Real-World Summarization: When Evaluation Reaches Its Limits

Patrícia Schmidtová
Charles University*
{schmidtova,odusek}@ufal.mff.cuni.cz

Ondřej Dušek
Charles University

Saad Mahamood
trivago N.V.
saad@saad.me.uk

Abstract

We examine evaluation of faithfulness to input data in the context of hotel highlights – brief LLM-generated summaries that capture unique features of accommodations. Through human evaluation campaigns involving categorical error assessment and span-level annotation, we compare traditional metrics, trainable methods, and LLM-as-a-judge approaches. Our findings reveal that simpler metrics like word overlap correlate surprisingly well with human judgments (r=0.63), often outperforming more complex methods when applied to out-of-domain data. We further demonstrate that while LLMs can generate high-quality highlights, they prove unreliable for evaluation as they tend to severely under- or over-annotate. Our analysis of real-world business impacts shows incorrect and non-checkable information pose the greatest risks. We also highlight challenges in crowdsourced evaluations.

1 Introduction

Instruction-tuned large language models (LLMs) have become ubiquitous in natural language processing (Qin et al., 2024). They proved very capable and versatile, with simple prompting achieving various tasks, avoiding the need to produce costly in-domain training datasets required for most previous approaches (Wei et al., 2022). However, LLMs are known to have issues with faithfulness of their outputs to the input, where they produce text not grounded in the input prompt (*hallucinations*; Ji et al., 2023).

Evaluating faithfulness is therefore crucial, especially without human-written references, which are expensive and risk leakage into LLMs' training data (Oren et al., 2024). However, traditional evaluation metrics for NLP show low correlation with human judgments (Novikova et al., 2017) and overt reliance on surface similarities (Gehrmann et al., 2023). Despite this, most works in the field still rely on them (Schmidtova et al., 2024). A new alternative is using LLMs themselves to evaluate generated outputs (Gu et al., 2024; Bavaresco et al., 2024). While promising, LLMs may show self-bias (Koo et al., 2024) and performance may vary across domains.

We focus on faithfulness evaluation of text summarization using a case study of generating short highlights from hotel descriptions and reviews (Kamath et al., 2024). We are concerned with the following research questions: (1) How well do faithfulness metrics generalize to unseen domains? (2) Can we evaluate faithfulness in a referenceless scenario? (3) Can LLMs be used as judges in this setting? and (4) What is the estimated business impact of these errors?

In response to these questions, we present the following contributions:

- We demonstrate that simple metrics (e.g., word overlap) outperform most trainable methods on out-of-domain data. Sophisticated methods may not generalize effectively.
- We validate multiple referenceless evaluation methods against human annotations, finding several methods that correlate well with human judgments (up to r=0.67).
- We provide empirical evidence cautioning against uncritical use of LLM-as-a-judge approaches, revealing systematic tendencies to either over- or under-annotate errors, depending on the model used.

2 Task Description

Hotel Highlights (Kamath et al., 2024) are brief LLM-generated summaries that capture unique features of hotels based on their descriptions and reviews. These highlights help travelers select appropriate accommodations without reading numerous reviews and lengthy descriptions. Consequently,

*Work done while on internship at trivago.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25014–25026
November 4-9, 2025

ensuring highlight accuracy is essential.

The highlights intentionally present hotels in a positive light, with subjective phrases such as 'a local gem' that, while not explicitly supported by the input text, are not considered errors. This characteristic presents additional evaluation challenges.

The objective of the task is not to produce a single gold-standard summary, but rather to generate a diverse set of approximately 10 highlights that can be ranked and sampled. We therefore focus primarily on verifying whether the information in each highlight is properly grounded in the source description. Table 1 demonstrates an example hotel description with two corresponding highlights.

3 Human Evaluation

3.1 Categorical Annotations

We utilized a dataset from prior work (Kamath et al., 2024), containing 120 description-highlight pairs. The summaries were generated using PaLM2 text-bison (PaLM Team, 2023). Each pair received annotations from 30 evaluators, categorized as: *no errors*, *hallucination* (absence of supporting evidence in the input), *contradiction* (of a statement in the output with respect to the input), or *both types* of errors. Each annotator completed one attention check – an example that contained a very prominent hallucination. For the purposes of this paper, we filtered out all of the annotators who did not pass this attention check. After filtering, 19 to 22 judgments were available for each example.

