




Investigating the Configurability of LLMs for the Generation of Knowledge Work Datasets

Desiree Heim^{1,2}^a, Christian Jilek¹^b, Adrian Ulges³ and Andreas Dengel^{1,2}^c

¹Smart Data and Knowledge Services Department, German Research Center for Artificial Intelligence (DFKI), Germany

²Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Germany

³Department DCSM, RheinMain University of Applied Sciences, Germany

{desiree.heim, christian.jilek, andreas.dengel}@dfki.de, adrian.ulges@hs-rm.de

Keywords: Knowledge Work Dataset Generator, Large Language Model, Configurability.


Abstract: The evaluation of support tools designed for knowledge workers is challenging due to the lack of publicly available, extensive, and complete data collections. Existing data collections have inherent problems such as incompleteness due to privacy-preserving methods and lack of contextual information. Hence, generating datasets can represent a good alternative, in particular, Large Language Models (LLM) enable a simple possibility of generating textual artifacts. Just recently, we therefore proposed a knowledge work dataset generator, called KnoWoGen. So far, the adherence of generated knowledge work documents to parameters such as document type, involved persons, or topics has not been examined. However, this aspect is crucial to examine since generated documents should reflect given parameters properly as they could serve as highly relevant ground truth information for training or evaluation purposes. In this paper, we address this missing evaluation aspect by conducting respective user studies. These studies assess the documents' adherence to multiple parameters and specifically to a given domain parameter as an important, representative. We base our experiments on documents generated with KnoWoGen and use the Mistral-7B-Instruct model as LLM. We observe that in the given setting, the generated documents showed a high quality regarding the adherence to parameters in general and specifically to the parameter specifying the document's domain. Hence, 75% of the given ratings in the parameter-related experiments received the highest or second-highest quality score which is a promising outcome for the feasibility of generating high-qualitative knowledge work documents based on given configurations.


1 INTRODUCTION


Knowledge workers in the organizational context are professionals who primarily do mental, knowledge-centered work (Sordi et al., 2020). Typical actions a knowledge worker does are acquiring information, authoring documents, sharing information, and networking with others (Reinhardt et al., 2011). Here, they often work with non-public, i.e. either personal or organizational files.

Knowledge workers can be supported by a range of artificial intelligence-based methods, for instance, the knowledge worker's current task can be identified and relevant information can be proactively delivered (see, e.g. Tung et al. (2017)). While a range

of support tools exist, there is typically no suitable, comprehensive data collection to evaluate these tools against since knowledge work (KW) data collections have inherent problems. Since KW often involves private or corporate documents, they are subject to censoring to preserve confidential or private information, which leads to incomplete data. Moreover, gathering comprehensive contextually relevant information would require data owners to annotate their data thoroughly, which would require a high manual effort from the owners and distract them from their actual work. These issues of KW data collections are so severe that with Gonçalves (2011) even meta-literature emerged. He states that existing datasets might not be usable when evaluating Personal Information Management (PIM; Jones (2008)) tools due to the lack of important background information about the authors, their documents, and insights into their PIM process that would represent important ground-truth data.

^a <https://orcid.org/0000-0003-4486-3046>

^b <https://orcid.org/0000-0002-5926-1673>

^c <https://orcid.org/0000-0002-6100-8255>

Hence, synthetic data can be a good alternative to data collections. In particular, since with the emergence of Large Language Models (LLM; Zhao et al. (2023)) generating various documents is possible by prompting the LLM with suitable instructions. Motivated by the issues of data collections and the document generation abilities of LLMs, we recently proposed a KW dataset generator, called KnoWoGen¹ (Heim et al., 2024). It simulates multiple agents completing KW tasks and generates corresponding documents with an LLM. These documents are the core artifacts of the created datasets. Alongside the documents also contextual information about their creation and use as well as information about the general simulation setting is stored and provided in the final dataset. This includes parameters used in the document generation prompt, such as involved persons, topics, or the document's type. One main advantage of this approach, compared to data collections, is that no data protection measures are necessary as the data is artificially generated. Moreover, it comprises comprehensive contextual information that can be processed as inputs or ground truth data when evaluating support tools.

