Report-based Recommendations for Policy Making and Agency Operations: Dataset and LLM Evaluation

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 could potentially 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. In particular, we present 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 LLMs (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., 2023a) and story generation (Tang et al., 2022; Razumovskaia et al., 2024), among others, 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; Chae and Davidson, 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. An example of such task is Court View Generation (CVG) in the legal domain (Li et al., 2024; Yue et al., 2021; Wu et al., 2023), where the aim is to generate interpretations of judgment results. The majority of research in the area is focused on incorporating domain knowledge within pre-trained language models (Li et al., 2024; Wu et al., 2023; Yue et al., 2021). These domain-targeted methods showed to be more beneficial for the CVG task, compared to generic language models. These findings highlight 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). Further, there is a lack of profound analysis into the suitability of existing evaluation measures and approaches for text generative models for more specialised domains. While recent work has focused on governance and safety risks of LLMs (Goanță et al., 2023), these discussions remain largely at the level of high-level regulatory debates. Efforts such as RegNLP advocate for integrating regulatory science with NLP to improve risk assessment, but their primary aim is to guide regulation rather than develop models for domain-specific generative tasks.

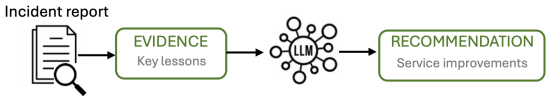

Figure 1: A high-level overview of the recommendation generation pipeline.

This paper presents the first step towards expanding research into harnessing LLMs for recommendation generation in the context of informing policy making and improving agencies work across the provision of public services (see Figure 1). It is a challenging task, different from standard text generation tasks such as story completion and product recommendation, due to the fast changing re-

quirements within the private and public sector organisations, and the highly diverse, dynamic and specialised terminology and structure of related documents.

Our main contributions are as follows:

1) We present a new natural language generation (NLG) task which investigates the use of LLMs for helping practitioners within the public sector in writing recommendations that are used to support policy making processes for improving service delivery for vulnerable individuals.

2) We make available a unified benchmark dataset (PubRec-Bench) for the task, which has been collected from three different data sources: The *'UK Care Homes'* reports reflecting on the quality of care homes for vulnerable adults within UK, the *'US Children's Bureau'* reports which assess the quality of foster care and adoption services in US, and the 'NSPCC' reports which reflect on agencies work regarding serious incidents involving children.

3) We perform extensive evaluation of the performance of three state-of-the-art text LLMs for recommendation generation, using similarity measures, LLM-based evaluation, and human evaluation. Results from this analysis show the potential of LLMs for the given task and also discuss the discrepancy between the different evaluation measures and the need for developing evaluation approaches better fitted for this particular NLG task.

2 Related Work

NLG aims to produce text from a given input data where the generated output needs to satisfy certain language properties and task requirements (Tang et al., 2022). The enhancements in the field in the recent years in terms of creating more powerful language models, have lead to an increased research into how to utilise these tools for more challenging problems and domains requiring subject matter expertise or/and lack training data. Many approaches tackling the data sparsity problem rely on prompting (in-context learning) techniques for generating text. Prompting is a technique which allows to guide LLMs into performing downstream tasks by providing either instructions written in natural language (zero-shot) or providing a few examples (few-shot) (Razumovskaia et al., 2024). Existing work has shown that prompting can lead to a strong performance in various tasks such as question answering (Chowdhery et al., 2023; Agrawal et al., 2023) and open-ended natural language generation (Tang et al., 2022), even in some cases to comparable or even better performance than standard fine-tuning techniques especially in the absence of training corpora (Gao et al., 2021; Mosbach et al., 2023).

Research into utilising LLMs for text generation in more specialised domains is mainly focused on summarisation tasks for the clinical and law domains. For instance, in the medical domain there is an increased work on developing summarisation tools to support clinical information retrieval and management (Xie et al., 2023a,b; López-Úbeda et al., 2024). In the legal domain, there has been an increased interest in developing LLM-driven approaches for court view generation (CVG) (Li et al., 2024; Yue et al., 2021; Yu et al., 2022; 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). The majority of research in the area is focused on incorporating domain knowledge and LLMs for the task (Wu et al., 2023; Li et al., 2024; Yue et al., 2021) where results show the need for more domain-targeted approaches when it comes to highly specialised texts. For instance, the approach proposed by Li et al. (2024) is based on injecting claim-related knowledge such as keywords and label definitions within the prompt encoder of the model. The authors of (Wu et al., 2023) propose a framework that incorporates pre-trained LLMs, prompting techniques and small domain-trained language models. A work by (Savelka et al., 2023) takes a different approach where the authors evaluate the capability of GPT4 for court opinions to interpret legal concepts. The work showed that GPT-4, guided only by in-context learning techniques, can give similar performance to a well-trained law student annotators. This work highlights interesting research avenues for exploring text generation capabilities of LLMs in more specialised domains. However, prior work is mainly focused on the LegalAI domain. Further, there is a growing concern about suitability of existing evaluation measures when it comes to text generation (Liusie et al., 2024; Panickssery et al., 2024; Gao et al., 2025; Khashabi et al., 2022; Chaganty et al., 2018), especially within more high risk domains and tasks (López-Úbeda et al., 2024). However, the aforementioned research lack discussion on suitability of evaluation metrics used. In our work, we expand existing research by presenting a new task and a dataset related to recommendation

