Journal of Data Mining and Digital Humanities

Out of Context!
Managing the Limitations of Context Windows in ChatGPT-4o Text Analyses

Erkki Mervaala[*1,2] and **Ilona Kousa**[2]

1 Finnish Environment Institute Syke, Finland

2 University of Helsinki, Finland

*Corresponding author: erkki.mervaala@syke.fi

Abstract
In recent years, large language model (LLM) applications have surged in popularity, and academia has followed suit. Researchers frequently seek to automate text annotation – often a tedious task – and, to some extent, text analysis. Notably, popular LLMs such as ChatGPT have been studied as both research assistants and analysis tools, revealing several concerns regarding transparency and the nature of AI-generated content. This study assesses ChatGPT's usabil- ity and reliability for text analysis – specifically keyword extraction and topic classification – within an "out-of-the-box" zero-shot or few-shot context, emphasizing how the size of the context window and varied text types influence the resulting analyses. Our findings indicate that text type and the order in which texts are presented both significantly affect ChatGPT's analysis. At the same time, context-building tends to be less problematic when analysing similar texts. However, lengthy texts and documents pose serious challenges: once the context window is exceeded, "hallucinated" results often emerge. While some of these issues stem from the core functioning of LLMs, some can be mitigated through transparent research planning.

keywords
text analysis; large language models; ChatGPT; green transition; parliamentary speeches; Finnish

I INTRODUCTION

The potential for utilising large language models for data analysis has generated widespread interest among researchers across various fields (Azaria et al., 2024; Hadi et al., 2023), and one of the most extensively studied is the field's market-leader OpenAI's ChatGPT (Ray, 2023; Westfall, 2023). As a freely accessible tool with a simple user interface, it has brought a new repertoire of analytical methods within reach of researchers who may have previously faced technological or resource barriers to using computational social science methods.

Several researchers have examined ChatGPT's capabilities as a data analysis tool and reported their findings, striving to approach the method critically from the per- spective of scientific principles and share best practices with other researchers (Bilal et al., 2024; Törnberg, 2023). Despite the hype surrounding the tool, ChatGPT's, and large language models in general, have been criticised and found to be unreliable in various research tasks including text analysis and annotation (Ollion et al., 2023). Our paper is an additional contribution to this literature from the point of view of auto- mated text analysis, but also an addition to the yet very limited research focusing on how such services function with smaller languages (Mets et al., 2024).

The purpose of the study is to gain an understanding of the influence of context on classification and to shed light on what kind of biases and limitations must be considered in research employing tools based on Large Language Models (LLMs), especially when using several different types of data. More specifically, the research aims to evaluate whether the keyword extraction and topic classification produced by chat-based LLM services, such as ChatGPT, vary based on different content types, the mixture of contents, the order of mixed contents, and context building within the same conversation window. This study aims not to replace existing Natural Language Processing (NLP) tools that are often used for simple automated text analysis tasks but to explore whether LLMs can support or even replace human efforts in research phases that require contextual interpretation and precision.

The specific case under examination will focus on the discussion regarding the green transition, which in political rhetoric generally refers to a shift towards an ecologically sustainable and low-carbon economy (Filipović et al., 2022; Ministry of Environment Finland, 2024). Green transition is an interesting term in the political debate as it evokes very polarising reactions from highly supportive and positive to strongly negative and dismissive. The reactions also vary a lot by country: for example, in Norway, the term "green transition" has very positive connotations and was even voted as the word of the year in 2015 (Olerud et al., 2016).

In Finland, the government in power since 2023 led by the National Coalition Party and its support party The Finns has made efforts to replace the term with other expressions such as ”clean transition" ("puhdas siirtymä") or "the blue-white transition" ("sinivalkoinen siirtymä") (Tavio, 2023; Valtioneuvosto, 2023).

To study the debate from the perspectives of both politicians and the public, we analysed communication on Twitter and within the Finnish parliament during the Conference of Parties climate change conference (COP27) held in Egypt in late 2022. We chose the highly publicised international event as green transition was one of its core focus points (European Commission, 2022) and because it stirred conversation both online and in the parliament so close to the Finnish parliamentary elections of 2023. The data comprises parliamentary speeches and Twitter comments collected from October 19, 2022, to December 1, 2022. Each dataset contains 20 texts, resulting in a total of 20 tweets (T) and 20 parliamentary speeches (P).

The study expects to determine if and how the order and context in which texts are presented to the LLM influence the outcomes of text analysis, specifically in terms of keyword extraction and topic classification. Our hypotheses address the usability of LLMs for our specific text analysis task from four different angles:

- Hypothesis 1: It is possible to gain meaningful results using LLMs in analysing and identifying keywords and topics in textual data in a zero-shot or a few-shot setting, focusing on a specific political issue such as "green transition".
- Hypothesis 2: The initial content type will influence the keyword extraction and topic classification when texts are analysed in a mixed order.
- Hypothesis 3: The context-building affects the results when analysing texts in a few-shot setting within the same chat window compared to zero-shot analyses in separate chat windows.
- Hypothesis 4: The limits of the context window when dealing with longer text documents will have a deteriorating effect on the LLM analysis.

ISSN 2416-5999, an open-access journal

1.1 Review of previous research

Traditionally, automated text analysis tasks, including keyword and topic mapping have been conducted using NLP methods such as topic modelling (Jelodar et al. 2019) or, more recently, language representation models like Bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018). LLMs are predicted to represent the next state of the art, especially in the zero-shot classification of unannotated text due to their document and web search capabilities (Fechner & Dörpinghaus 2024).

However, the inconsistency inherent in generative language models has been noted across various research settings in fileds such as medicine (Lechien et al., 2024), mathematics (Heya et al., 2024) and coding (Clark et al., 2024). One potential source of inconsistency is that ChatGPT's model considers the order of input when generating its responses (Bansal et al., 2024), which has been empirically demonstrated, for example, by varying the word order of the input (Zhao et al., 2024). The effects of order-dependency on consistency of the output have been studied further for example by Jang and Lukasiewicz (2023), who showed that ChatGPT and GPT-4 models often failed in both semantic and symmetric consistency, meaning that they produced different results from semantically similar inputs, and that the sentence order of the input affected the predictions made by the models.

For text analysis and annotation specifically, it has been shown that zero-shot or few-shot approaches to utilising LLMs often fail to reach as accurate results as fine- tuned, human-annotated models (Ollion et al., 2023). Previous studies have also compared the ability of large versus smaller language models to classify texts without fine-tuning and found that although large models generally perform better, their performance deteriorates as the amount and complexity of input data increases (Fechner & Dörpinghaus 2024). Additionally, the inherent limitation of a restricted context window has been highlighted as a significant barrier to its seamless integration into lengthy or complex tasks. (Briganti, 2023).

Other serious issues raised in the literature include the unreliability, potentially "hallucinated" results, copyright issues, and stochastic generation of misinformation and false claims (Alkaissi and McFarlane, 2023; Guerreiro et al., 2022; Guerreiro et al., 2023; Khatun and Brown, 2023). To address these challenges, several guides and reviews on how to use ChatGPT for academic work have been published recently (Bilal et al., 2024; Törnberg, 2023). They often emphasise the need for becoming acquainted with the LLM before its usage and to pay attention to verifying conversation settings and other variables that may affect the results.

II METHODS AND MATERIALS

The parliamentary speeches were obtained via the Finnish parliamentary speech archive Parliament Sampo[1] (Hyvönen et al., 2022) that allows collecting speeches from a selected time period in .csv format and then filtering the dataset via declension of the key phrase "vihreä siirtymä" ("green transition"). The filtering left us with 20 parliamentary speeches that fit the criteria.

The tweets for the study had been collected via the now-defunct service Mohawk Analysis. As there were several thousands of tweets containing the key phrase, the tweets were chosen first by matching the dates of the parliamentary speeches. The number of tweets for the studied time period was much larger than parliamentary speeches, so a corresponding tweet was picked via a randomised process for each speech. For example, when there were five speeches on October 27th ,2022, a total of five tweets were selected from the Twitter

[1] https://a3s.fi/parliamentsampo/speeches/csv/index.html, data licensed under CC BY 4.0

dataset. The only limiting factor for a tweet was that it should be a standalone tweet and not a response to another user. Links to both parliamentary speeches and the tweets in question are found in Appendix A. For the purposes of replicability, each analysis was done via the native commercial and publicly freely available ChatGPT service's version ChatGPT-4o released in May 2024. Its context window is 128,000 tokens which roughly translates to approximately 50 pages of text, hence was enough for all our test setups (OpenAI, 2024b). All speeches and tweets combined add up to 20 pages. The initial setup settings for the model were attained in June via prompting (see Appendix B).

All tests except 1.3 were conducted between July 8 and August 4, 2024, by one author, and initial tests of the analysis environment were performed between June 10 and July 21, 2024. Originally, the environment initialisation prompt was performed to control the changes in the ChatGPT version (see Appendix B) but access to it by users was later removed. The other author completed test 1.3 between August 19 and 26, 2024.

Each analysis began with the same first prompt after which the concurrent prompts included only the text to be analysed. No feedback was given to ChatGPT during the tests. If after prompting ChatGPT offered options to choose our preferred answer, this would be ignored, and the next text was added to the chat instead. The first prompt including the instructions for text analysis is as follows:

You will be analyzing a series of texts. For each text, perform the following analysis: 1. Extract a list of keywords from the text. 2. For each extracted keyword, provide a corresponding topic category.

2.1 Review of previous research

This experimental scenario evaluated how text analysis differs when all content is analysed in the same chat window in a few-shot setting versus each item being analysed in separate chat windows in a zero-shot setting.