So far, LLM-based KW dataset generation has been only evaluated by assessing the authenticity, i.e., realistic appearance, of the generated documents to ensure the feasibility of employing an LLM for the document creation (Heim et al., 2024). However, it is also important to verify that descriptive parameters used in the document generation prompt are correctly reflected in the document since they could be used as ground truth. As our contribution in this paper, we fill this gap and examine the faithfulness of documents, generated with KnoWoGen, concerning their parameters. Hereby, we address two research questions:

- RQ 1. Is it possible to generate KW documents adhering to given parameters with LLMs?
- RQ 2. Is it possible to generate KW documents for various domains with LLMs?

Besides RQ 1 assessing the reflection of parameters, RQ 2 aims to investigate the ability of LLMs to generate domain-specific documents which is highly relevant as KW spans a wide range of diverse domains.

This paper is structured as follows: Section 2 introduces works related to our evaluations. Section 3 particularizes the generation of the documents included in the evaluation and the experiment setup. Next, Section 4 describes the experiments results which are subsequently discussed in Section 5. The paper concludes, in Section 6, with a summary of our findings and suggestions of future research directions.

¹For more information also consult the website of the KnoWoGen: <https://purl.archive.org/knowogen>

2 RELATED WORK

Content generated by Large Language Models (LLM) can be evaluated either automatically or by humans (Chang et al., 2024). While for some common tasks like Question Answering multiple benchmarks exist, human evaluations often provide more comprehensive insights. In the special use case of synthetic data generation, generated data can be either assessed directly or indirectly (Long et al., 2024). Direct evaluations of, for instance, class labels generated for a text, can be realized by employing benchmarks. However, if open-ended texts are generated either human evaluators or auxiliary models are required to assess the data's correctness. For indirect evaluations, the generated data is used to train or fine-tune a model on downstream tasks. The trained model is then evaluated on the downstream tasks based on existing benchmarks or based on humans' or auxiliary models' judgments if no standardized answers exist.

There are some similar approaches to the generation of knowledge work documents from given parameters or descriptions using LLMs, that also generate open-ended texts from given conditions for synthetic datasets. While some approaches only use a single parameter (Ye et al., 2022), others use multiple parameters (Xu et al., 2024; Yu et al., 2023) to increase the documents' variability compared to single-parameter-approaches. The aforementioned approaches evaluate the correct usage of main parameters in the synthesized texts mainly by using the generated data as training data to train or fine-tune models for specific downstream tasks, like text classification (Ye et al., 2022; Xu et al., 2024; Yu et al., 2023) or Named Entity Extraction (Xu et al., 2024), in which the given parameters act as the results. These trained models are then evaluated on the downstream tasks against model versions fine-tuned with similar other generated, human-labeled data or data from benchmark datasets or, alternatively, against other, sometimes non-fine-tuned, baseline models. Besides, Ye et al. (2022) additionally tested a parameter representing a label for a classification task using a classifier and compared the accuracy against other common benchmark datasets. Furthermore, they also conducted a human evaluation to verify the correct reflection of the label parameter as well as the relevance of the generated text to the application use case, e.g., synthesizing a movie review text. Moreover, the mentioned approaches assessed the lexical diversity of generated texts, e.g., by their vocabulary size, (Yu et al., 2023; Xu et al., 2024) and their naturalness which was evaluated by Ye et al. (2022) based on a human evaluation. In this paper, we explore the use

case of generating knowledge work documents. Since we seek to get detailed, direct, and reliable insights into parameter adherence and domain adaptability as a specific parameter, a direct human evaluation fits our demands best.

3 METHODOLOGY

This paper investigates two research questions. Namely, whether a Large Language Model (LLM) respects given parameters in generated documents, and whether it can generate domain-specific documents. We used the KnoWoGen generator to produce documents and conducted user studies to assess them. This section explains the role of parameters in KnoWoGen and the design of our experiments.

3.1 The Role of Parameters in the KnoWoGen

In the KnoWoGen (Heim et al., 2024), parameters are either given in the configuration when specifying concrete tasks or defined at simulation time, i.e., agents are assigned to tasks or parameters are randomly sampled. When documents are generated, parameters are used to select the matching prompt template and inserted into the template to create the final prompt sent to an LLM to generate the specific document.