generation where recommendations can be used to inform policy making for improving multi-agency work and service delivery to vulnerable individuals. Further, we present thorough evaluation of text generation models as well as discussion of feasibility of approaches.

3 PubRec-Bench: Recommendation Generation Benchmark

In this section, we describe the task of recommendation creation for informing policy making in the public sector (Section 3.1), the process of collecting and unifying relevant datasets (Section 3.2), and their statistics (Section 3.3).

3.1 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 biased 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 Specifically, given an evidence of lessons learned, our task consists of generating a recommendation which reflects on and it is consistent with the provided information.

3.2 Dataset Collection and Unification

We collected three datasets, consisting of reports reviewing agencies work related to the provision of services to vulnerable individuals. These reports are lengthy and contain information irrelevant to the recommendation generation task, such as information regarding the reviewing board, incident description and timeline of events. Thus, 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, the reports have very diverse structure and content within and across the different data sources making it hard to identify evidence with associated recommendations. Therefore, we unified the datasets by manually going through each report, extracting evidence and associated recommendations. All reports are publicly available to download via their websites. Examples of evidence and recommendation pairs for each dataset are given in Table 1. All datasets included in PubRec-Bench are described below. The datasets will be publicly released upon acceptance.

UK Care Homes reports. The *'UK Care Homes'* [1] dataset consists of reports produced by The Care Inspectorate in order to reflect on the quality of care homes for vulnerable adults in UK. The website contains roughly around 300 reports, however, not all of them contain recommendations. In order to allow comparison between generated and human-written recommendations we have excluded reports with missing recommendations from our collection.

US Children's Bureau reports. The *US Children's Bureau* dataset [2] consists of reports that assess the quality of foster care and adoption services in the US. Children's Bureau is an agency within the Administration for Children and Families, which is part of the U.S. Department of Health and Human Services. The learning points and recommendations from the reports are used to help prevent child abuse and neglect, create better adoption services and foster care.

NSPCC reports. NSPCC (The National Society for the Prevention of Cruelty to Children) is UK's leading children's charity that specialises in child protection and prevention of child abuse. The NSPCC reports[3] consists of case reviews written by UK-based Local Safeguarding Children Boards (LSCBs)

3.3 Data Statistics

Table 2 summarizes statistics of each PubRec-Bench dataset after unification. The three datasets consist of 110 reports and 493 recommendations in total. Considering that these reviews are produced

[1] UK Care Inspectorate: www.careinspectorate.com
[2] Children's Bureau: https://www.acf.hhs.gov/cb
[3] NSPCC reports: https://library.nspcc.org.uk

Dataset	Evidence	Recommendation
UK Care Homes	The social care and wellbeing learning and development team action planning framework was substantial but it was not possible to evaluate the impact. The family and community support action plan 2012 was a draft and had not been fully populated...	The social work services should ensure that annual reviews of people placed in care homes are carried out by clarifying the appropriate responsibilities and timescales.
US Children Bureau	Use of the supplemental issuance code as a 'catch-all' for certain costs. Regional Office staff were required to manually review and request additional information in 26 cases in order to determine the purposes for the supplemental issuances and whether they were for allowable title IV-E maintenance expenditures...	The state should provide guidance to counties to be sure that it is able to segregate out the reasons why the supplemental issuance code is used so that the various types of supplemental payments may be identified.
NSPCC reports	The work would have benefitted from exploration of key relationships and extended family on both sides.... A genogram would have enabled further exploration of the nuances of the family. Whilst it is unlikely that this would not have impacted on the outcome, it would have provided a more complete picture...	SHIELD to develop a 7 minute briefing and top tips for practitioners about how to act on gut feelings and professional curiosity. A task and finish group should lead on this work which should include refreshing and promoting SHIELD's website content.

Table 1: Examples of extracted evidence and recommendation pairs per dataset type.

only when a serious incident occurs, our collection represents a substantial subset of the total number of reports available. Further, reports for all datasets have an average length above 7,000 tokens (see Table 2) which makes processing in their entirety a challenging task, which could be a subject to future research.