The annotators frequently did not fully agree on the presence of hallucination – we chose to interpret this as a signal and operate with the percentage of eligible annotators who marked an example as hallucination, contradiction, or both. Based on our manual analysis, this percentage was a good indicator of error saliency: Annotators agreed more on blatant errors and tended to disagree on more subtle errors, for example, those caused by ambiguity (swimming pool vs. pool table), or subjective interpretations of objective facts (is 10km close?).

3.2 Span Annotations

We implemented span-level error annotation to get more explainable and actionable feedback on highlight quality. For this experiment, we used a fresh set of description-highlight pairs generated by Gemini 1.5 Flash (Gemini Team, 2024). By manual inspection, we identified three error types: *non-checkable* (facts not present in the description), *misleading* (facts taken out of context), and *incorrect* (contradicting the description).

Annotators We gathered annotations from 124 crowd workers recruited via Prolific across 496 description-highlight pairs. The annotators were native English speakers from the United Kingdom or United States with >90% approval rates.

Before launching the evaluation, we conducted three pilot studies to assess guideline clarity, cognitive load, and expected completion time. The total annotation cost, including pilots, approached £800.

Method Each annotator evaluated 8 sampled pairs plus two manually selected attention checks, totaling 10 examples. The annotations were collected using the Factgenie interface (Kasner et al., 2024) Each example was annotated by two annotators (each belonging in either group A or B, with a subtle annotation instruction distinction explained in Appendix D). Based on pilot findings, we ran separate campaigns for shorter and longer descriptions, to provide consistent time estimates for completion. The annotators were positively motivated to focus on quality by a bonus payment to those passing our two attention checks.

Results The annotations (summarized in Figure 1) indicate 58% of highlights are error-free, while the remainder contain non-checkable (20%), misleading (19%), or incorrect (7%) content.[1]

Quality We deployed two attention checks to gauge the quality of the annotations. One of the them required spotting an incorrect error, such as H2 in Table 1, 76 % of annotators marked an error in the example with only 39 % correctly identifying the error as incorrect. The other check required the annotator to abstain from marking a span as an error given the example was carefully checked to be error-free. 51 % of annotators abstained and thus passed this check. 33% of annotators passed both of our attention checks.

Based on these insights, we anticipate that the recall of our annotators was acceptable (disregarding the error type). On the other hand, over-annotation emerged as a common issue. This was confirmed by internal domain experts who analyzed 20 annotations per error type and found slightly over half of the annotated spans contained no actual errors. This problem was more pronounced in group B. We suspect this stems from source text length – when

[1] Individual highlights may contain multiple error types.

Description: *Just a 5-minute walk from Mall of the Emirates, DoubleTree by Hilton Hotel and Residences Dubai offers modern accommodations. [...] The hotel is 7.0 km from Dubai Marina and 12.1 km from Dubai Mall. Dubai International Airport is 30 minutes away by car.*
H1: *Shop in the Mall of the Emirates thanks to the hotel's convenient location.*
H2: *Enjoy wonderful views across the Hudson River to New Jersey and Liberty Island from select suites.*

Table 1: Hotel description excerpt with two corresponding highlights. H1 is correct but H2 contains incorrect information (highlighted in red). H2 was used as one of the attention checks.

annotators struggled to quickly locate information in the text, they marked the spans containing this information as errors. This occurred despite explicit encouragement to use Ctrl+F for efficiently locating information.

3.3 Estimation of Real-World Impacts

On a sample of 60 error span annotations described above, we determined that 32 have *no business impact* (no actual error), 13 present *low business impact* (clients unlikely to complain about being misled), 13 show *medium business impact* (clients might complain without requesting compensation), and only 2 indicate *high business impact* (clients likely to request compensation or a significant reputation risk). Incorrect information most frequently causes a higher business impact, followed by non-checkable information.

4 Validating Quality Estimation Methods

4.1 Example-level Binary Classification

We experimented with automatically determining whether a given example contains a semantic error or not. Using data described in Section 3.1, we calculate the Spearman rank correlation between automatic metric scores and the percentage of annotators who believe there is an error in a given highlight. By doing so, we are aiming to capture the subjectivity and uncertainty as a signal – if most annotators agree there is (not) an error, then the automatic metric should agree with the majority to be considered reliable. We consider the following metric types:

Single Word Overlap The simplest method we consider – overlap of word forms between the highlights and descriptions – proved to reach the second highest correlation with the human annotation. We tested several variations, shown in Table 2. Word overlap can easily be confused by phenomena such as negation; however, it is cheap and quick to calculate.

