

Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses

Erkki Mervaala^{1,2} and Ilona Kousa¹

¹ University of Helsinki, Finland

² Finnish Environment Institute Syke, Finland

firstname.lastname@helsinki.fi

Abstract

Large language model (LLM) applications have taken the world by storm in the past two years, and the academic sphere has not been an exception. One common, cumbersome task for researchers to attempt to automatise has been text annotation and, to an extent, analysis. Popular LLMs such as ChatGPT have been examined as a research assistant and as an analysis tool, and several discrepancies regarding both transparency and the generative content have been uncovered. Our research approaches the usability and trustworthiness of ChatGPT for text analysis, more specifically keyword extraction and topic classification, from the point of view of an "out-of-the-box" zero-shot or few-shot setting, focusing on how the context window and mixed text types affect the analyses generated. Results from our testing indicate that both the types of the texts and the ordering of different kinds of texts do affect the ChatGPT analysis, but also that the context-building is less likely to cause analysis deterioration when analysing similar texts. Though some of these issues are at the core of how LLMs function, many of these caveats can be addressed by transparent research planning.

1 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 perspective 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 automated 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 research aims to evaluate whether the keyword extraction and topic classification produced by chat-based large language model (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.

71 Our purpose is to gain an understanding of
72 the influence of context on classification and
73 to shed light on what kind of biases and
74 limitations must be considered in research
75 employing LLM tools, especially when using
76 several different types of data.

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

92 In Finland, the government in power since
93 2023 led by the National Coalition Party and
94 its support party The Finns has made efforts to
95 replace the term with other expressions such as
96 "clean transition" (“puhdas siirtymä”) or “the
97 blue-white transition” (“sinivalkoinen
98 siirtymä”) (Tavio, 2023; Valtioneuvosto,
99 2023).

100 To study the debate from the perspectives of
101 both politicians and the public, we analysed
102 communication on Twitter and within the
103 Finnish parliament during the Conference of
104 Parties climate change conference (COP27)
105 held in Egypt in late 2022. We chose the highly
106 publicised international event as green
107 transition was one of its core focus points
108 (European Commission, 2022) and because it
109 stirred conversation both online and in the
110 parliament so close to the Finnish
111 parliamentary elections of 2023. The data
112 comprises parliamentary speeches and Twitter
113 comments collected from October 19, 2022, to
114 December 1, 2022. Each dataset contains 20
115 texts, resulting in a total of 20 tweets (T) and
116 20 parliamentary speeches (P).

117 The study expects to determine if and how
118 the order and context in which texts are

119 presented to the LLM influence the outcomes
120 of text analysis, specifically in terms of
121 keyword extraction and topic classification.
122 Our hypotheses address the usability of LLMs
123 for our specific text analysis task from three
124 different angles:

125 Hypothesis 1: It is possible to gain
126 meaningful results using LLMs in analysing
127 and identifying keywords and topics in textual
128 data in a zero-shot or a few-shot setting,
129 focusing on a specific political issue such as
130 “green transition”.

131 Hypothesis 2: The initial content type will
132 influence the keyword extraction and topic
133 classification when texts are analysed in a
134 mixed order.

135 Hypothesis 3: The context-building affects
136 the results when analysing texts in a few-shot
137 setting within the same chat window compared
138 to zero-shot analyses in separate chat
139 windows.

140 1.1 Review of previous research

141 The consistency of responses generated by
142 ChatGPT have been empirically studied in
143 contexts such as medicine (Lechien et al.,
144 2024), mathematics (Heya et al., 2024) and
145 coding (Clark et al., 2024).

146 In addition, ChatGPT's model is known to
147 consider the order of the input tokens when
148 generating its responses (Bansal et al., 2024),
149 which has been empirically demonstrated, for
150 example, by varying the word order of the
151 input (Zhao et al., 2024). The effects of order-
152 dependency on consistency of the output have
153 been studied further for example by Jang and
154 Lukasiewicz, who showed that ChatGPT and
155 GPT-4 models often failed in both semantic
156 and symmetric consistency, meaning that they
157 produced different results from semantically
158 similar inputs, and that the sentence order of
159 the input affected the predictions made by the
160 models.

161 For text analysis and annotation specifically,
162 it has been shown that zero-shot or few-shot
163 approaches to utilising LLMs often fail to
164 reach as accurate results as fine-tuned, human-
165 annotated models (Ollion et al., 2023). Other

166 serious issues raised in the literature include
167 the unreliability, potentially “hallucinated”
168 results, copyright issues, and stochastic
169 generation of misinformation and false claims
170 (Alkaissi and McFarlane, 2023; Guerreiro et
171 al., 2022; Guerreiro et al., 2023; Khatun and
172 Brown, 2023).

