

Applying computational approaches to energy discourse: a comparative methodological study of rule-based and large language model qualitative content analysis

Ilona Kousa^{1*}

1 University of Helsinki, Finland

*Corresponding author: Ilona Kousa ilona.kousa@helsinki.fi

Abstract

The objective of this study is to conduct a comparative analysis of a rule-based ontological classification tool and a large language model (LLM) chatbot as qualitative content analysis tools. The focus is on assessing their strengths and limitations when applied to the study of the discourse surrounding the energy transition. To achieve this, I used the tools to analyse two different types of corpora: citizens' social media discussions and politicians' parliamentary speeches. In the analysis, I evaluated the differences in the methods' recall and precision levels. Additionally, I assessed the extent to which these methods align with scientific principles, including reliability, transparency, and research integrity. The results reveal a classic trade-off: LLM's precision is high, but its recall is comparatively low, suggesting its strength in generating accurate but potentially incomplete analyses. On the other hand, the rule-based method outperforms in recall at the expense of precision, capturing more data points, but with varying accuracy levels. I discuss the implications of the results and outline ideas for leveraging the strengths of both methods in future studies. This article provides researchers with insights into the selection and application of computational tools in the social sciences and humanities, as well as in multifaceted research topics, such as energy transition.

keywords

LLM, rule-based NLP, qualitative content analysis, classification

I INTRODUCTION

"Investments in the green transition are essential for our environment, economy, and security... Investments enable, for example, purchase and conversion subsidies for low-emission vehicles. We prevent mobility poverty by providing support and alternative modes of transportation to people."

A member of a government party at the Finnish parliament in June 2022

"The green transition has driven energy prices up... the government is pouring taxpayers' money into electric vehicle support... it has ripped off the money for the green overhaul at the expense of citizens' purchasing power and daily survival and put the state in disastrous debt."

An anonymous commenter on an online news site in January 2023

The comments extracted from my research data are examples of the multifaceted nature of the energy discourse and the contradictions between the views of citizens and politicians. The first comment supports the government's policy, whereas the second opposes it. Studying discourses is an important part of understanding the energy transition from fossil fuels to renewable energy

sources, as they are intertwined with prevailing energy practices and are at the same time important factors in explaining change [Buschmann and Oels, 2019]. My goal in this article is to consider feasible approaches for studying energy discourse on a larger scale, through thousands or even millions of comments from citizens and politicians. Qualitative analysis of such large datasets is challenging, if not impossible, for humans; therefore, they require the adoption of automated text analysis methods [Grimmer and Stewart, 2013; Van Atteveldt et al., 2019]. In this article, I will employ two different approaches for the task and discuss their differences, as well as their strengths and limitations.

The utilisation and application of big data, artificial intelligence, and advanced computational methods continue to expand exponentially, reaching into new areas of research. Scholars from various fields, including social sciences and humanities, are exploring new methods to address complex societal challenges [Andreotta et al., 2019; Halford and Savage, 2017; Ziems et al., 2023]. Among these challenges, one area that has received considerable research attention is the energy transition. Researchers emphasise the significance of comprehensively studying all levels of decision-making in the energy sector, including sustainable policymaking and infrastructure planning [Radtke and Scherhauser, 2022]. Furthermore, they underscore the importance of recognising how energy-related decisions impact the daily lives and economic well-being of consumers [Ortiz et al., 2017; Steg et al., 2018]. The ability to conduct large-scale qualitative data analysis of the discourse surrounding the energy sector is important, given its multifaceted nature and influence on various stakeholders and societal structures.

Much of the previous NLP-powered energy transition research has focused on using unsupervised machine learning methods, specifically topic modeling, to uncover latent topics in the material [Dehler-Holland et al., 2021; Repo et al., 2021; Rizzoli et al., 2024; Saheb et al., 2022; Tie and Zhu, 2022]. Alternatively, dictionary-based NLP has been employed to analyse specific aspects of data, such as gender perspectives [Carroll et al., 2024], involvement of startups in renewable energy [Singh et al., 2021], and renewable energy investor sentiment [Herrera et al., 2022].

This study is part of a broader research project that involves extracting keywords and classifying a large dataset of the Finnish energy discourse. The dataset comprises two subsets: online consumer discussions, characterised by a more general vocabulary, and politicians' parliamentary speeches, which approach the subject from a macro-level perspective and contain a more specialised vocabulary. The purpose of using two different data sets in the research project is to gain a more comprehensive understanding of energy transition-related discussions, considering both the perspectives of everyday consumers and the rhetoric of policymakers. The objective of this article is to contribute a new methodological perspective to the field of energy transition research by evaluating the advantages and limitations of two NLP approaches to analysing energy transition discourse: a large language model (LLM) and a rule-based ontological classification. The NLP field is developing rapidly, and especially LLM tools are now heavily invested in and developed by several companies. For clarity, it is probably good to state that even though these two tools have been selected for the article, its ultimate purpose is not to focus on specific tools, but rather to reflect more generally on the characteristics of different approaches aimed at automating qualitative research work. I will now introduce these approaches in more detail.