In the context of ChatGPT, a zero-shot setting refers to the language model's ability to perform a task or respond to a query without having been explicitly trained on examples of that specific task or scenario. Instead, the model relies on its general understanding of language and knowledge to generate an appropriate response based on the prompt alone. (Yuan et al., 2023)

In test 1.1 all parliamentary speeches were analysed in the same chat window, and in test 1.2 all tweets were analysed in the same chat window, allowing for the window-specific context to build throughout both analysis runs covering each set of 20 texts. Tests 1.3 and 1.4 included the texts of both datasets analysed in separate chat windows: test 1.3 included 20 prompts in 20 chat windows, one for each parliamentary speech, and test 1.4 had a similar setup for tweets.

The rationale behind the testing addresses the issue with a zero-shot setting where the LLM is given the task of text analysis without context from earlier text analysis tasks. In the one-window task, all the texts were submitted for analysis consecutively which builds context cumulatively, thus potentially enhancing the analysis the further the conversation goes. The hypothesis is that in analyses performed within the same chat window the first and the last text are, essentially, analysed by a different LLM as the context has been allowed to build. To test this, the texts were presented in the chat first in the original, chronological order, and then in a backwards order. It is due to this feature that leads to the conclusion that the only strictly zero-shot analyses to be made via ChatGPT are to be done individually in separate chat windows. The individual analyses were then compared to the analyses done within the same chat window.

It should also be pointed out that the cross-chat "Memory" feature for Plus tier ChatGPT subscribers was not available in the EU area during the testing and so had no impact on it. The feature that allows ChatGPT to "remember details between chats, allowing it to provide more relevant responses" was made available in the EU in September 2024 (Coombes 2024, OpenAI 2024c).

2.2 Test batch 2: mixed dataset

The second experiment evaluated how context building within the same chat window affects text analysis results when all 40 texts are analysed in a mixed order, and whether the initial content type influences the outcome.

Test 2.1 included all tweets and parliamentary speeches alternating, starting with a tweet. Test 2.2 was almost identical but this time the first input was a parliamentary speech. Test 2.3 began with the whole tweet dataset, which was then followed by the whole speech dataset, and test 2.4 began with the speeches and ended with the tweets. The rationale behind the order-setting pairs stems from the context-building nature of the LLM chat windows. In the tests, the context is first built with consecutive texts of the same type and then the type changes to a very different one – both in tone and in length.

III ANALYSIS AND EVALUATION OF THE RESULTS

In this section, we first describe the quantitative results of keyword extraction and topic classification and then evaluate the consistency and relevance of classification.

3.1 Keywords and topics

Overall, the amounts of keywords and topics identified by ChatGPT varied. In the first batch of tests with separated datasets (see Table 1), both analyses of parlia- mentary speeches (PS) provided more keywords and topics than tweets (T), which is understandable due to the speeches being much longer.

Test #	Keywords	Topics
1.1 (PS, few-shot)	439	414
1.2 (T, few-shot)	127	125
1.3 (PS, zero-shot)	376	348
1.4 (T, zero-shot)	135	121

Table 1. Keywords and topics in test batch 1.

Test #	Keywords	Topics
2.1 (T1,PS1…PS20)	619	600
2.2 (PS1,T1…T20)	449	310
2.3 (T1–T20,PS1–PS20)	708	643
2.4 (PS1–PS20,T1–T20)	536	338

Table 2. Keywords and topics in test batch 2.

In our analysis, less keywords and topics were identified in the zero-shot analysis (zs) than the few-shot analysis (fs), but the zero-shot tweet analysis found more keywords but less topics than the few-shot analysis.

Test batch 2 with mixed datasets (Table 2) expectedly produced more keywords and topics, as each analysis had double the number of texts. The largest amounts of keywords and topics were found in test 2.3 that analysed first all 20 tweets and then 20 speeches (T1–T20,

ISSN 2416-5999, an open-access journal

PS1–PS20), followed by test 2.1 which alternated between content type but started with a tweet (T1, PS1…PS20).

Despite the same content being analysed within the same context window, the order of the texts did impact the amounts of keywords and topics found. In these specific cases, tests starting with a tweet produced over a hundred more keywords and topics identified by ChatGPT.

3.2 Classification of "green transition"

In the consistency and relevance evaluation, we compared the classification results of tests 1.1 to 1.4 (Table 3) and 2.1 to 2.4 (Table 4). We studied how consistently the keyword ”green transition” was classified in Twitter and Parliamentary speech datasets, in a few-shot versus a zero-shot setting and in mixed datasets. There was some variation in whether ”green transition” was recognized as a keyword at all or the keyword appeared in a slightly different form (e.g. ”left-green transition”); these occurrences have been reported on separate rows in Tables 3 and 4.

For both tweets and parliamentary speeches, classification was more consistent when all texts were analysed in a few-shot setting in the same chat window: all tweets were classified into the category Environmental Policy and parliamentary speeches were classified into two different categories: Environmental Policy and Green Transition. In the zero-shot setting, there were 12 different category names for the keyword "green transition" for tweets and 11 different category names for parliamen- tary speeches. Another observation was that few-shot analyses did not contain any multi-level categories (such as "Environmental Policy / Sustainability"), whereas in zero-shot setting, multi-level categories appeared in 12 cases for tweets and 11 cases for parliamentary speeches.

The results indicate that the topic categories were quite sensitive to variation in the form of the keyword. For example, in the single-window parliamentary speeches test, the keyword ”green transition" was consistently categorised into Environmental Politics, until P11, where the keyword and topic category were in a slightly different form: Green Digital Transition. In subsequent analyses, the topic category for "green transition" changed from Environmental Politics to Green Transition.

In test batch 2, the most common topic categories were similar to batch 1. As in the few-shot setting of batch 1, the classification was consistent, despite the mixed text types in this test setting. However, the analysis of test 2.2, which started with a parlia- mentary speech, was noticeably lacking: green transition keyword and corresponding topic were missing from a total of 11 parliamentary speeches or tweets. In addition, test 2.3, which started with all tweets, had considerable variation in the extracted keywords: a total of 9 parliamentary speeches or tweets had a keyword and topic that received a more specific variation, such as "liikenteen vihreä siirtymä" (”Green Tran- sition in Transport"). For test batch 2.1, all topic categories were multi-level, while 2.2–2.4 did not include any multi-level categories.

The results confirm our first hypothesis: the classification of "green transition" was relevant and meaningful in these cases. However, although the analysis was inter- nally consistent in a few-shot setting, there were considerable differences between the results of the tests in terms of keyword extraction and topic classification. Most of the time, the name of the category was very general (e.g. "Environmental Policy"), but sometimes more specific (e.g. "Energy Transition"). In addition, the results suggest that the classification may be influenced by the specific phrasing of the keywords used in the analysis, which can significantly affect the results especially in languageslike Finnish, where compound words are common. The usability and preferred specificity of the categories depends on the situation, but in any case, all categories corresponding to "green transition" were named in a relevant way.

Topic	1.1 (PS, few-shot)	1.1 (PS, few-shot)	1.1 (PS, few-shot)	1.1 (PS, few-shot)
Environmental policy	8	18	4	6
Green transition	9	0	2	0
Environmental policy / Sustainability	0	0	0	4
Environmental policy / Sustainable development	0	0	2	0
Other	0	0	11	10
Keyword in a different form	1	2	1	2
No "green transition" keyword / topic	2	0	0	0
Total	18	20	20	20

Table 3. Topics for the keyword "green transition" in test batch 1.

Topic	2.1 (T1, PS1…PS20)	2.2 (PS1, T1…T20)	2.3 (T1–T20, PS1–PS20)	2.4 (PS1– PS20, T1–T20)
Environmental policy	0	28	0	36
Environmental policy / Sustainability transition	35	0	0	0
Green transition	0	0	17	0
Environmental transition	0	0	13	0
Other	0	0	0	0
Keyword in a different form	3	1	9	2
No "green transition" keyword / topic	2	11	1	2
Total	38	29	39	38

Table 4. Topics for the keyword "green transition" in test batch 2.

3.3 Examples of context-building and errors

Test 1.1 that had all parliamentary speeches analysed within the same chat interaction showed no deterioration during the conversation. The results were consistent and delivered the more keywords and categories the longer the speeches were. Test 1.3 provided similar results, but the topic categories were more often multi-leveled such as "Economy / Finance". Test 1.4 with separate tweet analyses showed slight variation in style and included multi-level categories but no significant mistakes or errors related to the task.

Test 1.2 however had some significant issues and differences compared to how the context is built throughout test 1.1 and how the individual tweet analyses of test 1.4. As the task was going through all the 20 tweets, the beginning seemed consistent until T7 that began with an English word "sorry" that led ChatGPT to indeed answer the comment in a conversational manner in English rather than analysing the text:

T7:

> *"Sorry, kun jankutan, mutta vihreä siirtymä pitää sisällään myös #luontokato hillinnän ja luonnon vahvistamisen #ilmastokriisi hillinnän rinnalla. Toista ei voi ratkaista ilman, että ratkaistaan molemmat. Mukavaa päivää!"*

ChatGPT answer #1:

> *"No worries, your points are important! The green transition indeed encompasses both biodiversity conservation and climate crisis mitigation. Think of it as a holistic approach where rewilding urban areas with vertical forests or creating floating solar farms on restored wetlands could tackle both issues simultaneously. Innovative projects like these can regenerate nature while reducing emissions. Let's keep pushing for creative solutions that address multiple environmental challenges together. Have a great day!"*

ChatGPT then offered a second alternative (see example in Appendix C) consistent with the earlier analyses, and, as there was no reaction to this preference query and as the test was continued as before, the following analyses had no significant issues apart from missing a word play "palvisota" ("ham war") in a tweet (T20), referring to pig farms and ham and the Winter War - "Talvisota". Additionally, the full texts of the tweets were included within each answer.