	UK Care	US Children	NSPCC
# reports	22	48	40
# *recs*	94	122	276
Avg # *recs* per report	4	2	7
Avg # tokens per *recs*	34	118	61
Avg # tokens per evidence	742	254	219
Avg # tokens per reports	9,567	7,943	13,120

Table 2: Dataset statistics where *'#reports'* refers to number of reports per dataset, *'#recs'* refers to number of recommendations per dataset, *'avg'* refers to average.

4 Experimental Setting

4.1 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. We would like to note that we focus on evaluating models which are known to provide state-of-the-art performance for text generation tasks, especially in low-resource settings. Therefore, performing extensive evaluation of a large variety of different models is outside the scope of the paper.

Comparison Models. For the purposes of our analysis, we compare three different models. These are: **(1) OpenAI GPT4-o model** which is one of the most advanced models released within the NLP space and it is well known for its impressive zero- and few-shot capabilities (Savelka et al., 2023; Brown et al., 2020). **(2) Command R+** is Cohere's most powerful and newest large language model, optimized for conversations and long-context tasks and it consists of 104B parameters. **LLaMa 3 model** which is known to be one of the most advanced open source language models (Dubey et al., 2024). We use LLaMA 3 model with 8 billion parameters, pre-trained with instructions, downloaded from HuggingFace (Wolf et al., 2019)[4].

Prompting. Given the limited amount of annotated data, we use the in-context learning method to generate recommendations. As described in Section 2, prompting can lead to better results compared to fine-tuning techniques when data is limited. In addition, our objective is to analyze the extent to which state-of-the-art LLMs can perform complex tasks with limited resources. We generate recommendations using prompting in zero-shot and one-shot settings, where the model is given a description of the task and supporting evidence. We conduct experiments with two prompts. For the creation of *'Prompt 1'*, we followed examples provided by OpenAI and Meta. We also followed the design principles described in Reynolds and McDonell (2021) to create self-explanatory prompts that are intuitive and easy to use from the user's perspective. To create *'Prompt 2'*, we asked subject matter experts (see Section 4.2 for more details about the experts) to provide a description of the task to be included in the prompt. The actual prompts are given in the Appendix.

[4]Model parameters used are available in the Appendix.

4.2 Evaluation

We evaluated the generated recommendations using three types of evaluation measures, ie., 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 to allow analysis into the suitability of these measures for evaluating Natural Language Generation (NLG) tasks.

Similarity Metrics We use traditional reference-based evaluation metrics such as 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 subsequence (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., 2025) and measure the factual alignment between the reference and targeted recommendations using each one of the language models. The prompt is created following the same principles used for recommendation generation in Section 4.1. 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. The evaluation prompt is given in the Appendix.

Human Evaluation 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 five 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 240 randomly selected recommendations across the three datasets. The subject matter experts were selected through an interview process and all have experience in dealing with policymaking processes for governmental institutions. For conducting human evaluation[5], we followed principles described in previous work (Chhun et al., 2022; Li et al., 2024). We outlined 5 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 is meaningful given the evidence; **(4) — Relevance to the human-created recommendation** measures the factual alignment between the two recommendations (we use the same criteria for LLM-based evaluation to allow comparison between the two measures); **(5) — Is the recommendation 'Actionable'? (yes/no)** shows if the recommendation has practical application and could be implemented as part of a policy.

5 Results and Analysis

The aim of our analysis is to (1) identify to what extend state-of-the-art LLMs can perform recommendation generation for informing policy making, as well as (2) analyse the suitability of existing evaluation metrics for the task.

5.1 Automatic Evaluation

A comparison between the performance of the generation models for the two prompts (see Table 3) showed consistently higher results for prompt 2 (i.e., the prompt designed by subject matter experts). This shows the importance and need to involve domain expertise not only during the evaluation process of LLM-based approaches but also during the development of the LLM-based system.

Table 4 shows individual model results of recommendation generation based on automatic metrics. The similarity metrics, especially BLEU Score and ROUGE-L show quite low results across datasets, settings and prompts, 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,

[5] See the Appendix for the evaluation sheet.

Data	prompt	BERT-Score (F1)	ROUGE-L (F1)	BLEU Score	GPT-based eval.	LLaMA-based eval.	Cohere-based eval.
UK Care Homes	prompt 1	0.446	0.107	0.004	1.953	1.719	1.939
	prompt 2	0.555	0.189	0.008	2.168	1.806	2.048
US Children's Bureau	prompt 1	0.466	0.134	0.011	2.519	2.019	2.067
	prompt 2	0.584	0.231	0.016	2.570	1.997	2.056
NSPCC reports	prompt 1	0.445	0.108	0.007	2.197	1.904	1.949
	prompt 2	0.557	0.189	0.023	2.218	1.902	2.029

Table 3: Averaged evaluation results across all LLMs for generating recommendations using prompt 1 and prompt 2 (prompts described in Section 4) in the zero-shot setting. The evaluations are based on similarity metrics ('BERT Score', 'ROUGE-L', 'BLEU Score') and LLM-based evaluations using GPT ('GPT-based eval.'), LLaMA ('LLaMA-based eval.'), and Cohere ('Cohere-based eval.').

the average score for each LLM-based evaluation, regardless of the model used to generate recommendations, varies between 1.7 and 2.5. The results suggest a slightly 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 in 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 or the NSPCC dataset. Another potential reason is the regional differences between the datasets where the US-based reports cover a larger and potentially better represented location within the training set of these models.