One of the LLMs that has attracted the most interest among researchers and the public is the generative chatbot ChatGPT, developed by OpenAI. Since it was launched in November 2022, research articles using ChatGPT have been published at a rapid pace. Many studies have

compared the quality of ChatGPT's performance to human-generated work and found the two to be comparable. For example, 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 in answering 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]. [Ziems et al., 2023] did an extensive evaluation to measure the zero-shot performance of 13 language models on 25 representative English computational social science benchmarks and concluded that, except in a minority of cases, prompted LLMs did not match or exceed the performance of carefully fine-tuned classifiers, and the best LLM performance was often considered too low to fully replace human annotation. Some studies have found that ChatGPT's zero-shot performance is lacking, but prompt engineering and additional training have been shown to improve the results [Shi et al., 2023; Yuan et al., 2023].

Although 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 in the energy sector. In addition, Finnish-language data have not been used as research material.

In this study, I will compare ChatGPT's classification features with those of a traditional rule-based NLP tool called Etuma. The Etuma tool is based on dictionaries, grammar rules, and ontologies, performing basic NLP tasks such as lemmatization, as well as morphological, syntactic, semantic, and sentiment analysis [Etuma, 2023]. In this article, I will focus on the part of Etuma's analysis process that uses ontologies for classification. In the fields of information processing and artificial intelligence, ontology refers to the description and definition of existence in a form that can be understood by a computer [Gruber, 1995, p. 908]. Although ontologies may contain relatively generalisable information that allows them to be reused for different purposes [Spyns et al., 2002], they are often employed in NLP tasks that require specialised knowledge and vocabulary [Khadir et al., 2021]. Some studies have introduced ontologies developed for the energy sector; for example, [Booshehri et al., 2021], [Cuenca et al., 2020], and [Küçük and Arslan, 2014] for English, and [Glenc, 2022] for Polish. To the best of my knowledge, a dedicated energy ontology for the Finnish language has not yet been developed, although a general Finnish ontology¹ does exist.

This study addresses the following research questions:

- 1) How do the rule-based ontological classification tool (Etuma) and the LLM chatbot (ChatGPT) differ as qualitative content analysis methods?
 - a. What kind of classification do they produce without prior training specific to the context?
 - b. How do recall and precision differ between the two methods and data sets?
- 2) How well do these methods align with scientific principles such as reliability and scientific integrity?

¹ <https://finto.fi/yso/fi/>

In the following section, I describe the research materials, methods, and processes. I then present and discuss the results and their limitations. Finally, I conclude with insights and suggestions for future research.

II DATA

The corpus² used for this study was originally collected for a broader research project that focused on analysing the energy discourse from the perspectives of citizens and politicians. It consists of 110,295 social media comments from August 2022 to August 2023, and 25,872 parliamentary speeches from February 2022 to March 2023. The social media comments were collected using a web scraping tool called Mohawk Analytics [Legentic, 2023] with the search term “electric car” (“sähköauto” in Finnish) and the transcribed parliamentary speeches were downloaded from a database called Parliament Sampo³ [Hyvönen et al., 2022]. For this study, 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 one of the topics that have caused disagreements between citizens and politicians, and which will be qualitatively analysed in the project: electric car subsidies offered by the Finnish government. The subset corpus contained 44 social media comments and 29 parliamentary speeches.

The social media data included 21 tweets from Twitter (currently X), 19 online news comments, and four discussion forum posts. The parliamentary speech corpus consisted of 13 speeches from the Finns Party, three speeches from the Social Democratic Party, three speeches from the Centre Party, three speeches from the Green Party, and one speech each from the National Coalition Party and the Christian Democrats. In addition, the material included five 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 the keywords, topics, and text quotes were translated into English for this article. To clean the data, I removed mentions targeted to specific users (identified with the “@” character) in social media comments. I copied the original texts into an Excel file and recorded the analysis results obtained using the different methods in their respective columns. Additionally, I randomly selected a sample to validate the results. I describe the classification and validation processes in more detail in the following sections.

Social media comments were typically short, but their length varied between 20 and 155 words per comment. The comments were mostly critical towards the topic, as in statements like *"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 misspelt words. Parliamentary speeches were longer, their length varying between 72 and 662 words per speech. The speeches contained a more specialised technical and administrative vocabulary than the social media comments and were formal in style, 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.

² Available at [10.5281/zenodo.11208322](https://doi.org/10.5281/zenodo.11208322)

³ <https://a3s.fi/parliamentsampo/speeches/csv/index.html>, data licensed under CC BY 4.0

III METHODS

Due to the large amount of research material, I decided to use a combination of automatic text analysis and qualitative methods in the research project, following a similar approach as in [Grimmer and Stewart, 2013], [Guetterman et al., 2018] and [Jänicke et al., 2015]. For this study, I chose to employ two tools that both had Finnish language support, but were based on different analysis methodologies, namely Etuma and ChatGPT. These tools were used to extract keywords from the texts and classify them into topics. In this context, the term keyword refers to an expression in the corpus, and topic is a more general category into which keywords are classified. In the Etuma tool, the topics are predefined in the ontology, whereas in ChatGPT, they are based on the patterns learned from its training data. In this study, I used the paid versions of both tools. For ChatGPT, I had access to version 4 through a premium subscription on the Poe.com platform. For Etuma, I utilised a self-service researcher version that employs a basic ontology to which the user can make data or project-specific customisations. When comparing classification outputs, it is important to note that they are produced through very different data processing methods. For example, there are differences in considering the context: when creating a response, ChatGPT adapts to the context, taking into account the texts entered by the user and the model's own previous responses. In Etuma, such automatic context-adaptive learning does not occur.

