MDEval: Evaluating and Enhancing Markdown Awareness in Large Language Models

Zhongpu Chen [1] Yinfeng Liu [1] Long Shi [1] Zhi-Jie Wang [2] Xingyan Chen [1] Yu Zhao [1] Fuji Ren [3]

[1]Southwestern University of Finance and Economics [2]Chongqing University [3]University of Electronic Science and Technology of China

Motivation

Modern web chatbots using advanced LLMs often employ Markdown for better readability [1]. As shown in Figure 1, GPT-4o excels at producing well-structured Markdown output without explicit prompts, using headings, bolding, and lists to enhance long responses. While Claude-3.5-sonnet generally outperforms GPT-3.5-turbo, they're comparable in Markdown usage. This raises an important question: **How good are LLMs at Markdown output?**

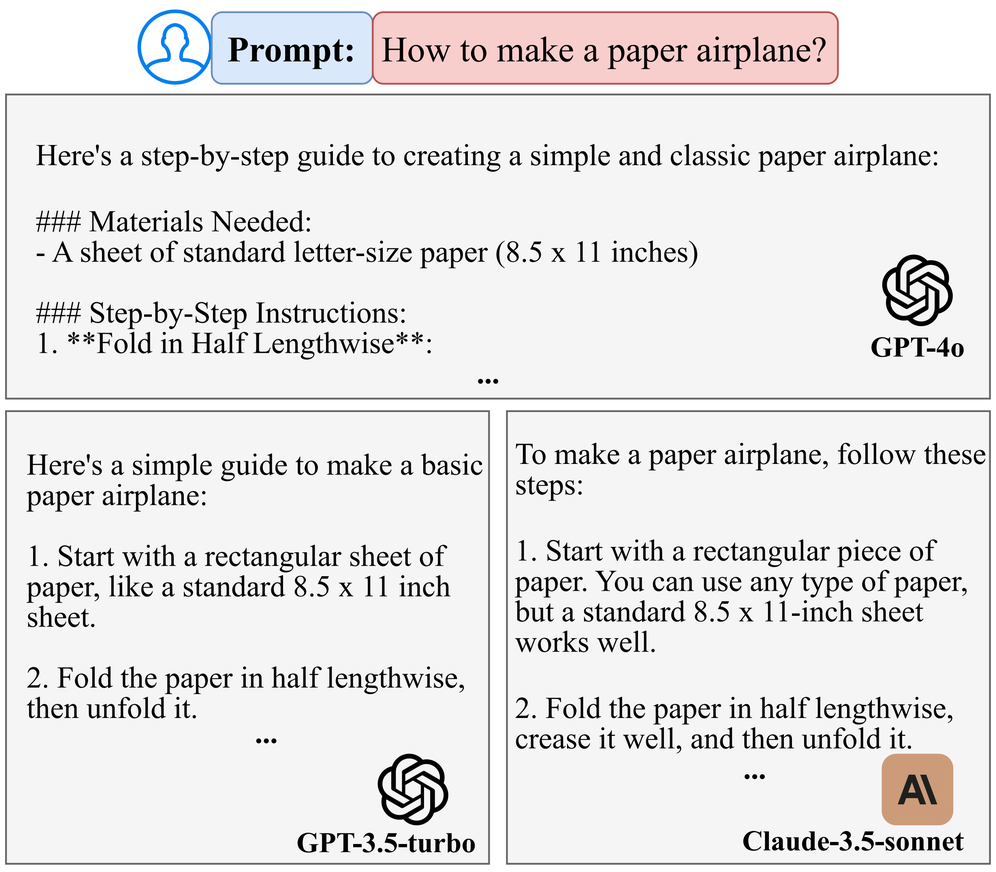

Figure 1. Differences of `Markdown Awareness` of LLMs under the same input prompt.

To this end, we introduce a novel LLM metric, termed `Markdown Awareness`, and then propose a comprehensive benchmark to evaluate the Markdown output capability of LLMs on a novel curated dataset with 20K instances covering 10 subjects in English and Chinese. Our results demonstrate that MDEval achieves a Spearman correlation of 0.791 and an accuracy of 84.1% with human, outperforming existing methods by a large margin. The code is available at `https://github.com/SWUFE-DB-Group/MDEval-Benchmark`.