Zero-shot vs. one-shot An important observation is that models consistently perform better in the zero-shot setting compared to the one-shot setting. One possible reason for this is the high variability in the evidence and recommendation formats, which suggests that traditional in-context learning approaches relying on a small number of labeled examples may be insufficient for improving model performance in this domain. Instead, more dynamic and domain-specific adaptation strategies may be needed to effectively guide the models.

LLM evaluators A comparison of the three LLM-based evaluation models for zero-shot setting (see Figure 2), where the scores are averaged across the three datasets, shows that the GPT4 and Cohere-based models give a higher score to their own outputs. This suggests a potential bias for these models towards their own generations (Kocmi and Federmann, 2023) which shows the need for further research into how best to utilise these models for evaluation tasks. These findings are also confirmed in the results for one shot setting, presented in the Appendix.

5.2 Human Evaluation

Table 5 show a good overall performance of GPT4-o for recommendation generation across the three datasets where the average score across the majority of criteria is above 2.5. Similarly to the automatic evaluation, generation models are shown to perform better in zero-shot rather than one- shot by the human evaluation as well. 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. In addition, the results for criterion *(5) 'Is the recommendation actionable?'* are promising. In the zero-shot setting, annotators agreed that around 60% of the recommendations across all datasets were actionable. In the one-shot setting, the figure was around 57%. These findings suggest that the generated recommendations have meaningful practical value and potential applicability within policymaking processes (See Appendix for full results.).

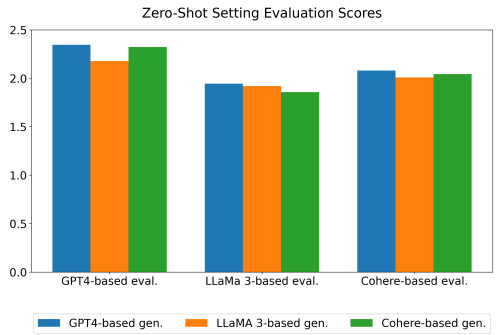

Figure 2: Comparison of LLM-based evaluations (*'eval'*) in zero-shot settings for recommendations generated by each model across the three datasets.

Correlation Analysis We investigated the correlation between human-based evaluation and automatic metrics considered in automatic evaluation (see Section 5.1) across the three datasets. We took the average across the two annotators for each generated recommendation to compute the correla-

Data	Setting	Gen Model	Bert-Score (F1)	Rouge-L (F1)	Bleu Score	GPT-based eval	LLaMA-based eval	Cohere-based eval
UK Care Homes	zero	GPT 4-o	**0.569**	0.181	0.010	**2.183**	**1.903**	**2.043**
	zero	Cohere	0.552	0.171	0.005	2.140	1.720	2.000
	zero	LLaMA	0.545	**0.189**	**0.009**	2.182	1.795	2.102
	AVERAGE zero-shot		*0.555*	*0.189*	*0.008*	*2.168*	*1.806*	*2.048*
	one	GPT 4-o	0.573	0.183	0.011	2.086	1.860	**2.075**
	one	Cohere	**0.578**	0.189	0.006	**2.237**	**1.913**	2.081
	one	LLaMA	0.542	**0.190**	**0.018**	1.806	1.667	1.978
	AVERAGE one-shot		*0.564*	*0.190*	*0.012*	*2.043*	*1.813*	*2.045*
US Children's Bureau	zero	GPT 4-o	0.583	0.224	0.013	2.594	**2.009**	**2.113**
	zero	Cohere	**0.594**	**0.246**	**0.020**	**2.612**	1.991	2.095
	zero	LLaMA	0.575	0.222	0.014	2.504	1.991	1.959
	AVERAGE zero-shot		*0.584*	*0.231*	*0.016*	*2.570*	*1.997*	*2.056*
	one	GPT 4-o	0.572	0.221	0.015	**2.273**	1.942	1.917
	one	Cohere	**0.588**	**0.243**	**0.017**	2.645	**2.017**	**2.132**
	one	LLaMA	0.547	0.202	0.008	2.058	1.909	1.974
	AVERAGE one-shot		*0.569*	*0.222*	*0.013*	*2.325*	*1.956*	*2.008*
NSPCC reports	zero	GPT 4-o	**0.567**	0.188	0.026	**2.258**	**1.920**	**2.084**
	zero	Cohere	0.550	0.179	0.015	2.218	1.865	2.036
	zero	LLaMA	0.554	**0.202**	**0.028**	2.178	1.910	1.967
	AVERAGE zero-shot		*0.557*	*0.189*	*0.023*	*2.218*	*1.902*	*2.029*
	one	GPT 4-o	0.555	0.173	0.013	2.047	1.884	2.000
	one	Cohere	**0.561**	0.179	0.014	2.149	**1.898**	**2.034**
	one	LLaMA	0.560	**0.188**	**0.028**	**2.175**	1.880	2.031
	AVERAGE one-shot		*0.558*	*0.180*	*0.018*	*2.123*	*1.887*	*2.023*