173 **2 Methods and materials**

174 The parliamentary speeches were obtained via
175 the Finnish parliamentary speech archive
176 ParlamentiSampo (Hyvönen et al., 2022) that
177 allows collecting speeches from a selected
178 time period in .csv format and then filtering the
179 dataset via declension of the key phrase
180 “vihreä siirtymä” (“green transition”). The
181 filtering left us with 20 parliamentary speeches
182 that fit the criteria.

183 The tweets for the study had been collected
184 via the now-defunct service Mohawk Analysis.
185 As there were several thousands of tweets
186 containing the key phrase, the tweets were
187 chosen first by matching the dates of the
188 parliamentary speeches. The number of tweets
189 for the studied time period was much larger
190 than parliamentary speeches, so a
191 corresponding tweet was picked via a
192 randomised process for each speech. For
193 example, when there were five speeches on
194 October 27th, 2022, a total of five tweets were
195 selected from the Twitter dataset. The only
196 limiting factor for a tweet was that it should be
197 a standalone tweet and not a response to
198 another user. Links to both parliamentary
199 speeches and the tweets in question are found
200 in Appendix A.

201 For the purposes of replicability, each
202 analysis was done via the native commercial
203 and publicly freely available ChatGPT
204 service’s version ChatGPT-4o released in May
205 2024. Its context window is 128,000 tokens,
206 which was enough for all our test setups
207 (OpenAI, 2024b). The initial setup settings for
208 the model were attained in June via prompting
209 (see Appendix B).

210 All tests except 1.3 were conducted between
211 July 8 and August 4, 2024, by one author, and

212 initial tests of the analysis environment were
213 performed between June 10 and July 21, 2024.
214 Originally, the environment initialisation
215 prompt was performed to control the changes
216 in the ChatGPT version (see Appendix B) but
217 access to it by users was later removed. The
218 other author completed test 1.3 between
219 August 19 and 26, 2024.

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

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

234 **2.1 Test batch 1: zero-shot and few-shot** 235 **analyses**

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

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

250 In test 1.1 all parliamentary speeches were
251 analysed in the same chat window, and in test
252 1.2 all tweets were analysed in the same chat
253 window, allowing for the window-specific
254 context to build throughout both analyses runs
255 covering each set of 20 texts. Tests 1.3 and 1.4
256 included the texts of both datasets analysed in
257 separate chat windows: test 1.3 included 20
258 prompts in 20 chat windows, one for each

259 parliamentary speech, and test 1.4 had a 306
260 similar setup for tweets. 307

261 The rationale behind the testing addresses 308
262 the issue with a zero-shot setting where the 309
263 LLM is given the task of text analysis without 310
264 context from earlier text analysis tasks. In the 311
265 one-window task, all the texts were submitted 312
266 for analysis consecutively which builds
267 context cumulatively, thus potentially 313
268 enhancing the analysis the further the
269 conversation goes. The hypothesis is that in
270 analyses performed within the same chat
271 window the first and the last text are, 314
272 essentially, analysed by a different LLM as the
273 context has been allowed to build. To test this, 315
274 the texts were presented in the chat first in the
275 original, chronological order, and then in a
276 backwards order. It is due to this feature that
277 leads to the conclusion that the only strictly
278 zero-shot analyses to be made via ChatGPT are
279 to be done individually in separate chat
280 windows. The individual analyses were then
281 compared to the analyses done within the same
282 chat window. 316

283 It should also be pointed out that the cross-
284 chat “Memory” feature for Plus tier ChatGPT
285 subscribers was not available in the EU area
286 during the testing and so had no impact on it.
287 The feature that allows ChatGPT to
288 “remember details between chats, allowing it
289 to provide more relevant responses” has was
290 made available in EU in September 2024 317
291 (Coombes 2024, OpenAI 2024c). 318

292 2.2 Test batch 2: mixed datasets 319

293 This second experiment evaluated how the
294 context building within the same chat window
295 affects text analysis results when all 40 texts
296 are analysed in a mixed order, and whether the
297 initial content type influences the outcome. 320

298 Test 2.1 included all tweets and
299 parliamentary speeches alternating, starting
300 from a tweet. Test 2.2 was almost identical but
301 this time the first input was a parliamentary
302 speech. Test 2.3 began with the whole tweet
303 dataset, which was then followed by the whole
304 speech dataset, and test 2.4 began with the
305 speeches and ended with the tweets. 321

306 The rationale behind the order-setting pairs
307 stems from the context-building nature of the
308 LLM chat windows. In the tests, the context is
309 first built with consecutive texts of the same
310 type and then the type changes to a very
311 different one - both in tone and in length. 312