For the comparison of the two methods, I adopted, with some modifications, the criteria established by [Hillard et al., 2008, p. 33] for an automated classification system that meets the requirements of social scientists. According to their framework, an ideal system for document classification and trend recognition should be 1) discriminatory: the topics should be mutually exclusive, ensuring clear differentiation between them; 2) accurate: the topic accurately should reflect the document's content; 3) reliable: classification should be consistent between documents; 4) probabilistic: it should classify secondary topics as well, i.e. have the capability to identify documents that may not primarily focus on the subject in question; and 5) efficient: it should achieve a high input-benefit ratio.

3.1 Research process

The initial phase of the study began in a setting where no training data or predefined 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.

Figure 1 shows the key phases of the research process. During the study, I conducted both distant reading and traditional close reading in parallel [Jänicke et al., 2015; Parks and Peters, 2023]. The concept of distant reading was introduced by [Moretti, 2013, p. 45], whose approach aimed to address the inherent challenge that humans face in effectively handling a large quantity of texts when limited to reading them on a sentence-by-sentence basis. A similar concept was also discussed by [Jockers, 2013] through the concepts of micro and macro analysis. The iterative process involves a distant reading phase of automated keyword extraction and topic classification and a close reading phase of examining the context in which the topics and keywords are discussed and validating the results, enabling a deeper understanding of the data. After these phases, the classification is refined to better align with the broader objectives of the research project, ensuring that it captures the relevant information.

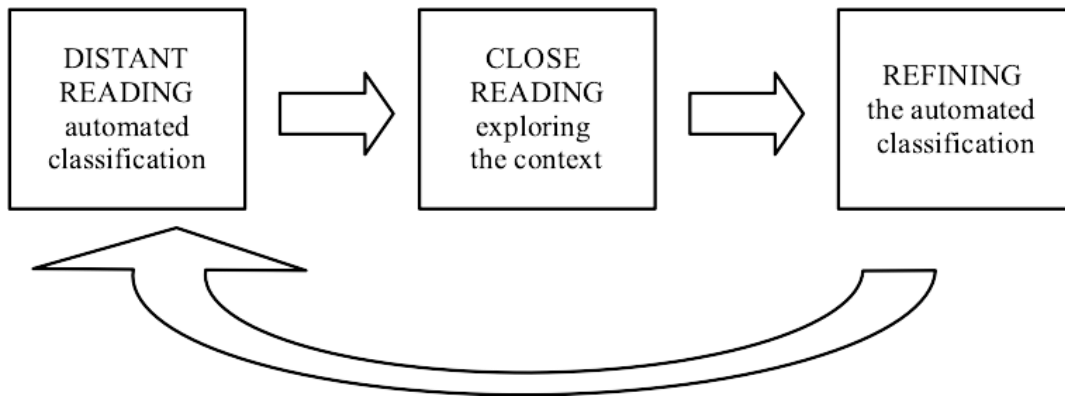


Figure 1. Research process.

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

In the following subsections, I describe in more detail the specific features of the research process for ChatGPT and Etuma separately.

3.2 ChatGPT

I selected the state-of-the-art large language model at the time, ChatGPT 4, for this study, as I wanted to assess a highly advanced language model to obtain a comprehensive understanding of the approach's capabilities. 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 and fine-tuned by humans using the Reinforcement Learning from Human Feedback (RLHF) method [OpenAI, 2023a].

I employed ChatGPT 4 through the Poe.com platform. My reason for purchasing a Poe.com subscription was that several other LLM tools, such as Claude, Llama, and Mistral AI, can also be used through the platform at the same price as a single tool premium subscription⁴, allowing me to cost-effectively extend my research to other tools as well. Users on the Poe.com platform can create their own chatbots and customise their settings according to their preferences. This includes configuring a default prompt, which serves as the initial message for the chatbot, and setting a temperature value. Increasing the temperature parameter allows the predictive model to take more risks, suggesting less likely alternatives, thereby reducing result consistency [OpenAI, 2023b]. The prompt plays a central role in determining the results generated by the ChatGPT-powered bot.

During the study, I used prompt engineering to minimise any potential impact that a poorly formatted prompt could have on the outcomes, 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. However, in this specific case, the prompt engineering techniques

⁴ <https://poe.com/about>

did not lead to improved results. A report detailing the example chat interactions of the prompt engineering experiment can be found in Annex 2. After 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.”.

During the distant reading phase (Figure 1), I input the texts individually into the same chat conversation and recorded the ChatGPT responses in an Excel table. I first analysed the social media material and then continued with an analysis of parliamentary speeches. The system autonomously identified 47 topics and 935 keywords in the data. Concurrently, I validated the classification by conducting a close reading of the original texts.

Furthermore, I explored how using more detailed prompts affected the results. This showed that with more precisely defined prompts, the model was able to extract additional relevant keywords and topics that had not been identified in the initial analysis. For example, with prompts like “extract relevant keywords and topics related to commuting” I was able to confirm some hypotheses about the themes present in the research material. However, in this article, I only report the results obtained with the initial prompt.