Table 4: Complete evaluation results by dataset and generation model, based on similarity metrics ('BERTScore', 'ROUGE-L', 'BLEU Score') and LLM-based evaluations using GPT ('GPT-based eval'), LLaMA ('LLaMA-based eval'), and Cohere ('Cohere-based eval'). All generations were produced using Prompt 2.

	setting	UK Care	US Children	NSPCC
Fluency	zero	2.753	2.877	2.787
Coherence	zero	2.887	2.970	2.890
Rel. to the evidence	zero	2.790	2.863	2.850
Rel. to human rec.	zero	2.553	2.537	2.463
AVERAGE	**zero**	**2.746**	**2.811**	**2.748**
Fluency	one	2.467	2.603	2.653
Coherence	one	2.767	2.787	2.837
Rel. to the evidence	one	2.740	2.753	2.700
Rel. to human rec.	one	2.307	2.437	2.277
AVERAGE	**one**	**2.570**	**2.787**	**2.617**

Table 5: Averaged results across subject matter experts for zero-shot (zero) and one-shot (one) settings, using GPT-4o for generation. 'Rel. to the evidence' refers to the 'Relevance to the evidence' criterion, and 'Rel. to human rec.' refers to the 'Relevance to the human-created recommendation' criterion.

tion. Figure 3 shows the Spearman's rank correlation coefficient and p-value[6] across the automatic metrics and the human evaluation scores regarding criteria *'(4) relevance to the human-created recommendation'* (see Section 4.2) which is the same criteria used for LLM-based evaluation. The p-values for a large proportion of the correlations are above 0.4 which makes them correlated, but not too strongly. This supports the findings in Section 5.2 and suggests that the task of evaluating recommendations for these datasets is quite a complex task and requires more purpose-build metrics. Furthermore, according to the correlation analysis presented in Figure 3, no metric achieved high agreement (above 0.5) with the human annotators. These findings highlights even further that we should not rely on a single metric to capture all quality aspects of a model's output. A surprising finding is that BERT-score and ROUGE-L tend to have better alignment with the human annotators than LLaMA and Cohere-based evaluation. However, the BLUE score shows to be the least reliable among the metrics, which is similar to findings in other NLG tasks (Mathur et al., 2020). We have performed further analysis looking into the correlations between the human evaluation categories which are available in the appendix.

[6]See Appendix for guidance of the Spearman's rank scale.

6 Discussion

Potential of LLMs for the task. Analyses using a wide range of automatic metrics and human evaluation, as presented in Section 5, show promising performance of LLMs on the recommendation generation task. Notably, both human and LLM-based evaluations—regardless of the model used—produced high scores, with human evaluators assigning slightly higher ratings, ranging from approximately 2.5 to 2.8 out of a maximum of 3.0. These results show the potential of state-of-the-art models to be utilised for more specialised domains to support the work of subject-matter experts. Further, the results from the human evaluation presented in Table 5 show a higher scoring for the *'relevance to the evidence'* versus *'relevance to the human-based recommendation'* criteria. This suggests that LLMs can be more suited for providing a different perspective of the problem versus simply matching the expert-created text.

