

How (Un)faithful are Explainable LLM-based NLG Metrics?

Alex Terentowicz¹ and Mateusz Lango^{1,2} and Ondřej Dušek²

¹Poznan University of Technology, Faculty of Computing and Telecommunications, Poznan, Poland

²Charles University, Faculty of Mathematics and Physics, Prague, Czechia

alex.terentowicz@student.put.edu.pl, {lango, odusek}@ufal.mff.cuni.cz

Abstract

Explainable NLG metrics are becoming a popular research topic; however, the faithfulness of the explanations they provide is typically not evaluated. In this work, we propose a testbed for assessing the faithfulness of span-based metrics by performing controlled perturbations of their explanations and observing changes in the final score. We show that several popular LLM evaluators do not consistently produce faithful explanations.

1 Introduction

Since large language models (LLMs) that exhibit strong instruction-following abilities became available, the NLG community has increasingly adopted LLMs as automatic text evaluators (Li et al., 2025). These LLM-based metrics often achieve good correlations with human judgments, even without using human-written references, and enable the evaluation of customisable aspects of generation quality (Hu et al., 2024b). Moreover, a growing number of these metrics aim to enhance interpretability by providing explanations alongside their scores, ranging from short rationales to rich, multi-component analyses (Liu et al., 2023; Jiang et al., 2023).

One particularly promising approach involves metrics that provide span-based annotations and explanations (Xu et al., 2023; Kartáč et al., 2025; Kasner et al., 2025). These systems highlight specific spans of the generated text where aspect-related errors occur, assess the severity of these issues, provide explanatory comments, and ultimately synthesize this information into an overall quality score. Such granular feedback is potentially valuable for system debugging, building user trust and supporting human annotators (Leiter et al., 2022).

Despite these advances, evaluation of explainable NLG metrics has largely focused on measuring the correlation with human judgments (Zhong et al., 2022; Xu et al., 2023; Liu et al., 2024). Only a few

studies have gone beyond this to assess metrics' robustness to perturbations (Zheng et al., 2023; Hu et al., 2024b), evaluate biases towards LLM-generated outputs (Hu et al., 2024a) or inspect the quality of the explanations themselves (Jiang et al., 2023). Crucially, existing evaluations of explanation quality typically rely on human judgments of plausibility, i.e., whether the explanation seems reasonable or convincing to a human (Jiang et al., 2023; Leiter et al., 2023; Kim et al., 2024; Kartáč et al., 2025). This, however, overlooks a key dimension of explanation quality: faithfulness.

A good explanation should satisfy two core properties: *plausibility*, meaning it is understandable and convincing to humans, and *faithfulness*, meaning it accurately reflects the reasoning process or internal behavior of the system it explains (Jacovi and Goldberg, 2020). Without an assessment of explanation faithfulness, we cannot determine whether these metrics truly provide insight into their computation of the score or merely generate plausible-sounding but potentially misleading rationales.

In this paper, we address this critical gap by proposing comprehensive experiments specifically designed to evaluate the faithfulness of explainable, span-based NLG metrics. Our contributions are threefold: First, we introduce a novel testbed¹ for systematically measuring the faithfulness of explanations provided by NLG metrics, based on perturbing the explanations and measuring changes in the final overall score. Second, we conduct extensive experiments evaluating the explainability of three state-of-the-art LLMs commonly employed in LLM-as-a-judge frameworks. Third, our empirical findings reveal surprisingly low faithfulness of LLM-based metrics, raising concerns about their interpretability and opening new avenues for future research in trustworthy NLG evaluation.

¹Code available at <https://github.com/langus0/faithfull-nlg>.

2 Related Work

There are only a few works that evaluate LLM-based NLG metrics beyond correlations with human judgments and, occasionally, the plausibility of explanations provided.

Liu et al. (2023) used pairs of human- and LLM-written summaries to show that an LLM metric (G-Eval) prefers LLM summaries over human summaries, regardless of their quality. Zheng et al. (2023) performed an analysis of closed-source LLMs using their evaluations of synthetically prepared pairs of generated texts. The experiments revealed that LLMs tend to favour outputs presented in a specific position in the prompt and favour longer or self-generated outputs. Zhang et al. (2024) designed special text perturbations of dialogue data to demonstrate a lack of robustness of LLM-based evaluators to many of them. To check whether LLMs are evaluating a given quality criterion, Hu et al. (2024a) designed special text perturbations affecting only one evaluation aspect. Their experiments showed that LLMs often confuse evaluation criteria despite explicit instructions. These analyses only concerned the final score and did not involve explanations.

Kasner et al. (2025) performed a comparison between human-annotated and LLM-generated spans when evaluating three NLG tasks, demonstrating a high level of agreement between the models and human annotators. However, the study only evaluated the correctness of the model’s textual explanations without further analysis.

3 Methodology

3.1 Motivation and overview

Following the setup of Kartáč et al. (2025), we assume that an explainable span-based NLG metric M takes as input a text generated by NLG system y for a given input x along with a selected quality criterion c . The metric outputs a final score s and a list of explanation triples: (t_i, e_i, a_i) , where each t_i is a text span containing a criterion-related issue, e_i is a textual explanation, and a_i is a severity assessment. Formally,

$$s, \{(t_i, e_i, a_i)\}_{i=1..n} = M(x, y, c)$$

Here, we assume that a_i and s are in $\{1, 2, 3, 4, 5\}$.

Since generating error explanations improves correlations with human judgment and provides the LLM additional reasoning steps, similarly to chain-of-thought (Chiang and Lee, 2023; Liu et al., 2023),

the explanations of NLG metrics are generated before the final score, allowing them to influence the final assessment performed by an autoregressive LLM.

$$s \sim M(s|x, y, c, \{(t_i, e_i, a_i)\}_{i=1..n})$$

Under the assumption that M is faithful, changes to the generated explanation triples should influence the final score s . This motivates our methodology for testing the faithfulness of span-based NLG metrics: **we apply controlled perturbations to the explanation triples and measure the effect on the regenerated score.**

In our experiments, we begin by evaluating an input-output pair (x, y) under a selected criterion c using $M(x, y, c)$, which yields a score s and a set of explanation triples. We then apply controlled perturbations to the generated explanations – for example, by increasing all severity scores a_i by 1 – and regenerate a new score s' while keeping the modified explanation triples fixed:

$$s' \sim M(s|x, y, c, \{(t_i, e_i, a_i + 1)\}_{i=1..n})$$

By comparing the new score s' to the original s , we assess how changes in the explanation content influence the final judgment. For example, in the case of a perturbation where the severity scores are increased (i.e. $a_i \rightarrow a_i + 1$), the evaluator identified the same errors in the generated text, but assessed them as more severe, so an overall lower quality score reflecting this is expected.

3.2 Perturbations

We designed the following types of perturbations targeting all elements of the provided explanations $\{(t_i, e_i, a_i)\}_{i=1..n}$:

- **Severity Score** – We increase or decrease the severity score a_i of the triples by a given value, assuming that identifying errors as more severe should lead to a lower final score and vice-versa. We perform experiments at the *individual* level where the severity of a single error is modified and when *all* severities are modified at once.
- **Textual Explanation** – The generated textual explanations e_i are modified by rewriting them to sound more/less severe using text style transfer (Jin et al., 2021). We prompt an LLM to rewrite the error explanation to match a given severity level. The prompt is available in App. B

Data	Model	Severity Score				Explanation				Both				Add/Rem. Errors				Crit. Err.
		-1	+1	-2	+2	-1	+1	-2	+2	-1	+1	-2	+2	-1	+1	-2	+2	
HANNA	Nemotron	0.24	0.13	0.35	0.22	0.47	0.19	0.59	0.20	0.81	0.25	0.93	0.26	0.06	0.07	0.07	0.08	0.19
	Gemma	0.19	0.27	0.29	0.49	0.32	0.14	0.34	0.19	0.51	0.37	0.65	0.55	0.28	0.26	0.51	0.27	0.22
	Qwen	0.15	0.24	0.21	0.57	0.5	0.16	0.56	0.19	0.68	0.41	0.83	0.75	0.07	0.13	0.10	0.16	0.44
Summ-Eval	Nemotron	0.13	0.15	0.22	0.28	0.33	0.37	0.39	0.44	0.44	0.49	0.56	0.61	0.22	0.23	0.41	0.30	0.66
	Gemma	0.15	0.31	0.48	0.68	0.43	0.35	0.47	0.60	0.63	0.62	0.86	0.79	0.49	0.32	0.30	0.49	0.71
	Qwen	0.18	0.26	0.31	0.51	0.35	0.27	0.40	0.33	0.47	0.47	0.55	0.68	0.29	0.38	0.37	0.50	0.50
QAGS	Nemotron	0.12	0.14	0.21	0.21	0.25	0.24	0.29	0.29	0.36	0.32	0.43	0.44	0.37	0.19	0.36	0.26	0.51
	Gemma	0.14	0.34	0.28	0.60	0.24	0.14	0.32	0.23	0.36	0.52	0.56	0.59	0.35	0.36	0.80	0.45	0.47
	Qwen	0.27	0.19	0.42	0.28	0.18	0.06	0.19	0.08	0.43	0.27	0.54	0.36	0.34	0.20	0.50	0.24	0.24

Table 1: The proportion of changed predictions after applying different perturbations to the explanation triples of the metric. The results are averaged over all aspects of a given dataset.

- **Adding/Removing Errors** – We modify the number of detected errors n , by removing detected errors from the list, making the generated texts seem more error-free. Conversely, we also experimented with making the list of errors longer by adding synthetic errors to the explanation triples. The synthetic errors had a randomly selected text span from the text being assessed y , a random severity, and an explanation generated by LLM (see prompt in App. B).
- **Critical Error** – One particular error addition is a critical error encompassing the whole text (text span $t = y$) with maximum severity assigned and explanation “This error completely compromises the quality of this text on the selected aspect”.

3.3 Metrics

To assess the change of introduced perturbations to explanation triples, we use two simple measures: 1) the proportion of changed predictions ($s \neq s'$) – we check whether the score s' obtained after modifying the explanation (e.g., increasing the severities of errors found) is different from the originally obtained score s ; 2) the average change of overall scores ($avg(s' - s)$) – we measure the extent of change caused by the perturbation.

4 Experiments

4.1 Experimental setup

Datasets We conduct experiments on three popular datasets for NLG metrics meta-evaluation, which range over two tasks: summarization – QAGS (Wang et al., 2020), SummEval (Fabbri et al., 2021); and story generation – HANNA (Chhun et al., 2022).

Dataset	Model	Spearman ρ	Kendall τ
QAGS	Gemma	0.630	0.571
	Nemotron	0.659	0.595
	Qwen	0.644	0.581
SummEval	Gemma	0.442	0.388
	Nemotron	0.451	0.397
	Qwen	0.478	0.415
HANNA	Gemma	0.403	0.347
	Nemotron	0.446	0.378
	Qwen	0.394	0.335

Table 2: Correlations of LLM metrics with human evaluations on studied datasets, averaged over used aspects.

QAGS contains binary judgments of factual consistency for 235 CNN/DailyMail (Hermann et al., 2015) summaries and 239 XSum (Narayan et al., 2018) summaries. SummEval provides Likert-scale annotations on four quality aspects (factual consistency, relevance, coherence, fluency) for summaries generated by 16 systems for 100 CNN/DailyMail articles. HANNA comprises 1,056 stories derived from 96 prompts: for each prompt, one human writer and ten automatic story generation systems produced one story each. Every story was then evaluated by three annotators on six criteria.

Since we do not expect the faithfulness of the metric to be correlated with evaluated aspects, we decided to evaluate only a subset of available aspects: *factual consistency*, *coherence* and *relevance* on SummEval; *coherence*, *relevance* and *complexity* on HANNA. The QAGS dataset contains evaluations of summary *factual consistency*.

Metrics Recently, Kartáč et al. (2025) demonstrated that open-source LLM models can achieve similar or even better results than specialized NLG metrics or closed-sourced LLMs. Following their experimental setup, we prompted an open-source

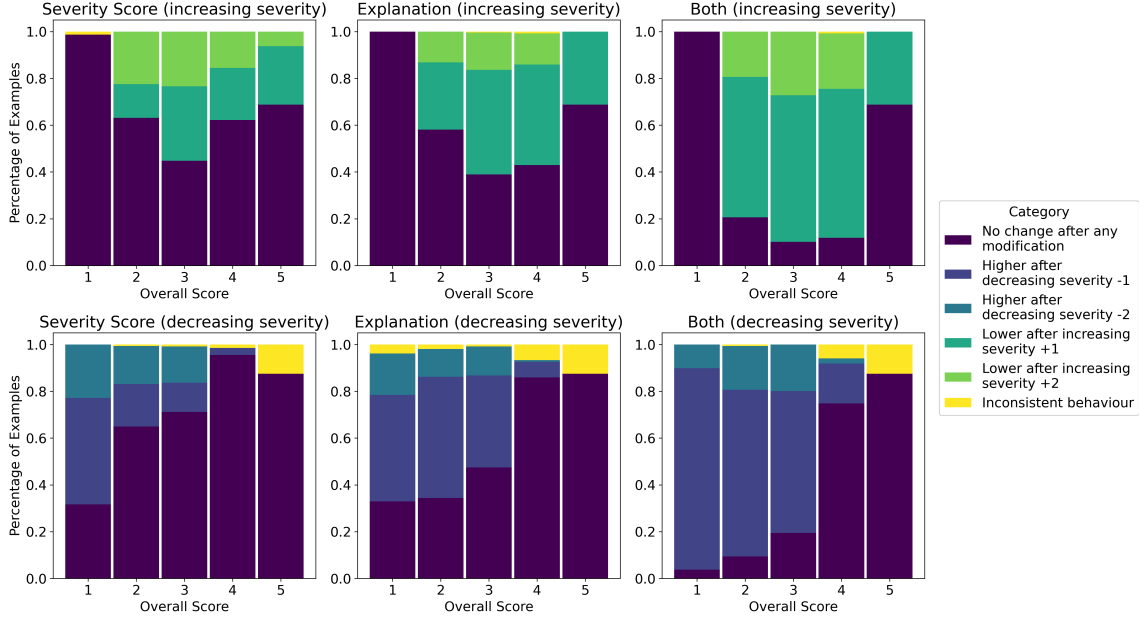


Figure 1: The percentage of examples whose overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. This and the following plots show representative results for factual correctness evaluated by Nemotron on SummEval (see the remaining plots in App. E).

LLM to analyse the given text, identify error spans related to a specific aspect and provide an output in the form of explanation triples and a final score. The prompt contained a description of the evaluated task and aspect, a template for the model’s response as well as the input and output of the NLG system (see full prompt in App. C). As evaluator models, three popular models were selected: Llama 3.1 Nemotron 70B (Wang et al., 2025), Qwen 2.5 72B (Yang et al., 2024) and Gemma 2 27B (Gemma Team et al., 2024). The meta-evaluation results of our metrics (i.e., correlations of final scores with human annotations) are presented in Tab. 2.

4.2 Results

We applied our perturbations to all evaluations with identified errors² and checked for score changes. The proportion of changed predictions for all perturbations is presented in Table 1.

The percentage of changed predictions for error severity modifications (both textual explanations and numerical severity assessments) applied simultaneously to all errors as a function of the original final score is presented in Fig. 1. As expected, increasing the error severity of examples that were al-

ready assigned the lowest possible score has no impact. However, surprisingly, decreasing the severity of all errors by 1 for about 10% of examples with the best overall score makes the final score *lower*, only for it to return to its highest value after decreasing it further (shown in yellow in Fig. 1). This result is consistent when modifying severity by changing the numerical severity assessment, the textual explanation, or both in combination.

There is also a large group of examples (74%) with a high evaluation of 4/5, which will not obtain the full score even if the severity of all errors is decreased by two. Similarly, the evaluation of over 90% of examples with the highest score remains unaffected by increasing error severity using any of the proposed perturbations.

Fig. 2 presents averaged change in final score as a function of total severity modification (sum of severity changes to all errors found for an example). In general, it seems that the models often disregard the assigned numerical values of error severity and are more sensitive to error explanations that sound more negative or positive. Combining two perturbations, i.e. changing both the numerical severity value and the explanation, has a more substantial effect. Different LLMs do not appear to be affected symmetrically by increasing or decreasing error severities, e.g. Nemotron’s decisions are more af-

²Examples with no errors found were excluded from the analysis, as the perturbations are not applicable.

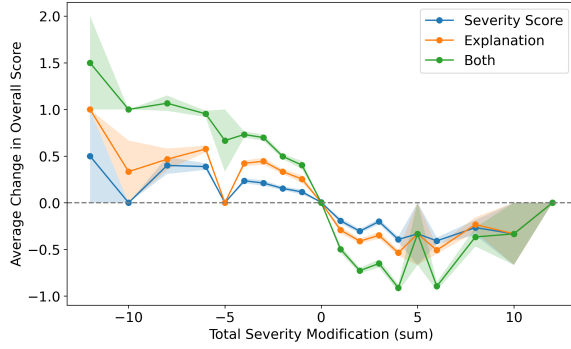


Figure 2: Average change in overall score in relation to total severity modification using different perturbations.

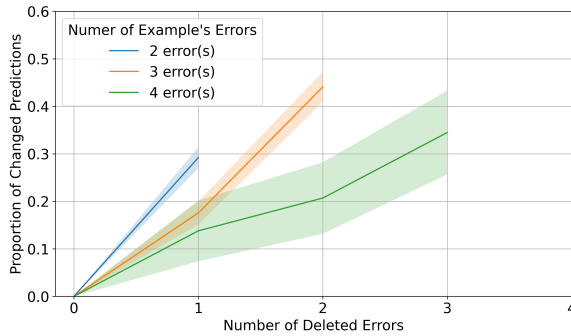


Figure 3: Proportion of changed predictions vs. the number of deleted errors for examples having 2, 3 or 4 errors.

ected by decreasing error severity than increasing it (93% vs 26% on HANNA).

Fig. 3 shows the impact of error deletion. As expected, deleting more errors leads to a higher proportion of changes in the final scores. However, the overall effect remains moderate: even when all but one error are removed, only 32–37% of predictions are affected. Interestingly, we also observed that removing errors from explanation triples may *decrease* the final score. While this typically occurs in about 1–3% of cases, it can reach up to 18% for certain models and aspect pairs. Quite surprisingly, this inconsistent metric behavior often occurs for examples that initially received the lowest possible score (see additional visualizations in App. E.2).