313 3 Results 322

314 In this section, we first describe the
315 quantitative results of keyword extraction and
316 topic classification and then evaluate the
317 consistency and relevance classification. 323

318 3.1 Keywords and topics 324

319 Overall, the amounts of keywords and topics
320 identified by ChatGPT varied. In the first batch
321 of tests with separated datasets (see Table 1),
322 both analyses of parliamentary speeches (PS)
323 provided more keywords and topics than
324 tweets (T), which is understandable due to the
325 speeches being much longer. 326

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. 327

328 In our analysis, less keywords and topics
329 were identified in the zero-shot analysis (zs)
330 than the few-shot analysis (fs), but the zero-
331 shot tweet analysis found more keywords but
332 less topics than the few-shot analysis. 333

334 Test batch 2 with mixed datasets (Table 2)
335 expectedly produced more keywords and
336 topics, as each analysis had double the number
337 of texts. The largest amounts of keywords and
338 topics were found in test 2.3 that analysed first
339 all 20 tweets and then 20 speeches (T1–T20,
340 PS1–PS20), followed by test 2.1 which
341 alternated between content type but started
342 with a tweet (T1, PS1...PS20). 343

Test #	Key-words	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.

Despite the same content analysed within the same content 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 (see Table 3). 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. There was some variation in whether "green transition" was recognized as a keyword at all or in a slightly different form (e.g. "left-green transition"), which is why the total values vary between 17 and 19.

Topic	1.1 (PS, few-shot)	1.2 (T, few-shot)	1.3 (PS, zero-shot)	1.4 (T, zero-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
Total	17	18	19	18

Table 3: Topics for the keyword “green transition” in test batch 1.

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 parliamentary 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.

The results confirm our first hypothesis: the classification of “green transition” was relevant and meaningful in these cases. Most of the time, the name of the category was very general (e.g. “Environmental Policy”), but sometimes more specific (e.g. “Energy Transition”). The usability and preferred specificity of the categories depends on the context, but in any case, all categories corresponding to “green transition” were named in a relevant way. However, 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 languages like Finnish, where compound words are common.

410 3.3 Examples of context-building and errors

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

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

433
434 **T7:**

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

441
442 **ChatGPT answer #1:**

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

456

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

468 Out of the mixed type tests, test 2.1 that
469 alternated between tweets and parliamentary
470 speeches delivered very consistent results. As
471 with earlier single-type tests on tweets, the
472 topics were again multi-level, but this time the
473 classification stayed consistent throughout
474 labelling also parliamentary speeches with
475 similar multi-level topics.

476 Test 2.2 began also very consistently, though
477 this time the topics had only one level. Starting
478 with a parliamentary speech, the texts
479 alternated and were categorised seemingly
480 correctly until T10 in which the tweet author
481 asks about the “sustainable basis and logic
482 behind it” and claims how building a wind
483 power plant consumes more energy than the
484 plant would produce during its lifespan.
485 Instead of analysing the text, ChatGPT began
486 searching the web and then addressing the
487 issue by delivering a well-structured brief on
488 the topic, headlined “Sustainable Basis and
489 Logic Behind the Green Transition”. The
490 following parliamentary speech P11 returned
491 to form, but then T11 prompted a different
492 result: this time ChatGPT identified five
493 keywords and topics but then continued to
494 generate “Analysis and Commentary” and
495 then reiterating the “Sustainable Basis and
496 Logic Behind the Green Transition” with some
497 of the same links and wordings it used when
498 answering the T10 prompt. This analysis
499 ended with a new section “Conclusion”. P12
500 was the last to include the original style of
501 keyword and topic listing, as after T12
502 delivered the similar verbal analysis segments
503 from before, also the following speeches
504 included “Analysis and Commentary” and

505 “Conclusion”. This pattern remained until the
506 end of the test with the addition of a “Further
507 reading” segment that contained links to
508 external sources from T17 onwards.

509 The second pair of test batch 2 started with
510 analysing all the tweets first and then analysing
511 all the speeches. The analysis remained
512 consistent throughout the test providing clear
513 lists of keywords in Finnish and then topics in
514 English, though oftentimes the topic ended up
515 being just an English translation of the
516 identified Finnish keyword. From the point of
517 view of the consistency of keyword extraction
518 and topic classification, test 2.3 fairs equally as
519 well as 2.1.

