A Human Assessment of Reference-Free and Reference-Based Evaluation Approaches in the HR Domain

Rajna Fani
Technical University Munich
rajna.fani@tum.de

Abstract

Addressing the complex challenge of Natural Language Generation (NLG) evaluation, this research embarks on an exploration within the Human Resources (HR) domain, specifically through an HR chatbot use case. It contrasts state-of-the-art, reference-free evaluation metrics against traditional reference-based metrics to discern a deeper understanding of text quality. Incorporating human evaluation for a comprehensive comparison, a correlation analysis between these metrics is conducted to determine the most efficacious evaluation method. In the evaluation of the HR Q&A Chatbot use case across three models (LongT5, GPT3.5, GPT4), employing 5 different evaluation metrics, the superior performance was consistently demonstrated by the GPT-4 model. Additionally, through expert analysis, we infer that reference-free evaluation metrics such as G-Eval and Prometheus demonstrate reliability closely aligned with that of human evaluation.

1 Introduction

In the era of Large Language Models (LLMs), assessing the quality of generated text presents an ongoing challenge. This study explores the effectiveness of reference-free metrics in evaluating text quality produced by advanced language models, comparing them with traditional evaluation methods. Our research finds its practical application in addressing prolonged waiting times for employees seeking information from the Human Resources department through SAP HR Chatbots.

We investigated the structure of the HR Q&A Chatbot across three distinct models: OpenAI's LLMs GPT-3.5-turbo, GPT-4, and the Language Model LongT5 (Guo et al., 2021), aiming to determine the most effective model for HR applications to achieve the goal of covering 30% of the HR tickets with the Chatbot application. The research evaluates these two approaches, the Fine-tuned Language Model (LM) Approach and the LLM-Powered Approach, using a question-answering dataset that includes FAQs and user utterances from Chatbot logs to gauge generative model performance.

Through a thorough analysis blending quantitative and qualitative methods, we seek to assess the effectiveness of automated metrics, leading to an investigation of the reliability of automatic metrics when compared to human evaluations by domain experts. Subsequently, we delve into newer metrics showing potential in NLG, exploring their comparative value against traditional ones. Our goal is to determine if reference-free evaluation metrics, particularly those utilizing advanced language models, provide more dependable assessments of generative model performance compared to traditional reference-based metrics.

Through human evaluation and various metrics, we identify new state-of-the-art evaluation methods for NLG, particularly within a HR Chatbot Use Case. We implemented and assessed a spectrum of metrics to provide a comprehensive evaluation framework.

Reference-based Metrics:

1. N-gram based Metrics: Traditional metrics like BLEU (Papineni et al., 2002a) and ROUGE (Lin, 2004a) were utilized for their simplicity and widespread adoption in the evaluation of text similarity to reference outputs.

2. Embedding-based Metrics: BERTScore (Zhang et al., 2019), an embedding-based metric that evaluates the semantic similarity between the generated text and reference texts.

Reference-free Metrics:

1. Prompt-based Metric: G-Eval (Liu et al., 2023) represents an innovative approach to NLG evaluation by leveraging the capabilities of large language models through carefully designed prompts.

2. Tuning-based Metric: Prometheus (Kim et al., 2023) extends the potential of reference-free evaluation by fine-tuning language models on labeled evaluation data.

Each of these metrics was rigorously compared to human evaluations conducted by domain experts within the HR field.

2 Related Work

Evaluating Natural Language Generation (NLG) systems remains a challenge due to the multifaceted nature of language and the diverse applications of NLG technologies. Traditional metrics like BLEU (Papineni et al., 2002b) and ROUGE (Lin, 2004b)have been widely used due to their simplicity and efficiency. However, these metrics often fail to capture the nuanced understanding of language quality, coherence, and relevance required in more sophisticated NLG applications, such as dialog systems, story generation, and summarization (Wei et al., 2021), (Reiter, 2018).

BERTScore, introduced by (Zhang et al., 2019), has gained widespread use across a variety of NLG tasks, including Text Summarization (Deutsch and Roth, 2021), and Dialogue Systems (Wei et al., 2021). However, task-agnostic metrics, despite their broad applicability, have shown only weak correlation with human judgment (Novikova et al., 2017).

Recent advancements in evaluation methodologies, such as the development of reference-free metrics, seek to address the shortcomings of traditional (Gao et al., 2024). These new metrics, including G-EVAL (Liu et al., 2023) and Prometheus (Kim et al., 2023) assess the quality of generated text based on its intrinsic properties rather than comparison to a reference text (Gao et al., 2024). This approach is particularly valuable for applications where the generation of reference texts is impractical or where valid outputs are highly diverse. Large pre-trained language models (LLMs) like the OpenAI Models have further propelled these innovations, enabling more sophisticated evaluation tools that show a higher correlation with human judgments (Li et al., 2024). In addressing these challenges, our goal is to refine and expand upon current methodologies in NLG evaluation, ensuring that future frameworks can more accurately and comprehensively reflect the nuanced complexities and contextual diversities intrinsic to generated texts across a spectrum of NLG applications.