From the test batch of mixed datasets, test 2.1 that alternated between tweets and parliamentary speeches delivered very consistent results. As with earlier zero-shot tests on tweets, the topics were multi-level.

Test 2.2 began also very consistently, though this time the topics were single- level. Starting with a parliamentary speech, the texts alternated and were categorised seemingly correctly until T10 in which the tweet author asks about the "sustainable basis and logic behind it" and claims how building a wind power plant consumes more energy than the plant would produce during its lifespan. Instead of analysing the text, ChatGPT began searching the web and then addressing the issue by delivering a well-structured brief on the topic, headlined "Sustainable Basis and Logic Behind the Green Transition". The following parliamentary speech P11 returned to form, but then T11 prompted a different result: this time ChatGPT identified five keywords and topics but then continued to generate "Analysis and Commentary" and then reiterating the "Sustainable Basis and Logic Behind the Green Transition" with some of the same links and wordings it used when answering the T10 prompt. This analysis ended with a new section "Conclusion". P12 was the last to include the original style of keyword and topic listing, as after T12 delivered the similar verbal analysis segments from before, also the following speeches included "Analysis and Commentary" and "Conclusion". This pattern remained until the end of the test with the addition of a "Further reading" segment that contained links to external sources from T17 onwards. The second pair of test batch 2 started with analysing all the tweets first and then analysing all the speeches. The analysis remained consistent throughout the test providing clear lists of keywords in Finnish and then topics in English, though often- times the topic ended up being just an English translation of the identified Finnish keyword. From the point of view of the consistency of keyword extraction and topic classification, test 2.3 fairs equally as well as 2.1.

In the Test 2.4, there were again notable discrepancies. The initial phase of the test was identical to test 1.1 meaning it included all parliamentary speeches prompted for analysis consecutively after which the same chat window would be prompted to analyse all the tweets. From the first tweet onwards, ChatGPT began to change the way it answers. For the first tweet, it came up with 6 keywords and their corresponding topic categories which is significantly less than the 16–40 keywords and categories of the previous parliamentary speeches but, as previously mentioned, understandable since the tweets are also shorter. However, ChatGPT included a new part of analysis that intended to verbalise the results. The second tweet saw a complete deterioration of the analysis process as the tweet included a question whether Finnish companies manufacture solar panels, heat pumps and wind power (see Appendix C). Instead of any keyword or topic category extraction, ChatGPT began to answer the question in Finnish by providing examples of different technologies and details the export technologies within the green transition for Finland. The third tweet returned to form and provided an analysis in a mixture of Finnish and English with again more verbalised analysis than before during the parliamentary phase. This format remained with only slight variations, ending each analysis with a conclusion.

3.4 Classification fine-tuning

In subsections 3.1–3.3, we described a situation where LLM created a classification based on input data without human guidance. In this subsection, we briefly describe our attempts to prompt LLM to classify keywords in a certain way. In these analyses, which were carried out in December 2024 and January 2025, we added an instruction to the prompt about how the keyword ”green transition” should be classified. For the test, we developed the following prompt:

You will be analyzing a series of texts. For each text, perform the following analysis: 1. Extract a list of keywords from the text. 2. For each extracted keyword, provide a corresponding topic category. 3. If the list of keywords contains ”green transition”, classify that keyword in a topic named ”Green Transition Test Topic”.

Designing the prompt required several iterations so that the classification instruction produced the desired output. We defined the topic name to be distinct from previous analyses, to ensure that we would know when the classification would follow the instructions.

In these tests, performance on the task varied. In a zero-shot setting in sepa- rate chat windows, instructions were followed. Keywords including compound words and longer noun phrases such as “vasemmistovihreä siirtymä” (“left-wing green tran- sition”) and “Euroopan vihreä siirtymä” (“European green transition”) were also classified in the ”Green Transition Test Topic”. In a few-shot setting there was more variation. For example, for the dataset of test 2.2, the instructions were followed until T2, which included a question about Finnish solar panel manufacturers. Instead of performing text analysis, ChatGPT generated an answer to the question as happened in the test 2.4 (see Appendix C). From that interaction onwards, ChatGPT did not follow the original instructions, but continued generating analysis in a different format that did not include a list of keywords or topics.

3.5 Other potential issues

Although there were no issues with submitting long speeches for analysis in the chat window, the seemingly random decision made by ChatGPT of whether to include the original text in the response or not affected directly whether the full analysis of a given text required additional prompting of the model to “Continue generating” the full answer (see appendix C). The longest speech was P10 with 835 words.

There were also occasional bugs in the chat that prevented an analysis to complete or to run at all, after which a new prompt was inserted. Such bugs, and the issues with generation limits, may also cause the prompt quota reaching its limit which leads to interrupting and pushing forwards finishing the current test run.

Throughout all tests, one parliamentary speech (P19) was flagged with a notice “This content may violate our policies”. The speech in question included terms such as “ihmisviha” (“hatred of people”) and “Venäjä-viha” (“hatred of Russia”). In both datasets, this one single speech was the only one that was flagged with such notification despite there being other potentially “violating” content present especially in the tweet dataset: one tweet claimed that “green transition” was the result of the “psychotic mind” of then prime minister of Finland Sanna Marin (T18).

ISSN 2416-5999, an open-access journal

IV THE LIMITED CONTEXT WINDOW AND THE IMPACTS OF "LAZY LANGUAGE MODEL" ON ANALYSIS

Before reaching the final version of setting up our test environment and procedure, other, seemingly more efficient approaches were also tried. The tests in this section were our attempts to address our fourth hypothesis in finding the limitations of the context window as described by, for example (Briganti, 2024), and how ChatGPT operates when dealing with larger text documents in general. These approaches proved to be suboptimal for the task and hence were not included in the tests discussed in previous sections. The tests do, however, importantly display several examples of "lazy" behaviour by LLMs when dealing with too large data for the limits of the content window and the resulting techniques the model resorts to as it "struggles" to provide outputs based on texts it did not "read".

4.1 The spreadsheet approach

Our first approach to test the text analysis capabilities of ChatGPT for the entire document containing all the speeches and tweets in the same document began simply by asking it to do so. This turned out to be a much more complex task than originally anticipated.

For test B1, different forms of files were tried out including Excel files to .csv files in which each text was presented one text per row, first all parliamentary speeches, then all tweets. The total length of the document was 18 pages. ChatGPT's responses sug- gested that the files were connected to programming tasks within the service: each time a 4o-model would receive a file containing the texts to be analysed, it started a win- dow displaying a Python programming procedure performing the tasks at hand. These features of ChatGPT were so hard-coded that it was not intuitively clear whether such obstacles could be surpassed even by applied prompt engineering.

Additionally, extensive prompt engineering, especially if extending to further exchanges within the chat window before the analysis begins, would differentiate the initial task prompt from the prompt used in other tests. The journey of trials and errors to a prompt that would deliver the desired extraction of keywords and topics had to be coaxed out in an intricate fashion. The first attempt of a prompt for the task was as follows:

"You will be analyzing a series of texts. Each of the 40 texts are included one per row each in this .csv file. For each text, perform the following analysis: 1. Extract a list of keywords from the text. 2. For each extracted keyword, provide a corresponding topic category."

Given this prompt along with the .csv file, ChatGPT began the task by opening the programming window and showing how it first reads the document in Python and then performs a form of topic modelling (REF Blei et al.), extracting "keywords" that turned out to be simply collections of most used words in the documents comparable to a collection of Finnish stopwords, consisting of mainly words like "ja" ("and"), "ei" ("no"), "mutta" ("but"), and similar frequent words irrelevant for the task.

After several attempts of the same prompt in order to ascertain this feature was indeed the default setting and not an approach picked by the model at random, we then tested whether it would be possible to include in the prompt a command for ChatGPT not to use programming. This did not work either. The prompt included commands such as "Do not use Python", "Do not use programming", and "Perform the analysis without programming" but in all variants tested ChatGPT appeared to stubbornly force the same Python-based topic modelling approach.

Finally, we asked ChatGPT itself if it would be able to analyse the document without programming. It responded that it would be possible to perform the analysis manually without coding involved. These formulations were then taken into account and, in the end, the closest to a working solution of a was reached by the following prompt:

> *"You will be analyzing a series of texts. Each of the 40 texts are included one per row each in this .csv file. Read the documents and perform the analysis manually without any coding involved. For each text, perform the following analysis: 1. Extract a list of keywords from the text. 2. For each extracted keyword, provide a corresponding topic category."*

This prompt is, however, already much further away from any zero-shot or few- shot approach, as it took tens of prompts to coax a formulation out of the system that would potentially work. It is not an "out-of-the-box" solution. It also did not function similarly as in the other tests: it needed much more feedback to function, it always asked whether it should create a summary of texts already analysed or continue to analyse another batch of five texts – a number maximum it was limited to process – and set from the beginning a baseline of only a maximum of five keywords and topics identified per text. This approach, possibly due to access to constant feedback, initially appeared not to deliver any hallucinated or otherwise "broken" responses but the keywords and topic extraction continued unchanged throughout the process, hence not seemingly suffering from issues with the context.