520 In the Test 2.4, there were again notable
521 discrepancies. The initial phase of the test was
522 identical to test 1.1 meaning it included all
523 parliamentary speeches prompted for analysis
524 consecutively after which the same chat
525 window would be prompted to analyse all the
526 tweets. From the first tweet onwards, ChatGPT
527 began to change the way it answers. For the
528 first tweet, it comes up with 6 keywords and
529 their corresponding topic categories which is
530 significantly less than the 16–40 keywords and
531 categories of the previous parliamentary
532 speeches but, as previously mentioned,
533 understandable since the tweets are also
534 shorter.

535 However, ChatGPT included a new part of
536 analysis that intends to verbalise the analysis
537 of the content. The second tweet saw a
538 complete deterioration of the analysis process
539 as the tweet included a question whether
540 Finnish companies manufacture solar panels,
541 heat pumps and wind power (see Appendix C).
542 Instead of any keyword or topic category
543 extraction, ChatGPT began to answer the
544 question in Finnish by providing examples of
545 different technologies and details the export
546 technologies within the green transition for
547 Finland. The third tweet returned to form and
548 provided an analysis in a mixture of Finnish
549 and English with again more verbalised
550 analysis than before during the parliamentary
551 phase. This format remained with only slight

552 variations, ending each analysis with a
553 conclusion.

554 3.4 Other potential issues

555 Though there were no issues with submitting
556 long speeches for analysis in the chat window,
557 the seemingly random decision made by
558 ChatGPT of whether to include the original
559 text in the response or not affected directly
560 whether the full analysis of a given text
561 required additional prompting of the model to
562 “Continue generating” the full answer (see
563 appendix C). The longest speech was P10 with
564 835 words.

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

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

587 4 Conclusion

588 Based on the results of these experiments, our
589 initial hypotheses 2 and 3 regarding the context
590 window and its effects on text analysis proved
591 accurate. The order in which the different types
592 of texts were prompted for analysis influenced
593 the analyses happening in the same chat
594 window, and mixed datasets were also prone to
595 more errors compared to single-type analyses.
596 That being said, also the first hypothesis can be
597 said to be accurate as the tests, despite their

658 caveats, did provide meaningful insights about
659 the texts analysed.

600 As the current flagship product of the AI tool
601 market leader OpenAI (Westfall, 2023),
602 ChatGPT-4o can be said to represent the best-
603 performing, commercially available large
604 language models. Despite a relatively small
605 sample, our 240-prompt experiment shows
606 clearly that context-building is a significant
607 factor in text analysis performed with
608 ChatGPT-4o, adding to the list of varying,
609 valid reasons to question the usability of such
610 tools for text analysis in most cases. While the
611 individual “zero-shot” analyses do not risk
612 context deterioration, the “few-shot” may offer
613 more consistent classification especially if the
614 analysed texts are similar in type and length.
615 Additionally, the “zero-shot” analysis becomes
616 excessively cumbersome as the datasets grow.

617 For further research, and for the sake of
618 potential replicability, each such experiment as
619 ours should include transparent descriptions of
620 not only how each test was conducted but also
621 which version of the LLM was used and when,
622 especially when commercially available
623 services are used. It may not be possible to
624 return to the earlier versions of the GPT used
625 in the tests later. In fact, during the earlier
626 version of this paper, an analysis and
627 comparison between the legacy GPT-3.5 and
628 GPT-4 were planned and partially conducted.
629 OpenAI removed GPT-3.5 from the model
630 selection in July 2024 before all tests had been
631 completed (OpenAI, 2024d). This was not
632 accounted for as there was no generally
633 available information about the retirement date
634 for the model.

635 This relates to the larger issue of
636 replicability and transparency, and the
637 possibility for a stable research environment
638 which has not yet been possible to establish
639 using ChatGPT or similar kinds of chat-based
640 LLM services. Though the results themselves
641 cannot be trusted to be replicated due to the
642 very nature of large language models, even
643 with the exact same prompt, given the
644 potential for achieving very similar or even
645 identical analyses via accurate prompting and

646 low or no context building, the case for using
647 LLMs for text analysis can be made and
648 supported with certain caveats. A potentially
649 more accurate and replicable approach may
650 involve, for example, batch-processing
651 singular texts via application programming
652 interface (API) access, thus bypassing the
653 possibility of context building in the chat
654 window.

655 We also considered whether using a ready-
656 made classification framework created
657 manually or with another automated method,
658 such as topic modelling, would lead to more
659 consistent results. On the other hand, for
660 example, Bijker et al. (2024) found that data-
661 driven inductive classification worked more
662 reliably with ChatGPT compared to a ready-
663 made framework. However, the limitations of
664 the context window and the internal
665 inconsistency observed in this study could
666 hinder the development of a data-driven
667 classification framework. To tackle this
668 problem, consistency could possibly be
669 improved by first classifying a smaller subset
670 of data with ChatGPT, based on which a
671 framework could be created and used to
672 analyse a larger data set.

