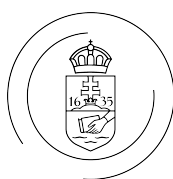


Investigating Bias in LLM Self-Evaluation

Thesis

Mathematics Expert in Data Analytics and Machine Learning



ELTE | FACULTY OF
SCIENCE

Eötvös Loránd University
Faculty of Science

Budapest, 2025

Contents

1	Introduction	3
1.1	A Brief Introduction to LLMs	3
1.2	LLM Evaluators, LLM-as-a-Judge	5
1.2.1	LLM-Judge Prompting Basics	6
1.2.2	Metrics	7
1.2.3	AutoCalibrate: Let the LLM Define the Criteria	8
	References	9

1 Introduction

1.1 A Brief Introduction to LLMs

A language model is a machine learning model designed to perform a wide range of tasks that involve natural language processing (NLP), including text summarization, translation, sentiment analysis, spam detection, content moderation, text generation, etc.

Significant advancements in deep learning [1, 2, 3] led to the emergence of **large language models** (LLMs) — particularly generative LLMs — which in the early 2020s became commercialized and widely adopted in both industry and popular discourse.

A generative LLM is a model which has a parameter count on the order of billions or more (hence "large"), and predicts the conditional probability [4]

$$P(w_m|w_0, \dots, w_{m-1}) \tag{1}$$

where $m \in \mathbb{N}$, w_0 is a special start symbol, and w_k is the k -th token (for $1 \leq k \leq m$) in a sequence of tokens that form a piece of text in some language, be it a natural language or a formal one like programming languages. The interpretation of the tokens depends on the exact tokenization strategy used, which may define tokens as words, word pieces, n-grams, or individual characters, and spaces, punctuation marks, etc.

Text generation is then an autoregressive process where given a sequence of tokens as a prefix — known as the **prompt** — the model estimates the probability distribution of the next token, takes a sample from that distribution, appends it to the sequence, and repeats the process with the extended sequence until certain stopping conditions are met.

The sampling can usually be controlled via numerical parameters; a frequently used one is the **temperature** [5]: the closer it is to 0, the more the sampling will lean toward the most probable token — making the algorithm more deterministic —,

while higher values increase the randomization, making the generated text feel more *creative* until, above a certain threshold, it becomes incoherent and semantically meaningless. If $v \in \mathbb{N}$ denotes the number of all possible tokens available for the model (vocabulary size), and $\mathbf{s} \in \mathbb{R}^v$ is an output vector of the model assigning a score to each token as the continuation of a given input, then the distribution for the sampling, with respect to the temperature $T \in \mathbb{R}$ can be calculated via the softmax function:

$$\text{softmax}\left(\frac{1}{T}\mathbf{s}\right) = \left[\frac{\exp(\frac{s_i}{T})}{\sum_{j=1}^v \exp(\frac{s_j}{T})}\right]_{i=1}^v \quad (2)$$

In practical implementations, if T is sufficiently close or exactly equal to 0, then the sampling is usually replaced with the deterministic argmax function in order to preserve numerical stability. Non-zero T values control the flatness of the distribution, leading to the aforementioned behavior.

With sufficiently large model complexity and training corpora size and diversity, LLMs start to exhibit capabilities which rival that of top performer humans in a broad class of problems [2, 3]. The versatility of the models is often utilized in a setting where the prompt is composed of two parts, each consisting of instructions given in natural language:

- the **system prompt** can instruct the model to behave in a certain way, for example, to act like a helpful AI assistant, an expert in a domain, or to generate its texts in the style of fictional 18th-century Caribbean pirates,
- and the **user prompt** which describes a task to be carried out by the model, ranging from text translation or summarization to solving complex programming problems or pointing out business risks in legal documents, and more.

Generative models with sufficient generalization capabilities can predict likely continuations of such prompts with such high accuracy that as an emergent phenomenon, the generated text will often contain an actual solution to the proposed

problem. This instruction-following paradigm enables models to perform **few-shot learning** [2] or even **zero-shot learning** by interpreting tasks directly from the natural language description, based on just a few or zero examples, respectively, without specific training or fine-tuning.