3 Corpus

The dataset used in the development of the HR chatbot was compiled using the company's internal HR policies with the help of domain experts. While each sample consisted of a Question, Answer, and Context triplet, additional metadata such as the user's region, company, employment status, and applicable company policies was also included. A snippet of such a sample is shown in Table 1. The dataset was compiled using two separate sources to have a mix of a gold dataset (FAQ dataset) and real-life noisy data (UT dataset). Both datasets follow the same structure and differences exist in the distribution of the questions.

We extracted all unique HR articles to form a knowledge base for answering new user questions. Additionally, an evaluation set of 6k samples was used to evaluate both the retriever and the chatbot as a whole.

DATA TRIPLET
Question: How can I apply for half a day of holiday?
Answer: Unfortunately, vacation days in your country can only be taken as full days.
Context: {Relevant Article}

META DATA
User Role: Employee
Name of KBA: Vacation
Company Name: {Company Name}
Company Code: {Company Code}
Region: {Region}
Country Code: {Country Code}
FAQ Category: {FAQ Category}
Process ID: {Process ID}
Service ID: {Process ID}

Table 1: HR Dataset Sample

3.1 Dataset Collection

FAQ Dataset N≈48k: This is a collection of potential questions, along with their corresponding articles and gold-standard answers. It is carefully created and curated by domain experts based on the company's internal policies.

UT Dataset (N≈41k): This is a collection of real user utterances (UT) gathered from previous iterations of the chatbot. Inspired by a semi-supervised learning approach, a simplistic text-matching approach was implemented, that mapped each user query to a question from the FAQ dataset. The chatbot logs from this development cycle were inspected and corrected by the domain experts.

4 Methodology

Our objective was to implement and evaluate completely new solutions for the retriever and NLG module of the RAG framework with the help of domain experts, improving the baseline version of the chatbot. An illustration of the RAG pipeline of the chatbot including the parts with human-in-the-loop can be observed in Figure 1.

In the NLG module, the fine-tuned Long-T5 model was replaced with OpenAI's more capable Large Language Models ChatGPT and GPT-4. These models leverage their advanced language generation capabilities and offer great versatility of their responses through flexible instruction prompting for varying requirements, instead of relying on fixed responses of a fine-tuned smaller model. The answers from the most optimized version of RAG pipeline were used for the evaluation of the respective models.

4.1 Baseline Models for Chatbot Evaluation

This section provides an introduction to the baseline models and an overview of the dataset employed in our study. It is important to acknowledge that the development and implementation of the Chatbot Pipeline were conducted by fellow students. I actively collaborated with these individuals, offering insights and staying informed about model improvements as we worked together.

4.1.1 LongT5 (Fine-tuning driven)

For evaluation, we primarily relied on the LongT5 model, which had already been fine-tuned with the SAP HR Dataset. This model was fine-tuned on a combination of the FAQ dataset and UT dataset for a generative question-answering task. To limit computational complexity, the model was filtered to an maximum input length of 7168 tokens and would require both question and corresponding context as input so it generates the answer.

During the model evaluation process, our goal was to generate random responses to presented questions, so the HR experts could evaluate the generated answers' performance. However, a significant challenge emerged when the HR department provided an updated dataset, while the LongT5-7168 model had been trained on an older version. Due to resource and time constraints, retraining the model with the new data was not possible.

This posed a dilemma: while the new Large Language Models (LLMs) could be fine-tuned using the latest HR dataset, the LongT5 remained aligned with the previous dataset. To address this issue, we extracted questions from the LongT5 model's test set and identified common questions shared with the new dataset. These overlapping questions formed the basis of our evaluation, ensuring a consistent and equitable assessment of the model's performance.

4.1.2 OpenAI Models (Prompt driven)

Advancements in Large Language Models (LLMs) have opened up new possibilities for exploration within the HR chatbot domain. To evaluate the potential benefits of an LLM-based HR chatbot, we employed OpenAI's GPT-3.5-turbo and GPT-4 models.

Extensive prompt engineering was conducted from the fellow student to tailor the responses of the LLMs to the company's requirements for an HR chatbot. This process included our qualitative analysis and multiple small evaluations from 10-100 sample responses by the company's HR experts. We analyzed feedback from these evaluation runs and addressed the main issues in the next iteration of the process. This continued until the responses of the LLM complied with the requirements in virtually all tested cases. These models were fine-tuned using the latest SAP HR dataset, ensuring they were updated with the most current data available. The final prompt used in our chatbot is shown in Table 4.

For a fair comparison with the previously implemented LongT5 model, we presented the same set of overlapping questions from the LongT5 evaluation phase to both GPT-3.5-turbo and GPT-4. This method allowed us to directly compare the answers generated by the new LLM-based chatbots with those from the LM-based LongT5, ensuring a level playing field for performance assessment.

