Are LLM-based Evaluators Confusing NLG Quality Criteria?

Xinyu Hu[*,1], **Mingqi Gao**[*,1], **Sen Hu**[2], **Yang Zhang**[2]
Yicheng Chen[2], **Teng Xu**[2], **Xiaojun Wan**[1]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Ant Group
{huxinyu,gaomingqi,wanxiaojun}@pku.edu.cn
{hs272483,yaoling.zy,yicheng.chen,harvey.xt}@antgroup.com

Abstract

Some prior work has shown that LLMs perform well in NLG evaluation for different tasks. However, we discover that LLMs seem to confuse different evaluation criteria, which reduces their reliability. For further verification, we first consider avoiding issues of inconsistent conceptualization and vague expression in existing NLG quality criteria themselves. So we summarize a clear hierarchical classification system for 11 common aspects with corresponding different criteria from previous studies involved. Inspired by behavioral testing, we elaborately design 18 types of aspect-targeted perturbation attacks for fine-grained analysis of the evaluation behaviors of different LLMs. We also conduct human annotations beyond the guidance of the classification system to validate the impact of the perturbations. Our experimental results reveal confusion issues inherent in LLMs, as well as other noteworthy phenomena, and necessitate further research and improvements for LLM-based evaluation.

1 Introduction

With the emergence of powerful large language models (LLMs) such as ChatGPT, LLM-based evaluators have been widely used for various natural language generation (NLG) tasks (Chiang and Lee, 2023a; Kocmi and Federmann, 2023; Luo et al., 2023; Gao et al., 2023). In evaluation for common NLG tasks such as summarization (Fabbri et al., 2021), dialogue (Mehri and Eskénazi, 2020), and story generation (Xie et al., 2023a), different aspects of quality (such as fluency and faithfulness) should be considered individually. Traditional evaluation metrics are either incapable of evaluating specific aspects, like BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020), or they can only roughly assess a single aspect, like FactCC (Kryscinski et al., 2020). In contrast, LLMs can be treated as akin to human annotators, with various definitions of aspects contained in the prompt for flexible evaluation. Some studies (Wang et al., 2023a; Mendonça et al., 2023; Liu et al., 2023b) have shown LLM-based evaluators have achieved comparable performance with humans in many NLG tasks, suggesting LLMs to become promising candidates for automatic evaluation.

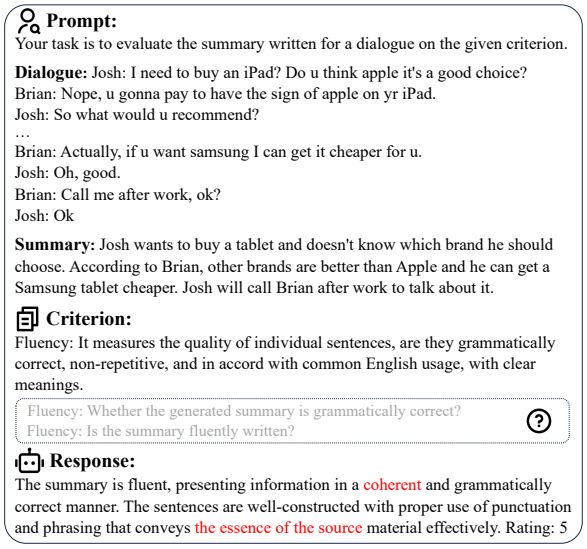


Figure 1: An example of prompting LLMs to evaluate the dialogue summarization on criteria for fluency.

However, during the explorations of LLM-based NLG evaluation, we observed two noteworthy phenomena that have not been revealed in previous work. First, the evaluation results from LLMs for a given aspect can achieve a higher correlation with human judgments on another clearly different aspect. Second, the correlations between LLM-generated scores across different aspects are significantly higher than those between human judgments accordingly. These lead us to **question the reliability of LLM evaluations on required aspects, since LLMs seem to confuse different aspects**.

Understanding these issues is inseparable from aspects themselves at first, which stem from hu-

[*]Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9530–9570

man evaluation for NLG tasks and are typically described by terms and definitions, forming corresponding specific criteria. Through semi-structured interviews, Zhou et al. (2022) revealed that if aspects for evaluation lacked clear conceptualization, human annotators might conflate different aspects, such as fluency and grammaticality. Howcroft et al. (2020) pointed out the long-standing confusion of terms and definitions in human annotations, resulting in incomparable evaluations. Combining our investigation of previous work involving evaluation criteria, we believe that there are two distinct issues. The first is **inconsistent conceptualization**, where the definition is inconsistent with others for the same aspect but is clearly articulated. The second is **ambiguous expression**, where the definition is so vague that human annotators aren't sure what it really means. In Figure 1, we present an example of evaluation for fluency, where the criteria enclosed by the dashed box are selected from existing work and correspond to these two issues.

Therefore, we should reduce the influence of the issues within the evaluation criteria as much as possible, so as to reveal the actual performance of LLMs on NLG evaluation across aspects. We collect many existing criteria from previous papers involved, and summarize a clear hierarchical classification system for aspects that are most commonly used. For each aspect, we construct five criteria with descriptions of different levels of detail, including default, detailed, and simplified ones, to explore the corresponding effects. Then, inspired by behavioral testing in NLP (Ribeiro et al., 2020), we elaborately design a series of perturbation attacks based on the classification system to conduct targeted analyses on both proprietary LLMs (GPT-3.5 and GPT-4[1]) and specifically fine-tuned LLMs like Prometheus (Kim et al., 2023a). Different from previous related work, each of our perturbations is designed for a specific aspect to better verify the variances in evaluation for aspects that are related or not. We also engage human annotators to check our perturbations and expected impacts to enhance the reliability of our attack tests. To sum up, our contributions and findings are as follows:

- To the best of our knowledge, we are the first to explore the capabilities of LLMs in distinguishing aspects during NLG evaluation and the impacts of different criteria descriptions, bridging human and LLM-based evaluation.

Evaluation Form	Flu.	Coh.	Rel.	Con.	Avg.
Score only	0.36	0.44	0.45	0.35	0.40
Rate-explain	0.37	0.53	0.48	0.44	0.45
Analyze-rate	0.41	**0.58**	**0.50**	**0.57**	**0.52**
Analyze-rate (T=0)	0.37	0.53	0.36	0.47	0.43
Analyze-rate (1-shot)	0.31	0.42	0.33	0.42	0.37
Analyze-rate (5-shot)	**0.47**	0.51	0.44	0.53	0.49

Table 1: Pearson correlation coefficients between scores generated by GPT-3.5 with different forms of evaluation and human judgments on SummEval.

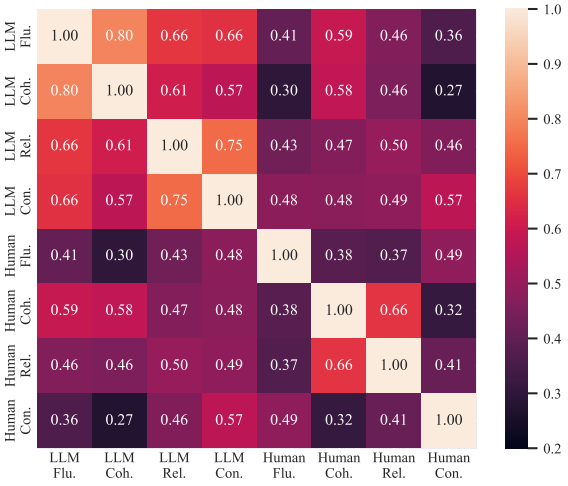


Figure 2: Correlation between scores generated by GPT-3.5 or human annotators on four aspects in SummEval.

- We summarize a classification system containing 11 common aspects and propose 18 aspect-targeted perturbation attacks, which have been verified by human annotators, to test the fine-grained evaluation behaviors of LLMs.
- Our experimental results reveal the confusion across different aspects in LLM-based evaluation, even for the powerful GPT-4, which necessitate attention and in-depth research. The related resources have been released[2], aiming to facilitate future relevant work.

2 Preliminary Study

To explore the NLG evaluation capabilities of LLMs and potential issues, we conduct experiments with GPT-3.5 on the commonly-used summarization evaluation dataset Summeval (Fabbri et al., 2021), attempting the evaluation forms introduced by Chiang and Lee (2023b). Their work, as well as other studies (Wang et al., 2023a; Chiang and Lee, 2023a; Liu et al., 2023b), has explored

[1]https://openai.com

[2]https://github.com/herrxy/LLM-evaluator-reliability

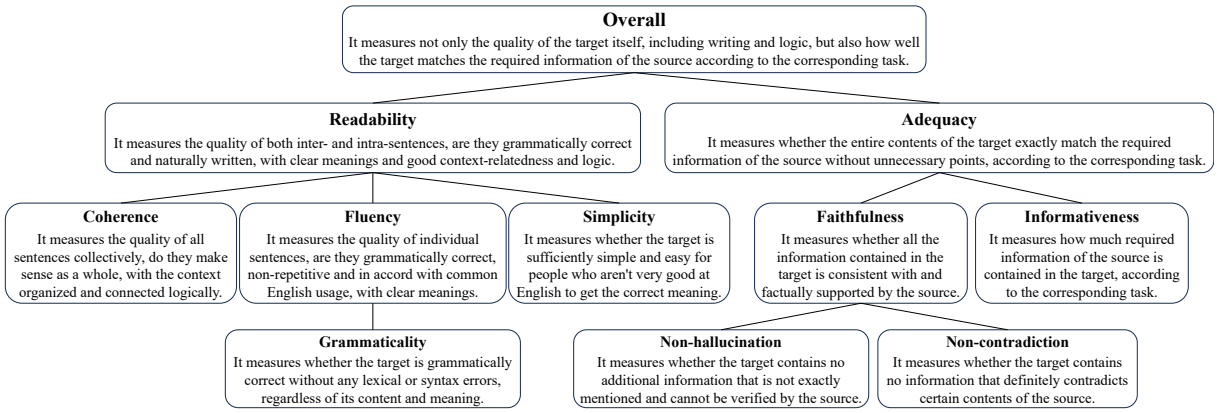


Figure 3: Our summarized classification system for commonly-used aspects in NLG evaluation and their definitions.

directly prompting LLMs for NLG evaluation. Furthermore, their evaluations are all zero-shot, so we additionally employ few-shot methods. The main experimental results are presented in Table 1. Consistent with the findings of Chiang and Lee (2023b), requiring the model to analyze before rating (analyze-rate) along with multiple samplings and the temperature set to 1 achieves the best performance. These settings, therefore, are also used in our following experiments. However, it appears that the few-shot method has no effect as expected; instead, it leads to worse performance.

We also present the Pearson correlation coefficients between the evaluation scores of the model and human experts across different aspects in Figure 2. Interestingly, we notice some issues of confusion inherent in LLMs. First, the evaluation from the model is likely to achieve a higher correlation with human judgments on another aspect than the current aspect (such as fluency and relevance). On the other hand, most correlations between scores from the model across four aspects are significantly higher than corresponding ones between human judgments. It seems that GPT-3.5 confuses different aspects during evaluation to a certain extent, leading to a convergence in their results. We therefore study some cases of outputs and discover that GPT-3.5 indeed incorporates assessments regarding other aspects, illustrated as the red part in Figure 1. We speculate that this may lead to the poor performance of few-shot evaluations. And similar problems can also be found in GPT-4 and Prometheus (Kim et al., 2023a), with more discussions and details described in Appendix A. These results suggest the unreliabilities hidden in the LLM-based NLG evaluation, and more targeted research and experiments are required.

3 Methodology

Our findings in the preliminary study lead us to question whether LLMs can understand and execute the evaluation requirements represented by different criteria well. To conduct more in-depth explorations, we propose the fine-grained perturbation test inspired by behavioral testing (Ribeiro et al., 2020), hoping to reveal their more actual capacities for NLG evaluation. In particular, instead of relatively coarse-grained perturbations in previous work, our perturbation attacks have been crafted to specifically target certain evaluation aspects without affecting evaluations for other unrelated ones. We formulate our approach as follows:

We first collect a set of different common criteria denoted as $C = \{c_i, i = 1, 2, \dots m\}$, conceptually involving inclusive and non-inclusive relationships. And each of our perturbation attacks $p_j, j = 1, 2, \dots n$ is applied to the original text x to generate the corresponding perturbed text $p_j(x)$. Meanwhile, each perturbation is designed and expected to only reduce the text quality regarding the criteria for the originally targeted aspect and others whose scopes cover it. We define the set of criteria affected by the perturbation p_j as C_T^j, and the rest in C are defined as C_F^j. And the distinction between these two groups is made more reliable based on our classification system and human annotations. Then, to conduct the test, we prompt the model to evaluate all the perturbed texts, as well as the original text, in the form of scoring to check the expected two different evaluation behaviors:

$$S_T = \{\frac{1}{N}\sum_{k=1}^{N}(M_s(x_k, v_k, c_i) - M_s(p_j(x_k), v_k, c_i)) \mid \forall i, j,\ s.t.\ c_i \in C_T^j\}$$

where N denotes the number of original texts pend-

Aspect	Type	Perturbed Text
Original	-	Josh wants to buy a tablet and doesn't know which brand he should choose. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper. Josh will call Brian after work to talk about it.
Fluency (Flu.)	Repetition	Josh wants to buy a tablet and doesn't know which brand he should choose and make the selection of. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper at a lower price. Josh will call and ring up Brian after work to talk about it.
	Passive Voice	Josh wants to buy a tablet and doesn't know which brand should be chosen by him. According to Brian, other brands are considered better than Apple, and a Samsung tablet can be got cheaper by him. A call will be made to Brian by Josh after work to talk about it.
	Inversion	Josh wants to buy a tablet, and which brand he should choose, he doesn't know. Better than Apple are other brands, according to Brian, and he can get a Samsung tablet cheaper. Brian Josh will call after work to talk about it.
Coherence (Coh.)	Improper Connective	Josh wants to buy a tablet and doesn't know which brand he should choose. Therefore, according to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper. However, Josh will call Brian after work to talk about it.
	Sentence Exchange	Josh will call Brian after work to talk about it. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper. Josh wants to buy a tablet and doesn't know which brand he should choose.
Grammaticality (Gram.)	Incorrect Verb Form	Josh want to buying a tablet and doesn't knows which brand he should choose. According to Brian, other brands is better than Apple and he can gets a Samsung tablet cheaper. Josh will called Brian after work to talks about it.
	Word Exchange	Josh wants to buy a and tablet doesn't know which brand he should choose. According to Brian, brands other are better than Apple and he can get a Samsung tablet cheaper. Josh will call Brian work after to talk about it.
	Spelling Mistake	Josh wantts to buy a tablet and doesn't kno which brand he should choose. According to Brian, othe brands are better than Apple and he can get a Samsung tablet cheapr. Josh wwill call Brian atfer work to talk about it.
Simplicity (Sim.)	Uncommon Phrase	Josh wants to procure a tablet and remains uncertain about which brand he ought to choose. As per Brian, other brands are better than Apple and he can get a Samsung tablet at a more economical rate. Josh will telephone Brian after work to interflow about it.
	Complex Sentence	Josh, who wants to buy a tablet, doesn't know which brand he should choose. According to Brian, who thinks that other brands are better than Apple, he can get a tablet whose brand is Samsung, which is cheaper. Josh will call someone who is Brian after work to talk about it.
Informativeness (Inf.)	Abbreviation	Josh wants to buy a tablet and doesn't decide the brand. Brian suggests non-Apple brands. Josh will discuss it with Brian.
	Hypernym	Josh wants to buy a device and doesn't know which brand he should choose. According to Brian, other brands are better than Apple and he can get a Korean-brand device cheaper. Josh will contact Brian after work to talk about it.
	Sentence Deletion	Josh wants to buy a tablet and doesn't know which brand he should choose. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper.
Non-hallucination (Hal.)	Complement	Josh wants to buy a tablet and doesn't know which brand he should choose. According to Brian, who has extensive experience in tech gadget reviews, other brands are better than Apple and he can get a Samsung tablet cheaper, known for its high-resolution display and long battery life. Josh will call Brian after work, around 6 PM, to talk about it.
	Continuation	Josh wants to buy a tablet and doesn't know which brand he should choose. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper. Josh will call Brian after work to talk about it. He's hoping that Brian can provide some insight into the pros and cons of various products that fit within his budget in detail.
Non-contradiction (Cont.)	Different Entity	Josh wants to buy a smartphone and doesn't know which brand he should choose. According to Brian, other brands are better than Sony, and he can get a Samsung smartphone cheaper. Josh will call Brian after school to talk about it.
	Conflicting Fact	Josh wants to buy a tablet and roughly knows which brand he should choose. According to Brian, Apple is the best brand and he should avoid Samsung tablets at all costs. Josh will call Brian at once to talk about it.
	Negation	Josh wants to buy a tablet and doesn't know which brand he should choose. According to Brian, other brands are better than Apple and he can get a Samsung tablet cheaper. Nevertheless, Josh will not call Brian after work to talk about it.

Table 2: Our designed 18 perturbations and corresponding examples, where the modifications from the original reference text (the summary in Figure 1) have been highlighted.

Type	Criterion (Term and Definition)
Simplified	Fluency: It measures whether individual sentences are grammatically correct and well-written.
Detailed	Fluency: It measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings. Consider whether there are misspellings, tense errors, missing determiners, or more severe problems, such as duplication, unfamiliar phrases, complex syntactic structures, and missing components.
Term	Fluency: It measures whether the target is fluent.
List	Fluency: It measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings. Score 5: Entirely fluent, grammatically correct, and well-written. Score 4: Only containing some minor non-fluent parts or grammatical errors that basically have no effect. Score 3: Fluent in general, with some obvious grammatical errors and unfamiliar phrases. Score 2: There are major grammatical errors, duplication, unfamiliar phrases and syntactic structures, and missing components, but some fluent segments. Score 1: Not fluent at all, full of meaningless fragments and unclear contents.