The impact of adding a critical error appears to depend on the task being evaluated (see Tab. 1). For story generation, the models are quite insensitive as the final score remains unchanged for the majority (> 56%) of examples. For the remaining tasks, this proportion still remains relatively high (> 29%), considering the severity of the perturbation. Additionally, a more detailed analysis presented in Fig. 4 reveals that, even if the prediction changes,

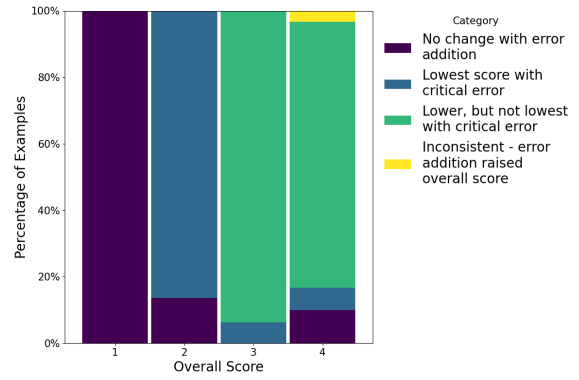


Figure 4: The percentage of examples whose overall score changed after adding a critical error. The results are presented for Nemotron on the QAGS dataset.

adding a critical error does not result in the lowest overall score if the initial assessment was high.

5 Summary

This paper draws attention to the issue of explanation faithfulness in span-based NLG metrics, which has not been addressed in previous research or included in metrics’ meta-evaluations. To bridge this gap, we have proposed a new testbed for evaluating explanation faithfulness using special perturbations of the explanations.

Our analysis revealed that the faithfulness of explanations provided by LLM-based metrics is limited, prompting a call for caution when interpreting their outputs. For example, interpretations such as “this text would receive a lower score if the errors were assessed as more severe” were found to be incorrect in most cases. The models are more likely to change their assessment if provided with explanations that *sound* as if errors are more/less severe than to an actual numerical severity assessment. Inconsistent metric behaviours with respect to severity changes or error deletions were also observed. The explanations provided by the current metrics were also found to be poorly calibrated, as they do not semantically react to an increase or decrease in error severity.

The reasons behind the low faithfulness of the explanations offered by NLG metrics require further investigation. One potential reason might be that LLMs evaluate each instance independently and lack clear points of reference. Another reason may be that, although error analysis serves as the reasoning outcome instead of the standard chain of thought in LLM metrics, it may not be sufficiently considered when predicting the final score.

Limitations

The presented methodology focuses on the faithfulness with respect to the final score, which is assumed to be predicted by an LLM alongside explanations. While this is true for the majority of existing NLG metrics, there are some notable exceptions like TIGERScore (Jiang et al., 2023). This metric performs error analysis with penalty scores (severity assessments) and computes the final score as a sum of it. This results in full faithfulness with respect to the final score, as it is not predicted by a machine learning model, at the cost of achieving lower correlations with human judgments.

While the faithfulness of NLG metrics is essential to make sure that the detailed error analysis is reliable and the researchers can trust that the provided reasons actually influenced the obtained scores, in some applications outside of NLG evaluation, like language learning, the obtained unfaithful explanations can still be useful (Naismith et al., 2023).

Ethics Statement

We do not anticipate any negative ethical implications arising from this study. The licenses for LLMs models permit our use of the model weights. We used AI-assisted coding (i.e. Copilot) with the bulk being human-written. For writing, AI was used to check grammar mistakes.

Acknowledgments

Co-funded by the European Union (ERC, NG-NLG, 101039303) and National Science Centre, Poland (Grant No. 2022/47/D/ST6/01770). This work used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101). The authors thank Ivan Kartáč for his feedback on an earlier version of the manuscript. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

- Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [A closer look into using large language models for automatic evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.
- Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Gemma Team Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *NIPS*, pages 1693–1701.
- Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024a. [Are LLM-based evaluators confusing NLG quality criteria?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570, Bangkok, Thailand. Association for Computational Linguistics.
- Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024b. [Themis: A reference-free NLG evaluation language model with flexibility and interpretability](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui Chen. 2023. [Tigerscore: Towards building explainable metric for all text generation tasks](#). *ArXiv*, abs/2310.00752.
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. [Deep Learning for Text Style Transfer: A Survey](#). *Computational Linguistics*, pages 1–51.
- Ivan Kartáč, Mateusz Lango, and Ondřej Dušek. 2025. Openlauge: An explainable metric for nlg evaluation with open-weights llms. *arXiv preprint arXiv:2503.11858*.
- Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek,

- and Simone Balloccu. 2025. [Large language models as span annotators](#). *Preprint*, arXiv:2504.08697.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. [Prometheus: Inducing fine-grained evaluation capability in language models](#). *Preprint*, arXiv:2310.08491.
- Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. [Towards explainable evaluation metrics for natural language generation](#). *Preprint*, arXiv:2203.11131.
- Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. [The Eval4NLP 2023 shared task on prompting large language models as explainable metrics](#). In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of llm-as-a-judge](#). *Preprint*, arXiv:2411.16594.
- Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024. [X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8560–8579, Mexico City, Mexico. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. [Automated evaluation of written discourse coherence using GPT-4](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.
- Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025. [Helpsteer2-preference: Complementing ratings with preferences](#). *Preprint*, arXiv:2410.01257.
- 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Chen Zhang, Luis Fernando D’Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. [A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. [Towards a unified multi-dimensional evaluator for text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Frequently Asked Questions (FAQ)

- *Do you deviate from the standard technique of measuring the correlation of automatic scores with human judgments?*

NLG metrics should be evaluated by measuring their correlation with human judgement in order to assess the resulting quality scores s . However, we find this evaluation insufficient in the context of explainable NLG metrics, and we therefore propose an *additional* evaluation of explanation faithfulness. As we believe that the standard evaluation should always be performed, Table 2 shows the results of the standard evaluation, demonstrating the correlation between the quality scores provided by the NLG metrics used and human judgement.

- *Is the proposed testbed suffering from data contamination bias?*

We propose a testbed that involves applying specific perturbations to the evaluators' output. This methodology can be applied to any dataset, including private ones not exposed to LLMs. In our experiments, we used three popular benchmarks that have potentially been leaked to LLMs. We conducted our experiments on three different LLM families and observed a lack of faithfulness to explanations on all of them.

- *Do your claims, e.g. "as expected, increasing the error severity of examples that were already assigned the lowest possible score has no impact" presuppose the low scores are accurate and would strongly correlate with human judgements?*

Our claim relates to the error severity assessment a_i which are provided in the metrics' explanations. As explanations of ML systems are offered to users to help them understand how the system operates, the severity assessments provided should behave in a way that makes sense to humans. The correlation between severity assessments and human expectations (or other aspects related to explanation faithfulness) has not been measured in the previous works on explainable NLG metrics. This paper proposes a new methodology to address this critical gap.

B Perturbation Prompts

The prompt for textual explanation perturbation is available in Listing 1. The prompt for generating an explanation for artificially introduced error is available in Listing 2.

C Evaluator prompts

Our evaluator prompts are adopted from (Kartáč et al., 2025). For self-completeness, we present the metric prompt for summarization task on Listing 3.

D Modification examples

Listing 6, 7, and 8 present modifications that increase the overall annotation score. These are modified versions of the annotation in Listing 5, which evaluated summarization in Listing 4. Conversely, Listings 11 illustrate modifications that reduce the severity of annotation in Listing 10, which is based on the generation shown in Listing 9.

E Additional visualizations

E.1 Changing severity and/or explanations

The visualizations of the proportions of prediction change in relation to final score are provided for all datasets, aspects, and models:

- HANNA dataset, respectively for Nemotron, Gemma, and Qwen
 - coherence – Figures 5, 6, 7.
 - complexity – Figures 8, 9, 10.
 - relevance – Figures 11, 12, 13.
- SummEval dataset, respectively for Nemotron, Gemma, and Qwen
 - coherence – Figures 14, 15, 16.
 - factual consistency – Figures 17, 18, 19.
 - relevance – Figures 20, 21, 22.
- QAGS dataset, respectively for Nemotron, Gemma, and Qwen
 - factual consistency – Figures 23, 24, 25.

The visualizations of the average prediction change in relation to total modification of error severity are provided for all aspects and models:

- HANNA dataset, respectively for Nemotron, Gemma, and Qwen:
 - coherence – Figures 26, 27, 28.

- complexity – Figures 29, 30, 31.
- relevance – Figures 32, 33, 34.
- SummEval dataset, respectively for Nemotron, Gemma, and Qwen:
 - coherence – Figures 35, 36, 37
 - factual consistency – Figures 38, 39, 40.
 - relevance – Figures 41, 42, 43.
- QAGS dataset, respectively for Nemotron, Gemma, and Qwen:
 - factual consistency – Figures 44, 45, 46.

E.2 Error addition and critical error

The visualizations of the proportions of prediction change in relation to final score after addition of errors are provided for selected datasets in all aspects, and models:

- HANNA dataset, respectively for Nemotron, Gemma, and Qwen:
 - coherence – Figures 47, 49, 51 for the addition of random errors, and 48, 50, 52 for the addition of a critical error.
 - complexity – Figures 53, 55, 57 for the addition of random errors, and 54, 56, 58 for the addition of a critical error.
 - relevance – Figures 59, 61, 63 for the addition of random errors, and 60, 62, 64 for the addition of a critical error.
- QAGS dataset, respectively for Nemotron, Gemma, and Qwen:
 - factual consistency – Figures 65, 67, 69 for the addition of random errors, and 66, 68, 70 for the addition of a critical error.

You are a Textual Style Transfer (TST) system, which changes the sentiment polarity of a given text.

You will be given an error explanation of a certain severity level, in the format:

Original Explanation: {{ explanation }}

Your task is to adjust it to make it sound {severity_change_str: more/less} severe. You will provide the error explanation with the style changed, in the format:

Modified Explanation: {{ explanation styled as {severity_change_str} severe }}

Do not add any additional text, comments, or severity mark. Provide only the modified explanation line.

There are five error severity levels:

Suggestion (1): optional improvement, not necessarily wrong. Example suggestion explanation:

'Explanation: This statement is out of context in the summary. The original article mentions the follower count as additional information about their online presence, but in the summary, it appears as a standalone fact without explaining its relevance to the main topic (their travels). However, this is more of a contextual issue, but since the numbers are accurate, the severity for factual consistency is relatively low.'

Minor (2): small error that doesn't hinder understanding. Example minor error explanation:

'Explanation: While not entirely inaccurate, this sentence lacks crucial contextual information present in the article (e.g., overcoming her father's death). However, since it doesn't introduce new inconsistent facts but rather omits them, its severity is lower. The primary issue here is more about completeness in conveying the article's intent rather than factual inconsistency.'

Moderate (3): noticeable error that may affect readability. Example moderate error explanation:

'Explanation: While this phrase is present in both the article and the summary, in the context of the summary, it lacks the preceding explanatory content that sets up the injustice being questioned. This omission makes the summary factually inconsistent by not providing the necessary background for the question's relevance.'

Major (4): serious error affecting meaning or clarity. Example major error explanation:

'Explanation: There is no information in the provided article that supports the claim about Indonesia's economic growth being its slowest pace since 2009. This additional, unsupported fact introduces a factual inconsistency.'

Critical (5): severe error that causes confusion or miscommunication. Example critical error explanation:

'Explanation: The summary introduces unrelated information not present in the article. There is no mention of children being involved in the accident or anyone suffering a broken wrist. This addition compromises factual consistency.'

Below you will find an error explanation of an error with severity level {severity}. Make it sound like a {severity_change_str} severe, {new_severity} severity error.

Listing 1: Text Style Transfer prompt to convert explanations.

There are five error severity levels:

Suggestion (1): optional improvement, not necessarily wrong. Example suggestion explanation:
 'Explanation: This statement is out of context in the summary. The original article mentions the follower count as additional information about their online presence, but in the summary, it appears as a standalone fact without explaining its relevance to the main topic (their travels). However, this is more of a contextual issue, but since the numbers are accurate, the severity for factual consistency is relatively low.'

Minor (2): small error that doesn't hinder understanding. Example minor error explanation:
 'Explanation: While not entirely inaccurate, this sentence lacks crucial contextual information present in the article (e.g., overcoming her father's death). However, since it doesn't introduce new inconsistent facts but rather omits them, its severity is lower. The primary issue here is more about completeness in conveying the article's intent rather than factual inconsistency.'

Moderate (3): noticeable error that may affect readability. Example moderate error explanation:
 'Explanation: While this phrase is present in both the article and the summary, in the context of the summary, it lacks the preceding explanatory content that sets up the injustice being questioned. This omission makes the summary factually inconsistent by not providing the necessary background for the question's relevance.'

Major (4): serious error affecting meaning or clarity. Example major error explanation:
 'Explanation: There is no information in the provided article that supports the claim about Indonesia's economic growth being its slowest pace since 2009. This additional, unsupported fact introduces a factual inconsistency.'

Critical (5): severe error that causes confusion or miscommunication. Example critical error explanation:
 'Explanation: The summary introduces unrelated information not present in the article. There is no mention of children being involved in the accident or anyone suffering a broken wrist. This addition compromises factual consistency.'

Now, you will be given a text span from the article which includes an error in the aspect of interest.
 The span with the error has been selected by an expert human annotator, and assigned a definite severity level.
 Your task is to generate an error explanation for this text span, which should be of the same severity level as the one assigned to it.

Do not add any additional text, comments, or severity mark. Provide only the explanation line.

Listing 2: Prompt to generate explanation for artificially introduced errors.

```

### Instructions
Your task is to evaluate a generated story. The model was instructed to write a
story based on a writing prompt.
Based on the given writing prompt and the generated story, identify errors in the
story with respect to {{ aspect_name }} (described below).
For each error, determine its severity on a scale from 1 to 5, where 1 are small
imperfections that only slightly affect the overall quality of the story, and 5
are critical issues that make the story unusable.
Never report more than 8 errors. If there are more, report only the 8 most
important ones.

Definition of {{ aspect_name|capitalize }}:
{{ aspect_definition }}

Rules:
Do not make assumptions and do not bring in external knowledge not present in the
provided context.
Identify only the errors related to the {{ aspect_name }} of the text. Do not
consider other aspects like {{ negative_aspect_examples|join(' or ') }}!
If there are no errors related to {{ aspect_name }} in the text, you should output
'No Error' and provide 'Excellent' score.

Steps:
1. Carefully read the writing prompt.
2. Read and try to understand the generated story. Then analyze it with
respect to coherence.
3. If the text contains any error that negatively affects its coherence,
identify its exact location (specific word or phrase), explain why it is
considered an error, and determine the severity of the error.
4. Finally, provide an overall score for the coherence of the text. The score
should be a label on the following scale (lowest to highest): 'Unacceptable',
'Poor', 'Fair', 'Good', 'Excellent'. The score 'Unacceptable' indicates that
the text is very incoherent, while 'Excellent' indicates that the text is
perfectly coherent.

### Writing Prompt
{{ input }}

### Story
{{ output }}

### Output format:
Generate your output exactly in this format:
```
Error 1:
Location: <location of the error - the exact word or phrase in the response>
Explanation: <explanation for the error, including the reason why it is considered
{{ aspect_name }} issue>
Severity: <integer from 1 to 5>

Error 2:
...

Overall score: <one of: Unacceptable, Poor, Fair, Good, Excellent>
Explanation of the score: <explanation of the score>
```

```

Listing 3: Annotator prompt template for the story generation (HANNA) task.

Input: "Two people promise their first born child to two different witches in return for a favour. These two people end up getting married and have a child together."

"Story": "**Prelude: ** Brujeria , they said, looking about the landscape as they said it. The one you seek is Brujeria. Travel past Three Rocks and you will find her hut. But speak to no one of your visit . Agueda was a young, but she knew what it meant if her mother were to die. Mother had been sick for quite a while, and the Ayuhuasca was n't working. The Urarina people were dependent upon their male shamans to delve into the spirit world, and seek out that which is causing the balance to break between the life of her mother and the illness that had struck her. But they could not make her well again. Agueda was desperate. At the tender age of 8, even she knew that if her mother had died, she would be alone in the village, and left to fend for herself. Agueda walked the forest, along the Chambira river, and spied the Three Rocks at the bend of the river. She went straight from there and saw a makeshift hut, with a fire burning with some fish being prepared. Agueda did n't know she was hungry until she saw the fish being cooked. She came up to the fire and plucked a fish from the stone that sat beside the flame. Without even thinking, she starting biting into the soft white meat of the fish, and savored every bite. She took another bite, and another. Before she knew it, the fish had almost been completely devoured. It tastes good, yes ? A voice said behind her. She jumped at the sound of the voice, dropping the fish in fear that she had been caught. She at once understood the fish was n't hers, and that she had done wrong in stealing it. However, an old woman now held the remains of the fish, having caught it before it hit ground. Examining it, she offered it back to Agueda. No use for me now. You take it, you finish it . Agueda nodded her head, and finished off what was left of the fish 's meat near the tail. The old woman eyed her, with an eye both sympathetic and yet still analyzing. It looked as if she had seen more than her age had let on, and her hair was so grey that it almost was white when the sun rays came through some of the trees to illuminate it. The sinews of her arm looked like a withered tree branch, but in them she felt there was a strength that was beyond that of the strongest of men in her village. This was the Brujeria. She spoke up to the strange woman, shyly given what she had just done. I 'm sorry elder. I was just so hungry The spirits told me , the Brujeria said, interrupting. And I was happy to oblige. It is a small thing for me to offer one so young, that has traveled so far from the village to seek one that should not be sought . As Agueda looked into her eyes deeper, she saw that there was a faint glow that almost whited her pupils. Was the Brujeria in a trance? Was she seeing both her and the spirit world, speaking to both at the same time? Agueda knew at once that the old woman knew what she had come to her for. Yes, I do . Replied the Brujeria. The old woman sat down on a log beside the fire, and began eating the other fish that was on the stone. Again, they told me. The Ayuhuasca does not work, they said. These are only plants that do not speak to the will of the the spirits that house them. Drink this . Agueda looked to the fire that the Brujeria had pointed to and saw a makeshift pot with a brew boiling. It was Ayuhuasca, but it smelled different than the ones that the village shamans prepared. This boiled with a certain glow to it, and it smelled sweeter. Was this an attempt to steal her energy? Her mother had always warned her about the dangerous Brujos that wanted nothing more than the life force of the ones that drank before them. I am not some fiendish Brujo as they say, child , said the old woman. The old ones know better, but I have been here because I need to be. I have need to study this book given by a man, and every eye in the village has sought to peek into its pages. Yet it is not for them, lest they summon curses to their souls that no healer can repair. So I stay here, for their sake, not just my own . Agueda looked at the book that the elder produced on her lap. The book was as tattered as the rags that she was wearing, but there was no mistaking that the air around this book brought a certain chill to her bones that the fire could hardly warm. Could this woman be the Warlock of Chiloe? Agueda once heard stories of the power of this"

Listing 4: Story generation example no. 7 from HANNA dataset.

```
Error 1:
Location: \"Agueda looked to the fire that the Brujeria had pointed to and saw a
makeshift pot with a brew boiling.\"
Explanation: This sentence lacks clarity and smooth transition. It jumps abruptly
from Agueda looking at the fire to focusing on a pot within the fire.
Severity: 3

Error 2:
Location: \"It was Ayuhuasca, but it smelled different than the ones that the
village shamans prepared.\"
Explanation: This sentence feels repetitive and clunky due to the phrasing \"the
ones that the village shamans prepared\". A more concise way to express this would
improve flow.
Severity: 2

Error 3:
Location: \"Was this an attempt to steal her energy?\"
Explanation: This question appears abruptly and without sufficient build-up. The
reader needs more context to understand why Agueda would think this.
Severity: 3

Error 4:
Location: \"I am not some fiendish Brujo as they say, child, said the old woman.\"
Explanation: The dialogue feels unnatural and overly explanatory. It tells rather
than shows the character's intentions.
Severity: 2

Overall score: Fair
Explanation: While the story generally follows a logical progression, several
instances of awkward phrasing and unclear transitions disrupt the smoothness of
the narrative. The dialogue also suffers from being somewhat clunky and expository.
```

Listing 5: Annotator evaluation for the example in Listing 4 generated using Gemma under the aspect of coherence.

