Do My Eyes Deceive Me?
A Survey of Human Evaluations of Hallucinations in NLG

Patrícia Schmidtová[1] **Eduardo Calò**[2] **Simone Balloccu**[3]
Dimitra Gkatzia[4] **Rudali Huidrom**[6] **Mateusz Lango**[1] **Fahime Same**[7]
Vilém Zouhar[5] **Saad Mahamood**[8] **Ondřej Dušek**[1]

[1]Charles University, Faculty of Mathematics and Physics [2]Utrecht University
[3]Technical University Darmstadt [4]Edinburgh Napier University [5]ETH Zürich
[6]ADAPT Research Centre, Dublin City University [7]trivago N.V. [8]Shopware

Abstract

Hallucinations are one of the most pressing challenges for large language models (LLMs). While numerous methods have been proposed to detect and mitigate them automatically, human evaluation continues to serve as the gold standard. However, these human evaluations of hallucinations show substantial variation in definitions, terminology, and evaluation practices. In this paper, we survey 64 studies involving human evaluation of hallucination published between 2019 and 2024, to investigate how hallucinations are currently defined and assessed. Our analysis reveals a lack of consistency in definitions and exposes several concerning methodological shortcomings. Crucial details, such as evaluation guidelines, user interface design, inter-annotator agreement metrics, and annotator demographics, are frequently under-reported or omitted altogether.

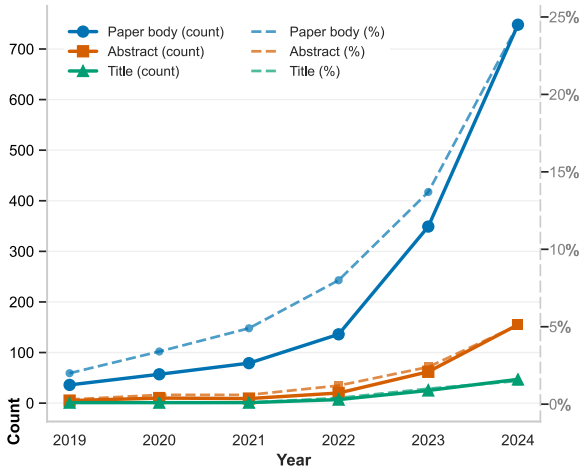

Figure 1: There is an exponentially growing trend of papers concerned with hallucination, both in absolute and in relative terms. The numeric values represented by this chart can be found in Table 2 (Appendix B).

1 Introduction

The popularity of large language models (LLMs) has led to an increase in human evaluations assessing the degree to which a model's outputs diverge from its inputs – in other words, the number of *hallucinations* or *confabulations* generated by the given language model. This is also reflected in the increased number of papers covering the topic (Figure 1). Human evaluations are commonly viewed as the more reliable way to evaluate natural language generation (NLG) systems (in contrast to, e.g., using automatic metrics).

Following on from recent NLP surveys that have looked at various human and automatic evaluation practices (Howcroft et al., 2020; van der Lee et al., 2021; Gehrmann et al., 2023; Balloccu et al., 2024; Schmidtova et al., 2024), this paper takes a more focused look at the challenge of evaluating the faithfulness of output from LLMs. We build on top of two earlier surveys that looked at hallucinations generally within NLG (Li et al., 2022; Ji et al., 2023). In contrast, we report in depth on how researchers are defining hallucination in their evaluations, as well as the current evaluation practices by looking at a broader set of papers over the past six years. Our goals are: (1) investigate and report on the current status quo of human evaluation for hallucinations, and (2) identify any shortcomings and recommend potential improvements. Our contributions in this paper are as follows:

- Based on a search over papers published in major NLP venues from 2019-2024, we identify 64 human evaluations of hallucination and extract key information on how evaluations are conducted (Section 3).
- We analyse our data and show the most common trends; we also conclude that important information is frequently not reported in the papers (Section 4).

Corresponding author: schmidtova@ufal.mff.cuni.cz

Proceedings of the 18th International Natural Language Generation Conference, pages 60–79
October 29–November 2, 2025.

- We discuss the main issues and make recommendations on how to address them and improve evaluation quality (Sections 5 and 6).

2 Related Work

Hallucinations The term *hallucination* is used in diverse and sometimes conflicting ways across the literature, making it difficult to assess model performance in a systematic way (Narayanan Venkit et al., 2024).

Recent works anchored in factuality frame hallucinations as content that fails to align with accepted truths. Ravichander et al. (2025) treat such outputs as content inconsistent with world knowledge or the user-provided context. Rawte et al. (2023b) highlight wholly fabricated or misleading details with respect to world knowledge. Tonmoy et al. (2024) describe the phenomenon broadly as the generation of ungrounded, factually erroneous text across varied domains. Luo et al. (2024) similarly emphasise outputs that appear correct but are not grounded in fact.

A complementary line of research emphasises faithfulness to the source input. Ji et al. (2023) characterise hallucination as text that diverges from the input, distinguishing intrinsic (contradicts the source) from extrinsic (unverifiable) cases, and contrasting the notions of faithfulness and factuality. Huang et al. (2025) extend this stance with a fine-grained taxonomy covering entity- and relation-level errors, factual fabrication, overclaims, and instruction, context, and logical inconsistencies.

Other studies focus on grounding in the model's own discourse. Zhang et al. (2023b) distinguish *input-*, *context-*, and *fact-conflicting* hallucinations, while Rawte et al. (2023a) differentiate *factual mirage* (distortions based on an otherwise correct prompt) from *silver lining* (elaborate narratives generated from an incorrect prompt). Huidrom and Belz (2023) similarly move away from external verification, framing hallucinations as meaning-level deviations where fluent outputs misinterpret or distort the intended content.

These definitions vary along three principal axes: (i) the grounding criterion (input context, external knowledge, or self-consistency), (ii) the verifiability standard (direct contradiction vs. unverifiability), and (iii) the granularity of error types (binary vs. multi-class taxonomies). This heterogeneity makes it difficult to achieve reproducible evaluation and impedes the development of comparable metrics. We discuss the prevalence of these different definition types in our surveyed papers in Sections 4 and 5.

Meta-Evaluations in NLG Over the past two decades, the NLG community has increasingly recognised inconsistencies in the evaluation of generated text. Howcroft et al. (2020) analysed 165 NLG papers employing human evaluation, documented the diversity of quality dimensions, and introduced standardised evaluation sheets and definitions to enhance consistency. Additionally, Belz et al. (2020) proposed an 18-property classification for human evaluation methods in NLG to support comparability, meta-evaluation, and reproducibility testing. Gehrmann et al. (2023) subsequently reviewed two decades of human and automatic evaluation practices, assessed the extent to which 66 contemporary studies adhered to recommended guidelines, and proposed concrete reporting standards and template evaluation reports to strengthen methodological rigour. Addressing the quality of the studies, Ruan et al. (2024) found that only 30% of NLG papers release human evaluation guidelines, and 77% of those contain vulnerabilities. We present similar findings, with a more specific focus on human evaluation of hallucinations, offering a deeper analysis within this scope.

