Never Truly Out of Fashion: A Retrospective Look at Evaluation in NLG

Patrícia Schmidtová[1] ✉, **Saad Mahamood**[2], and **Ondřej Dušek**[1]

[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[2]Shopware, Düsseldorf, Germany
✉ Corresponding author: schmidtova@ufal.mff.cuni.cz

Abstract

Human evaluation (HE) remains the gold standard for assessing natural language generation (NLG) systems, yet automatic metrics are cheaper and faster, creating mounting pressure to skip it. We ask how evaluation practices have changed as NLG research scales. We analyse 24,291 papers from the ACL Anthology (1952–2025) through regular-expression-powered keyword analysis. Before 1990, the majority of NLG papers reported no evaluation at all; today, evaluation is near-universal and HE has held broadly stable over the past decade, despite the rapid emergence of large language model (LLM) judges (referred to as LLM-as-a-judge) since 2023. However, while LLM judges currently serve predominantly as a complement rather than a full substitute for human evaluation, a substantial share of papers already use them without any human validation. Faithfulness has become the fastest-rising evaluation criterion since 2020, coming back into fashion after almost 15 years of decline, tracking the prominence of hallucination research, while criteria such as grammaticality and fluency are receding, suggesting these qualities may increasingly be taken for granted as model outputs improve. Our findings provide a longitudinal baseline for tracking where the field stands.

1 Introduction

The ACL Anthology[1] now comprises over 120,000 papers and keeps growing at an unprecedented rate. Amid this pressure for volume, evaluation, which is the primary mechanism to certify the field's progress, risks becoming a formality rather than a guarantee. This has led to researchers using a plethora of automatic metrics inappropriately (Schmidtova et al., 2024), while at the same time human evaluations remain a fraction of conducted evaluations and struggle with methodological shortcomings (Howcroft et al., 2020; van der Lee et al., 2021). With the ever increasing number of new large language models (LLMs) and the prevalence of using LLMs to evaluate generated text, the incentive to reach for automatic approaches to evaluation has never been stronger (Gehrmann et al., 2023). Crucially, model development and iterative optimization generally require cheap, scalable automatic metrics for rapid feedback in the development loop, since conducting human evaluations at each iteration is logistically and financially impractical. Automatic metrics are thus a structural necessity in modern NLG engineering, even if human evaluation through the use of task-based evaluations remains the ultimate arbiter of quality.

Yet the field has navigated similar tensions before. Debates over the adequacy of BLEU (Papineni et al., 2002) for machine translation (Freitag et al., 2022; Reiter, 2018), the broader question of metric validity in NLG (Stent et al., 2005; Belz and Reiter, 2006; Reiter and Belz, 2009; Novikova et al., 2017), and the gradual consensus around structured evaluation protocols such as Direct Assessment and MQM (Freitag et al., 2022; Belz et al., 2020) all point to a community that periodically pauses, reflects, and self-corrects. This work offers such reflection through a quantitative analysis of evaluation trends across seven decades of the ACL Anthology. Our perspective is informed by a long line of work arguing that evaluation in NLG must be taken more seriously – from early calls for human-centred and realistic assessment (Reiter and Belz, 2009; Reiter, 2011) to recent evidence that the community still devotes vanishingly little attention on real-world impact evaluations (Reiter, 2025).

We use a corpus of 24,291 NLG system papers from the ACL Anthology (1952–2025)[2] to exam-

[1]https://aclanthology.org/

[2]See Section 4 for details on data availability.

ine: (1) how human evaluation (HE) prevalence and annotation methodology have evolved across venues and years; (2) whether LLM-as-a-judge is displacing human evaluation; (3) which evaluation criteria have risen, faded, or returned after periods of absence; (4) how the balance between automatic, human, and LLM-judge evaluation has shifted; (5) whether evaluation practices differ systematically by NLG task; or (6) which evaluation methodologies were popular.

2 Methodology

We construct a large-scale analysis pipeline over a corpus of NLP research papers, using keyword-based signal detection over available paper text (title/abstract and, where available, full text). An LLM-based extraction experiment proved insufficiently reliable at scale and is reported only in Appendix A. Our regular expressions prioritise precision; pattern details are listed in Appendix A.