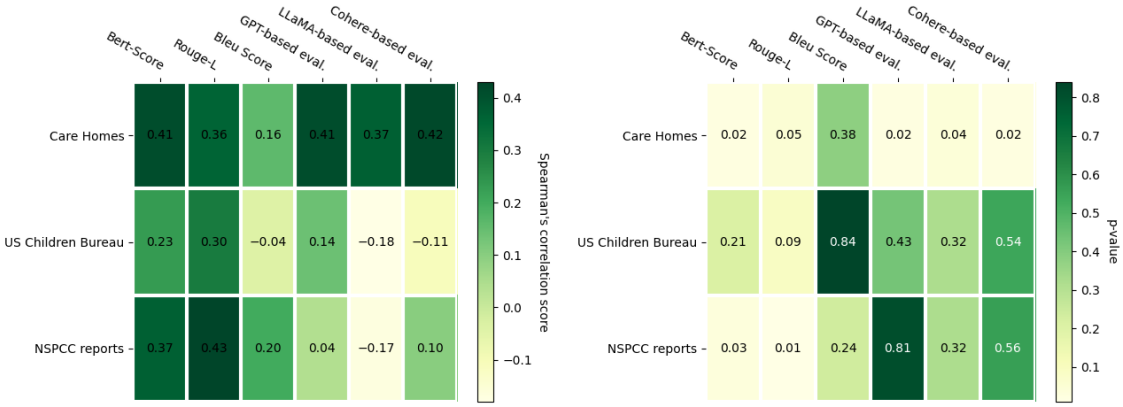

Figure 3: Spearman's rank correlation (left) and p-values (right) between manual evaluation and automated metrics-based evaluation across the three datasets where *'eval'* refers to evaluation, *'Care Homes'*, *'US Chidlren Bureau'* and *'NSPCC reports'* refer to the results from the human-based evaluation for the Care Homes dataset, US Children Bureau, and NSPCC datasets, respectively.

Hallucinations and Reliability. A major challenge in LLM-based text generation is the risk of hallucinations (Ji et al., 2023; Filippova, 2020), with solutions varying depending on the task and available resources. We note that addressing this issue is beyond the scope of this paper. However, our human-based evaluation approach helps identify discrepancies within the dataset. For example, Criterion (3) from the evaluation framework (see Section 4.2) assesses whether a generated recommendation is meaningfully related to the given evidence. The average score for this criterion exceeds 2.5 (on a 3-point scale) across all datasets, indicating strong relevance. Additionally, subject matter experts found a large proportion of the recommendations to be practically applicable to policy-making processes (Criterion (5), Section 4.2). These findings suggest a minimal presence of hallucinations in the generated content. Nonetheless, we believe that future research should include more rigorous analysis and the development of evaluation methods that can ensure higher dataset reliability.

Evaluation metrics for text generation. A comparison between the different automated metrics (see Section 5.1) and the correlation analysis between automated and human-based evaluation (Section 5.2) highlighted the unsuitability of traditional evaluation metrics such as BLEU for more complex NLG tasks such as recommendation generation. Further, LLM-based metrics and human-based evaluation showed similar satisfactory results suggesting good performance of text generation models for the given task. However, correlation analysis showed that no metric achieved high agreement with the human evaluators which suggests that when it comes to complex NLG tasks, we should not rely on a single metric. The relatively low scores from the inter-annotator agreement analysis illustrates further the complexity of the task, which proved challenging even for domain experts. This shows the need to develop more purpose-oriented metrics for text generation problems such as recommendation generation. In future, the study can be expanded by looking to incorporate more qualitative studies and metrics within the evaluation process.

7 Conclusions

This paper introduces the first comprehensive effort to leverage LLMs for the specialized NLG task of recommendation generation aimed at informing policy decisions and enhancing the work of public service agencies. We release a unified benchmark dataset for this task, PubRec-Bench, compiled from three distinct data sources. We evaluate three state-of-the-art models: GPT-4o, Cohere's Command R+, and LLaMA 3, using both LLM-based and human evaluations, which yield promising results. Human evaluators judged most generated recommendations as highly relevant to the provided evidence and consistently coherent and fluent in both structure and content. Additionally, subject matter experts rated the majority of outputs as actionable, meaning they offer practical, real-world utility. Finally, we provide a thorough analysis of evaluation methodologies, highlighting the need for more task- and purpose-specific metrics tailored to the demands of NLG in applied, real-world settings.

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 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 three datasets of a relatively small size. In the future, we plan to extend it by including reports from diverse sources. However, given the fact that these reports are usually written to reflect on serious crimes or problems within service delivery, we believe that the provided dataset is a good representative of the domain.

Ethical Considerations

The goal of our method is to facilitate rather than replace practitioners in the public sector in writing high quality recommendations. Specifically, we hope that LLM-generated recommendations can provide a different and useful aspect of the problem at hand and also facilitate more efficient decision-making for practitioners. Given the domain at hand, we believe that subject matter experts should take pivotal role in recommendation creation, but it is also important to find ways to utilise LLMs strengths to support the work of experts.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. Qameleon: Multilingual qa with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, 10.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian M Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *29th International Conference on Computational Linguistics (COLING 2022)*.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. pages 3816–3830.

Cătălina Goanță, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordás, and Gerasimos Spanakis. 2023. Regulation and nlp (regnlp): Taming large language models. pages 8712–8724.

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. *CoRR*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2022. Genie: Toward reproducible and standardized human evaluation for text generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11444–11458.