Type	Metric	Corr. (Spearman)
O	Form / Lemma overlap	**0.63** / 0.62
	Noun overlap	0.41
	Adjective overlap	0.55
N	BLEU (no BP)	0.51
	ROUGE-L (P / R / F)	0.38 / **0.56** / 0.41
T	NLI-entailment	**0.67**
	AlignScore – base NLI	0.58
	BertScore	0.57
	LaBSE	0.12

Table 2: Correlation for overlap-based methods (O), n-gram overlap methods (N), and trainable methods (T). This table includes a small selection of the explored methods, see Table 4 in the Appendix for full results.

N-Gram Overlap We measured BLEU (Post, 2018) (without brevity penalty) and ROUGE-L (Lin, 2004) between the highlights and the descriptions. They reach correlations with human judgments comparable to single-word methods while being more robust, because they consider a longer combination of n-grams.

Natural Language Inference (NLI) was demonstrated to work as a referenceless metric for semantic accuracy in data-to-text generation (Dušek and Kasner, 2020) and summarization (Maynez et al., 2020). If the generated summary is entailed by the source description, the intended meaning was likely preserved. On the contrary, if the summary is not entailed by the description, then it is likely there is a semantic error.

We performed initial experiments with older NLI models (He et al., 2021; Laurer et al., 2022) to measure the entailment likelihood between the source and the highlight. The results did not seem promising – for the vast majority of samples, the likelihood of entailment was very close to 0 or 1, with a maximum correlation of 0.18 with human judgments on hallucinations.

However, using ModernBERT (Sileo, 2024),[2] a

[2] https://huggingface.co/tasksource/ModernBERT-base-nli

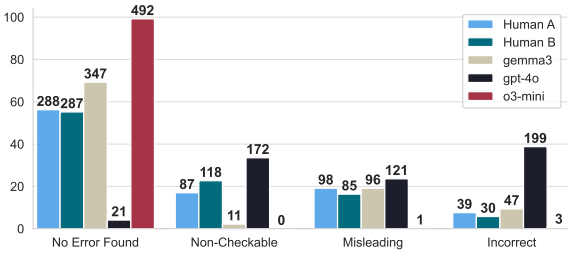

Figure 1: Dataset-level distribution of error types based on human and LLM-based span annotation campaigns. The y axis shows % of the test set marked with the given error type, with numbers of instances shown above.

newer model trained on more data, proved to be helpful as we reached a correlation of 0.67 with human judgment on the categorical set. AlignScore (Zha et al., 2023), a metric based on NLI reached a moderate correlation of 0.58.

Text Embedding Similarity We have experimented with 6 different embedding models and found that BertScore (Zhang et al., 2020) reached a moderate correlation with human judgment. By comparison, the next similarity-based measure – cosine similarity of LaBSE embeddings (Feng et al., 2022) reached an unsatisfying correlation of 0.12. The issue with these measures was that the similarity always stays high due to thematic closeness but is unable to reflect small pieces of unsupported information in the highlights.

4.2 Locating the Error and its Severity

LLM-as-a-Judge Many papers rely on LLMs for annotation, often replacing human annotation to save money. Following Kocmi et al. (2024) and Kasner et al. (2025), we used LLMs as span annotators to identify errors in the text and to provide a reasoning for each error.

We used GPT-4o (OpenAI Team, 2024), o3-mini,[3] and Gemma3 (Gemma Team, 2025). We present the prompt used in Appendix C. As shown in Figure 1, GPT-4o heavily overannotates, frequently citing a nonsensical reason. This is especially notable in the *incorrect* error category. Conversely, o3-mini severely under-annotates, although the few annotations it produces are accurate. Gemma3 most closely matches the human distribution of errors, but it under-utilizes the *non-checkable* category.

We used precision, recall, and F1 to measure agreement on the example level, shown in Table 3. The precision and F1 between the two human annotator groups set a reasonable baseline and we can see that none of the LLMs reaches this baseline yet. On the other hand, GPT-4o achieves a high recall due to its overannotation. We note that span annotation yields lower agreement scores than categorical ratings because annotators have significantly more choices when selecting text boundaries.

Curiously, prompting the models to be more lenient and explicitly showing an example of a correct highlight led to a higher count of annotated errors. In sum, we are able get more granular insights from human annotation, but LLMs are not yet to be fully trusted to evaluate this task.

5 Discussion

Gold Human Annotation? After investing £800 in crowdsourced span annotations and receiving poor-quality results despite multiple precautions (pre-filtering participants, conducting pilots, and offering payment bonuses for attention checks), we conclude that alternative approaches may be more cost-effective for difficult tasks. These include evaluating smaller output samples with larger annotator groups, or whitelisting trusted annotators who pass multiple qualification tests (Zhang et al., 2023). We refer to Schmidtová et al. (2025) for more observations on human evaluation of hallucinations.