2.1 Corpus Assembly

For papers published up to and partially including 2022, we use the ACL Anthology Open Research Corpus (ACL-OCL; Rohatgi et al., 2022),[3] comprising approximately 70,587 full-text articles across 215 venues. For 2022 onwards, we supplement with our own crawl of the ACL Anthology, restricted to nine core venues (ACL, EMNLP, NAACL, EACL, AACL, IJCNLP, INLG, TACL, CL) due to the cost of large-scale PDF processing.[4] After deduplication, the combined corpus comprises **85,792 papers** (1952–2025).

2.2 Paper Filtering

We apply a cascaded filter to retain papers that (a) are not meta-papers (surveys, tutorials, proceedings prefaces), (b) perform a generative NLP task such as MT, summarisation, dialogue, data-to-text, or question generation, matched against the full text of the paper, and (c) contain at least one mention of evaluation in the full text. Evaluation signals fall into three categories: *human evaluation* (mentions of human/manual evaluation, annotation studies, crowdsourcing, named protocols such as Direct Assessment and MQM); *automatic evaluation* (e.g., BLEU, ROUGE, BERTScore, COMET); and *LLM-as-a-judge* (mentions of situations an LLM evaluates system outputs). This procedure yields a final analysis corpus of **24,291 papers** (see Table 1 in the Appendix).

2.3 Analysis

All analyses are conducted at the paper level, with year and venue as primary grouping variables and task type as a secondary variable. Evaluation criteria are detected via keyword matching over full paper text using patterns derived from the taxonomy proposed by Howcroft et al. (2020).

3 Results

Human evaluation prevalence and trajectory. Human evaluation has been remarkably stable over the past decade: the proportion of NLG papers reporting at least one human evaluation has fluctuated between 34% and 42% every year since 2018, with no sustained decline despite the growing availability of automatic metrics and LLM judges. In absolute terms, the number of human-evaluated papers has grown substantially – from around 550 per year in 2018–2019 to over 650 in 2025 – but this tracks the overall growth of the field rather than a shift in evaluation culture. A logistic regression over the recent decade (2015–2025) confirms the absence of any statistically significant trend in the proportion of papers using human evaluation ($\beta = 0.0023, p = 0.69$), statistically validating this stability. The proportion of papers at core, general NLP venues closely mirrors the corpus-wide trend, though specialised generation venues such as the International Conference on Natural Language Generation (INLG) show consistently higher human evaluation rates: 53.1% overall, compared to 36.0% in core NLP venues (see Appendix B).

LLM-as-a-judge adoption. The adoption of LLM-as-a-judge has grown exponentially since 2022 (Bavaresco et al., 2025): 109 papers in 2024 and 224 in 2025, compared to 19 in 2023 and zero before. This rapid growth is highly statistically significant under logistic regression over 2020–2025 ($\beta = 0.9523, p = 4.25 \times 10^{-76}$, odds ratio $e^{0.9523} \approx 2.59$ per year). Earlier papers generally used either custom-finetuned encoder-only models, such as RoBERTa (Liu et al., 2019), for classification, or GPT-2 (Radford et al., 2019) to measure perplexity. Among papers using an instruction-tuned LLM judge, 61.4% report human evaluation as well, and 86.2% of these also explicitly mention validation or manual checking of the LLM outputs.

[3] https://huggingface.co/datasets/ACL-OCL/ACL-OCL-Corpus

[4] This venue asymmetry means that post-2022 comparisons involving smaller venues should be interpreted with caution.

This strong baseline suggests that LLM judges are currently viewed predominantly as complements to human evaluation rather than full substitutes. However, it also means that 38.6% of papers employing an LLM judge do so without any reported human evaluation, a share that bears watching as the paradigm matures. Adoption rates vary across tasks, with newer or long-form generation tasks showing the highest prevalence: story generation leads with 9.0% of its papers adopting LLM judges (9 papers), followed by code generation at 8.3% (15 papers), data-to-text at 3.2% (14 papers), and question generation at 3.1% (11 papers). In contrast, established tasks with massive historical paper volumes show much lower adoption rates, such as machine translation at 0.5% (51 papers), paraphrase generation at 1.4% (14 papers), and summarization at 2.2% (46 papers).

