Can LLMs provide Recommendations to support Policy Making and Agency Operations?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have provided incredible tools when it comes to text generation. These generative capabilities bring us to a point where LLMs can provide useful insights in policy making or agency operations. In this paper, we introduce a new task consisting of generating recommendations which can be used to inform future actions and improvements of agencies work within private and public organisations. The paper presents the first benchmark and coherent evaluation for developing recommendation systems to inform organisation policies. This task is clearly different from usual product or user recommendation systems, but rather aims at providing a basis to suggest policy improvements based on the conclusions drawn from reports. Our results demonstrate that state-of-the-art LLMs have the potential to emphasize and reflect on key issues and learning points within generated recommendations.

1 Introduction

Recent large language models (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) have shown exceptional abilities in text generation tasks such as summarisation (Zhang et al., 2024; Xie et al., 2023) and story generation (Tang et al., 2022; Razumovskaia et al., 2024), achieving results comparable to human-created text. Given the ability of LLMs to understand instructions written in natural language (*'prompts'*), the majority of work is focused on utilising prompt-based approaches for adapting pre-trained models to different domains and tasks (Viswanathan et al., 2023; Plaza-del Arco et al., 2023).

The continuous advancements in the creation of bigger and more powerful language models have led to further research into how these models can be utilised for more specialised tasks (Huang et al., 2024), usually performed by domain experts, such Court View Generation (CVG) in the legal domain (Li et al., 2024; Yue et al., 2021; Wu et al., 2023). CVG is a natural language generation (NLG) task, which aims to generate court views based on the plaintiff claims and the fact descriptions related to a given court case (Li et al., 2024). Research in the area have shown promising results of using pre-trained language models coupled with prompting techniques (Yue et al., 2021; Wu et al., 2023). Li et al. (2024) take this research further by proposing a method for incorporating domain knowledge and guidance within pre-trained language models. The method achieved better results for the CVG task, compared to generic language models. This work shows the need for further attention into developing approaches which harness the power of LLMs and the expertise of domain experts in order to improve text generation for more challenging and specialised domains. However, work in this area is still limited with the majority of research being related to the field of Legal Artificial Intelligence (LegalAI). This paper presents the first step towards expanding research into harnessing LLMs for more domain-specific and specialised NLG tasks such as recommendation generation for informing policy making and improving agencies work across the provision of public services. It is a challenging task, different from standard text generation tasks such as story completion and product recommendation, due to the fast changing requirements within the private and public sector organisations, and the highly diverse, dynamic and specialised terminology and structure of related documents.

Our contributions are as follows: (1) We present a new task within the field of NLG related to incorporating LLMs into the public services in order to support practitioners into writing a set of recommendations, related to a given incident or identified problem, which can be used to inform the design of better delivery services for vulnerable individuals. (2) We make available two datasets for the task.

The *'UK Care Homes'* reports reflecting on the quality of care homes for vulnerable adults within UK and the *'US Children's Bureau'* reports which assess the quality of foster care and adoption services in US. (3) We perform extensive evaluation of the performance of models for recommendation generation, using similarity measures, LLM-based evaluation, and human-based evaluation. Results from these analysis show the potential of LLMs for the given task and also highlight the discrepancy between the different evaluation measures and the need for developing evaluation approaches better fitted for this particular NLG task.

2 Recommendation Generation Dataset

Task Description. Local authorities and community safety partnerships often need to produce reports in order to reflect on public services or identify and describe related events that precede a serious incident, for example involving a child or vulnerable adult. A key role of these documents is to reflect on agencies' roles and the application of current practices in social care provision and crime prevention. These reports, despite being quite diverse in structure and topics, need to contain key lessons learned (***evidence)*** of good or bad practices that are used to derive a set of (***recommendations)***. These recommendations are disseminated (independent of the reports) across relevant institutions in order to inform the development of policy making for improving service delivery across different governmental sectors. The development of these recommendations can be bias and a resource- consuming task, resulting very often in the creation of bad quality content. In this paper, we explore if and how LLMs can be used to support practitioners in writing high quality recommendations (see example in Figure 1). Specifically, given an evidence of lessons learned, our task consists of generating a recommendation which reflect on and it is consistent with the provided information.

Dataset Creation. We collected two datasets, consisting of reports reviewing agencies work related to the provision of services to vulnerable individuals. The *'UK Care Homes'* [1] dataset consists of reports produced by The UK Care Inspectorate in order to reflect on the quality of care homes for vulnerable adults in UK. The *US Children's Bureau* dataset [2] consists of reports that assess the quality of foster care and adoption services in US.

