

Climate-Friendly Online Shopping Within the Green eCommerce Project: A Simple Syntactic Prompt Design Heuristic for ChatGPT; Using Single-Language-Models as Multi-Language Models

Alexander K. Seewald^A, Thomas Winter^D, Mario Platzer^B, Thomas Wernbacher^C, Alexander Pfeiffer^C

^A*Seewald Solutions GmbH, Lärchenstraße 1, 4616 Weißkirchen a.d. Traun, Austria
alex@seewald.at*

^B*yVerkehrsplanung GmbH, Brockmanngasse 55, 8010 Graz, Austria
mario.platzer@yverkehrplanung.at*

^C*Universität für Weiterbildung Krems, Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Austria
{thomas.wernbacher, alexander.pfeiffer}@donau-uni.ac.at*

^D*Liberacerta e.U., Am Lindenhof 37/11, 8043 Graz, Austria
thomas@liberacerta.com*

November 3, 2024

1 INTRODUCTION

In recent years, online shopping has increased rapidly. Recently, this trend was also being fuelled by the Covid 19 pandemic and, according to many experts, will continue unabated in the future. As a result, e-commerce in the B2C sector recorded record figures in 2021 in terms of turnover (9.6 billion euros, Knabl et al. (2021)) and in terms of postal parcels delivered. However, this flood of parcels goes hand in hand with the many negative consequences of a rapidly growing volume of goods transport on the last mile, which manifest themselves in traffic jams, noise pollution, air pollution and a decreasing quality of stay in public spaces.

Many study authors come to the conclusion that online shopping only has a better ecological balance than shopping in stationary retail stores under optimal framework conditions (e.g. promotion of collective orders, climate-friendly means of transport and avoidance of return shipments and same-day deliveries). However, the package delivery situation is currently characterized by frequent multiple deliveries, climate-damaging and underutilised means of transport and, above all, high return rates, which amount to up to 47% in the clothing sector (Knabl et al. (2021)). On the demand side, the situation is aggravated by the fact that end consumers in online shops are often offered no or only very limited climate-friendly delivery options, which contradicts the increasing sustainability awareness (Holtmann and Klitzsch (2021)) of many end customers.

This is where the preventive and customer-oriented approach of the *Green eCommerce* research project (Wernbacher et al., 2023) comes in. Within a continuation of the previous project *Think!First* (Wernbacher et al. (2019); Seewald et al. (2019)), contextually tailored add-ons based on behavioural, technology-based and logistical interventions for the existing online shops of the participating partners are designed, developed and tested in practice. With the help of a unique combination of a gamified loyalty system that rewards users for high compliance, persuasive design principles that are characterized by visually highlighting regional products with short delivery routes or collective orders, as well as AI-supported fitting tools and chat bots that automatically measure clothing sizes and point out and explain environmentally friendly delivery options, customers are encouraged to shop more consciously – in the sense of a traffic shift, traffic avoidance and traffic optimisation.

Through the active participation of the practical partners *Julius Meinl am Graben*, *kaufregional* and *ZERUM*, the innovative add-ons can be tested comprehensively and practically for different objectives, target groups and different product groups in real operations over several months. In addition, the integration of the innovative logistics service *Green to Home* from logistics partner *ERIVE* makes it possible to analyze the entire process between online shop operators, online end consumers, and package delivery service providers. Thus, this holistic approach generates new and in-depth insights into the acceptance, suitability and impact of innovative interventions in online shops.

Here, we primarily describe work done within this project on prompt design for personalized chatbots within the respective webshops of our partners. We intended the chatbot to enlighten and teach the customers on *green* topics such as environmentally-friendly delivery options and the need to reduce returns. The goal was to improve user acceptance of the reduced convenience often necessary for such options (e.g. longer delivery times, more time spent on choosing right sized garments)

Originally we had planned to train a text embedding model that chooses among a fixed set of static answers. However, when ChatGPT 4 came out (OpenAI (2022)), we deemed it far more useful to achieve the same task in a conversation-like setting. So we considered whether ChatGPT could be adapted to our task. In order to do so, all authors extensively tested its replies when asked questions about important *green* topics such as greenwashing, sustainable development, environmentally-friendly delivery options, and CO_2 reduction – which we would have had to add ourselves with the original approach – and found its performance in this area to be excellent. However, answers about our project and our partners were mostly wrong and additional information needed to be provided to the chatbot beforehand.¹ At the time of conducting this work, the only feasible way to adapt this chatbot was the design of suitable prompts.

During prompt design we made some observations on how to present reference question/answer pairs to ensure the best answer quality, which we will describe and interpret here. More specifically, we proposed a new prompt design method, *many-shot-prompting*, and also found a simple syntactic heuristic which predicts answer quality. This part of the work mainly concerns partners *Meinl am Graben (MG)*, *kaufregional (KR)*, and *ZERUM*

¹For *Meinl am Graben* question answering already worked moderately well without additional information. This is probably due to its long-time existence which made sufficient historical information available in the original training set for the chatbot.

(ZR). Evaluation of the chatbot’s effects is ongoing and no definitive results were available at the time of writing this paper.

We note that the question answering methodology behind state-of-the-art systems such as ChatGPT 4 is very flexible and may also be applied to decision support systems. In that case our simple syntactic heuristic could give valuable hints as to the appropriate amount of meta-information that should be provided.

Secondly, we propose and evaluate a simple method to use single-language large language models (LLMs) with any other language they were not trained for, using existing pretrained translation systems, and find that this process works remarkably well on our small corpus of German questions. Such an approach could also be useful for minority languages which often are not supported by LLMs.

Finally, we describe some limits observed during ChatGPT testing, speculate on the reasons for these limits, and estimate both power consumption and CO_2 -equivalent emissions for its training and inference.

This study was conducted between December 2022 and November 2023. During review in October 2024, we added additional information in Sec. 5.1 (last paragraph).

2 RELATED RESEARCH

Contrary to classical machine-learning based NLP processing systems which implicitly learn through their training data, when applying Large-Language-Models such as ChatGPT to a task, it must be explicitly told what to do (*initial* (1) and *pre-question meta-information* (2), see Sec. 3.1), followed by one (*one-shot prompting*) or several (*many-shot prompting*, proposed here in Sec. 3.1) solved examples of the NLP task itself. In some cases the initial meta-information is sufficient and no example needs to be given. This technique is called *zero-shot prompting*. In other cases the examples are not sufficient and we also need to prompt the intermediate steps to arrive at the solution. This is called *manual-chain-of-thought prompting*.