Human evaluation criteria. We detect evaluation criteria via keyword matching over full paper text using category names from the Howcroft et al. (2020) taxonomy, counting each criterion at most once per paper. As Howcroft et al. (2020) emphasise, the same term can carry different meanings across papers; our counts thus reflect terminology adoption rather than consistent operationalisation. Among 8,894 human-evaluation papers, fluency (22.5%), relevance (21.6%), and coherence (16.6%) are the most common. The apparent frequency of "accuracy" (52.8% of papers) is likely an artifact: the term is used loosely across many NLG papers to describe model performance broadly, exemplifying the terminological confusion highlighted by Howcroft et al. (2020). We retain it in Figure 1 but caution against interpreting its trend as reflecting deliberate evaluation design.

Figure 1 reveals genuinely cyclical patterns. Faithfulness (Maynez et al., 2020) follows a U-shaped trajectory: used in early NLG and MT evaluation, largely absent through the 2010s as automatic metrics displaced human assessment, and now rising sharply from 1.7% of HE papers in 2015 to 29.8% in 2025 ($\beta = 0.3329, p = 2.39 \times 10^{-94}$ under logistic regression over 2015–2025), driven by hallucination research (Schmidtova et al., 2025). Adequacy and grammaticality are on a declining arc, while relevance, consistency, and coherence show cyclical recoveries.

Evaluation modality mix. Automatic-only evaluation is by far the most common modality (57.8%) and has grown as a share of the corpus over time. Before 2000 – and largely before BLEU (Papineni et al., 2002) – automatic-only evaluation accounted for roughly 30–55% of papers per period, with human-only evaluation reaching 10–22%; by 2020–2025, automatic-only has stabilised around 55% while human-only has fallen to 2–3%, and combined human-and-automatic evaluation has risen from under 10% to over 30%. The near-absence of human-only evaluation today indicates that automatic metrics have become a de facto prerequisite: even papers that invest in human judgement almost always report automatic scores alongside them. LLM-judge-only papers remain a small fraction but are growing rapidly, rising from near-zero in the 2020 period to 5.5% of papers using an LLM judge without any human evaluation in 2025. Of those, roughly a half use API-only models from OpenAI, which so far seem to outperform open-weight models in correlation with human judgement (Huang et al., 2025); however, they raise questions about the reproducibility of such evaluations (Schroeder and Wood-Doughty, 2025). Evaluation modality also varies by venue group, with SIGGEN venues showing notably higher human evaluation rates than core NLP venues; we report the full breakdown in Appendix B.

Task-based differences. Human evaluation rates vary across NLG tasks, though less dramatically than one might expect: question generation has the highest rate at roughly 53%, while most other tasks cluster between 37% and 47%. Machine translation is the clear outlier at around 28% overall, reflecting its uniquely mature automatic metric ecosystem. However, MT's recent human evaluation rate is closer to 30%, with the lower overall average pulled down by early decades when human evaluation of MT was rarer. Indeed, the MT community has developed a notably rigorous meta-evaluation tradition – shared tasks such as WMT provide large-scale human annotations that calibrate automatic metrics against human judgements (Freitag et al., 2022; Zouhar et al., 2024). No comparable infrastructure exists for most other NLG tasks; establishing it would be a concrete step toward the same level of metric accountability.

Annotation methodologies. Among all 8,894 human-evaluated papers, Likert-scale rating is the most frequently detected annotation approach (19.0%), followed by post-editing (15.8%), binary and multi-class categorisation (13.9%), and ranking/pairwise comparison (13.2%). Error span an-

