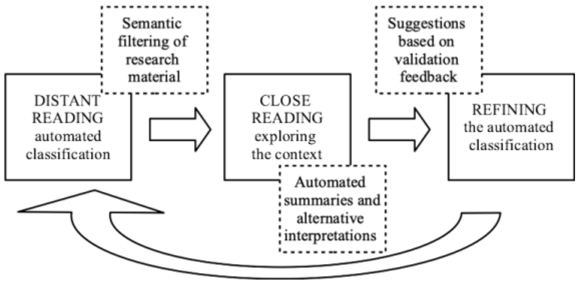

Figure 2: Research process combining rule-based and LLM approaches.

To address the question posed in the introduction about other NLP methods becoming obsolete, it is important to recognise that currently the principles of scientific research prevent ChatGPT from being a direct replacement for traditional NLP methods, at least in my research. However, its distinct advantages make it a potential complement to these methods, thereby enhancing my research toolkit.

Acknowledgments

I thank the Finnish Cultural Foundation, the Kone Foundation, and the Niilo Helander Foundation for supporting my research. I am also grateful to the anonymous reviewers and my colleagues for their comments on the manuscript.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Etuma (2023). *Etuma Natural Language Processing. Internal document.* Accessed on 26.9.2023.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. https://doi.org/10.48550/ARXIV.2303.15056

Grimmer, J., & Stewart, B. M. (2013). *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). *Augmenting qualitative text analysis with natural language processing: methodological study. Journal of medical Internet research,* 20(6): e231.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.* https://doi.org/10.48550/ARXIV.2301.07597

Huang, F., Kwak, H., & An, J. (2023). *Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. Companion Proceedings of the ACM Web Conference* 2023, 294–297. https://doi.org/10.1145/3543873.3587368

Hyvönen, E., Sinikallio, L., Leskinen, P., La Mela, M., Tuominen, J., Elo, K., Drobac, S., Koho, M., Ikkala, E., Tamper, M., Leal R. and Kesäniemi J. (2022). Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language. *Digital Parliamentary data in Action (DiPaDA 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference*, long paper, pp. 69-85, CEUR Workshop Proceedings, Vol. 3133, May, 2022.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. Eurographics Conference on Visualization (EuroVis) - STARs,* 21 pages. https://doi.org/10.2312/EUROVISSTAR.20151113

Lahtinen, T. (2000). *Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. Doctoral dissertation.* University of Helsinki.

Legentic (2023). *Legentic platform.* Accessed on 26.9.2023 at https://legentic.com/platform

OpenAI (2023a). *What is ChatGPT?* Accessed on 26.9.2023 at: https://help.openai.com/en/articles/6783457-what-is-chatgpt

OpenAI (2023b). *Quickstart. Adjust your settings.* Accessed on 25.9.2023 at: https://platform.openai.com/docs/quickstart/adjust-your-settings

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). *Linguistic ambiguity analysis in ChatGPT.* https://doi.org/10.48550/ARXIV.2302.06426

Ray, P. P. (2023). *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems,* 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark.* https://doi.org/10.48550/ARXIV.2304.11085

Shi, Y., Ma, H., Zhong, W., Tan, Q., Mai, G., Li, X., Liu, T., & Huang, J. (2023). *ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs.* https://doi.org/10.48550/ARXIV.2305.03513

Tapanainen, P. (1999). *Parsing in two frameworks: finite-state and functional dependency grammar. Doctoral dissertation.* University of Helsinki.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.* https://doi.org/10.48550/ARXIV.2302.11382

Yuan, C., Xie, Q., & Ananiadou, S. (2023). *Zero-shot Temporal Relation Extraction with ChatGPT.* https://doi.org/10.48550/ARXIV.2304.05454