Error 1:
Location: \"Agueda looked to the fire that the Brujeria had pointed to and saw a makeshift pot with a brew boiling.\"
Explanation: This sentence introduces a factually inconsistent detail by stating Agueda is focused on a pot within the fire, while the original text describes her looking at the fire itself. This alteration misrepresents the scene's context and creates a false narrative.
Severity: 4

Error 2:
Location: \"It was Ayuhuasca, but it smelled different than the ones that the village shamans prepared.\"
Explanation: This sentence inaccurately implies that the village shamans prepared the medicine, while the article states they were simply the ones who administered it. This omission creates a factual inconsistency by failing to distinguish between preparation and administration.
Severity: 3

Error 3:
Location: \"Was this an attempt to steal her energy?\"
Explanation: This question presents a fabricated scenario not supported by the article. The article does not suggest Agueda questioning anything related to this topic. This introduction of a baseless situation constitutes a factual inconsistency.
Severity: 4

Error 4:
Location: \"I am not some fiendish Brujo as they say, child, said the old woman.\"
Explanation: The dialogue lacks nuance and fails to organically convey the character's motivations, making the scene feel contrived and unconvincing.
Severity: 3

Overall score: Poor
Explanation: While the story attempts to follow the plot points of the prompt, several inconsistencies and illogical developments hinder its coherence. The presence of factual inaccuracies and forced scenarios significantly detract from the overall flow and believability of the narrative.

Listing 6: Evaluation in Listing 5 with errors modified to sound more severe by one severity category, and the overall score regenerated.

Error 1:
Location: \"Agueda looked to the fire that the Brujeria had pointed to and saw a makeshift pot with a brew boiling.\
Explanation: This sentence fabricates a detail not present in the source material, falsely claiming Agueda was focused on a pot within the fire when she was actually looking at the fire itself.
Severity: 5

Error 2:
Location: \"It was Ayuhuasca, but it smelled different than the ones that the village shamans prepared.\
Explanation: This sentence introduces a fabricated detail not found in the source material, claiming that the village shamans prepared something. This fabrication adds a layer of misinformation and undermines the factual accuracy of the summary.
Severity: 4

Error 3:
Location: \"Was this an attempt to steal her energy?\
Explanation: The summary fabricates a scenario where Agueda questions something without proper context, creating a misleading narrative.
Severity: 5

Error 4:
Location: \"I am not some fiendish Brujo as they say, child, said the old woman.\
Explanation: The dialogue fabricates interactions and motivations, introducing inconsistencies with the source material's portrayal of the characters.
Severity: 4

Overall score: Poor
Explanation: While the story attempts to follow the general plot points of the prompt, it suffers from significant inaccuracies and fabricated details, leading to a lack of coherence and faithfulness to the original source.

Listing 7: Evaluation in Listing 5 with errors modified to sound more severe by two severity categories, and the overall score regenerated.

```

Error 1:
Location: \"Agueda looked to the fire that the Brujeria had pointed to and saw a makeshift pot with a brew boiling.\"
Explanation: This sentence lacks clarity and smooth transition. It jumps abruptly from Agueda looking at the fire to focusing on a pot within the fire.
Severity: 3

Error 2:
Location: \"It was Ayuhuasca, but it smelled different than the ones that the village shamans prepared.\"
Explanation: This sentence feels repetitive and clunky due to the phrasing \"the ones that the village shamans prepared\". A more concise way to express this would improve flow.
Severity: 2

Error 3:
Location: \"Was this an attempt to steal her energy?\"
Explanation: This question appears abruptly and without sufficient build-up. The reader needs more context to understand why Agueda would think this.
Severity: 3

Error 4:
Location: \"I am not some fiendish Brujo as they say, child, said the old woman.\"
Explanation: The dialogue feels unnatural and overly explanatory. It tells rather than shows the character's intentions.
Severity: 2

Error 5:
Location: *full text*
Explanation: This error completely compromises the quality of this text on the selected aspect.
Severity: 5
Overall score: Poor

Explanation: While the story attempts to establish a narrative, it suffers from significant issues with coherence. The lack of clear transitions, abrupt shifts in focus, and confusing dialogue disrupt the flow and make it difficult to follow the plot.

```

Listing 8: Evaluation in Listing 5 with one critical error added at the end, and the overall score regenerated.

```

"inputs": {
  "Writing Prompt": "When the europeans discovered America they did n't colonize
  it. Instead they have been monitoring the Americans for centuries. Today, the
  first American ship reaches Europe."
},
"outputs": {
  "Story": "In the morn on the twenty-eighth day in the month of october in the
  year 1840 marks the date first ever an expedition from the blighted continent
  found it 's way into an English harbour. The American ships, as expected, were
  backward, but not as much so as one might expect. While the ships they boarded
  were no greater than those from the turn of the seventeenth century, they were
  marvelous in their craftsmanship: Long bows carved with ornate woodcraft ( their
  gods perhaps ), canon of unknown quality, large billowing sails, clean and
  professional rigging and the men aboard -- truly these men had sailed and had
  been sailing for a time. They were as the queen 's own. It gnawed upon me that
  when last sighted centuries earlier, many if not all the continent 's people 's
  were confined to freshwater craft; merely pilots of canoes and small rafts --
  Primitives. Indeed, in time, they had come to much the same conclusions that
  naval engineers past had come, no doubt guided by the smugglers ... or perhaps
  it was the many burnt hulls of the galleons left upon their shores by the
  Spanish or, and more unsettling, perhaps it was a hastily deconstructed remnant
  of our own navy that underlay the proud woodwork and bid them their voyage here.
  Spying them with a scope from the pier I spied the visage of one man, A stern
  corsair upon the prow. He was stone faced and dark haired bearing no weapon with
  one foot on the bow and the other dangling over the edge, a lit stick of a kind
  in his mouth. He seemed stoical and unconcerned -- no doubt this was not the
  first port he had seen on his voyage in. Perhaps they had been to France and
  found no welcome? At any rate, as I scanned the deck I saw more than one other,
  and perhaps to their nautical detriment more than one woman aboard as well --
  Though that perhaps spoke of other intentions. The admiralty has long dreaded
  this moment. It was, of course, inevitable. They would be curious as to why the
  world 's great colonial powers had stopped at their door, why ships long circled
  and barricaded their fledgling attempts at transport across the sea -- and more
  important why the lord 's anointed kings and our fair Queen held a line across
  an ocean for so long and then stopped.
  I felt ardour to call out a warning. I wished dearly to tell them that the few
  of ours who made it back from their sojourn died so miserably from having
  interacted with them. I would explain to them and tell them of the upheaval
  Europe and Asia faced, the death toll they brought upon us, la plaga nuevo . I
  would plead with them to leave. I would have if only I could but how does one
  recount this to them in their unknown tongue? How, in their language, can i say
  that they, without a single shot, had brought the once great majesty of our
  empire to it 's knees by that which is unaffactive of them? How? We were left
  with but one way: The only way. Our answer could only be delivered at the end of
  a gun. As the ship broached towards land I made a signal to the man at my right
  and he bid the gunners up upon the ramparts to take aim and fire above the prow.
  At my order they fired, but the response was as suspected. They did not go back.
  I saw them, faces alive. It was not fear amongst them though, but jubilation.
  The men and women broke out into shouts and waved their hands in the air ... and
  here I noticed the small ones racing up to their sides. There was no doubt in my
  mind, the cannon were inactive. We were the first Europeans to see them, and
  conversely, they to see us ( and I ask again what of France? ). We fired a
  second warning. This one seemed only to confuse them, as they did stop their
  jubilation, but did not away. I wrestled with the orders given. They could not
  be allowed to enter the harbour, but they were non-combatants, women, children,
  and lost sailors as eager to see our face as we were to see theirs -- but
  ignorant of the horror they inspired in us and unable to see our recoil. As it
  came within the boundary, and continuing even after a third shot fired we had no
  choice. The admiralty had made the order: The plague shall not again reach
  England. I ordered the full regiment to fire and not to stop until she was down.
  The ship and rendered it 's carved facade to splinters before it 's cannon could
  respond, or perhaps it did not matter. These wiley corsairs were no more then
  family men with children with no training or regiment to the skill. All were
  dead by the midday, drown or shot. I wonder though as to their origin and
  purpose. Why would they bring their families aboard? Did it not weigh upon them
  there might be danger? Thereupon I mired for the remaining"
}

```

Listing 9: Story generation example no. 16 from HANNA dataset.


```

Error 1:
Location: \"They were as the queen 's own.\"
Explanation: This sentence lacks clarity and logical connection to the preceding
sentences. It is unclear what is meant by \"as the queen's own\" and how it
relates to the description of the ships and their crew.
Severity: 3

Error 2:
Location: \"I felt ardour to call out a warning. I wished dearly to tell them...\"
Explanation: The sudden shift in perspective and tone feels jarring. The
narrator's desire to warn the Americans seems out of place considering the
previous descriptions of their advanced technology and potential threat.
Severity: 4

Error 3:
Location: \"How, in their language, can i say that they, without a single shot,
had brought the once great majesty of our empire to it 's knees by that which is
unaffective of them?\"
Explanation: This sentence is convoluted and grammatically incorrect. The meaning
is unclear and the phrasing is awkward.
Severity: 4

Error 4:
Location: \"There was no doubt in my mind, the cannon were inactive.\"
Explanation: This statement contradicts the previous descriptions of the ships'
advanced technology and the narrator's concerns about their potential threat.
Severity: 3

Error 5:
Location: \"We were the first Europeans to see them, and conversely, they to see
us ( and I ask again what of France? ).\"
Explanation: The abrupt introduction of France and the question mark feel
misplaced and lack context.
Severity: 2

Overall score: Poor
Explanation: While the story presents some interesting ideas, the numerous
inconsistencies and illogical connections between events significantly detract
from the overall coherence. The narrative jumps between perspectives and ideas
without clear transitions, making it difficult to follow the storyline.\"

```

Listing 10: Annotator evaluation for the example in Listing 9 generated using Gemma under the aspect of coherence.

```

Error 1:
Location: \"They were as the queen 's own.\"
Explanation: This sentence could benefit from clearer phrasing and stronger
connections to the surrounding context. The meaning of \"as the queen's own\"
regarding the ships and crew requires further clarification.
Severity: 2

Error 2:
Location: \"I felt ardour to call out a warning. I wished dearly to tell them...\"
Explanation: The transition between describing the alien civilization's
technological advancements and their perceived threat towards humans could be
smoother. The narrator's warning about the American response might feel slightly
abrupt given the prior context.
Severity: 3

Error 3:
Location: \"How, in their language, can i say that they, without a single shot,
had brought the once great majesty of our empire to it 's knees by that which is
unaffective of them?\"
Explanation: This sentence contains grammatical errors and awkward phrasing,
making its meaning unclear.
Severity: 3

Error 4:
Location: \"There was no doubt in my mind, the cannon were inactive.\"
Explanation: This statement presents a slightly different perspective on the
ships' capabilities compared to earlier descriptions, potentially lacking some
context regarding the narrator's concerns.
Severity: 2

Error 5:Location: \"We were the first Europeans to see them, and conversely, they
to see us ( and I ask again what of France? ).\"
Explanation: The mention of France feels slightly out of place and could benefit
from some additional context.
Severity: 1

Overall score: Fair
Explanation: While the story exhibits some inconsistencies and moments of unclear
transitions, it generally maintains a coherent narrative flow. The errors
primarily involve clarity and precision in expressing ideas rather than major
disruptions to the overall storyline.

```

Listing 11: Evaluation in Listing 10 with errors modified to sound less severe by one severity category, and the overall score regenerated.

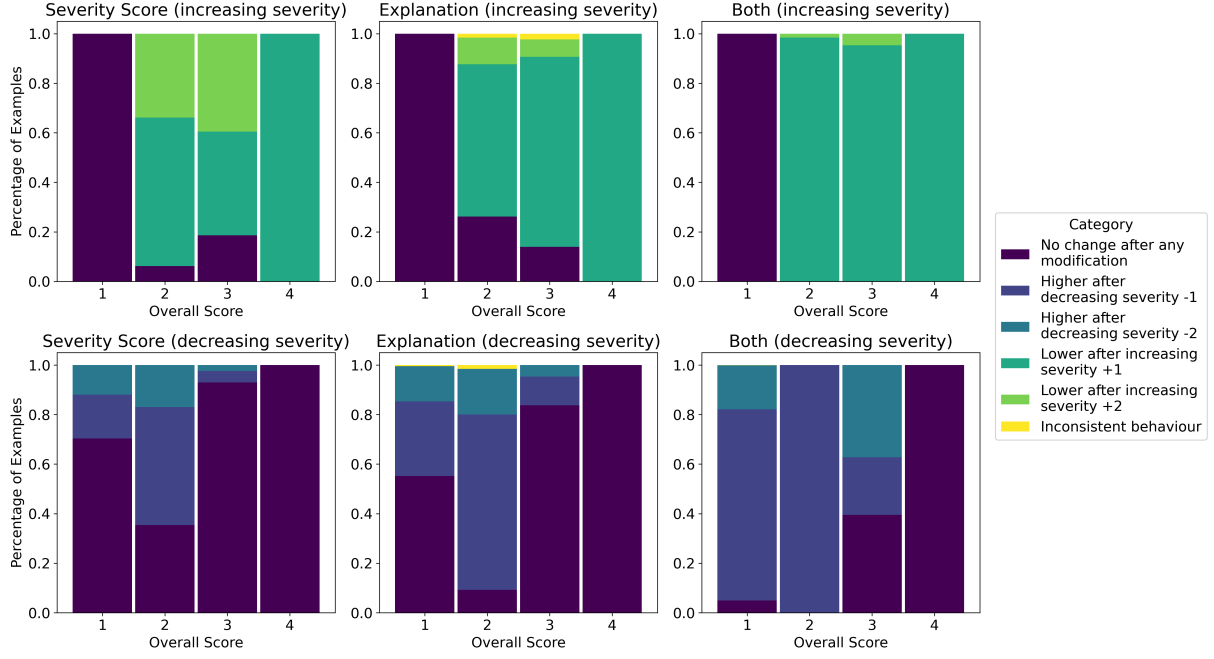


Figure 5: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

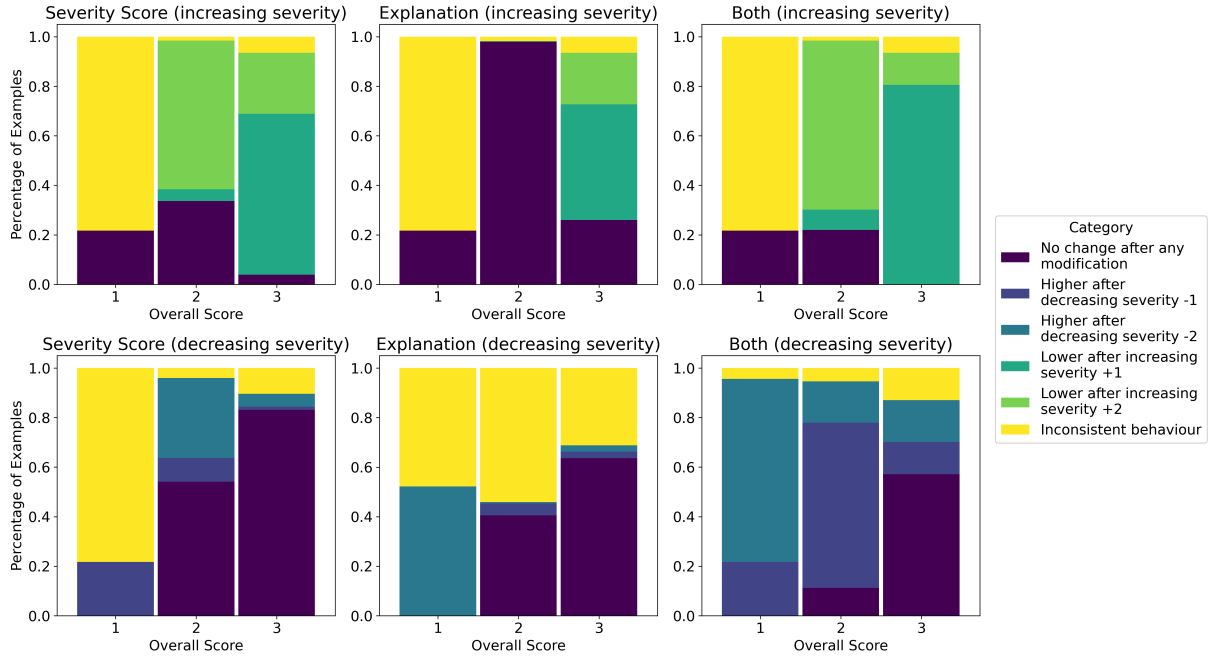


Figure 6: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating coherence on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