Surveys on Human Evaluation of Hallucinations Although hallucination in NLG has been the subject of numerous surveys (Zhang et al., 2023b; Rawte et al., 2023b; Sahoo et al., 2024; Agrawal et al., 2024; Tonmoy et al., 2024; Huang et al., 2025; Bai et al., 2025), the role of human evaluation is rarely explored in depth. Many surveys either omit this aspect entirely or only acknowledge that human evaluation remains the most reliable and commonly used method for assessing hallucinations (Zhang et al., 2023b; Huang et al., 2025). At the same time, reliable hallucination evaluation is often cited as an open research problem (Zhang et al., 2023b; Ji et al., 2023; Bai et al., 2025). A recent survey on automatic hallucination evaluation methods (Qi et al., 2025) underscores the need for unified annotation guidelines and stresses the importance of annotators possessing relevant domain expertise and proper training in evaluation criteria to obtain reliable human evaluation.

Only two surveys provide a more detailed discussion of human evaluation of hallucinations. Ji et al. (2023) identify two main types of hallucination annotation: scoring individual texts and comparing

multiple texts (e.g., against baselines or ground-truth references). A more in-depth analysis is presented by Li et al. (2022), who highlight the challenges associated with human evaluation, including the low inter-annotator agreement (IAA) reported in related studies. They also suggest ranking-based Best-Worst Scaling (Tang et al., 2022) as a more effective annotation framework for hallucination assessment. Nonetheless, this survey discussed only three works on human evaluation, all of which predate the introduction of LLMs. Our survey provides an extension of these results, in both depth and scope.

3 Methodology

3.1 Paper Selection

We considered papers published between 2019 and 2024 at the following conferences: ACL, NAACL, EACL, AACL, EMNLP, IJCNLP, and INLG. We also included papers from two journals: CL and TACL. 2019 was selected as the lower bound because it is the year when GPT-2 (Radford et al., 2019) was released, marking the beginning of the popularity of pre-trained Transformer language models in NLG.

In total, 12,418 papers from the selected venues and time period were automatically screened for the mention of terms 'hallucination' or 'confabulation' as well as mentions of 'human evaluation' or 'manual/qualitative analysis'.[1] 1,405 (11.3%) papers mentioned 'hallucination', 3,552 (28.6%) mentioned 'human evaluation', and 731 (5.9%) mentioned both. This means that 52% of papers concerned with hallucination also mention human evaluation. We ranked these 731 papers by the occurrence frequency of the terms of interest, prioritising those that mention either term in the abstract.

Then, we manually scanned the top 150 papers to confirm their relevance to the survey. A paper was considered relevant if it performed a human evaluation of hallucinations specifically. Moreover, we decided to limit our scope to text-only generation tasks, including structured input data in textual format (such as semantic triples or tables), excluding studies that evaluated multimodal tasks.

Applying these criteria led to the selection of 67 papers for the survey. Throughout the inspection of the surveyed papers, we found that six papers leveraged previously collected datasets with human annotations; thus, data collection was not described. We excluded those, resulting in 61 surveyed papers, performing 64 distinct human evaluations. Three papers performed two separate human evaluations, differing in annotation type (n=2) and the definition of hallucination and guidelines (n=1). The full list of annotated papers is in Appendix E.

3.2 Annotation Approach

Each of the surveyed papers was assigned to one of the authors of this paper,[2] who read the surveyed paper and annotated key information about the human evaluation of hallucinations performed in the paper (or lack thereof): **definition of hallucination**, **annotation type** (e.g., categorisation, Likert-scale), the NLP task in question, availability of annotated data, annotator demographics (including **number of annotators**, **annotator type/identity**, and any required **specific skills**), annotator **compensation**, annotation **quality assurance measures**, **inter-annotator agreement** (IAA) details, the annotation **user interface** used, and annotation **guidelines**. Table 1 in Appendix A includes the full list of the annotated features and their descriptions.

The majority of the attributes were identified at the beginning of the project as factors that could influence the quality and reliability of a human evaluation. For all attributes, our own annotation guidelines contained the description of the attribute and examples of values that could appear. For instance, under the attribute "who were the annotators?", possible categories included authors; PhD/Masters/Bachelors students; in-house, paid; in-house, volunteers; participants recruited through Prolific or Amazon MT; and other.

We noted even the most vague statements related to a given attribute, and any borderline cases were documented as comments and discussed during subsequent meetings to ensure consistent annotation. Due to the time-intensive nature of annotating the cohort of papers, no experiments with inter-annotator agreement were performed. Nevertheless, during the post-processing phase, two of the authors reconfirmed all annotations, especially focusing on the papers that failed to provide the majority of attributes.

[1] The search was performed using regular expressions and allowed for changes in form such as plurals or derivation.

[2] All authors are NLP researchers with at least three years of experience.

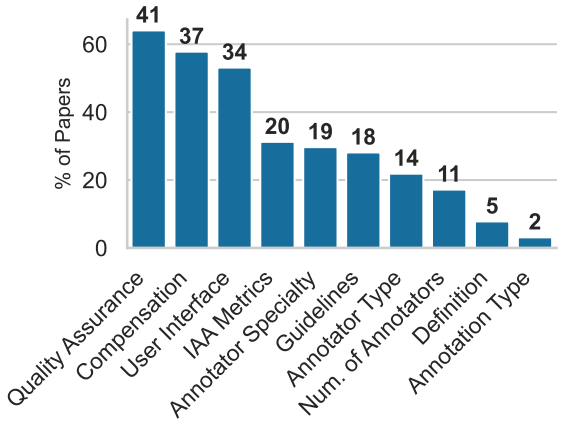

Figure 2: Percentage of papers lacking information, divided by key attributes outlined in Section 3 (absolute counts shown above each bar).

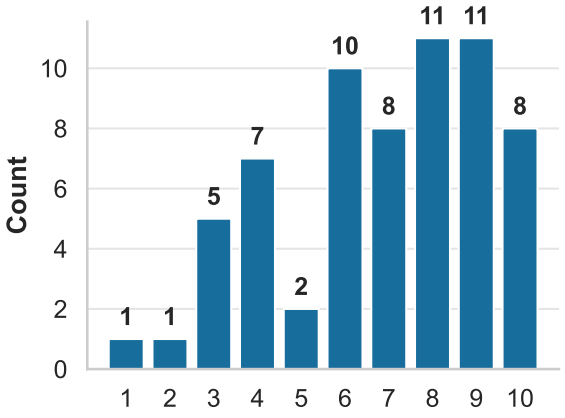

Figure 3: Amount of reported information as specified in Section 3.2. 10 (right-most column) is considered the best, where authors shared every necessary detail on their evaluation.