3.3 Etuma

Etuma’s technological foundation is rooted in NLP research conducted at the University of Helsinki [Lahtinen, 2000; Tapanainen, 1999], which has since been commercially continued by Etuma⁵. 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 extracted keywords referring to the same theme into more general categories called topics. For example, the keywords “electric car”, “e-car”, and “battery vehicle” are categorised into the same predefined topic called “Electric cars”. It is important to note that although Etuma refers to the classification categories 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 explicitly defined ontologies. I am not aware of any other tools⁶ based on the same approach that have customisable ontologies for the Finnish language.

In this study, I used Etuma’s Finnish ontology for analysis. Etuma's ontology has not been developed for one specific purpose; rather, it has been utilised over time to analyse diverse survey and online data, often catering to the research requirements of companies or public organisations. Through this iterative process, the ontology has gradually evolved. When employing this comprehensive ontology for different research objectives, some level of customisation is typically required. The extent of customisation depends on factors such as the nature of the data, analysis requirements, specificity of the research topic, and specialised vocabulary involved.

Using Etuma's tool, I followed the research process described in Figure 1. First, I uploaded 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

⁵ Etuma is my former employer and currently I act as Etuma’s advisor and board member.

⁶ Partially similar tools are Annif (<https://annif.org/>) and Finto AI (<https://ai.finto.fi/>).

within the sub-dataset. During the close reading phase, I reviewed the most frequently occurring topics and their corresponding keywords. I then reviewed the less frequent topics that seemed relevant to the research.

In the third phase of the process, the classification is refined to improve the relevance and precision by merging and splitting topics and transferring keywords between them. Etuma has a built-in user interface for these tasks because refining the classification is an integral part of the research process. The extent of this phase depends on the goals of the research, amount of material, and precision of the classification. After the classification-validation process is completed, new classification rules are updated to the 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 evaluating a situation in which no fine-tuning has been implemented.

IV EMPIRICAL ANALYSIS

In this section, I present the key findings of the comparison between an LLM and a rule-based approach to classification. Firstly, I describe the characteristics of keyword extraction and topic classification for both methods, along with relevant examples. Secondly, I present a comparison of the methods using a smaller sample, employing traditional metrics such as recall, precision, and F1 score. This analysis provided a quantitative evaluation of the features of each method. Additionally, Annex 1 contains a list of the most frequently occurring topics and keywords that were identified during the analysis.

4.1 Classification characteristics

Table 1 details the differences in the number of unique keywords and topics identified using each method. The ratio between methods was similar for both corpora (0.1 for topics and 0.6 for keywords), indicating that the text type had no significant effect on the results.

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

Table 1. Unique keywords and topics in corpora.

4.1.1 Keyword extraction

Both methods successfully analysed the Finnish language material without significant deficiencies or shortcomings. However, there were some differences in the keywords extracted 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 parliamentary speeches (e.g., MAL, KAISU) but did not classify them as exact topics. Initially, ChatGPT did not recognise 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).

Typographical errors and misspellings are common in social media materials. The Etuma tool provides a list of keywords that are not recognised, and among them, there were 31 unique keywords that were misspelt and thus left uncategorised. Based on my observations, ChatGPT analysed typographical errors correctly more frequently. ChatGPT also correctly classified more names of politicians (e.g., Li, Tynkkynen) than Etuma in the social media material. However, a detailed analysis of these features was not conducted in this study.

4.1.2 Topic classification

In terms of unique topics, the difference between Etuma and ChatGPT was even more notable, 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.

I also tested repeating identical prompts in new chat interactions, which revealed that the classification results for the same piece of text could change even though the content and prompt remained identical. In some cases, this was probably related to a feature of ChatGPT, where it considers both the tokens entered by the user and the tokens generated within a specific context window when formulating its response. The context window for ChatGPT 4 was 4,096 tokens during the analysis conducted in September 2023 and has been expanded to 8,192 tokens as of April 2024 [OpenAI, 2024]. 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 attempted to simplify the classification by grouping less precise keywords into a smaller set of topics, indicating an internal learning mechanism guiding the classification. As demonstrated in this example, this approach may result in lower precision.

On a few occasions, ChatGPT generated information that was not accurate or factual, a feature known as hallucination. For example, it stated that "Sulo Vileen" (Sulo Vilén, a character from a Finnish TV comedy series) is a colloquial expression used to refer to ordinary Finns, similar to the term "Joe Public" in English. Such misinformation may also point to deficiencies in the Finnish training data.

4.2 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 relevant keywords. 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, and no domain-specific teaching data or ontology were used for their configuration. As a result, I tagged a total of 151 keywords from the social media sample and 311 keywords from the parliamentary speech sample.

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

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

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

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

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

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

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

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

Table 2. Recall, precision, and F1 score.

4.2.1 Recall

The recall level of Etuma's classification was higher in the social media sample (0.85) than that 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.

