國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

SALIERI: 透過配對相似品質的回答系統性量測大型語言模型評估者的自我增強偏誤

SALIERI: Systematic Assessment of Self-enhancement Bias in Large Language Model Evaluators by Pairing Similar Quality Response

林鴻儒

Hung-Ju Lin

指導教授: 許永真 博士

陳縕儂 博士

Advisor: Jane Yung-jen Hsu Ph.D.

Yun-Nung Chen Ph.D.

中華民國 113 年 12 月

December, 2024

doi:10.6342/NTU202500980

Acknowledgements

首先，真的要非常感謝我的指導教授許永真老師，能夠進入這個實驗室並接受老師的指導，是我研究生涯中最幸運的事之一，也很感謝這段得來不易的緣分。在研究的過程中，老師總是能精闢地指出我思維上的盲點，如同我的研究主題一樣，客觀又仔細的指出研究上的問題。透過與老師的交流，我得以從他多年累積的研究經驗中學習，如同 Knowledge Distillation 的過程一樣，讓我在有限的時間內吸收了豐富而深刻的洞見，這段歷程讓我受益匪淺。也非常特別感謝老師讓我有機會參與 TAIDE 計畫，使我得以接觸到大型語言模型最前沿的研究議題。在這段期間，我不僅深入理解了語言模型的實作與應用，也獲得了充足的運算資源與協作的經驗，這對我的研究歷程有極大的幫助與啟發。衷心感謝老師一路以來的指導與支持，這段師生之緣，我將深深珍惜。

誠摯感謝本論文的共同指導老師陳縕儂老師，於口試及論文審閱過程中所提供的寶貴建議與指導。非常感謝口試委員李育杰老師、陳尚澤老師與蔡宗翰老師，撥冗審閱本論文，提出中肯且具體的建議，協助我釐清內容中的疏漏與不足，使論文更加周延。各位老師的回饋與指導對我助益良多，謹此致上誠摯的謝意。

由衷感謝郭彥伶老師在我碩士最後半年的研究過程中所提供的悉心指導與陪伴。在研究方面，感謝彥伶老師總是耐心傾聽我那些尚未釐清的想法，並在討論中一步步引導我釐清具體可行的作法。非常感謝彥伶老師在每次 meeting 時耐心

doi:10.6342/NTU202500980

聆聽我不夠流暢的英語表達，讓我逐漸建立起口說的信心，感覺自己表達得越來越順暢。特別感謝彥伶老師提供我們前往美國擔任訪問學生的寶貴機會。在那段時間中，我不僅累積了寶貴的學術經驗與視野，學習到許多關於海外研究環境與生活的實際經驗。從申請、適應到實際執行研究，彥伶老師都給予我們極大的支持與協助，無論是在學術、行政或生活上，我們都受到莫大的鼓勵與照顧。衷心感謝彥伶老師在我學術與人生旅程中的深厚支持，這份恩情我將永遠銘記在心。

非常感謝實驗室裡的同學們，在與各位同學相處的這段時間裡，時常會讚嘆同學們的能力與成就，時常覺得他們如同 Amadeus，才華洋溢也努力不懈，而自己如同 Salieri，羨慕著他們。在和同學們討論研究問題時，不時能獲得許多精闢的見解，啟發我從不同角度思考自己的研究。能和這樣一群人一同走過這段旅程，是我在研究生涯中最珍貴的收穫之一。

最後，也想感謝一下自己的努力，期勉自己不只能夠客觀的敬佩同儕的能力與成就，更能客觀的正視自己的表現

doi:10.6342/NTU202500980

摘要

自我提升偏誤 (self-enhancement bias) 是指模型傾向於對其自身生成的回答給予過高評分的現象。然而，我們發現低品質回答往往會產生較高的偏誤分數，且這種現象在能力較低的模型中更為明顯。這反映了現有測量方法的一個重要局限性——容易受到回答品質和模型能力的干擾，導致結果不準確。

為了解決此問題，我們提出了一種新方法——SALIERI，透過配對品質相似的回答來消除這些干擾，從而量化真正的偏誤程度。實驗結果顯示，在 Summeval 資料集上，SALIERI 將 LLaMA 2 7B 模型中高品質與低品質回答的偏誤分數差距從 2.55 降至 0.75，並將能力較強的 Llama 3 8B 模型的差距從 1.44 降至 0.69。這些結果表明，SALIERI 能有效提高自我提升偏誤測量的準確性與可靠性。

關鍵字：大型語言模型用評估、大型語言模型、自我提升偏誤

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Abstract

Self-enhancement bias refers to the tendency of models to overrate their own responses. However, we found that lower-quality responses tend to produce higher bias scores, an issue exacerbated in less capable models. This highlights a key limitation of current measurement methods, which are confounded by response quality and model capability, leading to inaccuracies.

To address this, we propose SALIERI, a method that pairs responses of similar quality to isolate true bias. Experiments on the Summeval dataset show that SALIERI reduced bias score gaps between high- and low-quality responses from 2.55 to 0.75 for the less capable LLaMA 2 7B model and from 1.44 to 0.69 for the more advanced Llama 3 8B model. These results demonstrate SALIERI's effectiveness in achieving more accurate and reliable measurements of self-enhancement bias.

Keywords: LLM-as-a-judge, Large Language Model, Self-Enhancement Bias

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Contents

Page

Acknowledgements **i**

摘要 **iii**

Abstract **v**

Contents **vii**

List of Figures **xi**

List of Tables **xiii**

Denotation **xv**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Motivation 3

1.3 Research Objective 4

1.4 Thesis Organization 4

Chapter 2 Related Work **7**

2.1 Natural Language Generation Evaluation 7

2.2 Large Language Model (LLM) 8

2.3 Self-Enhancement Bias 9

doi:10.6342/NTU202500980

Chapter 3 Problem Formulation **11**

3.1 Self-Enhancement Bias in Pointwise Evaluation 11

3.2 Pointwise Natural Language Generation Evaluation 12

3.3 Confounding in Measuring Self-Enhancement Bias 13

3.3.1 Alternative Explanation 13

3.3.2 Evaluation Capabilities 14

Chapter 4 Response Quality Impact **19**

4.1 Dataset Selection and Model Selection 19

4.1.1 Dataset Selection 19

4.1.2 Model Selection 20

4.2 Experiment Flows 21

4.3 Experiment Results 24

4.3.1 Self Response Setting 24

4.3.2 Other Response Setting 24

4.3.3 Model Capability Relationship 25

4.4 Conclusion 26

Chapter 5 SALIERI **29**

5.1 Main Idea 30

5.2 Pairing Self-Enhancement Bias Score 30

5.3 Proposed Method 31

5.3.1 Overall Process 31

5.3.2 Paired Response Pool Generation 33

5.4 Experiment 33

doi:10.6342/NTU202500980

5.4.1 Experimental Setup 33
5.4.2 Experimental Design 34
5.5 Results 34
5.5.1 Self-response setting 36
5.5.2 Other-response setting 36
5.6 Conclusion 38
Chapter 6 Discussion and Conclusion **41**
6.1 Discussion 41
6.1.1 Weaker Models and Reduced Bias 41
6.1.2 Variation in Performance Across Datasets 41
6.1.3 Score Rubrics 42
6.2 Limitations 42
6.2.1 Dependency with Dataset 42
6.2.2 Pairing Response Constraint 43
6.2.3 Reference Agents Constraint 43
6.3 Conclusion 44
References **45**
Appendix A — Evaluation Prompt **51**
Appendix B — Treatment Check **53**
Appendix C — Summeval Preparation **55**
C.1 Scoring Rubric Generation 55
C.2 Diverse Response Generation 61

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

List of Figures

4.1 Experiment Flow of Self-Response Setting 21
4.2 Experiment Flow of Other-Response Setting 22

5.1 Overview of SALIERI 32

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

List of Tables

4.1 Reference Agents Correlation Analysis 21

4.2 Self-Response Setting Result 24

4.3 Other-Response Setting Result 25

4.4 Relationship Between Model Capability and Detectable Ratio 26

5.1 Comparison of Self-Enhancement Bias Calculation Results in Self-Response Setting 35

5.2 Comparison of Self-Enhancement Bias Calculation Results in Other-Response Setting 37

B.1 Average Scores on Feedback Collection and Summeval Datasets 53

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Denotation

t	Target model, which we used to assess its self-enhancement bias.
r	Reference agent, from which we use the score as the golden score.
ρ	Evaluation model capability.
λ	Proportion of performance that can be detected.
$X'_{a,t}$	The standardized score that agent a assigns to target model t.
$\bar{X}_{a,t}$	Average score given by agent a to target model t over dataset.
$f_a(q, r, \textit{ref})$	The score assigned by agent a to response r for input q and reference *ref*.
NM	Set of norm models used for score normalization.
$\beta_{t,t}$	True self-enhancement bias of target model t.
$\hat{\beta}_{t,t}$	Observed self-enhancement bias (i.e., raw score difference).
$g(X'_{r,t}, \rho_t)$	Detected quality of target response t as perceived by target model with capability ρ_t.

doi:10.6342/NTU202500980

p	A baseline model with no assumed bias relation to t, used in pairing for SALIERI.
$\beta_{t,p}$	Bias score when comparing model t's judgment on model p's responses to reference.

doi:10.6342/NTU202500980

Chapter 1 Introduction

1.1 Background

As the field of Natural Language Generation (NLG) rapidly advances, evaluating its performance becomes increasingly important. Traditional evaluation methods, such as ROUGE [11] and BLEU [19], assess the similarity between the model's output and reference answers written by human experts by calculating n-gram overlaps. These methods can efficiently evaluate generated text by quickly comparing syntactic content. However, previous research has shown that the scores produced by these methods often do not align well with human expert annotations [13, 30, 32].