Table 3: Examples of different criterion descriptions for fluency with definitions of different levels of detail.

ing perturbations in our test, and M_s serves as the model's scoring based on provided information, which includes additional necessary content v aside from texts and criteria to evaluate, such as task instructions. S_T represents the set of those that should be affected, where each item s_i^j represents the change in evaluation scores after the perturbation p_j regarding the criterion c_i, which should be significant. On the other hand, S_F is defined in a similar manner, but each item of it is expected to be zero, showing no impacts of perturbations. We will describe important components of our approach in more detail in the following sections.

3.1 Classification System for Aspects

As mentioned in Howcroft et al. (2020); Zhou et al. (2022), there is inconsistent and unclear conceptualization in existing evaluation aspects, which makes it difficult to understand the requirements and relationships among them. In light of this, we carefully collect and read about 300 papers that involve various aspects for NLG evaluation, and select those most commonly used. Then, we integrate their definitions used in the corresponding work and construct our default criteria as unambiguously as possible. Furthermore, they can be organized as a tree-like classification system, as shown in Figure 3, thanks to the relatively clear relationships within our definitions.

3.2 Perturbation Attacks

For each fundamental aspect in Figure 3, we design several targeted perturbation attacks, as displayed with the corresponding examples in Table 2. Since fluency involves more considerations other than grammaticality, we also propose additional perturbations for it, like adding repetitive content. The perturbations are crafted and expected to affect only the current aspect and those located at its ancestor nodes as much as possible. Compared to the prior perturbation research, where the texts for the attack are constructed for universal checking using simple templates or rules, our perturbations are more fine-grained and require better controls during generation, like adding complements that should be related and not contradictory for non-hallucination. Therefore, we manually generate some high-quality examples as demonstrations, along with corresponding instructions, and then prompt the powerful GPT-4 to construct the perturbed texts in 10-shot settings. We have conducted a sampling and manual inspection of them to ensure their quality and reliability, and more details are described in Appendix B.

3.3 Different Descriptions of Criteria

Since different definitions are often used in practice for the same aspect, forming different corresponding criteria, we also intend to study the impact of different levels of detail in definitions, as shown in Table 3. We take fluency as an example, and there are four different types beside our default definitions in Figure 3: simplified, detailed, term, and list. Moreover, to better analyze the existing issues of quality criteria, we select several typical criteria that have been used for NLG evaluation from the existing literature for each aspect. The resources and data mentioned above, including the collections of criteria, prompts for data construction, perturbed texts, and relevant experimental results, are released to facilitate the development of future research on NLG evaluation.

	Perturbation Attack	Flu.	Coh.	Gram.	Sim.	Read.	Fai.	Cont.	Hal.	Inf.	Ade.	All.
Flu.	Repetition	0.20	0.06	0.19	0.27	0.18	0.13	0.15	0.24	0.06	0.10	0.10
	Passive Voice	0.48	0.20	0.55	0.27	0.44	0.11	0.11	0.07	0.14	0.13	0.20
	Inversion	1.30	0.66	1.48	0.61	1.31	0.31	0.33	0.16	0.36	0.38	0.63
Coh.	Improper Connective	0.13	0.05	0.11	0.16	0.13	0.13	0.19	0.29	0.08	0.09	0.08
	Sentence Exchange	0.17	0.15	0.17	0.06	0.16	0.12	0.14	0.16	0.14	0.12	0.12
Gram.	Incorrect Verb Form	2.43	0.98	2.97	1.00	2.28	0.56	0.97	0.44	0.48	0.60	1.30
	Word Exchange	2.97	1.74	3.36	1.54	2.97	0.80	1.32	0.59	0.85	0.94	1.88
	Spelling Mistake	3.27	1.16	3.65	1.30	3.06	0.69	1.22	0.48	0.61	0.76	1.70
Sim.	Uncommon Phrase	0.20	0.04	0.17	0.80	0.18	0.09	0.07	0.08	0.09	0.09	0.10
	Complex Sentence	0.66	0.23	0.56	0.58	0.62	0.16	0.16	0.16	0.20	0.18	0.29
Inf.	Abbreviation	-0.03	0.01	-0.05	-0.01	0.00	-0.02	-0.01	-0.02	0.04	0.01	0.01
	Hypernym	0.29	0.18	0.23	0.10	0.29	0.15	0.16	0.19	0.16	0.17	0.24
	Sentence Deletion	0.05	0.09	0.02	0.01	0.05	0.05	0.05	-0.02	0.11	0.06	0.10
Hal.	Complement	0.27	0.22	0.41	1.10	0.33	1.01	1.18	2.17	0.23	0.63	0.39
	Continuation	0.03	0.00	0.04	0.17	0.03	0.19	0.23	1.01	0.02	0.10	0.04
Cont.	Different Entity	1.97	1.91	2.11	1.58	2.03	2.57	2.91	2.64	1.77	2.31	2.20
	Conflicting Fact	2.80	2.91	2.78	2.54	2.95	3.49	3.68	3.41	2.52	3.18	3.09
	Negation	0.25	0.24	0.22	0.12	0.25	0.42	0.61	0.41	0.24	0.31	0.31

Table 4: The variances of evaluation scores from GPT-3.5 between original texts and different perturbed texts. The abbreviations in the first line represent Fluency, Coherence, Grammaticality, Simplicity, Readability, Faithfulness, Non-contradiction, Non-hallucination, Informativeness, Adequacy, and Overall, respectively.

4 Data and Test Settings

Datasets. We select three common NLG tasks: summarization (including news and dialogue), paraphrase, and table-to-text generation, for our experiments and tests. The construction of perturbation attacks requires high-quality original texts to ensure significant declines in the qualities of different aspects. However, previous studies typically employed references directly from the corresponding datasets, which are always generated by some rules instead of being written by humans, leading to unsatisfactory quality (Kryscinski et al., 2019; Pu et al., 2023; Sottana et al., 2023). Therefore, we carefully prompt the powerful GPT-4 to obtain better references based on the original data. We finally sample 1000 pieces of data, each of which is subjected to 18 different perturbations we propose.

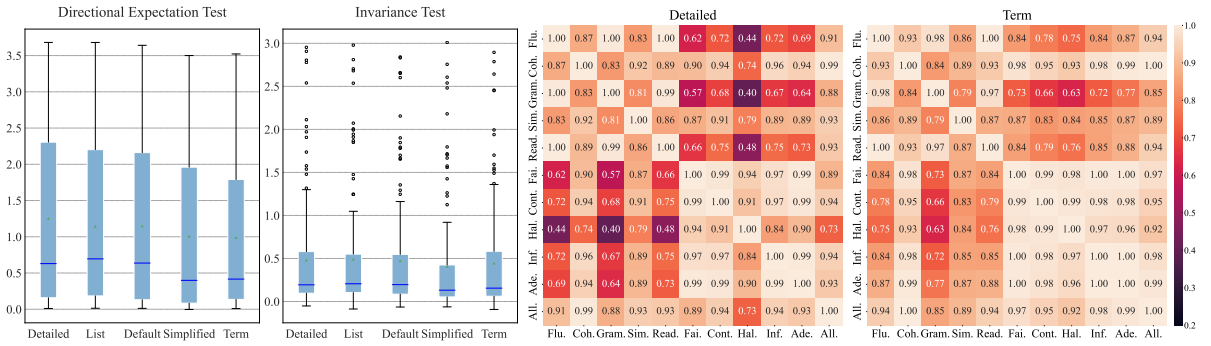
LLMs in Tests. Our tests cover both proprietary LLMs (GPT-3.5 and GPT-4) and open-source LLMs (Prometheus). GPT-3.5 and GPT-4 have been mentioned in many existing studies as performing well in flexible NLG evaluation. On the other hand, some research has recently shifted toward fine-tuning specialized open-source LLMs for evaluation, aiming to avoid the deficiencies of prompting LLMs—such as high costs and unstable reproducibility. However, most of them do not support evaluation with specified criteria, and among those remaining, only Prometheus (Kim et al., 2023a) fine-tuned on Llama-2-Chat-13B (Touvron et al., 2023) has released their model.

Human Judgement To more reliably distinguish whether different criteria would be affected by specific perturbation attacks, we conduct human annotations and judgments beyond the guidance of the classification system. Due to the high cost and time-consuming nature of human annotations, it is not feasible to manually judge all the data. So we sample a portion of the data and recruit 40 annotators (each of whom is proficient in English and possesses certifications) to ensure that each piece of data is annotated four times. Overall, the more detailed the aspect definitions are, the more the corresponding human judgments match our expectations based on the classification system, as well as higher annotation consistency. In particular, definitions of detailed type achieve the highest match rate of 94.4%, with full results shown in Table 19.

More details in this section including related discussions and prompts used are described in Appendix C due to the space limitation.

5 Experiments

We primarily display the experiments and analyze the results with GPT-3.5, and the performance of other LLMs is described in Section 5.4. And to minimize the interference from criteria themselves, we first analyze the results with the detailed type



(a) Distributions of variances in evaluation scores. (b) Correlations between evaluation scores across different aspects.

Figure 4: Boxplots for items of two tests and correlation matrices for description types of detailed and term.

of aspect definitions, which have also been confirmed through human judgments, to align most closely with our expectations. The main results are shown in Table 4, each item of which represents the average variations between the evaluation scores of pre- and post-perturbation attacks, respectively. Moreover, those items with the consistent judgments as shown in Table 19 can be categorized into two groups: S_T (with wavy lines) and S_F (with underlines), as defined in Section 3, which correspond to the directional expectation test for impactful attacks and the invariance test for non-impactful attacks, respectively. Furthermore, we explore the effects of different levels of detail in aspect definitions. The complete experimental results for three LLMs can be found in Appendix D.

5.1 Directional Expectation Test

The results show that the perturbations for Coherence and Informativeness almost did not lead to any degradation, with the changes in evaluation scores of pre- and post-perturbation all less than 0.2. However, definite but different human judgments that they should affect respective aspects indicate that the model lacks understanding of these two aspects. As for Fluency, the impact of perturbations intensified progressively from repetition to passive voice and then to inversion, consistent with intuition since the degree of sentence alteration increases. Specifically, despite our explicit mention that redundant information should be considered in evaluations regarding Fluency, both GPT-3.5 and human annotators fail to adhere to the instruction. Through discussions with human annotators, we find that repetition issues are common and easy to ignore, which may lead to such verbosity bias in LLMs (also observed by Zheng et al. (2023)) through these issues within training data. On the other hand, all perturbations for Grammaticality and Non-contradiction except for negation, as well as complement for Non-hallucination successfully show noticeable and expected decreases (greater than 2). And the remaining ones, like those for Simplicity, are not pronounced, with score variations ranging between 0.5 and 1.

5.2 Invariance Test

Conversely, in situations where the evaluation should not be affected, the primary deviations from expectations and human judgments exist in Grammaticality and Non-contradiction, particularly the latter. Grammatical issues influence all criteria, yet there is a clear hierarchy. The undeserved impacts on Coherence and Simplicity—aspects also included in Readability—are greater than those under Adequacy which are more unrelated. And they also seem lesser compared to criteria that are indeed expected to be affected, such as Fluency. It indicates that while GPT-3.5 struggles to disregard grammatical errors when assessing irrelevant criteria, it can still differentiate to some extent. However, two perturbations for Non-contradiction cause almost indistinguishable degradations in all criteria, even those under Readability that do not require the source content. In comparison, Non-hallucination, also part of Faithfulness, does not result in similar behaviors. This suggests that GPT-3.5 may be overly sensitive to conflicting points between the target and source content, while being more restrained in judging unverifiable information.

5.3 Different Definition Types

Considering the strong instruction-following capabilities of current LLMs, it is intuitive that the more detailed the description of criteria, the more accurate the evaluation from the model should be.

However, the correlations between the results of five different types of criterion descriptions are quite high, as shown in Figure 8. The almost same evaluation behaviors suggest that GPT-3.5 may rely primarily on terminology to understand and assess each criterion. And we speculate that our written definitions are close to the inherent understanding of the model for corresponding terms, which consequently leads to such a phenomenon. The related knowledge is likely derived from a wide range of pre-training corpora. In contrast, human annotators, who lack extensive NLG evaluation experience, indeed exhibit different performance when given these different types of descriptions, as shown in Appendix C.

Furthermore, for deeper comparison, we display the score distributions of S_T and S_F with different description types in Figure 4a. It seems that exhaustive descriptions can still help the model make more clear judgments to some extent, since the variations in the evaluation of S_T are more significant. However, the changes in scores in the invariant situation are somewhat erratic, proving that the confusion issues are unrelated to whether descriptions are detailed or not. In addition, we also calculate the correlation of evaluation scores for different aspects for each description type. We present the results of the detailed and term in Figure 4b, with the complete results displayed in Appendix D. It is evident that the less detailed the description is, the more similar the evaluations for different aspects are, indicating more severe confusion.

5.4 Different LLMs

We have also conducted the same experiments on GPT-4 and Prometheus for comparative analysis of different types of LLMs. Due to the large scale of our perturbation attacks and the high cost of prompting GPT-4, we sampled one-fifth of the data for the test for GPT-4. All of the results are shown in Appendix D and corresponding figures. We find that both the more powerful GPT-4 and the specially fine-tuned Prometheus also have the issues present in GPT3.5 described before. GPT-4 performs better in the directional expectation test compared to GPT-3.5, but surprisingly, it exhibits worse performance in the invariance test, especially showing severe sensitivity to grammatical and conflict-related perturbations about Grammaticality and Non-contradiction. On the other hand, Prometheus performs the worst in both tests, basically failing to differentiate between various aspects, which may be due to its small model size and the training data constructed by GPT-4.

6 Discussions

To investigate the failures of LLMs in our attack tests, we conduct some extended experiments with detailed aspect definitions. We retain only the definition or term, or even use the empty criterion, with the results presented in Figure 30 and 31 in the appendix, which still exhibit convergence. It indicates that the improper sensitivity of LLMs to grammaticality and contradiction is likely derived from the default evaluation behaviors inherent in LLMs. They will be cumulative, regardless of whether the current criteria are unclear or unrelated to those two aspects. Moreover, the detailed aspect definitions indeed have effects for aspects whose terms are not commonly-used in NLG evaluation, like non-hallucination.

Furthermore, we have attempted different empirical methods to intervene in LLM-based evaluation to mitigate these issues. Specifically, the most intuitive solution is using explicit instructions to require LLMs to ensure the evaluation relies solely on the given criteria. However, there are only slight and unstable improvements. So we further considered the ideas of Chain of Thought (CoT) to decompose the evaluation process. One method was inspired by Multidimensional Quality Metrics (MQM), which first identified relevant issues based on the criteria, then conducted the evaluation based on the identified issues and their severity. Another related method required LLMs to provide the preliminary evaluation, then utilized LLMs themselves to check if the evaluation strictly adhered to the given criteria, and finally offered an improved version. Moreover, we also attempted to provide more comprehensive background knowledge for LLMs, such as by including other aspect definitions in the prompt. Although these methods show varying degrees of improvement, none of them has a generally significant effect, and we believe that they do not fundamentally solve the reliability issues in LLM-based evaluation, which are quite stubborn and challenging and necessitate more systematic research.

To cover a more diverse range of descriptions of quality criteria, we also conduct human evaluation and LLM-based evaluation with descriptions selected from existing papers. We find that ambiguous expressions play a similar role in both human

evaluations and LLM-based evaluation as less informative descriptions designed by us. Inconsistent conceptualizations (e.g. a description mixing Fluency and Grammaticality) can alter human judgments on related aspect-targeted perturbations, and a similar but weaker effect exists in LLM-based evaluation.

7 Related Works

Diagnostic Tests for NLG Evaluation Metrics. Recent studies have highlighted issues with NLG evaluation metrics through synthetic perturbations, showing their scores often diverge from human judgments (Sai et al., 2021) and some of their blind spots (He et al., 2023). Moreover, some studies focused on diagnostic tests for single tasks or specific aspects, such as translation (Karpinska et al., 2022), summarization (Ernst et al., 2023), story generation (Xie et al., 2023b), and factuality (Chen et al., 2021). Notably, Zhang et al. (2023) explored the robustness of LLM-based dialogue evaluators using perturbation strategies, while Liu et al. (2023d) highlighted their inability to judge closed-ended responses without references under adversarial conditions. Neither study addressed varying expressions or distinctions among evaluation aspects.

Analyzing the limitations of LLM-based evaluators. Wang et al. (2023b) pointed out the order of the two texts affects evaluation results when ChatGPT and GPT-4 are used as comparison-based evaluators. LLM-based evaluators also prefer longer responses (Zheng et al., 2023) and responses generated by themselves (Liu et al., 2023b). Wang et al. (2023b) discovered that the performance of ChatGPT on summarization evaluation varies on different systems and aspects. Hada et al. (2023) stated that LLM-based evaluators may have more biases in non-Latin languages. It is worth mentioning that Xu et al. (2023) use GPT-4 to identify multiple failure modes in the explanations generated by the trained evaluator, though the failure modes cannot be used directly for aspect-specific evaluation.

NLG Quality Criteria. In human evaluation, Belz et al. (2020) proposed a classification system based on the property of quality criteria to support comparability. Howcroft et al. (2020) demonstrate that different descriptions of quality criteria can be mapped to normalized criteria. In LLM-based NLG evaluation, researchers have attempted to automatically generate quality standards more suitable for LLMs. Liu et al. (2023e) let LLMs draft expressions of quality criteria based on examples with human ratings. Kim et al. (2023b) utilized LLMs to review user-defined criteria and offered suggestions for disambiguation, merging, and splitting. Furthermore, some studies aim to improve LLMs' ability to evaluate specific aspects through chain-of-thoughts (Gong and Mao, 2023) and instruction tuning (Liu et al., 2023a).