Surprisingly, our work here finds an optimal value on the length ratio between meta-information and task examples (see Sec.3) for best answer quality, and therefore can be applied to all above techniques (except for *zero-shot prompting* where no task examples are given). We are not aware of any other work that relates such relatively simple syntactic measures to answer quality.

Here, we have primarily focussed on selecting papers concerned with evaluating the performance of ChatGPT at classical NLP tasks – including question answering (QA) which is most similar to our task of creating personalized chatbots – given appropriate prompts.

Panda and Kaur (2023) compared ChatGPT (OpenAI (2022)) to traditional chatbot systems in library and information centers. They described parts of the training process, noted advantages and limitations of ChatGPT and noted that it is worth exploring as an alternative. However, their work was purely theoretical in nature.

Yang et al. (2023) evaluated the performance of ChatGPT on aspect and query-based text summarization tasks. They described a simple prompt pattern to force ChatGPT to make such task-specific summarizations from text, and showed that this works remarkably well with similar ROUGE scores as state-of-the-art methods. They noted that input length restrictions are challenging and may be addressed by methods to retrieve significant sentences from lengthy inputs. They also noted that the generated summaries may include nonfactual or biased information. Lastly, they noted that they have not explored different prompts or multiple conversations to refine the summaries which could improve results. Our results concerning the ratio of meta-information to text-to-be-learned (see Sec.3) could be applied here to inform the approximate length of optimal prompts.

Zhong et al. (2023) compared ChatGPT with fine-tuned BERT and manually designed task-specific prompts. They also asked ChatGPT to provide its own prompts, which is an interesting idea we may consider in future work. They found that when comparing ChatGPT to BERT-base, it performed poorly on paraphrase and similarity tasks but surpassed all BERT-style models on natural language inference tasks, indicating its superiority in reasoning ability. It was also comparable to BERT-base on SST-2, CoLA and QA-related tasks such as QNLI. They gave samples of wrong outputs and found a correlation between less similar sentences and worse performance of ChatGPT in the paraphrase settings. They evaluated different prompting techniques and found that all of them improve ChatGPT performance. Our method to provide reference questions and answers may be described as an extended variant of *few-shot prompting* (Brown et al. (2020)) where a large set of questions and answers are provided and could thus be called *many-shot prompting*. With the help of *manual-chain-of-thought* which

gives the intermediate non-specific steps to be taken to arrive at the answer, ChatGPT exhibited state-of-the-art performance among all comparison models on some tasks, e.g. CoLA, SST-2 and RTE. However, this is not something we can easily do in our setting which is very diverse w.r.t. questions. They also found that fine-tuned RoBERTa-large still beat ChatGPT on some tasks, such as the paraphrase task (MRPC), by a clear margin.

Kocoń et al. (2023) automated ChatGPT evaluation – similar to what we did, only on a larger scale using multiple (free) accounts – and tested it on 25 diverse analytical NLP tasks. They showed that the average loss in quality of ChatGPT was about 25% for *zero-shot* (providing no task samples) and *few-shot* (providing a few task samples) evaluation, where a more difficult task resulted in a higher loss. They also noted some instability and non-determinism as the same runs a week apart had slightly different results. They found that prompt construction may affect the quality of the model’s performance, which we also found. Similarly to Zhong et al. (2023), they concluded that ChatGPT can solve most NLP problems well but loses to the best models currently available.

Concluding, ChatGPT seems to perform equally well than non-fine-tuned state-of-the-art systems at some tasks (e.g. QA, which is most similar to our task), better at tasks requiring inference or reasoning, and worse at paraphrase and similarity tasks. However, state-of-the-art systems fine-tuned to a specific task (e.g. paraphrase) still perform better by a clear margin. In effect ChatGPT can solve most NLP problems well but loses to the best models currently available while being far less efficient (see also Sec. 5.2 on efficiency). Prompt design remains an important part to determine performance. Here, our proposed heuristic (Sec. 3.2) helps to determine either an optimal set of task examples given a known prompt, or an optimal prompt length given a set of task examples.

Table 1: Dataset descriptions using short names for *Meinl am Graben (MG)*, *Zerum (ZR)*, *kaufregional (KR)*. There are two very similar variants of *KR* due to minor changes of website links and email addresses in the reference text and the removal of one duplicate question.

Data-set	#Q/A pairs	Q.length (chars)	A.length (chars)	Q+A length (chars)
<i>MG</i>	46	42.13 ± 14.64	175.43 ± 96.24	217.57 ± 98.27
<i>KR (1st)</i>	90	42.43 ± 14.19	192.07 ± 77.96	234.50 ± 79.81
<i>KR (2nd)</i>	89	42.64 ± 14.31	191.62 ± 78.01	234.26 ± 80.10
<i>ZR</i>	55	50.60 ± 20.15	325.51 ± 439.06	376.11 ± 441.61
<i>Comb</i>	170	46.96 ± 21.55	175.39 ± 164.72	222.66 ± 168.32

3 REFERENCE Q/A VS. META-INFORMATION

Within our project *Green eCommerce* we wished the chatbot to educate customers on important green topics as well as presenting the project and the various partners. Therefore we chose to use ChatGPT-3.5, which had at this point been available only for a short time, and later added additional experiments for ChatGPT-4 when it became available.

At the time of conducting this work, the only way to change the behaviour of ChatGPT was by manual instruction using prompts that are given *before* the customer can communicate with the chatbot. We are aware of other chatbots where fine-tuning or training with additional textual data is possible, and may look into this in the future.

To enable prompt design, we asked our partners to provide sets of questions with reference answers, either summarized from email conversations with customers, or taken from already available frequently asked questions, or from various other sources. The goal was to provide these reference question/answer pairs to the chatbot with additional meta-information as background knowledge to improve answer quality versus the unprompted version. As this can be seen as an extension of *few-shot prompting*, we have simply called it *many-shot prompting*.