A Appendices

Appendix 1 Most frequent topics and keywords

Ranking	ChatGPT 4 topic and frequency	ChatGPT 4 keywords (translated into English)	Etuma topic and frequency	Etuma keywords (translated into English)
1.	Economy (n=53)	“wealthy”, “electric car subsidy”, “public spending”	Subsidies (n=72)	“electric car subsidy”, “coronavirus aid”
2.	Politics (n=53)	“government”, “left-wing”, “vote”	Cars (n=42)	“electric car”, “car”
3.	Social Issues (n=41)	“rural areas, “economic hardship”, “social and health services”	Government Organizations (n=34)	“government”, “EU”, “IMF”
4.	Environment (n=27)	“forest conservation”, “nature restoration”, “swamps	Fuel (n=20)	“fuel”, “gasoline”, “diesel”
5.	Automotive Industry (n=18)	“electric car subsidy”, “internal combustion engine”, “electric car”	Prices (n=19)	“price”, “expensive”, “cheap”
6.	Energy (n=14)	“electricity prices”, “energy prices”, “fossil fuels”	Purchasing (n=18)	“buy”
7.	Transportation (n=12)	“fuel taxation”, “electric car use”, “diesel cars”	Energy (n=15)	“electricity”, “energy”
8.	Economics (n=8)	“fuel tax”, “transportation costs”, “costs”	Payments (n=14)	“pay”
9.	Society (n=8)	“mail delivery”, “public transportation”, “accessible cars”	Taxes (n=12)	“taxes”
10.	Agriculture (n=7)	“agriculture”, “food producers”, “grain”	Income (n=11)	“low income”, “high income”
11.	Infrastructure (n=7)	“basic route maintenance”, “development project costs”	Climate Change (n=9)	“climate change”
12.	Economics and Finance (n=6)	“acquisition support”, “date of purchase”	Debt (n=9)	“debt”, “national debt”
13.	Industry (n=6)	“game companies”	Social Security (n=8)	“child benefit”, “state aid”
14.	Energy and Environment (n=5)	“climate actions”, energy self-sufficiency”, “energy crisis”	Wealthy (n=6)	“wealthy”
15.	Policy and Planning (n=5)	“support procedure”, “steering mechanisms”, “goals”	Saving (n=5)	“saving”

Appendix 2 Examples of prompt engineering attempts with ChatGPT 4 in September 2023

Chat interaction 1

Researcher	*“In the scope of text classification, suggest a better question that would allow me to get the most accurate classification of all possible keywords and topics that appear in the texts I provide.”*
ChatGPT 4	*“Given the following text, could you please identify and categorize all the keywords and topics present? Please provide a detailed breakdown of each keyword and topic and explain why each keyword or topic falls into its respective category based on the context of the text.”*
Researcher	*“Here is the text to be classified: [a social media comment from the corpus]”*

Result: One less keyword was extracted, and two additional topics were uncovered in the results following prompt engineering. Similar results were obtained with another social media comment and two parliamentary speeches.

Chat interaction 2

Researcher	*“In the scope of text classification, suggest a better question that would allow me to get the most accurate classification of all possible keywords and topics that appear in the texts I provide.”*
ChatGPT 4	*“Sure, here's a revised version of your question:* *"Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category."”*
Researcher	*“Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category.”*
ChatGPT	*“I'm sorry for any confusion, but as a text-based AI model, I don't have the capability to process a set of texts in real-time or perform keyword extraction and topic classification on a large dataset. However, I can certainly guide you on how you might approach this task using various natural language processing (NLP) techniques.* *Here's a general approach you might consider:* *Text Preprocessing: This involves cleaning the text by removing punctuation, converting all text to lowercase, removing stop words (common words like 'and', 'the', 'is', etc.), and stemming/lemmatization (reducing words to their root form).* *Keyword Extraction: There are several techniques for keyword extraction, including TF-IDF (Term Frequency-Inverse Document Frequency), TextRank, and RAKE (Rapid Automatic Keyword Extraction). These methods can help identify the most important words in each text.* *Topic Classification: Once you have the keywords, you can use them to classify the topics. One common approach is Latent Dirichlet Allocation (LDA), which is a type of probabilistic model that assumes each document is a mixture of a certain number of topics.* *The output of this process could be a list where each entry includes the keyword and its associated topic category. For example:* *- Keyword: 'climate', Topic: 'Environment'* *- Keyword: 'election', Topic: 'Politics'* *- Keyword: 'AI', Topic: 'Technology'* *Please note that this is a simplified explanation of the process and actual implementation might require a deeper understanding of NLP and machine learning techniques.”*

Result: No keywords or topics.

Chat interaction 3

Researcher	*“Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category.”*
ChatGPT 4	*“Sure, I can certainly help with that. However, I need some example texts to work with. Could you please provide some?”*
Researcher	*“Here is the text: [a social media comment from the corpus]”*

Result: Two more keywords extracted, and two fewer topics compared to the previous classification.