Some LLMs pass attention checks Gemma3 and o3-mini both passed the attention check that was designed to be simple and visible, yet only 33% of crowdworkers passed it. This shows that LLMs can capture the more visible errors, but still have room for improvement in subtler errors.

Running Gemma3 without validation would underestimate the amount of non-checkable errors which have a higher business impact. Similarly, relying on GPT-4o without the context of validation would significantly overestimate the seriousness of incorrect errors.

Evaluation lags behind generation Tang et al. (2023) found that most trainable summarization metrics work more reliably on older model outputs compared to the new ones. Indeed, we can see that the best trainable metric is the ModernBert NLI model from 2024 (Sileo, 2024). On the other hand, BertScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), as well as the older NLI models (Laurer et al., 2022; He et al., 2021) performed worse.

[3] https://platform.openai.com/docs/models/o3-mini

Reference	Hypothesis	Ref Ct	Hyp Ct	Prec (H)	Rec (H)	F1 (H)	Prec (S)	Rec (S)	F1 (S)
Human A (longer)	Human B (longer)	176	154	0.179	0.223	**0.199**	0.307	0.382	**0.340**
Human A (longer)	Gemma3	176	85	0.082	0.060	0.069	0.185	0.136	0.157
Human A (longer)	GPT-4o	176	258	0.068	0.141	0.092	0.143	0.293	0.192
Human A (shorter)	Human B (shorter)	144	161	0.107	0.104	0.106	0.288	0.280	0.284
Human A (shorter)	Gemma3	144	93	0.085	0.066	0.074	0.226	0.175	0.197
Human A (shorter)	GPT-4o	144	298	0.108	**0.232**	0.147	0.198	0.425	0.270
Human B (longer)	Human A (longer)	154	176	**0.217**	0.175	0.194	**0.376**	0.302	0.335
Human B (longer)	Gemma3	154	85	0.052	0.031	0.039	0.218	0.129	0.162
Human B (longer)	GPT-4o	154	258	0.066	0.109	0.082	0.211	0.349	0.263
Human B (shorter)	Human A (shorter)	161	144	0.104	0.107	0.106	0.280	0.288	0.284
Human B (shorter)	Gemma3	161	93	0.107	0.085	0.095	0.253	0.202	0.225
Human B (shorter)	GPT-4o	161	298	0.076	0.167	0.104	0.210	**0.464**	0.289

Table 3: Agreement metrics for span-level annotation (for two human annotator groups and human vs. LLM annotation). In each row, a Hypothesis annotation campaign is compared to a Reference campaign. Ref and Hyp Ct show annotation counts for each campaign. Soft (S) metrics count any overlapping spans as matches, disregarding error types. Hard (H) metrics require both span overlap and matching error types. Highest values are shown in bold.

This insight would explain our observation that n-gram metrics outperformed the majority of trainable metrics: as models evolve and make different types of errors, metrics that have been trained on the outputs of past models fail to generalize to these new error types.

Correlation strength While correlations around $r = 0.63$ may seem moderate, they are high by current NLG standards, where many metrics achieve correlations below 0.3 with human judgments (Novikova et al., 2017) and even highly optimized LLM metrics are in the 0.4-0.7 range for well-known tasks (Hu et al., 2024).

Simpler methods to the rescue We observed that objective and quick to compute metrics, such as word overlap, correlate well with human judgment. We argue that they are a solid choice to be measured and reported in the absence of other evaluation metrics for estimating faithfulness. We do not condone using such simple metrics in isolation to claim state-of-the-art results, but rather emphasize their importance as complementary tools, especially when evaluating new tasks or domains in business contexts requiring scalable quality control without incurring overwhelming costs.

Moramarco et al. (2022) have observed a similar effect when evaluating generated consultation notes in the medical domain. This further supports our hypothesis that trainable methods have limited generalizability to out-of-distribution domains.

6 Related Work

The validation of automatic metrics against human judgment is an active area of research, with broad consensus that metrics cannot fully substitute for human evaluation (Belz and Reiter, 2006; Novikova et al., 2017). This challenge has intensified with recent models, whose outputs often fall into metrics' blind spots (Tang et al., 2023).

However, few works evaluate the real-world impact of NLP systems (Reiter, 2025). Moramarco et al. (2022) validated automatic metrics against human judgment for medical consultation note generation, finding – consistent with our work – that human insights are essential for assessing practical utility, and that simple metrics retain significant value. Similarly, Pu et al. (2024) employed auxiliary tasks such as question-answering to evaluate summary usefulness, showcasing alternative approaches to intrinsic metrics.