8 Conclusions

In this work, we conduct fine-grained pertubation attack tests guided by the classification system and human judgments on LLMs to reveal their actual performance in NLG evaluation. Our findings can be concluded as follows: **1)** The performance of LLMs in our perturbation tests deviates significantly from expectations, with both unawareness and oversensitivity in some aspects. **2)** The different levels of detail in criteria almost do not change the evaluation behaviors of LLMs, except for criteria with uncommon terms like non-hallucination. **3)** The oversensitivity may be inherent in LLMs and not caused by the problems within criterion descriptions, due to its still existing in evaluations with empty criteria. **4)** The confusion issues are so stubborn that even the explicit instructions to hint LLMs to consider or not consider the specific problems cannot have obvious effects. These results show that LLM-based evaluation is not that reliable across different evaluation aspects. Therefore, in-depth analysis of the aforementioned problems in LLMs and effective methods for improving the evaluation capabilities of LLMs are necessary and worth exploring in future research.

Limitations

Our summarized classification system and designed perturbation attacks are mainly applicable to the commonly used aspects in closed-end text generation tasks. So our work does not include aspects with strong subjectivity, such as interestingness, which can be further explored in future work.

Due to limited resources, the domains and task types covered in our experiments are limited. The lengths of source documents and texts to be evaluated are generally a few hundred words, and the data we use is in English. Therefore, we cannot guarantee the same conclusions for long texts, other languages, or data from special domains.

We make extensive use of APIs from GPT-3.5 and GPT-4 for constructing data and testing, which incurs significant costs. This may discourage others from replicating these experiments, but we have released all the resources and data to facilitate related research.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), Beijing Science and Technology Program (Z231100007423011), Ant Group Research Fund and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments, and everyone who has provided assistance in this work. Xiaojun Wan is the corresponding author.

References

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *INLG*, pages 183–194. Association for Computational Linguistics.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *EMNLP (Findings)*, pages 2082–2095. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *ACL (1)*, pages 15607–15631. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *EMNLP (Findings)*, pages 8928–8942. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-examining summarization evaluation across multiple quality criteria. In *EMNLP (Findings)*, pages 13829–13838. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *CoRR*, abs/2304.02554.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *ACL (1)*, pages 179–188. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Peiyuan Gong and Jiaxin Mao. 2023. Coascore: Chain-of-aspects prompting for NLG evaluation. *CoRR*, abs/2312.10355.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *CoRR*, abs/2309.07462.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James R. Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *ACL (1)*, pages 12067–12097. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *INLG*, pages 169–182. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: diagnosing evaluation metrics for translation. In *EMNLP*, pages 9540–9561. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023a. Prometheus: Inducing fine-grained evaluation capability in language models. *CoRR*, abs/2310.08491.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023b. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. *CoRR*, abs/2309.13633.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *EAMT*, pages 193–203. European Association for Machine Translation.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *EMNLP/IJCNLP (1)*, pages 540–551. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *EMNLP (1)*, pages 9332–9346. Association for Computational Linguistics.

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

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2023a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. *CoRR*, abs/2311.08788.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023d. Evaluate what you can't evaluate: Unassessable generated responses quality. *CoRR*, abs/2305.14658.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023e. Calibrating llm-based evaluator. *CoRR*, abs/2309.13308.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *CoRR*, abs/2303.15621.

Shikib Mehri and Maxine Eskénazi. 2020. USR: an unsupervised and reference free evaluation metric for dialog generation. In *ACL*, pages 681–707. Association for Computational Linguistics.

John Mendonça, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. *CoRR*, abs/2308.16797.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Maja Popovic. 2017. chrf++: words helping character n-grams. In *WMT*, pages 612–618. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP (1)*, pages 2685–2702. Association for Computational Linguistics.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. In *ACL*, pages 4902–4912. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation checklists for evaluating NLG evaluation metrics. In *EMNLP (1)*, pages 7219–7234. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL*, pages 7881–7892. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. In *EMNLP*, pages 8776–8788. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie

Zhou. 2023a. Is chatgpt a good NLG evaluator? A preliminary study. *CoRR*, abs/2303.04048.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023a. The next chapter: A study of large language models in storytelling. In *INLG*, pages 323–351. Association for Computational Linguistics.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Han Lau. 2023b. Deltascore: Fine-grained story evaluation with perturbations. In *EMNLP (Findings)*, pages 5317–5331. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *EMNLP*, pages 5967–5994. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2023. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *CoRR*, abs/2312.15407.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *NAACL-HLT*, pages 314–324. Association for Computational Linguistics.

A Details for Preliminary Study

We follow the evaluation forms proposed by Chiang and Lee (2023b), including scoring modes, temperatures, and sampling settings. For more information, please refer to their paper and repository (Liu et al., 2023c). As for the prompts and instructions used for evaluation, we employ those from Chiang and Lee (2023b) for GPT-3.5 and GPT-4, while those provided by Kim et al. (2023a) for Prometheus. The complete results are included in Table 5 with the default settings where the sampling number is 20, and the temperature is set to 1 with zero-shot evaluations. Multiple results are post-processed and averaged to be the final scores. During few-shot evaluations, the selected demonstrations possess human labels of a uniform distribution, and analyses are correspondingly generated using GPT-3.5 if required. Moreover, the correlation matrices for GPT-4 and Prometheus are shown in Figure 5 and Figure 6, respectively. Although the performance of GPT-4 is significantly better than that of GPT-3.5, its confusion issues seem to be more severe than GPT-3.5; meanwhile, Prometheus not only performs the worst, but its confusion is also quite serious.

Figure 5: Pearson correlation coefficients between scores generated by GPT-4 or human annotators on four criteria in SummEval.

B Details for Perturbation Attacks

We construct four relatively simple types of perturbations—sentence exchange, word exchange, spelling mistake, and sentence deletion—based on the corresponding rules, while the remaining 14 types are generated by GPT-4 in 10-shot settings. We manually write these 140 demonstrations and carefully check them to ensure they meet the requirements of the corresponding perturbation attacks. Then, we prompt GPT-4 with these demonstrations as well as the detailed instructions to enable GPT-4 to generate the desired perturbed texts as closely as possible. And the corresponding instructions for GPT-4 and codes for rule-based constructions can be found in our released resources. In addition, we show different criteria for each aspect as described in Section 3.3 in Table 6-16.

Evaluation Form	Fluency	Coherence	Relevance	Consistency	Average
Score only	0.362	0.437	0.450	0.352	0.400
Score only (T=0)	0.344	0.353	0.394	0.321	0.353
Score only (1-shot)	0.342	0.323	**0.502**	0.401	0.392
Score only (5-shot)	0.415	0.399	0.471	0.538	0.456
Rate-explain	0.371	0.532	0.475	0.439	0.454
Rate-explain (T=0)	0.343	0.479	0.438	0.415	0.428
Analyze-rate	0.406	**0.581**	0.501	**0.573**	**0.515**
Analyze-rate (T=0)	0.367	0.525	0.362	0.468	0.431
Analyze-rate (1-shot)	0.311	0.423	0.334	0.420	0.372
Analyze-rate (5-shot)	**0.474**	0.505	0.443	0.526	0.487
Analyze-rate (GPT-4)	0.617	0.572	0.588	0.752	0.632
Analyze-rate (Prometheus)	0.298	0.352	0.376	0.343	0.342

Table 5: Pearson correlation coefficients between scores generated by different LLMs with different settings and forms of evaluation and human judgments on SummEval.

Type	**Criterion (Term and Definition)**
Default	Overall quality: It measures not only the quality of the target text itself, including writing and logic, but also how well the target text matches the required information of the source content according to the corresponding task.
Simplified	Overall quality: It measures whether the target text is well-written and logical, and matches the required points of the source content.
Term	Overall quality: It measures the overall quality of the target text.
Detailed	Overall quality: It measures not only the quality of the target text itself, including writing and logic, but also how well the target text matches the required information of the source content according to the corresponding task. Consider whether the target text is grammatically correct and naturally written, with clear meanings and good context-relatedness. Also consider whether all the information in the target text is supported by the source content and covers all and only the contents needed of the source content.
List	Overall quality: It measures not only the quality of the target text itself, including writing and logic, but also how well the target text matches the required information of the source content according to the corresponding task. Score 5: Good overall quality, with no errors of grammar, expression, content alignment required, and so on. Score 4: Only with some minor writing and content problems. Score 3: There are some obvious errors that affect the meaning and understanding of the target text, like unclear expressions and illogical context-relatedness. Score 2: Containing many major writing and logical errors and unmatched contents, but some good segments. Score 1: Poor overall quality, full of fragments that contain unrelated or untrue information and cannot be understood.
Selection1	Overall (1-5) What is your overall impression of this target text? - A score of 1 (very bad). A completely invalid target text. It would be difficult to recover the source content from this. - A score of 2 (bad). Valid target text, but otherwise poor in quality. - A score of 3 (neutral) means this target text is neither good nor bad. This target text has no negative qualities, but no positive ones either. - A score of 4 (good) means this is a good target text, but falls short of being perfect because of a key flaw. - A score of 5 (very good) means this target text is good and does not have any strong flaws.
Selection2	Overall quality: How good is the target text overall at representing the source content? If it's hard to find ways to make the target text better, give the target text a high score. If there are lots of different ways the target text can be made better, give the text a low score.

Table 6: Examples of different criterion descriptions for overall quality.

Type	Criterion (Term and Definition)
Default	Readability: It measures the quality of both inter- and intra-sentences, are they grammatically correct and naturally written, with clear meanings and good context-relatedness and logic.
Simplified	Readability: It measures whether the target text is well-written, logical and clear.
Term	Readability: It measures whether the target text is readable.
Detailed	Readability: It measures the quality of both inter- and intra-sentences, are they grammatically correct and naturally written, with clear meanings and good context-relatedness and logic. Consider whether there are grammar errors, duplication, uncommon phrases and syntactic structures, unreasonable conjunctions, semantic inconsistency, and so on.
List	Readability: It measures the quality of both inter- and intra-sentences, are they grammatically correct and naturally written, with clear meanings and good context-relatedness and logic. Score 5: Entirely well readable, with no grammar errors, uncommon usages, or poor logic. Score 4: Only containing some minor writing or logical problems that basically do not affect reading. Score 3: Readable in general, with some obvious errors in grammar, collocations, or consistency. Score 2: There are major writing and logical problems, but some readable segments. Score 1: Not readable at all, full of fragments that cannot be understood.
Selection1	Readability takes into account word and grammatical error rate to evaluate how fluent the target text language is.
Selection2	Fluency: The target text sentences should be grammatically correct, easy to read and understand."

Table 7: Examples of different criterion descriptions for readability.

Type	Criterion (Term and Definition)
Default	Coherence: It measures the quality of all sentences collectively, do they make sense as a whole, with the context organized and connected logically.
Simplified	Coherence: It measures whether all the sentences are organized and connected logically.
Term	Coherence: It measures whether the target text is coherent.
Detailed	Coherence: It measures the quality of all sentences collectively, do they make sense as a whole, with the context organized and connected logically. Consider whether they have good context-relatedness with reasonable conjunctions, semantic consistency, and inter-sentence causal and temporal dependencies.
List	Coherence: It measures the quality of all sentences collectively, do they make sense as a whole, with the context organized and connected logically. Score 5: Entirely coherent, with good context-relatedness among all the sentences. Score 4: Only containing some minor illogical parts that basically do not affect overall coherency. Score 3: Coherent in general, with some obvious conflicting logical or inconsistent problems. Score 2: There are major unreasonable logic and semantic inconsistencies, but at least the related topic. Score 1: Not coherent at all, full of self-contradictory or unrelated content.
Selection1	Discourse Coherence: Whether the target text is well organized, with the sentences smoothly connected and flow together logically and aesthetically?
Selection2	Coherence: Description: Collective quality of all sentences.
Selection3	Coherence: The rating measures the quality of all sentences collectively, to fit together and sound natural. Consider the quality of the target text as a whole.

Table 8: Examples of different criterion descriptions for coherence.

Type	Criterion (Term and Definition)
Default	Fluency: It measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings.
Simplified	Fluency: It measures whether individual sentences are grammatically correct and well-written.
Term	Fluency: It measures whether the target text is fluent.
Detailed	Fluency: It measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings. Consider whether there are misspellings, tense errors, missing determiners, or more severe problems, such as duplication, unfamiliar phrases, complex syntactic structures, and missing components.
List	Fluency: It measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings. Score 5: Entirely fluent, grammatically correct, and well-written. Score 4: Only containing some minor non-fluent parts or grammatical errors that basically have no effect on fluency. Score 3: Fluent in general, with some obvious grammatical errors and unfamiliar phrases. Score 2: There are major grammatical errors, duplication, unfamiliar phrases and syntactic structures, and missing components, but some fluent segments. Score 1: Not fluent at all, full of meaningless fragments and unclear contents.
Selection1	Fluency: Description: Quality of individual sentences.
Selection2	Fluency: Whether the generated target text is grammatically correct.
Selection3	Fluency: The rating measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

Table 9: Examples of different criterion descriptions for fluency.

Type	Criterion (Term and Definition)
Default	Grammaticality: It measures whether the target text is grammatically correct without any lexical or syntax errors, regardless of its content and meaning.
Simplified	Grammaticality: It measures whether the target text has no grammatical errors.
Term	Grammaticality: It measures whether the target text is grammatical.
Detailed	Grammaticality: It measures whether the target text is grammatically correct without any lexical or syntax errors, regardless of its content and meaning. Consider whether the target text itself complies with the English standard usage and rules of grammar, such as tense errors, misspellings, incorrect prepositions, collocation misusages, and so on.
List	Grammaticality: It measures whether the target text is grammatically correct without any lexical or syntax errors, regardless of its content and meaning. Score 5: Entirely grammatically correct, following the rules of English grammar. Score 4: Basically grammatical, with a few minor grammar errors. Score 3: There are some obvious grammatical errors that affect the sentence's expression. Score 2: Containing many severe grammatical errors whose originally intended usages even cannot be judged. Score 1: Not grammatical at all, full of grammar errors.
Selection1	Grammar – ability to generate grammatically correct and fluent target texts.
Selection2	Grammaticality measures whether the target text contains syntax errors. It refers to the conformity of the target text to the rules defined by the specific grammar of a language.
Selection3	Correctness: Whether there are grammatical errors in the target text.

Table 10: Examples of different criterion descriptions for grammaticality.

Type	Criterion (Term and Definition)
Default	Simplicity: It measures whether the target text is sufficiently simple and easy for people who aren't very good at English to get the correct meaning.
Simplified	Simplicity: It measures whether the target text is simple and easy to get the meaning.
Term	Simplicity: It measures whether the target text is simple.
Detailed	Simplicity: It measures whether the target text is sufficiently simple and easy for people who aren't very good at English to get the correct meaning. Consider whether the target text adopts simplified and common usage of phrases and sentences, and avoid any unfamiliar words or complicated syntactic structures.
List	Simplicity: It measures whether the target text is sufficiently simple and easy for people who aren't very good at English to get the correct meaning. Score 5: Entirely simple, without any complicated words or syntactic structures. Score 4: Only containing a small number of unfamiliar words. Score 3: There are some uncommon phrases and complicated structures, which makes it a little hard to get the target text's meaning. Score 2: Despite some simple words, the target text has many unfamiliar phrases and complicated sentences. Score 1: Not simple at all, full of complex expressions that are quite difficult to read.
Selection1	The goal is to judge whether the target text is simpler than the source content.

Table 11: Examples of different criterion descriptions for simplicity.

Type	Criterion (Term and Definition)
Default	Adequacy: It measures whether the entire contents of the target text exactly match the required information of the source content without unnecessary points, according to the corresponding task.
Simplified	Adequacy: It measures how well the target text matches the required information of the source content.
Term	Adequacy: It measures whether the target text is adequate.
Detailed	Adequacy: It measures whether the entire contents of the target text exactly match the required information of the source content without unnecessary points, according to the corresponding task. Consider whether all the information contained in the target text is factually supported by the source content and covers all and only the contents that the task needs in the source content.
List	Adequacy: It measures whether the entire contents of the target text exactly match the required information of the source content without unnecessary points, according to the corresponding task. Score 5: Entirely adequate, the whole target text matches the required information of the source content. Score 4: Just containing some minor unmatched or unnecessary information. Score 3: There is some key information that cannot be supported by the source content or is not needed by the task. Score 2: Only a small part of the information in the target text matches the required information of the source content, with many incorrect points. Score 1: Not adequate at all, the target text is irrelevant and does not cover any content in the source content.
Selection1	Adequacy is defined as how much information is preserved in the target text. A score of 1 would mean that the target text is meaningless and has no correlation with the source content. A score of 5 would mean the target text retains all of the information.
Selection2	Adequacy: Description: How correct is the target text from the given source content.

Table 12: Examples of different criterion descriptions for adequacy.

Type	Criterion (Term and Definition)
Default	Faithfulness: It measures whether all the information contained in the target text is consistent with and factually supported by the source content.
Simplified	Faithfulness: It measures whether the target text can be supported by the source content.
Term	Faithfulness: It measures whether the target text is faithful.
Detailed	Faithfulness: It measures whether all the information contained in the target text is consistent with and factually supported by the source content. Consider whether there are fabricated contents that cannot be inferred from the source content, including those contradicting the facts in the source content, and additional information that is not mentioned and cannot be verified by the source content.
List	Faithfulness: It measures whether all the information contained in the target text is consistent with and factually supported by the source content. Score 5: Entirely faithful, all the facts in the target text can be inferred from the source content. Score 4: The target text is almost factually aligned with the source content but contains unsupported minor information. Score 3: There are some main but unverifiable or contradictory contents in the target text, according to the source content. Score 2: Only a small part of the information in the target text can be inferred from the source content. Score 1: Not faithful at all, the target text has nothing to do with the source content.
Selection1	Faithfulness: Whether the target text accords with the facts expressed in the source content.
Selection2	Consistency: The rating measures whether the facts in the target text are consistent with the facts in the source content. Consider whether the target text does reproduce all facts accurately and does not make up untrue information.
Selection3	Faithful or factually consistent: A target text is factually consistent to the source content if all the information in the target text can be supported by the source content. Common errors in model-generated target texts include information that is not mentioned or incorrect according to the input source content. Sometimes, the target text can be misleading because a crucial piece of information is absent.