4.2 Evaluation Framework

In our analysis, we utilize reference-based evaluation metrics including BERTScore (Zhang et al., 2019), ROUGE (Lin, 2004a), and BLEU (Papineni et al., 2002a). Additionally, we investigate the use of Large Language Models (LLMs) as evaluators. To evaluate the effectiveness of these automated metrics, we incorporate domain experts in a human-in-the-loop approach.

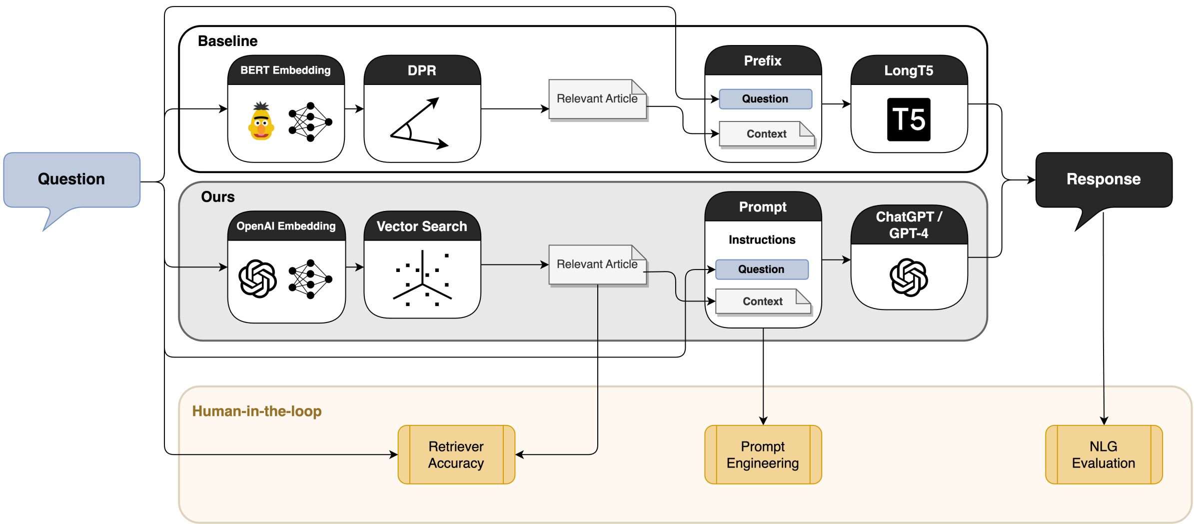

Figure 1: Block diagram of the methodology introduced in our paper, illustrating baseline and Open AI models, highlighting the role of the human-in-the-loop during development

4.2.1 Human Evaluation Setup

The human assessment phase of our study played a vital role, especially in comparing outcomes using different metrics. Focused on the HR Domain, our evaluators, all HR experts, brought a high level of precision and insight to the evaluation process. The approach employed in our study was extrinsic(van der Lee et al., 2021) due to its focus on evaluating how the text impacts within the HR domain. This method required significant resources but greatly enriched our analysis with expert perspectives. The primar goal was to have at least two HR domain experts as evaluators for unbiased evaluation (Ethayarajh and Jurafsky, 2022), but because of resources constraints, only one domain expert helped us evaluating 100 samples across the three previously mentioned models.

Criteria used for evaluating NLG systems The evaluation was carried out utilizing a 5-point Likert with a score between 1-5 (Likert, 1932) scale (Hämäläinen and Alnajjar, 2021). The criteria used for the evaluation framework was justified through the comprehensive survey (Liang and Li, 2021). This survey emphasizes these aspects as essential for evaluating linguistic quality, context appropriateness, user experience, and the human-likeness of chatbot responses. Initially, the selected criteria were:

1. Readability: This criterion assesses how easily the response can be understood.

2. Relevance: This criterion assess if the response connects well with the context of the question.

3. Truthfulness: This criterion evaluates the factual accuracy and reliability of each response. It assesses if the information is true and if it's missing any important details.

4. Naturalness: This criterion measures how closely the generated text resembles human-like speech or writing, focusing on fluency, coherence, and the appropriateness of expressions and style.

Experts Assessment Following feedback from HR Domain Experts on what they found most beneficial, we added *usability* as a criterion to evaluate the usefulness and practicality of responses, where high usability scores reflect clarity and the provision of actionable information. Following the initial iteration of samples, we chose to exclude Naturalness from the evaluation criteria in the final batch, as it was considered irrelevant to the HR Use Case and our ultimate objective of assessing the Chatbot's effectiveness.

Apart from manually curating the collected dataset, the domain experts also evaluated the performance of the retriever by verifying the correctness of the retrieved articles. They verified the accuracy of matched questions, contextual information (KBA), and correct answers, providing detailed feedback to ensure the integrity and relevance of our findings. The input from the HR Domain Experts made sure our evaluation was thorough and trustworthy, protecting sensitive information.