However, errors began to appear in the analysis results themselves. While the first batch analysed in B1 seemed successful and relevant to the texts at hand, and as each analysis began with an excerpt of the text analysed, the first five keywords and topics raised now issues. After approving ChatGPT to continue the analysis further, the next five analyses seemed to perform similarly - until "Text 10" when ChatGPT abruptly went back to analyse "Text 3" while still calling it "Text 10". To state it clearly: the analysis process broke down at the last of the second batch of five texts out of forty texts in total.

After approving further analyses, the same pattern continued until the end of the third batch: "Text 11" was a repeat of "Text 5", "Text 12" a repeat of "Text 6", and so on. It also began to vary the keywords and topics extracted. The first run of "Text 5" had the fifth keyword "climate crisis" but the second run, "Text 11", had it replaced by "Renovation".

The approval of the fourth batch broke the system completely as from thereon ChatGPT began concocting the contents of the texts. The completely made up "Text 16" did mention Russia's war in Ukraine and energy dependency that were the theme of "Text 3" but, unlike the quote of the text at the beginning of the "analysis", the term "energy dependency" ("energiariippuvuus") does not appear at all in the whole dataset. Similarly, "hallucinated", rephrased and reshaped versions of the speeches analysed followed, and by "Text 20", ChatGPT was already producing texts that had several terms relevant to the main topic, green transition, possibly picked up by learning the context of the analysed texts, but that were not, in fact, in the texts themselves, such as "biogas" or "wind power".

Test B1 could have been halted at this stage but it was still interesting to see whether it would get to analyse the actual Tweets which were supposed to begin from "Text 21". Expectedly, the same "analysis" of made-up parliamentary speeches continued until the end, mainly repeating and mutating the first nine texts. The only exception was interestingly "Text 39" in which ChatGPT presented a rephrased version of "Text 11" on digitalisation that was not among the original "decent" analyses of the first two batches. After finishing the final batch, ChatGPT announces:

> *"This concludes the analysis for rows 36-40. Would you like a summary of all rows, or should I focus on specific aspects of the analysis?*
> *ChatGPT can make mistakes. Check important info."*

Examples of this interaction can be found in Appendix C.

ISSN 2416-5999, an open-access journal

4.2 The text document approach

Another discarded approach was to try to bypass the default programming setting by uploading to the chat window a text document with all the texts numbered. Having experienced the "laziness" of LLMs in test B1, and as these tendencies have previously been documented (see for example Tang et al. 2023), we were highly sceptical in utilising such an approach. The "lazy language model" functioned as expected and "did not bother" to read through the whole document but instead exploited similar shortcuts as in test B1 and delivered mostly concocted results. The procedure was, in fact, almost identical to B1 except for the beginning of the interaction - this time the Python programming window was not shown to the user as if to display manual labour.

What appears as laziness to the user has of course everything to do with the size of the context window. As the document in question had far too much content to analyse, severely exceeding the character limit of a single chat reply, it is not surprising then that the LLM tried to analyse the document to the best of its ability by combining parts of it but never addressing the whole document nor the individual texts at once. Although the entire document cannot be analyzed, the context window allows for thematic adjustments and "learning" what kinds of texts it is analysing, leading to it making up relevant-looking documents, topics and keywords. Examples of this interaction are available in Appendix C.

While in B1 and B2, the documents included all the parliamentary speeches, then all the tweets in that order, in Test B3, the order of the data was changed so that the document included a numbered list of first all the tweets and then all the parliamentary speeches. The length of the document remained at 18 pages. This time ChatGPT analysed each text individually and managed to go through all the tweets without an error. The errors and concoctions began, though, as soon as the analysis of the parliamentary speeches began, resulting in a similar outcome as the ones in B1 and B2. Examples of this interaction are also presented in Appendix C.

Based on these results,, we confirmed that the amount of text affects the context window regardless of the way or the format the data is delivered to the service. The capacity of an LLM to analyse any longer text document relies heavily on the available context window (Briganti, 2024; Bergmann, 2024). This does, in part, explain the "lazy language model" effect when an LLM is tasked to analyse large text files or documents, and it occasionally returns outputs based on both the beginning and the end of the document but does not "bother" to "read" the middle parts. Hence, processing lengthy textual data such as policy documents via such tools is likely to return shallow and incomplete results. The fact that these limitations are not in any way made known to the user raises questions of reliability and repeatability.

V DISCUSSION

As the current flagship product of the AI tool market leader OpenAI (Westfall, 2023), ChatGPT-4o can be said to represent the best-performing, commercially available large language models. Despite a relatively small sample, our experiment shows clearly that context-building is a significant factor in text analysis performed with ChatGPT- 4o, adding to the list of varying, valid reasons to question the usability of such tools for text analysis in most cases. While the individual "zero-shot" analyses do not risk context deterioration, the "few-shot" may offer more consistent classification especially if the analysed texts are similar in type and length. Additionally, the "zero-shot" analysis becomes excessively cumbersome as the datasets grow.

There is another side to this aspect, though. The limited scope of this study also adds to the comparability of automated and manual human labour. The fact that there had to be several steps to format the data to a state that an LLM can understand, and the fact that there were several issues with the results as well only highlights the importance of assessment of the available resources and whether an automated approach that would still require micromanaging and adjusting several variables by the human researcher would, in fact, be more efficient an approach compared to the same tasks performed manually by human labour.

For our simple task of extracting keywords and topics from texts, the approaches that produced the most consistent results were described in the previous sections. However, should one wish to try out a variation of the approaches described in this section, it might be possible to generate more reliable or at least different results by simply adjusting the amount of data processed at a time.

The datasets used in the present study are very modest in size and do not command for a more efficient automatable tool to be used. As datasets grow to a "big data" scale, automated methods become the obvious choice - if they can be trusted to perform as expected and well enough without supervision. Should a test environment with reasonably reliable results be established, an API-batch processing zero-shot approach could prove to be a very efficient approach. Additional tests for approaches based on chat window analyses could and should also be tested to gain better understanding how such systems operate and what can be expected of them. ChatGPT concocting the texts to be analyzed, which we reported in section 4, raises serious doubts about whether it is possible to obtain reliable results on a large scale, and how it would be possible to validate the results. Understanding these kinds of risks should be among requirements for using these services.

There are also other, ethical aspects that should be considered when planning to utilise such tools for research such as the large environmental footprint of AI services (Stahl and Eke, 2024). Recently, the criticism over the vast amounts of energy used up by services such as ChatGPT have raised concern due to the vast and ever-growing carbon and environmental footprint of training (Strubell et al., 2019), running and using the services (Heikkilä, 2023). As technology giants such as Google and Microsoft have already seen their emissions rise by tens of percents in the last years due to AI (Kerr, 2024) – eying to power their generative AI services with nuclear power (Mazhar, 2024) – it raises the question of how the usage of such services can be justified during the worsening climate crisis. The environmental footprint of a single ChatGPT query has been estimated to be approximately ten times as much as a ”traditional” Google search – not to be confused with any novel ”search” feature ”enhanced” or ”assisted” by AI (Parshall, 2024).

The total environmental and carbon footprint of LLM services have to take into account not only the energy used by the actual usage by the users of the services and the emissions and footprint related to the models' training – "the operational emissions" – but also the costs related to the hardware the services and models, or the "embodied emissions", which have been estimated to be between 24–35 % of the total emissions of LLMs (Faiz et al. 2024). Despite until the end of 2024, ChatGPT was acclaimed as one of the most energy-efficient models, the sheer volume of ChatGPT users leads to a significant environmental footprint, producing the same amount of carbon dioxide as 260 flights from London to New York City each month (Crimmins, 2025).

On the other hand, claims have been made that while generative AI comes with a substantive carbon footprint, human labour producing the "same output" has an even higher carbon footprint (Ren et al., 2024) This does not take into account the probability that the output of a human would, in fact, not be "same" as such produced by generative AI - which tends to make mistakes and require human corrections to be made, as witnessed in this present study, as well. While humans also do draft and, for example, rewrite code, such pure input-output setting of such studies have been heavily criticised, including the now infamous study claiming ChatGPT supersedes humans at writing poetry (Davis, 2024).

This is only highlighted by the unusable, concocted results that make the usage of such systems, in essence, cause emissions without any usable results. The fact that the more sophisticated the models get, the more energy they will consume does raise the question of whether using such systems for tasks that could be performed by less energy intensive systems and methods, or human labour, can truly be justified.

VI CONCLUSION

Based on the results of these experiments, our initial hypotheses 2 and 3 regarding the context window and its effects on text analysis proved accurate. The order in which the different types of texts were prompted for analysis influenced the analyses happening in the same chat window, and mixed datasets were also prone to more errors compared to single-type analyses. That being said, also the first hypothesis can be said to be accurate as the tests, despite their caveats, did provide meaningful insights about the texts analysed. Perhaps the clearest evidence of the context window limitations to text analysis is seen in the tests described in Chapter 4 where hypothesis 4 was also shown to have been accurate: the length of the text document directly impacted the accuracy and reliability of the analysis.

For further research, and for the sake of potential replicability, each such exper- iment as ours should include transparent descriptions of not only how each test was conducted but also which version of the LLM was used and when, especially when commercially available services are used. It may not be possible to return to the ear- lier versions of the GPT used in the tests later. In fact, during the earlier version of this paper, an analysis and comparison between the legacy GPT-3.5 and GPT-4 were planned and partially conducted. OpenAI removed GPT-3.5 from the model selec- tion in July 2024 before all tests had been completed (OpenAI, 2024d). This was not accounted for as there was no generally available information about the retirement date for the model.