[1]UK Care Inspectorate:www.careinspectorate.com

[2]Children's Bureau: https://www.acf.hhs.gov/cb

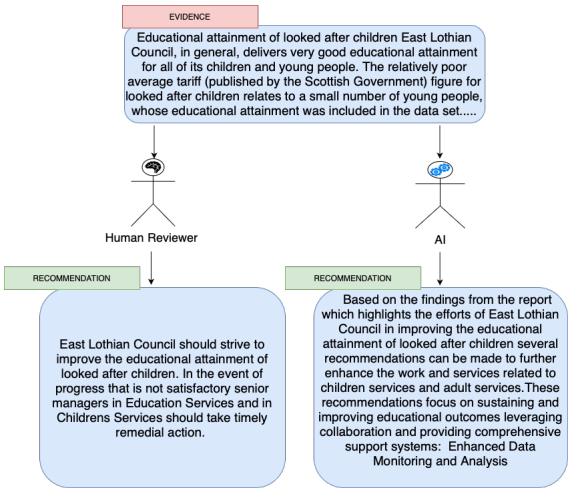

Figure 1: An example of human- and GPT- generated recommendations given an evidence.

Both datasets are publicly available to download via their websites.

The two datasets consist of 70 reports and 216 recommendations in total (see Table 1), which is a relatively small collection. However, considering that these reviews are produced only when a specific event occurs such as an incident, our collection represents a good subset of the total number of reports available. For the purposes of our analysis, we have extracted the evidence from the reports as these contain sufficient information for generating recommendations, and this setting can help prevent possible LLM hallucinations with irrelevant information from the reports. Further, reports for both datasets have an average length above 7,000 tokens (see Table 1) which makes processing in their entirety a challenging task, subject to future research. Both datasets will be publicly released upon acceptance.

	Care Homes	US Children Bureau
# reports	22	48
# *recs*	94	122
Avg number of *recs* per report	4	2
Avg tokens (*recs*)	34	118
Avg tokens (evidence)	742	254
Avg tokens (reports)	9,567	7,943

Table 1: Dataset statistics (*recs*=recommendations)

.

3 Experimental Setting

3.1 Recommendation Generation

The aim of the paper is to analyse the feasibility of incorporating LLMs within the process of writing recommendations for improving public services and agencies work based on evidence collected from previous good and bad practices. For these

Data	Model	Bert-Score (F1)	Rouge-L (F1)	Bleu Score	GPT-based eval.	LLaMA-based eval.
UK Care Homes	GPT4-o	0.497 (±0.035)	0.143 (±0.055)	0.004 (±0.015)	1.957	1.714
	LLaMA 3	0.525 (±0.038)	0.171 (±0.062)	0.007 (±0.02)	1.902	1.728
US Children's Bureau	GPT4-o	0.551 (±0.049)	0.204 (±0.053)	0.021(±0.033)	2.692	2.101
	LLaMA 3	0.542 (±0.049)	0.196 (±0.058)	0.012 (±0.023)	2.350	2.000

Table 2: Averaged evaluation results across generated recommendation per dataset based on similarity metrics ('eval.' refers to evaluation).

purposes, we use the OpenAI GPT4-o model as it is known to be one of the most powerful NLP models available. Further, we use LLaMA 3 model with 8 billion parameters, pre-trained with instructions, downloaded from HuggingFace (Wolf et al., 2019). We generate recommendations using prompting in zero-shot settings where the model is given a description of the task and an evidence. For creating the prompt, we followed examples provided by OpenAI and Meta. In addition, we followed design principles described in (Reynolds and McDonell, 2021) for creating self-explanatory prompts which are easy and intuitive to use from user perspective.

Prompt for generating recommendations

Provide a recommendation for improving agencies work and services related to children care and children services. The recommendation should cover topics mentioned in the given evidence without deviating from the topics mentioned and not writing any fact which is not present here.
Evidence:[Evidence]

3.2 Evaluation

We evaluate the generated recommendations using three types of evaluation measures, i.e., similarity metrics, LLM-based evaluation, and human-based evaluation. This allows us to capture different aspects of how well the models perform for recommendation generation as well as allow analysis into the suitability of these measures for evaluating Natural Language Generation (NLG) tasks.