3.1 EXPERIMENTAL SETUP

Since there was no official API for ChatGPT-3.5 when we started this work, we used code from the WhatsApp-GPT project by Daniel Gross.² We only used *server.py* and adapted it slightly to our needs. A minor challenge

²See <https://github.com/danielgross/whatsapp-gpt>

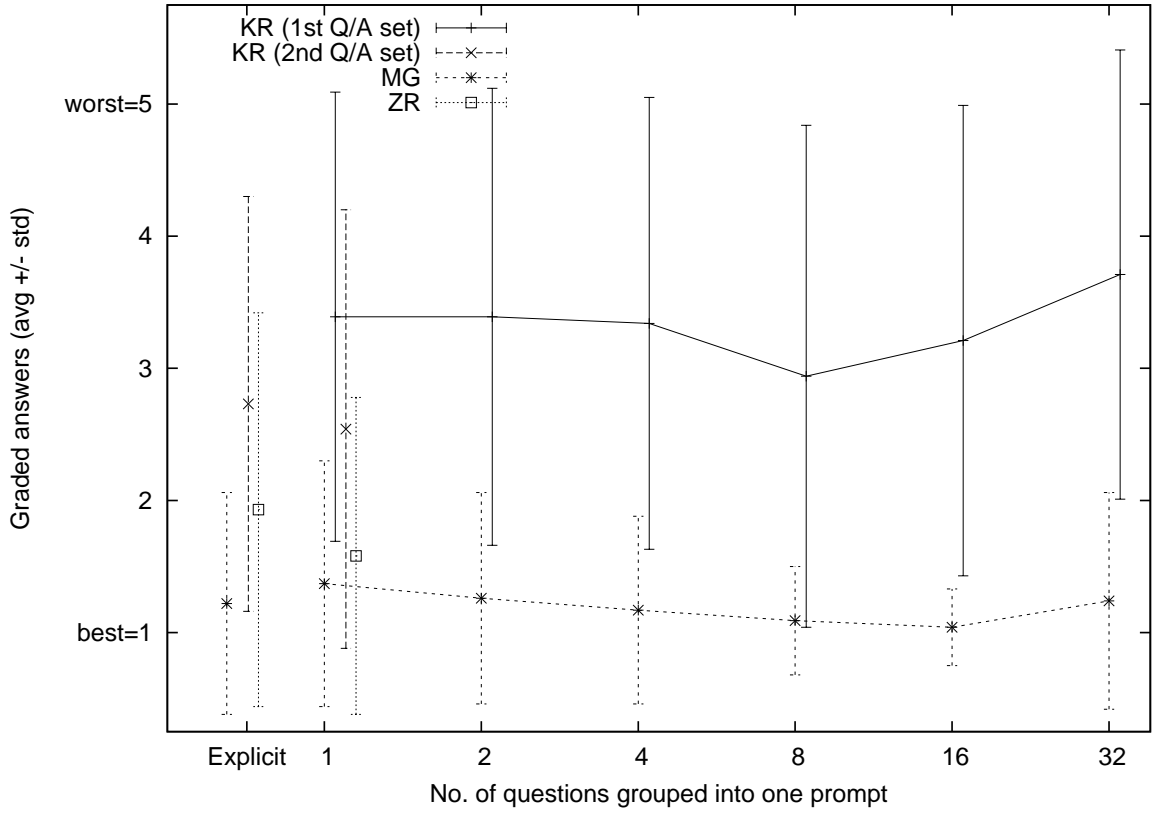


Figure 1: Grading of ChatGPT-3.5 answers when trained by prompts with different grouping values. Prompts were based on various reference question/answer sets from project partners. The separate *Explicit* setting explicitly marks questions, answers and meta-information in the prompts using the *role* parameter – more details see text. *Kaufregional* (KR) and *ZERUM* (ZR) were slightly offset for clarity, *Meinl am Graben* (MG) was kept at the correct position.

was to find a version of the Firefox browser which worked with OpenAI and Python playwright. Initially, OpenAI cut connections every 2h and then forced a new login plus solving a Captcha. After we upgraded to ChatGPT Plus, connections were held open for about 72h but there still was only a single conversational thread as we used only a single account.³ ChatGPT-4 was initially only available for 25 inputs every 3h, with maximum input length of about 16k.⁴ This limited the extent of our experiments, however it proved sufficient for our purposes at first.

Our partner *Meinl am Graben* (MG) provided 46 question/answer pairs, *kaufregional* (KR) provided 90 question/answer pairs and *Zerum* (ZR) provided 55 question/answer pairs. Statistics on these datasets can be found in Table 1 and sample question/answer pairs can be found in Table 4. We then designed a set of prompts as follows.

1. Initial meta-information: *You are a chatbot which answers questions about [partner name], relying on following information.*⁵

³Kocoń et al. (2023) used multiple accounts of the freely available version. However as we needed to use the paid ChatGPT Plus version both for accessing ChatGPT-4 as well as increased stability and reliability, this would have incurred additional costs of \$20 per month per account, which quickly becomes prohibitive when the number of accounts is increased.

⁴This was not a hard limit. It was possible to add information beyond this limit, however chatbot responses became idiosyncratic and erratic and answer quality suffered.

⁵Original: *Du bist ein Chatbot der Fragen über [Partnername] beantwortet, unter Berücksichtigung der noch folgenden Informationen.*

