Position: Scores Without Context? Rethinking the Role of Evaluation in the Era of LLMs

Jiawei Zhou
Stony Brook University
jiawei.zhou.1@stonybrook.edu

Abstract

Recent years have seen rapid growth in evaluation and benchmarking in NLP, driven by advances in large language models (LLMs). This growth has shifted evaluation from measuring generalization to tracking capability, often without reference to training assumptions. We argue that this creates a conceptual gap: results are frequently interpreted without considering what models could plausibly have learned, rendering many conclusions scientifically underdetermined. We propose an expectation-aware view, where the informativeness of evaluation depends on its relationship to training data, model design, and tasks. We further distinguish between evaluation for scientific understanding and capability tracking, and provide recommendations for aligning evaluation with its intended purpose in the LLM era.

1 Introduction

Evaluation has long been central to progress in natural language processing and machine learning, traditionally tied to generalization: we evaluate to infer what a model has learned beyond its training data. The rise of large language models (LLMs) has further elevated the role of evaluation—evident in growing publication trends—while altering its epistemic function. Modern models are trained on massive, partially undisclosed corpora and evaluated across diverse tasks, from question answering and coding to multimodal reasoning and agentic workflows. As a result, evaluation is often treated as a standalone measurement, increasingly decoupled from learning assumptions, bringing both new opportunities (e.g., capability discovery and application-driven testing) and risks (e.g., misinterpretation and weakened scientific conclusions).

This shift introduces a key ambiguity: when training data are unknown, what does an evaluation result actually tell us? Strong performance may reflect generalization, memorization, or distributional alignment, while failure may indicate model limitations or unrealistic expectations. Without grounding in what a model could plausibly have learned, evaluation is difficult to interpret beyond surface-level performance.

In this paper, we argue that evaluation should be understood relative to expectation. The informativeness of an evaluation result depends on its relationship to the model's training data, design, and the evaluation task. A benchmark score alone is not a scientific conclusion; its value lies in how it compares to what we would expect under these conditions. We refer to this perspective as **expectation-aware evaluation**.

From this viewpoint, we distinguish two roles of evaluation in the LLM era. *Evaluation for scientific understanding* aims to draw conclusions about generalization and model behavior, requiring training-aware analysis. In contrast, *evaluation for capability tracking* treats models as black-box systems and measures what they can do in practice. While both are valuable, they support different claims and require different standards of interpretation.

We analyze recent trends in LLM evaluation and propose a framework to re-ground evaluation in model exposure and expectation, along with recommendations for aligning evaluation design with its intended purpose.

2 The Paradigm Shift: From Generalization to Capability

Evaluation has become a central component of NLP research in recent years. Figure 1 shows the number and proportion of evaluation-focused papers in the ACL Anthology from 2015 to 2025,[1] based on keyword matching.[2] We observe a steady

[1] Data source: https://aclanthology.org/.

[2] We identify evaluation-related papers using keyword pattern matching on paper titles, including variants of `eval*`, `benchmark*`, `leaderboard`, `test set`, `test suite`, `assess*`, `metric*`, `scoring`, `annotation scheme`, `human`

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1048–1054

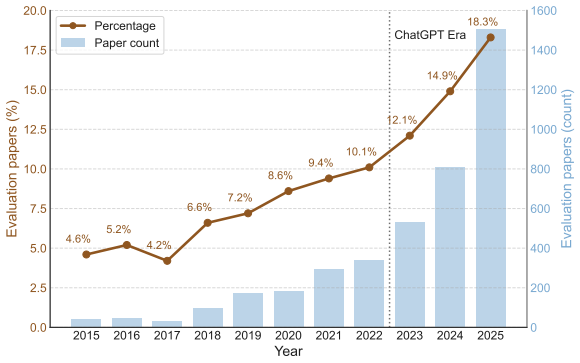

Figure 1: Trends in evaluation-related papers in the ACL Anthology from 2015 to 2025.

increase over time, with a sharp rise following the emergence of large language models. This reflects both the expansion of model capabilities and the growing need to evaluate systems across diverse tasks, including reasoning, coding, multimodal, and agentic tasks.