Parameters are stored in the knowledge graph to retain background information about generated documents. These parameters can serve as training data or ground truth data for evaluations, making it crucial to verify their correct reflection in the documents.

Among the parameters defining the documents, the domain parameter is particularly influential as it affects the whole content of the document. Moreover, having a high domain adaptability is important because knowledge work in general encompasses a variety of domains and KnoWoGen aims to be a configurable, generally employable generator.

Also beyond the KnoWoGen, when synthesizing knowledge work documents and using the given description included in the prompt as ground truth, the verification of those two aspects is essential.

3.2 Evaluation of Documents Concerning Given Parameters

Large Language Model Selection. The KnoWoGen requires an instruction-fine-tuned Large Language Model (LLM). At the time of the experiment execution, Version 0.2 of the Mistral-

7B-Instruct model (Jiang et al., 2023) was showing decent qualities on several benchmarks², supported a relatively large context window with 32k tokens enabling the generation of long documents, and had a good ratio between resource consumption and quality, therefore we decided to use this LLM to generate documents with the KnoWoGen.

General Experiment Setup. For both research questions, we conducted user studies to get detailed and reliable insights. Especially when examining various domain parameter values, we think human experts are indispensable since they do not only have specific factual knowledge but can also well-assess which aspects are relevant in a domain, how technical terminology is correctly used, and how documents in the domain are typically composed on a content and structural level. Moreover, although, in several test runs, we did not see that a given parameter was not respected in a generated document, we wanted to confirm this perception since parameter values can invoke varying subjective expectations.

The user studies were organized as an online questionnaire. In the questionnaire, we asked the participants to rate the quality of generated documents regarding the above-mentioned aspects on a 5-point Likert scale (Likert, 1932). Here, high values always corresponded to high-quality documents concerning specific aspects to keep the studies consistent. Moreover, participants always had the option to comment on their given ratings to get insights into the participants' reasoning. However, we did not want to make comments mandatory to avoid making the experiment too effortful since we targeted a larger number of participants. Moreover, we wanted to let participants decide when they find it appropriate to comment on their decision to get significant comments.

Furthermore, we told the participants that all documents have been generated to prevent that they are distracted from the actual questions having assumptions that the content might be generated. Moreover, knowing this also corresponds to the actual end user case in which the users also know that and also assess the generated documents with respect to their reflection of the given configuration.

Experiment Questions. Since we have not used Llama2-13B-Chat (Touvron et al., 2023) as was the case in a previously conducted experiment (Heim et al., 2024), we additionally included a similar user

²We consulted the Huggingface Leaderboard for Open LLMs (Fourrier et al., 2024) which summarizes the performance of a range of LLMs on several common benchmarks

study that also assesses the documents' authenticity to have a comparative value enabling a better classification of the results. In contrast to the previous experiment, we assessed the documents' naturalness by differentiating it into social and linguistic or content-related naturalness to investigate them separately. Here, social naturalness refers to all authenticity aspects related to the social contact among involved or addressed persons in the e-mail or the meeting such as how they greet or address each other. Linguistic naturalness refers to the authenticity of the language used, i.e. how human-like it is, e.g., whether the word choice seems plausible. Moreover, content-related naturalness means how natural it is that specific content-related structures or concrete contents appear in the documents. In the experiments, we asked two questions related to naturalness. The first question addressed social naturalness and the second other aspects of naturalness, including in particular linguistic and content-related aspects.

To examine the adherence to given parameters, we compiled descriptions similar to the documents' descriptions in the prompts used to generate the documents including relevant parameters that were also directly used in the prompts. This included, for instance, document types, involved persons, and topics that should have been covered in the document.

In the third experiment examining the adherence of documents to given domains, we asked the participants three questions. Each question targeted a different abstraction level of the documents' domain adherence to get fine-grained insights. The lowest abstraction level was the word level. Here, participants were asked whether appropriate terminology was used and whether technical terms were used correctly. The question on the statement level asked the participants to rate whether statements made in the documents were factually correct. The document level targeted the question of whether the documents' structures were appropriate in the domain and whether the mentioned contents seemed plausible for the respective document type in the domain.