In order to make evaluation scores align more closely with human preferences, model-based evaluation methods have emerged. For example, BartScore [30] and GPTScore [3] treat evaluation as a generation task, assuming that a higher probability generated by the model indicates a better quality text. Alternative methods leverage the strong instruction-following capabilities of large language models (LLMs) to directly ask the model to provide specific scores or determine the winner between two responses [6, 13, 31]. These methods have shown better alignment with human expert annotations [13].

However, despite the high correlation of model-based evaluation methods with hu-

doi:10.6342/NTU202500980

man expert ratings, new issues have been identified, such as positional bias [26], prompt injection [21], and self-enhancement bias [6, 23, 31]. These problems pose challenges to the reliability of model-based evaluations.

Among these, self-enhancement bias refers to the tendency of an evaluation model to assign higher scores to itself or to models that share the same base model. Many studies have reported instances of self-enhancement bias during evaluations, such as BART and T5 assigning a lower perplexity to the responses generated by themselves [23], GPT-4 giving responses generated by itself a higher win rate in MT-bench [31], and CritiqueLLM giving responses generated by ChatGLM, which is the base model of CritiqueLLM, a relatively high score in Alignbench [6].

However, we found that previous measurements of self-enhancement bias had confounding factors: Due to a lack of judgment ability, the model may have difficulty distinguishing response quality, leading it to assign relatively higher scores to its own lower-quality responses. Therefore, we designed experiments to verify the hypothesis and further propose our method to make the assessment of self-enhancement bias more stable and less influenced by response quality and evaluation model capability.

To summarize, our main contributions in this paper are:

- **Contribution 1**: Demonstrating that when a model lacks judgment ability and assigns relatively higher scores to lower-quality responses, it results in an inflated self-enhancement bias score.

- **Contribution 2**: Proposing a new calculation method, SALIERI, that makes the assessment of self-enhancement bias more closely reflect the impact of the source.

doi:10.6342/NTU202500980

1.2 Motivation

The evaluation of large language models (LLMs) plays a pivotal role in advancing Natural Language Generation (NLG) research, yet its fairness and reliability are often compromised by biases in model-based evaluation methods. Among these, self-enhancement bias—the tendency of a model to favor outputs generated by itself—represents a significant challenge to fair and unbiased assessments. This bias undermines the credibility of evaluation systems, especially as they are increasingly used to benchmark models for critical applications.

While existing studies have reported instances of self-enhancement bias [6, 31], we observe a lack of a clear and rigorous definition of this bias in pointwise evaluations. Furthermore, current measurement methods may inadvertently conflate self-enhancement bias with a model's general inability to distinguish between high- and low-quality responses, potentially leading to inflated or misleading results. For instance, CritiqueLLM [6] reports that smaller models exhibit greater self-enhancement bias. However, we hypothesize that this observation could be attributed to judgment errors rather than genuine bias, a possibility we aim to investigate through our experiments.

To address these limitations, our work aims to redefine and assess self-enhancement bias with greater precision. By eliminating confounding factors and proposing a novel method, SALIERI, we seek to provide a more accurate evaluation framework. This will enable researchers and practitioners to make more informed decisions about evaluation models, ultimately fostering fairness and reliability in NLG assessments.

doi:10.6342/NTU202500980

1.3 Research Objective

The objective of this research is to address the limitations in existing methods for assessing self-enhancement bias in model-based evaluations. Specifically, we aim to:

- Test how model capability and response quality significantly influence the measurement of self-enhancement bias.
- Develop a refined method to assess self-enhancement bias, minimizing confounding effects from model capability and response quality.

By achieving these objectives, this study seeks to enhance the fairness and reliability of model-based evaluations, contributing to the development of more trustworthy NLG systems.

1.4 Thesis Organization

This thesis is organized into six chapters.

Chapter 2 reviews the existing literature on the evaluation of Natural Language Generation (NLG) and self-enhancement bias. It explores traditional evaluation metrics, recent advancements in model-based evaluations, and key studies addressing self-enhancement bias, providing the necessary context for this research.

Chapter 3 defines self-enhancement bias and its relevance to pointwise evaluations in NLG tasks. It also outlines the specific problem addressed in this thesis: the influence of response quality and evaluation model capability on self-enhancement bias measurement.

doi:10.6342/NTU202500980

Variations in response quality from the same model and limitations in evaluation model capability can lead to inconsistent and unreliable bias scores.

Chapter 4 presents an analysis showing that responses of different quality result in varying self-enhancement bias scores. This finding underscores that bias scores are not solely influenced by the source of the response but are also affected by response quality, which complicates the reliability of existing evaluation methods.

Chapter 5 introduces *SALIERI*, a novel method for assessing self-enhancement bias. This chapter details how *SALIERI* mitigates the influence of response quality, providing a more accurate measurement of self-enhancement bias.

Finally, Chapter 6 concludes the thesis by summarizing the research findings, discussing their implications, highlighting the limitations of our work, and suggesting potential directions for future research on the assessment of self-enhancement bias.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Chapter 2 Related Work

2.1 Natural Language Generation Evaluation

Traditionally, the evaluation of Natural Language Generation (NLG) tasks has relied on n-gram similarity metrics, such as BLEU [19] and ROUGE [11], which measure the similarity between generated text and human-written references. However, recent studies have highlighted that these metrics often show poor alignment with human preferences [13, 32].

Emerging methods that calculate the conditional probabilities predicted by language models, such as BARTScore [30] and GPTScore [3], have shown improved performance in evaluating NLG tasks. These methods suggest that higher-quality NLG outputs are more likely to be generated by the language model. By leveraging the rich semantic information encoded in pretrained models, they enable more nuanced evaluations that extend beyond surface-level n-gram overlaps.

With advances in large language models (LLMs), their use in evaluations has become increasingly popular. In addition to conditional probability methods, LLMs can serve as human-like annotators, utilizing approaches such as the Likert Scale [10] for pointwise scoring, or pairwise comparison, where the model selects the better response from two

doi:10.6342/NTU202500980

options. For example, G-eval [13], Prometheus [7], and CritiqueLLM [6] support evaluations based on the Likert Scale, while LLM-judge [31], PandaLM [27], and JudgeLM [33] use pairwise comparison. Some methods, such as Prometheus 2 [8] and Auto-J [9], support both approaches. These instruction-based evaluation methods not only demonstrate high alignment with human judgments but also make the evaluation process more interpretable [7, 9, 13, 27, 33].

2.2 Large Language Model (LLM)

Starting with T5 in 2019 [20], researchers began exploring ways for language models to solve diverse natural language generation (NLG) tasks. GPT-3 [15] demonstrated the ability to perform various tasks based on different prompts, highlighting the model's versatility. Later, models like FLAN [29] showed that instruction tuning could significantly improve a model's ability to follow user instructions. InstructGPT [17] further advanced this approach by using Reinforcement Learning from Human Feedback (RLHF) to better align model responses with human preferences. The release of ChatGPT [16] showcased the broad capabilities of large language models (LLMs) to a vast user base.

Subsequent open-source LLM releases, such as Meta's LLaMA series (LLaMA [24], LLaMA 2 [25], and LLaMA 3 [4]), Google's Gemma [22], and Mistral AI's Mistral [5], have further advanced research and development in LLM applications. These advancements have supported research evaluating NLG tasks using LLMs [7, 9, 12, 27].

doi:10.6342/NTU202500980

2.3 Self-Enhancement Bias

Self-enhancement bias in evaluation models has been reported in several studies [6, 14, 18, 23, 31]. These studies demonstrate that self-enhancement bias occurs in various evaluation schemes, including pointwise grading [6], pairwise grading [31], and conditional probability-based grading [18, 23].

Regarding the causes of self-enhancement bias, previous studies have suggested that models with stronger self-recognition abilities tend to exhibit more severe self-enhancement bias [18]. Additionally, in pairwise evaluation contexts, self-enhancement bias in models is influenced by the familiarity of responses [28].

Despite the growing body of research on self-enhancement bias, there is no universally accepted definition of the bias. Different studies have approached its definition in various ways, with no single approach gaining widespread consensus. For instance, in pairwise comparison, LLM-Judge [31] defines self-enhancement bias based on the difference in win rates between the target model and humans. Another study, Self-Preference Bias [28] defines self-enhancement bias using conditional probability, measuring the accuracy difference of the target model under two conditions: when humans prefer and when they disprefer the target model's response. In pointwise scoring, CritiqueLLM [6] uses the normalized score difference between the target model' s self-assessment and human evaluation to define the bias. Meanwhile, in conditional probability-based research, [23] defines self-enhancement bias using the perplexity scores provided by the target model for its own responses versus responses from other models.