4.2.2 Reference-based metrics

In evaluating the effectiveness of reference-based metrics, we examine two distinct categories: N-gram based metrics and embedding-based metrics.

N-gram based metrics N-gram based metrics, such as BLEU (Bilingual Evaluation Understudy)(Papineni et al., 2002a) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(Lin, 2004a), assess the similarity between the generated response and the ground truth answer by analyzing the overlap of n-grams, with higher scores indicating superior performance. These metrics have been widely adopted in natural language generation (NLG) tasks due to their simplicity and effectiveness in capturing linguistic quality. BLEU, in particular, has been extensively used in machine translation evaluation and has shown strong correlations with human judgment in various studies (Papineni, 2002),(Mathur et al., 2020). Similarly, ROUGE has been favored for its ability to evaluate the quality of automatic summaries (Lin, 2004a). Recent studies have demonstrated that over 60% of NLG papers rely solely on ROUGE or BLEU for system evaluation (Kasai et al., 2022).

Embedding-based Metrics Embedding-based metrics, such as BERTScore (Zhang et al., 2019), leverage deep contextual embeddings from language models like BERT to assess the semantic similarity between generated and reference texts. This approach offers a nuanced evaluation of text quality, focusing on semantic rather than surface-level similarity. BERTScore, introduced by (Zhang et al., 2019), outperforms traditional metrics by aligning more closely with human judgment, as it accounts for the contextual usage of words. BERTScore's capability to accurately reflect text quality makes it an ideal choice for assessing chatbot responses in the HR domain, where semantic precision and relevance are crucial.

4.2.3 Reference-free metrics

In the evolving landscape of Natural Language Generation (NLG) evaluation, LLM-based metrics emerge as a compelling alternative, offering insights into model performance without the constraints of pre-defined reference responses.

Prompt-based Evaluation Prompt-based evaluation is at the forefront of NLG advancements, particularly with the utilization of LLMs (Li et al., 2024). This method integrates evaluation into prompt creation, using specialized hints to guide LLMs in assessing text quality and coherence. Typically, a prompt template acts as a structured framework containing instructions, aspects, criteria, and desired output formats, ensuring systematic evaluation of generated text. These templates enable precise articulation of evaluation requirements, ensuring consistency and reproducibility.

We followed the approach described by (Liu et al., 2023) and tailored the prompts to be suitable for the evaluation of a question-answering task. G-EVAL stands out because it uses GPT-4's advanced abilities, along with a method called chain-of-thought and a form-filling approach, to carefully judge how good the generated texts are. This method is proven to be more like how humans judge things, making it a unique and innovative tool for evaluation. The limitation of this metric is its lack of cost-effectiveness, as it operates through API calls that are subject to budget constraints. The prompt used for the G-Eval metric was tailored to each criteria and was conducted following the instructions from the official paper and the model implementation (Liu et al., 2023). One example of the prompt can also be found in Table 5. The implementation of the G-Eval metric for evaluating 100 samples across three models proved to be highly time-efficient, requiring only 2 hours to complete the evaluation of all samples.

Tuning-based Evaluation In the field of NLG evaluation, there is a significant shift toward leveraging open-source language models, such as LLaMA (Touvron et al., 2023), for fine-tuning purposes, moving away from the traditional reliance on proprietary models like GPT-3.5-turbo and GPT-4. This transition is driven by the need for cost-effective alternatives that allow for precise model evaluation on specific tasks without the financial constraints of expensive API usage associated with closed-based models.

This study utilizes Prometheus, a pioneering reference-free metric, to assess the quality of outputs from LongT5, GPT-3.5, and GPT-4 models within the HR chatbot domain. Prometheus stands out for its fine-tuned evaluation capability, which leverages a large language model to perform nuanced analysis based on customized score rubrics (Li et al., 2024). This unique approach enables Prometheus to evaluate text generation tasks comprehensively, considering factors such as creativity, relevance, and coherence without relying on reference texts. This evaluation metric demands careful crafting of prompts, which can greatly influence evaluation outcomes. A template of the final prompt used for Prometheus evaluation metrics is showcased in Table 6.

A significant limitation of this metric is its high demand for computational resources and its lack of time efficiency. For our study, it took approximately 8 hours to evaluate a mere 60 samples from a single model across four distinct criteria. Consequently, to assess 720 responses in total, we needed around 24 hours, underscoring the metric's extensive computational and time requirements.

5 Results

5.1 Models Performance Benchmark

In our analysis, we meticulously evaluate the performance of the GPT-3.5, GPT-4, and LongT5 models by examining their Readability, Relevance, Truthfulness, and Usability through a detailed evaluation process. This comprehensive evaluation leverages scores derived from human assessments, reference-free and reference-based automatic metrics, providing a holistic view of each model's capabilities in generating human-like text that aligns with these key performance indicators. An overview of all evaluation scores highlighting model performance across several dimensions is summarized in Table 2.