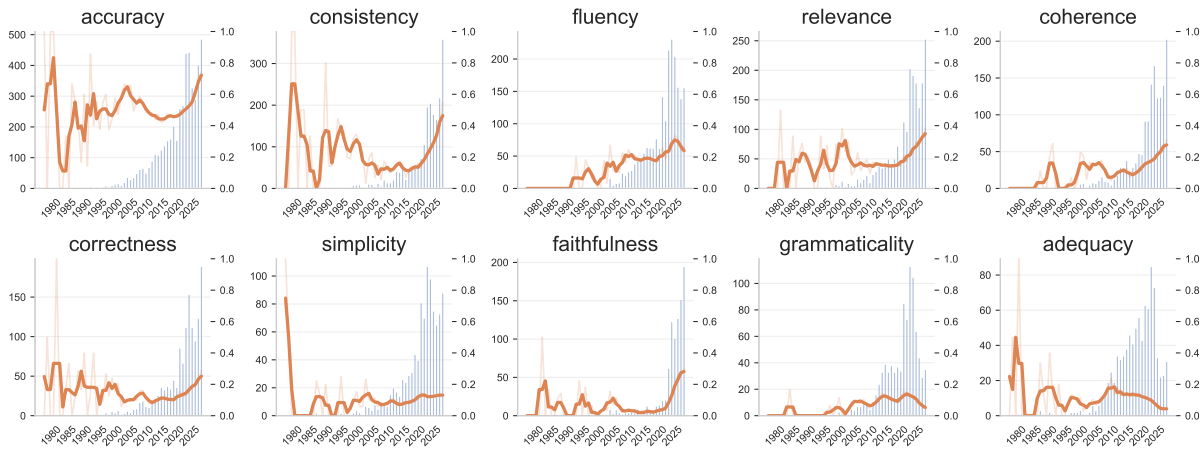

Figure 1: Prevalence of top 10 human-evaluation criteria over time. Left Y-axis: annual paper counts (bars); right Y-axis: proportion of HE papers mentioning the criterion (thick line: 3-year rolling mean; faint line: raw annual proportion).

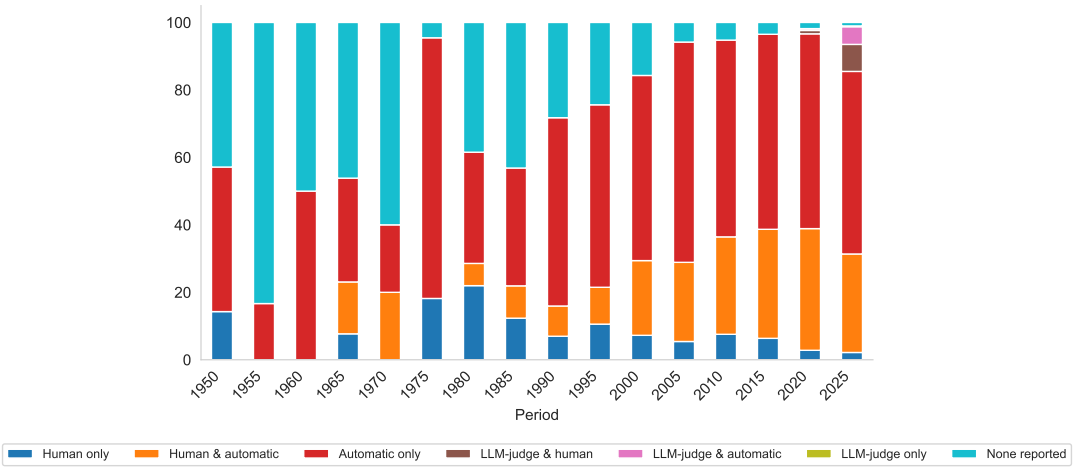

Figure 2: Evaluation modality mix by 5-year period (1950–2025) as a percentage of papers (Y-axis), showing the proportion relying on human-only, automatic-only, combined human-and-automatic evaluation, or LLM-as-a-judge approaches.

notation (4.0%), Direct Assessment (Graham et al., 2013) (2.3%), and Best-Worst Scaling (Kiritchenko and Mohammad, 2017) (0.8%) round out the detected methods. While we attempted to extract fine-grained metadata regarding reporting quality, specifically inter-annotator agreement (IAA), annotator counts, and rated item counts, these variables could not be reliably validated due to low recall and reporting inconsistencies, and are therefore omitted from our quantitative analysis. Best-Worst Scaling and error span annotation remain specialised, concentrated in MT and reference-free evaluation.

Method choices also shift over time (Figure 3): across all papers, post-editing was once the most common methodology, peaking in popularity around 2014–2015 when it was used in 28.6% of all human-evaluated papers (and 43.3% in machine translation). Over the last decade, however, its use has declined sharply, dropping to just 6.1% in 2025 (and 17.7% in machine translation specifically). In contrast, Likert-scale rating has experienced substantial growth, rising from 13.8% in 2015 to become the most prevalent approach at 28.7% in 2025. Additionally, error span annotation has steadily gained traction in recent years, growing from 3.7% in 2015 to 8.0% in 2025, reflecting a growing interest in fine-grained evaluation. We show the longitudinal breakdown of these methodology trends in Figure 3.