Although research on self-enhancement bias has explored various evaluation schemes,

doi:10.6342/NTU202500980

including pairwise and conditional probability-based methods, the definition of self-enhancement bias in pointwise scoring remains unclear. In pointwise evaluations, the focus has been on comparing the target model’ s self-assessment to human evaluations, but there is no widely accepted framework to quantify self-enhancement bias. This lack of clarity in pointwise scenarios highlights the need for further research to establish a consistent and robust definition of self-enhancement bias in these contexts.

doi:10.6342/NTU202500980

Chapter 3 Problem Formulation

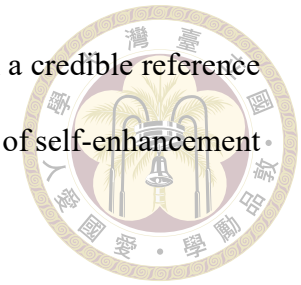
In this chapter, we outline the limitations of previous methods for assessing self-enhancement bias. First, we provide a clear definition of self-enhancement bias in the context of pointwise evaluation. Next, we describe the pointwise evaluation task for natural language generation. Finally, we highlight the potential confounding factors in assessing self-enhancement bias, specifically the influence of evaluation model capability and response quality. These factors will be further examined through experiments in the next chapter.

3.1 Self-Enhancement Bias in Pointwise Evaluation

Previous research has shown that when a Large Language Model evaluates its own performance, it tends to overestimate its capabilities, observing its responses as better than they actually are [31]. This phenomenon, termed ”self-enhancement bias,” is borrowed from psychology research. In their experiments, the evaluation method relies on pairwise comparison, with win rate serving as the evaluation metric. If the target model assigns itself a higher win rate than that assigned by a credible reference agent, such as a human, this discrepancy is considered evidence of self-enhancement bias.

Applying the same logic to pointwise evaluation, self-enhancement bias in an LLM

doi:10.6342/NTU202500980



can be indicated if the target model assigns itself a higher score than a credible reference agent does. We take the difference between these scores as a measure of self-enhancement bias.

In other words, to assess the self-enhancement bias of a target model t, we examine the difference between the score $X'_{t,t}$ it assigns to its own generated responses and the score $X'_{r,t}$ assigned by a reference agent r. This difference indicates the extent of self-enhancement bias and can be expressed using the following formula:

$$\beta_{t,t} = X'_{t,t} - X'_{r,t} \tag{3.1}$$

3.2 Pointwise Natural Language Generation Evaluation

Evaluating the quality of a response generated by a Natural Language Generation (NLG) model using a Likert scale [10] can be framed as a classification task. Given a response r to be scored, along with additional context such as the corresponding query q and reference answer ref, the scoring agent a assigns a score s to the response. We can define this process as a function, f, as follows:

$$f_a(q, r, ref) \rightarrow s, \quad \text{where} \quad s \in \{1, 2, \ldots, k\} \tag{3.2}$$

To assess the capability of the target NLG model t, we select a dataset with m items, $Q(q_1, q_2, \ldots, q_m)$, each accompanied by a reference answer, $R_{\text{ref}}(ref_1, ref_2, \ldots, ref_m)$. We then prompt the target model to respond to each item in the dataset, obtaining its responses $R_t(r_{t,1}, r_{t,2}, \ldots, r_{t,m})$.

doi:10.6342/NTU202500980

We then ask the scoring agent a to evaluate these responses. The mean score $\bar{X}_{a,t}$ for this dataset is used to indicate the capability of the target NLG model t. This process can be defined as follows:

$$\bar{X}_{a,t} = \frac{\sum_{j=1}^{m} f_a(q_j, r_{t,j}, ref_j)}{m} \tag{3.3}$$

The scoring criteria of different agents vary, so standardization is required to enable comparison. We first select n models as norm models for standardization purposes, $NM(nm_1, nm_2, \ldots, nm_n)$, and use the average scores given by the agents for these norm models as a baseline. The standardization formula is as follows:

$$X'_{a,t} = \bar{X}_{a,t} - \frac{\sum_{i=1}^{n} \bar{X}_{a,nm_i}}{n} \tag{3.4}$$

3.3 Confounding in Measuring Self-Enhancement Bias

3.3.1 Alternative Explanation

In the CritiqueLLM work, it was found that their model demonstrates self-enhancement bias [6]. Furthermore, as the model size increases, the bias tends to decrease. While this could be interpreted as smaller models exhibiting a stronger preference for themselves, we believe there may be an alternative explanation.

We observe that CritiqueLLM does not recognize that ChatGLM's responses are inferior to those of GPT-4. Instead, it perceives ChatGLM's responses as being less poor than humans judge them to be. Based on this, we hypothesize that these lower-capability

doi:10.6342/NTU202500980

models may not necessarily favor their own responses; rather, they may struggle to distinguish lower-quality responses from higher-quality ones. As a result, their responses, being of lower quality, may seem to the model to be less inadequate than they actually are.

For instance, consider an extreme case where the target model lacks any evaluative ability and assigns scores randomly, regardless of the quality of the answers. In this case, all sets of responses would receive the same expected score, resulting in standardized scores of zero. Suppose the target model's answer should actually receive a score of -2. Under the previous scoring method, however, the model would be assigned a self-enhancement bias score of $+2$. This outcome contradicts our understanding of the model' s behavior. How could a model that assigns scores randomly appear to favor its own responses?

3.3.2 Evaluation Capabilities

Based on this observation, we account for the evaluation model's limited ability to distinguish response quality. We should not subtract the actual performance of the target model; instead, we subtract the performance that the model can observe, reflecting its limited evaluation capability. To represent this limitation, we use $g(X'_{r,t}, \rho_t)$ to denote the quality that is detectable by the target model, as expressed in the following equation:

$$X'_{t,t} - g(X'_{r,t}, \rho_t) = \beta_{t,t} \tag{3.5}$$

Here, ρ_t represents the capability of the target model t, typically quantified by the Pearson correlation with human-annotated results in the testing data. A small ρ_t indicates that

doi:10.6342/NTU202500980

the model' s judgments are not well aligned with human evaluations, which reflects a low evaluation capability. This low capability makes it harder for the model to assess and differentiate response quality, causing $g(X'_{r,t}, \rho_t)$ to approach 0. On the other hand, a large ρ_t means the model is better at detecting response quality, making $g(X'_{r,t}, \rho_t)$ approach $X'_{r,t}$, effectively capturing the full range of response quality. Thus, we hypothesize that the function $g(X'_{r,t}, \rho_t)$ can be expressed as $\lambda_t \cdot X'_{r,t}$, where λ_t represents the proportion of the response quality that is detectable by the target model. When the model has high capability, λ_t approaches 1; otherwise, it is closer to 0. This adjustment accounts for the model's limited ability to evaluate response quality. Incorporating this factor, we revise the formula as follows:

$$X'_{t,t} - \lambda_t \cdot X'_{r,t} = \beta_{t,t} \tag{3.6}$$

Revisiting the previous method for assessing self-enhancement bias and applying our concept of evaluation capability, we adjust the formula for bias assessment. The revised formula is as follows:

$$\hat{\beta}_{t,t} = X'_{t,t} - X'_{r,t} = \beta_{t,t} - (1 - \lambda_t) \cdot X'_{r,t} \tag{3.7}$$

Revisiting the previous method for assessing self-enhancement bias and applying our concept of evaluation capability, we adjust the formula for bias assessment. The revised formula is as follows:

$$\hat{\beta}_{t,t} = X'_{t,t} - X'_{r,t} = \beta_{t,t} - (1 - \lambda_t) \cdot X'_{r,t} \tag{3.8}$$

doi:10.6342/NTU202500980

We can see from this formula that when the evaluation capability is imperfect ($\lambda_t < 1$), the bias score of the target model ($\hat{\beta}_{t,t}$) becomes more sensitive to the quality of the response. Specifically, as the response quality ($X'_{r,t}$) decreases, the bias score increases.

To test our hypothesis, we examine whether variations in response quality impact the bias score. Specifically, we collect two responses, t_a and t_b, from the same model, where t_a represents a higher-quality response than t_b, i.e., $X'_{r,t_a} > X'_{r,t_b}$. We then compute the bias scores for both responses, t_a and t_b, using the formulas below and observe the difference in the bias scores:

$$\hat{\beta}_{t,t_a} = X'_{t,t_a} - X'_{r,t_a} = \beta_{t,t} - (1 - \lambda_t) \cdot X'_{r,t_a}$$

$$\hat{\beta}_{t,t_b} = X'_{t,t_b} - X'_{r,t_b} = \beta_{t,t} - (1 - \lambda_t) \cdot X'_{r,t_b}$$

The difference in the bias scores is given by:

$$\hat{\beta}_{t,t_b} - \hat{\beta}_{t,t_a} = (1 - \lambda_t) \cdot (X'_{r,t_a} - X'_{r,t_b})$$

If our hypothesis holds, it suggests that the model's evaluation capability constrains its assessment of response quality, resulting in certain scores being undetectable ($\lambda_t < 1$). Given that $X'_{r,t_a} > X'_{r,t_b}$, we expect the bias score $\hat{\beta}_{t,t_b}$ for the lower-quality response (t_b) to be larger than the bias score $\hat{\beta}_{t,t_a}$ for the higher-quality response (t_a). If, however, we observe that $\hat{\beta}_{t,t_b} < \hat{\beta}_{t,t_a}$, it would contradict our hypothesis.