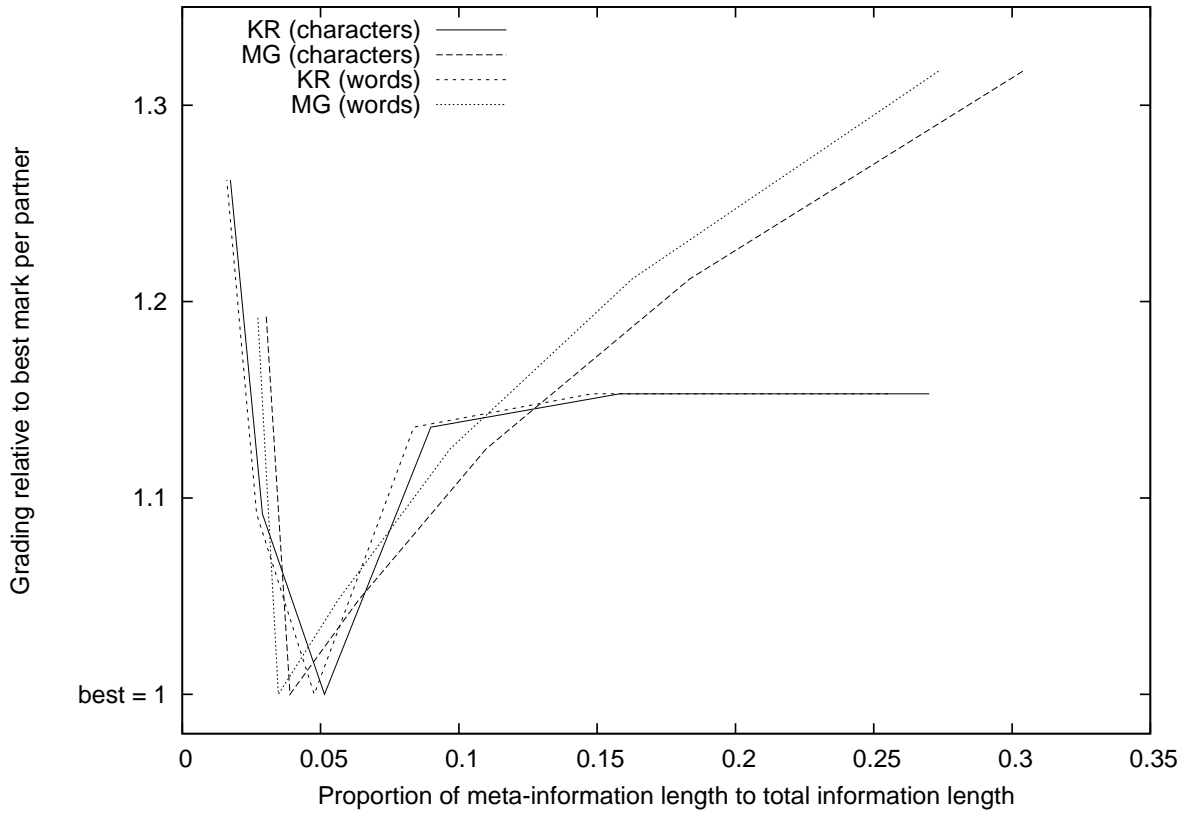


Figure 2: Grading of ChatGPT-3.5 answers (relative to best grade per partner) versus proportion of meta information length to total information length. Total information includes both reference question/answers and meta-information about how to use them (e.g. *Please do not answer questions but remember the following information.*) – more details see text. Information length was measured either by words or by characters.

2. Pre-question meta-information, i.e. *Please do not answer questions but remember the following information.*⁶ – followed by an empty line, and a set of 1, 2, 4, 8, 16, or 32 question/answer pairs in sequence with an empty line between them. We denote these numbers as *grouping values*.
3. Step 2. is repeated as often as necessary until all reference question/answer pairs were presented once.

We’ve referred to this process as *many-shot prompting*. All question/answer pairs and prompts were in German. Each prompt was presented to ChatGPT via the webbased interface in the order mentioned above. Afterwards all reference questions were asked again and the answers noted and manually graded by our partners with Austrian school marks (1 = best, 5 = worst). Finally, conversations were completely cleared⁷ before starting the next test.

Once it became available, we switched to the completion endpoint of the ChatGPT API. We then noted that – where before we had to rely on the chatbot to distinguish questions from answers – this API additionally allowed to define a role for each text part, thus explicitly defining which parts of the text refer to user question, assistant answer or general system instructions, by settings the tag *role* in the JSON input accordingly. We mapped the initial instruction (1 in enumeration above) to *system*, all questions to *user* and all answers to *assistant* (2,3 in enumeration above). This approach forces a grouping value of 1, i.e. providing each question and each answer separately as two messages. We tested this alternative as well, and named it *explicit* since in this case explicit

⁶Original: *Bitte keine Fragen beantworten, sondern merke dir einfach die folgenden Informationen.*

⁷First clearing the named conversation on the top left-hand-side, then using *Clear all conversation* in the context menu on the bottom left.

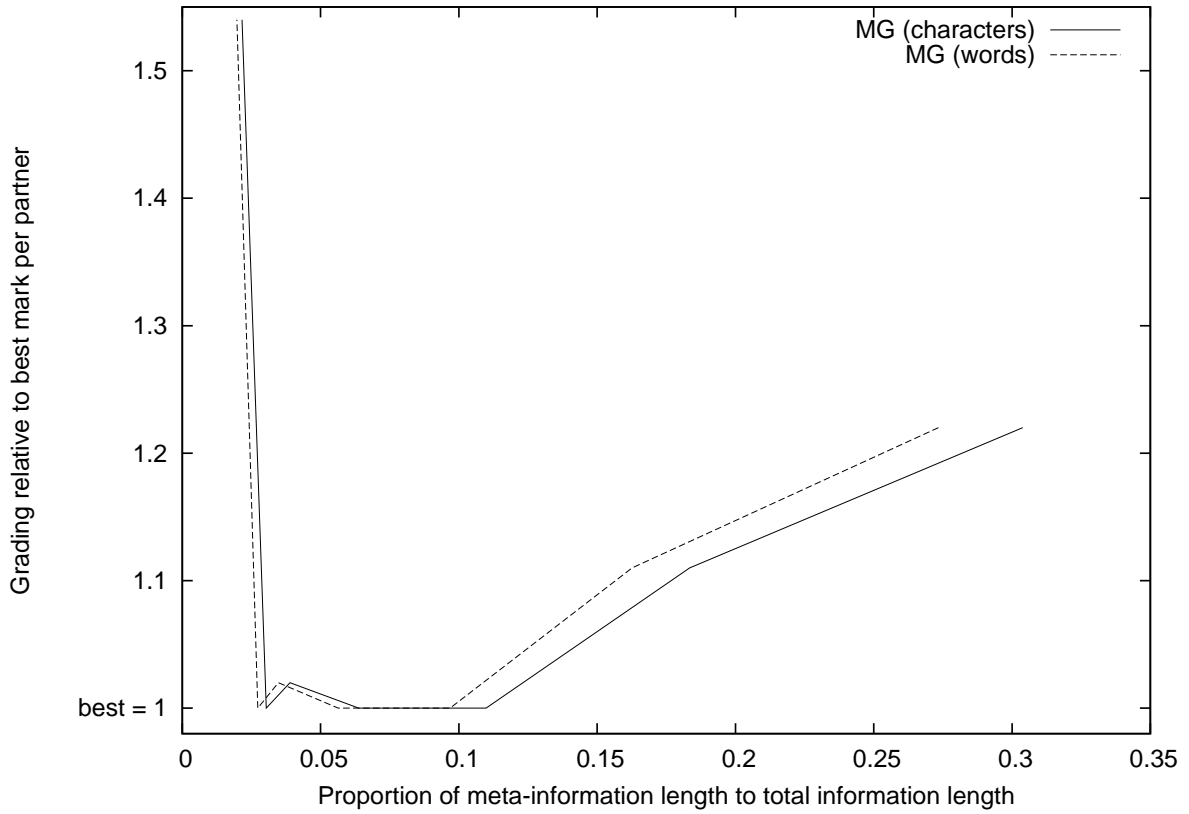


Figure 3: Grading of ChatGPT-4 answers (relative to best grade per partner) versus proportion of meta information length to total information length. Total information includes both reference question/answers and meta-information about how to use them (e.g. *Please do not answer questions but remember the following information.*) – more details see text. Information length was measured either by words or by characters.

information is given and no grouping takes place. It may be supposed that this explicit information yields better results than our ad-hoc grouping approach, however that is not the case as we shall see later.