4 Results

High-Level Statistics The results presented in this section are based on the 64 human evaluations we reviewed. Most papers came from EMNLP (n=26), followed by ACL (n=17), INLG (n=8), NAACL (n=7), EACL (n=4), and TACL (n=2). The majority of papers were published in 2024 (n=27), then 2023 (n=22), 2022 (n=11), and 2020 and 2021 were both represented by two papers. Summarisation was the most frequent task (n=31), followed by data-to-text generation (n=9), dialogue response generation and question answering (n=6 for both).

4.1 Missing Information

Figure 2 shows that a large number of the papers surveyed did not report key information. As we expected, most papers provided a definition of hallucination and described the annotation types used in their human evaluation. However, interestingly, five papers did not include a definition, and two did not specify the annotation type.

The situation gets worse with annotation details: annotator compensation (relevant for ethical reasons) and basic experimental and methodological information (e.g., the IAA metrics used, if any, and the guidelines provided to annotators) are often not reported. Furthermore, over 60% of the papers did not specify whether, and how, they implemented quality assurance methods to ensure the reliability of the collected annotations. Quality assurance includes topics such as whether the authors included a training phase, calibration, comprehension checks, and piloting in their experiment. See Appendix D for additional charts.

Completeness If we consider the missing information from the point of view of completeness, i.e., how many papers provide all key information, or its subset of a given size (considering the 10 key attributes set in bold font in Section 3.2), the situation looks somewhat less problematic. A large portion of papers report the majority of key information, as shown in Figure 3. However, only 8 papers provide all 10 key attributes. Notably, only one paper adopted a standardized reporting, which consisted of the human evaluation datasheet (HEDS; Shimorina and Belz, 2022) that aims to standardise reporting practices in human evaluations to ensure clarity and reproducibility. 14 human evaluations were poorly reported, with less than half of the key attributes mentioned.

More Prestige $\neq$ More Rigour We observed a concerning trend where papers published at the top-ranked conferences – ACL and EMNLP – more frequently omit key information (see Figure 4). This is particularly surprising given that multiple of the attributes we considered (annotator demographics, compensation, UI, and guidelines) are specifically requested by the Responsible NLP Checklist (Dodge et al., 2019; Rogers et al., 2021), which has been incorporated into reviewer guidelines or a mandatory part of every ACL Rolling Review (ARR) submission since NAACL 2022.

To see how this checklist is being honored, we filtered out papers published at venues where this checklist was in place, totalling 47 papers. Section

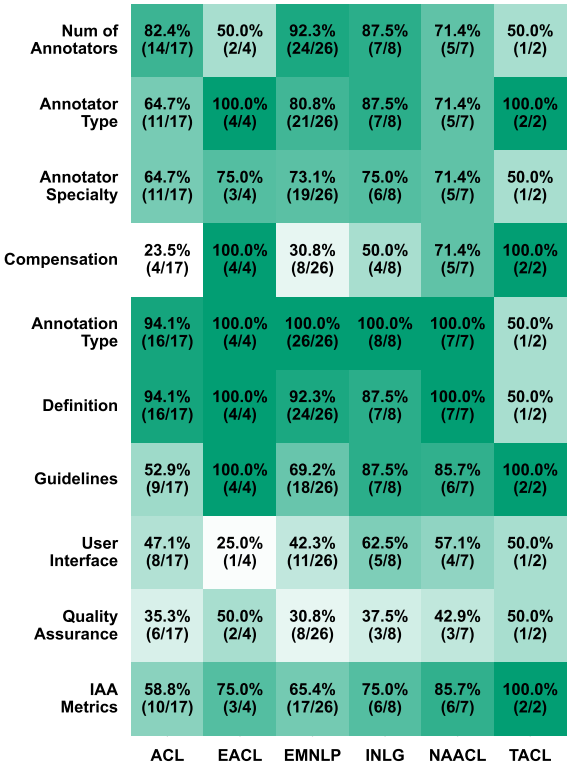

Figure 4: The proportion of papers from a given venue (X axis) that specify a given attribute (Y axis).

D of the Responsible NLP Checklist is concerned with reporting practices around human evaluation. Notably, Question D1 asks for "full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?", which maps to our criteria of providing guidelines and information on the user interface. Guidelines were not provided in 14 (30%) papers, and user interface details in 30 (64%) papers that should have adhered to the checklist. Question D2 is about how people were recruited (the identity of annotators) and how they were paid (compensation). This information was not mentioned by 13 (28%) and 31 (66%) papers, respectively.

4.2 Hallucination Definitions

We focused on the grounding criterion used to define hallucination, and the granularity of the definitions in our analysis (cf. Section 2). As we discuss in Section 5, verifiability standards (contradictions vs. non-verifiability) were generally vague.

Grounding Criterion In Figure 5, we examine the grounding source depending on the task. Faithfulness to the source input is by far the most commonly used criterion to assess the presence of hallucinations in generated outputs across all tasks.

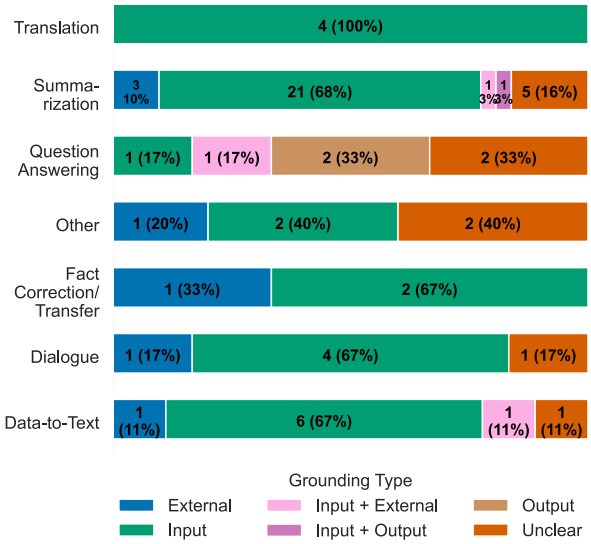

Figure 5: Distribution of grounding sources across the most prominent tasks.

Only seven papers evaluated outputs against external knowledge, suggesting that factuality plays a relatively minor role in current evaluation practices. This scarcity may be partly attributable to the type of generation task: for tasks such as data-to-text generation, summarisation, or translation, the standard practice is to compare the output with the input text rather than with external knowledge. Conversely, this does not hold for question answering, where the grounding criteria were the most diverse. For 11 papers, it is unclear which source was used to verify the outputs' veracity.

Granularity of Error Types The granularity of hallucination error types varied across papers. Most papers (24) treated hallucination as a singular phenomenon, providing only one definition or category. This was followed by multi-class approaches (20 papers), where hallucination was represented using multiple distinct types (e.g., severity scales). 12 papers adopted a binary classification, distinguishing between two possible outcomes (e.g., hallucinated vs. non-hallucinated). Finally, 11 papers were unclear about the granularity used.