In the research setting, I first analysed the shorter social media comments and then the longer parliamentary speeches. The order of the input within the context window may have affected the recall. Another explanation for the result comes from the fact that ChatGPT has been trained

with web-crawled material. For Etuma, a possible explanation for the difference is that the tool has been optimised for the analysis of relatively short customer feedback and survey responses, not for the analysis of longer texts [Etuma, 2023]. Additionally, social media posts used more common language terms, while parliamentary speeches had more specialised terms that the tools did not always identify as keywords. However, the scope of this study did not include determining the impact of these factors on the results.

4.2.2 Precision

Etuma's precision rate was 0.70 for both parliamentary speech texts and social media posts. For a single text, the precision ranged from 0.28 to 1.00. However, different factors affect the precision rate in the two samples. There were more misspelled words in social media posts, while in parliamentary speeches, there was a more specialised vocabulary that Etuma had recognised but left unclassified, referring to deficiencies in the general ontology.

The precision rate of ChatGPT was higher (0.96) for both samples, ranging from 0.60 to 1.00 for a single text. The results indicate that the precision of the results obtained is not significantly influenced by the type of text being analysed. Errors were typically related to the interpretation of the correct topic rather than keyword extraction. For example, from the sentence that embodies typical rhetoric 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.

4.2.3 F1 score

The F1 score, which considers both recall and precision, was slightly higher for Etuma in both the social media (0.77 compared to 0.75 for ChatGPT) and parliamentary speech (0.75 compared to 0.72 for ChatGPT) samples.

V DISCUSSION

In this section, I revisit the research questions presented in the Introduction. I examine the differences between the rule-based ontological classification tool and the LLM chatbot as qualitative content analysis methods and evaluate their characteristics in terms of recall, precision, and efficiency (Research Question 1). In addition, I discuss how well these methods align with scientific principles, with a particular focus on repeatability, transparency, and research integrity (Research Question 2). Table 3 summarises the results of the comparison according to the criteria of [Hillard et al., 2008], with slight modifications.

	ChatGPT 4	Etuma
Are the topics discriminatory?	No (at least in this setting)	Yes
Accuracy of the results	Higher precision, automatically adapts to context	Requires manual refining to improve precision and adapt to context

Reliability of the results	LLM approach has challenges with hallucination, repeatability, and transparency	Manually set rules are consistent and transparent, but classification may vary between researchers
Are there multiple topics identified per document?	Yes	Yes
Is it efficient to implement?	Yes (but LLM training requires data and computing power)	Yes (but ontology refining requires manual work)

Table 3. Summary of the method comparison.

5.1 Comparison of classification characteristics

5.1.1 Recall

In this study, Etuma's recall was higher than that of ChatGPT. The results revealed a fundamental difference between these approaches. ChatGPT focused on the main points and tended to overlook rhetorical expressions and topics that were mentioned less frequently and indirectly. A lack of detail was also observed in a previous study comparing ChatGPT responses to those of human experts [Guo et al., 2023]. In situations where the corpus contains a significant amount of noise or irrelevant data, the ability of ChatGPT to emphasise essential information can be beneficial. However, there are scenarios in which researchers specifically look for details and rhetorical language, which may not align with ChatGPT's primary focus unless prompted specifically. In addition, ChatGPT tended to use overlapping and inconsistent topic names. Without predefined topic names, the classification did not meet the discrimination criterion presented by [Hillard et al., 2008].

5.1.2 Precision

Precision measured in this study corresponded to the accuracy criterion proposed by [Hillard et al., 2008]. As anticipated from prior research [Ortega-Martín et al., 2023], ChatGPT performed well in semantic disambiguation and integrated cultural contexts into its classification. The adaptability of information related to cultural context stands out as a key 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. Although both methods may fail to classify certain terms, abbreviations, and misspelled words based on the vocabularies and training data used, this study revealed that ChatGPT outperformed Etuma in these regards.

In this study, there were no noticeable deficiencies in 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 because of the limited training materials available.

5.1.3 Efficiency

[Hillard et al., 2008] listed efficiency as one of the criteria for a useful classification. The F1 score can be seen as an indicator of the efficiency of classification. Often finding the optimal result requires balancing between recall and precision. In this case, because of the differences between the approaches, I do not consider the F1 scores to be the most important criterion for choosing between the two methods. Depending on the goals of the research as well as the time, funds and expertise available, it may make sense to prioritise recall, precision, or some

completely different aspects. For example, commercial solutions require funding to license them. In addition, large language models (LLMs) require significant computing power and the use of open-source alternatives requires GPUs capable of running them.

Integral to the research process outlined in this study is an iterative cycle of validating and fine-tuning the results. The workload involved in this step depends on the recall and precision of the initial analysis performed using the automated method. If they are low, the researcher faces a substantial workload in validating and fine-tuning classification, which may not be feasible in all cases. The Etuma tool has built-in tools designed to improve recall and precision as it is part of the standard process of the method. The constrained context window of ChatGPT limits the ability to fine-tune analyses using prompts and maintain consistent classification. However, employing ChatGPT differently may make fine-tuning its classification feasible, which could be a worthwhile topic for further research.

5.2 Compliance with scientific principles

5.2.1 Reliability

In this study, classification reliability criterion by [Hillard et al., 2008] was particularly related to repeatability and transparency. With the Etuma tool, the outcome of the analysis remains consistent unless the researcher alters the classification rules. By contrast, a characteristic of ChatGPT is that an identical input can yield different outputs. 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 the findings of 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, such as independently annotating the same material several times and then comparing the results.