Automatic Metrics. We use traditional reference-based evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) which measure the extent to which generated content matches the n-grams of the reference text. In particular, we use ROUGE-L to measure the longest common sub-sequence (LCS). In addition, we use BERTScore (Zhang et al., 2019), an embedding-based method which uses embedding representations of the reference and the target text to compute semantic similarity between them. This metric could be better suited to the varying size of recommendations. Nonetheless, we anticipate that these automatic metrics may have shortcoming when it comes to the evaluation and therefore, we propose both an additional automatic LLM-based metric and a human evaluation.

LLM-based Evaluation. We use a prompt-based approach (Gao et al., 2024) and GPT4-o model for measuring the factual alignment between the reference and target recommendations. The prompt is created following the same principles used for recommendation generation in Section 3.1. Within the prompt we specify the evaluation criteria based on a 3-point Likert scale where 1 refers to the lack of any factual alignment between the recommendations and 3 refers to a complete factual alignment between them. We use the same scale for the human evaluation to allow comparison between the evaluation approaches.

Prompt for evaluating recommendations

You are given two recommendations (Recommendation 1 and Recommendation 2). Your task is to measure the factual alignment between the two recommendations using a scale from 1 to 3 where 1 refers to the lack of any factual alignment between the recommendations and 3 refers to a complete factual alignment between them.
Evaluation Form: Answer by starting with 'Rating:' and then give the explanation of the rating on the next line by 'Rationale:

Human Evaluation. For conducting human evaluation, we followed principles described in (Chhun et al., 2022) and (Li et al., 2024). In this way we outlined 4 main criteria for conducting the evaluation: **(1) Fluency**— measures the quality of the text including grammatical errors and repetitions; **(2) Coherence** — measures whether the recommendation makes logical sense. **(3) Relevance to the evidence**— measures whether the recommendation matches the given evidence; **(4) Relevance to the human-created recommendation** — measures the factual alignment between the two rec-

ommendations (we use the same criteria for LLM-based evaluation to allow comparison between the two measures);

During evaluation, participants are given the generated recommendation, the evidence used to generate the recommendation, and the human-created recommendation. Each recommendation is evaluated by two subject matter experts using a 3-point Likert scale where 1 is worst and 3 is best. Finally, considering the highly specialised nature of the datasets which require domain experts for evaluation, we performed these experiments for 50 randomly selected recommendations across the two datasets.[3]

4 Results and Analysis

Automatic Evaluation. Table 2 shows results of recommendation generation based on automatic metrics. The similarity metrics, especially Bleu Score and Rouge-L show quite low results across the datasets and models in comparison to LLM-based evaluation. This highlights the limitations of these traditional automatic metrics to capture the factual correctness of generated text as well as semantic similarities for more complex NLG tasks. In contrast, LLM-based evaluation (regardless of model used) shows a good quality of generated recommendations regarding factual consistency with the gold standard. Specifically, the average score between GPT4-o and LLaMA for the UK Care Homes for recommendations generated using GPT4-o is 1.836 and for LLaMA-generated recommendations is 1.815. For the US Children's Bureau dataset, the average scores for GPT4 and LLaMA are 2.397 and 2.175, respectively. The results suggest a better performance for GPT4-o and thus we use recommendations generated with this model to perform human evaluation. Overall, evaluation results show a better performance for the US Children's Bureau dataset which can be attributed to the fact that the 'evidence' for these documents are shorter passages in comparison to the UK Care Home dataset. Another potential reason is the regional differences between the two datasets where the US-based reports cover a bigger and potentially better represented location within the training set of these models.

Human Evaluation. Figure 2 shows a good overall performance of GPT4-o for recommendation generation across both datasets where the average score across the majority of criteria is above 2. The finding, from the previous section, that GPT4-o performs better for the US-based dataset is also confirmed by the human evaluators. These results also show higher overall score for the 'relevance to the evidence'-based criteria versus 'relevance to the human-created recommendation' (0.5 difference in score). This suggests that a strength of LLMs in NLG is in providing a different perspective for the task/input which can be useful to users, versus simply recreating the human gold standard. This also highlights the need for more task-targeted and purpose-oriented evaluation metrics.

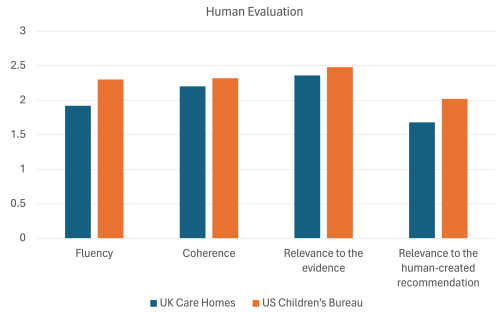

Figure 2: Results from human-based evaluation.