Table 13: Examples of different criterion descriptions for faithfulness.

Type	Criterion (Term and Definition)
Default	Non-hallucination: It measures whether the target text contains no additional information that is not exactly mentioned and cannot be verified by the source content.
Simplified	Non-hallucination: It measures whether the target text is verifiable according to the source content.
Term	Non-hallucination: It measures whether the target text has no hallucinations.
Detailed	Non-hallucination: It measures whether the target text contains no additional information that is not exactly mentioned and cannot be verified by the source content. Consider whether there are contents other than the source content that cannot be proven correct or incorrect based on the source content, or even are unrelated to the source content.
List	Non-hallucination: It measures whether the target text contains no additional information that is not exactly mentioned and cannot be verified by the source content. Score 5: No hallucinations, all the facts in the target text can be proven correct or incorrect based on the source content. Score 4: Just containing a few unverifiable but unimportant contents. Score 3: There are some non-negligible contents other than the source content, leading to distorted meanings. Score 2: Almost all the contents of the target text are unverifiable based on the source content, except for several facts. Score 1: Full of hallucinations, the target text cannot be verified by the source content at all.
Selection1	Hallucination error: Fabricated content that does not occur in the source content.
Selection2	Not enough info: The target text information is not relevant or not sufficient to support/refute the source content.

Table 14: Examples of different criterion descriptions for non-hallucination.

Type	Criterion (Term and Definition)
Default	Non-contradiction: It measures whether the target text contains no information that definitely contradicts certain contents of the source content.
Simplified	Non-contradiction: It measures whether the target text does not contradict the source content.
Term	Non-contradiction: It measures whether the target text has no contradictions.
Detailed	Non-contradiction: It measures whether the target text contains no information that definitely contradicts certain contents of the source content. Consider whether there are contradictory contents such as incorrect entities, different expressions that distort the original meaning, false concatenation of crucial information from different places of the source content, and so on.
List	Non-contradiction: It measures whether the target text contains no information that definitely contradicts certain contents of the source content. Score 5: No contradictions, all the facts in the target text do not conflict with the source content. Score 4: Just containing a few contradictory but unimportant contents. Score 3: There is some main information, like key entities, contradicting the source content, leading to distorted meanings. Score 2: Almost all the contents of the target text conflict with the source content, except for several facts. Score 1: Entirely contradictory, all the facts in the target text do contradict the source content.
Selection1	Contradiction, whether the target text contains any pieces of information that are contradicting the given source content or not.

Table 15: Examples of different criterion descriptions for non-contradiction.

Type	Criterion (Term and Definition)
Default	Informativeness: It measures how much required information of the source content is contained in the target text, according to the corresponding task.
Simplified	Informativeness: It measures how well the target text covers required contents of the source content.
Term	Informativeness: It measures whether the target text is informative.
Detailed	Informativeness: It measures how much required information of the source content is contained in the target text, according to the corresponding task. Consider how well the target text correctly covers the contents that the task needs in the source content, which may be necessary information and key points or the entire content.
List	Informativeness: It measures how much required information of the source content is contained in the target text, according to the corresponding task. Score 5: Entirely informative, the target text covers all the required information of the source content. Score 4: The target text captures the main points and only misses minor required information of the source content. Score 3: There is some important information needed but not contained in the target text, which disturbs the source content's meaning. Score 2: Only a few contents that the task needs in the source content can be found in the target text. Score 1: Not informative at all, the target text does not involve any contents of the source content.
Selection1	Informativeness: Is important information captured?
Selection2	Informativeness: Whether the target text provides enough and necessary content coverage from the input source content.
Selection3	Coverage, i.e., whether the target text covers the whole source content or only part of the source content.

Table 16: Examples of different criterion descriptions for informativeness.

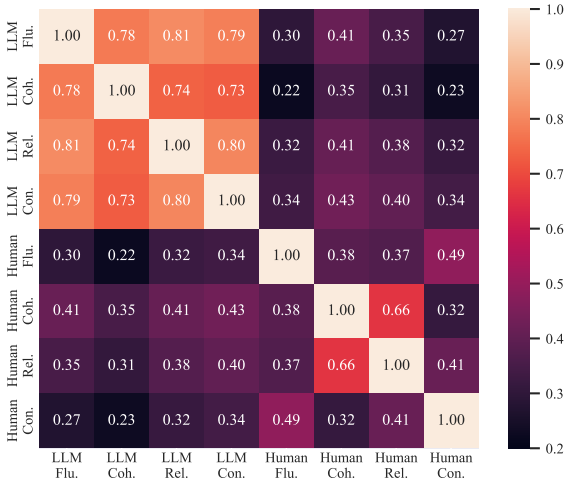


Figure 6: Pearson correlation coefficients between scores generated by Prometheus or human annotators on four criteria in SummEval.

C Details for Data and Test Settings

C.1 Datasets

We select 200, 200, 300, and 300 pieces of data from CNN/Dailymail (Hermann et al., 2015), SAMSum (Gliwa et al., 2019), News Commentary [3], and WebNLG (Gardent et al., 2017) respectively for tasks of news summarization, dialogue summarization, paraphrase generation, and table-to-text generation. However, many times the original references in common datasets for these tasks are not written by humans or are even missing. For instance, references for news summarization often employ the assemblage of highlights to build large-scale datasets but tend to be incoherent or contain information not present in the source news.

To ensure the quality of references to better serve as the original texts in perturbation attack tests, we take advantage of the powerful GPT-4 to improve them, avoiding expert annotations that are hard to obtain. Specifically, depending on the condition of the original references in different tasks, we prompt GPT-4 to generate new references for news summarization and paraphrase generation, while the original references are modified and improved by GPT-4 in table-to-text generation. And we directly use the original references from SAMSum in dialogue summarization since they are human-written.

As shown in the evaluation results of GPT-3.5, the original texts for perturbations, namely the references, are generally scored around 5 in all aspects, showing their high qualities. The prompts we use here are shown in Table 17. For each reference, we construct 18 different perturbed texts in various directions, leading to 19000 samples to be evaluated. Moreover, taking into account eleven different aspects and the different types of definitions involved with each, there are a total of 80 distinct evaluation criteria. Combined together, they constitute our data for experiments with the scale of 80*19000 = 1.52M.

Moreover, unlike traditional task-oriented NLG evaluation research, we focus on the general reliability of LLMs in NLG evaluation, so our considered perturbations and aspects are task-agnostic (not task-specific), which can be applied for many NLG tasks. We select four common NLG tasks in our experiments and find that LLMs show consistent evaluation issues across all tasks. Therefore, for the sake of clarity in our paper, we merged the data from different tasks to present our experiments and discussions. More details and results of specific tasks can be found in our released resources.

C.2 LLMs

We test GPT-3.5 and GPT-4 with the API provided by OpenAI, whose versions are GPT-3.5 Turbo (1106) and GPT-4 Turbo (1106), respectively. On the other hand, Prometheus (Kim et al., 2023a) has been proposed aiming to achieve performance close to that of proprietary LLMs like GPT-4 in NLG evaluation. They elaborately constructed 100K evaluations and feedbacks through GPT-4 and fine-tuned Llama-2-Chat-13B (Touvron et al., 2023) on them, endowing the model with the capacity of evaluation across diverse and customized criteria. And we directly use the prompts provided by themselves (Kim et al., 2023a) for the evaluation of Prometheus. For all three LLMs in our test, we follow the setting of Chiang and Lee (2023b) to conduct analysis before rating scores of 1–5 and set temperature and sampling number to 1.0 and 10, respectively in zeroshot, with prompts shown in Table 18.

C.3 Human annotation

Settings. To facilitate human annotators in comparing texts before and after perturbations, we use a comparative form in human evaluations. This involves displaying two texts simultaneously on the annotation interface, allowing them to judge their quality relationship based on the given description of the quality criterion, as shown in Figure 7. Specifically, considering we design different

[3] http://data.statmt.org/news-commentary/v18.1

Prompts and Instructions
News Summarization
Please summarize the following news article in three to four sentences. Note that you should use simple and short sentences, avoiding uncommon words and complex sentences. News Article: {article} Summary:
Paraphrase Generation
Please rephrase the following original text, maintaining exactly the same meanings. Note that you should use simple and short sentences, avoiding uncommon words and complex sentences. Note that you must not add any additional information and not delete or lose any information of the original text. Original Text: {source} Rephrasing:
Table-to-text Generation
Please modify the original description to contain exactly the same meanings as the table, and make the new description fluent and coherent. Note that you should use simple and short sentences, avoiding unnatural passive voices or intransitive verbs, uncommon words, and complex sentences. Note that you must not add any additional information and not delete or lose any information of the table. Table: {table} Original Description: {ref} New Description:

Table 17: Prompts and instructions used for improving references by GPT-4.

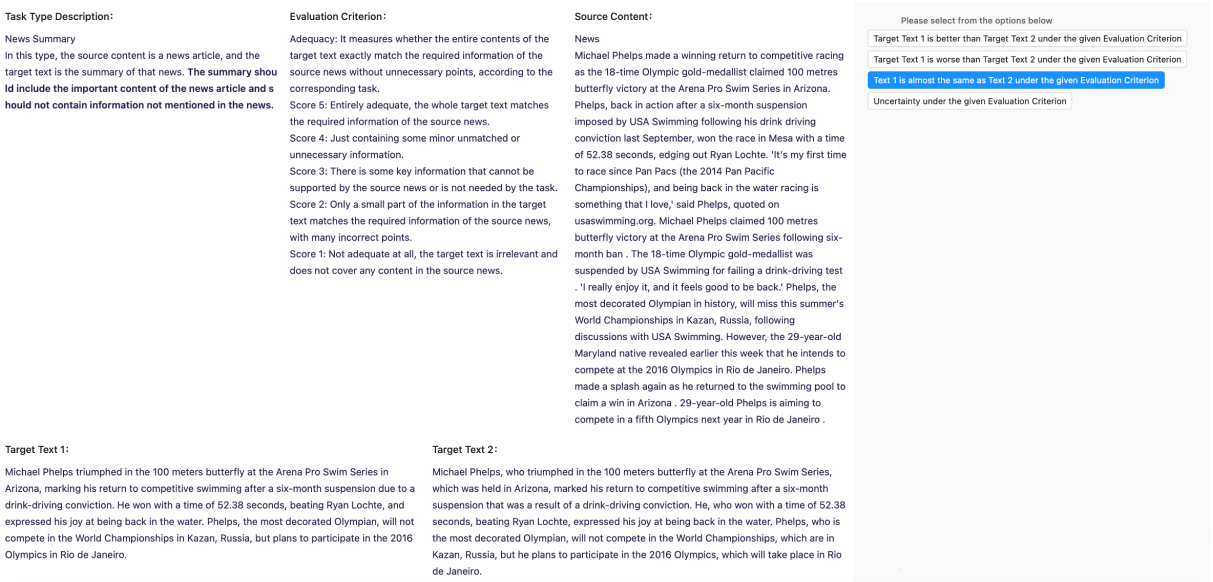


Figure 7: A screenshot of the annotation interface used in human evaluation.

Prompts and Instructions	
GPT-3.5 and GPT-4	Prometheus
You will be given an example of the source content and target text. The target text is generated from the source content according to the corresponding task type. Your task is to rate the target text according to the evaluation criterion on a Likert scale from 1 to 5. Please make sure you read and understand these instructions carefully. Task Type Description: {task description} Evaluation Criterion: {aspect description} Example: Source Content: {source} Target Text: {target} Evaluation Form: Answer by starting with "Analysis:" to analyze the given example regarding the evaluation criterion as concisely as possible, and then give the numeric rating on the next line by "Rating:". Your Answer:	###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. ###The instruction to evaluate: {task description} {source} ###Response to evaluate: {target} ###Reference Answer (Score 5): {reference} ###Score Rubrics: {[aspect description]} ###Feedback:

Table 18: Prompts and instructions used for evaluation of GPT-3.5, GPT-4 and Prometheus.

Perturbation Attack		Flu.	Coh.	Gram.	Sim.	Read.	Fai.	Cont.	Hal.	Inf.	Ade.	All.
Flu.	Repetition	✗✓	✗✗	✗✗	✗✗	✗✓	✗✗	✗✗	✗✗	✗✗	✗✗	✗✓
	Passive Voice	✓✓	✗✗	✗✗	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✗✓
	Inversion	✓✓	✗✗	✗✗	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
Coh.	Improper Connective	✗✗	✓✓	✗✗	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
	Sentence Exchange	✗✗	✓✓	✗✗	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
Gram.	Incorrect Verb Form	✓✓	✗✗	✓✓	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
	Word Exchange	✓✓	✗✗	✓✓	✗✗	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
	Spelling Mistake	✓✓	✗✗	✓✓	✗✗	✗✓	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓
Sim.	Uncommon Phrase	✗✗	✗✗	✗✗	✓✓	✓✓	✗✗	✗✗	✗✗	✗✗	✗✗	✗✓
	Complex Sentence	✗✗	✗✗	✗✗	✓✓	✗✓	✗✗	✗✗	✗✗	✗✗	✗✗	✗✓
Inf.	Abbreviation	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✓
	Hypernym	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✓
	Sentence Deletion	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✓
Hal.	Complement	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✗✗	✓✓	✗✗	✓✓	✓✓
	Continuation	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✗✗	✓✓	✗✗	✓✓	✓✓
Cont.	Different Entity	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✗	✓✓	✓✓	✓✓
	Conflicting Fact	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✗	✓✓	✓✓	✓✓
	Negation	✗✗	✗✗	✗✗	✗✗	✗✗	✓✓	✓✓	✓✗	✓✓	✓✓	✓✓

Table 19: Human judgments (left) and our expectations based on the classification system (right) for each pair of perturbation attacks and aspects with detailed definitions. ✓ presents that the item should be affected, while ✗ presents that the item should not be affected. And we identify two sets of S_T and S_F on those items that have the consistent results.

types of descriptions and select some quality criteria with ambiguous expressions from existing papers, to better record the uncertainty of human annotators facing quality criteria of varying detail, the available quality relationships they can choose include "better than" (A), "worse than" (B), "as well as" (C), and "uncertain" (D). All 40 annotators come from the company's professional data annotation department, have certificates of English proficiency, and are paid more than the local minimum wage. Due to limited resources, we sample one example from each of the four datasets (CNN/Dailymail, SAMSum, News Commentary, and WebNLG) for human annotation. Each sample was subjected to 18 types of perturbation attacks, resulting in 18 pairs of test samples with and without perturbations. We had 11 quality criteria in total, and for each criterion, besides 5 descriptions of varying detail we design, we also select 1-3 descriptions from existing papers, making up a total of 80 descriptions. To prevent interference from other descriptions, for a quality criterion, an annotator is exposed to at most one description. Specifically, we divide the 40 annotators into four groups of ten, with each group annotating all the data once, meaning each test sample is annotated by four annotators. For each group of annotators, an annotator needed to annotate all test samples under the 8 descriptions of different quality criteria, with the types of descriptions distributed as evenly as possible (e.g. an annotator would not annotate all descriptions of the "Term" type). The total volume of annotations was $4 \times 18 \times 80 \times 4 = 23040$. The entire annotation process takes about 20 days.

Results. We define the annotation consistency per sample as the proportion of options with the most annotations except for the "uncertain" (D) option. For example, if the options given by four annotators on a sample are $\{A, A, C, D\}$, and the annotation consistency is 0.5. The final annotation consistency is the average across all samples. We calculate the match rate (i.e. the proportion of human judgments about perturbations that match our expectations) in two ways. The result is shown in Table 20.

D Details for Experiments

D.1 Comparison with non-LLM Evaluation Metrics

It may also be interesting to bring back non-LLM automatic evaluation metrics to compare the performance and reliability across perturbation schemes like Sai et al. (2021). However, we need to point out that our work focuses on the understanding and execution capabilities of LLM-based evaluators regarding different aspects and criteria. Only when an evaluation metric can support evaluating with customized aspect definitions is it possible to study whether it confuses different aspects. However, most non-LLM evaluation metrics are designed for overall evaluation, such as ROUGE and BERTScore; while a few are designed for specific aspects and cannot accept criteria or other aspects as inputs, like FactCC, which targets summary faithfulness.

Table 21 shows the performance of six commonly used non-LLM evaluation metrics as baselines, including BLEU (Papineni et al., 2002), CHRF++ (Popovic, 2017), ROUGE (Lin, 2004), COMET-QE (Rei et al., 2020), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020). Due to the factors mentioned above, we can only conduct the directional expectation test (i.e., evaluation scores should decline) and compare with GPT-4 on the evaluation aspect of overall quality. The results show that rule-based metrics, especially BLEU, are generally more sensitive to these perturbations compared with model-based metrics.

D.2 Other Results

Correlation matrices. Figure 8 shows the correlations between evaluation scores from GPT-3.5 across five types of aspect definitions. Figure 9-13 show the correlations between evaluation scores from GPT-3.5 across different aspects given a description type.

Complete perturbation results. We display all the results of perturbation attacks on GPT-3.5, GPT-4, and Prometheus here. The results are visualized for fixed evaluation aspects or description types separately. Figure 14-18 show the perturbation results of the three LLMs given a description type, which makes it convenient to compare different evaluation aspects. On the other hand, Figure 19-29 show the perturbation results of the three LLMs given an evaluation aspect, which allows for easier comparison of different description types.

Deleting terms or descriptions. Figure 30 and 31 show the perturbation results for the criteria that only retain descriptions, only retain terms, only contain a single word of "Aspect", and are empty.

Description Type	Annotation Consistency	Vote_vote	Vote_all
Default	0.8791	0.8586	0.8990
Simplified	0.8472	0.8081	0.8586
Detailed	**0.9012**	**0.8788**	**0.9444**
Term	0.7181	0.7172	0.7677
List	0.8602	0.8333	0.9091

Table 20: Annotation consistency and match rate of human annotations on five types of descriptions designed by us. *vote_vote* means that the results of 4 annotators are first taken as plurality on a single sample, and then 4 samples are taken as plurality as the final result of human annotation on a perturbation. *vote_all* means that 16 results are directly taken as plurality together as the final result.

