Additional Research Questions to Be Included in the Camera-Ready Version

ANONYMOUS AUTHOR(S)

1 RQ6: Open-Source Large Language Model

The evaluation in the previous sections suggests that ChatGPT-based scores have a stronger correlation with human scores than lexical similarity-based metrics. In this section, we investigate whether we can replace the closed-source ChatGPT with open-source LLMs while keeping a strong correlation with human scores. To this end, in this section, we replace ChatGPT with DeepSeek-Coder [2], a well-known open-source large language model. Notably, we reuse all of the prompts for ChatGPT and employ the default settings of DeepSeek-Coder.

To have a quantitative and objective assessment of the correlation between DeepSeek-Coder-based scores and human scores, we compute the Spearman Rank Correlation between them as shown in Table 1. The Spearman Rank Correlation Coefficient is 0.49, much larger than those of the lexical similarity-based scores (e.g., 0.21 of METER, 0.22 of BLEU, and 0.25 of ROUGE). It is even slightly greater than that (0.47) of ChatGPT-based scores. The results may suggest that ChatGPT could be safely replaced with open-source LLMs like DeepSeek-Coder while keeping the strong correlation between LLM-based scores and human scores. It also suggests that the proposed approach can work well with various LLMs.

To visualize the consistency between DeepSeek-Coder-based scores and human scores, we present a heatmap of their overlap in Fig.1. From this figure, we observe that the data distribution is primarily concentrated along the diagonal, suggesting that DeepSeek-Coder-based scores are often aligned with human scores. We also use the Kolmogorov-Smirnov (K-S) statistic [1] to evaluate whether the distributions of DeepSeek-Coder-based scores significantly differ across reviews of different quality levels. Results are presented in Table 1. From this table, we observe that the K-S statistic of DeepSeek-Coder-based scores is very close to that of ChatGPT-based scores, suggesting that DeepSeek-Coder-based scores are comparable to ChatGPT-based scores in distinguishing high-quality reviews.

We conclude based on the preceding analysis that ChatGPT could be safely replaced with the state-of-the-art open-source LLMs concerning the proposed LLM-based scoring approach.

References

[1] Giovanni Fasano and Alberto Franceschini. 1987. A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society* 225, 1 (1987), 155–170.

[2] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming–The Rise of Code Intelligence. *arXiv preprint arXiv:2401.14196* (2024).

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

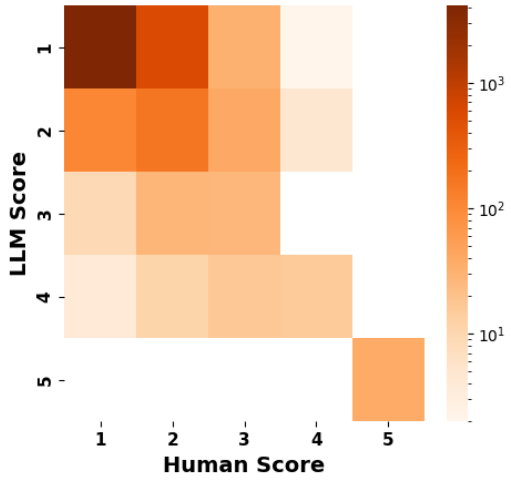

Fig. 1. Heatmap of Overlap Between LLM(deepSeek) Scores and Human Scores

Table 1. Kolmogorov-Smirnov Statistic Across Reviews of Varying Quality Levels

Quality Level	**2-Point**	**3-Point**	**4-Point**	**5-Point**
1-Point	BLEU: 0.26 Embedding: 0.55 ChatGPT: 0.55 DeepSeek: 0.53	BLEU:0.36 Embedding: 0.71 ChatGPT: 0.71 DeepSeek: 0.73	BLEU: 0.44 Embedding: 0.73 ChatGPT: 0.73 DeepSeek: 0.79	BLEU: 1.00 Embedding: 1.00 ChatGPT: 1.00 DeepSeek: 1.00
2-Point		BLEU:0.16 Embedding: 0.26 ChatGPT: 0.24 DeepSeek: 0.28	BLEU: 0.30 Embedding: 0.35 ChatGPT: 0.47 DeepSeek: 0.55	BLEU: 0.99 Embedding: 1.00 ChatGPT: 1.00 DeepSeek: 1.00
3-Point			BLEU: 0.20 Embedding: 0.22 ChatGPT: 0.28 DeepSeek: 0.33	BLEU: 1.00 Embedding: 1.00 LLM: 1.00 DeepSeek: 1.00
4-Point				BLEU: 0.97 Embedding: 1.00 ChatGPT: 1.00 DeepSeek: 1.00