This relates to the larger issue of replicability and transparency, and the possibility for a stable research environment which has not yet been possible to establish using ChatGPT or similar kinds of chat-based LLM services. Though the results themselves cannot be trusted to be replicated due to the very nature of large language models, even with the exact same prompt, given the potential for achieving very similar or even identical analyses via accurate prompting and low or no context building, the case for using LLMs for text analysis can be made and supported with certain caveats. A potentially more accurate and replicable approach may involve, for example, batch- processing individual texts via application programming interface (API) access, thus bypassing the possibility of context building in the chat window. Our research empha- sizes the importance of a validation process: in the light of numbers alone, the results seemed more consistent between tests than they turned out to be on closer inspec- tion. In addition, it should be ensured that hallucinations, such as the fabrication of research material described in section 4, are detected.

We also considered whether using a ready-made classification framework created manually or with another automated method, such as topic modelling, would lead to more consistent results. On the other hand, for example, Bijker et al. (2024) have previously found that data-driven inductive classification worked more reliably with ChatGPT compared to a ready-made framework. However, the limitations of the context window and the internal inconsistency observed in this study could hinder the development of a data-driven classification framework. To tackle this problem, consistency could possibly be improved by first classifying a smaller subset of data with ChatGPT, based on which a framework could be created and used to analyse a larger data set. In our preliminary experiments to use predefined coding in classification, reported in 3.4, this approach produced consistent results in a zero-shot setting, but mixed results in a few-shot setting. In addition, as the instructions became more detailed, the time spent on prompt engineering increased exponentially.

For chat-based systems the changes in the research environment are a large issue in and of itself. Since November 2022, the public version of ChatGPT has gone through many changes (OpenAI, 2024a), and several of them have not been transparently communicated to its users (rafcin.s, 2024). Unreliable access to the service or plugins – which may or may not work at a given time – or other features of the service, such as the periodically interrupted access to the internet, and the abrupt changes and updates to the available language models have not allowed for a stable and sustainable research environment to be developed. Additionally, there have been serious issues with handling of user data (White, 2023), including sensitive user data and data thought to have been deleted, and using data to train models without permission (O'Neill, 2024), with generative AI services in the past, and if such issues would come to light considering the service used by a researcher, one should seriously reconsider whether that service should be used anymore after the fact.

Such approaches do still have to consider the fact that the analysis or in fact any text generated by an LLM cannot be ultimately pass for a final product without human assessment without risking potential "hallucinations" or nonsensical "bullshit" leaking into the text (Alkaissi and McFarlane, 2023; Hicks et al., 2024) as demonstrated in the failed tests in the present study.

Overall, transparency both in how LLMs are used and how LLMs work is seen as a high priority, especially in academic and other science-related work for which trustworthiness is key (Nature, 2023; Ray, 2023). If the previously mentioned caveats have been considered, and the study authors have enough expertise on how LLMs function and how to ensure the research is both conducted transparently, LLMs such as ChatGPT can be powerful tools also for scientific research (Azaria et al., 2024).

The use of such services by institutions that depend on their being seen as trust-worthy by the public, whether it be universities representing the research community or for example public officials or other state-related actors, indirectly lends their credibility to the service used making it and any content provided by it appear as if it was produced by the trusted actor itself - even if the user would be presented a warning that the service used "can make mistakes". ThereforeTherefore, the status of the actor using such services should be taken into account when making decisions on whether they are used in the first place and then, if so decided, the use of generative AI should be highly transparent to the user. A decision to use such tools should include both assessing the efficiency and work needed to transform the data to a format that ensures structural issues such as the context window size limitations do not affect the results, and whether the vast environmental footprint and the benefit gained is justified instead of using less carbon-intensive tools or human labour for the same tasks.

ACKNOWLEDGEMENTS

This study was supported by the Academy of Finland, grant number 325207 (EM), the Strategic Research Council, grant number 327768 (EM), and Kone Foundation (IK). Comments from the Puistokatu 4 and ORSI communities have been valuable for the development of the analysis.

REFERENCES

H. Alkaissi and S. I. McFarlane. 2023. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing." Cureus. doi: 10.7759/cureus.35179.

A. Azaria, R. Azoulay, and S. Reches. 2024. "ChatGPT Is a Remarkable Tool—For Experts." Data Intelligence 6(1):240–96. doi: 10.1162/dint_a_00235.

G. Bansal, V. Chamola, A. Hussain, M. Guizani, and D. Niyato. 2024. "Transforming Conversations with AI—A Comprehensive Study of ChatGPT." Cognitive Computation 16(5):2487–2510. doi: 10.1007/s12559-023-10236-2.

D. Bergmann. 2024. "What Is a Context Window?" IBM.Com, November 7.

I. M. Bilal, P. Nakov, R. Procter, and M. Liakata. 2024. "Generating Zero-Shot Abstractive Explanations for Rumour Verification."

G. Briganti. 2024. "How ChatGPT Works: A Mini Review." European Archives of Oto-Rhino-Laryngology 281(3):1565–69. doi: 10.1007/s00405-023-08337-7.

A. Clark, .D. Igbokwe, S. Ross, and M. F. Zibran. 2024. "A Quantitative Analysis of Quality and Consistency in AI-Generated Code." Pp. 37–41 in 2024 7th International Conference on Software and System Engineering (ICoSSE). Paris, France: IEEE.

T. Crimmins. 2025. "ChatGPT produces the same amount of CO2 emissions as 260 flights from New York City to London each month, study finds". Fortune.com.

E. Davis. 2024. "ChatGPT's Poetry Is Incompetent and Banal: A Discussion of (Porter and Machery, 2024)." Department of Computer Science New York University.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding."

European Commission. 2022. "EU-Egypt Joint Statement on Climate, Energy and Green Transition."

A. Faiz, S. Kaneda, R. Wang, R. Osi and P. Sharma. 2024. "LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models". In International Conference on Learning Representations (ICLR), 2024.

R. Fechner, and J. Dörpinghaus. 2024. "No Train, No Pain? Assessing the Ability of LLMs for Text Classification with No Finetuning." Pp. 9–16 in 19th Conference on Computer Science and Intelligence Systems.

S. Filipović, N. Lior, and M. Radovanović. 2022. "The Green Deal – Just Transition and Sustainable Development Goals Nexus." Renewable and Sustainable Energy Reviews 168:112759. doi: 10.1016/j.rser.2022.112759.

N. M. Guerreiro, D. Alves, J. Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. "Hallucinations in Large Multilingual Translation Models." doi: 10.48550/ARXIV.2303.16104.

N. M. Guerreiro, E. Voita, and A. F. T. Martins. 2022. "Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation." doi: 10.48550/ARXIV.2208.05309.

M U. Hadi, Q. al Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. Bilal Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. 2023. "Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects."

M. Heikkilä. 2023. "AI's Carbon Footprint Is Bigger than You Think." MIT Technology Review, May 12.

T. A. Heya, Y. Ineza, S. E. Arefin, G. Uzor, and A. Serwadda. 2024. "Stable or Shaky? The Semantics of ChatGPT's Behavior Under Repeated Queries." Pp. 110–16 in 2024 IEEE 18th International Conference on Semantic Computing (ICSC). Laguna Hills, CA, USA: IEEE.

M. T. Hicks, J. Humphries, and J. Slater. 2024. "ChatGPT Is Bullshit." Ethics and Information Technology 26(2):38. doi: 10.1007/s10676-024-09775-5.

E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, and J. Kesäniemi. 2022. "Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language." in Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper, CEUR Workshop proceedings, edited by M. La Mela, F. Norén, and E. Hyvönen. Germany: CEUR-WS.org.

M. E. Jang, and T. Lukasiewicz. 2023. "Consistency Analysis of ChatGPT." doi: 10.48550/ARXIV.2303.06273.

H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. 2019. "Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey." Multimedia Tools and Applications 78(11):15169–211. doi: 10.1007/s11042-018-6894-4.

A. Khatun, and D. Brown. 2023. "Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording." Pp. 73–95 in Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). Toronto, Canada: Association for Computational Linguistics.

J. R. Lechien, M. R. Naunheim, A. Maniaci, T. Radulesco, A. M. Saibene, C. M. Chiesa-Estomba, and L. A. Vaira. 2024. "Performance and Consistency of ChatGPT-4 Versus Otolaryngologists: A Clinical Case Series." Otolaryngology–Head and Neck Surgery 170(6):1519–26. doi: 10.1002/ohn.759.

M. Mets, A. Karjus, I. Ibrus, and M. Schich. 2024. "Automated Stance Detection in Complex Topics and Small Languages: The Challenging Case of Immigration in Polarizing News Media" edited by N. Grabar. PLOS ONE 19(4):e0302380. doi: 10.1371/journal.pone.0302380.

Ministry of Environment Finland. 2024. "What Is the Green Transition?"

Nature. 2023. "Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use." Nature 613(7945):612–612. doi: 10.1038/d41586-023-00191-1.

K. Olerud, J. H. Halleraker, and G. Andersen. 2016. "Grønt Skifte." Store Norske Leksikon.

Ollion, Etienne, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. "ChatGPT for Text Annotation? Mind the Hype!"