4.3 Annotation Details

Annotation Type Figure 6 shows that categorisation (e.g., binary labels indicating whether a hallucination is present, or multi-class labels) is by far the most commonly used type of hallucination annotation. It is followed, at a distance, by Likert scale and span-based annotations, which appear in almost equal proportions.

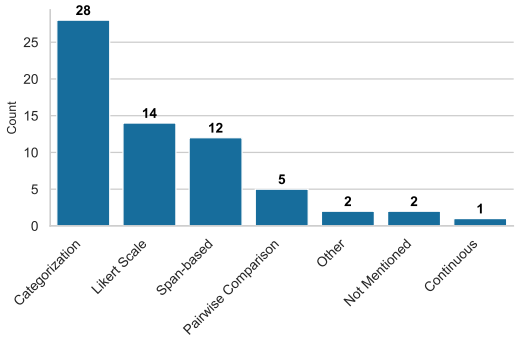

Figure 6: Number of papers per type of hallucination annotation used.

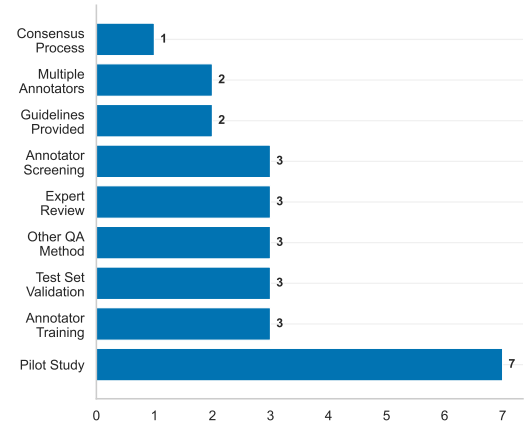

Figure 7: Number of studies containing quality assurance methods. One paper can use multiple methods.

Annotation Guidelines As identifying and categorising hallucinations is a complex task, clear and comprehensive annotation guidelines are critical for annotation. We found that guidelines were presented in 46 instances. Other key components, such as annotator instructions (present in only 25 cases) and contextual information (present in only 14 cases), were often omitted. Particularly concerning is the rarity of examples, critical in clarifying the tasks, which were provided in only six cases.

User Interface Information on details of user interfaces used during annotation is rarely reported in the reviewed papers. Less than 10 papers provided a screenshot, and 3 papers mentioned using Google Forms. However, the vast majority either described the user interface very vaguely, or did not comment on it at all. This confirms the broader trend identified by Calò et al. (2025) of overlooking this important factor in NLP evaluation.

Quality Assurance Quality assurance (QA) concerns the measures researchers take to ensure that human evaluation experiments produce reliable and consistent results (Belz et al., 2024). Common practices include annotator training, calibration, piloting, and providing guidelines and examples in the experiment. Of the 64 studies reviewed, 22 (34.4%) explicitly report their quality assurance strategy, with piloting being the most common method. Despite the importance of quality assurance, only five studies have reported the use of multiple methods. Figure 7 shows the different QA methods examined in this paper.

IAA Details The situation with reporting IAA information is concerning. 27 papers did not report any IAA. Of these, one study used only one annotator per example and therefore could not report the inter-annotator agreement. In contrast, 20 papers reported high IAA, 15 reported medium IAA, and 2 reported low IAA.[3] The small number of papers reporting low IAA may be because studies with poor agreement opted not to report it. Of the 27 papers that failed to report IAA, there are 8 for which it was unclear if this was even possible: one used only a single annotator, making IAA measurement impossible, while the other seven did not mention how many annotators they used.

4.4 Annotator Details

Compensation Almost 60% of papers (37 in total) do not report the compensation provided to the annotators. Among those that do, various papers only provide vague statements. Some cite compensation as "above the minimum wage", "fair payment according to our organisation's standards", and "a competitive hourly rate that is benchmarked against similar roles in the US".

Annotator Group Figure 8 shows that students and crowd workers are tied as the most commonly used annotator groups, followed closely by experts. Surprisingly, 12 papers did not mention who the annotators were. Of the 15 papers that reported using students as annotators, 11 gave no details on compensation, one relied on voluntary participation, and only three specified an hourly rate.

Specific Annotator Skills Figure 9 categorises annotators according to the skills or qualifications required for each experiment. Notably, 19 papers do not specify any qualifications, whereas 16 call

[3]The classification of IAA into high, medium, and low was done by the authors of this survey based on commonly considered thresholds.

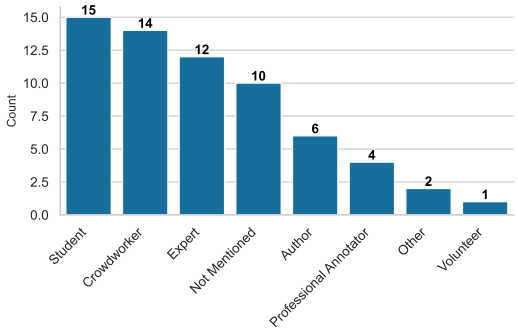

Figure 8: Number of papers per annotator group.

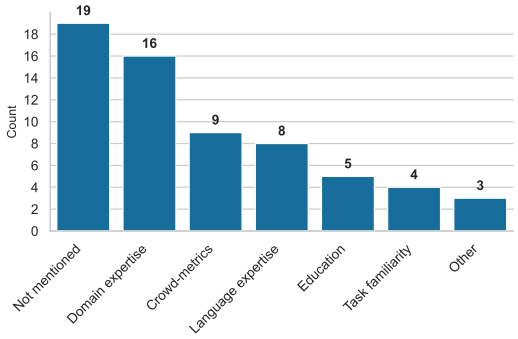

Figure 9: Distribution of annotator skills required by authors (note that more than one skill may be required for a given paper).

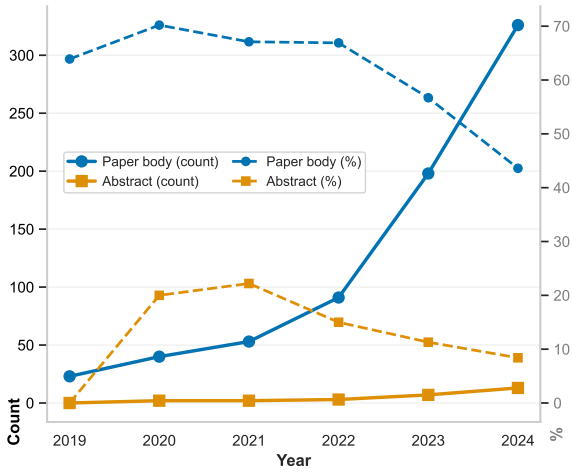

Figure 10: The count of hallucination papers that use human evaluation is growing (solid blue line). However, their relative proportion is falling, i.e., a larger % of papers do not consider human evaluation (dashed blue line). We observe a similar trend in the abstracts (yellow), where a lower proportion of authors deem human evaluation important enough to mention. This information is also available as Table 3.

for domain expertise (e.g., computer science or linguistics). The remaining studies apply criteria such as crowd-platform metrics (e.g., HIT approval rate), language proficiency, formal university education, task familiarity, or other bespoke requirements.

