Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00561

Generation, given the original question q and the answer Y to be evaluated, they first prompt the LLM to generate n possible questions q_i for Y. Then, the relevance of Y is represented by the average similarity between q_i and q, denoted as $\sum_{i=1}^{n} \text{sim}(q_i, q)$, where $\text{sim}(q_i, q)$ refers to the cosine similarity of the embeddings of q_i and q. The embedding is generated by OpenAI *text-embedding-ada-002*, which can efficiently convert text into a 1536-dimensional vector, capturing semantic information and ensuring that similar texts are positioned close to each other in the vector space. Furthermore, Sheng et al. (2024) developed a more sophisticated method based on embeddings from the open-source decoder-only LLM, utilizing Principal Component Analysis to adapt it for both pointwise scoring and pairwise comparison.

2.2 Probability-based Metrics

To better utilize the knowledge inherent in language models, probability-based methods like BARTScore formulate text generation evaluation as conditional probability comparison, positing that the better the quality of the target text, the higher the likelihood that models should be able to generate it. Recently, GPTScore (Fu et al. 2023a) has established tailored evaluation templates for each aspect to effectively guide multiple LLMs for NLG evaluation, including GPT3 (Brown et al. 2020), OPT (Zhang et al. 2022), and FLAN (Chung et al. 2022). The core idea of GPTScore is that a good generative language model is more likely to assign higher probabilities to high-quality text generated in response to a given instruction and context. Specifically, given a generative large language model θ, context information X (such as a source document), output text $Y = \{y_1, y_2, \ldots, y_m\}$ containing m tokens to be evaluated, and instruction I that specifies the requirement for the LLMs to generate text that can flexibly correspond to different evaluation aspects (e.g., *generating a factually consistent summary* for the aspect of consistency), GPTScore is defined as:

$$\text{GPTScore}(X, Y, I, \theta) = \sum_{i=1}^{m} \log P(y_i | y_{<i}, X, I, \theta)$$

Similarly, Murugadoss et al. (2024) score the task output Y to be evaluated by its perplexity under the corresponding large language model θ, given only the task context X. They believe this approach is unbiased by prompts, which transparently measures alignment with model training data. Furthermore, such methods have also been applied to the hallucination detection of the LLM-generated text (Varshney et al. 2023) with three different attempts for calculating the probability score.

On the other hand, some works leverage the variation in probabilities under changed conditions as the evaluation metric. FFLM (Jia et al. 2023) proposes to evaluate the faithfulness of the target text by calculating a combination of probability changes based on the intuition that the generation probability of a given text segment increases when more consistent information is provided, and vice versa. Similarly, DELTASCORE (Xie et al. 2023) measures the quality of different story aspects according to the likelihood difference between pre- and post-perturbation states with LLMs including GPT-3.5 (text-davinci-003) that provide logits. They believe that the sensitivity to specific perturbations indicates the quality of related aspects, and their experiments demonstrate the effectiveness of their approach.