Generation of Documents. For examining the domain authenticity, we generated two documents similar to the earlier experiments (Heim et al., 2024) to increase the comparability between the results. Hence, we generated an e-mail and meeting minutes. The e-mail was a reply to an invitation to join a planning committee for a company's annual party. The meeting was an interview for an administrative specialist position.

Furthermore, we generated meeting notes and a project proposal to examine the reflection of param-

eters. The two document types were consciously chosen to be fundamentally different. While meeting notes reflect an interactive discussion among multiple people and are tendentially rather informal, project proposals are usually more formal and go in-depth. In the generated meeting notes, the menu for a company's annual social event was discussed. The project proposal suggested a language learning hub as an innovative idea to support learning English as a foreign language. For all generated documents, we explicitly selected topics that are generally understandable and do not require expert knowledge.

To examine the domain coherence, we generated document types that typically consist mainly of domain-specific content. Hence, we generated exams, paper drafts, project plans, course planning sessions, job interviews, discussions about recent methods, and potential further research directions. All seven document types were generated for each participant individually using a domain they were acquainted with.

All documents with the prompts used to generate them and experiment questions can be found in the supplementary material repository³.

Participants. In the experiments, two participant groups were involved. One group consisted of people who were acquainted with a specific domain. We assembled the group so that its participants covered diverse expert domains. Among the eight participants were two females and six males and their ages ranged between 18 and 44. All of them had an academic background and together covered the following domains: Architecture, Chemical Engineering, Cognitive Science, Computer Science, Data Science, Financial Mathematics, and Mechanical Engineering. Four participants stated that they had advanced knowledge about the domain, two rated their knowledge as being medium, and two stated that they had basic knowledge about the chosen domain. Moreover, all used LLM regularly and had good English skills (B1-C1). This participant group participated in all experiments. The seven documents for the domain coherence experiment were generated individually for each participant using one of their domains of expertise.

In addition to the participants from the first group, 41 further participants took part in the document authenticity and general parameter adherence experiments. Most of them were also recruited in an academic environment. For this group, there were no special requirements, and all participants got the same

³Supplementary material containing generated documents with their prompts, concretely asked experiment questions and raw results: <https://zenodo.org/records/13975025>

documents to judge. In total, 49 participants aged between 18 and 54 completed the experiments. 34 of the participants were males and 15 females. The majority were students, researchers, or software engineers. 43% had a background in Computer Science and 14% in Mathematics. The domains Psychology, Social Sciences, Education, and Engineering were also represented multiple times. Except for one case, all participants stated that their English language proficiency was B1 or higher. 65% used LLMs regularly, 27% occasionally, and the rest had not used LLMs before.

4 EVALUATION RESULTS

In this section, we present the results of the experiments introduced in the last section. The raw results are available in the supplementary material repository².

4.1 Experiment 1: Naturalness of Generated Documents

In the first experiment, participants had to rate the naturalness of two documents according to their social and linguistic naturalness on a 5-point Likert scale (Likert, 1932). The main purpose of this experiment was to have a comparative baseline that facilitates the classification of the two other experiments' results presented in this paper since the Large Language Model utilized can have a high impact on the quality of the generated documents.

Figure 1 shows the quality scores assigned to the generated documents regarding their social and linguistic naturalness. Since the scores for social and linguistic naturalness were almost equally distributed (Wasserstein distance of 0.024), we merged them into one distribution by taking the average of the count respectively for all scores. Overall, the documents' naturalness was rated highly and with a high agreement as around 80% gave one of the two highest scores.

In addition to the rating distribution of this experiment, the figure also depicts the assigned score distributions of generated and real documents from the preceding experiment (Heim et al., 2024). In direct comparison, the distribution of the naturalness score assigned in the present experiment resembles the distribution of real documents more than the one of generated documents. However, it must be noted that the evaluation settings of the two experiments differ, not least because participants in our experiments have been only given generated documents to judge and were made aware of this situation.