5 Discussion

Human Evaluation Shows Absolute Growth, Relative Decline We analyzed 1,405 papers that mention hallucinations, tracking over time how many also discuss human evaluation, both in absolute numbers and relative proportions. Figure 10 shows that while the absolute number of hallucination papers mentioning human evaluation continues to grow, the proportion of such papers is declining dramatically. This decline is indicative of a rapid overall increase in hallucination research outpacing the use of human evaluation methods.

We offer our hypothesis of possible factors that could have contributed to this trend. First, running a properly designed human evaluation requires a considerable amount of effort (Thomson et al., 2024). As the absence of human insight into the model's errors is becoming standard, with less than 50% of hallucination papers providing it, many authors have very little incentive to undertake this effort.

Second, LLM-as-a-judge evaluations (Bavaresco et al., 2025; Kasner et al., 2025) have emerged as an alternative to human evaluation due to their lower cost. This approach, made popular in machine translation evaluation (Kocmi and Federmann, 2023), has grown in the hallucination literature from just one paper in 2023 to 40 papers (5% of hallucination papers) in 2024. While we cannot establish a direct causal relationship, this growth coincides with the relative decline in human evaluation usage.

Third, researchers may increasingly substitute benchmark evaluation for human evaluation, viewing automated metrics on standard datasets as sufficient. We discuss why this practice is problematic in our next point. Robust human evaluation methodologies matter as in the end "a careful and well-designed human evaluation is usually the best way to meaningfully evaluate an NLG system" (Reiter, 2024).

The Continued Importance of Human Evaluation Despite trends toward automated evaluation, human judgment remains essential for hallucination assessment. Human evaluation provides irreplaceable insights that automated approaches cannot capture: humans can apply specialized do-

main expertise (e.g., medical, legal knowledge), assess contextual appropriateness and cultural sensitivity, and evaluate real-world applicability from an end-user perspective (Reiter, 2025). Moreover, human evaluation often reveals novel failure modes and error patterns not captured by predefined metrics, which is crucial for understanding model limitations and improving systems. The persistent misalignment between automatic metrics and human judgment (Belz and Reiter, 2006; Novikova et al., 2017) has intensified as evaluation tasks have grown more complex – hallucinations require nuanced assessment that goes beyond simpler, previously evaluated factors (Howcroft et al., 2020). Current automated alternatives suffer from significant limitations: benchmarks are plagued by LLM training data contamination (Balloccu et al., 2024; Golchin and Surdeanu, 2025), dataset errors (Gema et al., 2025), and poor real-world applicability (Hardy et al., 2025; Lunardi et al., 2025), while LLM-as-a-judge approaches require human validation for new datasets or tasks (Schmidtová et al., 2025) and may inherit training data biases. Finally, developing better automated metrics itself requires human-annotated gold standards, and certain domains demand human validation for safety or regulatory compliance.

Hallucination Definition Hallucination definitions vary greatly in the papers we reviewed. Some authors used standardised definitions previously published in the literature, while others grounded their definitions in the specific context of the task, such as medical decision-making. For instance,[4] one paper defined hallucination in the context of clinical safety as:

> *factual accuracy, specifically looking for missing or incorrect information that could lead to errors in medical treatment after discharge*

However, we also found vague or difficult-to-interpret definitions, such as:

> *Hallucinations - 0: no stuff that is not factual. - 1: even if there is one stuff that is not correct, gibberish also gets this*

Lack of Reporting and Release of Research Data From the papers that we have surveyed, there is a significant gap in the number of details reported by researchers. In particular, details such as remuneration details, experimental details, IAA metrics, and the guidelines used. Efforts have been made to encourage researchers to fill in standardised human evaluation reporting sheets, such as HEDS (Shimorina and Belz, 2022), to greater evaluation reproducibility. However, our results indicate this is far from standard practice, with only a single paper filling in such a datasheet.

There is also an additional need for researchers to be more proactive in releasing research data. Only 20 papers (43%) in our survey have actually released any annotation data. Model outputs with error annotations would not only be useful for further error analysis, but also for the development and improvement of new automatic evaluation methods, such as COMET (Rei et al., 2020) was developed for machine translation.

The issues observed in our results are not new issues. The same lack of reporting was observed by Howcroft et al. (2020) in their survey of NLG human evaluations. However, it is disappointing, but not surprising, that no progress on this front has been made over the past five years.

Are Responsible NLP Checklists Used Responsibly? Our findings reveal a troubling inconsistency between the formal requirements of the Responsible NLP Checklist and actual reporting practices in papers published at premier NLP venues. Despite the checklist being integrated into reviewer guidelines and made mandatory for submissions to the ACL Rolling Review since NAACL 2022, a significant proportion of papers failed to report critical information about human evaluation practices. This pattern is surprising when contrasted with papers from INLG, a venue not formally bound by the checklist, which nonetheless reported these attributes more consistently. This discrepancy raises concerns about the seriousness with which authors and reviewers are taking the Responsible NLP checklist. It suggests that mere formal inclusion of reporting guidelines may not be sufficient; instead, stronger enforcement via higher peer review quality may be necessary to ensure broader compliance and more transparent reporting of information. The checklist itself may need simplification in order to increase its practical adoption.

Quality Assurance in Human Annotations Obtaining high-quality and consistent annotations is challenging, especially when annotators must look for divergences between input data and model outputs. For example, Thomson et al. (2023), in their methodology for evaluating the accuracy of data-to-

[4] See more examples of hallucination definitions in Table 5, Appendix C.

text systems, recommends the use of pilot studies to develop intuitive error categories for error-span annotations. Additionally, to ensure high agreement between recruited participants, careful design of the qualification tasks is needed to filter out subpar annotators (Zhang et al., 2023a).

From the results above, there is clear underreporting of key information in the majority of the surveyed papers, especially with respect to quality assurance information. This makes comparison between different papers challenging and prevents researchers from understanding whether a given set of results has methodological gaps or not. Furthermore, sharing lessons from quality assurance methods can enhance the reliability and reproducibility of future evaluations by setting a standard.

Reporting Inter-Annotator Agreement Although highly informative, IAA is often not reported. One possible reason is that peer review may discourage authors from including low IAA scores. However, low IAA does not necessarily indicate poor annotation quality (especially when quality assurance steps, such as piloting and attention or comprehension checks, have been taken). Instead, it may reflect the inherent subjectivity of the task, including factors such as ambiguity, implicit assumptions, or the difficulty of certain items (Plank, 2022). We argue that all of this information is valuable and worth reporting. As Reiter et al. (2003) already emphasised more than two decades ago, sharing negative results is important for scientific research; however, little progress has been made in this regard.