3.2 RESULTS FOR CHATGPT-3.5

It turned out that the grouping of questions – i.e. how many questions were combined into one prompt – had a significant effect on the quality of the given answers. Fig. 1 shows the average and standard deviation of chatbot answer grading. It can be seen that the best marks are around the grouping values of 8 and 16 with marks to both sides being worse all the way to the maximum and minimum values.

To better analyze this pattern, we’ve first normalized each partner’s grades to the best average grade. Then we measured the proportion of meta-information length – both initial meta-information (sent once) and pre-question meta-information (sent before each set of reference question/answer pairs) – to the total sent prompt length which additionally includes the actual reference questions and reference answers.⁸

Fig. 2 shows the results. As you can see the best grades are obtained in a narrow range between about 3.6% and 5.4% depending on the partner and on whether we measure length with characters or words. An average value of 4.5% seems to work well. Note that the exponentially increasing question grouping value will due to its

⁸We also tested computing the proportion of meta-information length to length of actual reference questions and answers (since it only depends on the question set), but results were very similar and the best grades were at almost exactly the same position.

sparsity tend to diffuse the actual optimal value of this proportion so it may be possible to get a smaller range with additional experiments.

What was somewhat surprising is that providing explicit information (label *Explicit*) about text types via *role* performed universally worse. For *Meinl am Graben (MG)* we evaluated this setting with the original question set and plotted it as part of the original analysis; however *kaufregional (KR)* had by then switched to a new question/answer set (*2nd Q/A set*) where only a few settings were graded, and *ZERUM (ZR)* only provided evaluations for a subset of grouping values due to time constraints, so for these two partners we only have two data points each: grouping value *1* and the *Explicit* setting. These values are also shown in Fig. 1.

3.3 RESULTS FOR CHATGPT-4

Can we also observe the same patterns for ChatGPT-4? Although testing it was quite tricky⁹ and so we could only test *Meinl am Graben (MG)* here, Fig. 3 shows that the pattern is similar but the range of best proportions is less narrow. ChatGPT-4 seems to accept a larger range of proportions with the best possible grading of 1.0 ± 0.0 , however lumping all 46 questions into a single prompt (leftmost point) yields much worse answer quality performance. It could also be that the much higher answer quality of ChatGPT-4 prevented us from resolving smaller differences between proportions and that the actual pattern is much more similar to 3.5. So we look forward to repeating this experiment with a larger question set in future work.

We may still tentatively conclude that the intention-determining ability of ChatGPT-4 works better than that of ChatGPT-3.5 for a larger range of meta-information length, but worse for the minimum meta-information length – the latter probably because then insufficient meta-information is given to enable ChatGPT-3.5 to find out what needs to be done.

Concerning *Explicit*, the same pattern as in ChatGPT-3.5 is observed here: it performs worse than the best grouping value: *MG*: 1.07 ± 0.44 (*Explicit*) vs. 1.00 ± 0.00 (grouping value 8); *KR*: 2.00 ± 1.47 (*Explicit*) vs. 1.90 ± 1.37 (grouping value 1); *ZR* did not provide a ChatGPT-4.0 evaluation.

All the detailed numerical results concerning grading of the different question/answer sets for all grading values, including the *Explicit* setting, can be found in Table 2. Table 3 shows the character and word counts used earlier for evaluation of the proposed heuristic.

Table 2: Grading (1 = best, 5 = worst) for ChatGPT, averaged over all graded answers (mean \pm standard deviation)

Partner	V.	Explicit	Grouping					
			1	2	4	8	16	32
KR (1st)	3.5		3.39 ± 1.70	3.39 ± 1.73	3.34 ± 1.71	2.94 ± 1.90	3.21 ± 1.78	3.71 ± 1.70
KR (2nd)	3.5	2.73 ± 1.57	2.54 ± 1.66					
MG	3.5	1.22 ± 0.84	1.37 ± 0.93	1.26 ± 0.80	1.17 ± 0.71	1.09 ± 0.41	1.04 ± 0.29	1.24 ± 0.82
MG	4.0		1.22 ± 0.63	1.11 ± 0.53	1.00 ± 0.00	1.00 ± 0.00	1.02 ± 0.15	1.00 ± 0.00
ZR	3.5	1.93 ± 1.49	1.58 ± 1.20					

4 USING SINGLE-LANGUAGE MODELS AS MULTI-LANGUAGE MODELS

Initially we thought that ChatGPT would only understand English, so we built a backend to translate questions and answers back-and-forth between German and English. However, after noting that it is possible to pose questions in German as well, we were interested in determining whether commonly used translation systems have become sufficiently good to enable to use single-language large language models (LLM) also in other languages, essentially wrapping them in a translation system to enable their use in a multi-language setting. Here, we explicitly

⁹At the time of conducting this work, the limitation for ChatGPT-4 was 25 prompts every 3h, meaning that most experiments had to be run for quite long. Additionally every ChatGPT crash – of which there were several – necessitated restarting the experiment from the first question/answer pair.

Table 3: Character and word counts for all relevant Q/A sets and partners.

Partner	Type	Explicit	Grouping					
			1	2	4	8	16	32
KR (1st)	Meta (chars)	122	8041	4081	2145	1177	649	385
	Q/A (chars)	21728			21728			
	Meta (words)	15	1005	510	268	147	81	48
	Q/A (words)	2934			2934			
MG	Meta (chars)	124	4171	2147	1179	651	387	299
	Q/A (chars)	9556			9556			
	Meta (words)	17	523	270	149	83	50	39
	Q/A (words)	1390			1390			

defined the source language, however using automated language detection this approach could easily be adapted to transparently support multiple languages.