Finally, Table 3 shows a decent similarity and correlation between human and LLM-based evaluation measures, using both GPT and LLaMA as evaluators. Nonetheless, this is limited to a few samples and there may still be biases such as model preferring their own generations (Kocmi and Federmann, 2023), which may emerge in a larger settings that we have not analysed in this work.

Dataset	GPT-based eval.	LLaMA-based eval.	Human eval.
UK Care Homes	1.957	1.714	1.656
US Children's Bureau	2.692	2.101	2.000

Table 3: Comparison between LLM- and Human-based evaluation in reference with criteria (4) (*relevance to human recommendation*.

5 Conclusions

This paper presents the first work towards incorporating LLMs for more domain-specific and specialised NLG tasks such as recommendation generation for informing policy making and improving agencies work related to the provision of public services. We present two datasets relevant to the task and perform an evaluation of the performance of GPT4-o and LLaMA 3 across the two datasets using zero-shot prompting. LLM-based and human-based evaluations of GPT4-o's output show promising results where human evaluators found the majority of generated recommendations to be relevant to the given evidence as well as coherent and fluent in their structure and content.

[3]Instructions available in the appendix.

6 Limitations

This study was the first approximation to use LLMs for recommendation generation to support policy making and agency work. As such, it comes with its own limitations. First, the datasets are available in English only which limits their usage to only English based tasks. Second, analyses are performed for two models in zero-shot settings. As future work we plan on extending these analysis to understand how the performance of models can be improved for the given task. Finally, the corpus consists of two datasets of a relatively small size. In future, we plan to extend it by including reports from diverse sources.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. How good are low-bit quantized llama3 models? an empirical study. *arXiv preprint arXiv:2404.14047*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.

Ang Li, Yiquan Wu, Yifei Liu, Kun Kuang, Fei Wu, and Ming Cai. 2024. Enhancing court view generation with knowledge injection and guidance. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5896–5906.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2023. Leveraging label variation in large language models for zero-shot text classification. *arXiv preprint arXiv:2307.12973*.

Evgeniia Razumovskaia, Joshua Maynez, Annie Louis, Mirella Lapata, and Shashi Narayan. 2024. Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10616–10631, Torino, Italia. ELRA and ICCL.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Context-tuning: Learning contextualized prompts for natural language generation. *arXiv preprint arXiv:2201.08670*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. In *arXiv e-prints*, pages arXiv–2307, online. arXiv.

Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu, and Graham Neubig. 2023. Prompt2model: Generating deployable models from natural language instructions. *arXiv preprint arXiv:2308.12261*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. *arXiv preprint arXiv:2310.09241*.

QIANQIAN Xie, ZHEHENG Luo, BENYOU Wang, and SOPHIA Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *arXiv preprint arXiv:2304.08763*.

Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021. Circumstances enhanced criminal court view generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1855–1859.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Appendix

B Human Evaluation

The instructions, given to the subject matter experts, who participated in the human evaluation are illustrated in Figure 3. Further, Figure 4 shows results from GPT-based evaluation categorised by score.

Your task is to evaluate AI generated recommendations ('Generated Recommendation') following the given criteria:

(1) Fluency: measures the quality of the text ('Generated Recommendation') including grammatical errors and repetitions;

(2) Relevance to the evidence: measures whether the 'Generated Recommendation' matches its clue;

(3) Relevance to the human-created recommendation measures the factual alignment between the two recommendations ('Generated Recommendation' and the 'Human Recommendation') -- Note: if the recommendation is "good" related to the 'Relevance to the clue' criteria but very different to the human recommendation, the score should be 1.

(4) Coherence: measures whether the recommendation makes logical sense;

Please evaluate each 'Generated Recommendation' within the given excel file using a 3-point scale where 1 is worst and 3 is best.

Figure 3: Instructions for human evaluation.

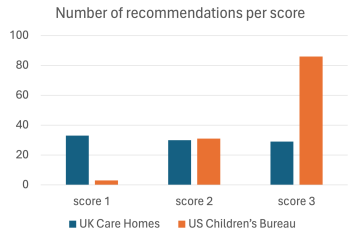

Figure 4: Comparison of results from GPT-based evaluation between the two datasets, ie, Care home reports and US reports (‘US data’).