However, while evaluation is more prevalent than ever, its role has become less clearly defined. This growth is not merely quantitative, but also reflects a qualitative shift in how evaluation is designed and interpreted.

From Generalization to Capability Traditionally, evaluation in NLP is tied to *generalization*: models are trained on well-defined datasets and evaluated on held-out test sets, with performance interpreted as what the model has learned beyond its training data. For example, widely used datasets pre-LLMs such as GLUE (Wang et al., 2018) for language understanding, Squad (Rajpurkar et al., 2016) for question answering, Wikitext-103 (Merity et al., 2017) for language modeling, WMT (Specia et al., 2020) for machine translation, etc. all have explicit training and testing data split and follow a controlled evaluation protocol.

In the LLM era, evaluation increasingly serves a different role. Models are trained on large-scale, heterogeneous, and often undisclosed data (Brown et al., 2020; Achiam et al., 2023; Team et al., 2023), making training distributions difficult to characterize. At the same time, evaluation spans a wide range of tasks, from standard benchmarks to open-ended generation (Chiang et al., 2024), dynamic evaluation (Jain et al., 2024; Li et al., 2025), and system-level applications (Xie et al., 2024; Patwardhan et al., 2025). As a result, evaluation is often treated as a black-box measure of capability—assessing whether a model can perform a task in practice, rather than whether it generalizes beyond training.

`evaluat*`, `automatic evaluat*`, `shared task`, and `dataset`. This heuristic is approximate and intended to capture general trends.

Metric	2018	2025
Train-eval Discussion	17/20 (85%)	3/20 (15%)
Capability-focused eval	0/20 (0%)	13/20 (65%)

Table 1: Small sample audit showing a clear paradigm shift before and after LLM era for the role of evaluation. Results are obtained from manual inspection of 20 randomly sampled evaluation-focused papers from ACL & EMNLP from respective years.

Empirical Illustration of the Shift To illustrate how evaluation practices have changed, we conduct a small-scale audit of evaluation-focused papers from the ACL Anthology. Using the keyword `evaluat` for title matching, we sample 20 papers from 2018 and 20 papers from 2025 by selecting the first matching papers from ACL and EMNLP. We then manually annotate each paper based on whether it explicitly discusses the relationship between training and evaluation data, and its primary evaluation purpose.

We observe a clear shift, illustrated in Table 1. In 2018, 17 out of 20 papers (85%) explicitly frame evaluation in terms of generalization, with consideration of training–test splits or controlled settings. In contrast, only 3 out of 20 papers (15%) in 2025 do so. Instead, most recent papers (approximately 13 out of 20) treat evaluation as a measure of capability, focusing on whether models can perform tasks without reference to training exposure.

While limited in scope, this comparison highlights a clear trend: evaluation is increasingly decoupled from training assumptions and is now more often used to characterize model capabilities rather than to study generalization. Perhaps another well-known example is the evolution of the GPT model evaluations. GPT-2 evaluation (Radford et al., 2019) clearly states "...so we report results on the validation set which has no significant overlap...", GPT-3 (Brown et al., 2020) conducts experiments "Preventing Memorization Of Benchmarks", but GPT-4 (Achiam et al., 2023) no longer includes any mention of "overlap" or "memorization". This reflects a broader shift toward capability-oriented, training-agnostic evaluation.

3 What Can We Learn from Evaluation? An Expectation-Aware Perspective

As evaluation shifts away from being grounded in learning conditions and increasingly becomes a standalone object, it introduces challenges for scientific interpretation. From a knowledge acquisition perspective, evaluation should enhance our understanding of the learning process and guide scientific progress, rather than merely reporting a model's performance on a task.

For example, GPT-5.4 from OpenAI, released in March 2026 (OpenAI, 2026), is reported to achieve "state-of-the-art computer-use capabilities," with a "75.0% success rate" on OSWorld-Verified, "surpassing human performance at 72.4%" (Xie et al., 2024). While impressive, such results primarily indicate that GPT-5.4 is a leading system on this task. What we learn from these scores depends on what we know about the model's training. Computer-use capabilities rely on specific interaction patterns and training tailored to this setting.[3] If such training data and procedures are known, strong performance may not be surprising, and the informativeness of the evaluation depends on how it compares to our expectations. In this sense, evaluation scores often reflect product capability, but provide limited insight into the underlying scientific advances of large language models.