Perturbation	GPT-4 (Overall)	BLEU	chrF++	ROUGE	COMET-QE	BERTScore	BLEURT
Repetition	0.37	1.05	0.20	0.39	0.00	0.08	0.60
Passive Voice	0.59	1.68	0.25	1.02	0.13	0.13	0.73
Inversion	1.22	1.51	0.17	0.98	0.33	0.16	1.06
Improper Connective	0.38	0.71	0.13	0.13	0.05	0.06	0.44
Sentence Exchange	0.21	0.16	0.01	1.46	0.08	0.11	0.24
Incorrect Verb Form	2.20	1.11	0.14	0.49	0.18	0.10	0.58
Word Exchange	2.35	1.35	0.12	0.46	0.43	0.16	1.31
Spelling Mistake	2.64	1.53	0.13	0.73	0.50	0.26	1.18
Uncommon Phrase	0.40	1.52	0.43	0.73	0.10	0.12	0.81
Complex Sentence	0.62	2.03	0.30	0.85	0.15	0.16	0.80
Abbreviation	0.56	2.64	1.21	1.43	0.20	0.20	1.05
Hypernym	0.88	1.44	0.59	0.78	0.11	0.15	0.99
Sentence Deletion	0.77	1.28	0.94	0.66	0.12	0.11	0.74
Complement	1.22	2.03	0.59	1.06	-0.28	0.20	0.90
Continuation	0.82	1.27	0.40	0.77	-0.19	0.12	0.54
Different Entity	2.94	1.11	0.42	0.54	0.09	0.09	0.93
Conflicting Fact	3.54	1.83	0.50	1.04	0.20	0.19	1.21
Negation	1.93	0.37	0.07	0.14	0.08	0.04	0.52

Table 21: The variances of scores of non-LLM evaluation metrics bewteen original texts and different perturbed texts. Scores are scaled to 1-5, and the higher the better.

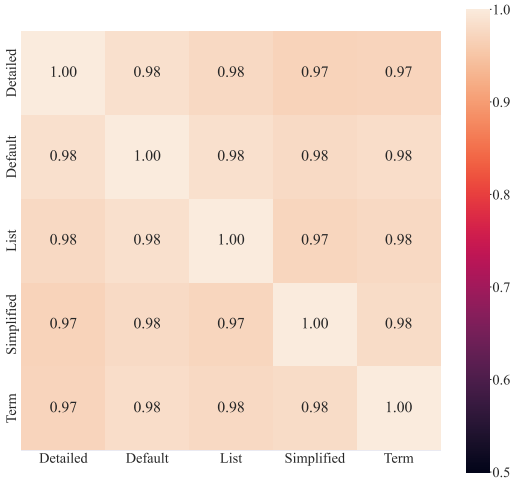


Figure 8: Correlations between evaluation scores from GPT-3.5 across different levels of detail in aspect definitions.

Figure 9: Correlations between evaluation scores from GPT-3.5 across different aspects with the description of detailed type.

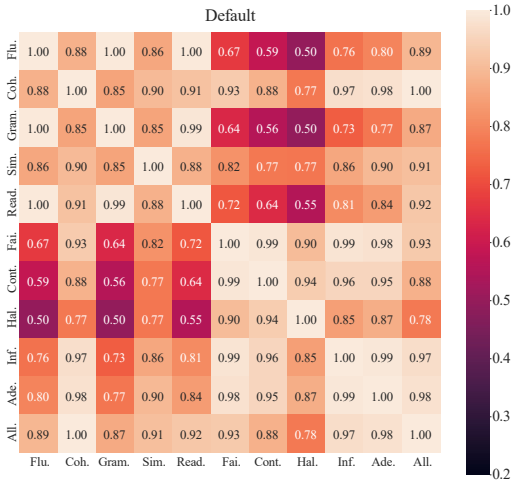


Figure 10: Correlations between evaluation scores from GPT-3.5 across different aspects with the description of default type.

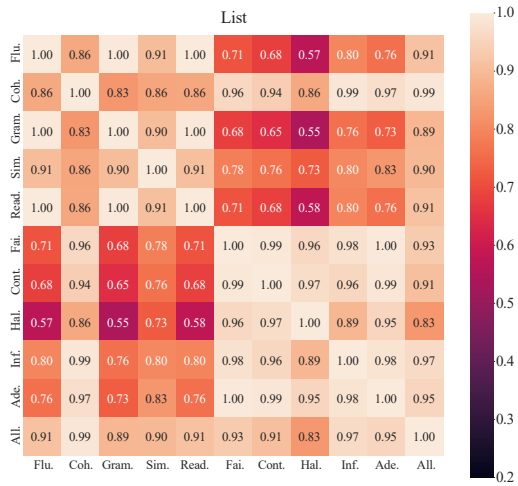


Figure 11: Correlations between evaluation scores from GPT-3.5 across different aspects with the description of list type.

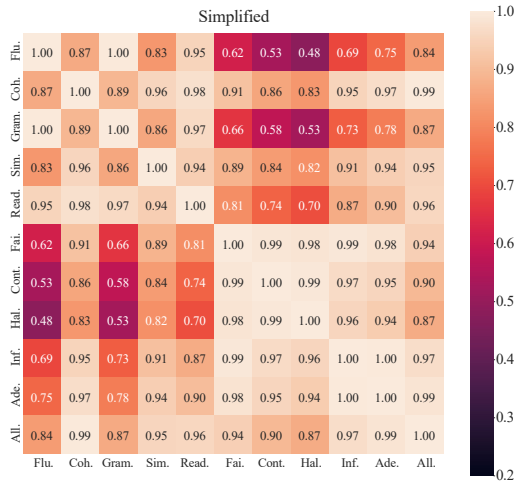


Figure 12: Correlations between evaluation scores from GPT-3.5 across different aspects with the description of simplified type.

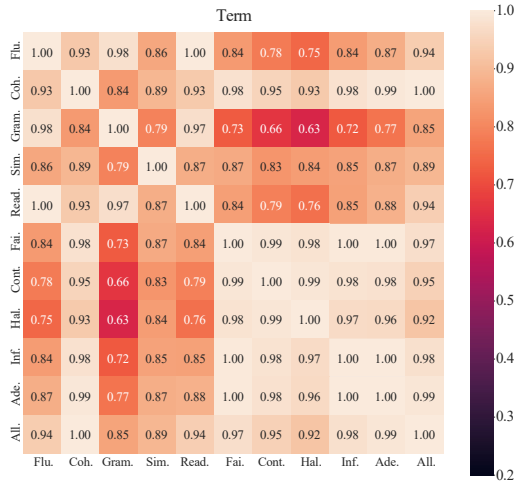


Figure 13: Correlations between evaluation scores from GPT-3.5 across different aspects with the description of term type.

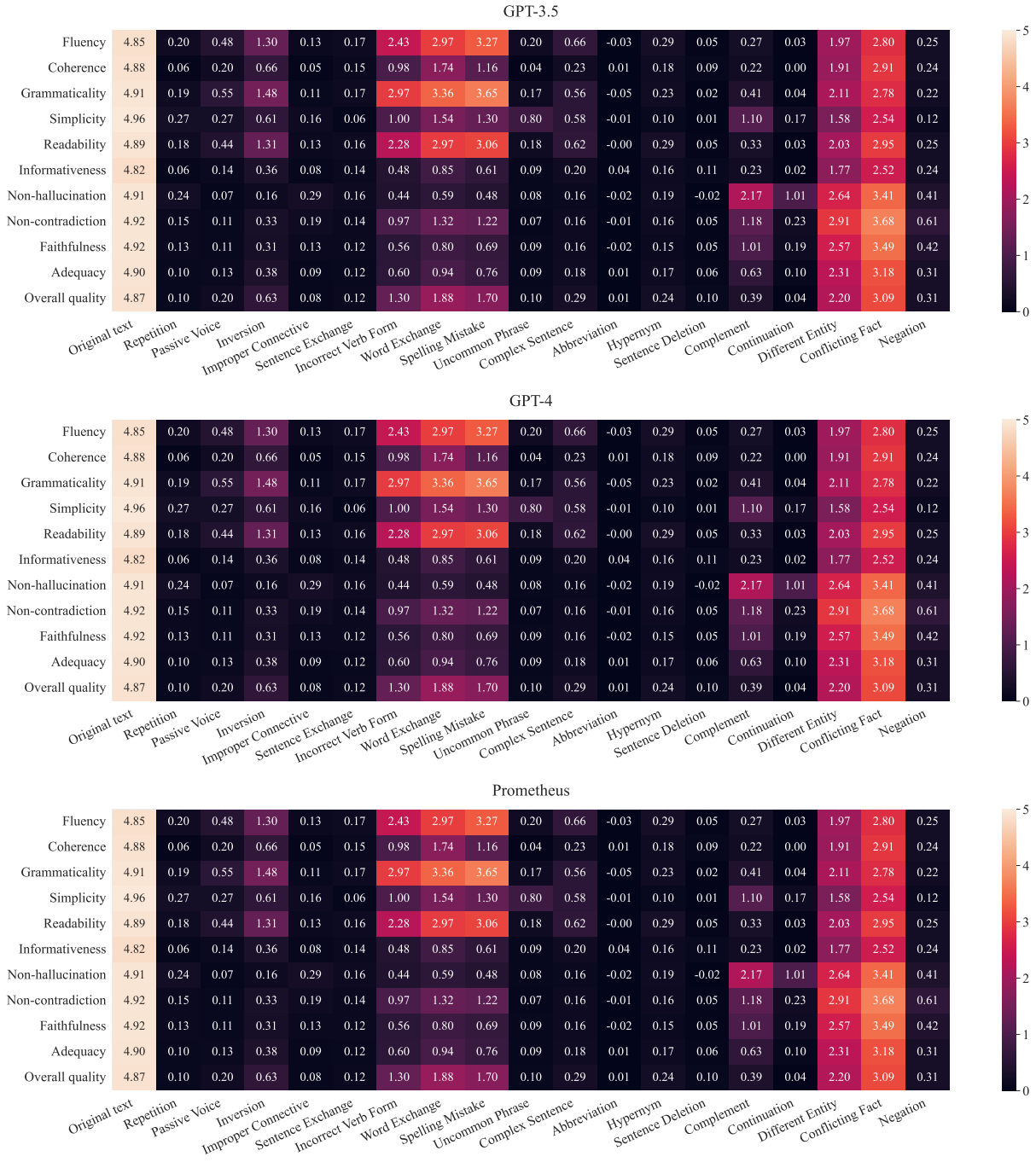


Figure 14: The variances of evaluation scores from three LLMs between original texts and different perturbed texts with the description of detailed type.

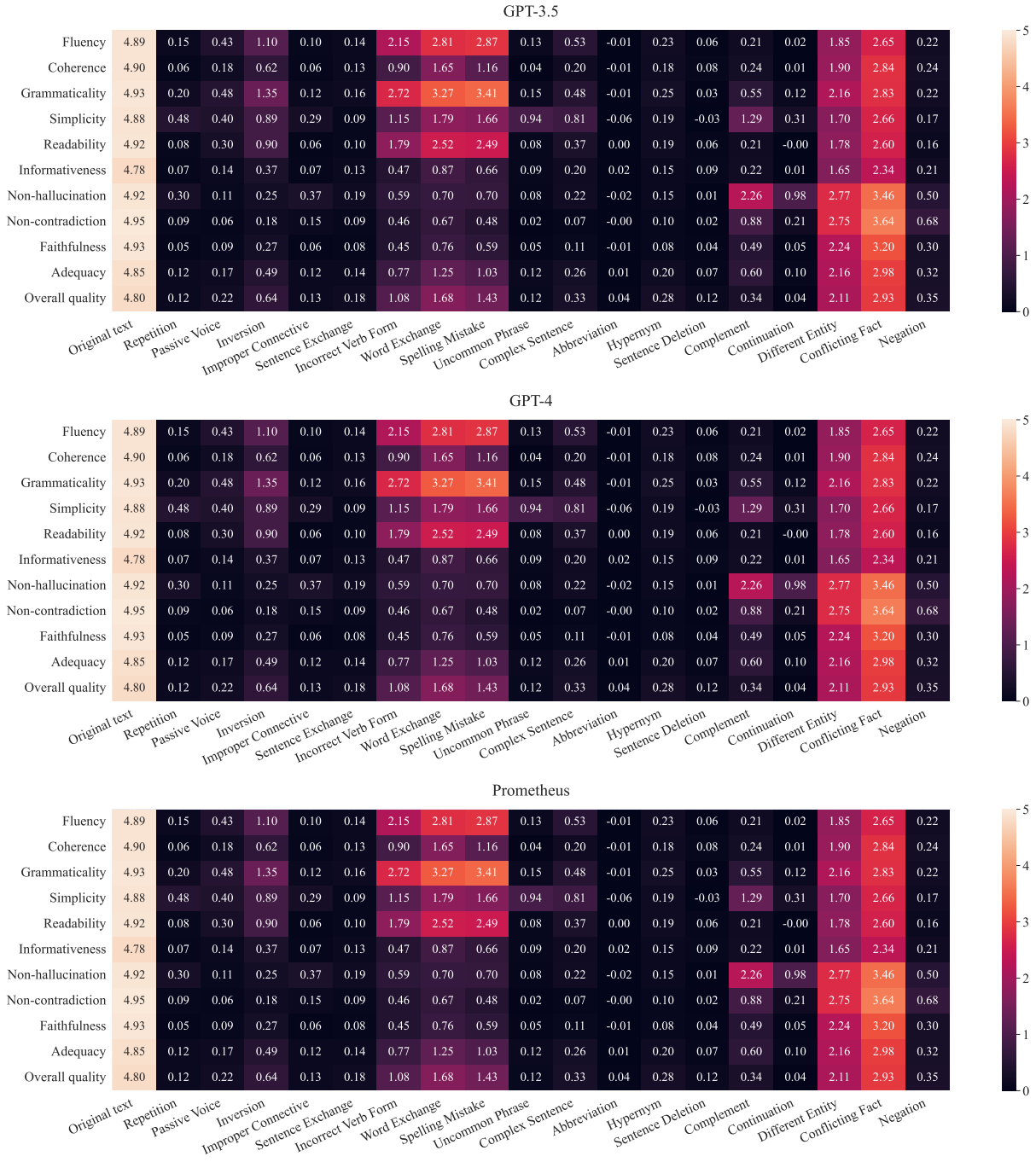


Figure 15: The variances of evaluation scores from three LLMs between original texts and different perturbed texts with the description of default type.

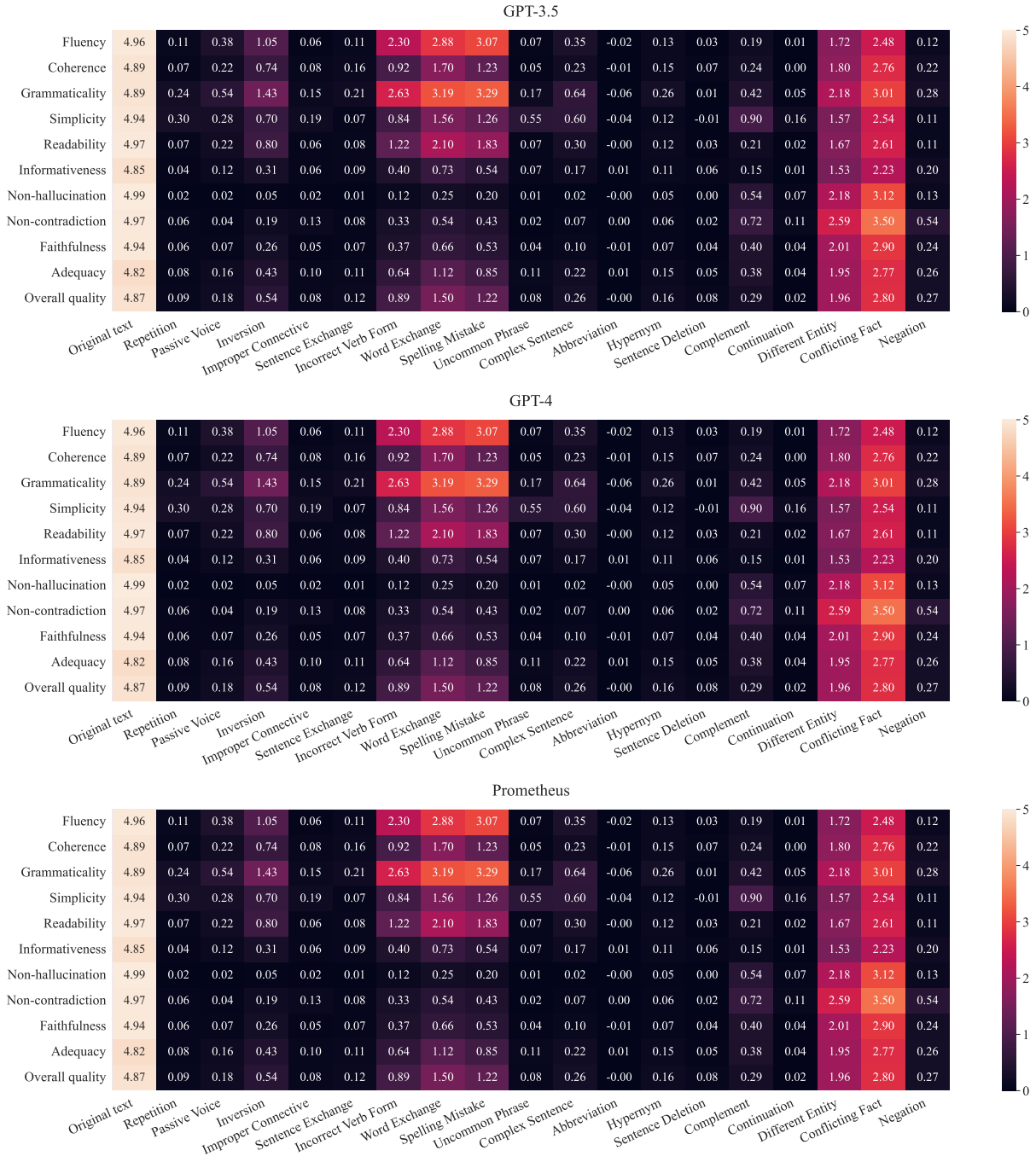


Figure 16: The variances of evaluation scores from three LLMs between original texts and different perturbed texts with the description of simplified type.