6 Recommendations

Defining Clear Hallucination Definitions Authors should use standardised definitions from existing literature whenever possible to promote consistency and facilitate comparison. Definitions must be clear, unambiguous, and should include concrete examples. In order to avoid ambiguity, authors should explicitly specify the grounding criterion, verifiability standard, and granularity of error types, as discussed in Section 2. When introducing new definitions, authors should justify why existing definitions are inadequate and pilot them with annotators to confirm understanding.

Inter-Annotator Agreement Reporting Authors should consistently report IAA results from their human evaluations, as this information is crucial for assessing the reliability and reproducibility of findings. In addition to reporting IAA scores, authors should specify the quality assurance steps taken, in order to provide proper context for interpreting these results. We urge reviewers to avoid rejecting papers with low IAA when proper quality assurance measures were implemented and the agreement is adequately addressed. Some tasks are inherently subjective, and using low IAA as a "rejection shortcut" reduces transparency and loses valuable insights about task difficulty and subjectivity that benefit the broader research community.

Use Evaluation Reporting Sheets Standardised evaluation reporting sheets, such as HEDS (Shimorina and Belz, 2022), allow for evaluation comparability and reproducibility by ensuring that all relevant evaluations are recorded. While such reporting sheets may seem overwhelming at first, they can help practitioners to better understand what details should be reported. At the very least, we urge authors to read through them to understand which information should always be reported.

7 Conclusion

Our survey of 64 human evaluation studies uncovered key insights across multiple aspects of how LLMs' hallucinations are assessed. Although human evaluation remains the gold standard for NLG researchers (Zhou et al., 2022), we observed a concerning decline in the proportion of papers conducting such evaluations. Moreover, methodological reporting is often lacking; critical details, such as inter-annotator agreement, annotator demographics, and annotation guidelines, are frequently omitted. Definitions and categorisations of hallucinations vary widely across tasks and papers. We argue that adopting standardised definitions addressing the three axes we propose in Section 2 would support a more unified understanding of hallucinations. Finally, it is troubling that even papers published at top NLP venues often fail to report essential information about their human evaluation procedures, despite being prompted to include the Responsible NLP Checklist. These findings underscore the urgent need for more rigorous, transparent, and standardised practices in human evaluations of hallucinations.

Limitations

In this survey, we only looked at papers published in well-known NLP conferences and journals over

the past six years. This means that it is possible that there are earlier papers with human evaluations that assess LLMs' for hallucination that may have been excluded from our analysis. Additionally, we only annotated papers in English and did not include papers that may have been published in other languages.

Ethics Statement

The focus of this work is to gain better insights into how human evaluations of hallucinations are performed. The annotations made in this paper were made by the authors and therefore, we did not recruit any external annotators nor process any personal data. We utilized a colourblind-friendly palette to improve the accessibility of our paper.

Supplementary Materials Availability Statement: Our annotation of the papers and the analysis scripts can be found on GitHub at: https://github.com/patuchen/human_eval_of_hallucinations.

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).

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can knowledge graphs reduce hallucinations in LLMs? : A survey. 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 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

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.

Anya Belz, Simon Mille, and David M Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. Association for Computational Linguistics (ACL).

Anya Belz, João Sedoc, Craig Thomson, Simon Mille, and Rudali Huidrom. 2024. The INLG 2024 tutorial on human evaluation of NLP system quality: Background, overall aims, and summaries of taught units. In *Proceedings of the 17th International Natural Language Generation Conference: Tutorial Abstract*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

Eduardo Calò, Lydia Penkert, and Saad Mahamood. 2025. Lessons from a user experience evaluation of NLP interfaces. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2915–2929, Albuquerque, New Mexico. Association for Computational Linguistics.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. 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.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Shahriar Golchin and Mihai Surdeanu. 2025. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel John Kochenderfer. 2025. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1032–1047, New York, NY, USA. Association for Computing Machinery.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Rudali Huidrom and Anja Belz. 2023. Towards a consensus taxonomy for annotating errors in automatically generated text. In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 527–540.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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 and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *Preprint*, arXiv:2203.05227.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. On robustness and reliability of benchmark-based evaluation of llms. *Preprint*, arXiv:2509.04013.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An audit on the perspectives and challenges of hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6528–6548, Miami, Florida, USA. 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.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siya Qi, Lin Gui, Yulan He, and Zheng Yuan. 2025. A survey of automatic hallucination evaluation on natural language generation. *Preprint*, arXiv:2404.12041.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM_Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023a. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ehud Reiter. 2024. *Natural Language Generation*. Springer Nature.

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

Ehud Reiter, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable NLG evaluation. 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 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

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, Ondřej Dušek, and Saad Mahamood. 2025. Real-world summarization: When evaluation reaches its limits. *Preprint*, arXiv:2507.11508.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):795–805.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

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

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023a. 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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 314–324, Seattle, United States. Association for Computational Linguistics.

A Annotated Features

In Table 1, we provide a detailed list of all features that were recorded for the surveyed papers.

B Yearly Trend Tables and Charts

In this section, we provide additional charts and tables. Data supporting the yearly trend charts can be found in Table 2 (trend of hallucination mentions in papers) and Table 4 (human evaluation and LLM-as-a-judge trends in hallucination papers). For completeness, analogous data for papers mentioning human evaluation is shown in Table 3 and Figure 11, and shows that while the absolute number of human evaluations started growing in 2022, this trend is not reflected when related to the number of papers considered.

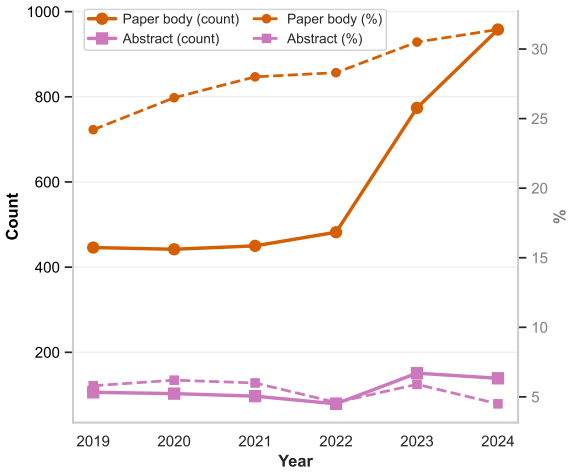

Figure 11: Yearly trend of human evaluation in papers from 2019-2024.

C Examples of Hallucination Definitions

In Table 5 we present examples of definitions found in the surveyed papers.

D Information Frequently Reported Together

Figure 12 reveals patterns in how information is reported or omitted.