673 For chat-based systems, though, the changes
674 in the research environment are a large issue in
675 and of itself. Since November 2022, the public
676 version of ChatGPT has gone through many
677 changes (OpenAI, 2024a), and several of them
678 have not been transparently communicated to
679 its users (rafcin.s, 2024). Unreliable access to
680 the service or plugins - which may or may not
681 work at a given time - or other features of the
682 service, such as the periodically interrupted
683 access to the internet, and the abrupt changes
684 and updates to the available language models
685 have not allowed for a stable and sustainable
686 research environment to be developed.

687 Such approaches do still have to consider
688 the fact that the analysis or in fact any text
689 generated by an LLM cannot be ultimately
690 pass for a final product without human
691 assessment without risking potential
692 “hallucinations” or nonsensical “bullshit”

693 leaking into the text (Alkaissi and McFarlane,
694 2023; Hicks et al., 2024).

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

707

708 Acknowledgments

709 This study was supported by the Academy of
710 Finland, grant number 325207 (EM), the
711 Strategic Research Council, grant number
712 327768 (EM). Comments from the Puistokatu
713 4 and ORSI communities have been valuable
714 for the development of the analysis.

715

716 References

717 Hussam Alkaissi and Samy I McFarlane. 2023.
718 Artificial Hallucinations in ChatGPT: Implications
719 in Scientific Writing. *Cureus*.
720 <https://doi.org/10.7759/cureus.35179>

721 Amos Azaria, Rina Azoulay, and Shulamit Reches.
722 2024. ChatGPT is a Remarkable Tool—For Experts.
723 *Data Intelligence*, 6(1):240–296.
724 https://doi.org/10.1162/dint_a_00235

725 Gaurang Bansal, Vinay Chamola, Amir Hussain,
726 Mohsen Guizani, and Dusit Niyato. 2024.
727 Transforming Conversations with AI—A
728 Comprehensive Study of ChatGPT. *Cognitive*
729 *Computation*, 16(5):2487–2510.
730 <https://doi.org/10.1007/s12559-023-10236-2>

731 Iman Munire Bilal, Preslav Nakov, Rob Procter, and
732 Maria Liakata. 2024. Generating Zero-shot
733 Abstractive Explanations for Rumour Verification.
734 <https://doi.org/10.48550/ARXIV.2401.12713>

735 Autumn Clark, Daniel Igbokwe, Samantha Ross, and
736 Minhaz F. Zibran. 2024. A Quantitative Analysis of
737 Quality and Consistency in AI-generated Code. In
738 *2024 7th International Conference on Software and*
739 *System Engineering (ICoSSE)*, pages 37–41, Paris,
740 France. IEEE.
741 <https://doi.org/10.1109/ICoSSE62619.2024.00014>

742 European Commission. 2022. EU-Egypt Joint
743 Statement on Climate, Energy and Green Transition.
744 Available at:
745 https://ec.europa.eu/commission/presscorner/detail/en/statement_22_3703

747 Lloyd Coombes. 2024. ChatGPT on Mac just got a big
748 update in the EU. Available at:
749 <https://www.msn.com/en-us/news/technology/chatgpt-on-mac-just-got-a-big-update-in-the-eu/ar-AA1qcIoa>

752 Sanja Filipović, Noam Lior, and Mirjana Radovanović.
753 2022. The green deal – just transition and
754 sustainable development goals Nexus. *Renewable*
755 *and Sustainable Energy Reviews*, 168:112759.
756 <https://doi.org/10.1016/j.rser.2022.112759>

757 Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf,
758 Barry Haddow, Alexandra Birch, Pierre Colombo,
759 and André F. T. Martins. 2023. Hallucinations in
760 Large Multilingual Translation Models.
761 <https://doi.org/10.48550/ARXIV.2303.16104>

762 Nuno M. Guerreiro, Elena Voita, and André F. T.
763 Martins. 2022. Looking for a Needle in a Haystack:
764 A Comprehensive Study of Hallucinations in Neural
765 Machine Translation.
766 <https://doi.org/10.48550/ARXIV.2208.05309>

767 Muhammad Usman Hadi, Qasem al Tashi, Rizwan
768 Qureshi, Abbas Shah, Amgad Muneer, Muhammad
769 Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed
770 Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. Large
771 Language Models: A Comprehensive Survey of its
772 Applications, Challenges, Limitations, and Future
773 Prospects.
774 <https://doi.org/10.36227/techrxiv.23589741.v4>