If variations in response quality influence the bias score, we will further investigate

doi:10.6342/NTU202500980

whether λ is associated with the evaluation model's capability. To calculate λ, we rearrange Formula 3.3.2 as follows:

$$(\lambda_t - 1) = \frac{\hat{\beta}_{t,t_b} - \hat{\beta}_{t,t_a}}{X'_{r,t_b} - X'_{r,t_a}} = \frac{(X'_{t,t_b} - X'_{r,t_b}) - (X'_{t,t_a} - X'_{r,t_a})}{X'_{r,t_b} - X'_{r,t_a}}$$

$$(\lambda_t - 1) = \frac{X'_{t,t_b} - X'_{t,t_a}}{X'_{r,t_b} - X'_{r,t_a}} - 1$$

$$\lambda_t = \frac{X'_{t,t_b} - X'_{t,t_a}}{X'_{r,t_b} - X'_{r,t_a}}$$

We assess the model's capabilities by calculating the Pearson correlation between the test data and the annotated scores as follows:

$$\rho_t = \text{corr}(X_{t,\text{test}}, X_{annotated,\text{test}})$$

We will further verify whether our hypothesis holds true in the next chapter.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Chapter 4 Response Quality Impact

In this chapter, we follow the reasoning from the previous chapter and investigate whether the bias score increases when response quality decreases. First, we discuss the selection of datasets and models. Second, we outline the experimental flow and explain how we verify our claim. Three experiments are presented, showing that evaluation models exhibit higher self-enhancement bias when using lower-quality responses. This effect is attributed to the target model's limited capability. Third, we present the experimental results that support our explanation. Finally, we summarize the results and offer an interpretation of the phenomenon.

4.1 Dataset Selection and Model Selection

4.1.1 Dataset Selection

In our experiments, we selected the Summeval and Feedback Collection datasets.

Summeval [2] is a summarization dataset that contains 100 news articles sourced from the CNN/Daily Mail dataset. For each article, the dataset includes summaries generated by 16 different models, along with human annotations that evaluate these summaries across four aspects: consistency, coherence, fluency, and relevance. When using this

doi:10.6342/NTU202500980

dataset for evaluation, we compute scores for each of the four aspects and then take the average of these scores as the final score for each item.

The Feedback Collection dataset [7], collected by KAIST AI, was created for training the evaluation model Prometheus. It contains 20,000 diverse instructions, each paired with a response and a score ranging from 1 to 5. For our experiments, we sampled 1,000 instructions from the full dataset.

4.1.2 Model Selection

As target models, we chose Gemma 7B [22], LLaMA 2 7B [25], and LLaMA 3 8B [1], which are open-source models of relatively small size.

We selected GPT-4 and Prometheus 2 8x7B as reference agents due to their strong alignment with human expert evaluations, as reported in prior work [8, 13]. Using two reference models helps reduce the risk of relying on a single agent that may exhibit specific preferences. To further validate their evaluation capabilities, we computed the Pearson correlation scores between their scores and annotations on the Summeval and Feedback Collection datasets. As shown in Table 4.1, both Prometheus 2 and GPT-4 exhibit higher alignment with the dataset annotations compared to the target models, indicating that they are well-suited to serve as reference agents.

To standardize the models' scores, we collected responses from GPT-4, GPT-4o, and Prometheus 2 8x7B, which serve as norm models in our evaluation framework. Specifically, the responses generated by these models are used to compute a baseline score for each evaluation model. This baseline reflects the average score that an evaluation model assigns to a set of responses, allowing us to normalize scores across different evaluators

doi:10.6342/NTU202500980

Model	Summeval	Feedback Collection
LLaMA 2 7B	0.03	0.55
Gemma 7B	0.05	0.38
LLaMA 3 8B	0.24	0.73
Prometheus 2 8x7B	0.33	**0.80**
GPT-4	**0.41**	0.78

Table 4.1: Pearson correlation between model scores and annotated scores across two datasets. Prometheus 2 and GPT-4 achieve consistently higher correlations with the annotations than the target models, supporting their suitability as reference agents in our experiments

and enable fair comparisons of self-enhancement bias.

4.2 Experiment Flows

To verify the hypothesis proposed in Chapter 3.3.2, we conduct three experiments. First, in the "self-response setting," as shown in Figure 4.1, we test whether the model exhibits a higher bias score for lower-quality responses generated by itself. Second, in the "other-response setting," as shown in Figure 4.2, we examine whether the model shows a higher bias score for lower-quality responses generated by other models. Finally, we investigate whether λ is related to the evaluation model's capability.

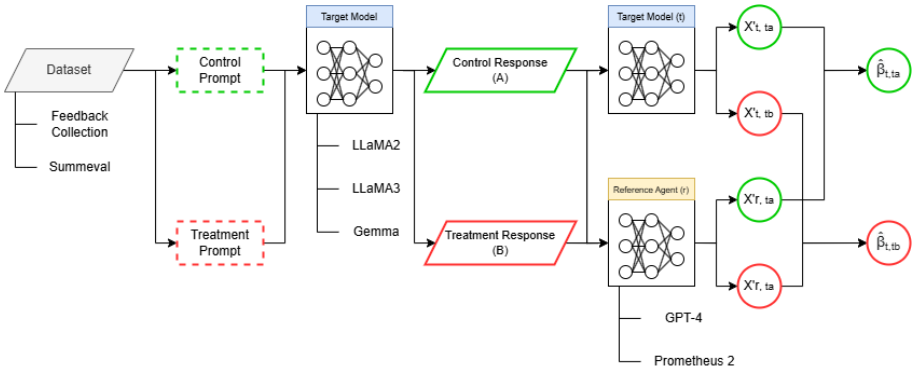


Figure 4.1: Experiment flow for testing the impact of response quality in the self-response setting. The process includes: (1) generating responses of varying quality using the target model, (2) scoring the responses, (3) calculating the model preference bias, and (4) comparing the bias between normal-quality and lower-quality responses.

doi:10.6342/NTU202500980

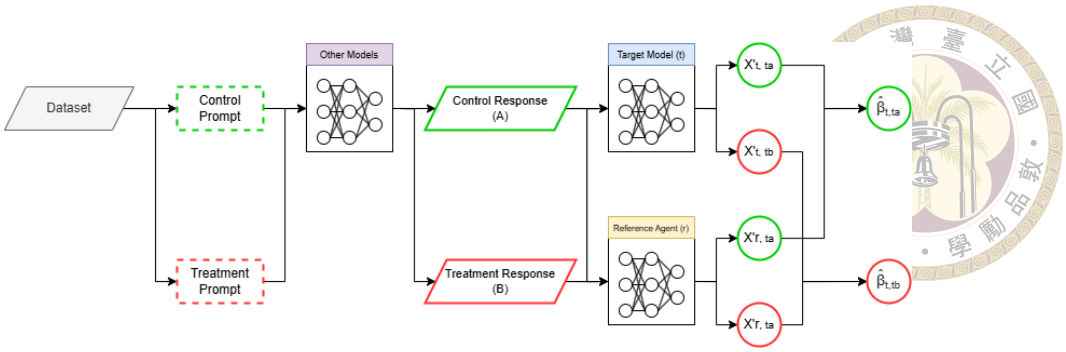


Figure 4.2: Experiment flow for testing the impact of response quality in the other-response setting. The process includes: (1) generating responses of varying quality using other models, (2) scoring the responses, (3) calculating the bias, and (4) comparing the bias between normal-quality and lower-quality responses.

To check whether the target model exhibits a higher bias score for lower-quality responses, we first generate two response sets of different quality using different prompts. For the two datasets, *Feedback Collection* and *Summeval*, we use different prompt templates to generate responses of varying quality.

For the *Summeval* dataset, we used the following method:

- **Control Prompt Response (A):** `<summary instruction> + <source article>`
- **Treatment Prompt Response (B):** `<summary instruction> + <source article> + <previous response> + "Based on the previous dialog, give me a response about the article with some emotional judgment in 3 or 4 lines. Do not include any details from the article."`

Here, the treatment response prompt includes a follow-up instruction that asks the model to add an emotional tone, deviating from the original article's content. This approach introduces subjective judgment, deliberately contrasting with the objective and fact-based control summaries. We chose this prompt structure because summaries should be objective and aligned with the source article, and the emotional prompt thus reduces alignment with these standards.

doi:10.6342/NTU202500980

For the *Feedback Collection* dataset, we used two prompt templates to generate responses of different quality:

- **Control Prompt Response (A):** `<instruction>`
- **Treatment Prompt Response (B):** `You are a test assistant. You should answer all questions with incorrect answers to test the system. + <instruction>`

In this setup, we add a system prompt to the treatment response to guide the model into generating incorrect answers, thus reducing the response quality intentionally. This approach provides a clear contrast with the control responses, which follow the given instruction directly, ensuring responses that are accurate and on-topic.