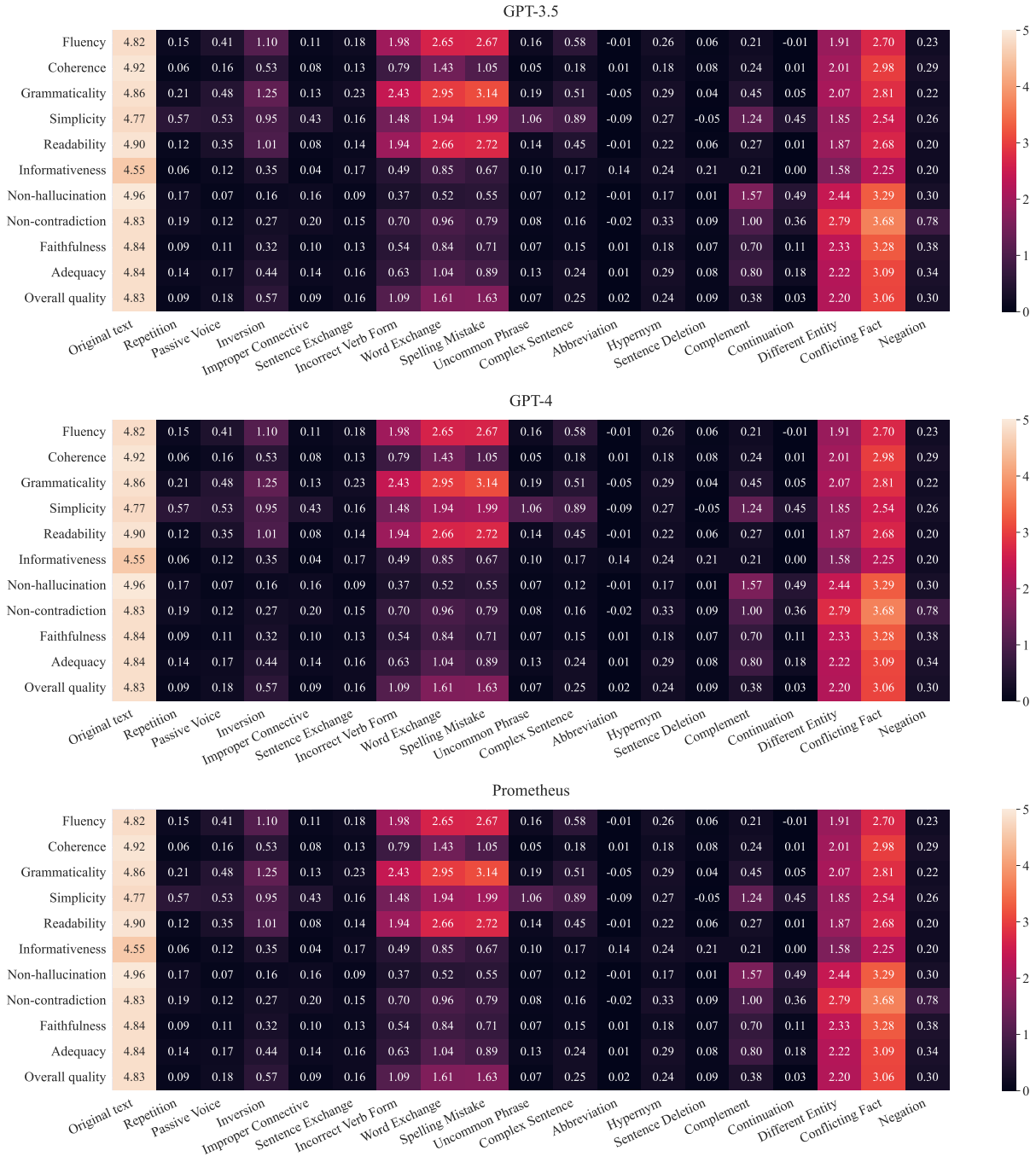


Figure 17: The variances of evaluation scores from three LLMs between original texts and different perturbed texts with the description of list type.

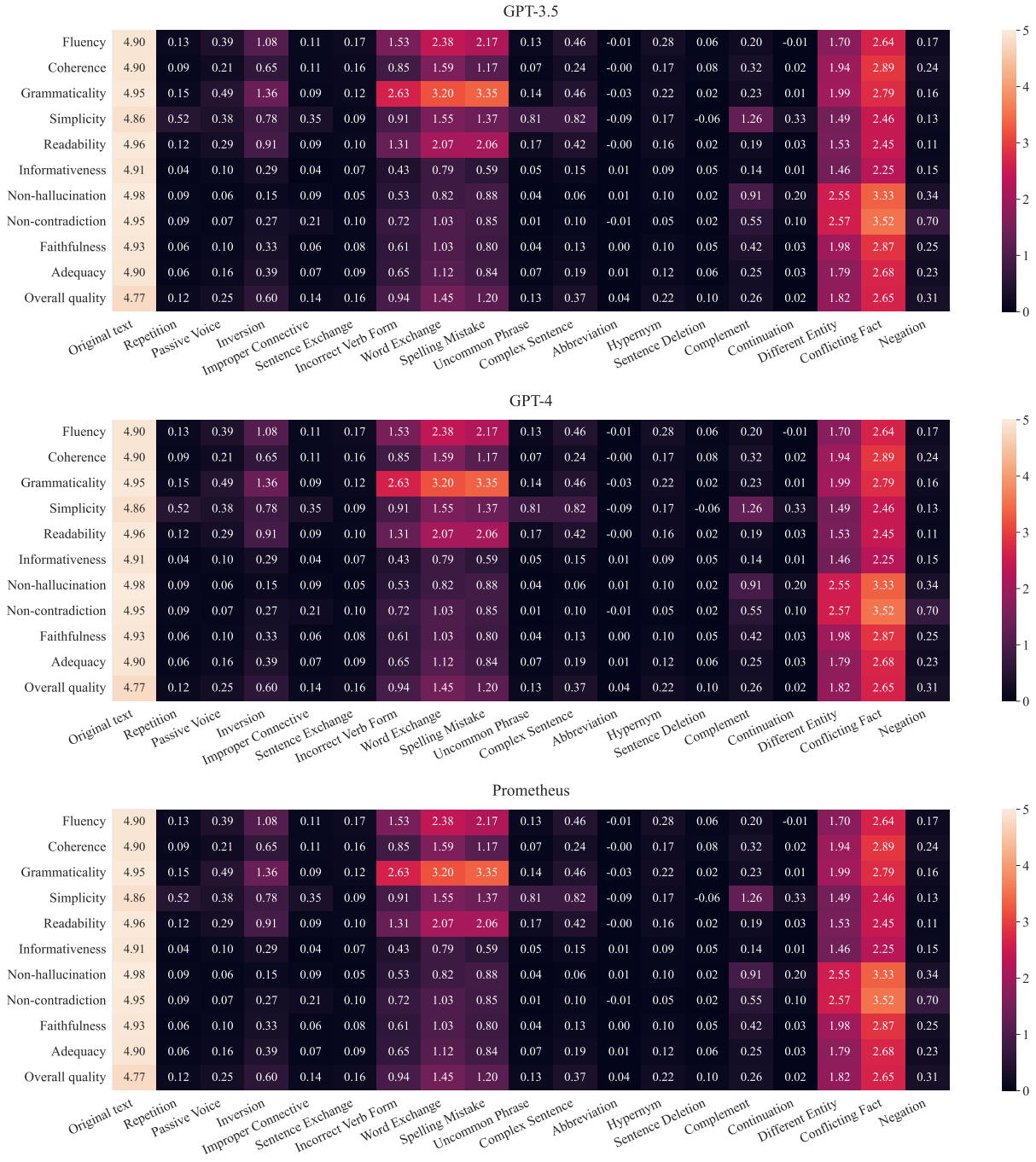


Figure 18: The variances of evaluation scores from three LLMs between original texts and different perturbed texts with the description of term type.

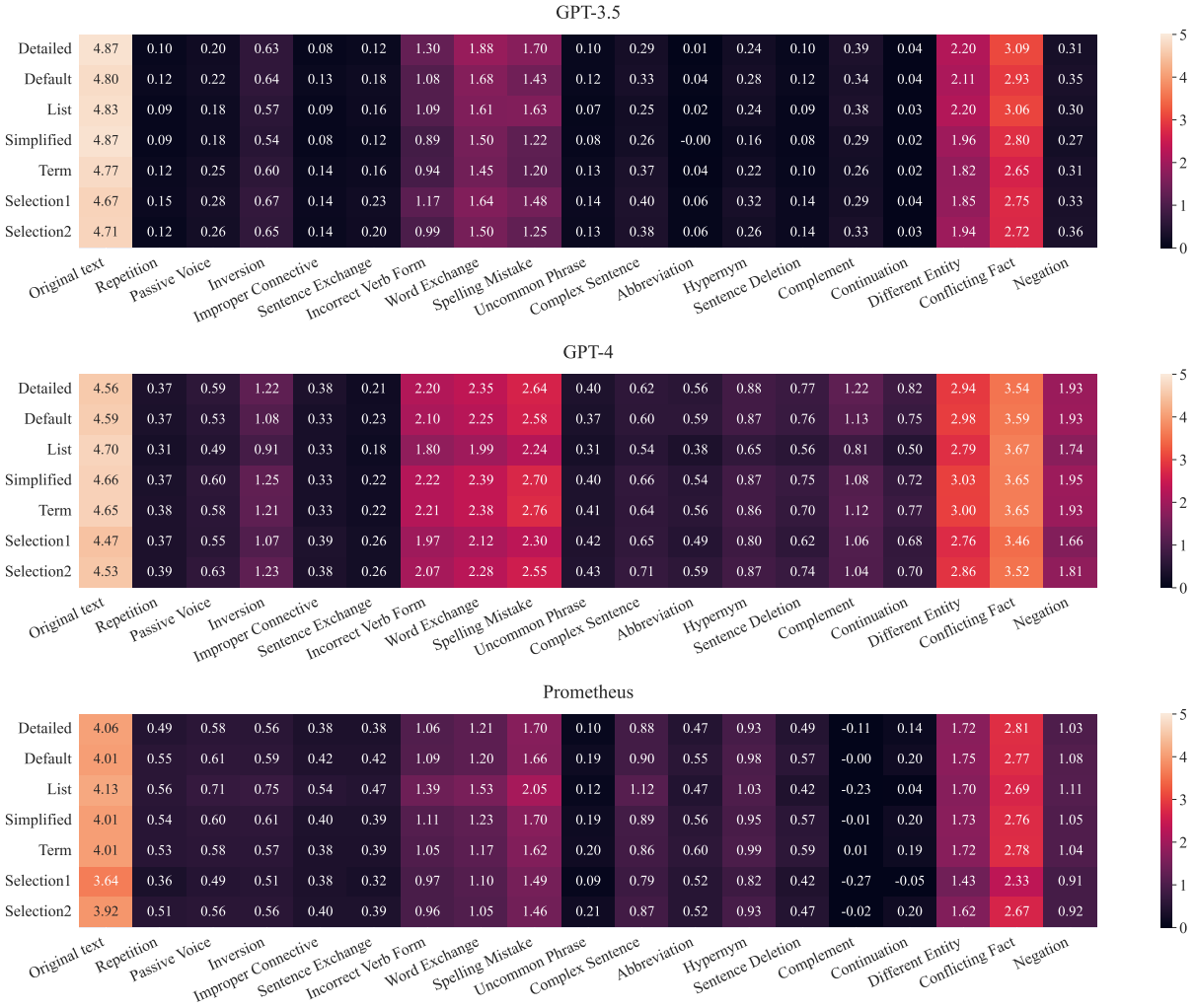


Figure 19: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Overall quality.

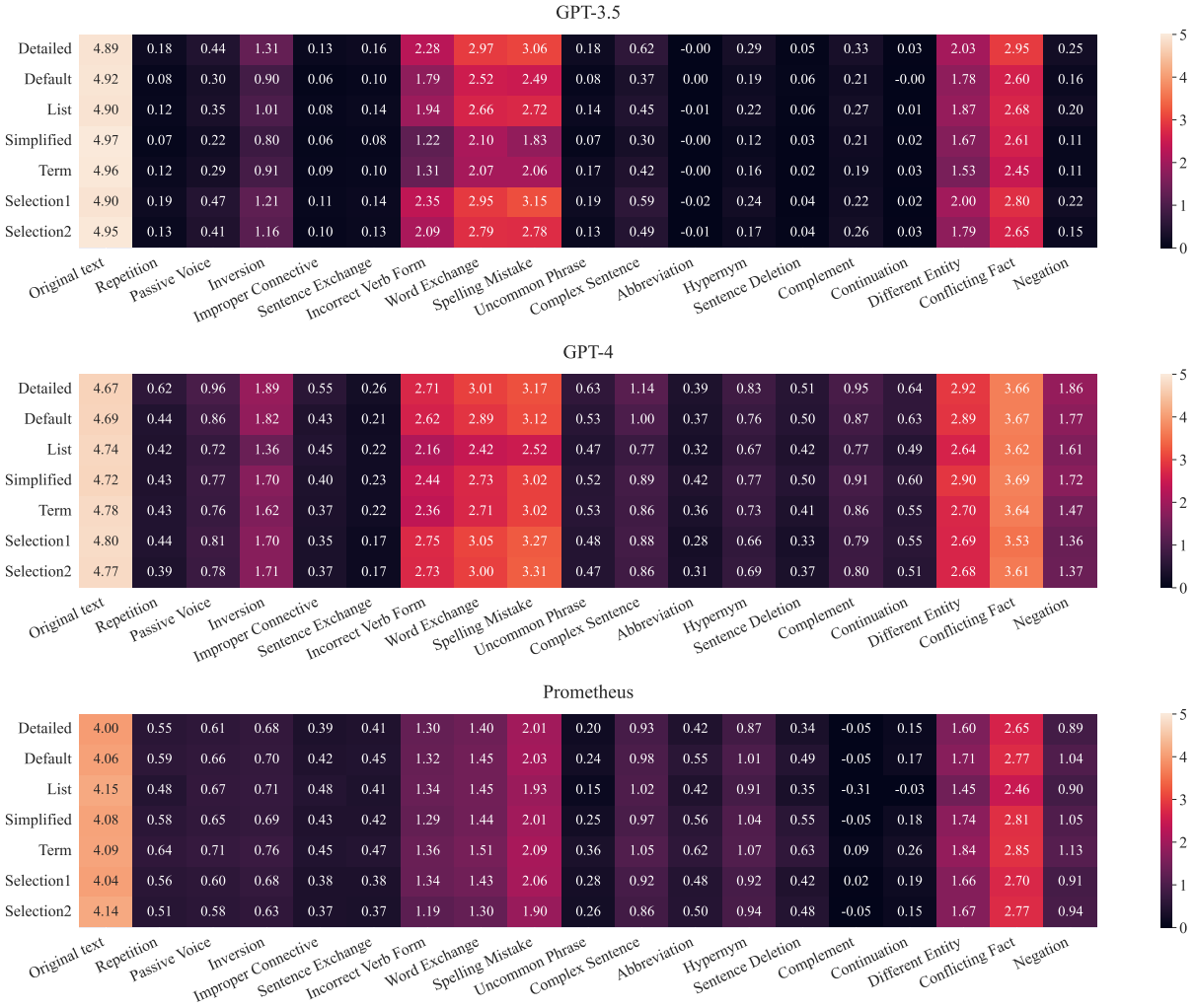


Figure 20: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Readability.

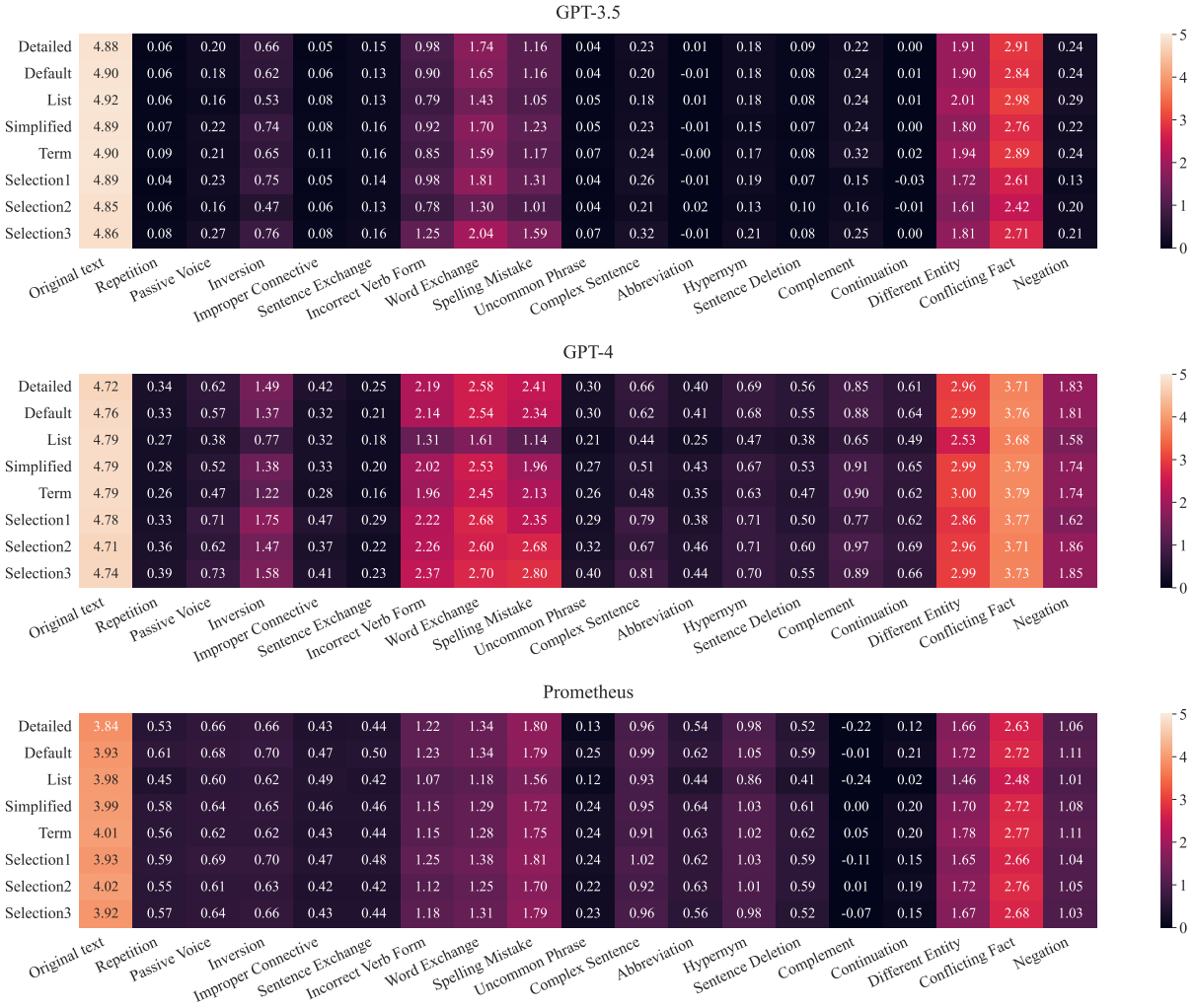


Figure 21: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Coherence.