Feature	Description
Number of annotators	The total number of individuals who conducted the annotations.
Identity of annotators	The group or affiliation of those who performed annotations (e.g., authors, students from the authors' lab, in-house employees, volunteers, Prolific workers, Amazon MTurk workers).
Specific quality of annotators	Additional, more specific qualifications of annotators, such as familiarity with the task, domain expertise, pre-filtered conditions on crowdworkers, native-speaker status, or residence in a specific region.
Compensation	The form and amount of payment provided to annotators, including currency and whether paid per task, per hour, or per HIT.
Annotation type	The kind of annotations collected, including word-level (e.g., span or span + category) or text-level (e.g., Likert scale, categorization, continuous score) annotations.
Task type	Type of the task addressed in the article, specifying whether it was data-to-text or text-to-text generation.
Was input or output annotated?	Indicates whether annotations were made on the input, the output, or both.
Task	The specific generation task of the study (e.g., summarization, machine translation, question answering, dialogue generation, data-to-text generation, style transfer, error correction).
Definition	Precise definition of hallucination presented in the paper.
Guidelines	Whether guidelines are available and, if so, which format they take (e.g., free text, tutorial, in-person briefing).
User Interface	The platform or tool used for annotation (e.g., Google Forms, Microsoft Forms, Label-Studio, Argilla, or a custom platform), or "NM" if not mentioned.
Quality Assurance	Whether the authors mention measures like training, calibration, comprehension checks, piloting, or golden label acquisition phase.
Is data available?	Whether the annotated data is publicly available.
IAA Metrics	The metrics used to assess the quality of annotations (e.g., inter-annotator agreement, accuracy, entropy).
Kappa Score (Cohen or Fleiss)	The reported score, if Kappa was used.
Krippendorff's Alpha Score	The reported score, if Krippendorff's Alpha was used.
Other IAA measure value	Mention any other IAA measure used.
Overall IAA assessment	A summary rating from the reported inter-annotator agreement. Low: 0–0.35, Medium: 0.36–0.60, High: 0.61–1.

Table 1: Description of the annotated features in the surveyed papers.

Year	Total Papers	Paper Body #	Abstract #	Title #	Paper Body %	Abstract %	Title %
2019	1841	36	5	1	2.0%	0.3%	0.1%
2020	1671	57	10	1	3.4%	0.6%	0.1%
2021	1606	79	9	1	4.9%	0.6%	0.1%
2022	1706	136	20	7	8.0%	1.2%	0.4%
2023	2539	349	62	25	13.7%	2.4%	1.0%
2024	3055	748	155	47	24.5%	5.1%	1.5%

Table 2: Evolution of hallucination mentions in academic papers from 2019-2024. Shows both absolute counts and percentages of papers mentioning hallucinations in PDF content, abstracts, and titles.

Year	Total Papers	Paper Body #	Abstract #	Title #	Paper body %	Abstract %	Title %
2019	1841	446	106	4	24.2%	5.8%	0.2%
2020	1671	442	103	3	26.5%	6.2%	0.2%
2021	1606	450	97	7	28.0%	6.0%	0.4%
2022	1706	482	79	4	28.3%	4.6%	0.2%
2023	2539	774	151	5	30.5%	5.9%	0.2%
2024	3055	958	139	5	31.4%	4.5%	0.2%

Table 3: Evolution of human evaluation mentions in academic papers from 2019-2024. Shows both absolute counts and percentages of papers mentioning human evaluation in PDF content, abstracts, and titles.

Year	Papers	HumEval	HumEval %	LLM Judge	LLM %
2019	36	23	63.9%	0	0.0%
2020	57	40	70.2%	0	0.0%
2021	79	53	67.1%	0	0.0%
2022	136	91	66.9%	0	0.0%
2023	349	198	56.7%	1	0.3%
2024	748	326	43.6%	40	5.3%

Table 4: Evolution of the use of human evaluation and LLM-as-a-judge in hallucination papers.

Definition	Comment
Hallucinations are cases in which the model generates output that is partially or completely unrelated to the source sentence, while omissions are translations that do not include some of the input information (Dale et al., 2023).	Use of a definition previously published in the literature.
Factual accuracy, specifically looking for missing or incorrect information that could lead to errors in medical treatment after discharge.	Definition grounded in the concrete task at hand (i.e., medical treatment after discharge).
Hallucinations - 0: no stuff that is not factual. - 1: even if there is one stuff that is not correct, gibberish also gets this.	Extremely unclear and vague definition.
The output text contains word span(s) for which there is no corresponding part of the input that they render. In other words, some content that is not present in the input and should not be rendered in the output is nevertheless rendered by some word span(s) in the output. Moreover, there is no content in the input that the word span(s) are intended to render, but render wrongly. i.e. this type of error can be fixed by removing something from the output.	Clear and extensive definition, giving specific details on how to handle various cases that can occur in the annotation.
Major errors: Readers knowledgeable in the space would likely recognise the error in the blue sentence. If printed in a newspaper, the newspaper would have to print a correction or retraction to maintain its reputation. Minor errors: Most readers would not notice the error or find it less important. If printed in a newspaper, the newspaper may not need to print a correction.	Interesting definition, rooting the different categories in something the annotators should be familiar with (i.e., reading newspapers).

Table 5: Selected definitions of hallucination from the surveyed papers.

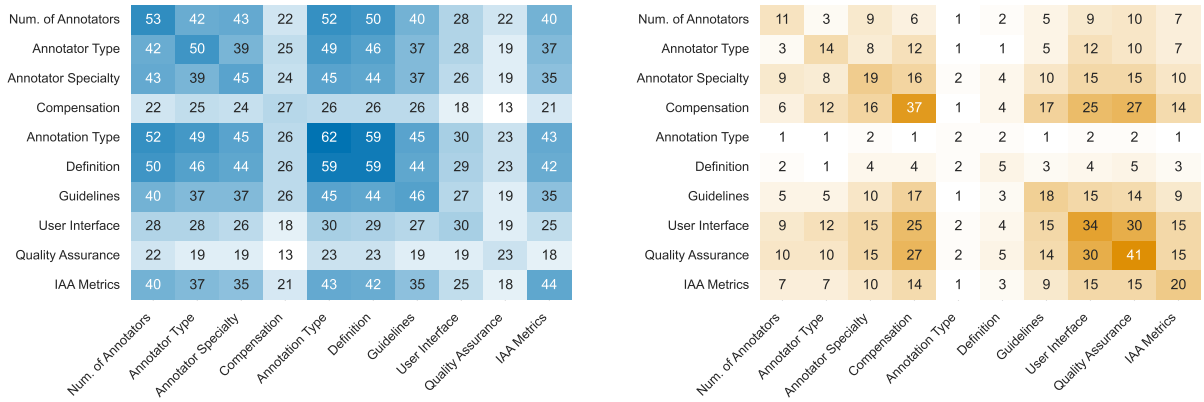

Figure 12: Co-occurrence of key information reported (left) or omitted (right). Please note that the values in the two charts do not have to sum up to the total amount of evaluations, because they do not account for cases when a paper reports exactly one of two given attributes.

E Full List of Papers Reviewed

In this section, we list all the work we reviewed and classified as relevant for our survey.