7 Conclusion

Addressing our research questions, we demonstrated that: (1) NLI entailment and simple statistical metrics achieve moderate correlation with human judgments and are thus the best out-of-the-box options for measuring faithfulness; (2) referenceless evaluation can be effective when validated properly; (3) while LLMs excel at generating hotel highlights, they prove unreliable as evaluators of content faithfulness; and (4) non-checkable and incorrect information have the highest potential for negative business impact.

We believe that real-world evaluation of tasks that emerged with the rise of LLMs and few-shot prompting should be studied carefully. Current evaluation methods are insufficient for automated quality assurance, and the errors that go unnoticed are likely to cause a negative business impact.

Limitations

Despite our precautions, the span annotations from crowdworkers were of a less-than-satisfactory quality. We still used this data for the validation of LLM-as-a-judge, because it was not feasible for us to annotate sufficient quantity of data in-house. We believe there are still signals to be learned from this noisy data. We highlight that the span annotations were not used for the main automatic correlation analysis – for this, we used the categorical data with 19 to 22 judgments of annotators who passed the attention check.

Ethical Consideration

Human Annotations The payment structure included an £8 per hour base rate paid out to all annotators after finishing the task – regardless of their annotation quality. We paid out a bonus of £4.60 per hour to workers who passed our attention check. This ensured compliant workers received the UK living wage of £12.60 per hour.[4] For comparison, the Prolific minimum wage is £6.00 per hour and the recommended wage is £9.00 per hour.[5]

Model Inference The total cost to run the LLMs for span annotations (2-3 runs on 500 examples per model to optimize the prompt) through APIs was less than $100.

Use of AI We used AI-assisted coding (i.e. Copilot) with the bulk being human-written. For writing, AI was used to check grammar mistakes.

Acknowledgments

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV 260 698. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

We would also like to thank all members of the insights and accommodation profile teams for their support and helpful feedback on this work. Specifically, we thank Agata Abram, Srinivas Ramesh Kamath, and Tatenda Matika for their help with business impact estimation and Ricardo Vega for his help with deploying Factgenie within trivago. Last, but not least, we thank our reviewers and area chair for their helpful comments and suggestions that led to the improvement of this manuscript.

[4] https://www.livingwage.org.uk/

[5] https://researcher-help.prolific.com/en/article/2273bd

References

Miriam Anschütz, Diego Miguel Lozano, and Georg Groh. 2023. This is not correct! negation-aware evaluation of language generation systems. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 163–175, Prague, Czechia. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on LLM-as-a-Judge. *CoRR*, abs/2411.15594.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. 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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248. ArXiv: 2202.03629.

Srinivas Ramesh Kamath, Fahime Same, and Saad Mahamood. 2024. Generating hotel highlights from unstructured text using LLMs. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 280–288, Tokyo, Japan. Association for Computational Linguistics.

Zdeněk Kasner, Ondrej Platek, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. factgenie: A framework for span-based evaluation of generated texts. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 13–15, Tokyo, Japan. Association for Computational Linguistics.

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.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI Team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving Test Set Contamination in Black-Box Language Models. In *ICLR*, Vienna, Austria.

PaLM Team. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2024. Is summary useful or not? an extrinsic human evaluation of text summaries on downstream tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9389–9404, Torino, Italia. ELRA and ICCL.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, pages 1–13.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Patrícia Schmidtová, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondřej Dušek. 2025. Do my eyes deceive me? A survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, Hanoi, Vietnam. Association for Computational Linguistics.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. Online.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Full Spearman Rank Correlation Results

In Table 4, we present Spearman rank correlations of all automatic methods with categorical human judgment (column Hallucination %) as well as the presence of a certain error category in the span-annotated data. In the span-annotated data, the correlations are lower, because they are compared to discrete values:

- 0 of the two annotators found the given error type
- 1 of the annotators found the given error type
- 2 – both annotators found the given error type

B Implementation Details

B.1 Metrics

We used the sacrebleu package to compute the frequencies of 1- to 4-grams. We then computed the harmonic mean ourselves to avoid the brevity penalty, that was undesirable in our case. To compute Rouge scores, we used the rouge_score Python package. For trainable metrics, we used models available on HuggingFace:

- msmarco-distilbert-base-v4 (Reimers and Gurevych, 2019)[6]
- msmarco-distilbert-base-tas-b (Reimers and Gurevych, 2019)[7]
- sentence-transformers/all-mpnet-base-v2 (Reimers and Gurevych, 2019)[8]
- tum-nlp/NegMPNet (Anschütz et al., 2023)[9]
- sentence-transformers/LaBSE (Feng et al., 2022) [10]
- tasksource/ModernBERT-base-nli (Sileo, 2024)[11]
- MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (Laurer et al., 2022)[12]
- cross-encoder/nli-deberta-v3-base (He et al., 2021) [13]

We use the implementation of BertScore from HuggingFace's evaluate library. For part-of-speech tagging and named entity recognition, we used SpaCy (Honnibal et al., 2020). For AlignScore (Zha et al., 2023), we used their GitHub repository directly.[14]

B.2 LLM Judges

We accessed the most recent version of OpenAI models via their API in April 2025. We ran Gemma3 locally using the Ollama library [15] with the 'gemma3:27b' checkpoint.

C LLM Judge Prompts

C.1 Main Prompt

The main prompt was trialled by Kasner et al. (2025) and found to work well for machine translation and data-to-text generation.

Given the hotel descriptions: `{data}`

Annotate all the errors in the following summary: `{text}`

Output the errors as a JSON list "annotations" in which each object contains fields "reason", "text", and "annotation_type". The value of "text" is the text of the error. The value of "reason" is the reason for the error. The value of "annotation_type" is one of {0, 1, 2, 3} based on the following list:

- 0: Not checkable: The fact in the text cannot be checked in the data.
- 1: Misleading: The fact in the text is misleading in the given context.
- 2: Incorrect fact: The fact in the text contradicts the data.

The list should be sorted by the position of the error in the text. Make sure that the annotations are not overlapping.

Example:

Data: "The closest major airports to Bomontist Suit are: Istanbul (SAW-Sabiha Gokcen Intl.) - 17.5 km / 10.9 mi Istanbul (IST-Ataturk Intl.)."

Summary: "Schiphol Airport is just a 15-minute drive from the hotel."

Output:

[6] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4

[7] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[9] https://huggingface.co/tum-nlp/NegMPNet

[10] https://huggingface.co/sentence-transformers/LaBSE

[11] https://huggingface.co/tasksource/ModernBERT-base-nli

[12] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

[13] https://huggingface.co/cross-encoder/nli-deberta-v3-base

[14] https://github.com/yuh-zha/AlignScore/

[15] https://ollama.com/

Type	Metric	Hal. %	Non-Checkable	Misleading	Incorrect	Any Error
Trainable	BERTScore F1*	-0.39	-0.11	0.06	-0.04	-0.06
Trainable	BERTScore Precision*	-0.58	-0.13	0.06	-0.02	-0.08
Trainable	BERTScore Recall*	-0.24	-0.09	0.05	-0.05	-0.04
Trainable	MBERT Entailment*	**-0.67**	-0.23	-0.06	-0.04	**-0.22**
Trainable	MBERT Neutral	0.66	**0.24**	0.05	0.04	**0.22**
Trainable	MBERT Contradiction	0.36	0.11	0.06	0.05	0.15
Trainable	mDeBERTa Entailment*	-0.07	-0.04	0.10	0.00	0.02
Trainable	mDeBERTa Neutral	0.10	0.03	-0.10	-0.02	-0.06
Trainable	mDeBERTa Contradiction	-0.18	-0.01	-0.01	0.05	0.01
Trainable	DeBERTa v.3 Entailment*	-0.08	-0.05	0.05	0.00	-0.01
Trainable	DeBERTa v.3 Neutral	0.03	0.08	-0.01	0.00	0.05
Trainable	DeBERTa v.3 Contradiction	0.04	0.01	-0.03	-0.01	-0.02
Trainable	AlignScore Base NLI SP*	-0.57	-0.14	-0.06	0.03	-0.13
Trainable	AlignScore Base NLI*	-0.58	-0.12	0.05	-0.05	-0.07
Trainable	AlignScore Base Bin SP*	-0.27	-0.05	-0.10	-0.00	-0.10
Trainable	AlignScore Base Bin*	-0.28	-0.03	-0.03	-0.05	-0.07
Trainable	AlignScore Large NLI SP*	-0.47	-0.16	-0.08	-0.02	-0.17
Trainable	AlignScore Large NLI*	-0.48	-0.13	-0.01	-0.05	-0.11
Trainable	AlignScore Large Bin SP*	-0.41	-0.06	-0.07	-0.03	-0.10
Trainable	AlignScore Large Bin*	-0.42	-0.08	-0.08	-0.06	-0.13
Trainable	Cosine Similarity*	0.00	-0.02	**0.11**	-0.04	0.04
Trainable	Dot Score*	-0.04	-0.04	0.07	-0.03	0.01
Trainable	MPNet Similarity*	-0.00	0.04	0.07	0.03	0.08
Trainable	NegMPNet Similarity*	-0.01	0.05	0.01	-0.05	0.01
Trainable	LaBSE Similarity*	-0.12	0.02	0.06	0.02	0.05
Word Overlap	Noun Coverage*	-0.41	-0.13	-0.05	0.02	-0.10
Word Overlap	Adjective Coverage*	-0.55	**-0.17**	**-0.09**	-0.01	**-0.18**
Word Overlap	Verb Coverage*	-0.18	-0.01	-0.04	-0.05	-0.08
Word Overlap	Form Coverage*	**-0.63**	-0.16	**-0.09**	-0.05	**-0.17**
Word Overlap	Lemma Coverage*	-0.62	-0.16	-0.07	-0.01	-0.15
Word Overlap	Entity Coverage*	-0.05	0.01	0.04	**-0.10**	-0.02
Other	Num. of Entities*	0.01	0.01	0.08	0.09	0.09
N-gram Overlap	BLEU (no brevity penalty)*	-0.51	-0.08	-0.03	-0.04	-0.09
N-gram Overlap	ROUGE-1 P*	-0.34	-0.07	0.06	-0.04	-0.05
N-gram Overlap	ROUGE-1 R*	**-0.58**	**-0.16**	**-0.09**	-0.04	**-0.18**
N-gram Overlap	ROUGE-1 F*	-0.37	-0.07	0.06	-0.04	-0.05
N-gram Overlap	ROUGE-L P*	-0.39	-0.10	0.06	-0.04	-0.06
N-gram Overlap	ROUGE-L R*	**-0.56**	**-0.18**	-0.06	-0.02	**-0.15**
N-gram Overlap	ROUGE-L F*	-0.41	-0.11	0.06	-0.04	-0.06