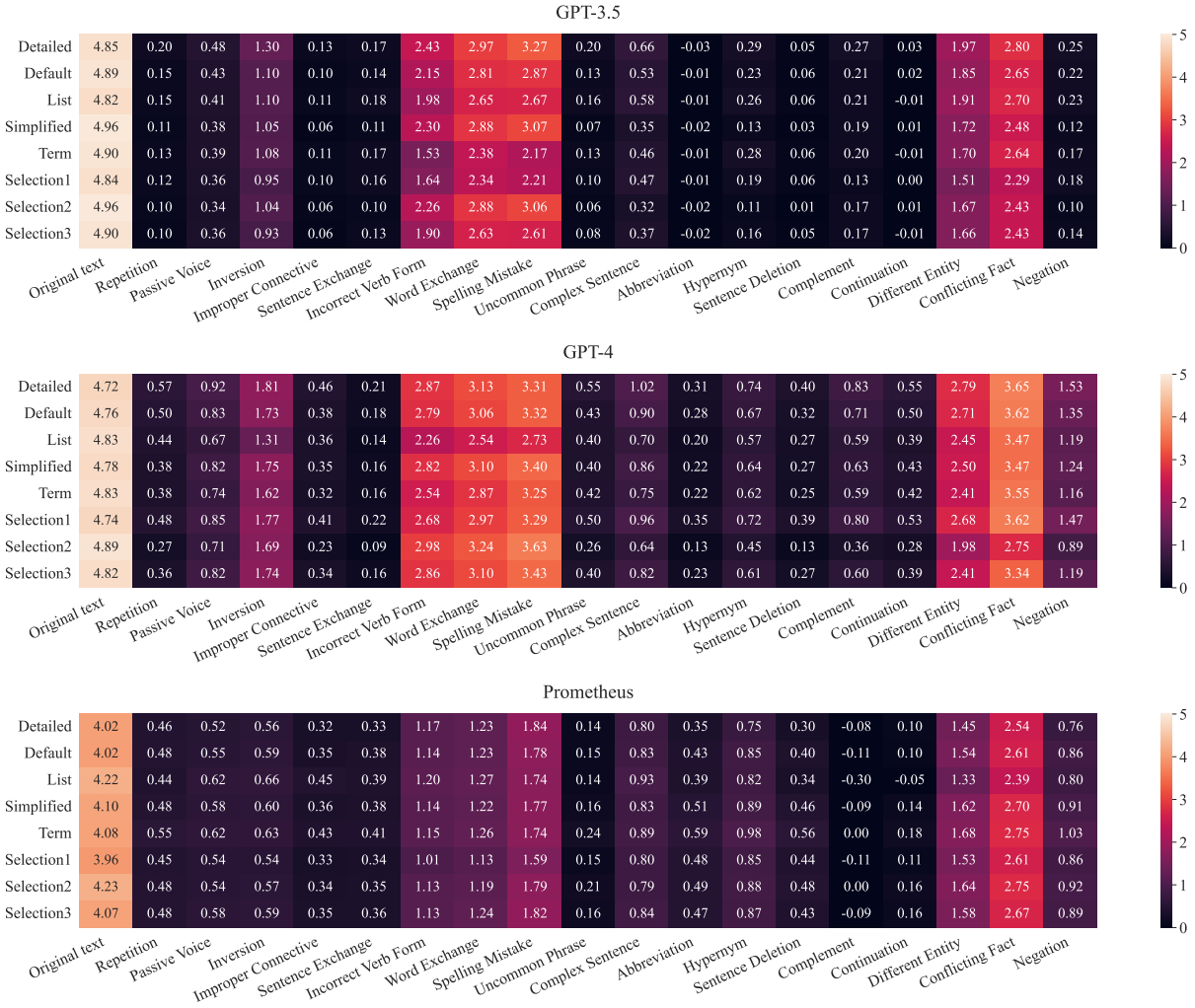


Figure 22: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Fluency.

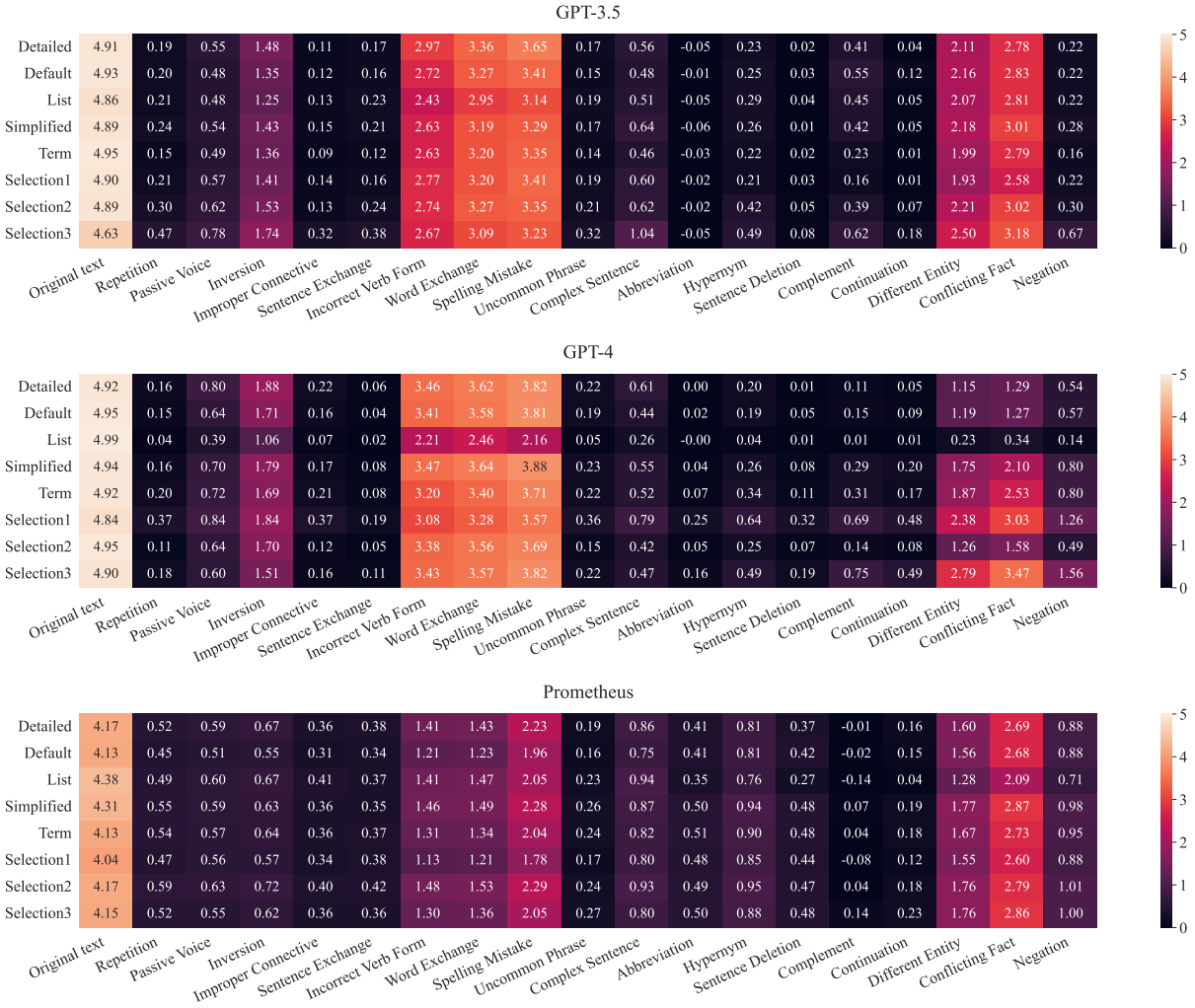


Figure 23: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Grammaticality.

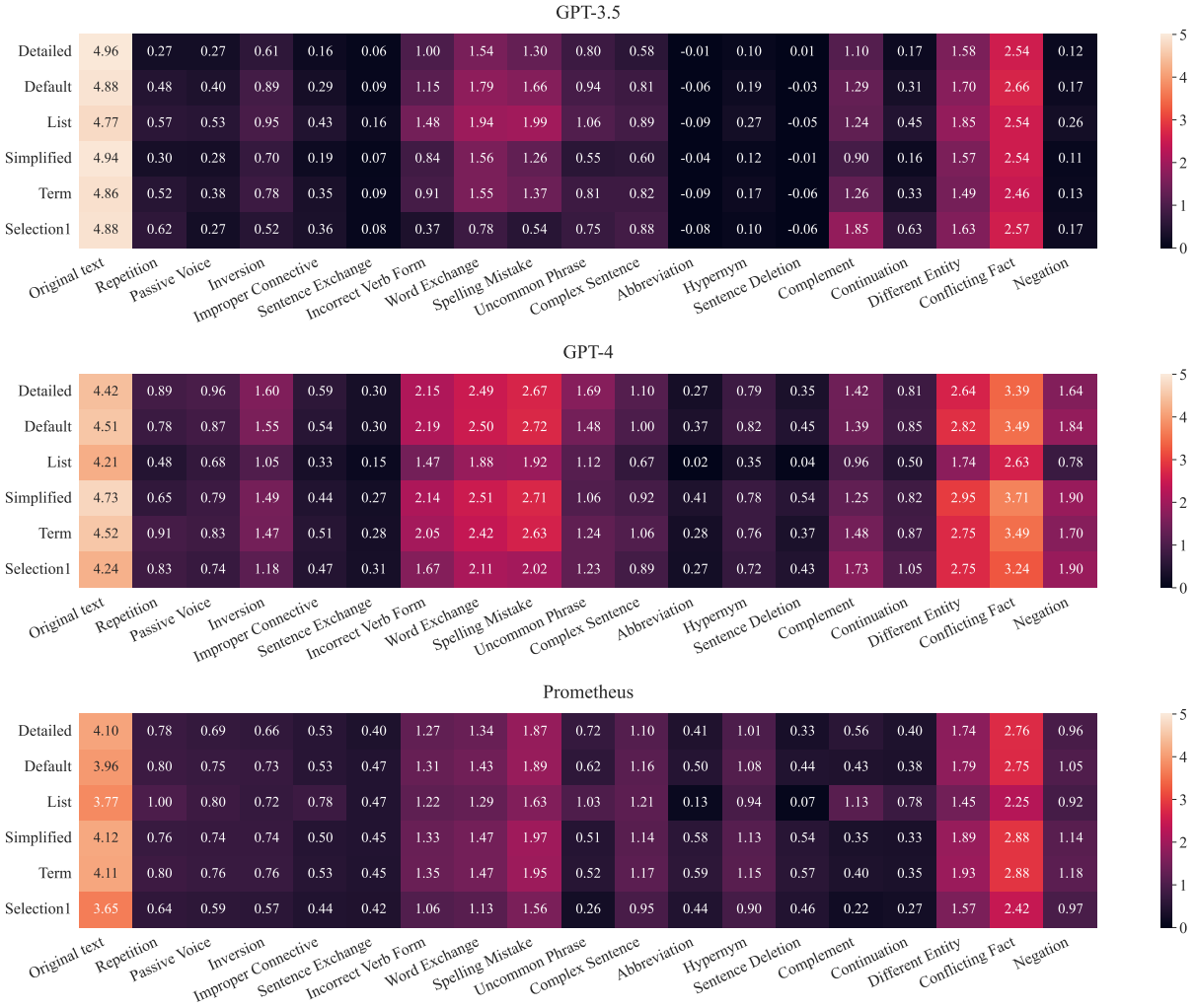


Figure 24: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Simplicity.

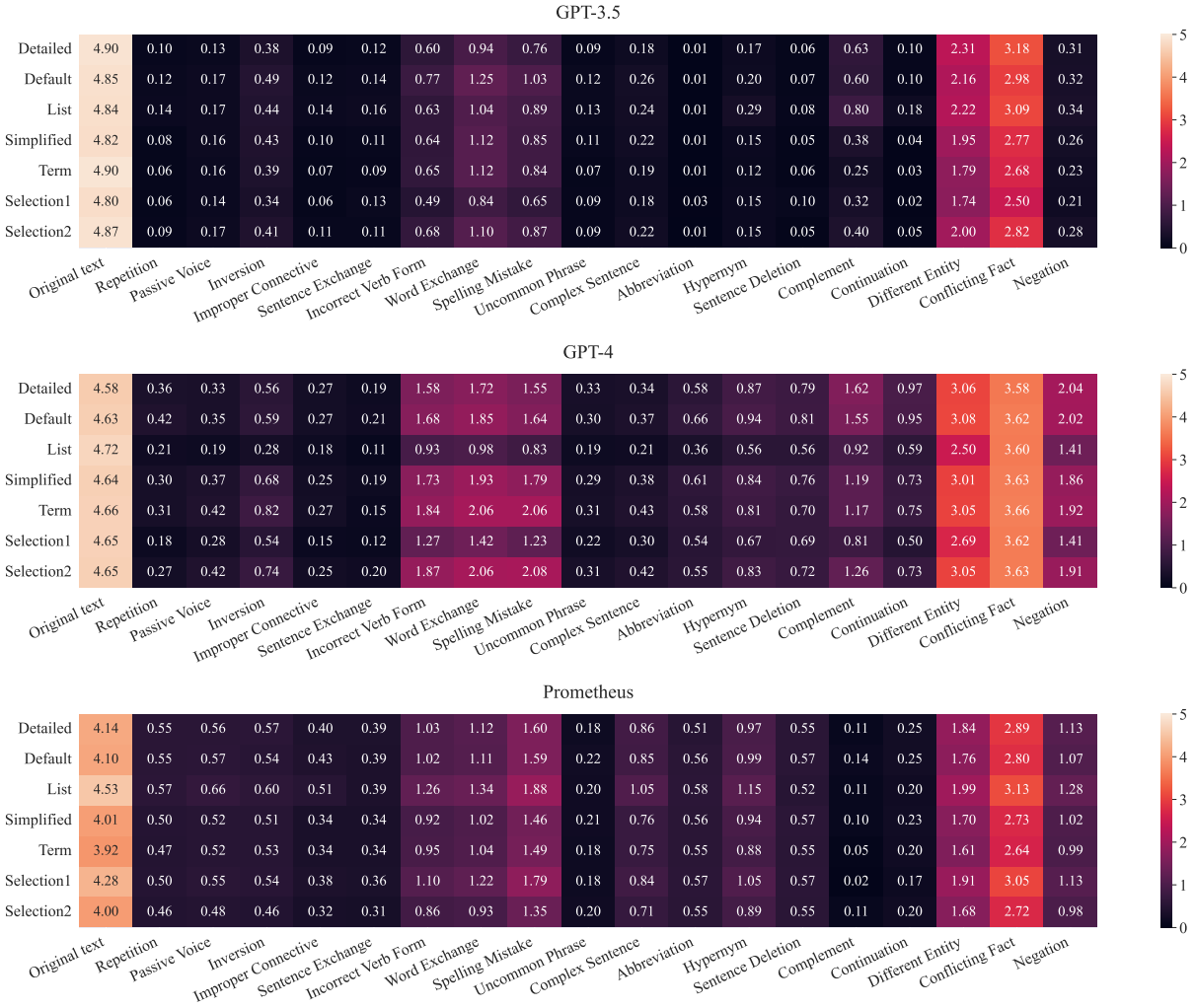


Figure 25: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Adequacy.