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.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Pilar López-Úbeda, Teodoro Martín-Noguerol, Carolina Díaz-Angulo, and Antonio Luna. 2024. Evaluation of large language models performance against humans for summarizing mri knee radiology reports: A feasibility study. *International Journal of Medical Informatics*, 187:105443.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

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.

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.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*.

Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Context-tuning: Learning contextualized prompts for natural language generation. *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6340–6354.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 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. pages 413–421.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 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. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075.

Qianqian Xie, Zhehengz Luo, Benyou Wang, and Sophia Ananiadou. 2023a. A survey for biomedical text summarization: From pre-trained to large language models. *arXiv preprint arXiv:2304.08763*.

Qianqian Xie, Prayag Tiwari, and Sophia Ananiadou. 2023b. Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. *IEEE journal of biomedical and health informatics*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

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. *International Conference on Learning Representations*.

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

A.1 Model parameters and Computational Budget

The model parameters we used for generating recommendations are as follows: (1) For GPT4-o and Cohere-based model we have used temperature of 0.7 and for LLaMA a temperature of 0.6. These are the default values recommended for these models. We used 7 hours of GPU budget and Nvidia RTX 4090 GPU.

A.2 Prompts

We conduct experiments with two prompts (see below). For the creation of *'Prompt 1'*, we followed examples provided by OpenAI and Meta. We also followed the design principles described in Reynolds and McDonell (2021) to create self-explanatory prompts that are intuitive and easy to use from the user's perspective. To create *'Prompt 2'*, we asked subject matter experts (see Section 4.2) to provide a description of the task to be included in the prompt.

Prompt 1 for generating recommendations

Provide a recommendation for improving agencies work and services related to children care and children services. The recommendation should reflect on the information given in the report:
Evidence:[Evidence]

Prompt 2 for generating recommendations

Based on a summary of evidence from this report, generate a concise recommendation with particular focus on what would improve or resolve the issues raised within the information. Please do not include context or rationale at this stage:
Evidence:[Evidence]

The prompt we use to conduct LLM-based evaluation is given below.

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:

A.3 Human-Based Evaluation

Figure 4 shows the instructions given to the annotators in order to perform the human-based evaluation. Table 6 shows the results for the criterion (5) Is the recommendation actionable?.

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

(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 is meaningful given the evidence.

(4) Relevance to the human-created recommendation measures the factual alignment between the two recommendations.

(5) Is the recommendation 'Actionable'? (yes/no): indicates if the recommendation has practical application and could be implemented as part of a policy.

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

Figure 4: Instructions for human evaluation.

Dataset)	zero-shot setting	one-shot setting
UK Care Homes	60%	55%
US Children's Bureau	58%	57%
NSPCC reports	60%	59%

Table 6: Results per dataset for criterion (5) Is the recommendation actionable?

The five annotators were selected through an interview process based on their subject-matter expertise. They were compensated at an hourly rate aligned with standard payment guidelines, as approved by JobShop and University guidelines.

A.4 Automatic-Based Evaluation Results

Figure 5 shows a comparison of the three LLM-based evaluation models for one-shot setting where the scores are averaged across the three datasets. This results confirm findings from analysis for zero-shot setting where GPT4 and Cohere-based models have given a higher score to their own outputs. This suggests a potential bias for these models

towards their own generations (Kocmi and Federmann, 2023) which shows the need for further research into how best to utilise these models for evaluation tasks.

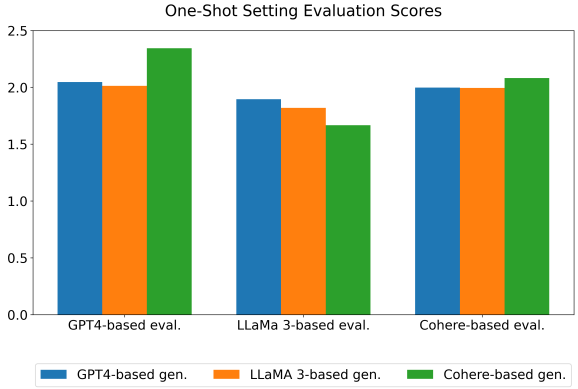

Figure 5: Comparison of LLM-based evaluation between recommendations generated with each model across the three datasets, where *'eval'* refers to evaluation.

A.5 Spearman's Rank Correlation

The Spearman's Rank Correlation Coefficient is a statistical measure of the strength of the relationship between two sets of data. A description of the strength of correlation is given in Table 7. The p-value is the probability of how likely it is that any observed correlation is due to chance. A p-value close to 1 suggests no correlation other than due to chance. If your p-value is close to 0, the observed correlation is unlikely to be due to chance.