Overall, GPT-4 shows clear domination in terms of generation capabilities for an HR chatbot use case. N-gram-based evaluation scores such as ROUGE and BLEU are quite low because given the generative nature of the (Large) Language Models, the answer may contain words different than the reference answers. Nonetheless, these results establish GPT-4 as the leading model, effectively combining advanced language skills with the demands of content accuracy and user engagement. On the other hand, the fine-tuned LongT5's performance is observed to be inferior when benchmarked against the OpenAI models. This outcome is consistent with the anticipated advancements in LLMs, which are progressively outpacing the capabilities of fine-tuning-driven models. The performance of GPT-3.5-turbo has been notably strong, trailing marginally behind GPT-4 in only a few scoring categories. Its close performance to GPT-4 raises important considerations for the trade-offs between computational efficiency and output quality.

Metric	**GPT-3.5**	**GPT-4**	**LongT5**
Reference-based Evaluation			
BLEU Score	0.27	0.28	**0.41**
ROUGE-1	0.48	**0.52**	0.51
ROUGE-2	0.36	0.35	**0.43**
ROUGE-L	0.46	**0.50**	0.49
BERTScore_P	0.88	0.90	**0.91**
BERTScore_R	**0.96**	0.93	0.91
BERTScore_F1	0.90	**0.91**	0.90
Reference-free Evaluation (LLM-based)			
G-Eval: Relevance	4.03	**4.51**	3.17
G-Eval: Readability	4.26	**4.49**	3.52
G-Eval: Truthfulness	4.12	**4.80**	3.36
G-Eval: Usability	4.67	**4.79**	3.29
Prometheus: Relevance	3.25	**3.70**	2.83
Prometheus: Readability	3.07	**4.22**	3.73
Prometheus: Truthfulness	3.20	**3.75**	3.32
Prometheus: Usability	3.98	**4.32**	2.83
Domain Expert Evaluation			
Human Eval: Readability	4.31	**4.76**	4.02
Human Eval: Relevance	4.31	**4.67**	3.46
Human Eval: Truthfulness	4.09	**4.41**	3.67
Human Eval: Usability	3.32	**4.11**	2.59

Table 2: Average Evaluation Scores. BLEU (0 to 1), ROUGE (0 to 1) and BERTScore (-1 to +1) were computed on 200 samples, Prometheus (1 to 5) on 60 samples, and Domain Expert Evaluation (1 to 5) & G-Eval (1 - 5) on 100 samples.

5.2 Correlation Analysis

Following the precedent set by (Zhong et al., 2022), we employ Spearman (Myers and Sirois, 2004) and Kendall (Abdi, 2007) correlation analyses to evaluate the relationship between automated metrics and human judgments in our dataset, which is not normally distributed. These non-parametric tests are chosen for their robustness in assessing monotonic and rank-based relationships, providing a comprehensive view of how well automated evaluations align with human assessments. The results analysed in the following section are showcased in Table 3.

5.2.1 Correlation Human Evaluation and Reference-based Metrics

The Spearman and Kendall correlation tests are conducted to examine the alignment between automated metrics and human evaluations across three models: LongT5, GPT-3.5, and GPT-4. The findings reveal a moderate correlation for all models, indicating that traditional automated scoring methods like BLEU, ROUGE, and BERTScore, despite providing some insights, only moderately align with the nuanced human judgment. Specifically, the BLEU metric across models demonstrates an

Criteria	LongT5		GPT-3.5		GPT-4	
	Spearman ρ	Kendall τ	Spearman ρ	Kendall τ	Spearman ρ	Kendall τ
BLEU	**0.459**	0.337	0.345	0.263	0.146	0.116
ROUGE-1	**0.435**	0.321	0.364	0.284	0.113	0.091
ROUGE-2	**0.462**	0.341	0.332	0.258	0.056	0.044
ROUGE-L	**0.433**	0.324	0.353	0.274	0.093	0.075
BERTScore_P	**0.457**	0.347	0.304	0.234	0.156	0.122
BERTScore_R	**0.466**	0.305	0.085	0.064	-0.022	-0.018
BERTScore_F1	**0.455**	0.332	0.246	0.192	0.097	0.077
G-Eval						
Usability	0.675	0.584	0.217	0.198	0.346	0.327
Relevance	0.569	0.499	0.339	0.304	0.325	0.306
Readability	0.208	0.181	0.395	0.373	0.139	0.137
Truthfulness	**0.726**	0.651	0.694	0.667	0.452	0.432
Prometheus						
Usability	**0.723**	0.675	0.386	0.351	0.516	0.495
Relevance	0.467	0.439	0.419	0.371	0.382	0.357
Readability	0.493	0.468	0.378	0.358	0.225	0.213
Truthfulness	0.541	0.521	0.439	0.402	0.454	0.427