For 11 of the 98 given ratings, comments were provided. Most comments regarding linguistic naturalness addressed lexical recurrence. Hence, three comments mentioned that single words or word groups were repeated. In one case, multiple words of the same word family, culture, cultures, and cultural, appeared in the document. Moreover, one participant stated that the answers to all three questions in the interview notes started with the interviewee's name and a verb. Besides, two comments mention that inapt words were used. So, the answers to interview questions were labeled as "solutions" instead of "answers" and in the planning of a company's annual party a "keynote address" was suggested instead of a "keynote". In addition, one comment indicated that the e-mail making suggestions for a company's party was linguistically too complex.

Furthermore, participants mentioned content-related aspects impacting the documents' naturalness. In the case of the interview meeting notes, one participant perceived the answers to interview questions as noted down in too much detail for meeting notes. Another comment stated that in the e-mail about a company party, the sender has made too many suggestions. Here, they expected fewer suggestions and a follow-up meeting. Moreover, one comment remarked that in the interview notes, the author referred to herself in the third person.

Regarding social naturalness, one comment stated that it seemed unnatural that the interviewee was referred to by her first name in the meeting notes. Another comment indicated that a suggestion for the company party sounded like an advertisement for the catering service and not like a suggestion.

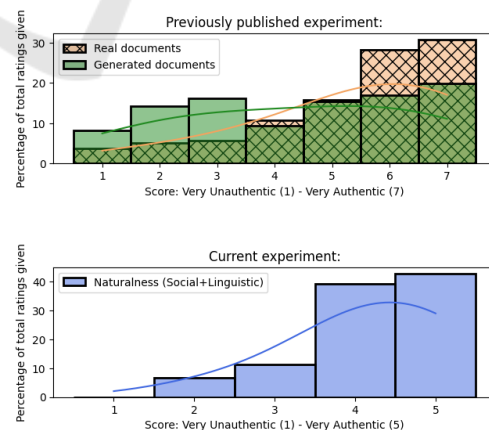


Figure 1: Plots showing the dispersion of given ratings of an earlier experiment assessing the perceived document authenticity of real and generated documents (Heim et al., 2024) (top) and the presently conducted experiment assessing the naturalness of generated documents (bottom).

4.2 Experiment 2: Parameter Adherence

In the second experiment, participants had to rate two documents regarding how well they respected descriptions utilized in the generation prompt that defined the document to be generated by key parameters. Figure 2 depicts the results. Again, overall, the participants assigned high scores. Moreover, around 75% of the participants agreed on a high score of four or five.

In the 13 comments given for the 98 ratings, no participant stated that a parameter, i.e. a part of the given description, was completely ignored in the generated document. Nevertheless, one comment indicated that the meeting notes did not properly address the occasion of the company's annual meeting when discussing the menu.

Two other comments stated that the generated documents did partially not meet their expectations relating to the description. Both comments addressed the discussion about the menu for the company party. One participant expected that the people involved in the discussion either discuss concrete dishes and beverages for the menu while the generated document rather discussed general suggestions. Another participant noted that the suggestion of collaborating with local businesses and farms seemed too exaggerated.

Moreover, three comments mentioned some technical content-related issues specific to the given document types. For the menu discussion notes, one participant perceived the document as being too focused since there were no remarks or further comments given alongside agenda points. In addition, they remarked that using bullet points instead of using whole sentences would have been more authentic for meeting notes. The other two comments referred to the generated project proposal and noted that the objectives and outcomes sections overlapped too much and suggested giving more details about the gap to existing works and potential risks.

Additionally, it was mentioned that the cost calculation in the project proposal was added up wrongly.

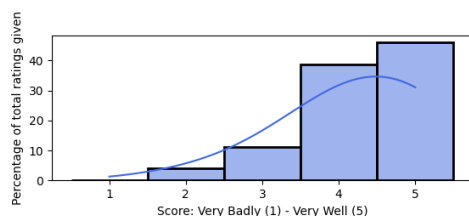


Figure 2: Bar chart depicting the dispersion of scores assigned for the parameter adherence of documents.

4.3 Experiment 3: Domain Adaptability

In the last experiment, a subgroup of the participants of the two previously introduced studies, the participants acquainted with specific domains, had to rate seven documents of various types generated using respectively on of their fields of expertise as the domain parameter. Here, participants could indicate their uncertainty if they were asked to judge the correctness of a document with unfamiliar facts or sub-fields.