Evaluation without expectation is uninterpretable. We argue that the meaning of an evaluation result depends on an implicit notion of *expectation*: what we anticipate a model to achieve given its training data and design. Formally, we view expectation as a function of training exposure and learning procedure, and the *value of evaluation* lies in how observed performance compares to this expectation:

$$\begin{aligned} V_{\text{eval}} &= f(\mathcal{D}_{\text{train}}, \mathcal{A}, \mathcal{D}_{\text{eval}}) \\ &= \left| P(\mathcal{D}_{\text{eval}}) - E(\mathcal{D}_{\text{train}}, \mathcal{A}) \right| \end{aligned}$$

where $\mathcal{D}_{\text{train}}$ denotes the training data distribution and content, $\mathcal{A}$ the learning algorithm or architecture, and $\mathcal{D}_{\text{eval}}$ the evaluation data. P denotes the observed evaluation performance (e.g., a score or set of judgments), while E represents the expected performance given $\mathcal{D}_{\text{train}}$ and $\mathcal{A}$.[4] The value V_{eval} thus reflects the knowledge gained from the evaluation. In short,

> *The takeaway of evaluation lies in the surprisal of results, not in the scores themselves.*

From this perspective, a benchmark score is not, by itself, a scientific conclusion. Instead, it is a signal whose *informativeness depends on its deviation from expectation*. To illustrate this mental model:

- Case 1, Low Value—Tracking: if $P \approx E$, i.e., results match expectation, evaluation serves primarily as "tracking." It confirms that the model behaves as anticipated (e.g., recalling Wikipedia knowledge for an LLM trained on web-scale data). While useful for monitoring progress and deployment, such results carry low scientific surprisal.
  - Case 1.1, Success matches expectation: evaluation scores are high but expected, confirming known effects. For example, it is expected that fine-tuning improves performance on domain-specific data (Devlin et al., 2019).
  - Case 1.2, Failure matches expectation: evaluation scores are low but expected, reflecting known limitations rather than model failure. For instance, LLMs trained primarily on Wikipedia perform poorly on coding tasks due to domain mismatch. For another example, novel instructions lead to degraded instruction following performance compared to observed ones during training (Sun et al., 2024), which is reasonable.
- Case 2, High Value—Scientific Gain: if $P \gg E$, i.e., performance significantly exceeds expectation, this may provide evidence of generalization or emergent capabilities. For example, GPT-3's strong few-shot generalization (Brown et al., 2020) was not anticipated based on smaller language models, yielding high scientific value and advancing understanding of LLM behavior.
- Case 3, High Value—Diagnostic Insight: if $P \ll E$, i.e., performance falls far below expectation, this may reveal flaws in experimental design, evaluation setup, or underlying assumptions, and can lead to important diagnostic insights. For instance, models aligned for safety may exhibit unexpected unsafe behavior under certain conditions (Greenblatt et al., 2024), highlighting gaps

[3] See https://developers.openai.com/api/docs/guides/tools-computer-use.

[4] For simplicity, we omit the model itself, which is implicit in both the evaluation process and the expectation formulation.

in alignment methods and motivating more robust approaches.

Expectation as a Function of Knowledge The above categorization highlights a scientific versus functional divide in the role of evaluation. Beyond this, the expectation E is itself grounded in our current state of knowledge. Formally, we can express:

$$E(\mathcal{D}_{\text{train}}, \mathcal{A}) = E(\mathcal{D}_{\text{train}}, \mathcal{A} \mid \mathcal{K}(t, p))$$

where $\mathcal{K}(t, p)$ denotes the available conceptual knowledge at time t for a given observer p. As such, expectations are both time-dependent and observer-dependent. For instance, the emergence of in-context learning (Brown et al., 2020) was largely unanticipated around 2019, but is widely expected today. Similarly, advances such as DeepSeek-R1 (Guo et al., 2025), which show that reinforcement learning can induce longer and more structured reasoning, may provide new insights for academic researchers, while being less surprising to industry practitioners such as OpenAI, who have reported similar effects (Chen, 2025) in their o-series models (OpenAI, 2024).

