Evaluation and Assessment as Complementary Frameworks

Elie Antoine
DIRO, RALI, Université de Montréal
Montréal, Québec, Canada
elie.antoine@umontreal.ca

Abstract

Language model capabilities have advanced faster than the methods used to evaluate them, particularly since the move from task-specific systems to general-purpose models which are deployed across an ever-widening range of tasks. When models were built for a single task, evaluation sat in a tight relationship between the task, the data, and the model. General-purpose models have weakened this relationship, and the evaluation practices that were built around it have not adjusted. This paper argues that addressing this gap requires treating *evaluation*, understood as quantitative performance measurement, and *assessment*, understood as the analysis of mechanisms and real-world behavior, as complementary rather than interchangeable. This distinction matters because *evaluation* is now often asked to stand alone in settings where a benchmark score cannot tell us what a model is doing, or how its behavior will hold up outside the benchmark.

1 Introduction

Natural Language Processing (NLP) has in recent years experienced an unprecedented expansion and societal impact. Until recently, it was less visible as a field, reaching the general public mainly through specific applications such as machine translation, autocorrect, or autocomplete. Today, a large share of the population[1] has heard of NLP through commercial models such as ChatGPT or Claude and a growing part use them in their professional and personal life or in even more personal contexts such as a substitute for consulting a physical or mental health professional.

[1]Numbers vary depending on region and sources : 68% to 90% of European and American workers have heard of generative AI according to (Bick et al., 2026), while roughly 90% of Americans have (Kennedy et al., 2025). These figures are centered around North America and Europe, thereby creating a picture that is likely inflated and not representative of the global population, a pattern reflected in the data on investment and impact (Microsoft AI Economy Institute, 2026).

This broad adoption has outpaced the field's ability to characterize what these models can and cannot do, as well as how and why they succeed or fail in the tasks for which they are used. Benchmarks remain the main framework through which the community tracks progress and compares models, and they have grown considerably in size and scope, now covering capabilities ranging from general knowledge and reasoning to instruction following, coding, tool use, and more abstract dimensions such as alignment and safety. Their logic, however, has mostly stayed the same: model outputs are compared against gold references, and the result is compressed into a single figure per task or benchmark. This is inherited from an era of task-specific systems, where the task was fixed and the reference was the right or at least reasonably attainable answer, and it was well suited to both the tasks and capabilities of the models at the time. Applied to general-purpose models that now saturate those benchmarks (Akhtar et al., 2026), it continues to produce rankings, but they tell us neither how a model handles a given input, nor how its evaluated behavior relates to its behavior in the open-ended tasks they are used for in practice. For example, ROUGE (Lin, 2004) was built for n-gram overlap on short extractive news summaries, where a reference output was close to the system output by design; it was stretched within summarization to longer and more abstractive cases, and is now reported as evidence of factual correctness in hallucination detection (Janiak et al., 2025) and of retrieval quality in RAG pipelines (Yu et al., 2025). These are settings where many outputs can be correct and similarity to one reference cannot tell them apart. The score keeps being produced, but is asked to carry the weight of a much broader claim about the model.

The position this paper takes is that filling this gap does not require replacing benchmarks, nor

scaling up the same logic. The question benchmarks answer, *how do models compare against a reference*, is not the only one worth asking, and the question of *how and why a model behaves as it does* calls for a different methodology. This is done here by clearly separating the concept of "examining" models into two frameworks, *evaluation* and *assessment*, which serve different purposes and answer different questions, and by arguing that we need to treat them as complementary rather than interchangeable, in the sense that one is often asked to do the work of the other.

The rest of this paper proceeds as follows. Section 2 sketches how the relationship between tasks, their associated data, inference methods, and evaluation has evolved with the rise of general-purpose models, and identifies the structural shift that motivates the rest of the argument. Section 3 defines and presents the difference between *evaluation* and *assessment*, and argues that the two are complementary rather than interchangeable. Section 4 discusses this distinction in relation to existing work, from linguistic probing and fine-grained evaluation to user-centered studies and more recent attempts to formalize evaluation by "vibes", and situates it alongside adjacent methodological proposals.