The problem solving performance of LLMs can be improved further by prompt engineering techniques like **chain-of-thought** prompting [6], where the model is provided with step-by-step example solutions to related problems in the prompt, encouraging it to also articulate intermediate reasoning steps before arriving at its final answer. It is worth emphasizing that — recalling equation 1 and the autoregressive text generation process — the chain-of-thought is only effective if it is placed *before* the final answer.

1.2 LLM Evaluators, LLM-as-a-Judge

The continuing development of LLMs and their integration into more and more systems to support a growing number of use cases necessitates regular measurement of their capabilities and monitoring their alignment with human preferences.

While evaluating the quality of LLM-generated text by utilizing human labor does not scale well, may suffer from human error or subjective personal preference bias, and can be expensive, traditional algorithmic metrics which often rely on surface-level similarities to reference examples (like BLEU for machine translation [7] or ROUGE for summarization and translation [8]), often fall short of achieving acceptable correlation levels with human judgement.

In recent years, in order to overcome these problems, the **LLM-based evaluation** or **LLM-as-a-judge** paradigm has been proposed [9, 10, 11, 12], where — taking advantage of the instruction following and the zero-shot and few-shot learning capabilities of LLMs — a model is instructed to act as a fair judge and generate a quality assessment for a piece of generated text either in the form of a single score, or one accompanied by an explanation or a list of problems. An advantage of the latter approach — besides easier interpretability — is that enumerating evidences before giving a final result can influence the score via the autoregressive generation

process, similarly to the improvements achieved by making large models include a chain-of-thought breakdown of complex problems before the final answer.

1.2.1 LLM-Judge Prompting Basics

There are numerous strategies to implement LLM-judges in practice [13], but a robust LLM-judge prompt usually includes the following elements:

- **Instructions** which clearly specify the evaluation task.
- Evaluation **aspects**, e.g. clarity, consistency, coherence, factuality, fluency, grammaticality, informativeness, structure, understandability, etc.
- Scoring **criteria** to specify the definitions for each score or score range.
- **Output format** specification so that the output of the judge can be programmatically parsed and interpreted.
- The **sample** itself to be evaluated or a pair of samples to be compared against each other.

Depending on the chosen evaluation strategy and aspect, additional elements may be included as well:

- Human-annotated **example** samples and their associated scores in few-shot evaluation scenarios.
- A **reference** answer for comparison with the evaluated sample, e.g. a human expert made translation, text summary, trivia answer, etc.
- The **source** data from which the evaluation sample was derived. (The original text to be translated, summarized, or the question to be answered, etc.)
- **Guidelines**, for example to help an LLM resolve the confusion that may arise in reference answer-based evaluations where some of the provided reference answers seem to contradict the model’s own knowledge, e.g. ”*Don’t worry*

about factuality with respect to the real world, just judge the example based on what you see. No need to overthink this task, it really comes down to just soft matching.” [14].

The judging model may also be fine-tuned using evaluation data constructed either manually or with the assistance of advanced models like GPT-4.

1.2.2 Metrics

Popular choices for scoring strategy include:

- **Binary classification:** the judge is expected to provide a ”yes” vs. ”no”, or a 0 vs. 1 verdict.
- **Pairwise comparison:** the judge is given two candidate answers, and has to select the one that is a better fit for the evaluation criteria, or declare a tie between them.
- **Multiclass classification:** the judge has to place the candidate on a discrete scale, usually between 1 and 5 points where 1 is the worst and 5 is the best.
- **Likert-style:** the judge has to rank the candidate answer along multiple dimensions using discrete scores, usually between 1 and 3 points where a higher score is better, then provide an overall 1 to 5 rating based on these scores.
- **Continuous score:** the candidate answer is scored with a number between 0 and 100.

If the judge LLM’s interface makes the raw token probabilities available, then they can be used for refining discrete scores and making them into continuous ones by taking the sum of the discrete score values weighted by the probabilities of the respective tokens, as seen in the G-EVAL framework [12]:

$$\text{score} = \sum_{i=1}^n p(s_i) \times s_i$$

where $S = \{s_1, s_2, \dots, s_n\}$ is the set of scores predefined in the prompt, and $p(s_i)$ are the probabilities of the respective tokens for the score values, as calculated by the model.