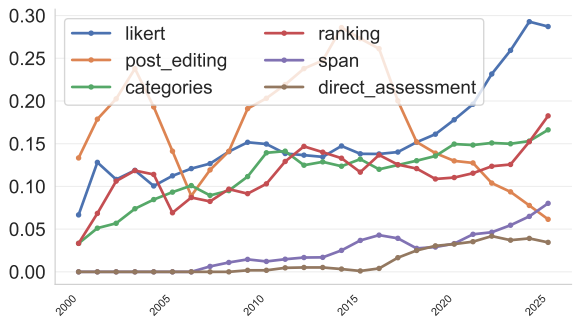

Figure 3: Prevalence of top human annotation methodologies over time, expressed as the fraction of human-evaluated papers on the Y-axis (3-year rolling mean).

4 Conclusions

We have presented a large-scale longitudinal analysis of NLG evaluation practices across 24,291 papers from the ACL Anthology spanning seven decades. The field has come a long way: before 1990, between 38% and 83% of NLG papers in each five-year period reported no evaluation at all, and the share only dropped below 20% after 2005. Today, evaluation is near-universal and human evaluation has held roughly steady at 35–45% of NLG papers over the past decade, defying predictions of decline (van der Lee et al., 2021; Reiter, 2025) even as it shows no upward trend despite repeated calls for more rigorous assessment.

Three findings stand out as actionable:

LLM judges need calibration infrastructure. LLM-as-a-judge has become standard practice in under two years: 109 papers in 2024 and 224 in 2025, up from near-zero before 2023. While 61.4% of these papers also report human evaluation, the remaining 38.6% rely on LLM judgements without any reported human validation. This mirrors earlier episodes the field later had to revisit, such as the widespread adoption of BLEU without adequate validation or the reliance on single-reference translation evaluation, and runs the risk of treating LLM judgements as a trusted signal before their failure modes are understood. We recommend that venues encourage, and reviewers expect, explicit calibration of LLM judges against human judgements, particularly for tasks and domains where no prior calibration exists.

Faithfulness evaluation needs standardisation. Faithfulness has become the fastest-growing evaluation criterion in our data and reflecting the centrality of hallucination as a research problem (Maynez et al., 2020; Schmidtova et al., 2025). Yet the broader picture is less encouraging: "accuracy" appears in over 50% of human-evaluation papers, used so loosely that it functions less as a defined criterion than as a catch-all for correctness (Howcroft et al., 2020). The risk is that faithfulness, as it matures, drifts toward the same terminological vagueness. Establishing shared annotation tasks for faithfulness – analogous to what WMT provides for translation quality via Direct Assessment and MQM (Freitag et al., 2022) – would help prevent this and provide the calibration infrastructure that other NLG tasks currently lack.

Some things never go out of fashion. The cyclical patterns in our data – faithfulness returning after a decade of absence, coherence and relevance waxing and waning with shifts in task popularity – suggest that the field's evaluation vocabulary is smaller and more stable than its rapid growth might imply. These questions recur because they reflect enduring properties of language use, not passing methodological fashions. In a similar spirit, Reiter (2024) argues that rule-based NLG systems retain lasting value even in an era dominated by neural methods – a reminder that what matters is fitness for purpose, not novelty. This principle extends to evaluation: human judgement, automatic metrics, and now LLM judges each have their place, and the field's task is not to choose among them but to understand when each is trustworthy.

One purpose of this paper is to document the current state of affairs and provide a baseline against which future shifts can be measured – whether toward a deeper LLM-judge reliance, new evaluation criteria, or a renewed emphasis on real-world impact (Reiter, 2025).

Our code and dataset (containing paper IDs, metadata, and extracted evaluation signals) are publicly available at `https://github.com/patuchen/trends_in_nlg_eval`. To respect publisher copyrights, the shared dataset does not include the raw full texts of the papers; instead, we provide the metadata, IDs, and extracted signals in a CSV file, along with utility scripts in the repository to automatically download and reconstruct the full-text corpus directly from the official ACL Anthology using the paper IDs.