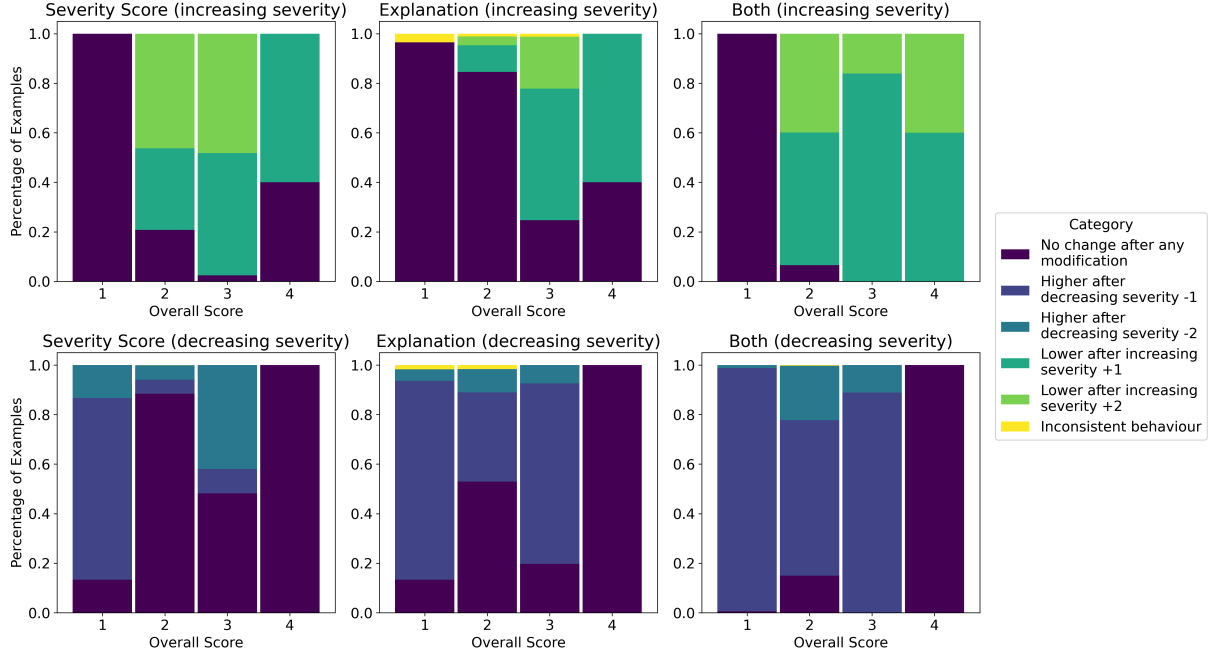


Figure 7: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating coherence on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

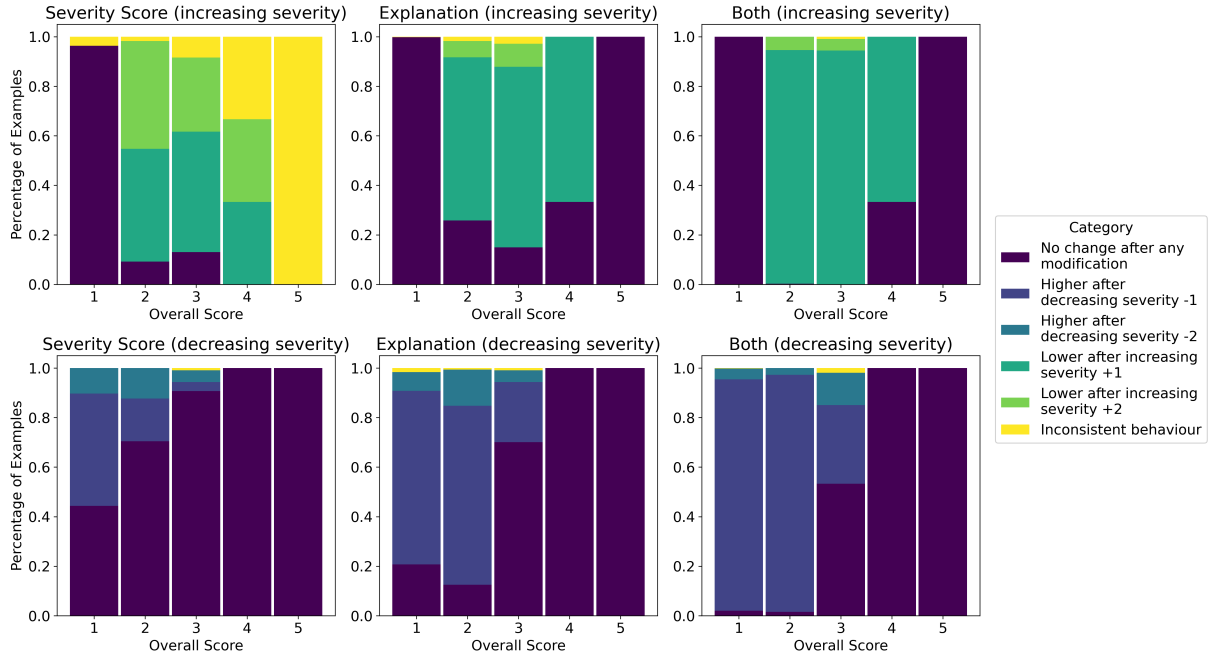


Figure 8: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating complexity on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

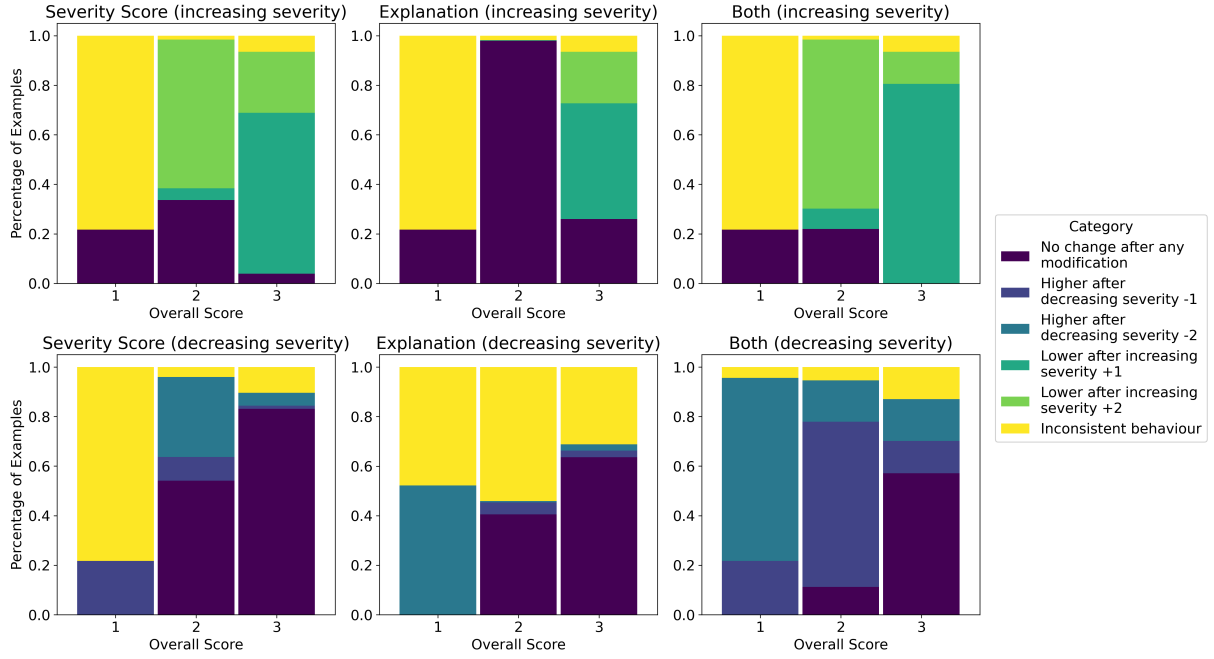


Figure 9: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating complexity on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

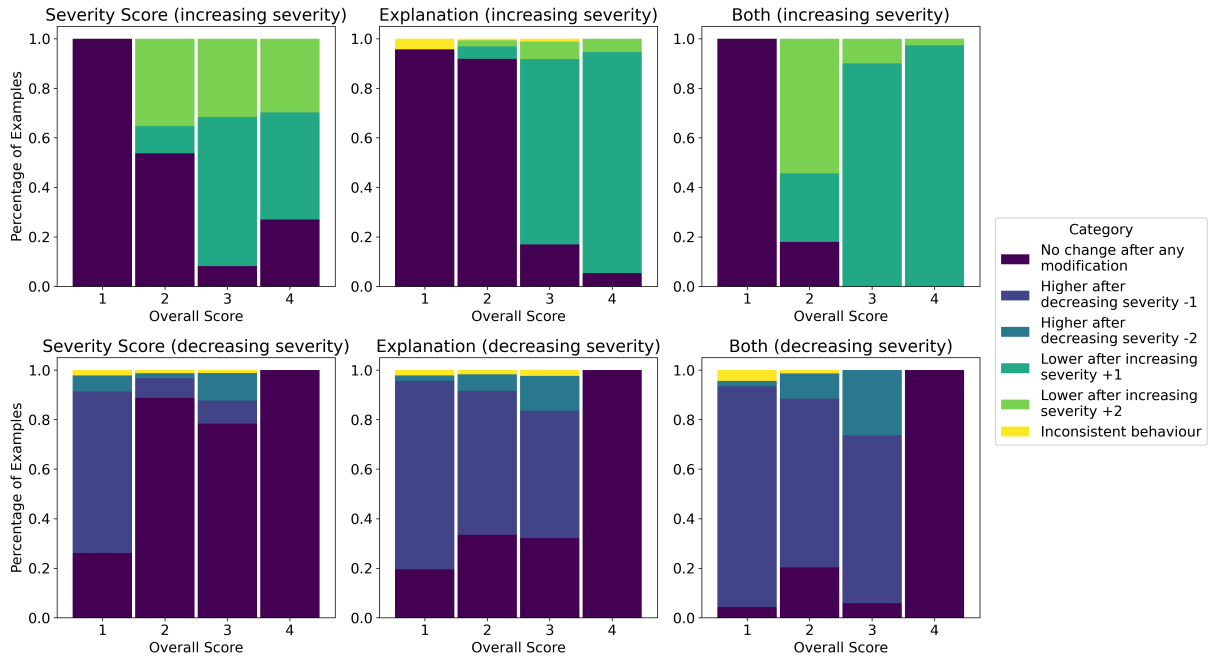


Figure 10: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating complexity on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

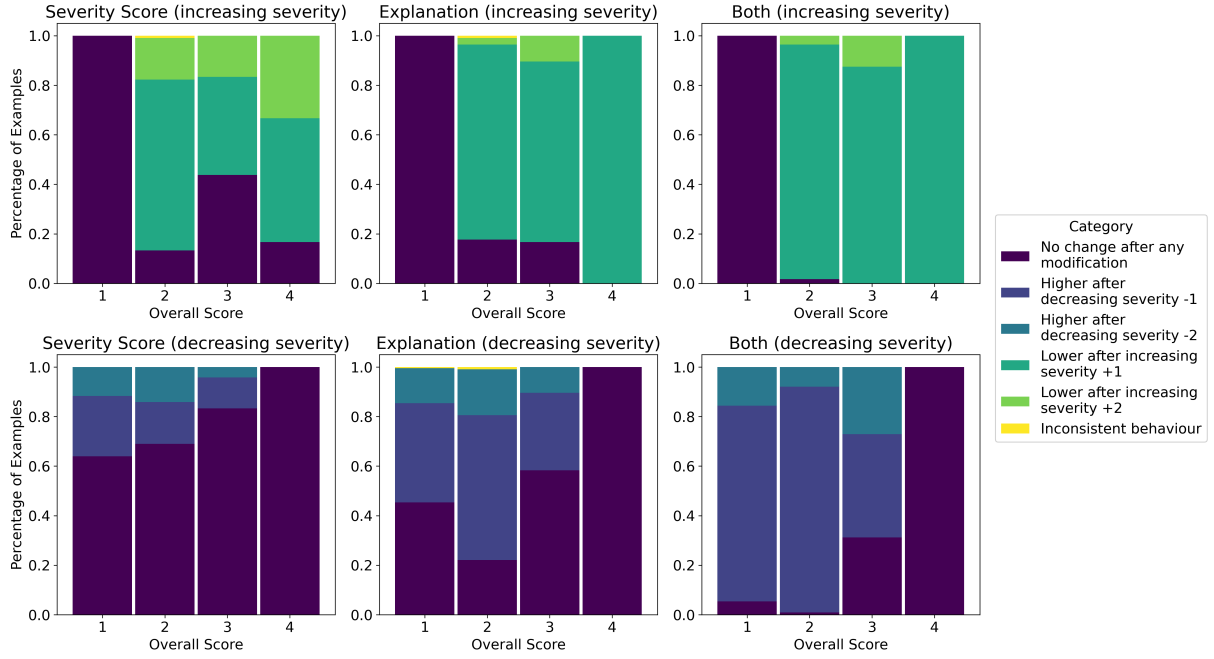


Figure 11: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

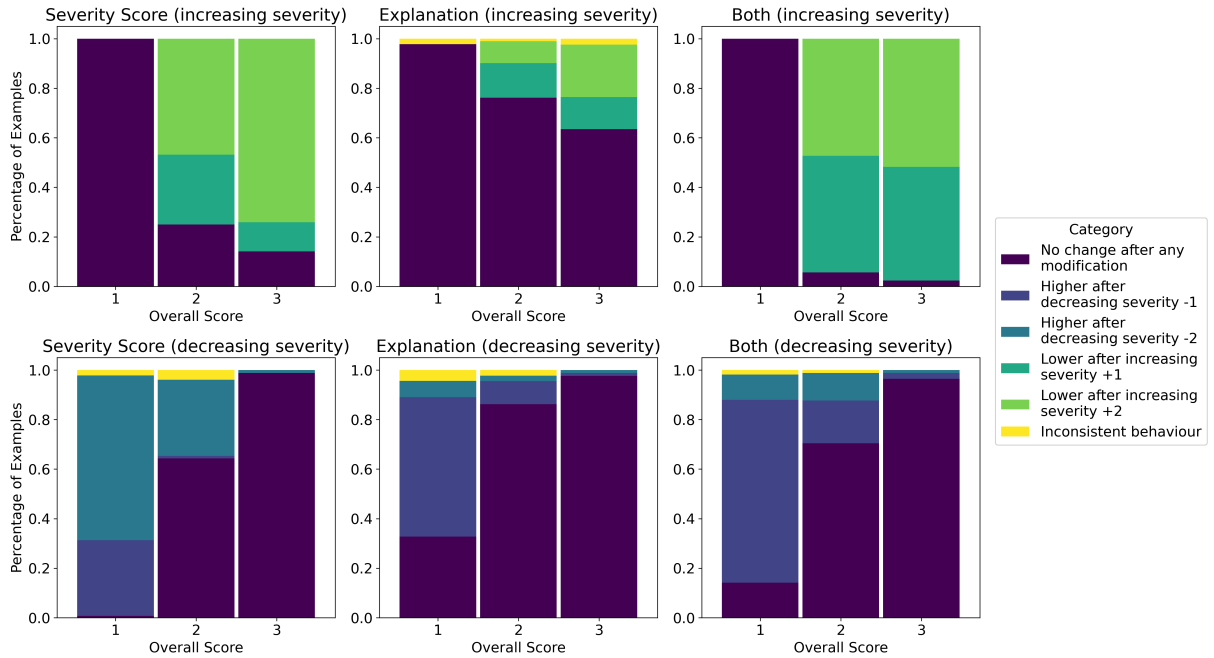


Figure 12: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating relevance on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

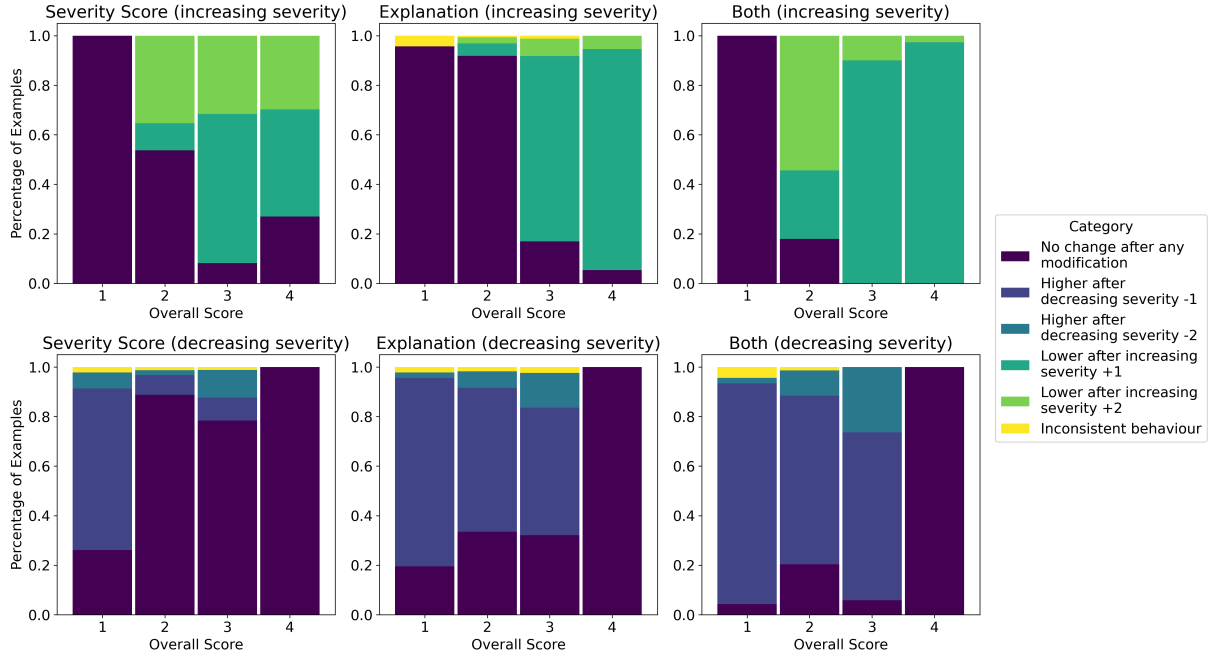


Figure 13: The percentage of examples which overall score changed after perturbing the severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating relevance on HANNA. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

