Neither Valid Nor Reliable? Investigating the Use of LLMs as Judges

Evaluating natural language generation (NLG) systems remains a core challenge of natural language processing (NLP), further complicated by the rise of large language models (LLMs) that aims to be general-purpose. Recently, large language models as judges (LLJs) have emerged as a promising alternative to traditional metrics, but their validity remains underexplored. We argue that the current enthusiasm around LLJs may be premature, as their adoption has outpaced rigorous scrutiny of their reliability and validity as evaluators. Drawing on measurement theory from the social sciences, we identify and critically assess four core assumptions underlying the use of LLJs: their ability to act as proxies for human judgment, their capabilities as evaluators, their scalability, and their cost-effectiveness. We examine how each of these assumptions may be challenged by the inherent limitations of LLMs, LLJs, or current practices in NLG evaluation.

A1: LLMs as a Proxy for Human Judgment. Work on LLJs [2] has primarily validated their use through convergent validity and showed high correlation between LLJs judgement and human judgment. This approach assumes that a metric is valid if it correlates with an existing, already validated metric for the same construct. However, prior work [1] has revealed significant inconsistencies in how human judgments in NLG evaluation are being elicited and collected, casting doubt on their validity as a benchmark. Moreover, the literature on LLJs seems to reproduce and even exacerbate many of the same issues found in previous NLG evaluation research. These inconsistencies in both human and LLM judgment collection practices raise important concerns about the validity of LLJs.

A2: LLMs as Capable Evaluators. Inherent limitations in LLMs' capabilities may affect their validity and reliability as evaluators across four key dimensions: (1) *instructions adherence*: LLMs frequently rely on their own interpretations of evaluation criteria, rather than following the instructions provided in the prompts, particularly across popular quality criteria used in the NLG literature; (2) *explainability*: LLJs are often presented as a more interpretable alternative of traditional evaluation metrics, however, none of these studies examined the faithfulness of the generated explanations; (3) *robustness*: studies have shown that LLJs are susceptible to a wide range of biases [3] (e.g., position bias, verbosity bias, etc.) and to adversarial attacks which question their reliability as evaluators; and (4) *expertise*: an LLM performance on a given task impact its content validity for that same task. Interestingly, LLJs have been proposed as substitutes for humans in highly subjective and contested tasks, such as hate speech detection. These constructs are highly subjective, and relying on LLJs—who tend to yield higher inter-annotator agreement and present their own inherent biases—risks overlooking the valuable diversity found in human disagreement.

A3: LLMs as Scalable Evaluators. Beyond their "traditional" role in evaluation, LLJs are increasingly being used for model enhancement. They can assume various roles throughout the training pipeline, including data generation and annotation, reward modeling, and verification. While these applications have led to notable improvements in utility, the generalization of these performance gains remains to be rigorously validated, especially as such practices blur the boundary between training and testing. Considering that LLJs are mostly validated using publicly available benchmarks, this raises the issue of data contamination: several studies have shown evidence of memorization of popular benchmarks in various state-of-the-art LLMs. Furthermore, the issue of competitive benchmarking has gained increasing attention in recent years, particularly in light of the rapid advancement of capabilities and considering that benchmarks are both testing instrument and testing material. Studies have demonstrated how easily evaluation frameworks can be manipulated, and we argue that automatizing the pipeline can only facilitate such malpractices.

A4: LLMs as Cost-Effective Evaluators. The long-term implications and non-financial costs of LLJs adoption are largely ignored in the literature and rarely discussed in critical depth, despite being crucial to establishing the validity of the framework. For example, the popularity of LLJs as annotators, raises concerns about the future of crowdworkers—an already vulnerable population. Another element to consider is the potential environmental cost of LLJs: the environmental impact of large language models during inference remains considerable—whether in terms of energy consumption, carbon emissions, or water usage—and continues to grow as model sizes increase. Finally, LLJs are not exempt from the well-documented societal biases present in large language models, even if such biases have not yet been extensively studied in this context.

In this work, we have explored how various pitfalls in NLG evaluation practices and the inherent limitations of LLMs can impact LLJs as evaluators. We argue that fully realizing the potential of LLJs depends on our ability to critically and systematically address these challenges. When properly implemented, LLJs offer a valuable opportunity to advance NLG evaluation—whether by enabling more realistic, interactive, and long-term evaluation pipelines that better reflect real-world usage, or by alleviating the burden of problematic annotation tasks involving harmful or traumatic content. Therefore, leveraging LLJs effectively will require a careful balance: improving efficiency without disregarding their broader societal impact.

[1] David M. Howcroft et al. "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions". In: *Proceedings of the 13th ICNLG*. 2020, pp. 169–182.

[2] Tom Kocmi and Christian Federmann. "Large Language Models Are State-of-the-Art Evaluators of Translation Quality". In: *Proceedings of the 24th EAMT*. 2023, pp. 193–203.

[3] Haitao Li et al. *LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. 2024.