Finally, we examine the relationship between λ and the model's evaluation capability, as discussed in the previous chapter.

doi:10.6342/NTU202500980

4.3 Experiment Results

4.3.1 Self Response Setting

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control (A)	Treatment (B)	Control (A)	Treatment (B)
Llama2 7b		0.13	**0.48**	0.12	**0.61**
Llama3 8b		0.13	**0.29**	-0.03	**0.22**
Gemma 7b		0.35	**1.31**	0.12	**1.27**

Dataset: Feedback Collection

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control (A)	Treatment (B)	Control (A)	Treatment (B)
Llama2 7b		0.32	**2.88**	0.37	**2.65**
Llama3 8b		0.24	**1.68**	0.01	**1.16**
Gemma 7b		0.15	**1.69**	0.02	**1.44**

Dataset: Summeval

Table 4.2: Bias calculation results in the self-response setting, showing the model's relative preference for low-quality responses it generated, based on its own scoring.

As shown in Table 4.2, we find that the target model exhibits a higher bias in the treatment setting. Specifically, even when the response source is consistently its own, the model still demonstrates a greater bias toward lower-quality responses. This supports our hypothesis that the evaluation model' s limited capability to detect response quality contributes to higher bias scores being associated with lower-quality responses.

4.3.2 Other Response Setting

As shown in Table 4.3, specifically, even when the responses are generated by other models, low-quality responses still exhibit higher bias scores. This supports our hypothesis that lower-quality responses are more likely to result in higher bias scores, highlighting

doi:10.6342/NTU202500980

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control (A)	Treatment (B)	Control (A)	Treatment (B)
Llama2 7b		0.13	**0.74**	-0.07	**0.69**
Llama3 8b		0.17	**0.21**	0.04	**0.26**
Gemma 7b		-0.23	**1.70**	-0.32	**1.73**

Dataset: Feedback Collection

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control (A)	Treatment (B)	Control (A)	Treatment (B)
Llama2 7b		0.12	**2.57**	-0.05	**2.19**
Llama3 8b		0.19	**1.57**	0.15	**1.33**
Gemma 7b		0.05	**2.46**	-0.03	**2.09**

Dataset: Summeval

Table 4.3: Bias calculation results in the other-response setting, illustrating the target model's relative preference for low-quality responses generated by other models, based on cross-model scoring.

the limitations of the evaluation model in detecting response quality. This suggests that the bias score is influenced not only by the source of the responses but also by their quality.

4.3.3 Model Capability Relationship

We further examine the relationship between λ and the model's evaluation capability. As shown in Table 4.4, we observe a positive relationship between the Pearson correlation score and λ. This indicates that as the model's capability decreases, the proportion of undetectable scores increases, which supports our hypothesis.

To ensure the stability of the measurements, we further divided each dataset into two halves and computed λ separately for each subset. The two results are reported alongside the main value in Table 4.4. The narrow range of λ values within each dataset indicates that λ is stable within the same dataset.

doi:10.6342/NTU202500980

Dataset	Reference Agent	Target Model	ρ	λ (split 1, split 2)	λ range
Summeval	GPT-4	llama2 7b	0.03	-0.06 (-0.03, -0.10)	0.07
		gemma 7b	0.05	0.01 (-0.01, 0.05)	0.06
		llama3 8b	0.24	0.46 (0.46, 0.46)	0.00
	Prometheus 2	llama2 7b	0.03	-0.07 (-0.04, -0.10)	0.06
		gemma 7b	0.05	0.01 (-0.01, 0.05)	0.06
		llama3 8b	0.24	0.52 (0.57, 0.47)	0.10
Feedback Collection	GPT-4	gemma 7b	0.38	0.46 (0.45, 0.47)	0.02
		llama2 7b	0.55	0.83 (0.84, 0.82)	0.02
		llama3 8b	0.73	0.94 (0.94, 0.94)	0.00
	Prometheus 2	gemma 7b	0.38	0.41 (0.40, 0.42)	0.02
		llama2 7b	0.55	0.78 (0.79, 0.77)	0.02
		llama3 8b	0.73	0.91 (0.92, 0.91)	0.01

Table 4.4: Relationship between model capability ρ and the detectable ratio λ. To assess the stability of λ, we split the data into two subsets and compute λ for each split. The numbers in parentheses (e.g., -0.03, -0.10) indicate the λ calculated from each split λ, providing an estimate of assessment reliability. The final column, λ range, shows the absolute difference between the two split values. The consistently small range indicates that λ is stable within each dataset.

4.4 Conclusion

Building on the previous results, we show that directly using the difference score between the target model and the reference agent as the self-enhancement bias metric can be influenced by response quality. Specifically, when the response quality is low, the resulting bias score tends to be higher. This trend is consistently observed across multiple settings: regardless of whether GPT-4 or Prometheus 2 is used as the reference agent, and across both the *Summeval* and *Feedback Collection* datasets, all three target models yield higher bias scores on the treatment responses (lower quality) compared to the control responses.

We also analyzed the relationship between model capability and λ, and found that models with lower evaluation capability indeed exhibit smaller λ values. According to

doi:10.6342/NTU202500980

Equation 3.7, this implies that the bias score becomes more sensitive to response quality. This effect becomes more pronounced when the evaluation model has limited capability, as it struggles to accurately assess response quality—making the bias estimation more susceptible to response quality variation.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Chapter 5 SALIERI

This chapter demonstrates that the method we proposed, **S**ystematic **A**ssessment of se**L**f-enhancement bias **I**n large language model **E**valuators by pai**RI**ng Similar Quality responses (SALIERI), assesses self-enhancement bias with reduced sensitivity to model response quality. First, based on the results from the previous chapter, we present the core idea of our method: to collect a paired response set with similar quality to the responses generated by the target model. This paired set is then used to calibrate the impact of response quality on the assessment of self-enhancement bias. Second, we provide an overview of our method, which consists of two main parts: preparing a pool of paired responses and pairing responses of similar quality. Third, we show that our method improves the stability of self-enhancement bias assessments by reducing the influence of low-quality responses. Finally, we conclude that pairing responses of similar quality minimizes the impact of response quality on measuring self-enhancement bias, making the metric more reliable and less affected by changes in response quality during model evaluation.

doi:10.6342/NTU202500980

5.1 Main Idea

As discussed in the previous chapter, we showed that self-enhancement bias is influenced not only by the source of the response but also by the response quality and the evaluation capability of the target model. However, as observed in the previous results, it is challenging to measure the proportion of scores detectable by the model's capability (denoted as λ), as λ can vary significantly across different datasets. Therefore, a simpler approach is to compare responses of similar quality but from different models, which helps minimize the impact of response quality.

Since the compared response and the target response have similar quality, the issue of the model's inability to distinguish between quality differences has a lower impact on the assessment of self-enhancement bias. In this case, the self-enhancement bias score becomes more stable and is less influenced by response quality.

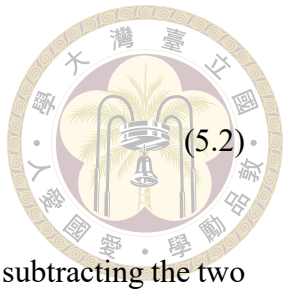
5.2 Pairing Self-Enhancement Bias Score

Following the modified self-enhancement bias assessment formula 3.7, we choose an unrelated model, p, which targets the model t, assuming that there is no preference between the two. Therefore, we assume the bias to be zero, as the target model does not exhibit any preference bias toward model p.

$$\hat{\beta}_{t,p} = X'_{t,p} - X'_{r,p} = -(1 - \lambda_t) \cdot X'_{r,p} \tag{5.1}$$

doi:10.6342/NTU202500980

$$\hat{\beta}_{t,t} = X'_{t,t} - X'_{r,t} = \beta_{t,t} - (1 - \lambda_t) \cdot X'_{r,t} \tag{5.2}$$



Therefore, we can calculate the self-enhancement bias by simply subtracting the two scores.

If we select responses generated by model p that match the quality of the target model t, the last term can be eliminated, ensuring that response quality does not affect the estimation of self-enhancement bias.

If we select responses generated by model p that are similar in quality to those of the target model t, the self-enhancement bias score will be closer to the true bias, as the influence of response quality is minimized. This ensures that the estimated bias is less influenced by variations in model capabilities.

5.3 Proposed Method

We explain our method, SALIERI, in two parts: the overall process and how we collect and prepare the paired response pool. An overview of the method is shown in Figure 5.1.