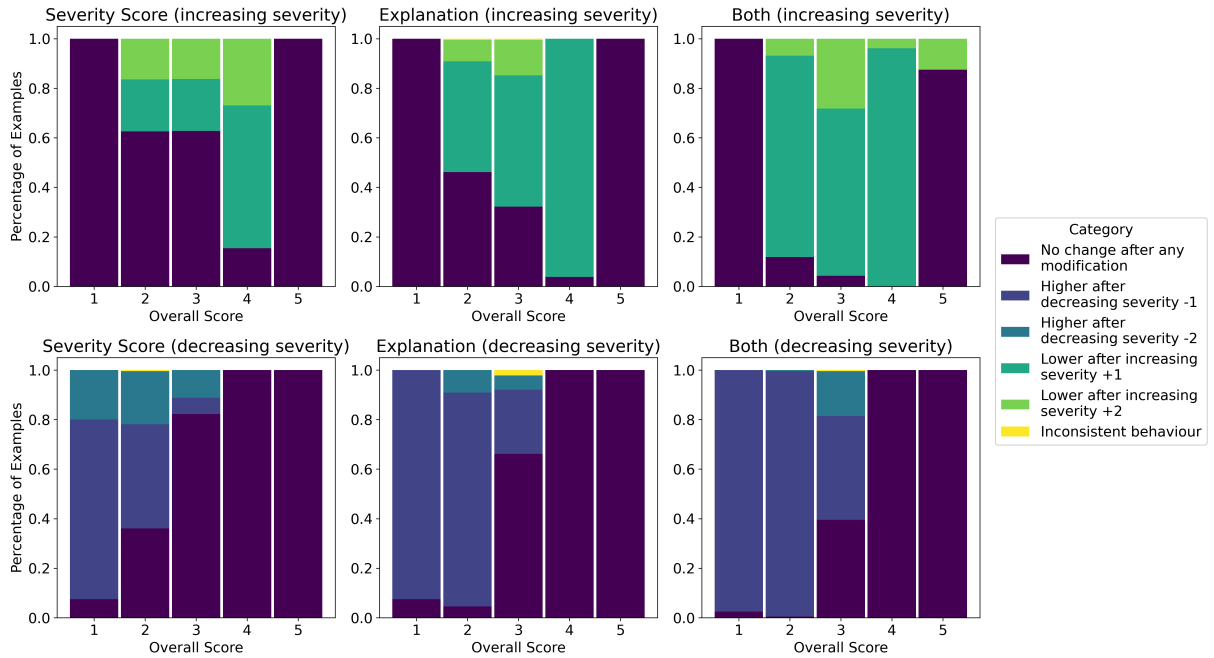


Figure 14: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

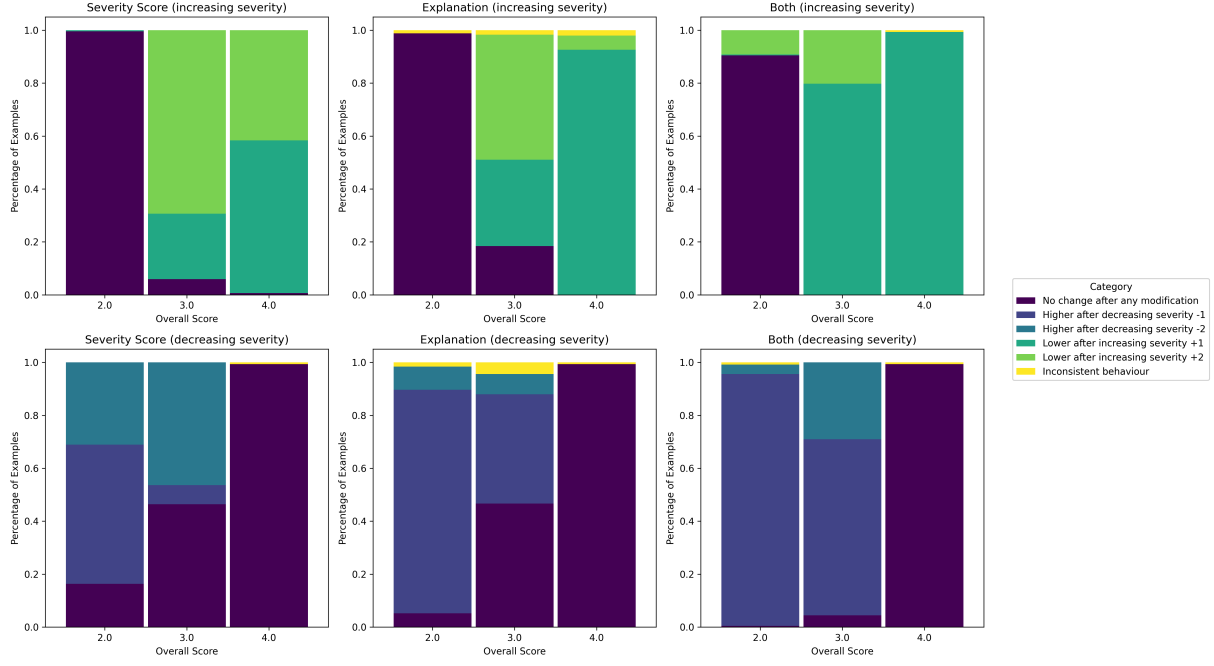


Figure 15: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating coherence on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

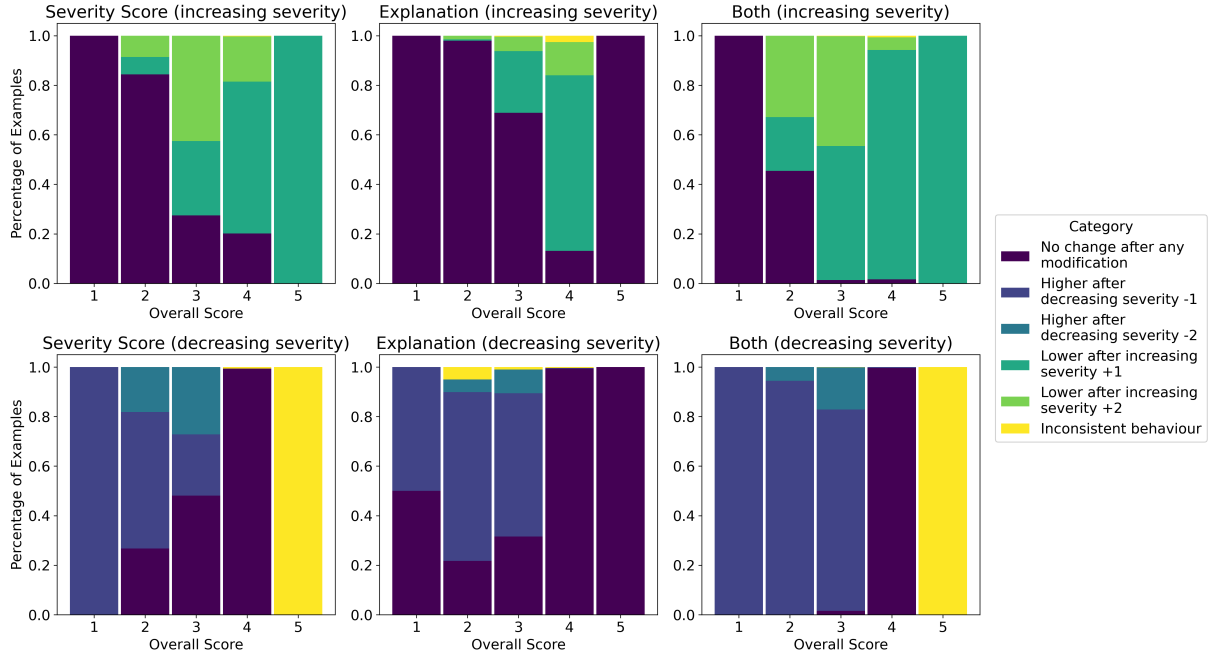


Figure 16: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating coherence on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

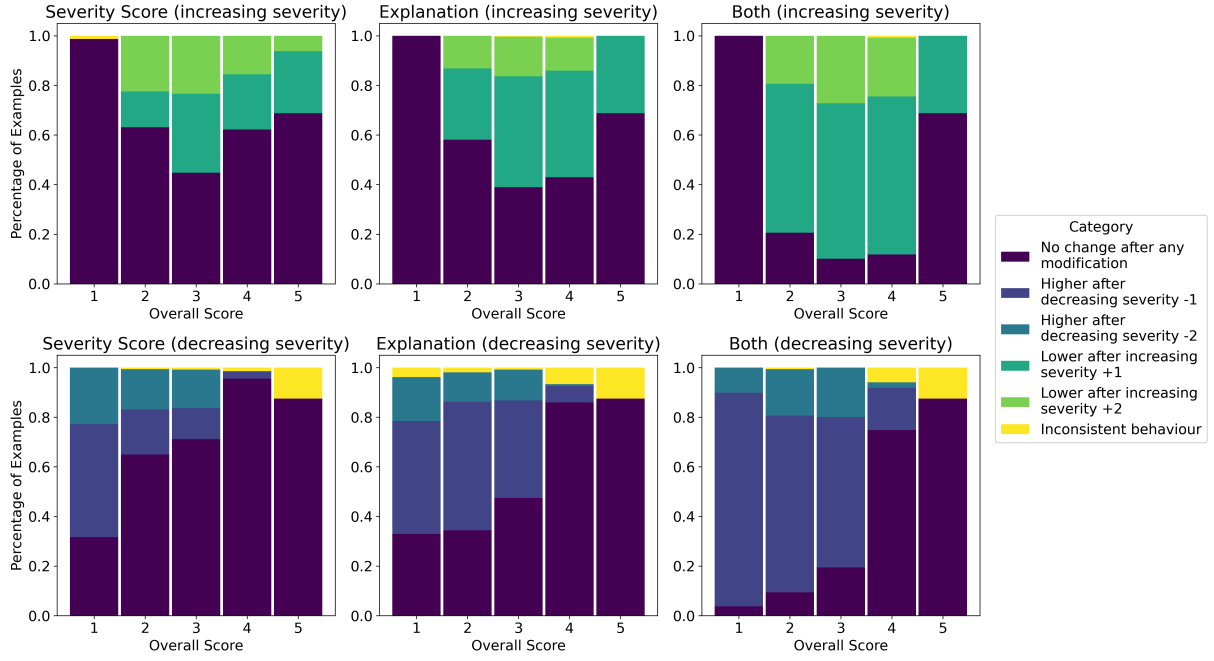


Figure 17: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

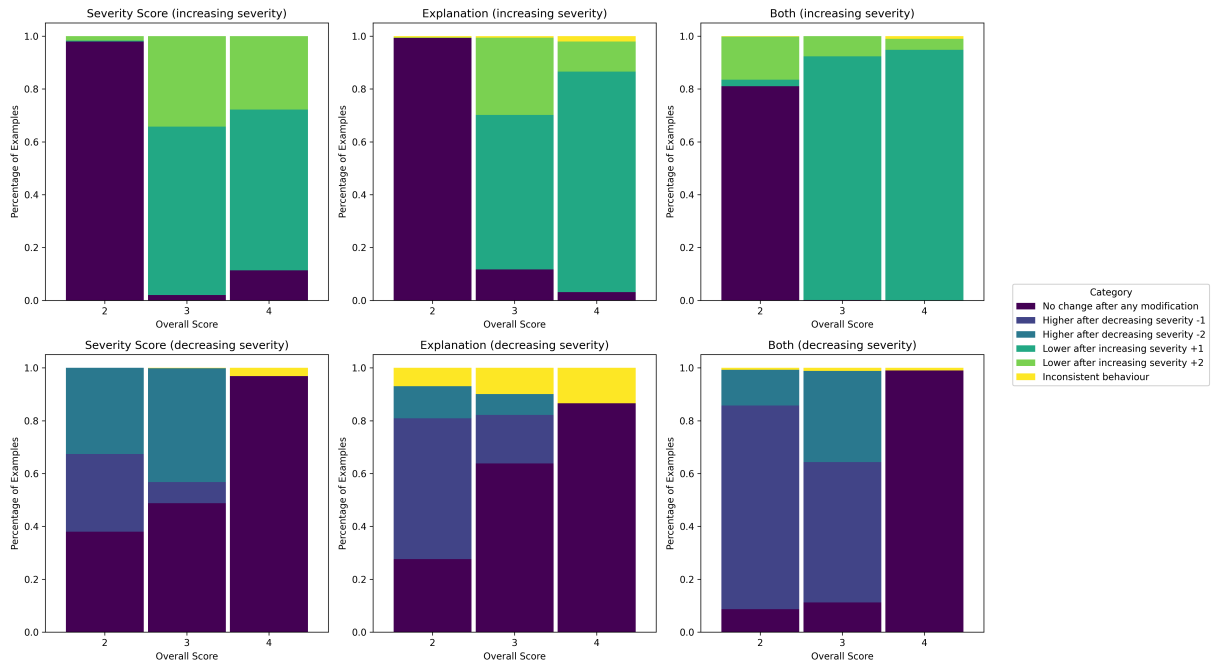


Figure 18: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating factual consistency on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

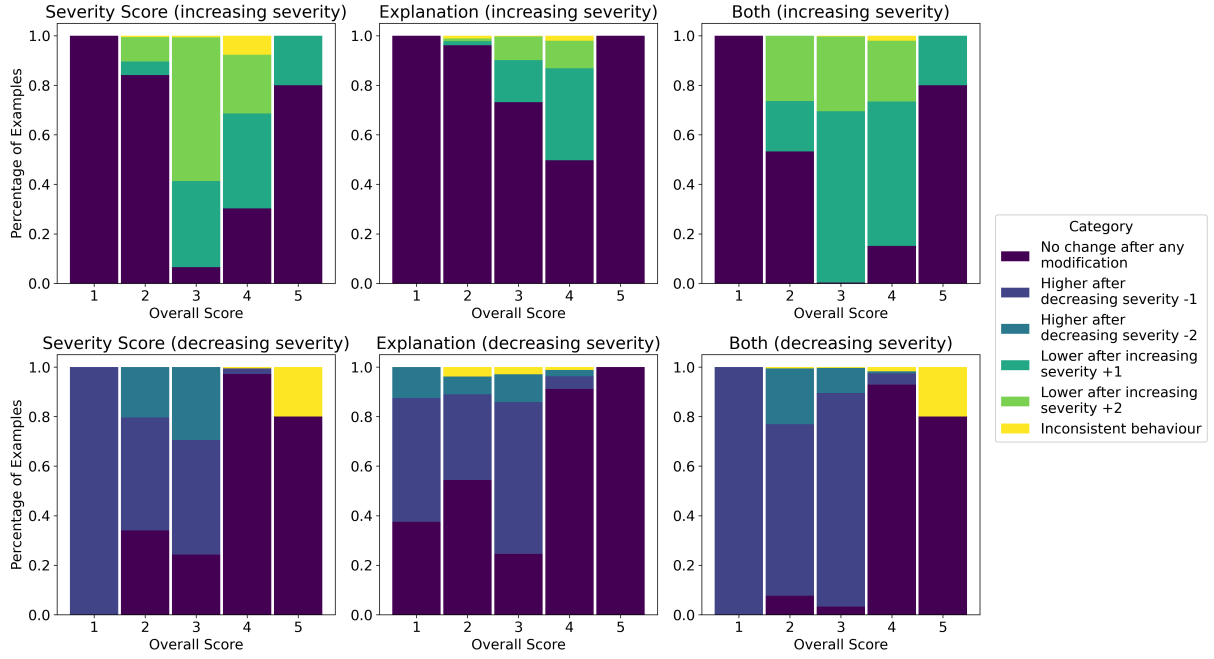


Figure 19: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating factual consistency on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

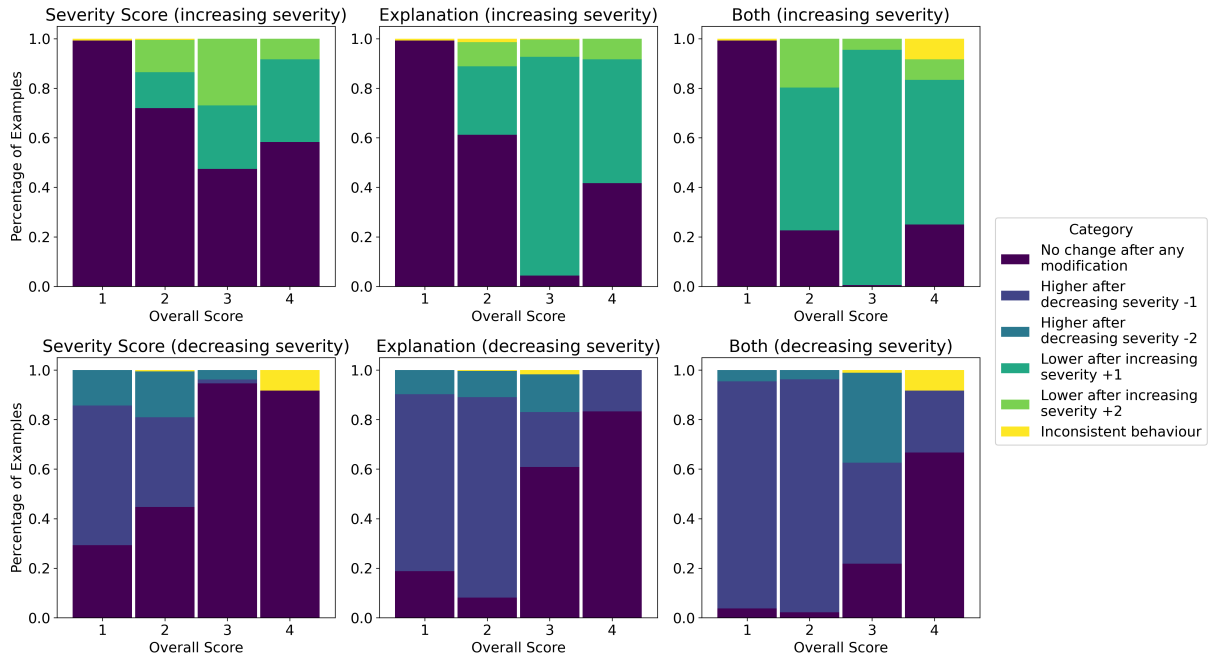


Figure 20: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

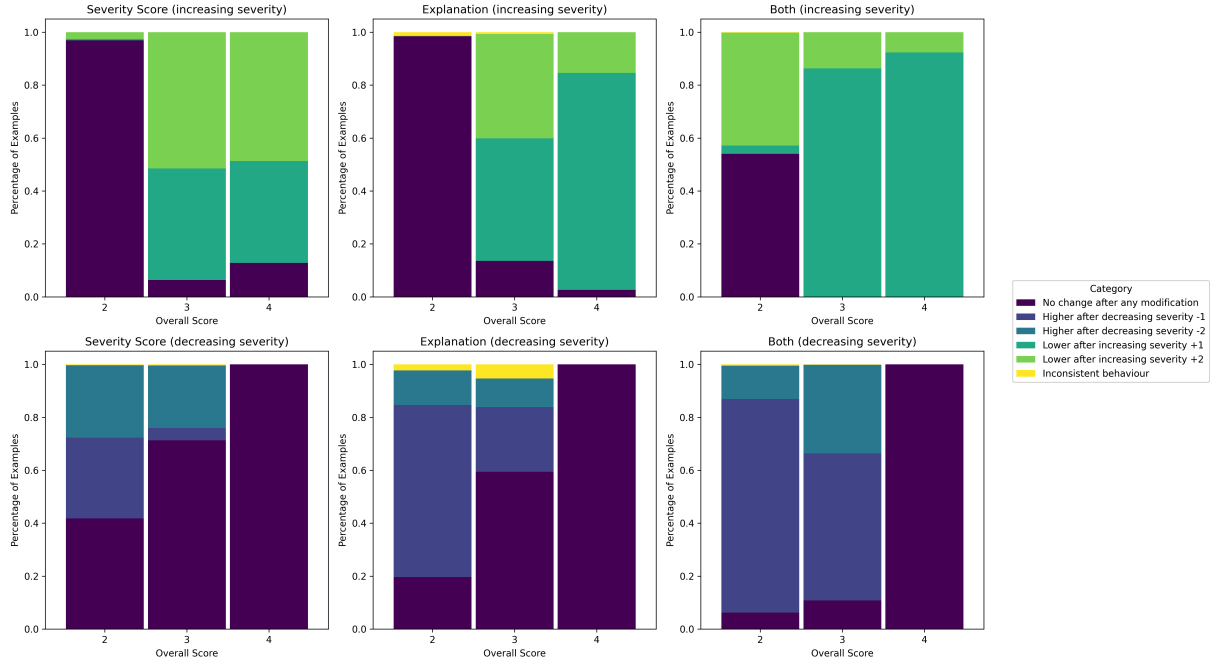


Figure 21: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating relevance on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