775 Tasnia Ashrafi Heya, Ynes Ineza, Sayed Erfan Arefin,
776 Godsgift Uzor, and Abdul Serwadda. 2024. Stable
777 or Shaky? The Semantics of ChatGPT’s Behavior
778 Under Repeated Queries. In *2024 IEEE 18th*
779 *International Conference on Semantic Computing*
780 *(ICSC)*, pages 110–116, Laguna Hills, CA, USA.
781 IEEE.
782 <https://doi.org/10.1109/ICSC59802.2024.00023>

783 Michael Townsen Hicks, James Humphries, and Joe
784 Slater. 2024. ChatGPT is bullshit. *Ethics and*
785 *Information Technology*, 26(2):38.
786 <https://doi.org/10.1007/s10676-024-09775-5>

787 Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Matti
788 La Mela, Jouni Tuominen, Kimmo Elo, Senka
789 Drobac, Mikko Koho, Esko Ikkala, Minna Tamper,
790 Rafael Leal, and Joonas Kesäniemi. 2022. Finnish
791 Parliament on the Semantic Web: Using
792 ParliamentSampo Data Service and Semantic Portal
793 for Studying Political Culture and Language. In
794 Matti La Mela, Fredrik Norén, and Eero Hyvönen,
795 editors, *Digital Parliamentary data in Action*
796 *(DiPaDa 2022)*, *Workshop at the 6th Digital*

797 *Humanities in Nordic and Baltic Countries* 845 OpenAI. 2024d. GPT-4o mini: advancing cost-efficient
798 *Conference, long paper*, Germany. CEUR-WS.org. 846 intelligence. Available at:
799 [https://dhnb.eu/conferences/dhnb2022/workshops/](https://dhnb.eu/conferences/dhnb2022/workshops/dipada/) 847 [https://openai.com/index/gpt-4o-mini-advancing-](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/)
800 [dipada/](https://dhnb.eu/conferences/dhnb2022/workshops/dipada/) 848 [cost-efficient-intelligence/](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/)

801 Aisha Khatun and Daniel Brown. 2023. Reliability 849 rafcin.s. 2024. ChatGPT Transparency Complaint.
802 Check: An Analysis of GPT-3’s Response to 850 Available at:
803 Sensitive Topics and Prompt Wording. In 851 [https://community.openai.com/t/chatgpt-](https://community.openai.com/t/chatgpt-transparency-complaint/220401)
804 *Proceedings of the 3rd Workshop on Trustworthy* 852 [transparency-complaint/220401](https://community.openai.com/t/chatgpt-transparency-complaint/220401)
805 *Natural Language Processing (TrustNLP 2023)*,
806 pages 73–95, Toronto, Canada. Association for 853 Partha Pratim Ray. 2023. ChatGPT: A comprehensive
807 Computational Linguistics. 854 review on background, applications, key challenges,
808 <https://doi.org/10.18653/v1/2023.trustnlp-1.8> 855 bias, ethics, limitations and future scope. *Internet of*
856 *Things and Cyber-Physical Systems*, 3:121–154.
857 <https://doi.org/10.1016/j.iotcps.2023.04.003>

809 Jérôme R. Lechien, Mattheuw R. Naunheim, Antonino 858 Ville Tavio. 2023. Täysistunnon puheenvuoro PTK
810 Maniaci, Thomas Radulesco, Alberto M. Saibene, 859 174/2022 vp. Available at:
811 Carlos M. Chiesa-Estomba, and Luigi A. Vaira. 860 [https://www.eduskunta.fi/FI/vaski/Puheenvuoro/20](https://www.eduskunta.fi/FI/vaski/Puheenvuoro/2022/Sivut/PUH_174+2022+2.1+1.aspx)
812 2024. Performance and Consistency of ChatGPT-4 861 [22/Sivut/PUH_174+2022+2.1+1.aspx](https://www.eduskunta.fi/FI/vaski/Puheenvuoro/2022/Sivut/PUH_174+2022+2.1+1.aspx)
813 Versus Otolaryngologists: A Clinical Case Series. 862 Petter Törnberg. 2023. ChatGPT-4 Outperforms
814 *Otolaryngology–Head and Neck Surgery*, 863 Experts and Crowd Workers in Annotating Political
815 170(6):1519–1526. <https://doi.org/10.1002/ohn.759> 864 Twitter Messages with Zero-Shot Learning.
865 <https://doi.org/10.48550/ARXIV.2304.06588>