In this research setting, ChatGPT lacked a structured classification system, which hindered its transparency. This made it difficult to clearly understand how the model worked and how it made decisions. An LLM operates as more of a black box, while a rule-based approach offers greater transparency because its classification is based on predefined dictionaries. 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, the amount of manual work in this approach often requires researchers to narrow their focus to, for example, a smaller subset of the corpus or the most prevalent topics.

5.2.2 Research integrity

Amid the surrounding technological hype, researchers have a responsibility to ensure that new technologies and tools are not adopted uncritically for scientific use. For example, biases and distortions in the training data and processes of these tools need to be discussed. While the current analysis may not show evident bias, they can still emerge in other types of content. Additionally, ChatGPT's tendency to produce hallucinations and its vulnerability to adversarial attacks underscore the need for cautious evaluation of the data it generates. Manually validating the analysis of a vast data set can be challenging, potentially allowing misinformation or biases to go unnoticed. The utilisation of these tools can also be constrained by the fact that certain methods, such as public language models, may not be suitable for analysing sensitive data.

Alongside the technological risks, it is important to consider the implications of these tools on the work of researchers. The findings of this study indicate that especially LLMs assume a significant portion of decision-making on behalf of researchers, leaving much of the research material unclassified. 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 results. For example, an attempt to summarise complex information into broad topics may inadvertently overlook nuances or lead to incorrect interpretations.

Furthermore, it can be a challenge to study new emerging phenomena or to analyse data from new perspectives with tools whose training data or defined rules and ontologies correspond to existing structures and prevailing concepts. 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.

5.3 Strengths and limitations

This study has several strengths. It addresses an existing knowledge gap by exploring the application of an LLM chatbot as a qualitative analysis tool for the study of energy discourse in Finnish using two distinct corpora and comparing the results with a rule-based method. This study offers insights for researchers who are considering employing either a large language model or a rule-based NLP approach for their analysis.

A limitation of this study is that the empirical material is relatively narrow and focuses on a 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. Another limitation is that this study focused only on a setting without special training data or ontologies, which would likely have produced different analysis results. In addition, in future studies, it would be worth investigating how much the results obtained with ChatGPT are influenced by the context window within which the analysis is performed. A comparison of this research setting with a setting in which context is not provided in the same chat interaction would provide more information about the impact of context on classification.

VI CONCLUSIONS

In this research setting without prior context specific training, ChatGPT had higher precision and could adapt to the context, which can be advantageous in certain research tasks. However, the issues of hallucination, limited repeatability, and lack of transparency pose significant challenges to the reliability of the results generated by ChatGPT. Researchers need to be cautious when using ChatGPT and thoroughly validate its outputs, as errors or biases can easily go unnoticed. Furthermore, a low recall rate can excessively restrict the researcher's autonomy in decision-making. In this study, my aim was to conduct a comprehensive classification that allows for qualitative analysis of the material from various perspectives. Therefore, I prefer not to rely on the tool to make decisions about what is important or interesting in the text on my behalf.

In contrast, a rule-based method, while requiring more manual effort in refining the precision and adapting to context, offers a more transparent and consistent approach. The reliance on manually set rules provides a clear understanding of the decision-making process, which can be beneficial for research transparency and reproducibility. However, the method's repeatability

may be affected by human factors, such as the expertise of the researchers involved in the refinement process.

6.1 Implications for future research

Repeatability and transparency are important features in scientific research, and the classification of qualitative content should be consistent and accurate. While ChatGPT may not yet surpass rule-based NLP methods in all these respects, it has undeniable strengths, such as accurate semantic analysis and information of cultural contexts.

In future research, it could be beneficial to employ the methods in parallel and leverage the strengths of both. This approach could provide a more comprehensive and reliable analysis. During this study, I developed some preliminary ideas for integrating the methods, as depicted by the dashed line in Figure 2.

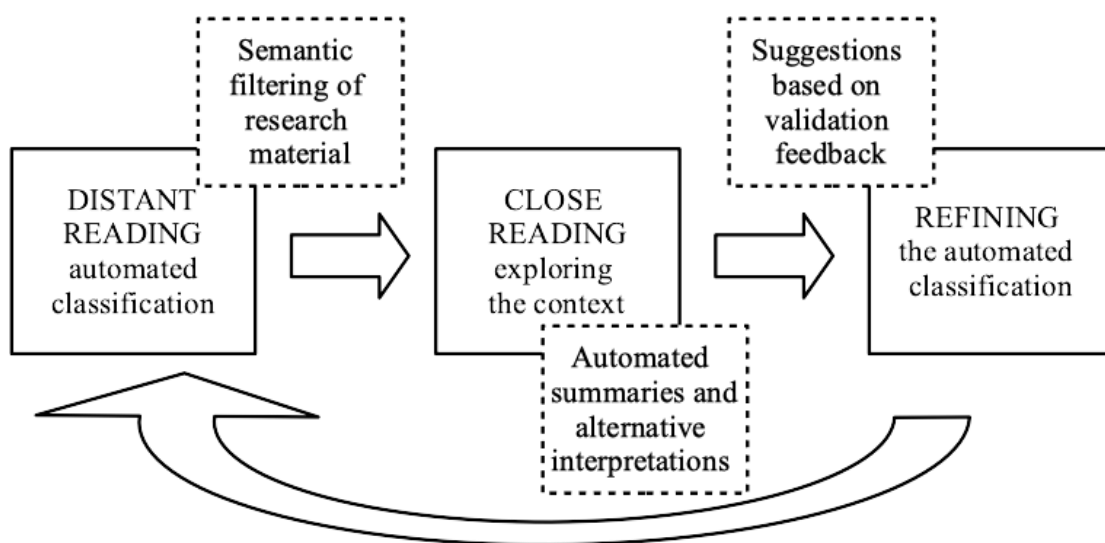


Figure 2. Research process combining rule-based and LLM approaches.