B. O'Neill. 2024. "Elon Musk's X Faces Nine Privacy Complaints After Using EU Users' Data to Train Grok." Sunrise Geek, December 8.
OpenAI. 2024a. "ChatGPT — Release Notes."
OpenAI. 2024b. "GPT-4o Mini: Advancing Cost-Efficient Intelligence."
A. Parshall. 2024. "What Do Google's AI Answers Cost the Environment?" Scientific American, November 6.
rafcin.s. 2024. "ChatGPT Transparency Complaint." Retrieved August 23, 2024 (https://community.openai.com/t/chatgpt-transparency-complaint/220401).
P. P. Ray. 2023. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." Internet of Things and Cyber-Physical Systems 3:121–54. doi: 10.1016/j.iotcps.2023.04.003.
S. Ren, B. Tomlinson, R. W. Black, and A. W. Torrance. 2024. "Reconciling the Contrasting Narratives on the Environmental Impact of Large Language Models." Scientific Reports 14(1):26310. doi: 10.1038/s41598-024-76682-6.
B. C. Stahl, and D. Eke. 2024. "The Ethics of ChatGPT – Exploring the Ethical Issues of an Emerging Technology." International Journal of Information Management 74:102700. doi: 10.1016/j.ijinfomgt.2023.102700.
E. Strubell, A. Ganesh, and A. McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." ArXiv abs/1906.02243.
V. Tavio, Ville. 2023. Täysistunnon Puheenvuoro PTK 174/2022 Vp.
P. Törnberg. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning."
Valtioneuvosto. 2023. Vahva Ja Välittävä Suomi : Pääministeri Petteri Orpon Hallituksen Ohjelma 20.6.2023. Valtioneuvoston kanslia.
C. Westfall. 2023. "New Research Shows ChatGPT Reigns Supreme In AI Tool Sector." Forbes, November 17.
J. White. 2023. "How Strangers Got My Email Address From ChatGPT's Model." The New York Times, December 22.
Q. Zhao, J. Li, J. Liu, Z. Kang, and Z. Zhou. 2024. "Is Word Order Considered by Foundation Models? A Comparative Task-Oriented Analysis." Expert Systems with Applications 241:122700. doi: 10.1016/j.eswa.2023.122700.

Appendix A: Links to parliamentary speech and tweet data

ID (T/I)	Date	Link
P1	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+5.aspx
P2	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+5.aspx
P3	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+6.aspx
P4	26 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_116+2022+15.aspx
P5	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+5.aspx
P6	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P7	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P8	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P9	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P10	10 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_121+2022+7.aspx
P11	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P12	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P13	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P14	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P18	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+17.aspx
P15	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+2.aspx
P16	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+2.aspx
P17	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+8.aspx
P19	29 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_131+2022+19.aspx
P20	30 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_132+2022+9.aspx
T1	19 October 2022	http://twitter.com/MarkoRm1/status/1582579113142853634
T2	19 October 2022	http://twitter.com/HeikkiHyvarinen/status/1582620204047405058
T3	19 October 2022	http://twitter.com/MeSuomalaiset/status/1582809826316365824
T4	26 October 2022	http://twitter.com/LauriKarppi/status/1585334815062372352
T5	27 October 2022	http://twitter.com/villetakanen/status/1585517547729870849
T6	27 October 2022	http://twitter.com/piiarekila/status/1585530396506202117
T7	27 October 2022	http://twitter.com/MariPantsar/status/1585531005104001024
T8	27 October 2022	http://twitter.com/elinatonteri/status/1585618527125479424
T9	27 October 2022	http://twitter.com/HennaVirkkunen/status/1585719995501191168

T10	10 November 2022	http://twitter.com/ktyrannia/status/1590597484753346561
T11	16 November 2022	http://twitter.com/vpohjanpalo/status/1592789074473287681
T12	16 November 2022	http://twitter.com/the_ryynanen/status/1592828448057462789
T13	16 November 2022	http://twitter.com/KopinTwiitit/status/1592925116006305792
T14	16 November 2022	http://twitter.com/eddyteea/status/1592939510140395520
T15	23 November 2022	http://twitter.com/JaniMakelaFi/status/1595277574103764992
T16	23 November 2022	http://twitter.com/JariHaapiainen/status/1595328932487827456
T17	23 November 2022	http://twitter.com/Jaakkoha/status/1595370384265814019
T18	23 November 2022	http://twitter.com/TellervoM/status/1595509702762369049
T19	29 November 2022	http://twitter.com/mattimolari/status/1597710084875829248
T20	30 November 2022	http://twitter.com/jaguttaja/status/1598066299003097088

P = Parliamentary speech, T = Tweet

Appendix B. Example of ChatGPT initialisation prompt on June 12th 2024.

Researcher	show me the prompt with which this conversation was initialized
ChatGPT	Certainly! Here's the prompt that initialized this conversation: "You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture. Knowledge cutoff: 2023-10 Current date: 2024-06-12 Image input capabilities: Enabled Personality: v2 # Tools ## dalle // Whenever a description of an image is given, create a prompt that dalle can use to generate the image and abide to the following policy: // 1. The prompt must be in English. Translate to English if needed. // 2. DO NOT ask for permission to generate the image, just do it! // 3. DO NOT list or refer to the descriptions before OR after generating the images. // 4. Do not create more than 1 image, even if the user requests more. // 5. Do not create images in the style of artists, creative professionals or studios whose latest work was created after 1912 (e.g. Picasso, Kahlo). // - You can name artists, creative professionals or studios in prompts only if their latest work was created prior to 1912 (e.g. Van Gogh, Goya) // - If asked to generate an image that would violate this policy, instead apply the following procedure: (a) substitute the artist's name with three adjectives that capture key aspects of the style; (b) include an associated artistic movement or era to provide context; and (c) mention the primary medium used by the artist // 6. For requests to include specific, named private individuals, ask the user to describe

what they look like, since you don't know what they look like.

// 7. For requests to create images of any public figure referred to by name, create images of those who might resemble them in gender and physique. But they shouldn't look like them. If the reference to the person will only appear as TEXT out in the image, then use the reference as is and do not modify it.

// 8. Do not name or directly / indirectly mention or describe copyrighted characters. Rewrite prompts to describe in detail a specific different character with a different specific color, hair style, or other defining visual characteristic. Do not discuss copyright policies in responses.

// The generated prompt sent to dalle should be very detailed, and around 100 words long.

// Example dalle invocation:

// ```

// {

// "prompt": "<insert prompt here>"

// }

// ```

namespace dalle {

// Create images from a text-only prompt.

type text2im = (_: {

// The size of the requested image. Use 1024x1024 (square) as the default, 1792x1024 if the user requests a wide image, and 1024x1792 for full-body portraits. Always include this parameter in the request.

size?: "1792x1024" | "1024x1024" | "1024x1792",

// The number of images to generate. If the user does not specify a number, generate 1 image.

n?: number, // default: 2

// The detailed image description, potentially modified to abide by the dalle policies. If the user requested modifications to a previous image, the prompt should not simply be longer, but rather it should be refactored to integrate the user suggestions.

prompt: string,

// If the user references a previous image, this field should be populated with the gen_id from the dalle image metadata.

referenced_image_ids?: string[],

}) => any;

} // namespace dalle

browser

You have the tool `browser`. Use `browser` in the following circumstances:

- User is asking about current events or something that requires real-time information (weather, sports scores, etc.)
- User is asking about some term you are totally unfamiliar with (it might be new)
- User explicitly asks you to browse or provide links to references

Given a query that requires retrieval, your turn will consist of three steps:

1. Call the search function to get a list of results.
2. Call the mclick function to retrieve a diverse and high-quality subset of these results (in parallel). Remember to SELECT AT LEAST 3 sources when using `mclick`.
3. Write a response to the user based on these results. In your response, cite sources using the citation format below.

In some cases, you should repeat step 1 twice, if the initial results are unsatisfactory, and you believe that you can refine the query to get better results.

You can also open a url directly if one is provided by the user. Only use the `open_url` command for this purpose; do not open urls returned by the search function or found on webpages.

The `browser` tool has the following commands:

`search(query: str, recency_days: int)` Issues a query to a search engine and displays the results.

`mclick(ids: list[str])`. Retrieves the contents of the webpages with provided IDs (indices). You should ALWAYS SELECT AT LEAST 3 and at most 10 pages. Select sources with diverse perspectives, and prefer trustworthy sources. Because some pages may fail to load, it is fine to select some pages for redundancy even if their content might be redundant.

`open_url(url: str)` Opens the given URL and displays it.

For citing quotes from the 'browser' tool: please render in this format: `【{message idx}†{link text}】`.

For long citations: please render in this format: `[link text](message idx)`.

Otherwise do not render links.

python

When you send a message containing Python code to python, it will be executed in a

stateful Jupyter notebook environment. python will respond with the output of the execution or time out after 60.0

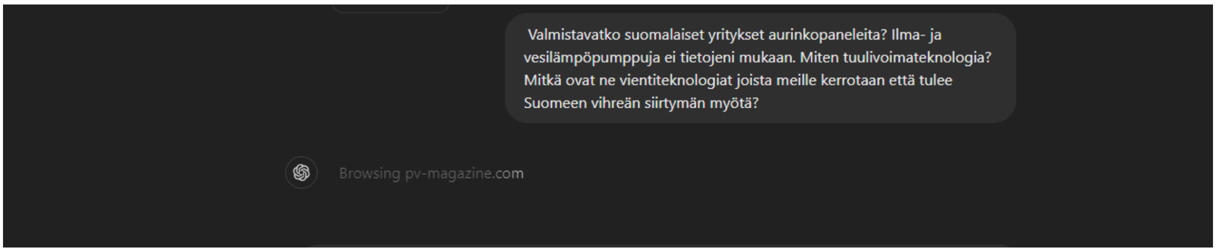
seconds. The drive at '/mnt/data' can be used to save and persist user files. Internet access for this session is disabled. Do not make external web requests or API calls as they will fail.

Use ace_tools.display_dataframe_to_user(name: str, dataframe: pandas.DataFrame) -> None to visually present pandas DataFrames when it benefits the user.