Challenges

However, it is non-trivial to design a benchmark to evaluate the `Markdown Awareness` for LLMs due to three major challenges:

- **(C1) Insufficient Datasets**: Difficulty in defining single "expected" outputs when prioritizing style/structure over content in generated Q&A pairs.
- **(C2) Metric Validity**: Traditional metrics (e.g., BLEU [2]) require unavailable ground-truth (due to C1); LLM-based evaluation (e.g., G-Eval [3]) lacks stability and explainability.
- **(C3) Metric Quantification**: `Markdown Awareness` is *structure-oriented*, not *content-oriented*, and thus it is difficult to be quantitative.

Contributions

1. To the best of our knowledge, MDEval is the first benchmark to evaluate the quality of LLMs' Markdown output, and we proposed a novel structure-oriented metric, named `Markdown Awareness`.
2. MDEval provides a dataset with 20K instances covering 10 subjects in Chinese and English. And we reported `Markdown Awareness` performance across 9 mainstream LLMs based on the dataset.
3. The proposed `Markdown Awareness` is of validity by combining model-based generation tasks and statistical methods seamlessly in an evaluation pipeline, and we also showed that it is able to align with humans' preference.
4. We demonstrated that through fine-tuning over the dataset constructed above, less performant open-source models can achieve comparable performance to GPT-4o in terms of `Markdown Awareness`.

Methods

As illustrated in Figure 2, given a specific task (e.g, "How to make a paper airplane") and a target LLM (e.g., Llama), the initial output generated by the target LLM may lack proper Markdown formatting (Phase 1). To this end, MDEval subsequently invokes an advanced LLM (e.g., GPT-4o) to rewrite the initial response to generate a well-structured Markdown counterpart while preserving the original textual content (Phase 2). The rewritten response serves as a model-dependent reference because it should be constructed on-the-fly, rather than from a pre-collected dataset. Since `Markdown Awareness` is structure-oriented, MDEval further converts the two responses above into HTML contents and then extracts the HTML tags, respectively (Phase 3). Finally, the edit distance is computed between the extracted HTML tags, and a lower edit distance often indicates a superior `Markdown Awareness` (Phase 4).

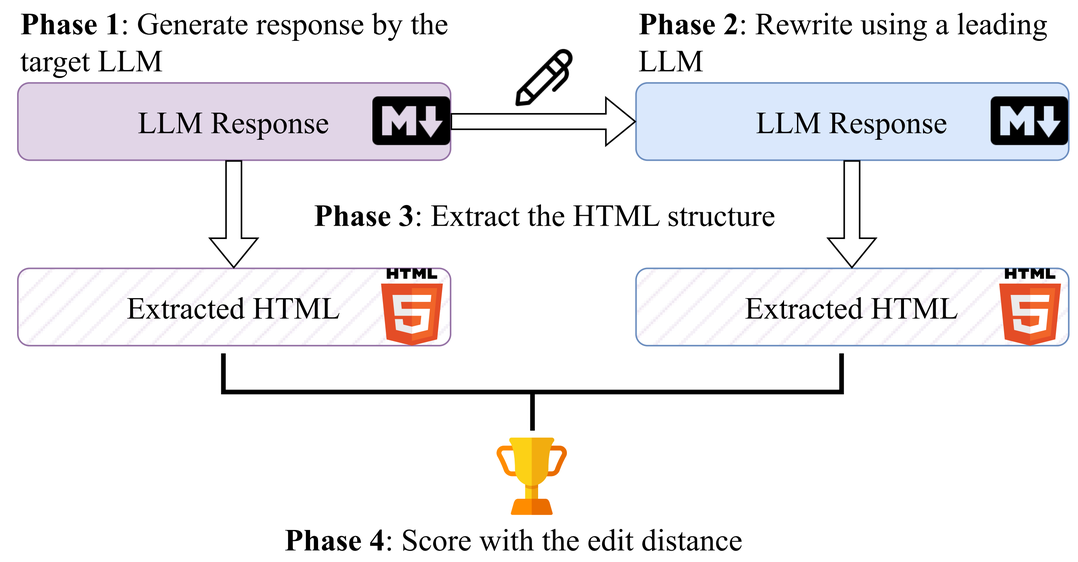

Figure 2. The overall framework of MDEval.

Inspired by Chatbot Arena [4], we also build a human alignment evaluation system using a pairwise comparison approach, and a test system can be publicly visited at `https://md-eval-human.pages.dev`.