2 From task-specific to generic models

The development of NLP methods can be described through three main components, strongly interacting with each other: **the task**, together with the annotated or curated data that supports it, **the inference method** used to produce outputs, which today is largely a Large Language Model, and **the evaluation methods and metrics** through which outputs are judged. What is interesting is how the interactions between these components have evolved. Earlier in the development of NLP as a field, research was primarily focused on the task. When a new method was developed, the task was fixed and relatively narrow, and the data was annotated specifically for that task, both for training the inference method and for evaluating it. The architectures themselves were often designed around the task, with specific inductive biases for sequence labeling, parsing, or machine translation, partly because model capacity at the time did not permit a unified approach. The evaluation methods and metrics were tied just as directly to the task: a tagger for named entity recognition, a translation system, and a summarizer would not be evaluated in the same way. This made sense both because each metric measured something specific to the task and because model performance was lower, so comparing systems against a few canonical references was already challenging enough to produce meaningful differences.

The first shift came with the arrival of early transformer (Vaswani et al., 2017) models on a moderate scale, such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020) and RoBERTa (Liu et al., 2019). Rather than training a model directly on the final task, a pre-training stage was introduced on a much simpler objective, typically predicting words hidden from their context, which was not itself a task of interest but proved remarkably effective as a "foundation" (Bommasani et al., 2021) for the real tasks. On top of this foundation, what was done was essentially the same as before, task-specific training, now in the form of fine-tuning a model.

The major change came with the idea of general-purpose models. T5 (Raffel et al., 2020) proposed that every task, regardless of its nature, could be cast in a single text-to-text format, while GPT-3 (Brown et al., 2020) showed that tasks could be specified at inference time through natural language prompts and a few examples, without any fine-tuning. Pushed to its limit through training models on conversational data and human feedback (Ouyang et al., 2022), this idea moved from a specialist-facing prompting paradigm to a general interaction mode, where tasks are handled through natural language interactions with a model never explicitly trained on most of them.

This last shift, sketched in Figure 1, marks an important change in which component is now driving the others. Research was previously organized around the task: the data was annotated for it, the architecture was designed for it, and the evaluation was tied to it. It is now organized around the model: the same system handles a wide range of tasks through prompting, and the tasks, data, and evaluations are defined in relation to what the model can do. One consequence is that because the model is generative, evaluation and task must be generative-compatible, regardless of their underlying structure. A classification or extraction problem becomes, through those models, a text-generation task, and the metric operates on the generated text rather than directly on the format and context it was originally about.

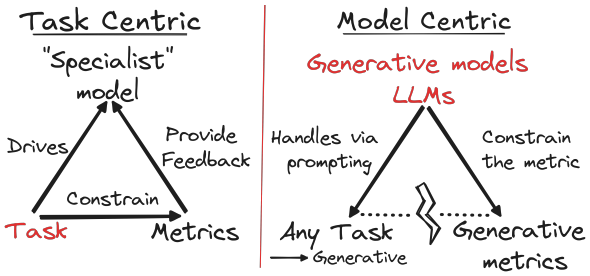

Figure 1: The shift from task-centric to model-centric NLP: the component that drives the others moves from the task to the model, and task and metric are recast in generative form to fit it.