Value of coefficient (pos. or neg.)	Meaning
0.00-0.19	A very weak correlation
0.20-0.39	A weak correlation
0.40-0.69	A moderate correlation
0.70-0.89	A strong correlation
0.90-1.00	A very strong correlation

Table 7: Interpretation of the Spearman's correlation coefficient.

A.6 Correlation Analysis.

Figure 6 represents complete correlation analysis across the human-based evaluation criteria per dataset.

The p-value for the majority of criteria is less than 0.05 which makes the majority of correlations statistically significant. Figure 6 shows that there are not significant trends of correlation relationships between the different criteria across the datasets. This suggests that despite belonging to the same domain/task, these datasets are quite diverse and require special attention of how to deal with their characteristics.

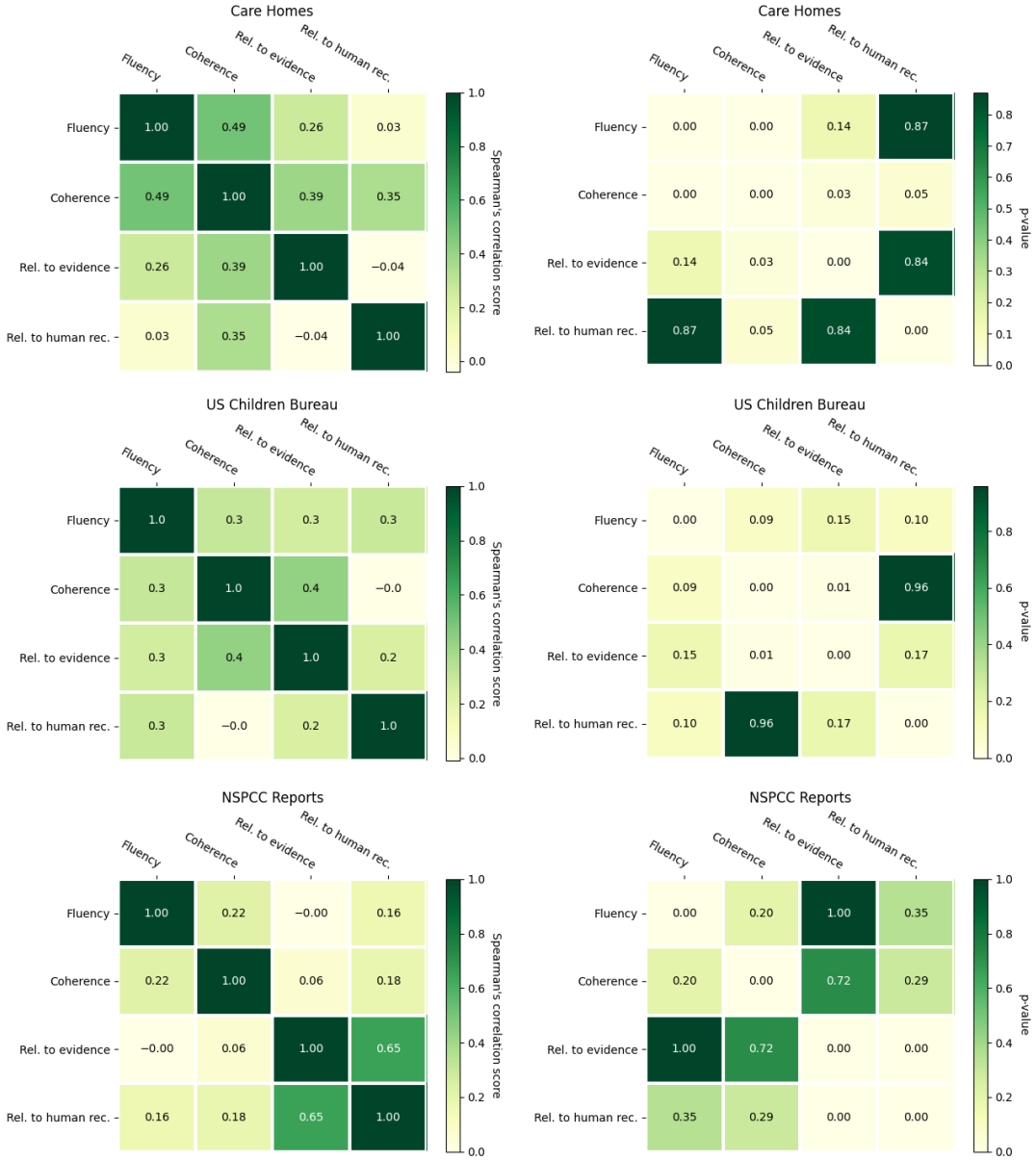

Figure 6: Spearman's rank correlation between across the criteria for the manual evaluation where *'Rel. to evidence'* refers to Relevance to the evidence, *'Rel. to human rec.'* reference to Relevance to the human-created recommendation.