Limitations

Methodological limitations Our measurements rely on keyword-based detection, which can yield

false positives when terms occur in related work or background sections and false negatives when papers use unusual terminology. We apply the same patterns to all available text (title/abstract and, when available, full text).

We deliberately use keyword-matching over LLM-based extraction due to initial experiments showing that LLMs had problematically low recall when extracting fine-grained methodological metadata at scale (see Appendix A). Consequently, we developed an extensive quality assurance procedure for our regex patterns, manually auditing samples to prevent false positives while balancing recall (100 test sentences for every regular expression). By aggressively stripping meta-references and generic terminology, our reported methodology counts necessarily represent a conservative lower bound of actual evaluation practices. We validated this regex-based approach on a manually curated sheet of 60 human-annotated papers, where 51 were successfully mapped to our corpus, yielding a recall of 82.4% after applying targeted pattern overrides.

Finally, our LLM-as-judge classification relies on model name regex matching. Pre-2023 uses of large models as evaluators (e.g. GPT-2 perplexity scoring, adversarial BERT discriminators) are conceptually distinct from the modern instruction-tuned paradigm; we distinguish these as *ml_evaluator* vs. *instruction_llm*, but it is possible that borderline cases remain.

Venue coverage asymmetry. Our post-2022 corpus covers only nine major venues, excluding workshops and smaller conferences that are present in the pre-2022 ACL-OCL data. Since workshop papers tend to report human evaluation less frequently, the post-2022 human evaluation rates may be slightly overestimated relative to earlier years where the full breadth of venues is represented. The most important venues for NLG research are covered throughout, but longitudinal trends should be interpreted with this asymmetry in mind.

Task coverage. Our generative task filter covers a curated set of named NLG tasks (see Appendix A) rather than all possible forms of natural language generation. While this set includes the largest task categories – machine translation alone accounts for over half the analysis corpus – emerging or niche generation tasks may have been omitted from the analysis.

Acknowledgments

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV project 260 821.

References

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).

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Ehud Reiter. 2011. Task-based evaluation of NLG systems: Control vs real-world context. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 28–32, Edinburgh, Scotland. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

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

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

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2022. The acl ocl corpus: advancing open science in computational linguistics. arXiv.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 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*, pages 60–79, Hanoi, Vietnam. 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.

Kayla Schroeder and Zach Wood-Doughty. 2025. Can you trust llm judgments? reliability of llm-as-a-judge. *Preprint*, arXiv:2412.12509.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351, Mexico City, Mexico. Springer.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Pipeline Details

Corpus statistics. Table 1 summarises the filtering stages.

Stage	Remaining	Dropped
Raw corpus (combined)	85,792	–
Meta-paper exclusion	81,953	3,839
Generative task filter	20,593	61,360
Evaluation signal filter	16,587	4,006
Full-text task pass	24,291	–

Table 1: Corpus size at each pipeline stage.

Task categories. The task filter recognises: *machine_translation*, *summarization*, *dialogue*, *data_to_text*, *question_generation*, *paraphrase*, *simplification*, *captioning*, *general_nlg*, *question_answering* (open-ended/generative QA), *instruction_following*, *counterspeech*, and *highlight_generation*. All of the code and the regular expressions are available in our public repository at `https://github.com/patuchen/trends_in_nlg_eval`.

Evaluation signal patterns. Human evaluation is detected by expressions such as *human eval**, *manual eval**, *annotators*, *inter-annotator*, *crowdsourc**, *Mechanical Turk*, *Direct Assessment*, *MQM*, and *Likert*. Automatic evaluation is detected by metric names including *BLEU*, *ROUGE*, *METEOR*, *chrF*, *BERTScore*, *BLEURT*, and *COMET*. LLM-as-a-judge is detected by patterns matching an LLM name followed by evaluation-adjacent verbs (e.g. *GPT-4 evaluates*, *Claude rates*).

LLM extraction experiment. To test whether fine-grained metadata could be extracted automatically, we prompted Qwen2.5-14B-Instruct (Team, 2024) (served via vLLM; Kwon et al., 2023) to extract structured evaluation records from isolated evaluation sections. Comparison against expert annotations revealed low recall for fields such as annotator counts and rated item counts, motivating our reliance on keyword-based signals for all reported results.

B Venue-Group differences.