Table 4: Spearman rank correlations between various metrics and hallucination percentage obtained from categorical human data. Note that the values of many metrics are inversely correlated with hallucination likelihoods, we mark those with an asterisk (*). Also shown are correlations with span-annotated human data: Non-Checkable, Misleading, Incorrect, and Any Error. The categorical data are more granular, allowing for a greater variety of ranks. On the other hand, with the span data, we only have three options for ranking a given summary depending on whether neither, one, or both annotators mark the presence of a given error type.

```
{
  "annotations": [
    {
      "reason": "Schiphol Airport is incorrect
                 as the accommodation
                 is in Istanbul",
      "text": "Schiphol Airport",
      "annotation_type": 2
    },
    {
      "reason": "15-minute drive cannot
                 be verified as we only know
                 distance, not journey time",
      "text": "15-minute drive",
      "annotation_type": 0
    }
  ]
}
```

C.2 "More Lenient" Prompt

After seeing gpt4o over-annotate, we trialled a "more lenient" prompt. The changes compared to the main prompt are the introduction of an example with no errors and a paragraph containing further instructions about which phenomena to not annotate as errors. However, this only made the model annotate more.

Given the hotel descriptions: `{data}`

Annotate all the errors in the following summary: `{text}`

Output the errors as a JSON list "annotations" in which each object contains fields "reason", "text", and "annotation_type". The value of "text" is the text of the error. The value of "reason" is the reason for the error. The value of "annotation_type" is one of {0, 1, 2, 3} based on the following list:

- 0: Not checkable: The fact in the text cannot be checked in the data.
- 1: Misleading: The fact in the text is misleading in the given context.
- 2: Incorrect fact: The fact in the text contradicts the data.

The list should be sorted by the position of the error in the text. Make sure that the annotations are not overlapping.

Example:

Data: "The closest major airports to Bomontist Suit are: Istanbul (SAW-Sabiha Gokcen Intl.) - 17.5 km / 10.9 mi Istanbul (IST-Ataturk Intl.)."
Summary: "Schiphol Airport is just a 15-minute drive from the hotel."
Output:

```
{
  "annotations": [
    {
      "reason": "Schiphol Airport is incorrect
                 as the accommodation
                 is in Istanbul",
      "text": "Schiphol Airport",
      "annotation_type": 2
    },
    {
      "reason": "15-minute drive cannot
                 be verified as we only know
                 distance, not journey time",
      "text": "15-minute drive",
      "annotation_type": 0
    }
  ]
}
```

Example:
Data: "Immerse yourself in Florida's culinary heritage with Latin fusion flavors at our restaurant, Blue Matisse, or sip craft cocktails at Nau Lounge."
Summary: "Experience the vibrant flavors of Latin cuisine with a modern twist at Blue Matisse restaurant."
Output:

```
{
  "annotations": []
}
```

Note that some details may not be mentioned in the text: do not count omissions as errors. Also do not be too strict: some facts can be less specific than in the data (rounded values, shortened or abbreviated text, etc.), do not count these as errors. Sometimes, stronger adjectives will be used to make the summary more exciting, these are also not errors. If there are no errors in the text, "annotations" will be an empty list.