Firstly, using the LLM’s capacity for semantic interpretations could enhance the semantic classification of another NLP method in the distant reading phase, assisting in semantic filtering of research material based on the studied phenomenon.

Secondly, in the close reading phase, LLM could aid researchers by generating automated summaries or in interpreting ambiguous or complex texts, suggesting alternative meanings and contexts to researchers in the validation process – a similar application was also discussed by [Ziems et al., 2023]. The knowledge within the LLM based on the vast training data could extend beyond the corpus, aiding in interpreting the discourse in a broader social context.

Finally, in the classification refinement phase, LLM's ability to identify semantic meanings and its creative capabilities could be used to formulate new topics or ontologies based on feedback from the validation process. This approach aligns with the partially automated ontology learning process, which has been discussed for example by [Khadir et al., 2021]. An ontology that combines concepts and terms from diverse sources like social media and parliamentary speeches, could serve as a versatile analytical framework and improve the effectiveness of the classification.

Acknowledgements

I would like to thank the anonymous reviewers for their valuable feedback, as well as the Finnish Cultural Foundation, Kone Foundation, and Niilo Helander Foundation for supporting this research.

References

- Andreotta, M., Nugroho, R., Hurlstone, M. J., Boschetti, F., Farrell, S., Walker, I., and Paris, C. (2019). Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, 51(4), 1766–1781. <https://doi.org/10.3758/s13428-019-01202-8>
- Bender, E. M., Gebru, T., McMillan-Major, A., and 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>
- Booshehri, M., Emele, L., Flügel, S., Förster, H., Frey, J., Frey, U., Glauer, M., Hastings, J., Hofmann, C., Hoyer-Klick, C., Hülk, L., Kleinau, A., Knosala, K., Kotzur, L., Kuckertz, P., Mossakowski, T., Muschner, C., Neuhaus, F., Pehl, M., ... Stappel, M. (2021). Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis. *Energy and AI*, 5, 100074. <https://doi.org/10.1016/j.egyai.2021.100074>
- Buschmann, P., and Oels, A. (2019). The overlooked role of discourse in breaking carbon lock-in: The case of the German energy transition. *WIREs Climate Change*, 10(3), e574. <https://doi.org/10.1002/wcc.574>
- Carroll, P., Singh, B., and Mangina, E. (2024). Uncovering gender dimensions in energy policy using Natural Language Processing. *Renewable and Sustainable Energy Reviews*, 193, 114281. <https://doi.org/10.1016/j.rser.2024.114281>
- Cuenca, J., Larrinaga, F., and Curry, E. (2020). DABGEO: A reusable and usable global energy ontology for the energy domain. *Journal of Web Semantics*, 61–62, 100550. <https://doi.org/10.1016/j.websem.2020.100550>
- Dehler-Holland, J., Schumacher, K., and Fichtner, W. (2021). Topic Modeling Uncovers Shifts in Media Framing of the German Renewable Energy Act. *Patterns*, 2(1), 100169. <https://doi.org/10.1016/j.patter.2020.100169>
- Etuma (2023). Etuma Natural Language Processing. Internal document. Accessed on 26.9.2023.
- Gilardi, F., Alizadeh, M., and Kubli, M. (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. <https://doi.org/10.48550/ARXIV.2303.15056>
- Glenc, P. (2022). Poland's Energy Transition: Towards an Ontology. *Przegląd Organizacji*, 24–31. <https://doi.org/10.33141/po.2022.03.03>
- Grimmer, J., and 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>
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5–6), 907–928. <https://doi.org/10.1006/ijhc.1995.1081>
- Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., and Vydiswaran, V. V. (2018). Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *Journal of Medical Internet Research*, 20(6), e231. <https://doi.org/10.2196/jmir.9702>
- Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection*. <https://doi.org/10.48550/ARXIV.2301.07597>
- Halford, S., and Savage, M. (2017). Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research. *Sociology*, 51(6), 1132–1148. <https://doi.org/10.1177/0038038517698639>
- Herrera, G. P., Constantino, M., Su, J.-J., and Naranpanawa, A. (2022). Renewable energy stocks forecast using Twitter investor sentiment and deep learning. *Energy Economics*, 114, 106285. <https://doi.org/10.1016/j.eneco.2022.106285>
- Hillard, D., Purpura, S., and Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology and Politics*, 4(4), 31–46. <https://doi.org/10.1080/19331680801975367>
- Huang, F., Kwak, H., and 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., Leskinen, P., Sinikallio, L., 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, CEUR Workshop Proceedings*, 3133, May 2022, 69–85.
- Jänicke, S., Franzini, G., Cheema, M. F., and 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>
- Khadir, A. C., Aliane, H., and Guessoum, A. (2021). Ontology learning: Grand tour and challenges. *Computer Science Review*, 39, 100339. <https://doi.org/10.1016/j.cosrev.2020.100339>
- Küçük, D., and Arslan, Y. (2014). Semi-automatic construction of a domain ontology for wind energy using Wikipedia articles. *Renewable Energy*, 62, 484–489. <https://doi.org/10.1016/j.renene.2013.08.002>
- 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 (2024). Models. Accessed on 2.5.2024 at: <https://platform.openai.com/docs/models/overview>