Vidhisha Balachandran, Hannaneh Hajishirzi, William W. Cohen, and Yulia Tsvetkov. 2022. Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nishant Balepur, Jie Huang, and Kevin Chen-Chuan Chang. 2023a. Expository Text Generation: Imitate, Retrieve, Paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Nishant Balepur, Jie Huang, and Kevin Chen-Chuan Chang. 2023b. Text Fact Transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4745–4764, Singapore. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S. Chao. 2023. Can LMs Generalize to Future Data? An Empirical Analysis on Text Summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16205–16217, Singapore. Association for Computational Linguistics.

Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. 2023. Fidelity-Enriched Contrastive Search: Reconciling the Faithfulness-Diversity Trade-Off in Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 843–851, Singapore. Association for Computational Linguistics.

Yongrui Chen, Haiyun Jiang, Xinting Huang, Shuming Shi, and Guilin Qi. 2024. DoG-Instruct: Towards Premium Instruction-Tuning Data via Text-Grounded Instruction Wrapping. 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 4125–4135, Mexico City, Mexico. Association for Computational Linguistics.

Jaepill Choi, Kyubyung Chae, Jiwoo Song, Yohan Jo, and Taesup Kim. 2024. Model-based Preference Optimization in Abstractive Summarization without Human Feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18837–18851, Miami, Florida, USA. Association for Computational Linguistics.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2023. Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673, Miami, Florida, USA. Association for Computational Linguistics.

Angela Fan and Claire Gardent. 2022a. Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

Nigel Fernandez, Alexander Scarlatos, and Andrew Lan. 2024. SYLLABUSQA: A Course Logistics Question Answering Dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10344–10369, Bangkok, Thailand. Association for Computational Linguistics.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2022. Evaluating Legal Accuracy of Neural Generators on the Generation of Criminal Court Dockets Description. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 73–99, Waterville, Maine, USA. Association for Computational Linguistics.

Nicolas Garneau and Luc Lamontagne. 2023. Guided Beam Search to Improve Generalization in Low-Resource Data-to-Text Generation. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 1–14, Prague, Czechia. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and Modeling Fine-grained Factuality in Summarization.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language Models Hallucinate, but May Excel at Fact Verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 1500–1517.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Knowledge-Centric Hallucination Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA. Association for Computational Linguistics.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. 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.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What Have We Achieved on Text Summarization?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Rudali Huidrom, Anya Belz, and Michela Lorandi. 2024. Differences in Semantic Errors Made by Different Types of Data-to-text Systems. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 609–621, Tokyo, Japan. Association for Computational Linguistics.

Saad Obaid ul Islam, Iza Škrjanec, Ondrej Dušek, and Vera Demberg. 2023. Tackling Hallucinations in Neural Chart Summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 414–423, Prague, Czechia. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

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, Prague, Czechia. Association for Computational Linguistics.

Mateusz Lango and Ondrej Dušek. 2023. Critic-Driven Decoding for Mitigating Hallucinations in Data-to-text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2853–2862, Singapore. Association for Computational Linguistics.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity Calibration for Opinion Summarization. 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 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

Yanyang Li, Jianqiao Zhao, Michael R. Lyu, and Liwei Wang. 2022. Eliciting Knowledge from Large Pre-Trained Models for Unsupervised Knowledge-Grounded Conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10551–10564, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Laura Mascarell, Ribin Chalumattu, and Julien Heitmann. 2023. Entropy-based Sampling for Abstractive Multi-document Summarization in Low-resource Settings. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 123–133, Prague, Czechia. 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.

Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur P. Parikh, and Emma Strubell. 2022. Improving Compositional Generalization with Self-Training for Data-to-Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4205–4219, Dublin, Ireland. Association for Computational Linguistics.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prakamya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, and Hong Yu. 2024. SYNFAC-EDIT: Synthetic Imitation Edit Feedback for Factual Alignment in Clinical Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20061–20083, Miami, Florida, USA. Association for Computational Linguistics.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-Source Multi-Type Knowledge Exploration and Exploitation for Dialogue Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12522–12537, Singapore. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The Patient is More Dead than Alive: Exploring the Current State of the Multi-Document Summarization of the Biomedical Literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Sanjana Ramprasad, Kundan Krishna, Zachary C. Lipton, and Byron C. Wallace. 2024. Evaluating the Factuality of Zero-shot Summarizers Across Varied Domains. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 50–59, St. Julian's, Malta. Association for Computational Linguistics.

Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2023. Hallucination Mitigation in Natural Language Generation from Large-Scale Open-Domain Knowledge Graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12506–12521, Singapore. Association for Computational Linguistics.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute First, then Generate: Locally-attributable Grounded Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo, Marcio Lima Inácio, and Thiago Alexandre Salgueiro Pardo. 2024. Investigating Paraphrase Generation as a Data Augmentation Strategy for Low-Resource AMR-to-Text Generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 663–675, Tokyo, Japan. Association for Computational Linguistics.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained Summarization Evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2024. Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17775–17789, Miami, Florida, USA. Association for Computational Linguistics.

Bin Sun, Yitong Li, Fei Mi, FanHu Bie, Yiwei Li, and Kan Li. 2023. Towards Fewer Hallucinations in Knowledge-Grounded Dialogue Generation via Augmentative and Contrastive Knowledge-Dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Erik Cambria. 2024. Prompted Aspect Key Point Analysis for Quantitative Review Summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W. Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. 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 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, and Scott Sanner. 2024. Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6601–6633, Miami, Florida, USA. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022. Evaluating and Improving Factuality in Multimodal Abstractive Summarization. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023a. Faithfulness-Aware Decoding Strategies for Abstractive Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

David Wan, Shiyue Zhang, and Mohit Bansal. 2023b. HISTALIGN: Improving Context Dependency in Language Generation by Aligning with History. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2960, Singapore. Association for Computational Linguistics.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022. Improving Faithfulness by Augmenting Negative Summaries from Fake Documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuoer Wang, Marcus Collins, Nikhita Vedula, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. 2023. Faithful Low-Resource Data-to-Text Generation through Cycle Training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2847–2867, Toronto, Canada. Association for Computational Linguistics.

Haiyang Wang, Yuchen Pan, Xin Song, Xuechen Zhao, Minghao Hu, and Bin Zhou. 2024. F^2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4457–4470, Miami, Florida, USA. Association for Computational Linguistics.

Jędrzej Warczynski, Mateusz Lango, and Ondrej Dušek. 2024. Leveraging Large Language Models for Building Interpretable Rule-Based Data-to-Text Systems. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 622–630, Tokyo, Japan. Association for Computational Linguistics.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 944–964, Malta. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024. Word Alignment as Preference for Machine Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3223–3239, Miami, Florida, USA. Association for Computational Linguistics.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. KILM: Knowledge Injection into Encoder-Decoder Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5013–5035, Toronto, Canada. Association for Computational Linguistics.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. 2024. Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, Tokyo, Japan. Association for Computational Linguistics.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024. TaPERA: Enhancing Faithfulness and Interpretability in Long-Form Table QA by Content Planning and Execution-based Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840, Bangkok, Thailand. Association for Computational Linguistics.