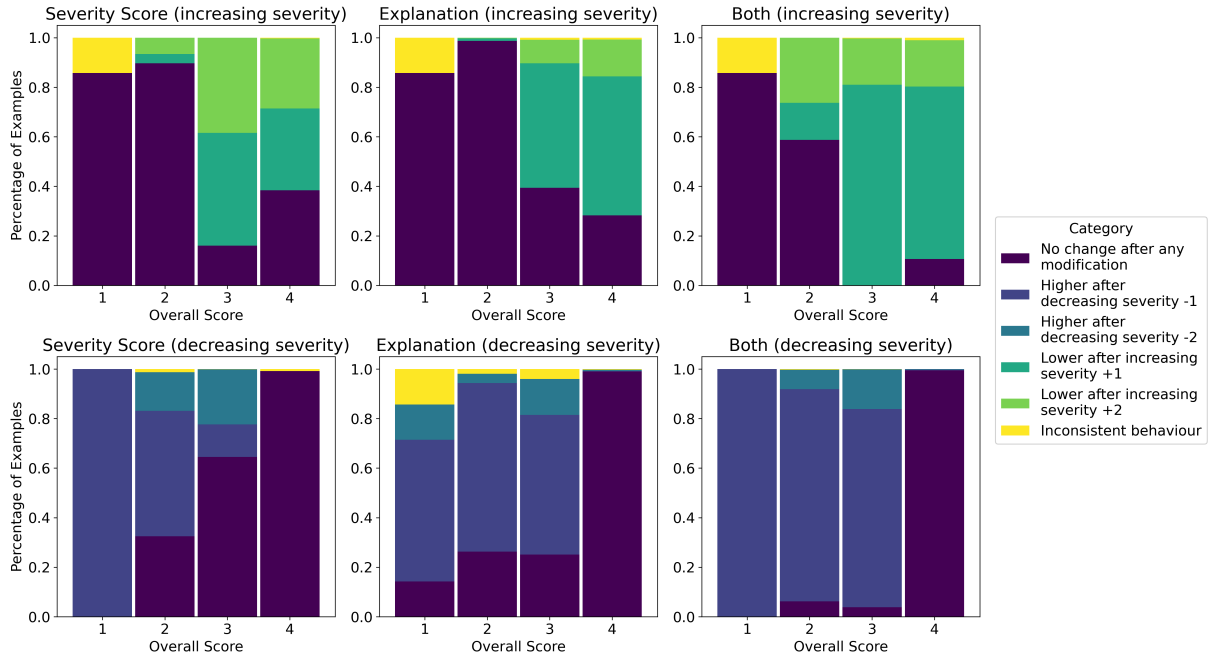


Figure 22: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating relevance on SummEval. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

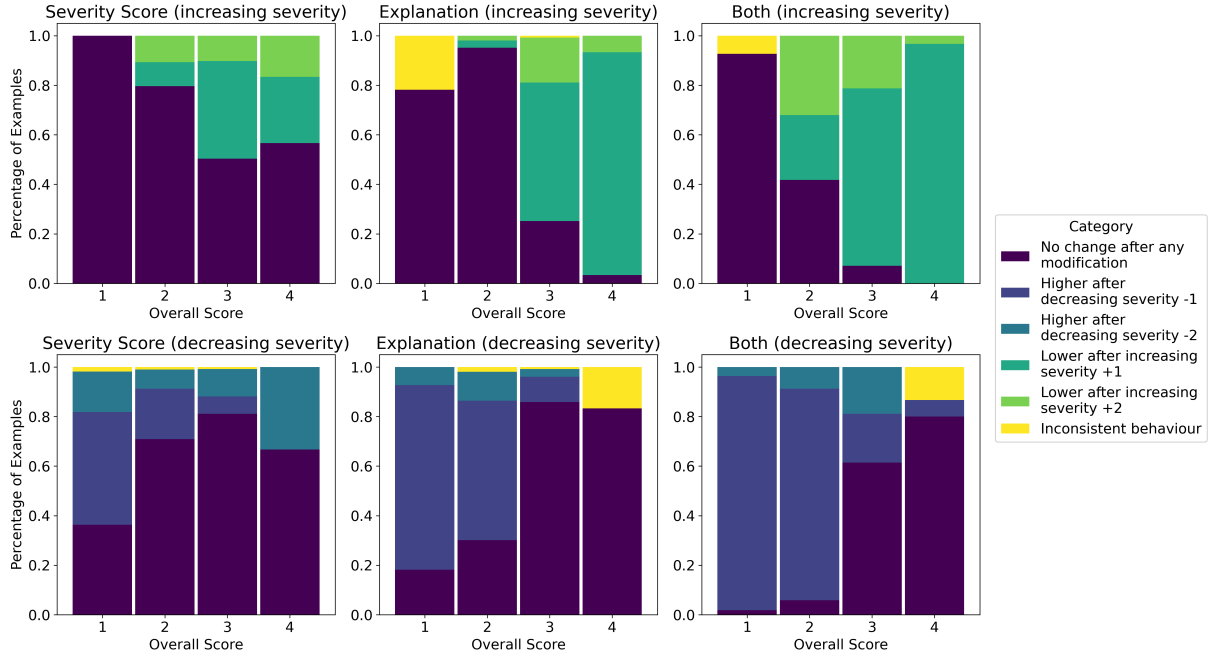


Figure 23: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on QAGS. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

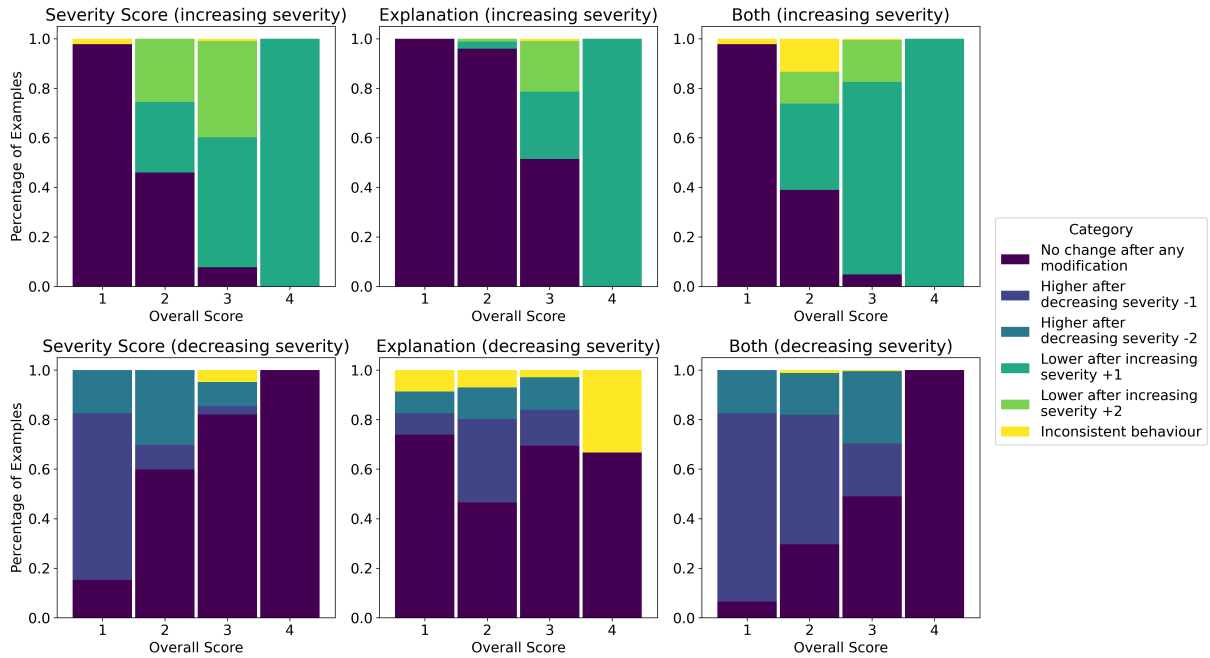


Figure 24: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Gemma while evaluating factual consistency on QAGS. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

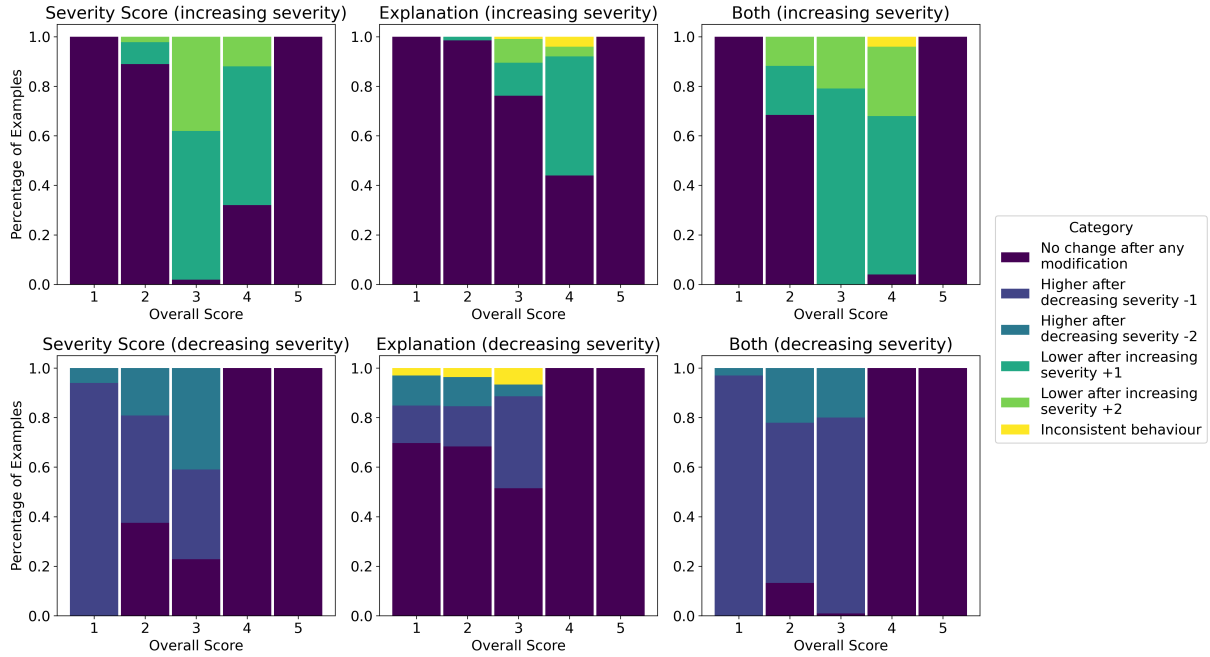


Figure 25: The percentage of examples which overall score changed after perturbing severity assessment (left), textual explanation (middle) or both (right) by increasing (top) or decreasing (bottom) error severity. The results are presented for Qwen 2.5 while evaluating factual consistency on QAGS. The 'Inconsistent behaviour' category measures situations in which changing the severity of one causes a change in the score, whereas changing the severity of two does not.

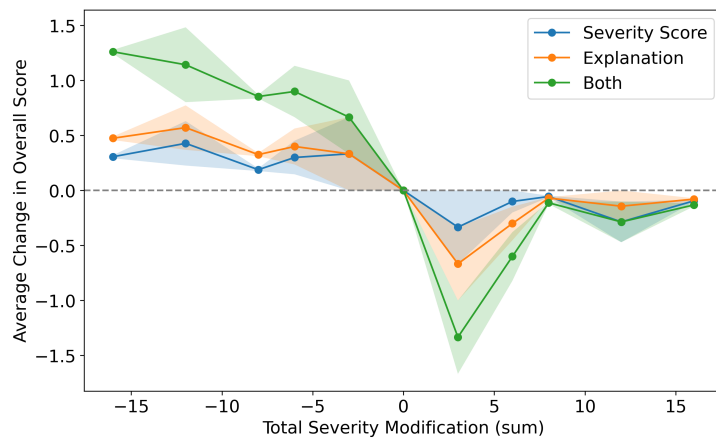


Figure 26: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on HANNA. Shade around lines shows standard deviation.

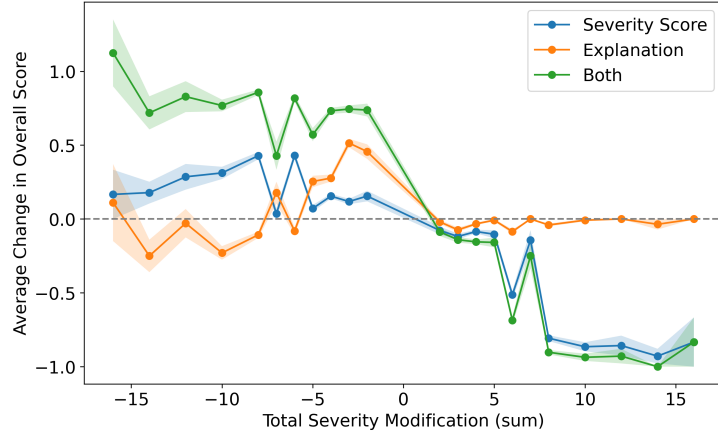


Figure 27: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating coherence on HANNA. Shade around lines shows standard deviation.

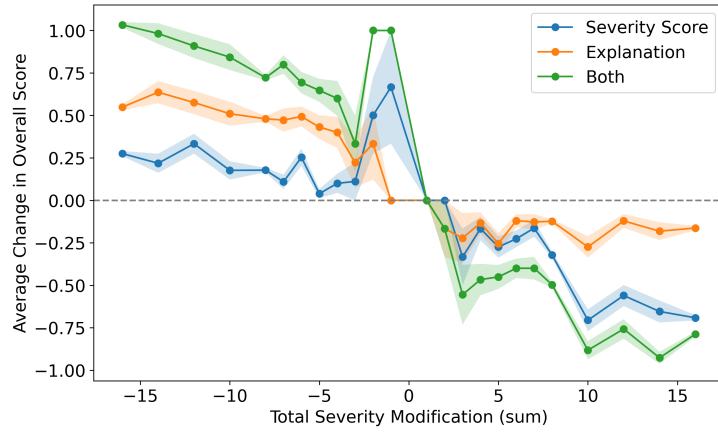


Figure 28: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating coherence on HANNA. Shade around lines shows standard deviation.

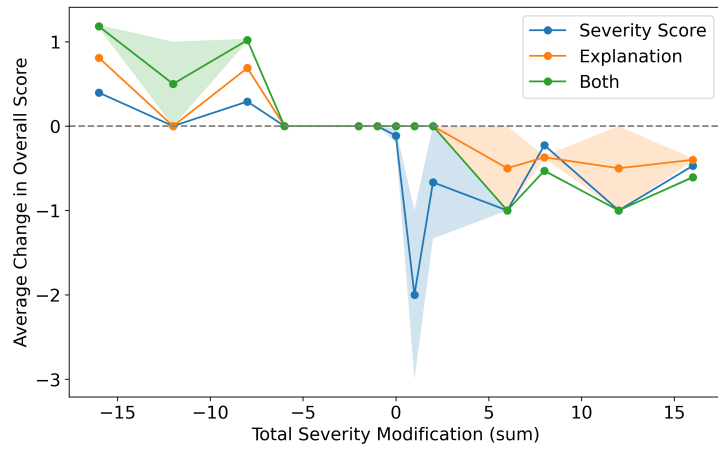


Figure 29: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating complexity on HANNA. Shade around lines shows standard deviation.

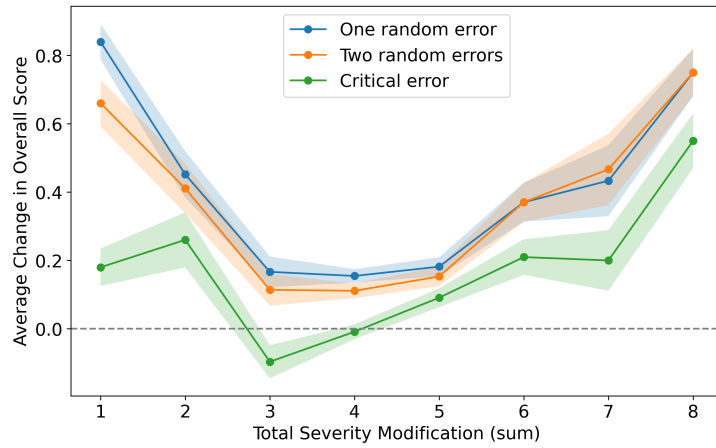


Figure 30: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating complexity on HANNA. Shade around lines shows standard deviation.