A second consequence of this shift is the increased diversity and abstraction of the tasks that these models can now plausibly be asked to perform. Asking a model to read a long financial document, identify the companies mentioned in it, extract the relevant performance indicators, and produce a summary in the form of slides with supporting visualizations is not an unrealistic request today (Egg et al., 2025; Zheng et al., 2025). Evaluating such a task is another matter entirely. What should be evaluated? Each atomic component? The named entity recognition and table extraction steps can probably be formalized, but are not as simple to test in practice as this framing suggests, and more abstract or open-ended components, such as data cleaning, normalization, the choice of which indicators are most relevant (relevant for whom, and for what use?), the use of code tools to produce visualizations, and the judgment of those visualizations, raise a deeper question: what would a gold reference even look like? Even with expert human annotators, it is unclear what should be annotated, against what criteria, and whether two equally competent annotators would agree on the answer. *What makes a visualization good*, for instance, depends on whether it is grounded in real data, readable, communicative, and well-designed, criteria spread across factual, aesthetic, and communicative dimensions that sit beyond the reach of the grading formats we currently rely on. Cases like this are part of what motivates looking more carefully at what we mean when we talk about "evaluation".

3 Evaluation and Assessment

Two practices are often discussed together under the general heading of evaluation, but they serve different purposes. The first, *evaluation* in a stricter sense, is the established practice of quantitative performance measurement, and its canonical question is: *Does my model perform better than some other model?* Given a model and a reference, *evaluation* produces a score, or a set of scores, that allows models to be compared at a given point in time on one or several criteria. Its primary function is comparative. The second, *assessment*, is broader, and its canonical question is: *How and why does my model or method work, or fail to work?* In the definition adopted here, it covers any method aimed at answering that question, including linguistic probing, behavioral testing, human-centered studies, and mechanistic interpretability, among others.

Beyond the definitions just given, "assessment" as a term is also increasingly common in recent NLP writing. A keyword search across the ACL Anthology, covering paper titles and abstracts of the main *ACL venues, shows that the share of papers using *both* terms has grown sharply in recent years, from roughly 1.7% in 2020 to 10.6% in 2025, a nearly sixfold increase in five years.[2] Over the same period, the share of papers using only "assessment" has stayed flat at 3–4%.

How to read this trend is not obvious. One reading is that "assessment" is simply a more modern term, adopted under the pull of topics like capability assessment, risk assessment, or safety evaluation (Shevlane et al., 2023), without any underlying change in practice. Another is that the community is not replacing one term with the other but adding "assessment" alongside "evaluation" in contexts where the latter alone no longer seems to cover what authors want to say: on this reading, something closer to a second practice is in fact taking shape. This data alone cannot decide between these readings. Following the argument of this paper, settling the question would itself require *assessment* rather than pure quantitative description. The aim in the rest of this section is therefore different: to argue that the two questions just set out are different in kind, and that clearly separating *assessment* and *evaluation* in the way proposed here is useful for thinking about model analysis, whatever the vocabulary ends up doing in community practice.

Evaluation and *assessment*, on this distinction, are not competing practices, and the argument of this paper is not that one should replace the other. *Evaluation* produces comparable numbers, which

[2] More detail and figures on this can be found in Appendix A

is what lets a field track progress and decide between methods. *Assessment*, by contrast, takes a model as an object of study rather than a point on a scale, and asks what it is doing, where it breaks, and how its behavior looks in the settings where it is actually used. Many of the questions about a model are not comparative at the level of aggregate performance, even when the methods used to answer them are quantitative, as in probing accuracies or behavioral test pass rates. A rounded account of what a model is and does typically draws on both.

In current practice, however, the two are far from balanced. *Evaluation* remains the dominant one, inherited from task-specific systems where comparing scores against a reference was both the natural thing to do and a reliable indicator of progress. It has continued largely unchanged, even as the conditions that supported it have weakened. This is visible in benchmark scores reported for commercial models that do not reliably return the same output, or in capability claims staked on a single aggregate number. The imbalance is not that *evaluation* is done too much, but that it is often asked to stand alone, in settings where a score on its own simply cannot tell us what the model is doing or how its behavior will hold up outside the benchmark. This imbalance is visible in the literature itself: Reiter (2025) reports that roughly 0.1% of ACL Anthology papers evaluate real-world impact, and that even these typically treat the impact finding as secondary to a metric-based one. When *assessment* is done, it is often positioned as a supplement to *evaluation* rather than a finding in its own right.