This dependence introduces an information gap in evaluation. When training data and development processes are not fully disclosed, common in industry settings, different groups may form misaligned expectations, leading to divergent interpretations of the same evaluation results. Consequently, evaluation may reflect not only model behavior, but also disparities in knowledge access. To support scientific progress, evaluation and its interpretation should be grounded in shared assumptions and more transparent contexts, enabling more aligned comparisons and contributing to the broader democratization of scientific understanding.

Discussion on Operationalization Given that expectations depend on individual knowledge, experience, and access to information, the proposed expectation-aware evaluation framework is inherently more conceptual than fully operational. In practice, it is difficult to define or estimate expectations consistently across different observers.

Nevertheless, the framework serves as a critical mental model for grounding evaluation in *personalized* scientific knowledge acquisition. It enables individuals to interpret results relative to their own understanding and to integrate new findings into their existing knowledge systems. At the same time, to support *knowledge sharing* and partial operationalization, researchers should explicitly state the assumptions and known facts used to form expectations—even if these are not universally shared—and interpret results relative to them. Such transparency enables more systematic inspection of evaluation outcomes, better contextualization of results, and more aligned knowledge exchange, contributing to the broader democratization of scientific understanding.

4 A Purpose-Based View of Evaluation

In the LLM era, evaluation has shifted from measuring generalization to tracking capability, but without accounting for training data and expectations, such evaluations are difficult to interpret scientifically. As evaluation now serves multiple roles with differing goals and assumptions, a key source of confusion is the conflation of these roles. We distinguish between two primary regimes below.

Evaluation for Scientific Understanding. This regime aims to advance model design and learning theory by uncovering what a model has learned and why it performs as it does. It is inherently *white-box*: evaluation is grounded in the relationship between training data, model architecture, and evaluation tasks, and interpreted through the expectation-aware framework in Section 3. The goal is not merely to measure performance, but to explain it and derive scientific insight. For example, early instruction-tuning work (Wei et al., 2021) demonstrates zero-shot generalization in following novel task instructions.

Evaluation for Capability Tracking. In contrast, this regime focuses on measuring what a model can do in practice, often without reference to its training data. It is fundamentally *black-box*: the model is treated as an opaque system—an "alien intelligence"—and evaluation asks whether it can perform tasks such as coding (Merrill et al., 2026), reasoning (Sun et al., 2025), or end-to-end workflows (Xie et al., 2024). This form of evaluation is particularly prevalent in the LLM era, where training data are large-scale and partially undisclosed. Representative examples include benchmarks such as SWE-Bench (Jimenez et al., 2023) and GPQA (Rein et al., 2024).

Discussion These regimes answer different questions: capability tracking asks what models can do, while scientific evaluation asks what we can learn from them. Both are essential but require

different standards and support different claims. Confusion arises when black-box results are taken as evidence of generalization, or when white-box evaluations are judged solely by benchmark performance. A concrete consequence of this distinction—the changing role of memorization—is discussed in Appendix A.

5 Implications for Evaluation Practice

These shifts suggest that evaluation should be designed and interpreted with explicit awareness of its purpose. Scientific evaluation requires training-aware, white-box designs that support conclusions about generalization and model behavior, while capability evaluation should be framed as black-box measurement, with claims limited to observable performance. Conflating these regimes risks over-interpreting results and misaligning evaluation with scientific claims.

More broadly, evaluation should be interpreted relative to what a model could plausibly have learned. Accounting for training data and model design helps distinguish memorization from generalization and avoid unsupported conclusions about reasoning or learning. Aligning evaluation purpose, methodology, and interpretation is essential to keep evaluation both scientifically meaningful and practically useful in the LLM era.