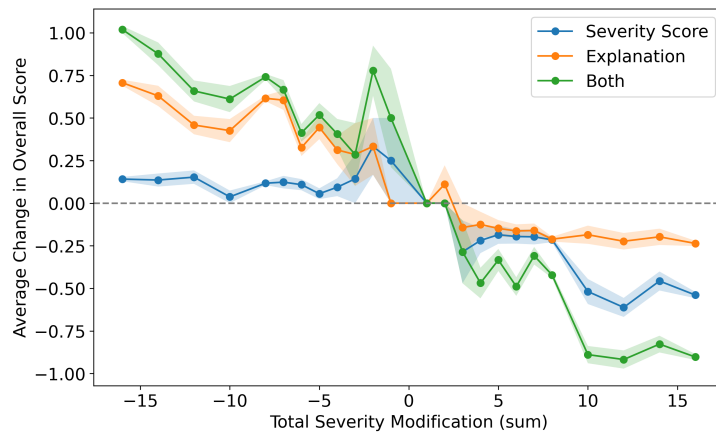


Figure 31: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating complexity on HANNA. Shade around lines shows standard deviation.

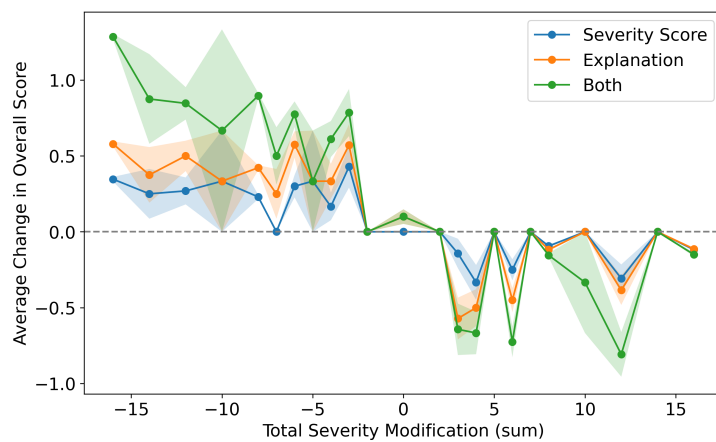


Figure 32: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on HANNA. Shade around lines shows standard deviation.

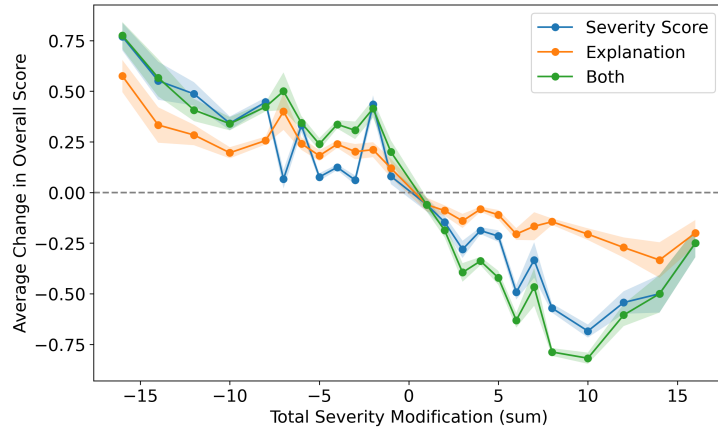


Figure 33: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating relevance on HANNA. Shade around lines shows standard deviation.

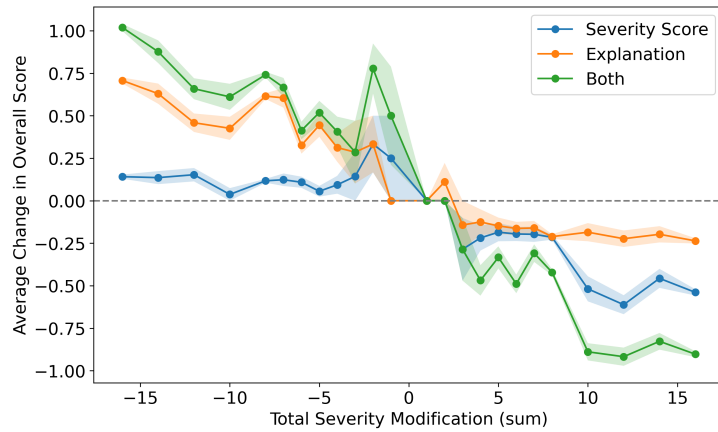


Figure 34: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating relevance on HANNA. Shade around lines shows standard deviation.

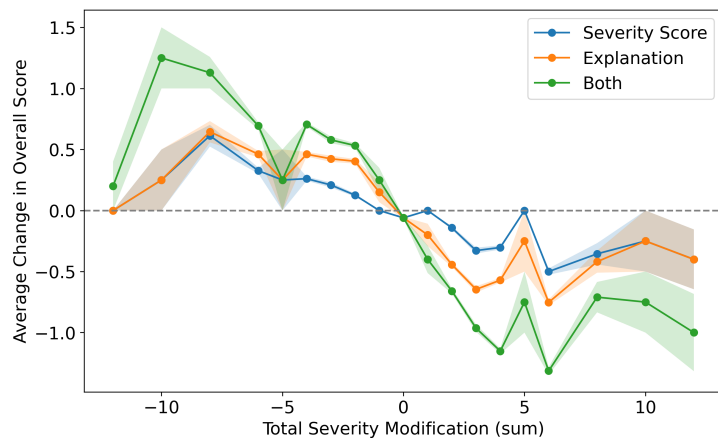


Figure 35: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on SummEval. Shade around lines shows standard deviation.

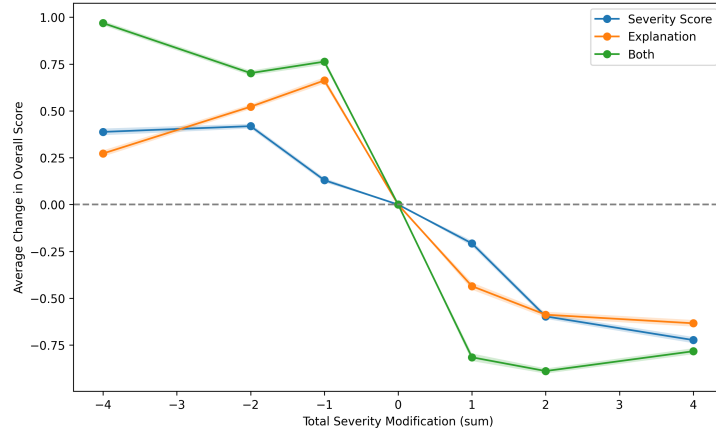


Figure 36: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating coherence on SummEval. Shade around lines shows standard deviation.

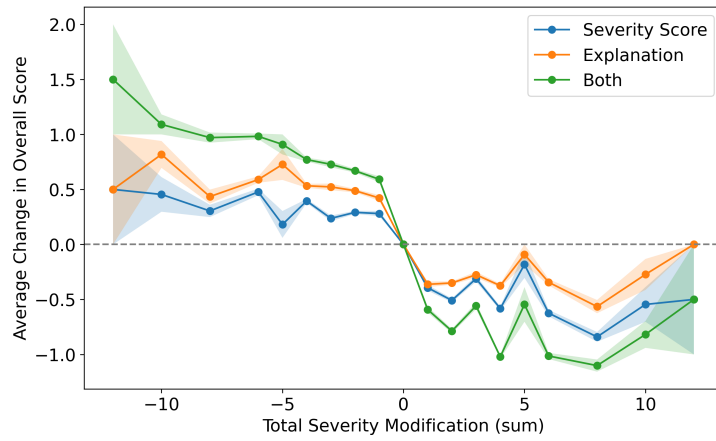


Figure 37: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating coherence on SummEval. Shade around lines shows standard deviation.

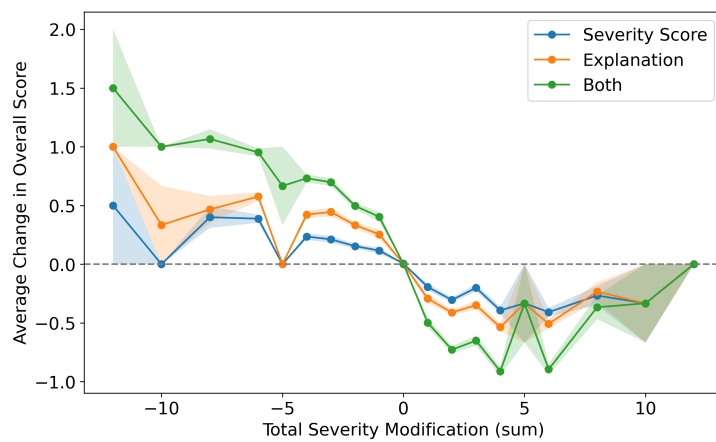


Figure 38: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on SummEval. Shade around lines shows standard deviation.

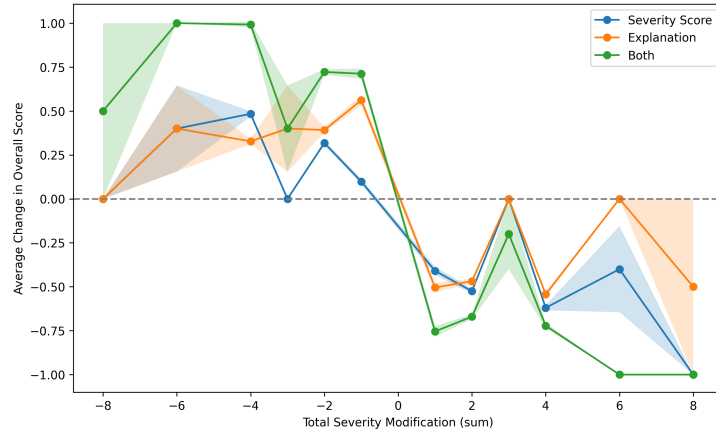


Figure 39: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating factual consistency on SummEval. Shade around lines shows standard deviation.

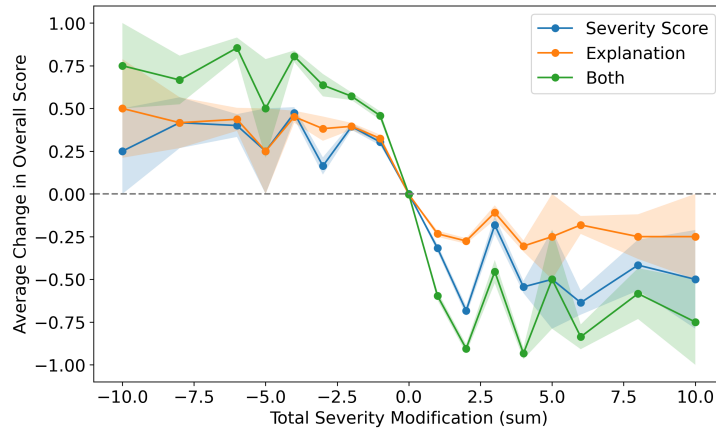


Figure 40: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating factual consistency on SummEval. Shade around lines shows standard deviation.

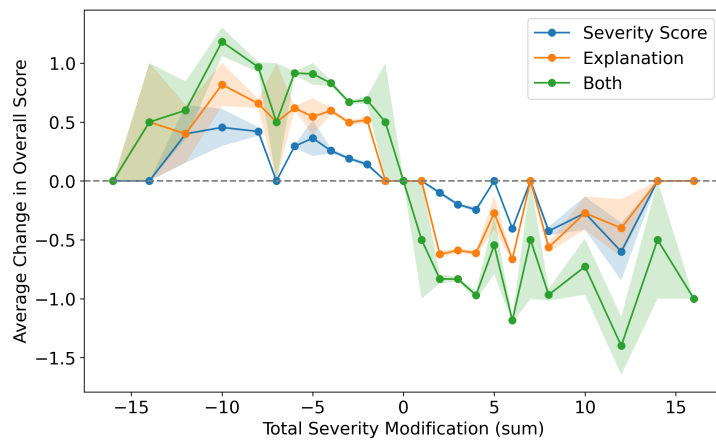


Figure 41: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on SummEval. Shade around lines shows standard deviation.

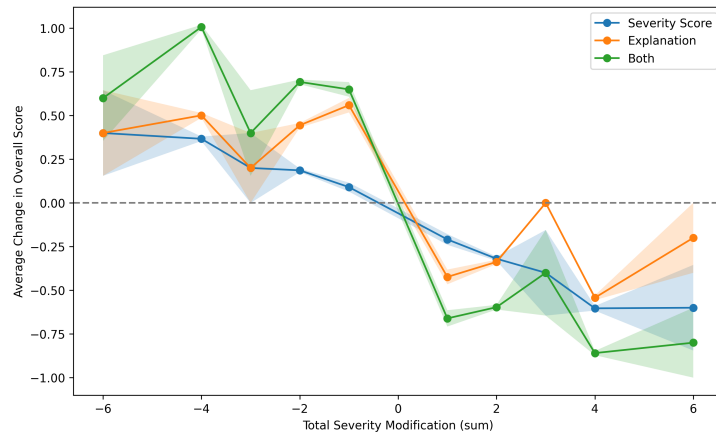


Figure 42: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating relevance on SummEval. Shade around lines shows standard deviation.

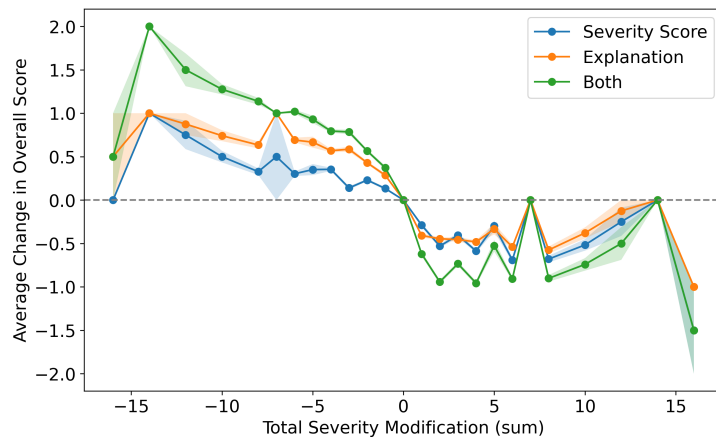


Figure 43: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating relevance on SummEval. Shade around lines shows standard deviation.

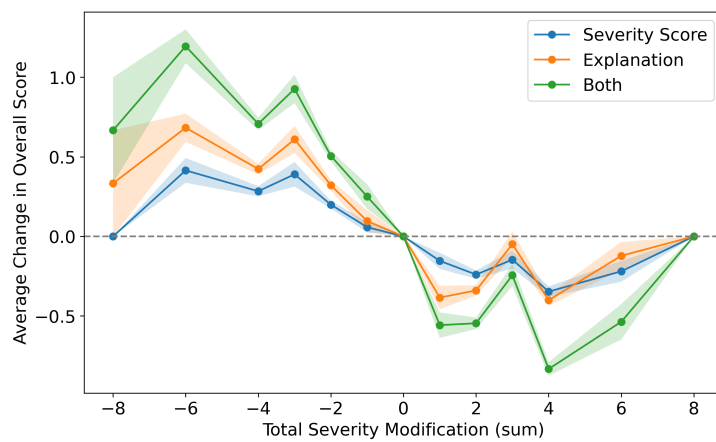


Figure 44: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on QAGS. Shade around lines shows standard deviation.

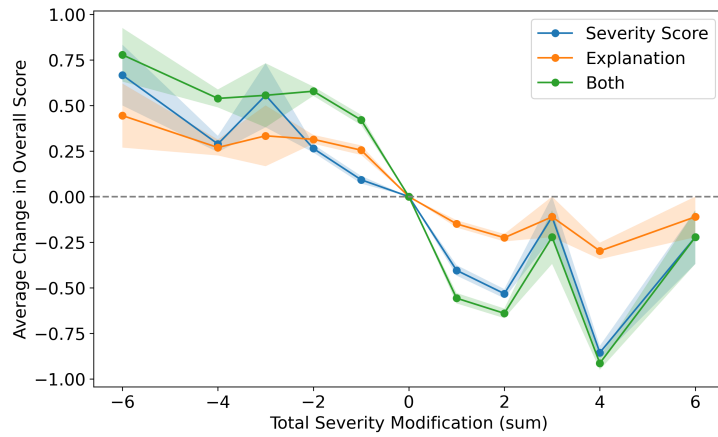


Figure 45: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Gemma while evaluating factual consistency on QAGS. Shade around lines shows standard deviation.

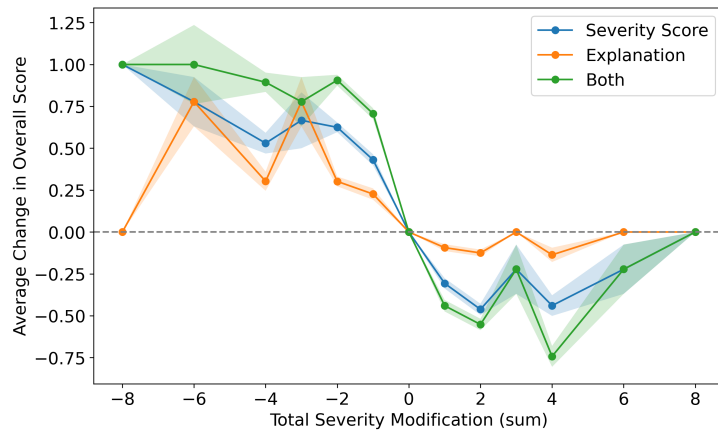


Figure 46: Average change in overall score in relation to total severity modification using different perturbations. The results are presented for Qwen 2.5 while evaluating factual consistency on QAGS. Shade around lines shows standard deviation.

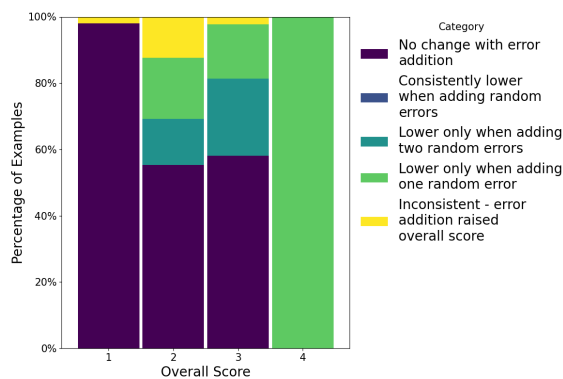


Figure 47: The percentage of examples which overall score changed after adding random errors. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on HANNA.