4 Existing Work

The distinction between *evaluation* and *assessment* as drawn here is not a new one in the sense that the practices it names already exist, and have for some time. Assessment of NLP systems, and of generative systems in particular, has a long history. The STOP system, which generated tailored smoking-cessation letters, was evaluated in the early 2000s through a randomised controlled clinical trial with over 2,500 participants (Reiter et al., 2001, 2003). The trial measured whether smokers receiving tailored letters were more likely to quit than smokers receiving a generic letter, rather than how the generated letters scored against a reference. It was not, however, the dominant practice even then, and mostly operated as a complement to metric-based evaluation rather than driving the analysis. This distinction is even more necessary now, given that modern models are expected to handle a wide range of open-ended tasks through a single interface. In fact, a substantial body of existing work is already doing *assessment*, even when it is not labeled as such, and the question is less how to build the practice from scratch than how to recognize it as a coherent framework.

The clearest cases are "evaluations" that sit closer to *assessment* than to *evaluation* in the narrow sense, such as BLiMP (Warstadt et al., 2020) and Holmes (Waldis et al., 2024), and more generally the fine-grained evaluation tradition (Gehrmann et al., 2023; Ribeiro et al., 2020), where what matters is the detail of what is being examined rather than the overall ranking of models: no one really seems to care, in practice, about a model's rank on BLiMP. Human evaluation of natural language generation outputs belongs here as well, along with user studies and deployment reports that track how systems behave once they leave the lab, a concern shared with the broader human-computer interaction community.

The same holds for probing and mechanistic interpretability, where the goal is again not to rank models but to understand how they function, this time by looking at the model's internals rather than its behavior. Probing is routinely described as evaluation, but what it actually does is closer to *assessment* in the sense defined here: the goal is not to rank models but to characterise what they have learned (Rogers et al., 2020). Mechanistic interpretability is a different case: the field already positions it as reverse-engineering rather than evaluation, which makes it a limit case for the framing in a different way, not because it has to be reclassified, but because it raises its own validity questions. In the line of work surveyed by Feldhus and Kopf (2025), which focuses on generating natural-language concept descriptions for neurons, attention heads, and SAE features, automation now operates at two distinct layers: the descriptions themselves are generated by other language models, and their quality is evaluated mostly through automatic measures. They note that "concept descriptions are for humans, making human judgment essential for validating the meaningfulness of automated metrics", and yet observe that human evaluation remains comparatively rare in this part of the field. The practice fits our definition of *assessment*, but the move to automate it should be

approached with care: the automated judges and metrics are themselves measurement instruments, and the validity questions Wallach et al. (2025) raise for evaluation apply to them too.

One further approach deserves its own mention: evaluation by "vibes". Unlike most of the practices just mentioned, it did not originate in research and move outward to users, but the opposite: it grew out of informal discussions among users on 𝕏/Twitter, where people shared their own practical tests alongside a more diffuse sense of a model's competence. This includes the more classical dimensions of code and writing, as well as less tangible qualities: the model's capacity to interact in ways that feel useful rather than flattering, avoiding what work on language models describes as *sycophancy*: the tendency of a model to adapt its answer to the user's stated beliefs or preferences, including when this leads it to endorse incorrect claims (Perez et al., 2023; Sharma et al., 2024). It also includes the ability to avoid something closer to a textual uncanny valley, where the output reads as almost-right but not quite. Recent work has begun attempting to formalize this (Dunlap et al., 2025; Itzhak et al., 2026). These attempts raise a question: what makes the practice interesting is that it is grounded in the user's own intuitive and contextual judgment, and proposals to automate it have to decide how much of that grounding to preserve as it is turned into something reproducible at scale.