Limitations

This paper presents a conceptual framework for understanding evaluation in the LLM era, but it has several limitations. First, our purpose-based taxonomy is intentionally coarse-grained. While we distinguish between scientific (white-box) evaluation and capability (black-box) evaluation, many important subcategories are not explicitly modeled. For example, evaluations aimed at probing internal model representations or behaviors, as well as meta-evaluation studies that analyze or improve evaluation metrics, may not fit cleanly into this binary categorization.

Second, our empirical analysis is limited in scope. The paper audit is based on a small sample of papers and simple annotation criteria, and the trend analysis relies on keyword-based identification of evaluation-related work. These analyses are intended to be illustrative rather than comprehensive, and a more systematic study would be needed to precisely characterize broader trends in the field.

Finally, our framework emphasizes the role of training data and expectation in interpreting evaluation results, but in practice, training data for large language models are often unavailable or only partially disclosed. As a result, applying expectation-aware evaluation may be challenging in real-world settings, and further work is needed to operationalize this perspective under realistic constraints.

Ethical Considerations

This paper presents a conceptual analysis and involves no human subjects or sensitive data. However, misinterpretation of evaluation results may lead to overstated claims about model capabilities or safety. By advocating for expectation-aware evaluation and clearer alignment between evaluation purpose and interpretation, this work aims to support more transparent and responsible use of evaluation in AI systems.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We also thank our colleagues and collaborators for helpful discussions that informed this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen. 2025. On reasoning improvements in openai o-series models. https://x.com/markchen90/status/1884303237186216272. Accessed: 2026-03-21.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An

open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

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.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, and 1 others. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Yanhong Li, Tianyang Xu, Kenan Tang, Karen Livescu, David McAllester, and Jiawei Zhou. 2025. Okbench: Democratizing llm evaluation with fully automated, on-demand, open knowledge benchmarking. *arXiv preprint arXiv:2511.08598*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, and 1 others. 2026. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*.

OpenAI. 2024. Introducing OpenAI o1. Accessed: 2026-03-21.

OpenAI. 2026. Introducing GPT-5.4. Accessed: 2026-03-20.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, and 1 others. 2025. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, and 1 others. 2020. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the fifth conference on machine translation*, pages 76–91.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Jiuding Sun, Chantal Shaib, and Byron Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *International Conference on Learning Representations*, volume 2024, pages 48103–48141.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

A A Consequence of Evaluation Paradigms: The Changing Role of Memorization

A direct consequence of the distinction between white-box (scientific) and black-box (capability) evaluation is a shift in how memorization is viewed. In the white-box regime, where evaluation is tied to training data and generalization, memorization has traditionally been treated as a confounding factor to be minimized through careful dataset design. In contrast, in the black-box regime, where the goal is to assess what a model can do in practice, memorization becomes a functional capability: large language models are expected to store and retrieve substantial amounts of knowledge to support downstream tasks.

For example, the 2018 ACL paper "Split and Rephrase: Better Evaluation and Stronger Baselines" (Aharoni and Goldberg, 2018), which we include in our small-sample audit (Section 2), explicitly states that "The original data-split is not suitable for measuring generalization, as it is susceptible to 'cheating' by fact memorization," and proposes a revised data split to mitigate this issue. Such practices were standard prior to the LLM era. In contrast, this concern is less frequently addressed today, partly due to limited access to training data and development details. Nevertheless, awareness of the memorization–generalization distinction persists: among the 20 evaluation papers we sampled from ACL and EMNLP 2025, although only 3 explicitly discuss the training–evaluation relationship, 8 papers (40%) acknowledge potential data contamination or memorization and attempt to mitigate it, for example by constructing evaluation data from later time periods.

This shift complicates the interpretation of evaluation results. Strong performance on knowledge-intensive benchmarks may reflect either generalization or exposure to similar data during training, and distinguishing between the two is often difficult when training data are opaque. From the perspective of expectation-aware evaluation, this reinforces the need to interpret results relative to what a model could plausibly have learned. More broadly, modern evaluation increasingly measures both learning and knowledge capacity, blurring a distinction that was central in earlier paradigms of machine learning.