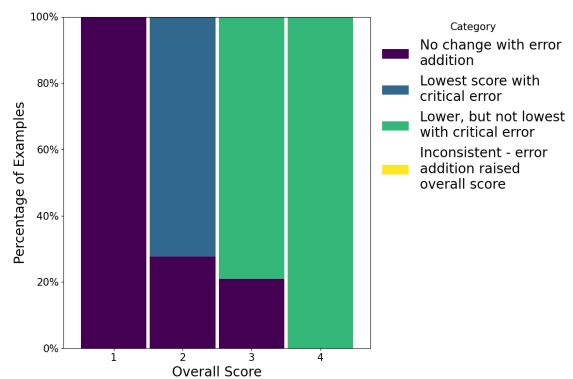


Figure 48: The percentage of examples which overall score changed after adding a critical error. The results are presented for Llama 3.1 Nemotron 70B while evaluating coherence on HANNA.

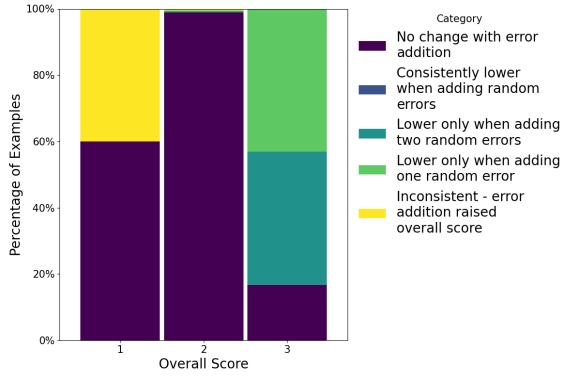


Figure 49: The percentage of examples which overall score changed after adding random errors. The results are presented for Gemma while evaluating coherence on HANNA.

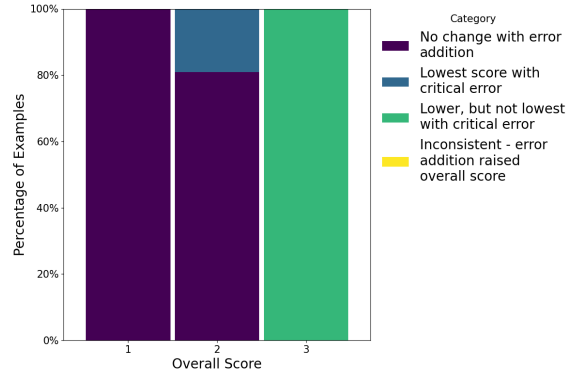


Figure 50: The percentage of examples which overall score changed after adding a critical error. The results are presented for Gemma while evaluating coherence on HANNA.

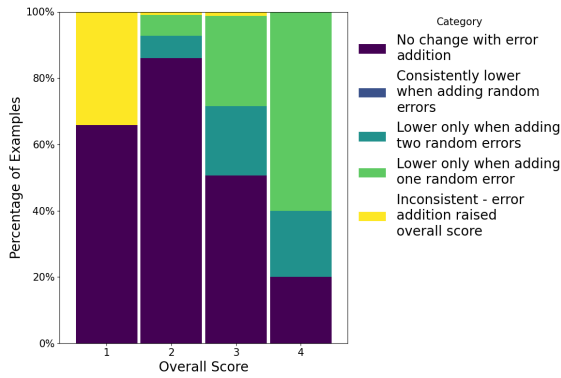


Figure 51: The percentage of examples which overall score changed after adding random errors. The results are presented for Qwen 2.5 while evaluating coherence on HANNA.

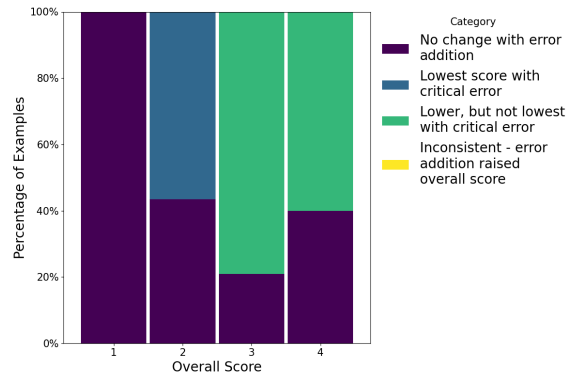


Figure 52: The percentage of examples which overall score changed after adding a critical error. The results are presented for Qwen 2.5 while evaluating coherence on HANNA.

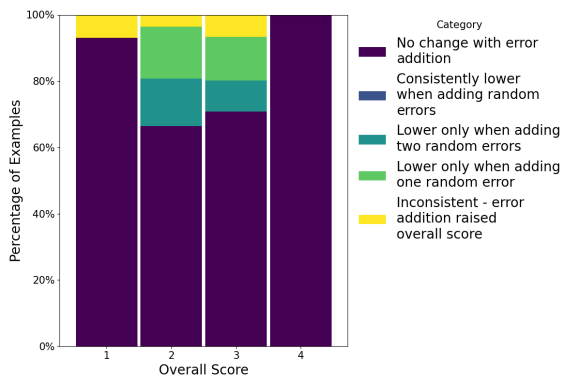


Figure 53: The percentage of examples which overall score changed after adding random errors. The results are presented for Llama 3.1 Nemotron 70B while evaluating complexity on HANNA.

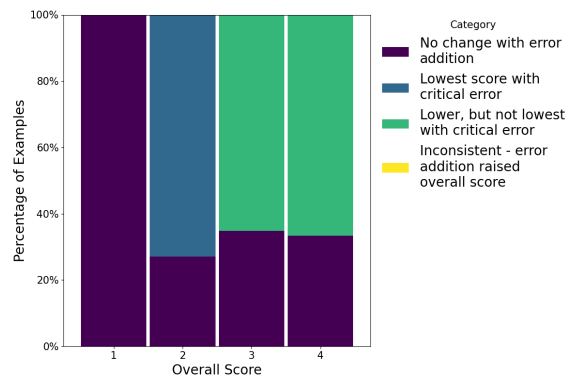


Figure 54: The percentage of examples which overall score changed after adding a critical error. The results are presented for Llama 3.1 Nemotron 70B while evaluating complexity on HANNA.

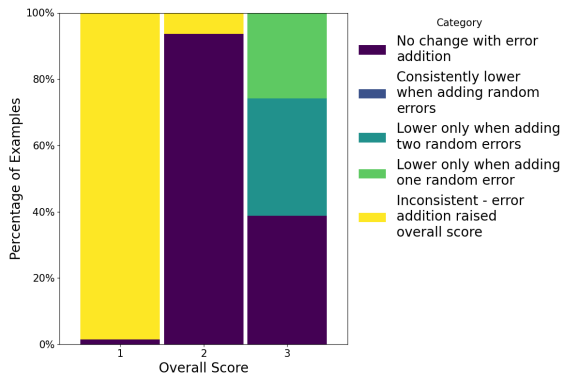


Figure 55: The percentage of examples which overall score changed after adding random errors. The results are presented for Gemma while evaluating complexity on HANNA.

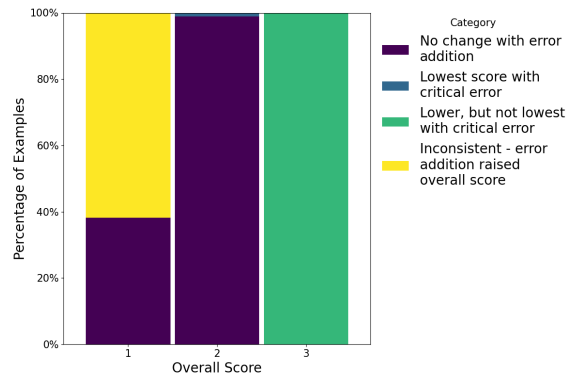


Figure 56: The percentage of examples which overall score changed after adding a critical error. The results are presented for Gemma while evaluating complexity on HANNA.

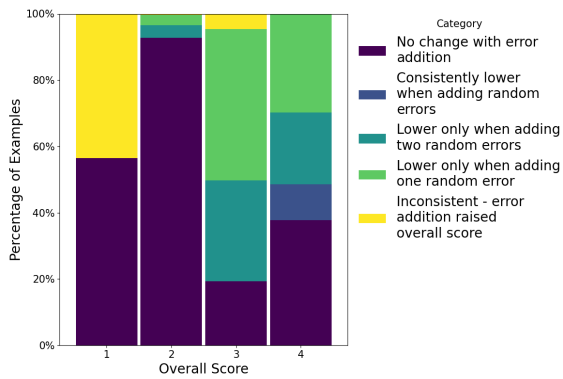


Figure 57: The percentage of examples which overall score changed after adding random errors. The results are presented for Qwen 2.5 while evaluating complexity on HANNA.

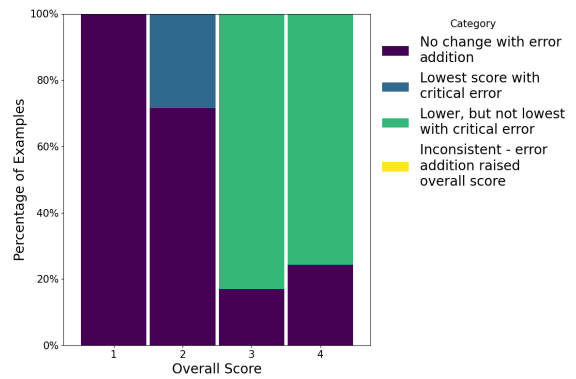


Figure 58: The percentage of examples which overall score changed after adding a critical error. The results are presented for Qwen 2.5 while evaluating complexity on HANNA.

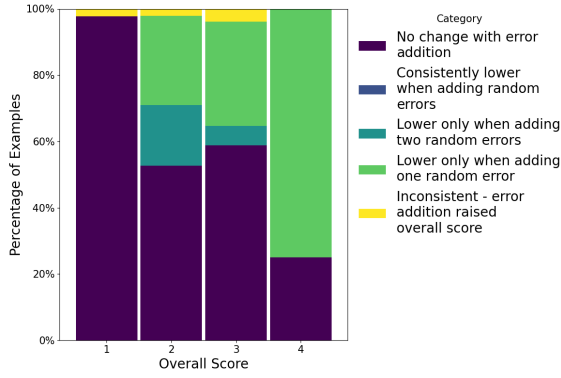


Figure 59: The percentage of examples which overall score changed after adding random errors. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on HANNA.

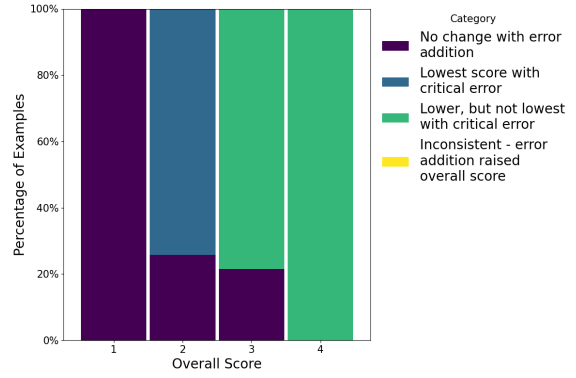


Figure 60: The percentage of examples which overall score changed after adding a critical error. The results are presented for Llama 3.1 Nemotron 70B while evaluating relevance on HANNA.

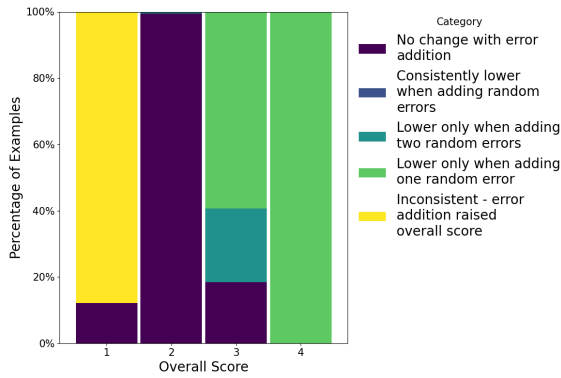


Figure 61: The percentage of examples which overall score changed after adding random errors. The results are presented for Gemma while evaluating relevance on HANNA.

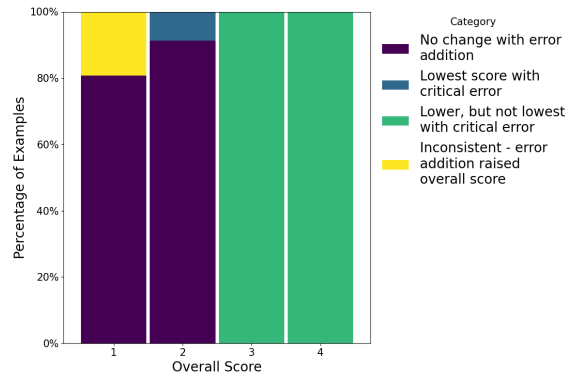


Figure 62: The percentage of examples which overall score changed after adding a critical error. The results are presented for Gemma while evaluating relevance on HANNA.

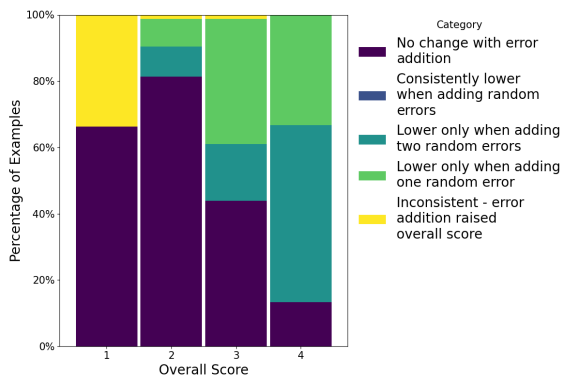


Figure 63: The percentage of examples which overall score changed after adding random errors. The results are presented for Qwen 2.5 while evaluating relevance on HANNA.

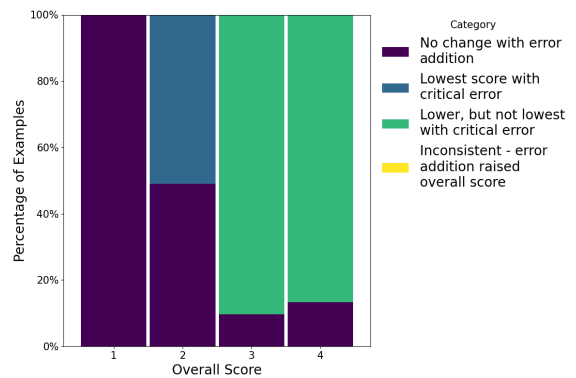


Figure 64: The percentage of examples which overall score changed after adding a critical error. The results are presented for Qwen 2.5 while evaluating relevance on HANNA.

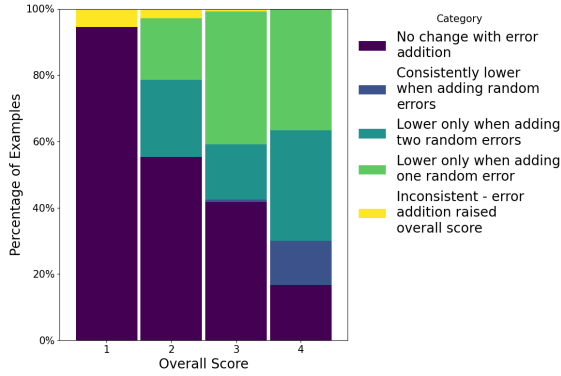


Figure 65: The percentage of examples which overall score changed after adding random errors. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on QAGS.

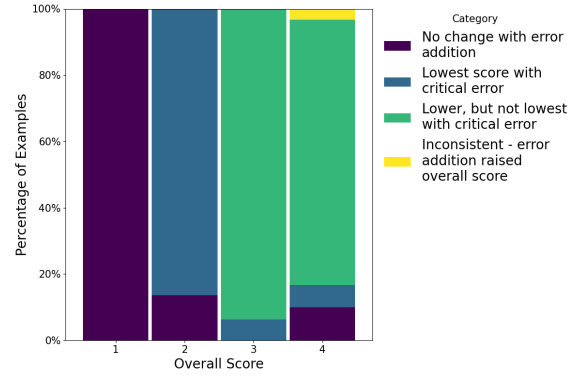


Figure 66: The percentage of examples which overall score changed after adding a critical error. The results are presented for Llama 3.1 Nemotron 70B while evaluating factual consistency on QAGS.

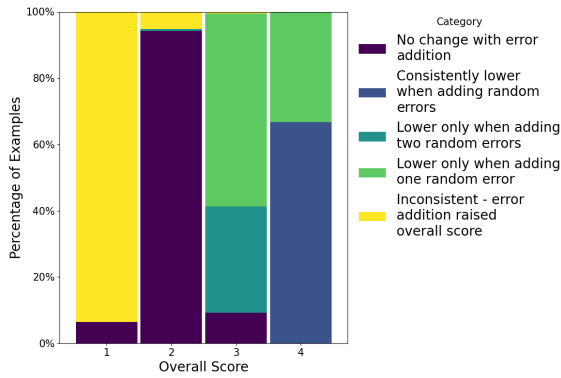


Figure 67: The percentage of examples which overall score changed after adding random errors. The results are presented for Gemma while evaluating factual consistency on QAGS.

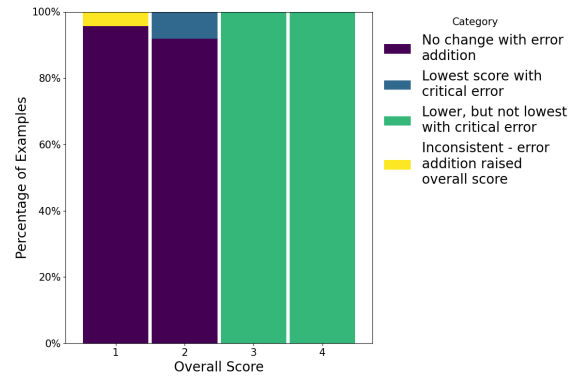


Figure 68: The percentage of examples which overall score changed after adding a critical error. The results are presented for Gemma while evaluating factual consistency on QAGS.

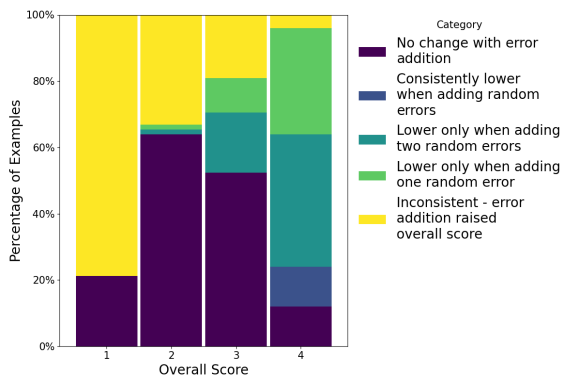


Figure 70: The percentage of examples which overall score changed after adding a critical error. The results are presented for Qwen 2.5 while evaluating factual consistency on QAGS.