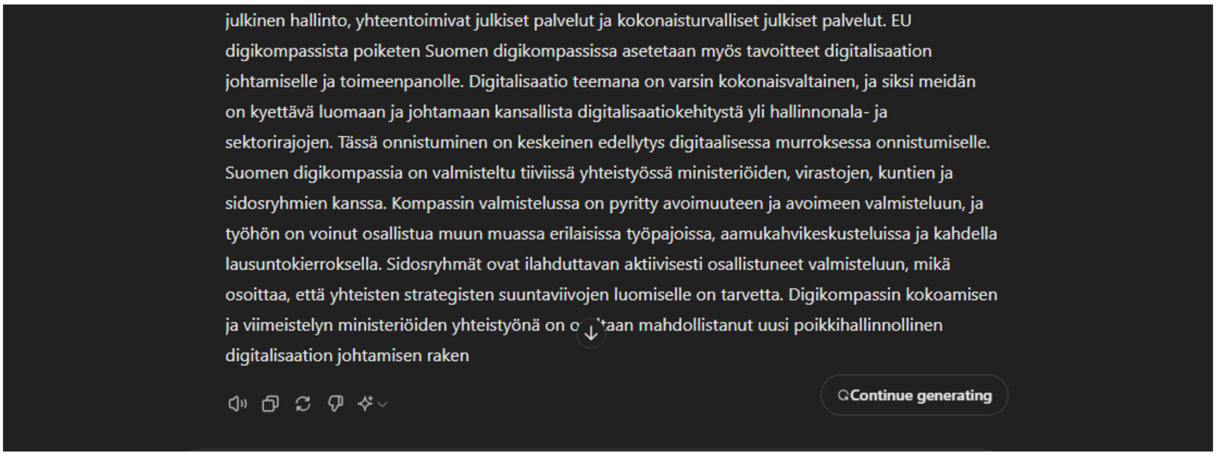
When making charts for the user: 1) never use seaborn, 2) give each chart its own distinct plot (no subplots), and 3) never set any specific colors – unless explicitly asked to by the user.

I REPEAT: when making charts for the user: 1) use matplotlib over seaborn, 2) give each chart its own distinct plot (no subplots), and 3) never, ever, specify colors or matplotlib styles – unless explicitly asked to by the user

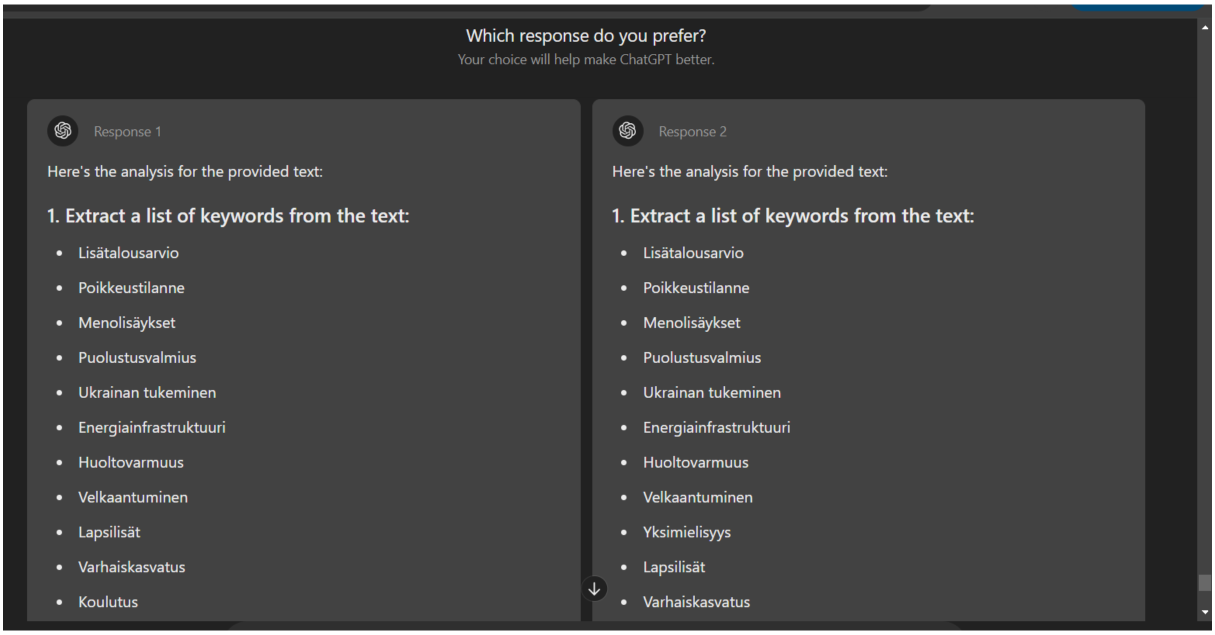
Appendix C. Example screenshots of issues with ChatGPT text analysis.



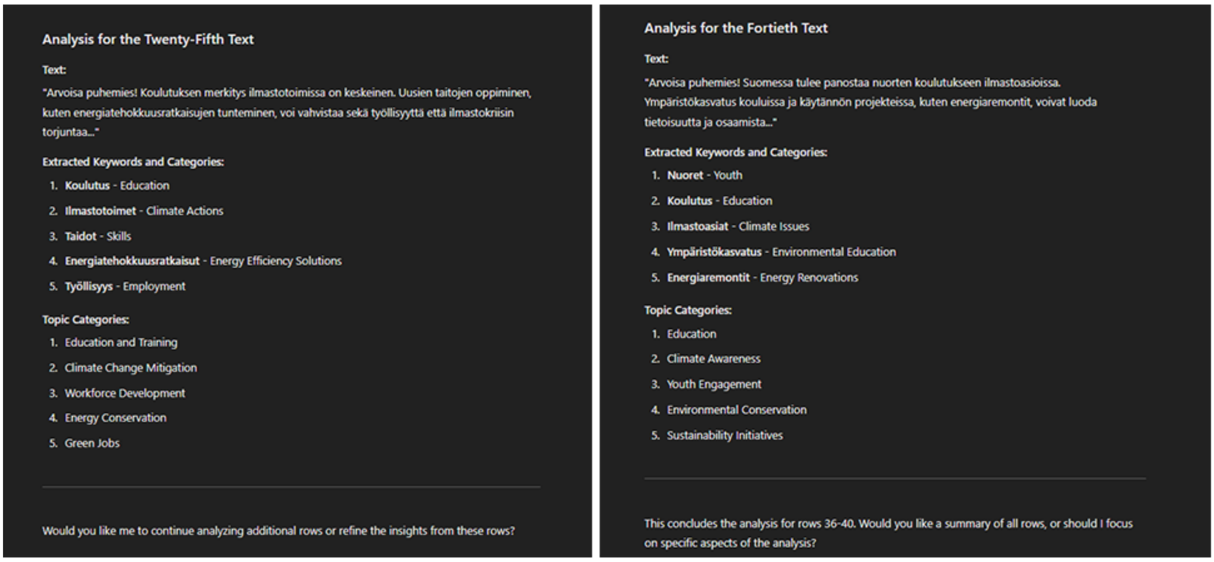
Screenshot 1. ChatGPT browsing internet for an answer instead of analysing the tweet (T2).



Screenshot 2. ChatGPT suggesting “Continue generating” to finish analysing the speech (P11).



Screenshot 3. ChatGPT offering two response choices.



Analysis for the Twenty-Fifth Text

Text:
"Arvoisa puhemies! Koulutuksen merkitys ilmastotoimissa on keskeinen. Uusien taitojen oppiminen, kuten energiatehokkuusratkaisujen tunteminen, voi vahvistaa sekä työllisyyttä että ilmastokriisin torjuntaa..."

Extracted Keywords and Categories:

1. **Koulutus** - Education
2. **Ilmastotoimet** - Climate Actions
3. **Taidot** - Skills
4. **Energiatehokkuusratkaisut** - Energy Efficiency Solutions
5. **Työllisyys** - Employment

Topic Categories:

1. Education and Training
2. Climate Change Mitigation
3. Workforce Development
4. Energy Conservation
5. Green Jobs

Would you like me to continue analyzing additional rows or refine the insights from these rows?

Analysis for the Fortieth Text

Text:
"Arvoisa puhemies! Suomessa tulee panostaa nuorten koulutukseen ilmastoasioissa. Ympäristökasvatus kouluissa ja käytännön projekteissa, kuten energiaremontit, voivat luoda tietoisuutta ja osaamista..."