4.1 EXPERIMENTAL SETUP

We obtained a set of 170 questions in German language from all project partners, dataset *Comb(ined)*.¹⁰ Reference answers and initial prompts were not used since we were only interested in determining whether it is feasible to use purely English-language models in other languages. The reason to use a multi-lingual model for this purpose is to have a reasonable comparison of ground truth by directly inputting questions in the source language, thus enabling at least a semi-automated evaluation.

We sent each question separately to ChatGPT-3.5, noting each answer. Additionally, we translated each question using the free DeepL¹¹ translator to English, sent each translated question separately to ChatGPT-3.5, and translated the answer back to German again using the same translator. Afterwards, we compared the obtained two answers – one directly obtained, one obtained by back-and-forth translation – both qualitatively and quantitatively. For brevity, we define the *non-translated version* to be the one where the original German question was posed to ChatGPT-3.5 and where a German answer was obtained; and the *translated version* where ChatGPT received only the English translation of each question and accordingly responded in English.

4.2 QUALITATIVE RESULTS

A German native speaker analyzed all 170 answer pairs by hand and graded them into a five-valued ordinal scale based on the preferred version according to syntax, grammar and comprehensible formulation of each answer. Such comparisons are out of necessity more subjective but also more informative than quantitative comparisons.

- **-2:** Strongly prefer non-translated version (i.e. translated version has major grammatical errors, uses words inconsistently or has other issues which the non-translated version does not have)
- **-1:** Weakly prefer non-translated version (i.e. translated version has some minor style issues but both are grammatically and syntactically correct)
- **0:** No preference (i.e. both versions are practically identical or have roughly the same amount of minor issues such that a clear preference cannot be obtained)
- **+1:** Weakly prefer translated version (i.e. non-translated version has some minor style issues but both are grammatically and syntactically correct)
- **+2:** Strongly prefer translated version (i.e. non-translated version has major grammatical errors, uses words inconsistently or has other issues which the translated version does not have)

¹⁰These include most of the questions from the previous sections. We only removed duplicates – some questions were present in Q/A sets from multiple partners – and those questions where the unprompted answer was "I am just an AI chatbot...". See Table 1 for statistics on this dataset.

¹¹<https://deepl.com>, DeepL GmbH, Cologne, Germany

We originally intended to also compare the semantic textual content. However, the content of both answers from the same pair was always the same.

93.52% of answer pairs were graded without $\{0\}$ or with a weak preference $\{-1,+1\}$. Of these answer pairs, 46.54% – almost half – were graded with no preference. Of the 6.47% with strong preference, all 11 were graded as a preference for the translated version $\{+2\}$. No answer pair had a strong preference for the non-translated version.

The grader noted that the word and sentence part order of the non-translated answers were more similar to English (insofar possible without breaking grammar rules) which led sometimes to a cumbersome reading and that this was not observed for the translated version. He also noted that the words used by the translated version more often matched the usual translation that would be obtained in German. In very rare cases the translated version had one or two additional sentences versus the non-translated version. This was not observed for the non-translated version.

We may conclude that with respect to syntax, grammar, semantic content and style, there is very little difference between the non-translated and the translated version, so either version could be used.

4.3 QUANTATIVE RESULTS

We also compared the answers of each answer pair using Levenshtein edit distance Levenshtein et al. (1966) normalized to the length of the longer string, thus ensuring the difference is between 0 (identical) and 1 (maximally different).

On average, the 170 answer pairs had an average normalized edit distance of 0.2022 ± 0.0761 . If we exclude those likely distant pairs with *strong preference* $\{+2,-2\}$, the average edit distance reduces slightly to 0.1977 ± 0.0702 . When we only consider those 74 pairs graded as *no preference* $\{0\}$, the average edit distance is again smaller at 0.1708 ± 0.0533 . When we only consider those pairs graded as *strong preference* $\{-2,+2\}$, the edit distance is much higher at 0.2667 ± 0.1221 . These patterns indicate that the edit distance reflects the qualitative scale values.

We may conclude that a more simplistic edit distance comparison also supports the conclusion that the answer pairs are quite similar. Furthermore, the edit distance shows the expected patterns w.r.t. differently manually graded pairs, which supports the qualitative evaluation as well.

As overall conclusion we note that it clearly seems feasible to extend a single language Large-Language-Model (LLM) to multiple languages using such an approach, perhaps extended with automated language detection to enable transparent use of input languages similar to multi-language LLMs.

5 DISCUSSION

Here we will discuss some additional observations obtained during the course of our investigations, and give some estimations on environmental effects of the utilized ChatGPT-4 model.

5.1 OBSERVED LIMITATIONS

While at first glance ChatGPT seems quite capable, a more detailed analysis indicates that it needs constant supervision, guidance and correction, perhaps likened to an enthusiastic but scatter-brained intern. More specifically, we found that ChatGPT sometimes found it hard to perform calculations correctly such as those necessary to estimate delivery costs, or to estimated CO_2 footprint.

To determine the failure modes of ChatGPT's arithmetic ability, we tested several simple exercises from a fourth year elementary school math book. Most of them worked reasonably well, however there was one exercise which always introduced errors: a simple conversion of area values, followed by sorting them in ascending order. We observed three different kinds of errors of which always at least one – and sometimes two – appeared:

- **Conversion errors:** E.g. $1\text{ dm}^2 = 0.1\text{ m}^2$
- **Missing one value:** E.g. not putting 10 dm^2 into the sorted order at all – it's just left out without comment. This is probably the error that is most likely to be overlooked.

- **Wrongly sorted values:** Values were correctly converted but wrongly sorted (e.g. considering 200 cm^2 above resp. larger than 10 dm^2)

When we informed the chatbot of the wrong sort order the correct order was often obtained. In some cases it was necessary to give the correct order or converted value, or both. A translation into English did not help and we observed the same types of errors there. The same types of errors always appeared for ChatGPT-3.5 and for ChatGPT-4, in several tests with slightly different formulations of the questions. We also tried *manual-chain-of-thought* to explicitly force conversion, then sorting – however, sorting was still almost always wrong.¹²

We did most of the tests in German, but the same issue can easily be reproduced in English as well. For reference, this is a typical dialogue we obtained (taken from ChatGPT-3.5 but would not look much different with ChatGPT-4). **Bold** text shows the ChatGPT answers. Note that this time just noting which value is wrongly sorted does not help but gives exactly the same sort order as before, and it is necessary to explicitly give the correct conversion.¹³

Arrange these areas in order of size.