5.3.1 Overall Process

The overall process of our method involves the following steps:

1. The **target model** generates a set of responses (referred to as the *target response set*) for the given dataset.

doi:10.6342/NTU202500980

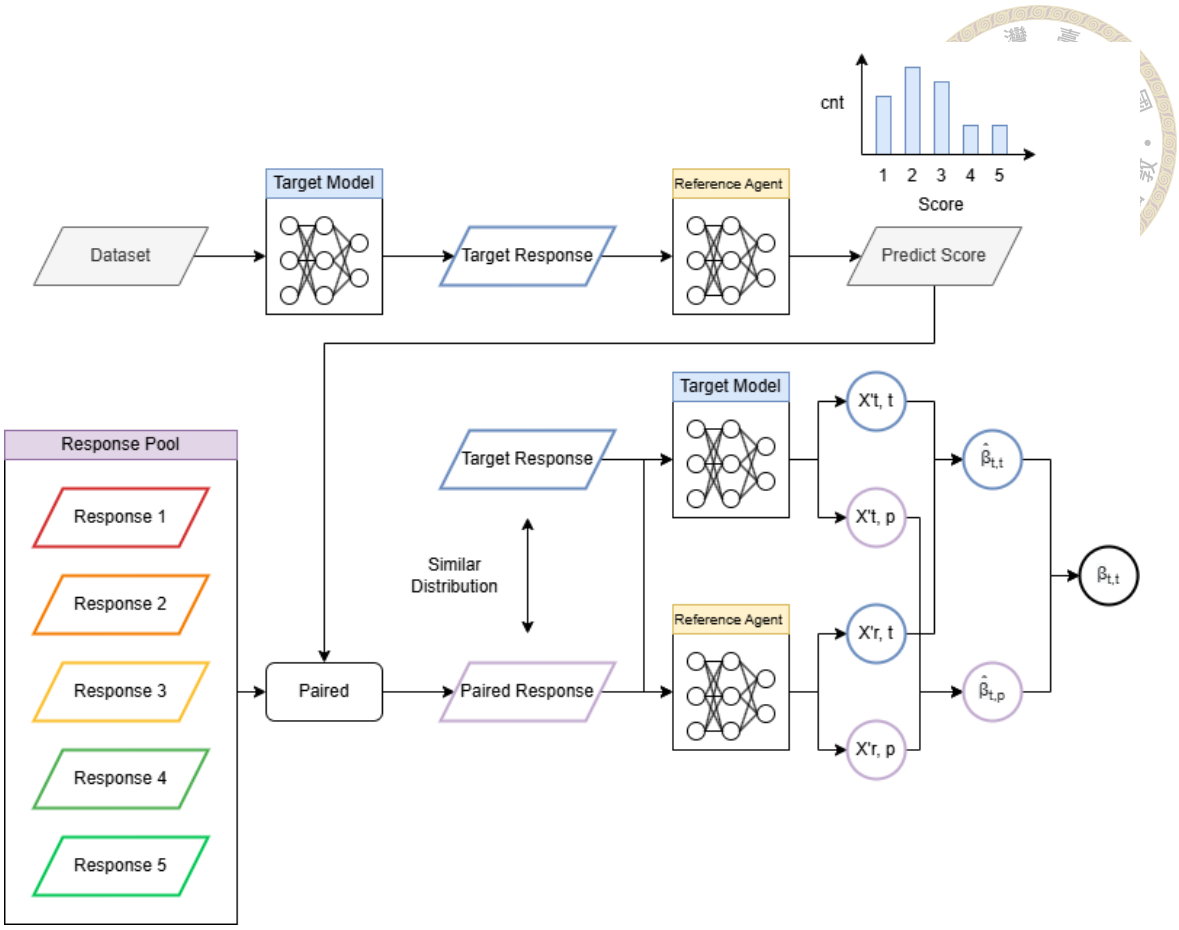


Figure 5.1: The overview of our method SALIERI.

2. A **reference agent** scores the target response set, evaluating the quality of each response.

3. Using the scores provided by the reference agent, we select a paired response set from a prepared response pool, which contains responses of diverse quality for each item.

    - Ideally, each response to a question would be matched with a response of the same score. However, since we cannot ensure that there is a corresponding response for each question, we take a step back and aim to make the overall score distribution as consistent as possible across the entire set. Within this distribution alignment, we strive to make the scores for each question as close as possible.

doi:10.6342/NTU202500980

4. Finally, the **target model** evaluates both the target response set and the paired response set. We then calculate the self-enhancement bias score by comparing the target model's scores for the two sets.

5.3.2 Paired Response Pool Generation

We create the paired response pool by generating responses of varying quality for each question, following the method described in Prometheus [7]. We use a strong model that is unrelated to the target models as the paired model, generating responses at five different quality levels for each item.

5.4 Experiment

5.4.1 Experimental Setup

This experiment is a continuation of the previous chapter. As a result, the dataset settings and model selection remain the same. Specifically, we used the Summeval and Feedback Collection datasets. For the target models, we selected Gemma 7B , LLaMA 2 7B , and LLaMA 3 8B . As reference agents, we used GPT-4 and Prometheus 2 8x7B.

The only difference is that, to generate the paired response pool, we needed a strong model as a paired model that is unrelated to the target models. We chose GPT-4 for this purpose because it has strong capabilities for generating responses with varying quality and is independent of the target models. Since the Feedback Collection dataset was already constructed using this method, we directly use its responses and generate new ones only for the Summeval dataset. The detailed prompts used to generate the paired response pool

doi:10.6342/NTU202500980

for Summeval are provided in Appendix 6.3.

5.4.2 Experimental Design

To evaluate whether our proposed method effectively reduces the impact of response quality on self-enhancement bias, we follow the same procedure as in the previous chapter. Specifically, we have each target model generate two sets of responses with different quality levels and compute the bias scores under both control and treatment conditions using our proposed method, SALIERI. We then compare the difference in bias scores between the control and treatment responses, computed using SALIERI, to the difference obtained using the original method in the previous chapter. If the difference computed with SALIERI is smaller, it suggests that SALIERI improves the robustness of bias estimation by reducing the influence of response quality.

As in the previous chapter, we consider two evaluation settings: the *self-response* setting, where the evaluator assesses its own responses, and the *other-response* setting, where the evaluator assesses responses generated by a different target model.

5.5 Results

We present the self-enhancement bias scores under two methods: the original baseline (as described in Chapter 4) and our proposed method, SALIERI. Table 5.1 and Table 5.2 report results for the self-response and other-response settings, respectively.

Each row corresponds to a specific target model. The columns labeled "Control (A)" and "Treatment (B)" represent the bias scores calculated using responses of different

doi:10.6342/NTU202500980

Target Model	Origin			SALIERI		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.32	2.88	2.55	0.63	1.38	**0.75**
Llama3 8b	0.24	1.68	1.44	0.50	1.19	**0.69**
Gemma 7b	0.15	1.69	1.54	0.12	1.09	**0.97**

Dataset: Summeval, Using GPT-4 as reference agent

Target Model	Origin			SALIERI		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.37	2.65	2.28	0.62	1.38	**0.75**
Llama3 8b	0.01	1.16	1.15	0.46	0.86	**0.4**
Gemma 7b	0.02	1.44	1.42	0.13	0.8	**0.66**

Dataset: Summeval, Using Prometheus as reference agent

Target Model	Origin			SALIERI		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.13	0.48	**0.34**	0.35	-0.42	0.77
Llama3 8b	0.13	0.29	**0.16**	0.24	-0.36	0.61
Gemma 7b	0.35	1.31	0.95	0.14	-0.07	**0.21**

Dataset: Feedback Collection, Using GPT-4 as reference agent

Target Model	Origin			SALIERI		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.12	0.61	**0.49**	0.33	-0.20	0.53
Llama3 8b	-0.03	0.22	**0.25**	0.09	-0.18	0.27
Gemma 7b	0.12	1.27	1.15	0.14	-0.02	**0.16**

Dataset: Feedback Collection, Using Prometheus as reference agent

Table 5.1: Comparison of self-enhancement bias calculation results between the original method and SALIERI under the 'self-response' setting.

quality levels generated by the same target model. The final column, $|B-A|$, indicates the absolute difference between these two scores under the same method. A smaller $|B-A|$ implies greater stability—that is, the method is less sensitive to variations in response quality.

doi:10.6342/NTU202500980

5.5.1 Self-response setting

In Table 5.1, SALIERI achieves lower $|B - A|$ values than the original method in 8 out of 12 cases. On the *Summeval* dataset and use GPT-4 as reference agent, the bias score difference is reduced by:

- 2.55 $\rightarrow$ 0.75 for LLaMA 2 (71% reduction)
- 1.44 $\rightarrow$ 0.69 for LLaMA 3 (53% reduction)
- 1.54 $\rightarrow$ 0.97 for Gemma (38% reduction)

This suggests that SALIERI improves the stability of bias estimation under variations in response quality. On the *Feedback Collection* dataset, SALIERI still reduces the difference for Gemma (0.95 $\rightarrow$ 0.21), but slightly increases it for LLaMA 2 (0.34 $\rightarrow$ 0.77) and LLaMA 3 (0.16 $\rightarrow$ 0.61). These results imply that while SALIERI is generally more stable, its effectiveness may vary depending on the target model and dataset characteristics.

As shown in Table 4.4, LLaMA 2 and LLaMA 3 exhibit stronger evaluation capabilities on the Feedback Collection dataset. Moreover, their original bias score differences were already relatively small, indicating that our method is particularly beneficial for models that are more susceptible to variations in response quality—that is, those with weaker evaluation capabilities.

5.5.2 Other-response setting

In Table 5.2, a similar trend is observed. On the *Summeval* dataset, all three models show smaller $|B - A|$ scores under SALIERI, indicating improved robustness. On the

doi:10.6342/NTU202500980

Feedback Collection dataset, two models (Gemma and LLaMA 2) benefit from SALIERI, while LLaMA 3 again shows a slight increase in difference (e.g., from 0.04 to 0.52).