Results

Table 1. `Markdown Awareness` Scores

Ranking	LLM	Score
#1	**Deepseek-v2-chat**	**0.946**
#2	GPT-4o	0.865
#3	GPT-4o-mini	0.830
#4	Gemini-1.5-pro	0.812
#5	Claude-3.5-sonnet	0.795
#6	Llama-13B	0.780
#7	Llama-2-13B	0.765
#8	Mistral-7B	0.745
#9	GPT-3.5-turbo	0.730

Table 2. Accuracy and Correlation

Method	Accuracy	Spearman	Pearson	Kendall
MDEval	**84.1%**	**0.791**	**0.844**	0.670
P-LLM	73.8%	0.779	0.783	**0.674**
R-LLM	70.7%	0.710	0.705	0.600
D-Rule	83.4%	0.757	0.685	0.625

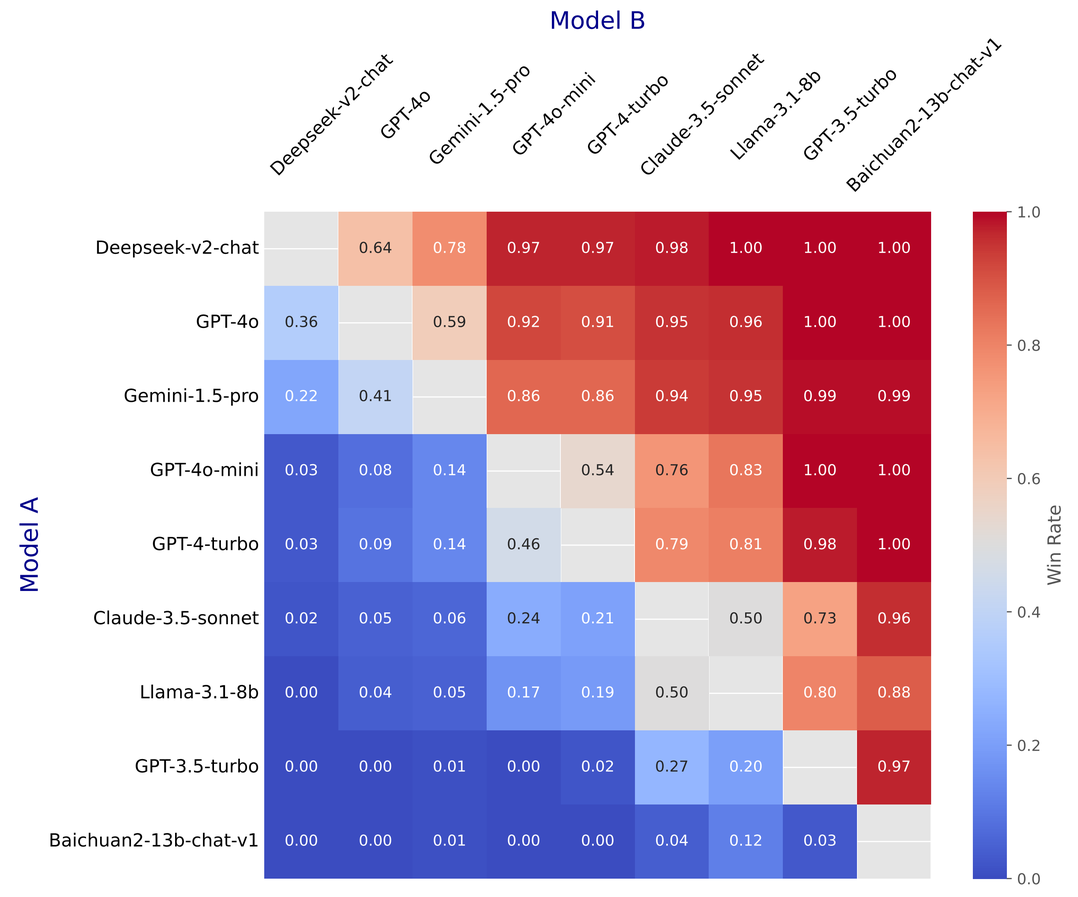

Figure 3. Pairwise win rate summary.

References

[1] Tensmeyer, D. et al. "Building Web Chatbots with Markdown." ACM WWW (2023).

[2] Papineni, K. et al. "BLEU: A Method for Automatic Evaluation of Machine Translation." ACL (2002).

[3] Liu, Y. et al. "G-Eval: Generalized Evaluation for Generation Models." EMNLP (2023).

[4] Chiang, L. et al. "Chatbot Arena: A Comprehensive Evaluation Platform." ICML (2024).

https://doi.org/10.1145/3696410.3714674 Proceedings of the ACM Web Conference 2025; April 28–May 2, 2025; Sydney, NSW, Australia. zpchen@swufe.edu.cn