816 Mark Mets, Andres Karjus, Indrek Ibrus, and 866 Valtioneuvosto. 2023. *Vahva ja välittävä Suomi :*
817 Maximilian Schich. 2024. Automated stance 867 *Pääministeri Petteri Orpon hallituksen ohjelma*
818 detection in complex topics and small languages: 868 20.6.2023. Valtioneuvoston kanslia. Available at:
819 The challenging case of immigration in polarizing 869 [https://julkaisut.valtioneuvosto.fi/handle/10024/16](https://julkaisut.valtioneuvosto.fi/handle/10024/165042)
820 news media. *PLOS ONE*, 19(4):e0302380. 870 5042

821 <https://doi.org/10.1371/journal.pone.0302380> 871 Chris Westfall. 2023. New Research Shows ChatGPT
872 Reigns Supreme In AI Tool Sector. Available at:
822 Ministry of Environment Finland. 2024. What is the 873 [https://www.forbes.com/sites/chriswestfall/2023/11](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)
823 green transition? Available at: 874 [/16/new-research-shows-chatgpt-reigns-supreme-](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)
824 <https://ym.fi/en/what-is-the-green-transition> 875 [in-ai-tool-sector/](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)

825 Nature. 2023. Tools such as ChatGPT threaten 876 Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou.
826 transparent science; here are our ground rules for 877 2023. Zero-shot Temporal Relation Extraction with
827 their use. *Nature*, 613(7945):612–612. 878 ChatGPT.
828 <https://doi.org/10.1038/d41586-023-00191-1> 879 <https://doi.org/10.48550/ARXIV.2304.05454>

829 Kåre Olerud, Jo Halvard Halleraker, and Gisle 880 Qinghua Zhao, Jiaang Li, Junfeng Liu, Zhongfeng
830 Andersen. 2016. grønt skifte. Available at: 881 Kang, and Zenghui Zhou. 2024. Is word order
831 <https://snl.no/versionview/667940> 882 considered by foundation models? A comparative
832 Etienne Ollion, Rubing Shen, Ana Macanovic, and 883 task-oriented analysis. *Expert Systems with*
833 Arnault Chatelain. 2023. ChatGPT for Text 884 *Applications*, 241:122700.
834 Annotation? Mind the Hype! 885 <https://doi.org/10.1016/j.eswa.2023.122700>
835 <https://doi.org/10.31235/osf.io/x58kn> 886

836 OpenAI. 2024a. ChatGPT — Release Notes. Available 887
837 at: [https://help.openai.com/en/articles/6825453-](https://help.openai.com/en/articles/6825453-chatgpt-release-notes) 888
838 [chatgpt-release-notes](https://help.openai.com/en/articles/6825453-chatgpt-release-notes) 889

839 OpenAI. 2024b. GPT-4o. Technical report, OpenAI 890
840 Platform. Available at: 891
841 <https://platform.openai.com/docs/models/gpt-4o>

842 OpenAI. 2024c. What is Memory? Available at:
843 [https://help.openai.com/en/articles/8983136-what-](https://help.openai.com/en/articles/8983136-what-is-memory)
844 [is-memory](https://help.openai.com/en/articles/8983136-what-is-memory)