Start with the smallest.

1 m^2
 10 cm^2
 10 dm^2
 1 dm^2
 99 dm^2
 200 cm^2
 50 mm^2

Starting with the smallest area and moving to the largest, the order is:

50 mm^2
 10 cm^2
 1 dm^2
 10 dm^2
 200 cm^2
 99 dm^2
 1 m^2

200 cm^2 is sorted wrongly. Please repeat.

I apologize for the mistake.

Here is the corrected order:

50 mm^2
 10 cm^2
 1 dm^2
 10 dm^2
 200 cm^2
 99 dm^2
 1 m^2

200 cm^2 is 2 dm^2 and should be before 10 dm^2 .

¹²Amongst the dozens of prompt variants we tried, we had a single run with ChatGPT-4 with an explicit mention of converting and then sorting the converted values which was correct, however after 5 runs with various errors of the type mentioned above we gave up to reproduce it. Clearly, the effort to find a prompt to make ChatGPT output the right answer on the first try vastly exceeds the actual work solving this rather simple question.

¹³In several previous runs, it was usually sufficient to note the wrong value without giving the explicit conversion.

You are correct, I apologize for my mistake.
 Here is the corrected order:
 50 mm^2
 10 cm^2
 1 dm^2
 200 cm^2 (which is equivalent to 2 dm^2)
 10 dm^2
 99 dm^2
 1 m^2

It might be interesting to repeat this exercise with a larger set of mathematical exams with known solutions, using randomly generated numbers, and for several other chatbots to see whether this is a general issue with similarly trained systems. We speculate that it may be due to the two different tasks which must be performed here – conversion followed by sorting – and that the system may not be able to sequentially segment the tasks in the proper way but uses intermediate results too early (e.g. using the value before conversion for sorting, believing 10 cm^2 and 10 dm^2 are the same value so only outputting one of them, etc.), thus explaining these bad results. Should the example no longer work, try with other values – again preferably randomly-generated ones.

A newer model provided by OpenAI in October 2024, ChatGPT o1-preview, solves this problem correctly – albeit at about an order of magnitude higher computational effort. This version has been specifically trained to solve mathematical problems better using reinforcement learning on reference mathematical exercises including solutions.

5.2 ENERGY CONSUMPTION AND CO_2 EMISSIONS OF CHATGPT-4.0

A paper such as this would be somewhat incomplete without an estimate on the energy consumption and CO_2 eq emission of ChatGPT-4.0.

For training, we have an estimate of 30,000 petaflop/s-days¹⁴ as reported by ChatGPT-4.0 itself.¹⁵ On a Nvidia A100 with 312 TFlops/s peak performance (for FP16/32 Tensor, i.e. half precision) this translates to 2,3 million GPU hours. However, it is unlikely that peak TFlops can be sustained for such a long training due to thermal overload, synchronization overhead, and speed of data input delivery, so we will multiply this estimate by three to get 6,9 million GPU hours.¹⁶ It is generally assumed that ChatGPT-4 was trained in the Microsoft Azure Central US computing center in Iowa. In a blog post Walsh, Noelle (2022) a Power-Usage-Effectiveness (PUE) of 1,17 is quoted for the US Azure computing centers. So according to Lacoste et al. (2019), assuming Nvidia A100s with a TDP of 250W were used, correcting for PUE, the total emissions for a single training run are estimated to be 1,493,505kg (1,5t) CO_2 eq with a total power consumption of 2.02 GWh. We are probably safe in assuming that at least ten different training runs were needed to arrive at the final system, so we may increase the estimates by a factor of 10.

For testing, we have a rough estimate from Brown et al. (2020) where it was mentioned that ChatGPT-3 – which has 175 billion parameters – uses 0.4 kWh to generate 100 pages of text.¹⁷ If we scale this linearly to 285 billion parameters – this number again reported by ChatGPT-4 itself¹⁸ – we get 0.65 kWh for 100 pages

¹⁴One petaflop/s day is one day of one-petaflop operations, corresponding to 8.64×10^{19} floating point operations.

¹⁵It was obtained by one-shot prompting and may of course be spurious, but as far as we know no other reliable information is available at the time of writing this paper.

¹⁶Shital Shah @syteelus gives an upper bound of 10,000 A100 x 11 months which corresponds to 79,2 million GPU hours. However it was noted that ChatGPT-4 was available relatively fast after ChatGPT-3.5, and it also seems that external feedback was less both of which should significantly reduce the actual training time. We may however assume that ten training runs were needed to obtain the final systems in which case the estimates are quite similar.

¹⁷This is neither a precise nor a very reliable estimate but there was none available at the time of writing this paper.

¹⁸Again obtained by one-shot prompting.

of text, or $3.6 \cdot 10^{-6}$ kWh per character, or $8 \cdot 10^{-4}$ kWh per question pair¹⁹. Assuming 485g CO_2eq per kWh (Umweltbundesamt (2022)), this corresponds to 1.7 milligrams per generated or processed character or 0.39 grams CO_2eq per question. However, given that new versions of ChatGPT are provided every few months, and these incur an unknown number – perhaps more than ten – training runs, this value is only reliable if a sufficient number of people are using ChatGPT, and thus the cost of training is appropriately widely distributed and therefore essentially amortized. At the time of writing this paper this is clearly the case.

So, while training is somewhat costly an application of the chatbot for a single user is not.²⁰ Unfortunately, until finetuning becomes available in the API it will be necessary to send the complete prompt on each question, incurring an additional 10-20k characters, or additional 17-34 grams CO_2eq , per question.

6 CONCLUSION

We have introduced the Green eCommerce project, which follows a preventative and customer-oriented approach to encourage customers to shop more environmental-consciously.

We have proposed a new prompt design method, *many-shot prompting*, to enable using a set of reference question/answer pairs to improve chatbot answer quality. We have analyzed this method and found that a simple syntactic heuristic – the proportion of meta-information to total information (measured either by characters or by words) – predicts best answer quality within a narrow range with the optimal value around 4.5% on two sets of reference questions created within our project. This heuristic may also be useful to other prompting methods. Surprisingly, explicitly providing the chatbot with information on text types via *role* always performed worse than the best grouping value.