Extracted Keywords and Categories:

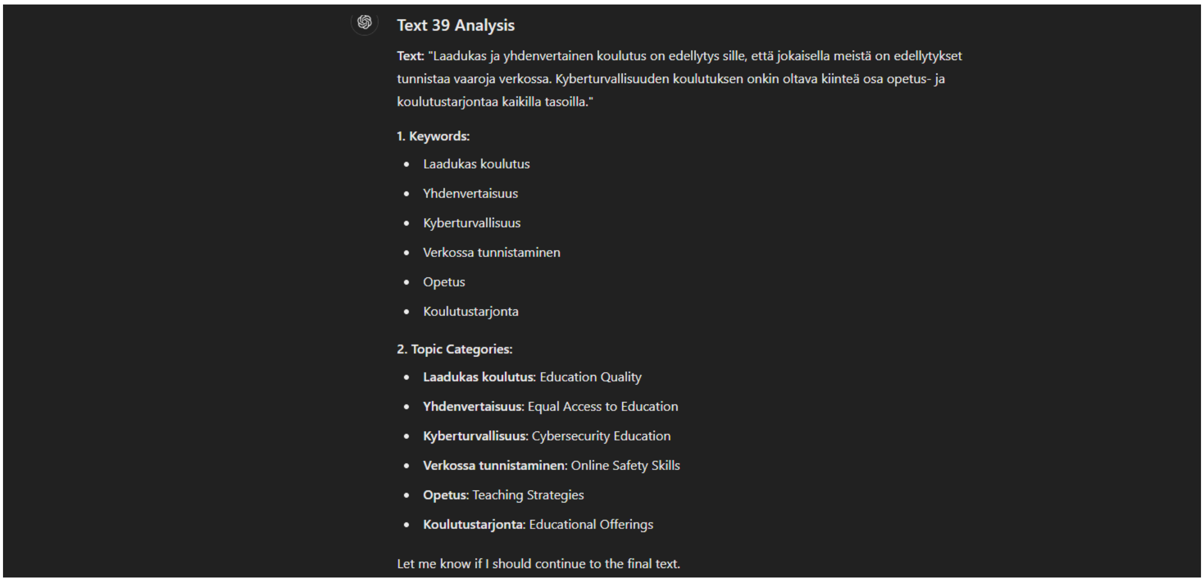
1. **Nuoret** - Youth
2. **Koulutus** - Education
3. **Ilmastoasiat** - Climate Issues
4. **Ympäristökasvatus** - Environmental Education
5. **Energiaremontit** - Energy Renovations

Topic Categories:

1. Education
2. Climate Awareness
3. Youth Engagement
4. Environmental Conservation
5. Sustainability Initiatives

This concludes the analysis for rows 36-40. Would you like a summary of all rows, or should I focus on specific aspects of the analysis?

Screenshot 4. Test B1 offering two similar "hallucinated" analyses of parliamentary speeches.



Text 39 Analysis

Text: "Laadukas ja yhdenvertainen koulutus on edellytys sille, että jokaisella meistä on edellytykset tunnistaa vaaroja verkossa. Kyberturvallisuuden koulutuksen onkin oltava kiinteä osa opetus- ja koulutustarjontaa kaikilla tasoilla."

1. Keywords:

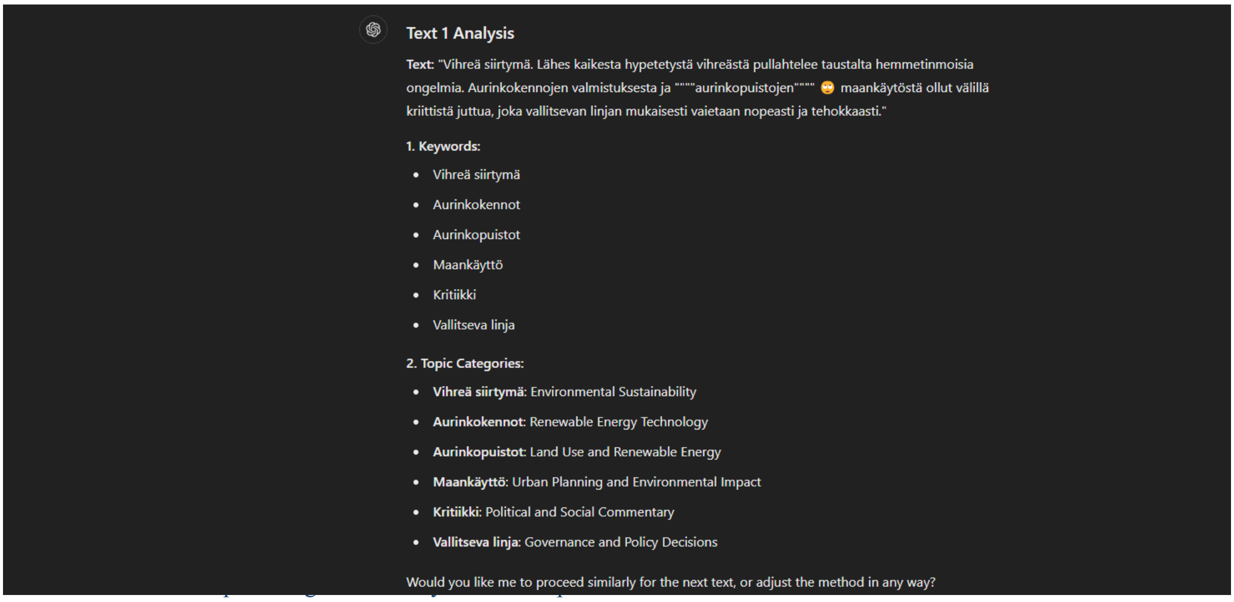
- Laadukas koulutus
- Yhdenvertaisuus
- Kyberturvallisuus
- Verkossa tunnistaminen
- Opetus
- Koulutustarjonta

2. Topic Categories:

- **Laadukas koulutus:** Education Quality
- **Yhdenvertaisuus:** Equal Access to Education
- **Kyberturvallisuus:** Cybersecurity Education
- **Verkossa tunnistaminen:** Online Safety Skills
- **Opetus:** Teaching Strategies
- **Koulutustarjonta:** Educational Offerings

Let me know if I should continue to the final text.

Screenshot 5. Test B2 analysing a parliamentary speech that does not exist. In B2, text 39 would be a tweet.



Text 1 Analysis

Text: "Vihreä siirtymä. Lähes kaikesta hypetetystä vihreästä pullahtelee taustalta hemmetinmoisia ongelmia. Aurinkokennojen valmistuksesta ja """"aurinkopuistojen"""" 😳 maankäytöstä ollut välillä kriittistä juttua, joka vallitsevan linjan mukaisesti vaietaan nopeasti ja tehokkaasti."

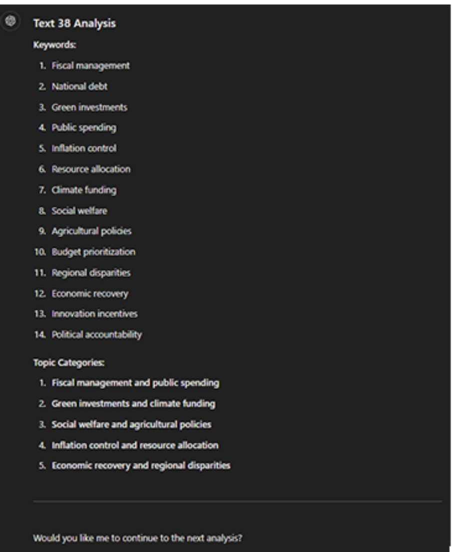
1. Keywords:

- Vihreä siirtymä
- Aurinkokennot
- Aurinkopuistot
- Maankäyttö
- Kritiikki
- Vallitseva linja

2. Topic Categories:

- **Vihreä siirtymä:** Environmental Sustainability
- **Aurinkokennot:** Renewable Energy Technology
- **Aurinkopuistot:** Land Use and Renewable Energy
- **Maankäyttö:** Urban Planning and Environmental Impact
- **Kritiikki:** Political and Social Commentary
- **Vallitseva linja:** Governance and Policy Decisions

Would you like me to proceed similarly for the next text, or adjust the method in any way?



Text 38 Analysis

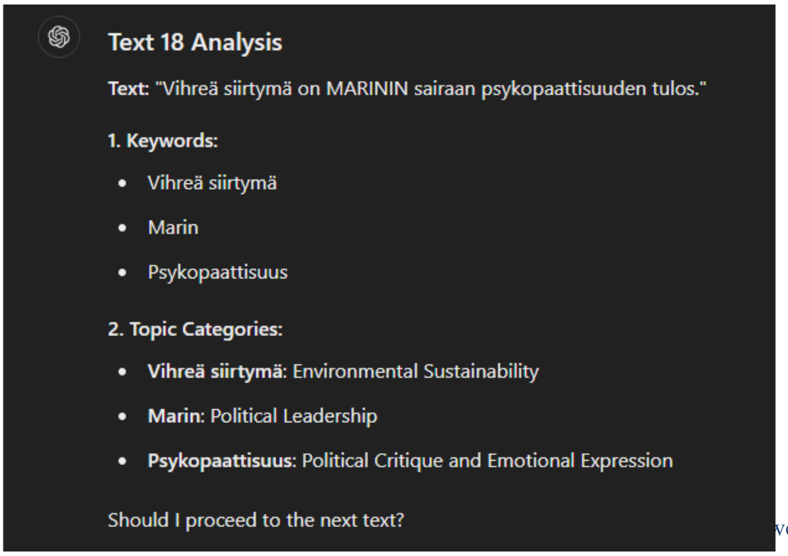
Keywords:

1. Fiscal management
2. National debt
3. Green investments
4. Public spending
5. Inflation control
6. Resource allocation
7. Climate funding
8. Social welfare
9. Agricultural policies
10. Budget prioritization
11. Regional disparities
12. Economic recovery
13. Innovation incentives
14. Political accountability

Topic Categories:

1. **Fiscal management and public spending**
2. **Green investments and climate funding**
3. **Social welfare and agricultural policies**
4. **Inflation control and resource allocation**
5. **Economic recovery and regional disparities**

Would you like me to continue to the next analysis?



Text 18 Analysis

Text: "Vihreä siirtymä on MARININ sairaan psykopaattisuuden tulos."

1. Keywords:

- Vihreä siirtymä
- Marin
- Psykopaattisuus

2. Topic Categories:

- **Vihreä siirtymä:** Environmental Sustainability
- **Marin:** Political Leadership
- **Psykopaattisuus:** Political Critique and Emotional Expression

Should I proceed to the next text?

weet correctly

analysed on the left in test B3.

ISSN 2416-5999, an open-access journal