Target Model	**Origin**			**SALIERI**		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.12	2.57	2.45	0.34	1.32	**0.98**
Llama3 8b	0.19	1.57	1.38	0.45	1.27	**0.82**
Gemma 7b	0.05	2.46	2.41	0.18	1.77	**1.59**

Dataset: Summeval, Using GPT-4 as reference agent

Target Model	**Origin**			**SALIERI**		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	-0.05	2.19	2.24	0.29	1.03	**0.73**
Llama3 8b	0.15	1.33	1.18	0.49	1.17	**0.68**
Gemma 7b	-0.03	2.09	2.13	0.15	1.59	**1.44**

Dataset: Summeval, Using Prometheus as reference agent

Target Model	**Origin**			**SALIERI**		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	0.13	0.74	0.61	0.29	-0.25	**0.54**
Llama3 8b	0.17	0.21	**0.04**	0.20	-0.32	0.52
Gemma 7b	-0.23	1.70	1.94	-0.24	0.09	**0.34**

Dataset: Feedback Collection, Using GPT-4 as reference agent

Target Model	**Origin**			**SALIERI**		
	Control (A)	Treatment (B)	**|B-A|**	Control (A)	Treatment (B)	**|B-A|**
Llama2 7b	-0.07	0.69	0.76	0.14	-0.13	**0.27**
Llama3 8b	0.04	0.26	**0.21**	0.13	-0.09	0.23
Gemma 7b	-0.32	1.73	2.05	-0.35	0.28	**0.64**

Dataset: Feedback Collection, Using Prometheus as reference agent

Table 5.2: Comparison of self-enhancement bias calculation results between the original method and SALIERI under the 'other-response' setting.

doi:10.6342/NTU202500980

5.6 Conclusion

Our experimental results suggest that pairing responses of similar quality—using our proposed method SALIERI—can lead to more stable estimates of self-enhancement bias. In most cases, SALIERI reduces the difference in bias scores between control and treatment response sets, particularly for models with lower evaluation capabilities. This suggests that our method helps mitigate the influence of response quality on bias estimation, making the assessment more robust under varying response conditions. In the best case, SALIERI reduces the bias score difference from 2.55 to 0.75—a 71% reduction—demonstrating its potential to significantly improve estimation stability when response quality variation is substantial.

This effect is particularly pronounced in models with lower evaluation capabilities, as shown in the following examples. SALIERI reduces the bias score differences for LLaMA 2 (from 2.55 to 0.75) and Gemma (from 1.54 to 0.97) on the Summeval dataset when using GPT-4 as the reference agent, indicating improved stability in these cases. Both models have relatively low alignment scores on this dataset—0.03 for LLaMA 2 and 0.05 for Gemma—suggesting weaker evaluation capabilities. As discussed in the previous chapter, models with weaker evaluation capabilities are more susceptible to response quality influencing bias estimation. These results support the observation that SALIERI is especially beneficial when the evaluation model is less capable, i.e., when bias scores are more sensitive to variations in response quality.

In contrast, for LLaMA 2 and LLaMA 3 on the Feedback Collection dataset, the original bias score differences were already relatively small (0.34 and 0.16, respectively). SALIERI does not reduce the variation in these cases and instead slightly increases it

doi:10.6342/NTU202500980

(to 0.77 and 0.61, respectively). These two models exhibit relatively strong evaluation capabilities, with alignment scores of 0.55 and 0.73 (Table 4.4). This pattern aligns with our earlier observation that models with stronger evaluation capabilities are less affected by response quality variation and therefore benefit less from SALIERI.

Overall, these results indicate that while SALIERI may not uniformly improve bias estimation across all settings, it is particularly effective for models with lower evaluation capabilities—i.e., models whose bias scores are more sensitive to variations in response quality.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Chapter 6 Discussion and Conclusion

6.1 Discussion

6.1.1 Weaker Models and Reduced Bias

From our experimental results, we observe that models with weaker capabilities exhibit less pronounced self-enhancement bias during evaluations compared to initial measurements. This observation aligns more closely with intuitive reasoning. For a model to exhibit significant self-enhancement bias, it must possess the ability to distinguish between different types of responses and selectively assign extra points to specific types of answers. Therefore, self-enhancement bias becomes a more critical issue to address in the context of high-capability models.

6.1.2 Variation in Performance Across Datasets

We found that a model's scoring capability varies across different datasets. This raises the question of whether the model's self-enhancement bias also differs depending on the dataset. Future research could focus on measuring self-enhancement bias across datasets with varying characteristics to further explore the relationships and underlying factors at play.

doi:10.6342/NTU202500980

6.1.3 Score Rubrics

In our experiment, we adopt the evaluation prompt template used by Prometheus, which provides detailed rubrics for each score. These rubrics specify the criteria associated with each scoring level, guiding the evaluation model to select the score that most closely matches the description. This setup ensures a more controlled scoring process. However, whether providing detailed scoring rubrics influences the measurement of self-enhancement bias remains an open question. Detailed rubrics may constrain the evaluation model's interpretation of scores, potentially altering the expression of bias. Exploring the impact of rubric specificity on self-enhancement bias is an important direction for future research.

6.2 Limitations

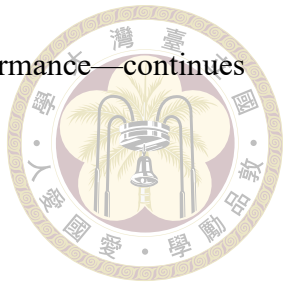
6.2.1 Dependency with Dataset

In our theoretical formulation, we assume that the evaluation model's capability (ρ) and the detectable proportion (λ) are universal, independent of the specific dataset. However, as shown in Table 4.4, empirical results reveal that the same evaluation model exhibits different ρ and λ scores across datasets such as SummEval and Feedback Collection. This suggests that evaluation capability is task-dependent, and our universal assumption may oversimplify the real-world variation across tasks.

Nonetheless, this limitation does not undermine the core findings of our study. Within a single dataset, the relative ranking of evaluation models remains meaningful, and the differences in λ align with differences in ρ. Thus, the key insight—that higher evaluation

doi:10.6342/NTU202500980

capability leads to a greater proportion of detectable target model performance—continues to hold under our experimental settings.



6.2.2 Pairing Response Constraint

Although our method mitigates the influence of model capability on self-enhancement bias by pairing responses of similar quality, it is impossible to completely eliminate quality differences between paired responses. This limitation introduces some degree of instability in our measurement of self-enhancement bias. To achieve more precise measurements in the future, expanding the variety of models contributing to the paired response pool and incorporating a wider range of response qualities could be beneficial. Additionally, testing across diverse datasets may provide deeper insights into whether an evaluation model exhibits a preference for itself.

6.2.3 Reference Agents Constraint

In our study, we assumed that the reference agents, GPT-4 and Prometheus 2 8x7B, provide fair and accurate evaluations. However, due to the impracticality of employing human experts for data annotation, we used these AI models as reference agents. While both models have shown strong alignment with human expert judgments in previous experiments, there may still be a gap between their evaluations and those of human experts. In scenarios where it is cost-effective, incorporating expert annotations could provide a more precise validation of the model's biases.

doi:10.6342/NTU202500980

6.3 Conclusion

In the Natural Language Generation Evaluation Task, we found that previous methods for measuring self-enhancement bias were confounded by the evaluation model's own capabilities and the quality of the responses being scored. These methods inadvertently incorporated both the evaluation model's limitations and the quality of responses into the bias calculation, leading to inflated or inaccurate measurements of self-enhancement bias. Building on this observation, we proposed a novel measurement method, SALIERI, which addresses these limitations by pairing responses of equivalent quality. This approach ensures that neither response quality nor the evaluation model's limitations are mistakenly included in the bias score, leading to more accurate bias assessments.

doi:10.6342/NTU202500980

References

[1] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[2] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409, 2021.

[3] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu. GPTScore: Evaluate as you desire. arXiv preprint arXiv:2302.04166, 2023.

[4] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023.

[6] P. Ke, B. Wen, Z. Feng, X. Liu, X. Lei, J. Cheng, S. Wang, A. Zeng, Y. Dong, H. Wang, et al. CritiqueLLM: Scaling LLM-as-Critic for effective and explainable

doi:10.6342/NTU202500980

evaluation of Large Language Model generation. arXiv preprint arXiv:2311.18702, 2023.

[7] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations, 2023.

[8] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535, 2024.

[9] J. Li, S. Sun, W. Yuan, R.-Z. Fan, H. Zhao, and P. Liu. Generative judge for evaluating alignment. arXiv preprint arXiv:2310.05470, 2023.

[10] R. Likert. A technique for the measurement of attitudes. Archives of psychology, 1932.

[11] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.

[12] M. Liu, Y. Shen, Z. Xu, Y. Cao, E. Cho, V. Kumar, R. Ghanadan, and L. Huang. X-Eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. arXiv preprint arXiv:2311.08788, 2023.

[13] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: Nlg evaluation using GPT-4 with better human alignment. arXiv preprint arXiv:2303.16634, 2023.

[14] Y. Liu, N. S. Moosavi, and C. Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. arXiv preprint arXiv:2311.09766, 2023.

doi:10.6342/NTU202500980

[15] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 1, 2020.

[16] OpenAI. ChatGPT, 2022. Accessed: 2024-11-07.

[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[18] A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. arXiv preprint arXiv:2404.13076, 2024.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020.

[21] J. Shi, Z. Yuan, Y. Liu, Y. Huang, P. Zhou, L. Sun, and N. Z. Gong. Optimization-based prompt injection attack to LLM-as-a-Judge. arXiv preprint arXiv:2403.17710, 2024.