Figure 3 shows the ratings given for the domain coherence on a word, statement, and document level. Overall, the given ratings are high having a median ranging between 4.5 and 5. For the word and statement level, participants had a high agreement and 75% of the participants gave a rating of 4 or 5. The highest variance occurred in the document-level case. Moreover, the participants were sure about their ratings as the checkbox indicating uncertainty about a rating was only ticked in approx. 10% of the cases.

In total, participants commented on seven of the 56 ratings. In two comments, they expressed their astonishment about the quality of the generated documents. One comment, referring to a generated exam, emphasized that especially the structure and classification of the content were outstanding.

Besides, three comments stated some domain-related issues in generated documents. In one case, in interview notes for a position as a Robotics engineer, a participant indicated that wrong information was provided which is why the participant gave a low rating on the statement level. The other two comments accompanied still high ratings but remarked that the contents of the documents did not entirely fit the given domain. Hence, one document detailing multiple interesting fields in the domain of chemical engineering contained a section about the field of "process automation and control" which the participant indicated as being not as relevant to the domain as the other mentioned fields. For another document, a project proposal about chemical manufacturing, a participant argued that the domain adherence was not ideal since the document's topic mixes two domains, namely chemical engineering and biotechnology.

Moreover, one comment, accompanying a low rating, stated that the document was about planning a course but did barely address the given domain. Furthermore, one domain-unrelated comment mentioned that while the document indicated that e-mail and phone number would be given as contact details, only a phone number was provided.

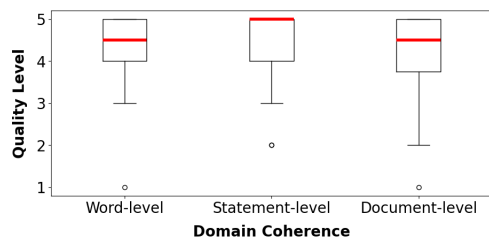


Figure 3: Boxplots of the results of Experiment 3. Each question is depicted in a separate box plot. The quality level corresponds to the given scores. A quality level of one represents the lowest quality score while a level of five represents the highest achievable rating.

5 DISCUSSION

Overall, participants gave predominantly high scores for all three experiments. In the first experiment, participants rated the naturalness with high scores regarding social and linguistic or content-related aspects. In comparison with the experiment published previously (Heim et al., 2024), the score distribution resembled the distribution of scores given for real documents more than the one given for generated documents. However, the experiment settings were different as the previous experiment was conducted based on real and generated documents, and our experiment only assessed generated documents. Moreover, in the previous experiment, participants did not know whether documents were real or generated and here we explicitly made participants aware that all documents had been generated. Nevertheless, the observations could indicate that the authenticity of generated documents improved since the previous KnoWoGen version, e.g., due to improved prompt instructions, and that documents generated with Mistral-7B-Instruct (Jiang et al., 2023) are more authentic than ones created with the Llama-13B-Chat model (Touvron et al., 2023). Moreover, the most commented issue was the linguistic repetitiveness that might be improved by adapting the Large Language Model (LLM) parameters accordingly. Additionally, participants mentioned that meeting notes seemed unauthentic because they contained full sentences instead of note-like structures. Hence, it might be meaningful to add for a specific document type more instructions regarding its typical form in the prompt. Besides, the stated problem of an author referring to herself in the third person might be solved by including an instruction in the prompt that asks the LLM more explicitly to take the role of the author.

The second experiment showed that overall, all parameters were respected. Nevertheless, one comment indicated that the influence of some parameters

could have been higher, and two others remarked that the produced content did not entirely match their expectations. Thus, for instance, a specific instruction could be added to the document generation prompt encouraging that parameters should also substantially influence the content of the respective document. Additionally, participants provided some further hints on issues in the comments that are unrelated to the parameter adherence. Like in the first experiment, again some suggestions were provided regarding the participants' expectations related to specific document types which could be considered in additional document type-specific instructions. Moreover, a few participants identified an erroneous cost calculation in a document which is likely not easily resolvable by adapting the prompt. However, such errors do not impact the objectives of generating a knowledge work dataset. Besides, real knowledge work documents might also include similar errors.