Table 3: Correlations between Automated Metrics and Human Evaluation across Models

average Spearman correlation score around 0.46 for LongT5, which underscores a consistent yet limited correlation with human evaluations. Due to its limited innovation, LongT5 typically produces text with fewer novel sentences, resulting in more favorable scores from n-gram-based metrics like BLEU and ROUGE. The analysis of GPT-3.5 and GPT-4, in particular, illuminates a significant gap between automated metrics and human judgment. As these models generate more varied and longer sentences, their outputs increasingly diverge from the patterns recognized by word-overlap metrics, such as BLEU and ROUGE. For instance, GPT-4's BLEU score correlation (Spearman's $\rho = 0.146$, Kendall's Tau $\tau = 0.116$) marks a clear disconnect, indicating that as text generation becomes more complex, the less effective traditional metrics are in evaluating it. This discrepancy calls into question the reliance on current automated metrics for assessing the creativity and nuance of outputs from advanced language models, highlighting the need for more sophisticated evaluation frameworks that can better align with human judgment.

5.2.2 Correlation Human Evaluation and Reference-free Metrics

Despite similar average scores between Reference-free metrics and Domain Expert evaluations shown in Table 2, their correlations are low. Since these methods measure linear and ordinal relationships, similar averages in evaluations do not imply a strong correlation as depicted in Table 3.

While G-Eval excels in assessing truthfulness, its capability in evaluating readability and usability lags behind, highlighting the need for further refinement. These findings suggest that while G-Eval is fairly reliable for gauging factual accuracy, it is less adept at capturing the subjective nuances as judged by humans. Prometheus outperforms G-Eval in assessing usability across all models, demonstrating its strength in evaluating the practical application of text. However, G-Eval tends to have a steadier performance across different models, particularly with LongT5, suggesting its robustness inaccurate evaluations. These findings suggest that while G-Eval is fairly reliable for gauging factual accuracy, Prometheus is better at assessing the practical application of the generated text. Both metrics show weak alignment in assessing readability, reflecting the inherent challenge of one LLM evaluating another's ability to produce easily understandable text. Overall, while Prometheus and G-Eval both serve as proxies for human evaluation, their effectiveness varies by model and evaluated criteria.

G-Eval: In evaluating the correlation between G-

Eval scores and human judgment across LongT5, GPT-3.5, and GPT-4 models on criteria such as relevance, readability, truthfulness, and usability, our analysis reveals distinct patterns:

Truthfulness stands out as a strong point for G-Eval across all models, with Spearman correlations ranging from 0.452 (GPT-4) to 0.726 (LongT5), indicating G-Eval's effective assessment of factual accuracy in generated content. Relevance shows a varied correlation, with a higher correlation in LongT5 models (Spearman: 0.569) compared to GPT-3.5 and GPT-4, where it drops to around 0.339 and 0.325, respectively. This suggests G-Eval's performance in evaluating relevance may depend heavily on the specific characteristics of the NLG model. Readability correlation is consistently low across models, with the highest Spearman correlation at 0.395 for GPT-3.5, pointing to a potential gap in G-Eval's capability to capture human perceptions of text readability. Usability also shows lower correlations, especially for GPT-3.5 (Spearman: 0.217) and GPT-4 (0.346), indicating challenges in G-Eval's assessment of the practical applicability of the generated text, as perceived by humans.

These results underscore the nuanced effectiveness of G-Eval in NLG evaluation. While it excels in assessing truthfulness, its capability in evaluating readability and usability lags behind, highlighting the need for further refinement. These findings suggest that while G-Eval is fairly reliable for gauging factual accuracy, it is less adept at capturing the subjective nuances as judged by humans.

Prometheus: For Prometheus, the correlation with human judgment exhibits a moderate strength in truthfulness across all models, with the highest Spearman correlation observed for LongT5 at 0.541, suggesting its relative reliability in assessing the factual content of NLG outputs. However, similar to G-Eval, readability assessments by Prometheus show weak alignment with human evaluations, reflecting the inherent challenge of one LLM evaluating another's ability to produce easily understandable text. In terms of relevance, the correlation is modest, with LongT5 again leading (Spearman: 0.4672), indicating that while Prometheus can gauge topical alignment to some extent, it is not entirely in sync with human perceptions. In contrast to G-Eval Performance, usability sees the strongest correlation, particularly for LongT5 (Spearman: 0.723), which implies that Prometheus can effectively judge the practical application of generated text, although this capability varies among different models.

6 Discussion

6.1 Implications

This section of the study explores the implications of our findings for the NLG domain and its utilization in Human Resources Domain.

1. **Advancements in Language Models for HR Applications**

   Our analysis showcases the supremacy of GPT-4 over the two other models in generating HR-related content, underlining its potential to significantly enhance the responsiveness and reliability of the HR chatbot. The implications of this finding suggest that the incorporation of more advanced LLMs could lead to improved employee experiences and operational efficiencies.