Three recent proposals name a related gap, each reaching for different vocabulary. Wallach et al. (2025) argue that evaluating generative AI systems should be understood as a *measurement* problem using the tools of social-science measurement theory. Weidinger et al. (2025) call for a mature *evaluation science* for NLP, and in particular for a *behavioral approach* that overlaps with what is here called *assessment*. Where these two are broad methodological reframings, Reiter (2025) is narrower, drawing a sharper binary between *metric evaluation* and what fits here as one specific kind of *assessment*, his *impact evaluation*: the measurement of real-world performance indicators in deployed usage rather than performance on a test set. Together with the vocabulary shift documented in Section 3, these proposals suggest that the community is actively trying to articulate and develop a practice that current methods do not quite cover.

5 Conclusion

This paper argues that *evaluation* and *assessment* should be treated as complementary practices rather than as a single one. The argument is not about vocabulary: other words could be used in place of these two, and several recent proposals already rely on different vocabularies to describe related concerns. The point is that *evaluation* is often used as a wide term, stretched to cover practices ranging from scoring against references to probing and behavioral testing, grouped together without clear distinctions about what each is set up to do.

Naming the two apart matters because *evaluation* carries a lot of weight in NLP and in benchmark-driven research more broadly. Comparing methods and tracking progress is what it is set up to do, but it is also one of the forces that direct research attention. What can easily be scored against a reference becomes the natural target of new work, and what resists that format is harder to argue for in publications. The open-ended cases sketched in Section 2 tend, on those grounds, to be taken up through *evaluation* rather than *assessment*, with the task reshaped until it produces a comparable score even where that framing does not really fit. *Assessment*, by contrast, is slower to set up and produces findings that do not reduce to a single comparable number, what Gehrmann et al. (2023) call an "incentive mismatch between conducting high-quality evaluations and publishing new models or modeling techniques". The asymmetry between the two is not the problem in itself; what follows from it is that the field's overall direction, what gets researched and optimized, ends up shaped largely by what *evaluation* can take up.

Recent proposals, from measurement theory to evaluation science to impact evaluation, are visibly trying to reach beyond *evaluation* as it is currently used. Naming the two apart is a small move toward that, and a reminder that not every question worth asking about a model is one a benchmark score can answer.

Limitations

The one piece of empirical material in this paper is the keyword search reported in Section 3, and it is too coarse and only quantitative to settle the question, as already noted in Section 3. A search over titles and abstracts cannot tell whether the rise of *assessment* alongside *evaluation* reflects a

change in what authors are doing or a change in what they call it. Settling what the trend actually reflects would itself require the kind of close reading the paper places under *assessment*.

Acknowledgments

I am particularly grateful to Frédéric Béchet, whose conversations during my PhD shaped how I think about NLP and evaluation, and which are at the root of the ideas developed here. I also thank Guy Lapalme for pointing me toward this workshop and encouraging me to formalize my ideas on the topic into a paper, and for his feedback on early versions, and Eliot Maes for his comments on later draft. Thanks also to the reviewers, whose feedback pushed me to clarify several parts of this paper. I am also grateful to Jian-Yun Nie, my postdoc advisor, for generously supporting my participation in this workshop.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *Preprint*, arXiv:2602.16763.

Alexander Bick, Adam Blandin, David J Deming, Nicola Fuchs-Schündeln, and Jonas Jessen. 2026. Mind the gap: Ai adoption in europe and the u.s. Working Paper 34995, National Bureau of Economic Research.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Dunlap, Krishna Mandal, Jacob Steinhardt, Joseph E Gonzalez, and 1 others. 2025. Vibecheck: Discover and quantify qualitative differences in large language models. In *International Conference on Learning Representations*, volume 2025, pages 69177–69205.

Alex Egg, Martin Iglesias Goyanes, Friso Kingma, Andreu Mora, Leandro von Werra, and Thomas Wolf. 2025. Dabstep: Data agent benchmark for multi-step reasoning. *ArXiv preprint*, abs/2506.23719.

Nils Feldhus and Laura Kopf. 2025. Interpreting language models through concept descriptions: A survey. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 149–162, Suzhou, 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. *J. Artif. Int. Res.*, 77.