Like in the two previous experiments, participants also perceived the coherence of generated documents to a specific given domain as high. In particular, the factual correctness of statements was rated with exceptionally high scores. Only for one document out of the generated 56 documents, one participant remarked that a wrong statement appeared. Moreover, there were no comments regarding wrongly used terminology. Concerning the overall contents of the documents, the ratings were a bit lower than for the other two questions. Two comments remarked that a document contained contents that also belonged to another, related domain, and in one case a mentioned sub-topic was not highly relevant to the domain. In summary, the domain adaptability of generated documents was high. Our study indicates that factually correct documents with suitable terminology can be generated. The biggest challenge was compiling content that matches the core of the domain. If suitable, it might be helpful to address this issue by sampling a more specific sub-topic in the domain and generating documents for this more focused topic. Alternatively, instructions to focus on core topics of the domain could be tested. Nevertheless, all mentioned problems about the overall content did not state that the contents were unrelated to the domain which is already a sign of high quality.

To conclude, the experiments show that the generated documents respect given parameters and domains and thus, we can answer both of our research questions positively. Hence, generating knowledge work documents based on descriptive parameters that can be later used as ground truth or training data is feasible.

6 CONCLUSION

In this paper, we examined how well Large Language Models (LLM) can generate knowledge work documents with respect to a specified topical domain and descriptions comprising multiple parameters. Our studies were conducted on documents generated by our knowledge work dataset generator KnoWoGen and the Mistral-7B-Instruct LLM. Overall, the experiments show that the generated documents were perceived as natural, fitting their intended domain and other parameters, making the parameters reliable ground truth data.

In future experiments, it would be meaningful to also examine multiple, related documents and inspect whether the generated documents are coherent regarding their common task description and their contents as this information can also serve as relevant ground truth. Moreover, regarding topics of generated documents, it would be also interesting to assess the content-related variability of documents in a larger set of documents targeting the same topic or domain. Moreover, since our experiments showed that parameters are well-respected, in follow-up work, it is now also meaningful to examine whether synthesized documents are valuable as training data to improve the performance of machine learning models on downstream tasks.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education and Research (BMBF) in the project SensAI (grant no. 01IW20007).

REFERENCES

- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.
- Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. (2024). Open LLM leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.
- Gonçalves, D. (2011). Pseudo-desktop collections and PIM: The missing link. In *ECIR 2011 workshop on evaluating personal search*, pages 3–4.
- Heim, D., Jilek, C., Ulges, A., and Dengel, A. (2024). Using large language models to generate authentic multi-agent knowledge work datasets. In *INFORMATIK 2024*, pages 1347–1357. Gesellschaft für Informatik e.V., Bonn.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b.
- Jones, W. (2008). *Keeping Found Things Found: The Study and Practice of Personal Information Management*. Morgan Kaufmann.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., and Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the ACL, ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11065–11082. ACL.
- Reinhardt, W., Schmidt, B., Sloep, P. B., and Drachsler, H. (2011). Knowledge worker roles and actions—results of two empirical studies. *Knowledge and Process Management*, 18:150–174.
- Sordi, J. O. D., de Azevedo, M. C., Bianchi, E. M. P. G., and Carandina, T. (2020). Defining the term knowledge worker: Toward improved ontology and operationalization. *Academy of Management Proc.*
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2307.09288.
- Tung, V. T., Jacucci, G., and Ruotsalo, T. (2017). Watching inside the screen: Digital activity monitoring for task recognition and proactive information retrieval. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):109:1–109:23.
- Xu, R., Cui, H., Yu, Y., Kan, X., Shi, W., Zhuang, Y., Wang, M. D., Jin, W., Ho, J., and Yang, C. (2024). Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the ACL, ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15496–15523. ACL.
- Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., and Kong, L. (2022). Zerogen: Efficient zero-shot learning via dataset generation. In *Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11653–11669. ACL.
- Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A. J., Krishna, R., Shen, J., and Zhang, C. (2023). Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems 36: Annual Conf. on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv*, 2303.18223.