- 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., and Alonso, A. (2023). *Linguistic ambiguity analysis in ChatGPT*. <https://doi.org/10.48550/ARXIV.2302.06426>
- Ortiz, M. A., Kurvers, S. R., and Bluysen, P. M. (2017). A review of comfort, health, and energy use: Understanding daily energy use and wellbeing for the development of a new approach to study comfort. *Energy and Buildings*, 152, 323–335. <https://doi.org/10.1016/j.enbuild.2017.07.060>
- Parks, L., and Peters, W. (2023). Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach. *International Journal of Social Research Methodology*, 26(4), 377–389. <https://doi.org/10.1080/13645579.2021.2018905>
- Radtke, J., and Scherhafer, P. (2022). A social science perspective on conflicts in the energy transition: An introduction to the special issue. *Utilities Policy*, 78, 101396. <https://doi.org/10.1016/j.jup.2022.101396>
- 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>
- Repo, P., Matschoss, K., and Mykkänen, J. (2021). Examining outlooks on sustainability transitions through computational language analysis. *Environmental Innovation and Societal Transitions*, 41, 74–76. <https://doi.org/10.1016/j.eist.2021.10.028>
- Rizzoli, V., Biddau, F., and Sarrica, M. (2024). The identity-attitude nexus in the representation of energy transition in a coal region (Sulcis, Italy): An exploration through the Structural Topic Model. *European Journal of Social Psychology*, 54(1), 118–135. <https://doi.org/10.1002/ejsp.3004>
- Saheb, T., Dehghani, M., and Saheb, T. (2022). Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. *Sustainable Computing: Informatics and Systems*, 35, 100699. <https://doi.org/10.1016/j.suscom.2022.100699>
- Shi, Y., Ma, H., Zhong, W., Tan, Q., Mai, G., Li, X., Liu, T., and Huang, J. (2023). *ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs*. <https://doi.org/10.48550/ARXIV.2305.03513>
- Singh, M., Jiao, J., Klobasa, M., and Frietsch, R. (2021). Making Energy-transition headway: A Data driven assessment of German energy startups. *Sustainable Energy Technologies and Assessments*, 47, 101322. <https://doi.org/10.1016/j.seta.2021.101322>
- Spyns, P., Meersman, R., and Jarrar, M. (2002). Data modelling versus ontology engineering. *ACM SIGMOD Record*, 31(4), 12–17. <https://doi.org/10.1145/637411.637413>
- Steg, L., Shwom, R., and Dietz, T. (2018). What Drives Energy Consumers?: Engaging People in a Sustainable Energy Transition. *IEEE Power and Energy Magazine*, 16(1), 20–28. <https://doi.org/10.1109/MPE.2017.2762379>
- Tapanainen, P. (1999). *Parsing in two frameworks: Finite-state and functional dependency grammar*. [Doctoral dissertation]. University of Helsinki.
- Tie, M., and Zhu, M. (2022). Interpreting low-carbon transition at the subnational level: Evidence from China using a Natural Language Processing approach. *Resources, Conservation and Recycling*, 187, 106636. <https://doi.org/10.1016/j.resconrec.2022.106636>
- Van Atteveldt, W., Welbers, K., and Van Der Velden, M. (2019). Studying Political Decision Making With Automatic Text Analysis. In W. Van Atteveldt, K. Welbers, and M. Van Der Velden, *Oxford Research Encyclopedia of Politics*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190228637.013.957>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and 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., and Ananiadou, S. (2023). *Zero-shot Temporal Relation Extraction with ChatGPT*. <https://doi.org/10.48550/ARXIV.2304.05454>
- Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). *Can Large Language Models Transform Computational Social Science?* <https://doi.org/10.48550/ARXIV.2305.03514>

ANNEX 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”

ANNEX 2

Examples of prompt engineering attempts with ChatGPT 4 in September 2023.

Chat interaction 1

Researcher	<i>"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."</i>
ChatGPT 4	<i>"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."</i>
Researcher	<i>"Here is the text to be classified: [a social media comment from the corpus]"</i>

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	<i>"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."</i>
ChatGPT 4	<i>"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.""</i>
Researcher	<i>"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."</i>
ChatGPT	<i>"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."</i>

Result: No keywords or topics.

Chat interaction 3

Researcher	<i>"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."</i>
ChatGPT 4	<i>"Sure, I can certainly help with that. However, I need some example texts to work with. Could you please provide some?"</i>
Researcher	<i>"Here is the text: [a social media comment from the corpus]"</i>

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