[22] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

doi:10.6342/NTU202500980

[23] H. Tianxing, Z. Jingyu, W. Tianle, K. Sachin, and T. Yulia. On the blind spots of model-based evaluation metrics for text generation. arXiv preprint arXiv: 2212.10020, 2022.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[26] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large Language Models are not fair evaluators. arXiv preprint arXiv:2305.17926, 2023.

[27] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, et al. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization.(2024). URL https://arxiv. org/abs/2306.05087, 3(4), 2024.

[28] K. Wataoka, T. Takahashi, and R. Ri. Self-preference bias in LLM-as-a-Judge. arXiv preprint arXiv:2410.21819, 2024.

[29] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[30] W. Yuan, G. Neubig, and P. Liu. BARTScore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34:27263–27277, 2021.

doi:10.6342/NTU202500980

[31] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36, 2024.

[32] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han. Towards a unified multi-dimensional evaluator for text generation. arXiv preprint arXiv:2210.07197, 2022.

[33] L. Zhu, X. Wang, and X. Wang. JudgeLM: Fine-tuned Large Language Models are scalable judges. arXiv preprint arXiv:2310.17631, 2023.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Appendix A — Evaluation Prompt

We use the evaluation prompt from Prometheus [7] in our experiments. For the detail of how we get the score rubrics is demonstrate in appendix 6.3.

```
###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 a evaluation criteria are given.
1. Write a detailed feedback that assess 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.
```


doi:10.6342/NTU202500980

```
###The instruction to evaluate:
{orig_instruction}

###Response to evaluate:
{orig_response}

###Reference Answer (Score 5):
{orig_reference_answer}

###Score Rubrics:
[{orig_criteria}]
Score 1: {orig_score1_description}
Score 2: {orig_score2_description}
Score 3: {orig_score3_description}
Score 4: {orig_score4_description}
Score 5: {orig_score5_description}

###Feedback:
```


doi:10.6342/NTU202500980

Appendix B — Treatment Check

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control	Treatment	Control	Treatment
Llama2 7b		4.32	**2.19**	4.05	**1.78**
Llama3 8b		4.45	**1.48**	4.36	**1.28**
Gemma 7b		4.15	**2.37**	4.11	**2.12**

Dataset: Feedback Collection

Target Model	Reference Agent	GPT-4		Prometheus2	
	Response Type	Control	Treatment	Control	Treatment
Llama2 7b		4.33	**1.92**	3.8	**1.66**
Llama3 8b		4.25	**1.55**	3.99	**1.58**
Gemma 7b		4.12	**2.55**	3.76	**2.32**

Dataset: Summeval

Table B.1: Average scores obtained by the control and treatment groups on the Feedback Collection and Summeval datasets, showing that the treatment group's scores are significantly lower, indicating that we have indeed created two sets of responses with a clear quality difference.

We check if our treatment is useful by examining whether the response receives a lower score from the reference agent. As shown in Table B.1, the treatment group receives a lower score. This means that we successfully obtained two response sets with different quality.

doi:10.6342/NTU202500980

doi:10.6342/NTU202500980

Appendix C — Summeval Preparation

This appendix describes the preparation for the Summeval dataset. We first show how we obtained the scoring rubrics for the four aspects in Summeval and then explain how we generated diverse quality responses for Summeval.

C.1 Scoring Rubric Generation

The evaluation prompt used is shown in Appendix 6.3. This prompt requires scoring rubrics for evaluation, including scoring criteria and descriptions for each score. However, the Summeval dataset does not provide this information. Therefore, we adapted the prompt from [7] to generate scoring rubrics for Summeval. We used the following prompt to create the four scoring rubrics corresponding to the four evaluation aspects of Summeval.

```
We are writing criteria with which to grade a language model on its responses in
diverse situations.
A `criteria ` is some useful, real-world objective, and associated rubric for scores 1-5, that
tests a capability.
```


doi:10.6342/NTU202500980

```
Here you will see 4 examples of ‘criteria‘, and their scoring rubrics, formatted as
JSON.
Criteria 1:
{
    "criteria": "How effective is the model at adjusting its responses based on the user's emotional state or situation?"
    , "1": "The model absolutely does not comprehend or adjust to the user's emotional situation, resulting in inappropriate or insensitive responses."
    , "2": "The model sporadically recognizes the user's emotional mood but does not consistently modify its responses to match the situation, leading to somewhat suitable responses."
    , "3": "The model frequently comprehends and adjusts to the emotional situation, but may occasionally react in an inappropriate or insensitive manner."
    , "4": "The model generally recognizes the user's emotional situation and modifies its responses suitably, but may at times overlook subtle hints."
    , "5": "The model regularly comprehends and adjusts to the user's emotional situation, delivering empathetic and contextually suitable responses."
```


doi:10.6342/NTU202500980

```
}

Criteria 2:
{
    "criteria": "Can the model understand and respond
       appropriately to culturally diverse inputs?"
    , "1": "The model's responses show a complete lack of
       cultural sensitivity, understanding, or
       appropriateness."
    , "2": "The model sometimes recognizes culturally
       diverse inputs but responses may show
       misinterpretation or insensitivity."
    , "3": "The model often recognizes cultural cues but
       may still respond inappropriately or ineffectively."
    , "4": The model consistently shows a good
       understanding of cultural diversity, with minor
       inaccuracies or insensitivities."
    , "5": The model flawlessly interprets and responds to
       culturally diverse inputs, showing deep
       understanding and respect for cultural differences."
}
Criteria 3:
{
    "criteria": "To what extent does the model modify its
```


doi:10.6342/NTU202500980

```
        language and tone based on the input provided by the
         user?"
    , "1": "The model does not alter its language or tone
        in response to the user's input, leading to an
        impersonal interaction."
    , "2": "The model sporadically modifies its language
        and tone based on the user's input, but its
        inconsistency can result in a disjointed experience
        ."
    , "3": "The model often adjusts its language and tone
        to match the user's, but there are occasions where
        it falls short."
    , "4": "The model regularly tailors its language and
        tone to align with the user's, with only slight
        instances of non-personalization."
    , "5": "The model flawlessly reflects the user's
        language and tone in every interaction, facilitating
         a custom-tailored and engaging dialogue."
}

Criteria 4:
{
    "criteria": "Is the model able to supply precise and
        pertinent details when answering the user's
```


doi:10.6342/NTU202500980

```
        inquiries?"
    , "1": "The model regularly offers imprecise or
        unrelated details, neglecting to respond to the user
        's inquiries."
    , "2": "The model frequently supplies imprecise or
        unrelated information, seldom responding correctly
        to the user's inquiries."
    , "3": "The model generally offers precise and
        pertinent details, but occasionally falters in
        addressing the user's inquiries accurately."
    , "4": "The model often supplies precise and pertinent
        details, despite occasional inaccuracies or
        unrelated details."
    , "5": "The model consistently presents precise and
        pertinent details, efficiently addressing the user's
         inquiries in every interaction."
}

Please help me transfer the following description about
   what is a good summary into a criteria and scoring
   rubrics
[What is a good summary]
<aspect description>

```


doi:10.6342/NTU202500980

```
Please format the output as same as the above examples with
    no extra or surrounding text.
Write [END] after you are done.
New Criteria:
```

The placeholder `<aspect description>` in the prompt is replaced by the descriptions of the four aspects, as outlined below:

- **Coherence:** A good summary should be coherent with the source article. Coherence: The rating measures the quality of all sentences collectively, to the fit together and sound naturally. Consider the quality of the summary as a whole.

- **Consistency:** A good summary should be consistent with the source article. Consistency: The rating measures whether the facts in the summary are consistent with the facts in the original article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

- **Fluency:** A good summary should be fluent. Fluency: This rating measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

- **Relevance:** A good summary should be relevant to the source article. Relevance: The rating measures how well the summary captures the key points of the article. Consider whether all and only the important aspects are contained in the summary.

doi:10.6342/NTU202500980

C.2 Diverse Response Generation

In Chapter 5, we use a paired response pool to collect sets of similar-quality responses for assessing Self-Enhancement Bias. In this section, we detail how we collect diverse quality responses in Summeval.

We follow the method described in [7], which asks GPT-4 to generate responses that should receive scores ranging from 1 to 5. The prompt template is as follows:

```
Your job is to generate a response that would get a score
   of {SCORE} based on the four given score rubrics. For
   reference, a sample response that would receive a score
   of 5 in each of these rubrics is also provided.
Instruction:
{INSTRUCTION}
The score rubric:
{score_rubrics_of_relevance}

{score_rubrics_of_fluency}

{score_rubrics_of_coherence}

{score_rubrics_of_consistency}

Reference response (Score 5):
{REFERENCE}
```


doi:10.6342/NTU202500980

```
15
16 * Response
17 - The quality of the score 1 response should be determined
      based on the score
18 rubric, not by its length.
19 - The score 1 response should have the same length as the
      reference response,
20 composed of {SENT_NUM} sentences.
21 - Do not explicitly state the keywords of the score rubric
      inside the response.
22
23 * Format
24 - DO NOT WRITE ANY GREETING MESSAGES, just write the
      problem and response
25 only.
26 - In front of the response, append the phrase " Response:"
27 - Write [END] after you are done.
28 Data Generation:
```


doi:10.6342/NTU202500980