2. **Impact of Reference-Free Metrics on NLG Evaluation**

   The demonstrated correlation of reference-free metrics with human judgment signifies a shift towards more autonomous, consistent, and nuanced NLG assessments. This advancement could lead to creating better evaluation methods, reducing the need for time-consuming human checks and making sure NLG systems are of high quality faster.

3. **Human Judgment as the Gold Standard**

   Despite technological advances, our findings reiterate the importance of human judgment, particularly in tasks that require understanding of complex, nuanced human interactions. This observation emphasizes the necessity to maintain human oversight in NLG applications, especially in sensitive fields like HR, to ensure the generated content meets the highest standards of quality and relevance. Although the reference-free metrics yielded promising results, there is a risk of inaccuracies in handling HR-sensitive topics, as these metrics may not account for the company's confidential internal information that lies beyond the model's knowledge base.

6.2 Challenges

Throughout the course of this study, several challenges were encountered that required strategic problem-solving and adaptation.

1. A significant challenge presented itself when the Human Resources department updated the dataset. Given that the LongT5 model had been pre-trained on an earlier version, this required creative workarounds so we could conduct a fair evaluation across all models. We opted to extract overlapping questions from the LongT5's test set that corresponded with the new dataset, thus ensuring consistency in our evaluation despite the discrepancy in training data.

2. Furthermore, computational costs posed a significant challenge, particularly with reference-free metrics. Prometheus, for example, proved to be exceptionally resource-intensive, taking upwards of 20 hours to complete the evaluation process for the set number of samples.

6.3 Future Work

The progression of this research lays the groundwork for several avenues of future exploration in the NLG domain.

Given the promising results of reference-free metrics, further refinement and development of these metrics are necessary. Future research could explore ways to integrate organizational knowledge bases and proprietary information to enhance the accuracy and relevancy of reference-free evaluations in specialized domains like HR.

Another milestone that could be further improved is the human evaluation from the HR Domain Experts. Having more than one person evaluating the samples would be a good strategy for unbiased evaluation. That could lead to more effective correlation analysis between the automated metrics and human evaluation as well.

Additionally, ongoing examination and addressing of ethical aspects, such as privacy issues and data biases, are essential focuses for future studies in AI-powered HR support systems.

7 Conclusion

By optimizing retrieval techniques and benchmarking state-of-the-art LLMs with the help of domain experts, we show how LLM-based applications could benefit from a domain expert as human-in-the-loop within various iterations of the development. Our comprehensive study on evaluating GPT-3.5-turbo, GPT-4, and LongT5 within an HR chatbot context highlighted GPT-4's superiority in generating coherent, relevant, and accurate responses, making it the preferred choice for enhancing HR efficiency through reduced ticket volumes. The investigation into n-gram-based metrics like BLEU and ROUGE revealed their declining effectiveness in accurately evaluating text from more complex models, suggesting a mismatch between traditional metrics and the evolving capabilities of language models.

Additionally, our exploration into reference-free metrics, notably G-Eval and Prometheus, demonstrated their potential in aligning closely with human judgment, offering a more reliable assessment of NLG quality. These findings underscore the shift towards employing advanced LLM-powered metrics for more effective NLG evaluations.

Essentially, this research supported the integration of GPT-4 in SAP HR Q&A Chatbot systems to enhance operational efficiency and the adoption of innovative evaluation metrics. These advancements are important for guaranteeing the quality and efficacy of not only the HR Chatbot that we integrated, but also NLG technologies in real-world scenarios, marking a substantial step towards more autonomous and precise NLG assessment methods.

8 Acknowledgements

I would like to extend my heartfelt gratitude to my supervisor, Anum Afzal, for her invaluable guidance, advice, and support throughout this research project. Her profound expertise and knowledge have been very helpful, and our weekly meetings have been extremely beneficial.

I am also thankful to Prof. Dr. Florian Matthes and the SAP team, with whom this study was collaboratively conducted. Their cooperation, contributions, and access to resources were crucial in the successful completion of this project. Specifically, I would like to acknowledge Patrick Heinze, Christopher Pielka, and Rudolph Heidecker, whose expertise and support played a pivotal role in the development, evaluation and refinement of the HR-specific QA chatbot.

References

Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. *arXiv preprint arXiv:2205.11930*.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human evaluation of creative nlg systems: An interdisciplinary survey on recent papers. *arXiv preprint arXiv:2108.00308*.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Hongru Liang and Huaqing Li. 2021. Towards standard criteria for human evaluation of chatbots: a survey. *arXiv preprint arXiv:2105.11197*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004a. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni. 2002. Machine translation evaluation: N-grams to the rescue. In *LREC*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Wei Wei, Bo Dai, Tuo Zhao, Lihong Li, Diyi Yang, Yun-Nung Chen, Y-Lan Boureau, Asli Celikyilmaz, Alborz Geramifard, Aman Ahuja, et al. 2021. The first workshop on evaluations and assessments of neural conversation systems. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

A Appendix

A.1 Prompts used for OpenAI Models in the NLG Module

The optimized prompt used for ChatGPT and GPT-4 during our experiments is shown in Table 4.