Itay Itzhak, Eliya Habba, Gabriel Stanovsky, and Yonatan Belinkov. 2026. From feelings to metrics: Understanding and formalizing how users vibe-test llms.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. 2025. The illusion of progress: Re-evaluating hallucination detection in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34728–34745, Suzhou, China. Association for Computational Linguistics.

Brian Kennedy, Eileen Yam, Emma Kikuchi, Isabelle Pula, and Javier Fuentes. 2025. How americans view AI and its impact on people and society. Chapter: "Americans' awareness of AI and views of use in daily life, control over it." Available at https://www.pewresearch.org/science/2025/09/17/ai-in-americans-lives-awareness-experiences-and-attitudes/.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

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. *ArXiv preprint*, abs/1907.11692.

Microsoft AI Economy Institute. 2026. AI diffusion data.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

Ehud Reiter, Roma Robertson, A. Scott Lennox, and Liesl Osman. 2001. Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, Toulouse, France. Association for Computational Linguistics.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, and 1 others. 2023. Model evaluation for extreme risks. *ArXiv preprint*, abs/2305.15324.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of retrieval-augmented generation: A survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu,

Xianpei Han, and Le Sun. 2025. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14402–14418, Suzhou, China. Association for Computational Linguistics.

A Anthology Search: Method and Data

The numbers reported in the main text come from a keyword search run directly on the ACL Anthology XML dump[3], parsed at the per-volume level rather than queried through the website. The corpus covers 117,273 papers from the main *ACL venues (ACL, EMNLP, NAACL, COLING, TACL, CL, LREC, and associated workshops) between 1952 and 2025; 2026 was excluded as only partially indexed at the time of the search.

For each paper, the title and abstract were concatenated and matched against two case-insensitive regular expressions with word boundaries:

```
\b(?:re)?(assess|assesses|assessed|
  assessing|assessment|assessments)\b
\b(?:re)?(evaluate|evaluates|evaluated|
  evaluating|evaluation|evaluations)\b
```

Each paper was then assigned to one of three mutually exclusive buckets, *evaluation only*, *assessment only*, or *both terms*, and yearly shares were computed against the total number of papers indexed for that year. An optional *re-* prefix is allowed (matching *reassess*, *reevaluate*, and their inflections). Other surface variants such as *evaluator*, *evaluative*, or *assessor* were not included.

Abstract coverage in the Anthology XML is uneven before roughly 2016. Several early volumes carry titles only, with the abstract field empty or missing, so pre-2000 rates under-count all three buckets, since only titles contribute to the match. Years before 1990 sit close to zero across all three buckets and were trimmed from the figure to keep the post-1990 trend readable. The post-2016 trend, on which the main argument rests, is not affected.

Figure 2 shows the full series from 1990 onward as the share of papers per year falling into each bucket. Shares are the comparable view across years given that the absolute number of papers grows by a factor of roughly twelve over the period.

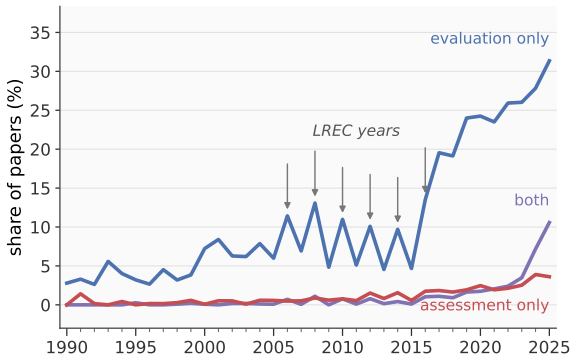

Figure 2: Share of papers in the ACL Anthology using the terms *evaluation*, *assessment*, or both, by year. Buckets are mutually exclusive per paper. Based on a keyword search over titles and abstracts of 117,273 papers across the main *ACL venues (1952–2025). The share of papers using both terms has grown from roughly 1.7% in 2020 to 10.6% in 2025.

[3] https://github.com/acl-org/acl-anthology