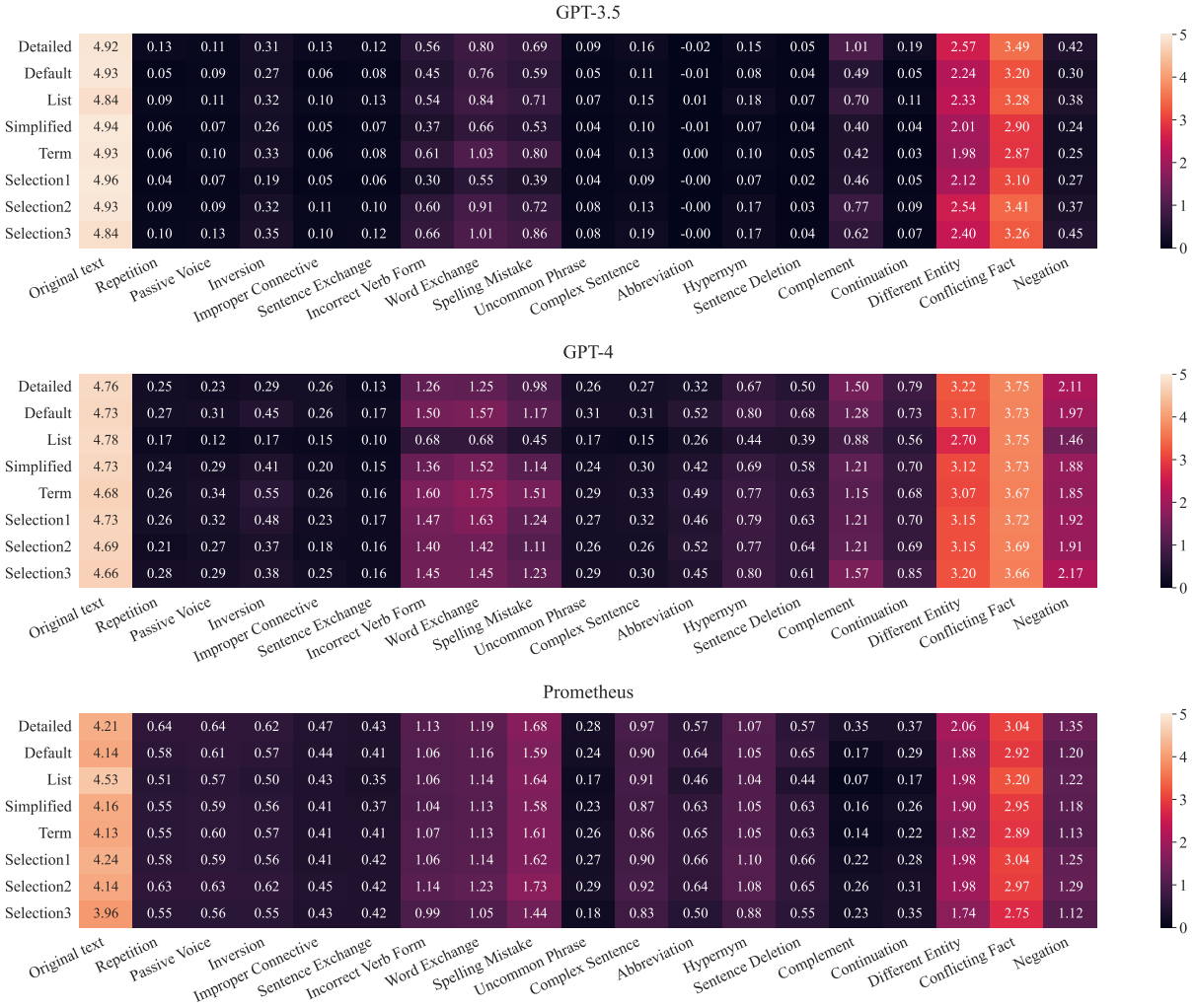


Figure 26: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Faithfulness.

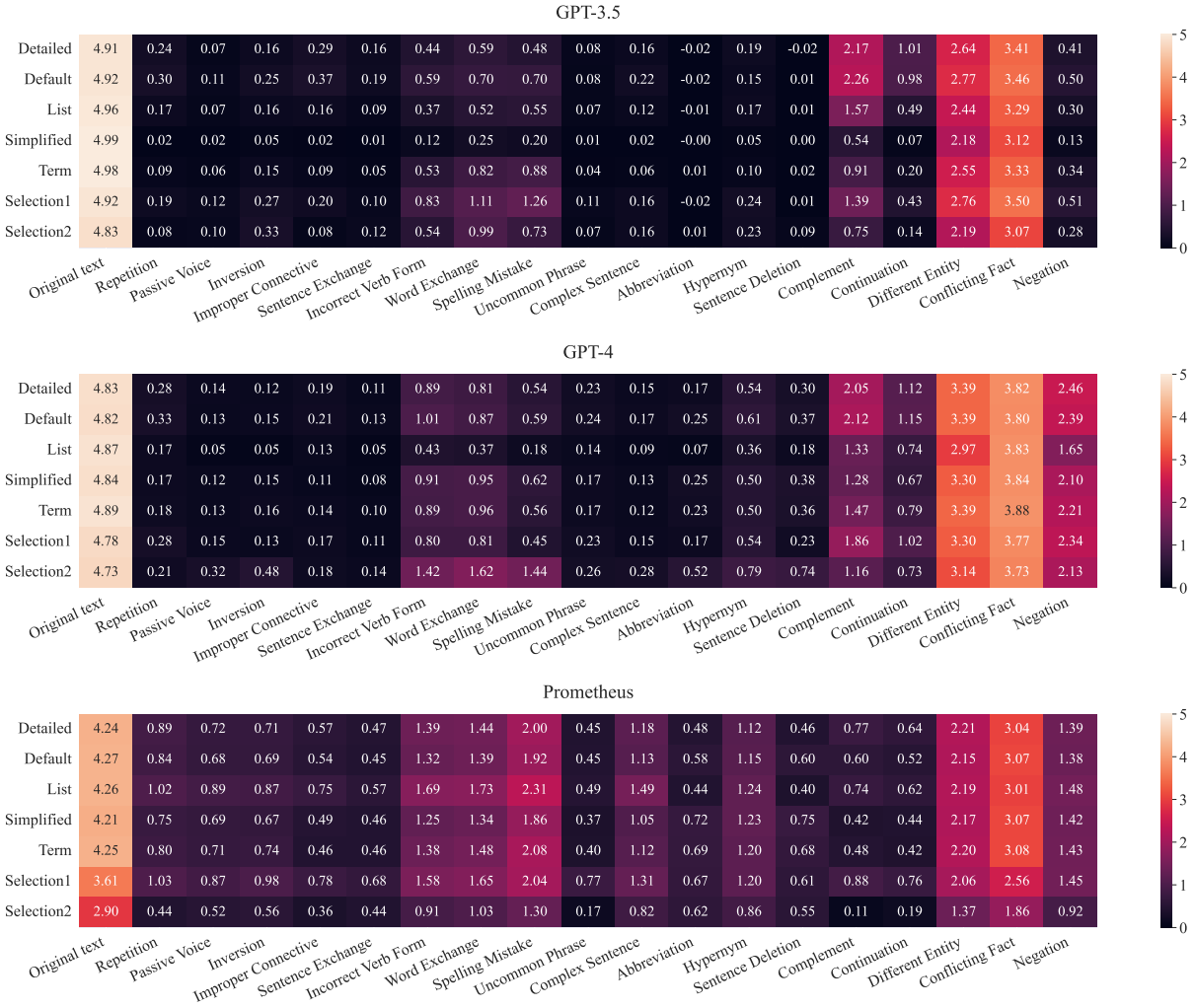


Figure 27: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Non-hallucination.

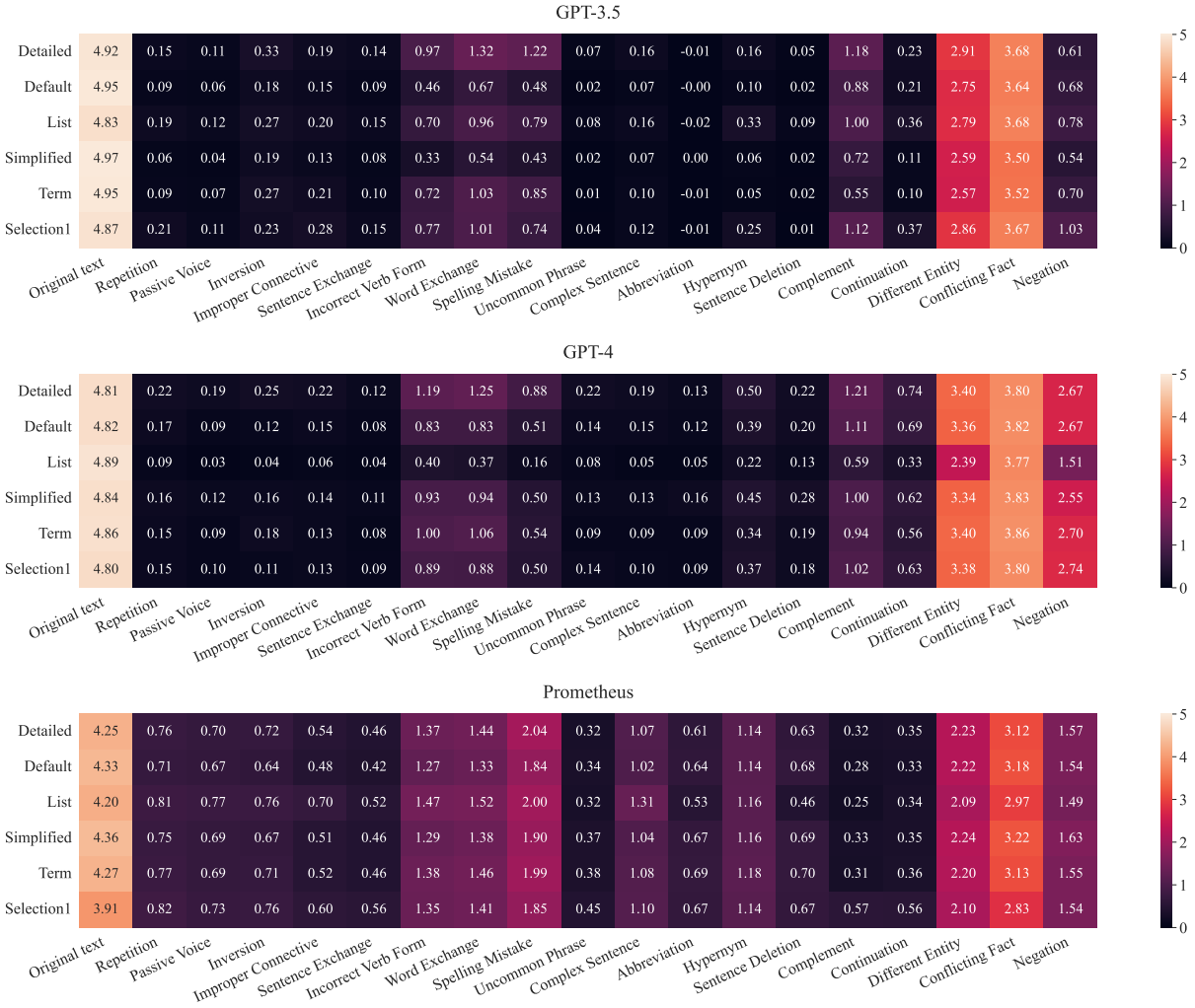


Figure 28: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Non-contradiction.

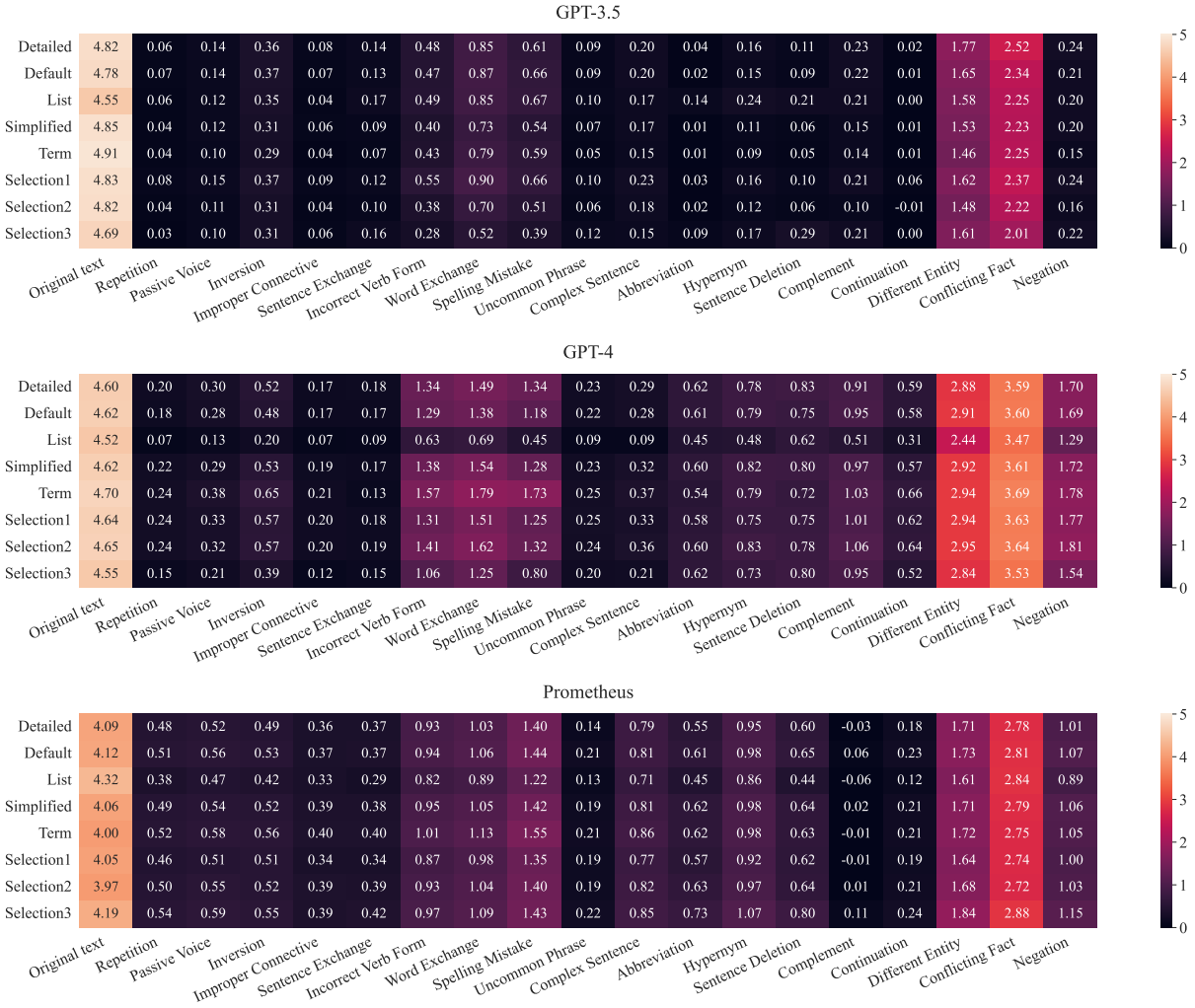


Figure 29: The variances of evaluation scores from three LLMs between original texts and different perturbed texts on Informativeness.

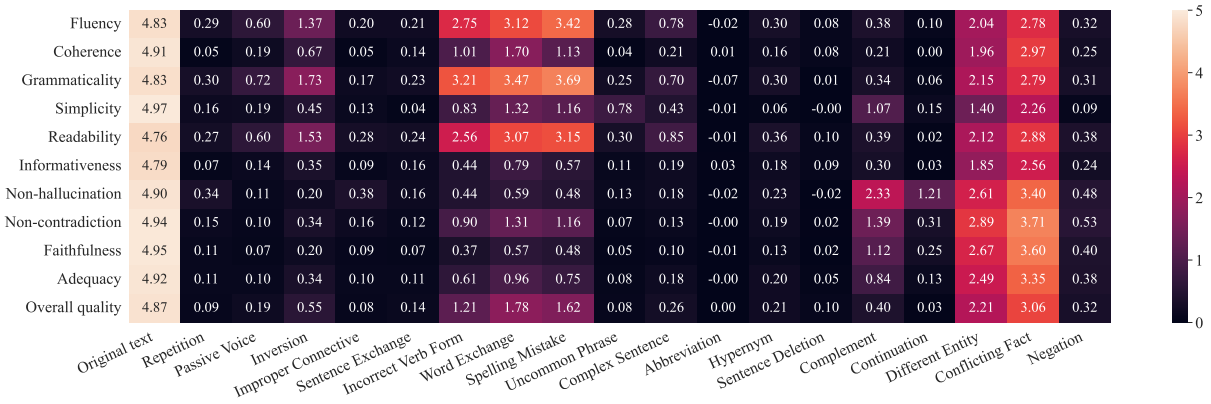


Figure 30: Results of perturbation attacks for the criteria that only retain descriptions.

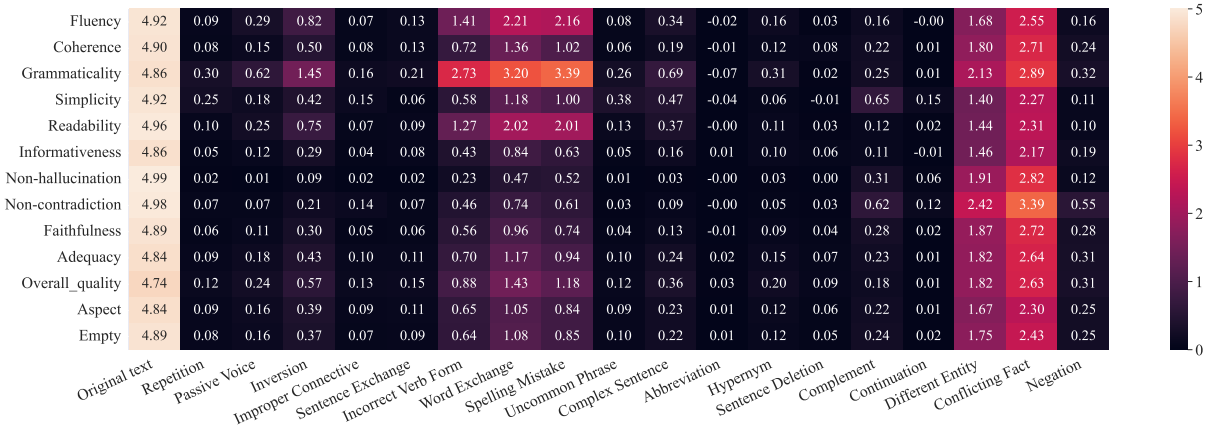


Figure 31: Results of perturbation attacks for the criteria that only retain terms, including the empty criterion and the meaningless criterion with a single word of "Aspect".