One practical application for this heuristic is to determine the appropriate amount of meta-information given a certain amount of information to be learned by the chatbot, or vice versa. It can also be used to optimize existing prompts to obtained the best answer quality by shortening or lengthening meta-information – or information – as necessary. From a theoretical viewpoint it is interesting that such a simple syntactic measure should have such a large effect on answer quality, and additional research will be needed to make such systems more robust w.r.t. the meta-information length.

We have also evaluated the feasibility of using single-language Large Language Models (LLMs) as multi-language LLMs using pretrained translation systems and found both on a qualitative and a quantitative level that the differences would be very small, thus indicating the feasibility of such an approach. It may well be that existing multi-language LLMs already take advantage of such translation systems during training – seeing as, for example, the totality of German texts is several magnitudes smaller than English texts – however it was still surprising that a direct application on short texts of just a few sentences works so well. Concerning our observation on word order it may be useful to inspect the deep learning translation model used and add appropriate levels to ChatGPT to enable it to also output a more intuitive word order for implicitly translated text.

Finally, we discussed limitations of ChatGPT as we observed them, gave a short example, and roughly estimated the energy and CO_2eq emission of ChatGPT-4 training and inference.

ACKNOWLEDGEMENTS

This project was funded by the Austrian Research Promotion Agency (FFG) and by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the Mobility for the Future (Mobilität der Zukunft (MdZ)) research program as a project for the 2021 call on M-Era as project *Green eCommerce*. We would especially like to thank Kristina P. and Lisa H. for many precise and consistent gradings of chatbot answers, often on short notice.

¹⁹We computed an average length of 223 characters from the reference questions/answers shown later and assumed processing input and output is equally complex from a computational standpoint, which seems reasonable.

²⁰For hundreds of millions or even billions of users it will of course be different.

References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Holtmann, U. and Klitzsch, C. (2021). *Talk about it - Nachhaltigkeit als PR-Thema*. EHI Retail Institute, Köln. <https://www.ehi.org/produkt/whitepaper-talk-about-it-nachhaltigkeit-als-pr-thema/>.
- Knabl, W., Köb, M., Meszaros, G., Prenger, C., Rischaneck, U., Segal, D., and Weigl, A. (2021). *Bundesweite eCommerce Studie Österreich*. Handelsverband. <https://www.handelsverband.at/publikationen/studien/e-commerce-studie-oesterreich/e-commerce-studie-oesterreich-2021>.
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniec, J., Gruza, M., Janz, A., Kancierz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, S., and Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.
- Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. <https://mlco2.github.io/impact/>.
- Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- OpenAI (2022). ChatGPT: Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt/>. (accessed 27th April 2023).
- Panda, S. and Kaur, N. (2023). Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers. *Library Hi Tech News*.
- Seewald, A. K., Wernbacher, T., Pfeiffer, A., Denk, N., Platzer, M., Berger, M., and Winter, T. (2019). Towards minimizing e-commerce returns for clothing. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, volume 2, pages 801–808.
- Umweltbundesamt (2022). CO₂-Emissionen pro Kilowattstunde Strom steigen 2021 wieder an. <https://www.umweltbundesamt.de/themen/co2-emissionen-pro-kilowattstunde-strom-steigen>. (accessed 28th April 2023).
- Walsh, Noelle (2022). How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability. <https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability>. (accessed 28th April 2023).
- Wernbacher, T., Platzer, M., Seewald, A. K., Winter, T., Wimmer, S., and Pfeiffer, A. (2023). Green eCommerce. In *IEEE 9th International Conference on Computing, Engineering und Design (ICCED 2023)*, pages 1–5. IEEE. In print.
- Wernbacher, T., Seewald, A., Denk, N., Pfeiffer, A., Platzer, M., and Winter, T. (2019). Think! first: Inducing behavioural change through gamification, persuasive design principles and machine learning. In *European Conference on Games Based Learning*, pages 793–XXIV. Academic Conferences International Limited.
- Yang, X., Li, Y., Zhang, X., Chen, H., and Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.
- Zhong, Q., Ding, L., Liu, J., Du, B., and Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

A APPENDIX

Table 4: Example questions and answers from the datasets used in this paper (three from each dataset, randomly selected), translated from the original German.

Data-set	Example questions (translated)	Example answers (translated)
MG	<p>What products does Meinel sell at the Graben?</p> <p>Can I rate products?</p> <p>Do you offer express delivery?</p>	<p>We offer regional and international specialties and delicacies. You can find a selection of our range in our online store</p> <p>This option is not currently available. But we are working on it.</p> <p>If you would like express delivery, please contact us at hauszustellung@meinlamgraben.eu or webshop@meinlamgraben.eu or +43 1 532 33 34 4000, stating your order number. We will find a customized solution for you.</p>
KR(1st)	<i>Due to the high similarity between KR (1st) and KR (2nd), examples are only shown for the latter</i>	
KR(2nd)	<p>Where are the products produced?</p> <p>How do I increase my visibility?</p> <p>Where can I find the ingredients of the food?</p>	<p>The companies must generate 50% of their value added in Austria in order to become partners in kauftregional.</p> <p>We advise you on the topics of search engine optimization, reach campaigns, performance marketing, online presence and other relevant topics relating to regional marketing. Contact us for more information and arrange a free consultation.</p> <p>You will find the "Properties" section under each product. Here you will find both the content characteristics and the ingredients.</p>
ZR	<p>Can I return clothes that don't fit well?</p> <p>How can I make sure I buy the right size underwear?</p> <p>Who is Zerum?</p>	<p>Clothing purchased online that is unworn, unwashed and undamaged can be returned to us within 4 weeks at your own expense and the purchase price will be refunded. Hygiene items such as underwear, earrings, socks and tights are excluded from exchange.</p> <p>Take a look at our size chart and compare your measurements with it. If you are not sure, please contact us at onlineshop@zerum.at</p> <p>ZERUM is a Graz-based fashion label that has been producing fair and ecological fashion since 2008. The man behind ZERUM is Sigmund Benzinger, who also founded the label. ZERUM not only has its own label, but also 5 concept stores in Austria where fair fashion from various designers is sold. ZERUM makes fair, timeless and ecological fashion with the greatest respect for the environment and all those involved in production. We want to give clothing back its value. We do not follow trends, but want to create companions for many years to come. Our products are fair - our production routes are short.</p>
Comb	<i>Since this dataset is essentially a combination of all other datasets, no samples are shown here</i>	