D Annotator Instructions

We present the instructions given to annotators below. We note that there were two groups A and B, with the only distinction in the final checkbox wording: the checkbox said "I did not find any errors in this summary" for group A and "There were no errors in this summary" for group B. We found no significant difference that could be attributed to this factor.

! For technical reasons, please use a different browser than Safari !

You will see a **collection of texts describing a hotel** and a **short summary of the texts focusing on a specific aspect of the hotel**. Your task will be to read both the texts and the summary and

identify parts of the summary that contain the errors described below.

Definitions and Examples of the Errors We present these in Table 5. In the original interface, they were presented to the annotators using Markdown formatting.

Guidelines for identifying the parts that contain an error
To mark a part of the sentence that contains the error, drag your cursor to highlight the text. **Aim to select the smallest span** that, if removed or replaced, would correct the error while allowing the rest of the sentence to remain intact.
Some summaries will not contain any errors, in such case you are expected to not annotate any spans and instead check the box saying "*I did not find any errors in this summary*", rate your overall impression and move on to the next example.

Example
Text: *Immerse yourself in Florida's culinary heritage with Latin fusion flavors at our restaurant, Blue Matisse, or sip craft cocktails at Nau Lounge.*
Summary: Experience the vibrant flavors of Latin cuisine with a modern twist at Blue Matisse restaurant. *(No span is selected)*
Explanation: This summary contains no errors, so instead of selecting any spans, just confirm "*I did not find any errors in this summary*" in the checkbox below the text.

Error Type	Definition	Example
Not Checkable	The summary contains information that is not mentioned anywhere in the original text. This information could either be objective (such as the presence of a swimming pool) or subjective (such as quietness).	**Text:** *A fun-filled vacation or relaxing business trip awaits you at the Holiday Inn Express & Suites Tampa Airport nestled on the beautiful waters of Tampa Bay at Rocky Point. Our hotel is minutes from the beautiful waterfront views of Tampa's Famous Riverwalk featuring miles of shops, artists and Tampa's premier dining. Our friendly and knowledgeable staff invite you to relax in the outdoor pool.* **Summary:** Enjoy stunning views of Tampa Bay and the beautiful waterfront from this pet-friendly hotel. **Explanation:** It was not mentioned whether the hotel is pet-friendly, thus this information is Not Checkable.
Misleading	The summary presents information that appears in the original text, however, it does so in a way that changes the perceived meaning. This can be due to subjective judgments (is an attraction 10 km away "close"?) or due to a word that can have multiple meanings (pool as in swimming pool or the game requiring a pool table).	**Text:** *Sheraton Düsseldorf Airport hotel is directly connected with the Terminal - in the unique location on the roof of car park P3, surrounded by 10,000m² greenery. [...] Relax from your travels or prepare for your meeting with green views.* **Summary:** Enjoy breathtaking views from the rooftop terrace and garden, offering a relaxing escape. **Explanation:** Terrace and garden are Misleading. The hotel seems to be on the roof, but there is no mention of a terrace. At the same time, 10,000m² seems unlikely to be a garden.
Incorrect	The summary contains information that either contradicts a statement from the original text (i.e the text mentioning the hotel is NOT pet-friendly, but the summary stating it is) or contains a severe error, such as using a wrong entity (e.g. place or a person), or a wrong number (for example confusion of different numbers or kilometers vs miles).	**Text:** *The closest major airports to Bomontist Suit are: Istanbul (SAW-Sabiha Gokcen Intl.) - 17.5 km / 10.9 mi Istanbul (IST-Ataturk Intl.).* **Summary:** Schiphol Airport is just a 15-minute drive from the hotel. **Explanation:** Schiphol Airport in Amsterdam is Incorrect, since the accommodation is clearly in Istanbul. In addition, 15-minute drive is Not Checkable in this context, because even though we know the distance, we don't know the expected speed of the journey.

Table 5: Definitions and Examples of Error Types