Another way to turn a discrete score into a continuous one is used in the GEMBA metric [15] for assessing translation quality: it requires the candidate answer to be dividable into smaller segments which are then evaluated one-by-one, and the resulting scores are averaged.

1.2.3 AutoCalibrate: Let the LLM Define the Criteria

Constructing the prompt template for a consistent, reproducible, and unbiased LLM-judge which also aligns well with human preferences is usually an iterative process, where the prompt is refined step-by-step until the LLM-judge can reliably produce evaluations that are sufficiently close to a set of human-labeled examples. A crucial part in this process is to come up with well-defined evaluation criteria.

The AUTOCALIBRATE method [16] attempts to automate this process by utilizing a sufficiently large model:

- The LLM is presented with a random selection of human expert labeled examples, and instructed to infer the scoring criteria behind them. This is repeated multiple times with different samples, producing a set of draft candidate criteria.
- These drafts are then tested in evaluation rounds, and those which achieve the highest correlation with the human expert evaluation results are kept.
- Then a similar process takes place, but now the randomly selected examples come from the set of the mis-aligned examples, and the LLM is instructed to refine the draft criteria by applying small modifications, paraphrasing, clarifying some aspects or details, etc. instead of coming up with new ones from scratch.
- Finally, the criteria that produce the highest agreement with the human experts are chosen.

References

- [1] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). DOI: <https://doi.org/10.48550/arXiv.1706.03762>.
- [2] OpenAI. “Language Models are Few-Shot Learners”. In: *CoRR* abs/2005.14165 (2020). DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- [3] OpenAI. “GPT-4 Technical Report”. In: (2023). DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
- [4] Tong Xiao and Jingbo Zhu. *Foundations of Large Language Models*. 2025. DOI: <https://doi.org/10.48550/arXiv.2501.09223>.
- [5] Enrique Manjavacas et al. “Synthetic Literature: Writing Science Fiction in a Co-Creative Process”. In: *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. Santiago de Compostela, Spain: Association for Computational Linguistics, Sept. 2017, pp. 29–37. DOI: <https://doi.org/10.18653/v1/W17-3904>.
- [6] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *CoRR* abs/2201.11903 (2022). DOI: <https://doi.org/10.48550/arXiv.2201.11903>.
- [7] Kishore Papineni et al. *BLEU: a method for automatic evaluation of machine translation*. 2001. DOI: <https://doi.org/10.3115/2F1073083.1073135>.
- [8] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013/>.
- [9] Jinlan Fu et al. *GPTScore: Evaluate as You Desire*. 2023. DOI: <https://doi.org/10.48550/arXiv.2302.04166>. arXiv: 2302.04166 [cs.CL].
- [10] Jiaan Wang et al. *Is ChatGPT a Good NLG Evaluator? A Preliminary Study*. 2023. DOI: <https://doi.org/10.48550/arXiv.2303.04048>. arXiv: 2303.04048 [cs.CL].

- [11] Yi Chen et al. *Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study*. 2023. DOI: <https://doi.org/10.48550/arXiv.2304.00723>. arXiv: 2304.00723 [cs.CL].
- [12] Yang Liu et al. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. 2023. DOI: <https://doi.org/10.48550/arXiv.2303.16634>. arXiv: 2303.16634 [cs.CL].
- [13] Zhen Li et al. *Leveraging Large Language Models for NLG Evaluation: Advances and Challenges*. 2024. DOI: <https://doi.org/10.48550/arXiv.2401.07103>. arXiv: 2401.07103 [cs.CL].
- [14] Pat Verga et al. *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*. 2024. DOI: <https://doi.org/10.48550/arXiv.2404.18796>. arXiv: 2404.18796 [cs.CL].
- [15] Tom Kocmi and Christian Federmann. *Large Language Models Are State-of-the-Art Evaluators of Translation Quality*. 2023. DOI: <https://doi.org/10.48550/arXiv.2302.14520>. arXiv: 2302.14520 [cs.CL].
- [16] Yuxuan Liu et al. *Calibrating LLM-Based Evaluator*. 2023. DOI: <https://doi.org/10.48550/arXiv.2309.13308>. arXiv: 2309.13308 [cs.CL].
- [17] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. DOI: <https://doi.org/10.48550/arXiv.2306.05685>. arXiv: 2306.05685 [cs.CL].