A.2 G-Eval Prompt

Table 5 shows the prompt used specifically for the Readability Criteria. The prompts for other criteria (Truthfulness, Usability, Relevance) follow similar instructions as the one shown for the Readability prompt.

A.3 Prometheus Evaluation Metric Prompt

The prompt for the Prometheus Evaluation Metric outlined in Table 6 was based on the official paper's guidelines (Kim et al., 2023) for Feedback Collection. This specific prompt illustrates the Readability Criteria and was similarly adapted for other criteria such as Truthfulness, Relevance, and Usability. In general, both LLM-based metrics follow similar evaluation criteria in the prompts.

A.4 G-Eval Output Example

G-Eval Readability: "rating": 4, "explanation": "The generated answer is quite detailed and provides a comprehensive guide on how to raise a leave request. However, it might be a bit overwhelming due to the amount of information provided, which could potentially confuse some readers. The sentences are clear and there's no use of jargon, but the explanation isn't very straightforward due to its length and complexity."

A.5 Prometheus Output Example:

Prometheus Readability: The response is very clear and straightforward, making it easy to understand. It directly answers the question by stating that the *** are visible on the *** and that a certain amount is deducted every month. The response also explains *** which adds to the clarity of the answer. The language used is simple and there is no jargon or convoluted explanations, making it very easy to understand. So the overall score is 5. [RESULT] 5 [CRITERIA] Readability

SYSTEM PROMPT

You are an HR chatbot for SAP and you provide truthful and concise answers to employee questions based on provided relevant HR articles.
1. Stay very concise and keep your answer below 150 words.
2. Do not include too much irrelevant information unrelated to the posed question.
3. Keep your response brief and on point.
4. Include URLs from the relevant article if it is important to answer the question.
5. If the answer applies to specific labs/countries/companies, include this information in your response.
6. Refer to the employee directly as "you" and not indirectly as "the employee".
7. If the provided HR article does not include the answer to the question, tell the employee to create an HRdirect ticket.
8. Answer in a polite, personal, user-friendly, and actionable way.
9. Never make up your response! If you do not know the answer to the question, just say so and ask the user to create an HRdirect ticket!

USER PROMPT

Question: {question}
Relevant Article: {article}

Table 4: Chatbot Prompt for OpenAI Models

SYSTEM PROMPT

You will be given a generated answer for a given question. Your task is to act as an evaluator and compare the generated answer with a reference answer on one metric. The reference answer is the fact-based benchmark and shall be assumed as the perfect answer for your evaluation. Please make sure you read and understand these instructions very carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria: {criteria}
Evaluation Steps: {steps}

USER PROMPT

Example: {example}
Question: {question}
Generated Answer: {generated_answer}
Reference Answer: {reference_answer}
Evaluation Form: Please provide your output in two parts separate as a Python dictionary with keys rating and explanation. First the rating in an integer followed by the explanation of the rating.
{metric_name}

METRIC SCORE CRITERIA

{The degree to which the generated answer matches the reference answer based on the metric description.}
Readability(1-5) - Please rate the readability of each chatbot response. This criterion assesses how easily the response can be understood. A response with high readability should be clear, concise, and straightforward, making it easy for the reader to comprehend the information presented. Complex sentences, jargon, or convoluted explanations should result in a lower readability score.

METRIC SCORE STEPS

{Readability Score Steps}
1. Read the chatbot response carefully.
2. Assess how easily the response can be understood. Consider the clarity and conciseness of the response.
3. Consider the complexity of the sentences, the use of jargon, and how straightforward the explanation is.
4. Assign a readability score from 1 to 5 based on these criteria, where 1 is the lowest (hard to understand) and 5 is the highest (very easy to understand).

Table 5: G-Eval Prompt Example for Readability Criteria

SYSTEM PROMPT
Task Description: An instruction (might include an input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing an evaluation criterion is given. 1. Write a detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general. 2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric. 3. The output format should look as follows: Feedback: [write a feedback for criteria] [RESULT] [an integer number between 1 and 5]. 4. Please do not generate any other opening, closing, and explanations.
Question to Evaluate: {instruction} **Response to Evaluate:** {response} **Reference Answer (Score 5):** {reference answer} **Score Rubrics:** {criteria description} Score 1: {Very Low correlation with the criteria description} Score 2: {Low correlation with the criteria description} Score 3: {Acceptable correlation with the criteria description} Score 4: {Good correlation with the criteria description} Score 5: {Excellent correlation with the criteria description} **{criteria description}**: Readability(1-5) - Please rate the readability of each chatbot response. This criterion assesses how easily the response can be understood. A response with high readability should be clear, concise, and straightforward. Complex sentences, jargon, or convoluted explanations should result in a lower readability score.

Table 6: Prometheus Prompt Example for Readability Criteria