Core NLP venues are the most automatic-only (60.6% auto-only; 33.1% human+auto), while SIGGEN venues show a more mixed profile (35.8% auto-only; 42.4% human+auto; 10.8% human-only). Journals resemble core venues in relying primarily on automatic metrics (50.8% auto-only; 40.3% human+auto), but show zero cases of LLM-as-a-judge evaluation without human evaluation, consistent with slower uptake of the judge-only pattern. See Figure 4 for the complete visual breakdown across all venue groups.

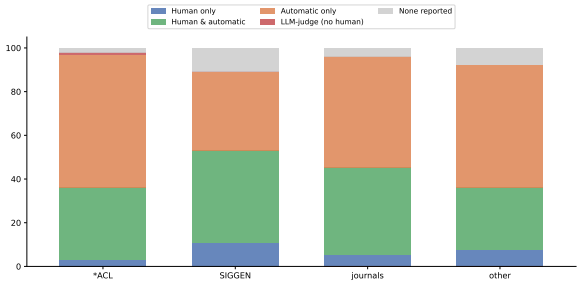

Figure 4: Evaluation modality profile by venue group, shown as a percentage of papers within each group on the Y-axis.

C Phrasing Examples of Historical NLG Papers Without Human Evaluation

For papers in earlier decades that reported no formal human evaluation or relied on informal, qualitative assessments of system output, the following examples illustrate typical phrasings:

- **1976:** *"We choose the interpretation showing the preferable matching of nouns and case by using an evaluation function below which has been established empirically"*
- **1984:** *"This translation is called by its authors as word-by-word, turn-by-turn one; several years have already passed in a complete satisfaction of the customers."*
- **1986:** *"The implementations of the whole system has already been completed and the translation results (10,000 sentences) are now being evaluated by professional translators and native speakers of English. The evaluation results obtained by now are quite satisfactory."*

D Human Evaluation Criteria by Task

Figure 5 shows a detailed heatmap breakdown of the top 10 human evaluation criteria across the ten most common NLG tasks in single-task papers.

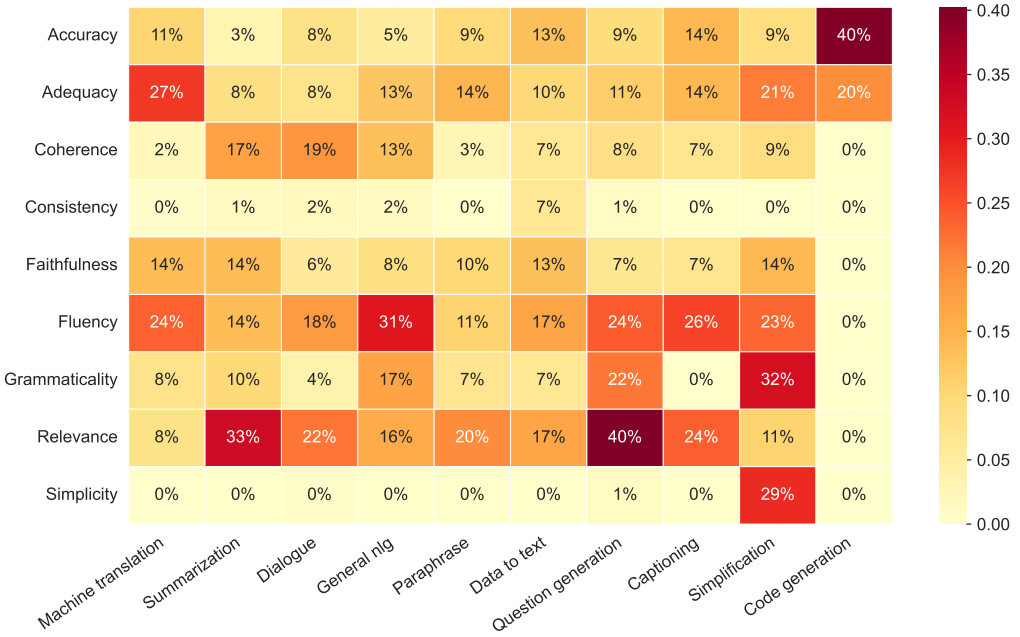

Figure 5: Prevalence of human evaluation criteria (listed on the Y-axis) across the top 10 single-task NLG tasks (listed on the X-axis), expressed as a percentage of papers per task that perform human evaluation.