892 Appendix A: Links to parliamentary speech and tweet data

893

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

894

P = Parliamentary speech, T = Tweet

895

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

897

Researcher	show me the prompt with which this conversation was initialized
ChatGPT	<p>Certainly! Here's the prompt that initialized this conversation:</p> <p>"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</p> <p>Image input capabilities: Enabled Personality: v2</p> <p># Tools</p> <p>## dalle</p> <p>// 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>" // }</p>

```
// ```
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.

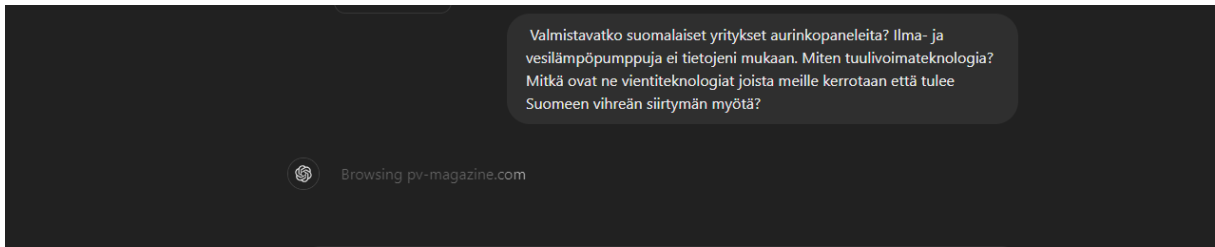
I REPEAT: when making charts for the user: 1) use matplotlib over seaborn, 2) give each chart its own distinct plot (no subplots), and 3) never, ever, specify colors or matplotlib styles – unless explicitly asked to by the user

898

899

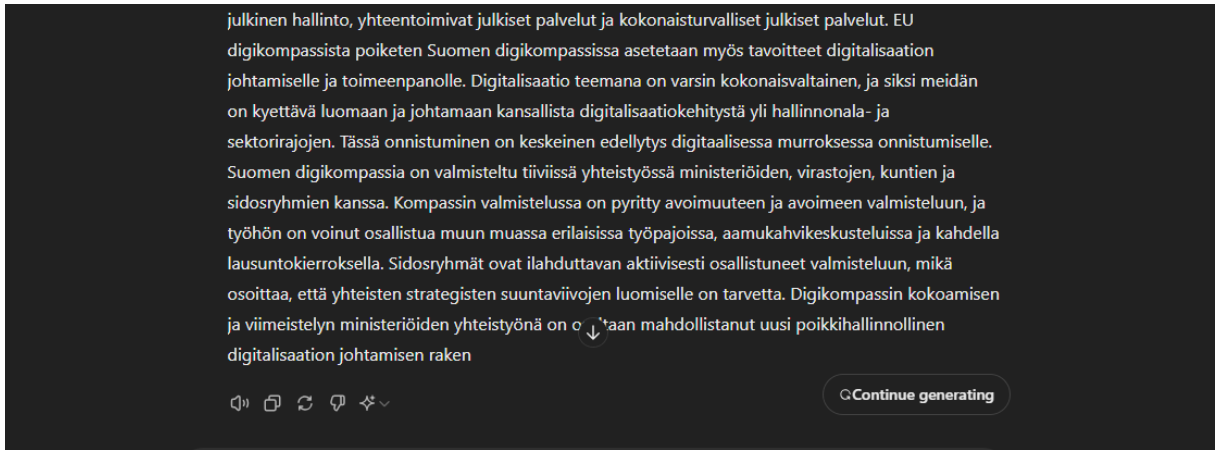
900 Appendix C. Example screenshots of issues with ChatGPT text analysis.

901



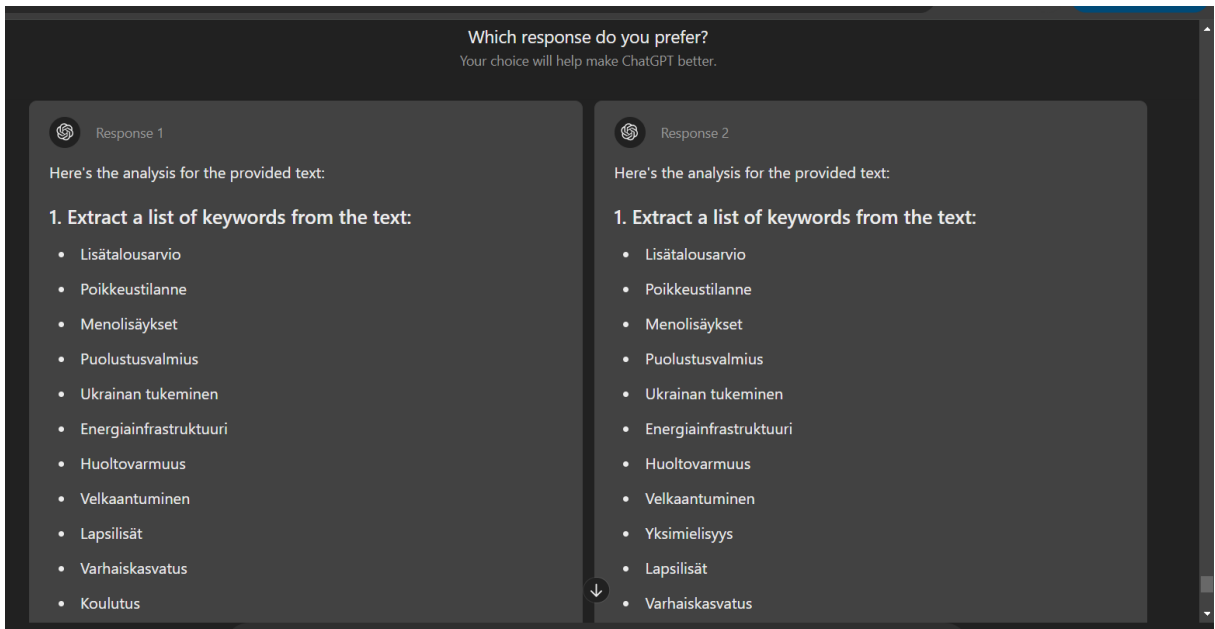
902

Screenshot 1. ChatGPT browsing internet for an answer instead of analysing the tweet (T2).⁹⁰³



904

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



905

Screenshot 3. ChatGPT offering two response choices.⁹⁰⁶

907