MIT Press • Open Encyclopedia of Cognitive Science

AI Model Evaluation

Jennifer Hu[1]

[1]Johns Hopkins University

MIT Press

Published on: Feb 05, 2026

DOI: https://doi.org/10.21428/e2759450.b0757e8d

Artificial intelligence (AI) model evaluation is the practice of measuring whether an AI model has a particular kind of ability or knowledge. The goals of evaluation range from assessing performance on a concrete task (e.g., spam email classification) to gathering evidence for a high-level cognitive capacity (e.g., understanding what another person is thinking). Regardless of the goal, model evaluation involves computing performance using a task, dataset, and metric. When evaluating a cognitive capacity of an AI model, performance is then typically used to make inferences about whether the model possesses the ability or knowledge of interest. The nature of AI models' cognitive abilities, as inferred through model evaluation, has profound implications for cognitive science, informing questions about the representation and learning of concepts, categories, linguistic structure, and meaning. It remains debated how best to evaluate models' cognitive abilities and compare them to those of humans.

History

Early AI researchers often used toy datasets to evaluate whether artificial neural networks could capture signature patterns of human learning (Rumelhart et al., 1987). Although not guided by cognitive questions, the UCI Machine Learning Repository provided one of the first centralized resources for *benchmarking*, or the use of standardized datasets and metrics to compare different systems. The repository compiled benchmarks for classic machine learning algorithms such as *classification* (sorting data points into predefined categories) and *clustering* (grouping similar data points together).

As deep learning models grew more sophisticated, there was a need for benchmarks to systematically measure more complex cognitive abilities. A well-known example of an early modern benchmark is ImageNet (Deng et al., 2009), which enabled large-scale evaluation of image classification. With the success of neural networks in natural language processing, a wave of benchmarks then began targeting more complex linguistic abilities such as language understanding (e.g., Wang et al., 2019) in the late 2000s and 2010s. These benchmarks typically involved training or fine-tuning a model on a *training set* and then evaluating the model on an identically distributed *test set* that the model had not yet been exposed to.

Researchers quickly identified several limitations of this general approach (Bowman & Dahl, 2021; Linzen, 2020; Schlangen, 2021). As AI models rapidly advanced during this time, several major benchmarks were "solved" (e.g., a model achieving human-level performance) within a year after their release (Kiela et al., 2021). In addition, models could often succeed at benchmarks using heuristics learned through statistical association instead of genuine generalization (McCoy et al., 2019). The early 2020s saw some attempts to address these issues—for example, through dynamically constructed or larger benchmarks (Kiela et al., 2021; Srivastava et al., 2023). However, some researchers remain skeptical about the general enterprise of benchmarking (Raji et al., 2021).

In contrast to prior AI models that were optimized for a specific task, modern large language models can be evaluated on any task that can be represented in natural language [see Large Language Models]. As such, evaluation has shifted away from formal linguistic tasks (e.g., labeling word classes or analyzing

sentence structure) toward tasks involving more general high-level cognition (e.g., reasoning by analogy) and domains involving human expert-level knowledge (Phan et al., 2025) [see Analogy].

Core concepts

Key definitions

A cognitive *construct* is an abstract ability or form of knowledge that the researcher seeks to measure (e.g., the ability to do math). The construct of interest cannot be measured except indirectly through performance on a task. A *task* specifies a concrete mapping from inputs to outputs that recruits the construct of interest (e.g., multiplication of three-digit integers).

A *stimulus* is a single instance of an input to the model (e.g., a string containing a multiplication problem). A *dataset* is a collection of stimuli. The core of a dataset is the test set, which contains the stimuli intended for evaluating the model. It is typically assumed that the model has not encountered the test set before evaluation (but see below for discussion of data contamination). Some datasets may also include a training set (to allow the model to learn the task before evaluation) and/or a validation set (to measure progress during training). Most datasets used to evaluate modern AI models only include a test set. Such benchmarks are said to test *emergent* abilities: that is, abilities that arise during training on a different objective (such as next-token prediction) instead of being learned through direct training.

A *metric* is a way of summarizing a model's outputs on a test dataset into a single score. Metrics can be intrinsic or extrinsic. Intrinsic metrics are those that can be computed using the dataset itself, such as accuracy, which measures correctness, or perplexity, which measures how well a probability distribution predicts a new sample. Extrinsic metrics are those that require information from an external source, such as a rating of the model's output obtained from human or AI annotators.

As a simple example, suppose a researcher wants to evaluate whether a particular model can do math. The researcher might decide to measure this *construct* through the *task* of computing the product of three-digit integers. The researcher would then build a *dataset* by creating 1,000 *stimuli*, each of which is a string of the form "Calculate X * Y =," in which X and Y are three-digit integers. The model's outputs might then be scored through two intrinsic *metrics*: (1) accuracy, or the proportion of stimuli in which the model's output matches the correct answer, and (2) mean error, or the difference between the model's output and the correct answer (averaged across all stimuli).

Evaluation design

Designing an evaluation involves several key decision points. One factor is choosing a task that faithfully measures the construct of interest. Another important factor is deciding how to obtain outputs from the model, for example, through measurements of internal values versus high-level prompting or through free response versus forced choice. These methodological decisions can lead to drastically different conclusions about model abilities (Hu et al., 2024; Lampinen, 2024) and may also serve different goals,

such as adversarially probing the limits of models' abilities versus revealing the potential of a model in ideal settings.

Inferring model abilities from evaluations

Evaluating whether an AI model has a certain construct involves making inferences about a model's *competence* (underlying ability or knowledge) based on a model's *performance* (demonstration of that knowledge through a specific task; Firestone, 2020; Millière & Rathkopf, 2025). A key challenge of this approach is that models can succeed at evaluations without necessarily capturing the cognitive ability being tested (McCoy et al., 2019). Conversely, models can fail at evaluations due to the auxiliary challenges associated with performing the task instead of a genuine lack of ability (Hu & Frank, 2024). Researchers also use evaluations to understand how factors like a model's training data or architecture support different cognitive abilities, but these inferences are only feasible with openly accessible models (Frank, 2023).

Questions, controversies, and new developments

Comparing models to humans

Model evaluation often involves comparing outputs from a model to behavior from humans. It remains debated how these model–human comparisons should be made. Some argue that models and humans should be compared under identical evaluation paradigms (Leivada et al., 2024), whereas others argue that evaluation paradigms should be designed according to models' and humans' unique computational mechanisms (Hu et al., 2024; Lampinen, 2024). Another topic of debate is whether models need to use human-like strategies for performing a task in order to genuinely possess the ability of interest (*anthropocentrism*; Millière & Rathkopf, 2025).

Data contamination

Data contamination occurs when evaluation stimuli are present in the model's training data (Dodge et al., 2021). If a stimulus has already been seen by a model, then the model's performance during testing may not reflect genuine generalization but rather a form of memorization (Magar & Schwartz, 2022). Because modern models are trained on massive amounts of data and/or proprietary datasets, it is often infeasible to know whether a particular stimulus was present in the training data. Private corporations also frequently evaluate models on nonpublic datasets without disclosing how such evaluations were designed.

New horizons of evaluation

As AI models are increasingly used by the general public in open-ended settings, there has been a movement towards *vibes-based evaluation* instead of formal, standardized evaluation. Some researchers have also advocated for using AI models to automate various parts of the evaluation process, for example,

to annotate another model's outputs (Zheng et al., 2023) or to synthetically generate stimuli for evaluation (Kim et al., 2025).

Broader connections

The ability of deep neural networks to learn intelligent behavior has wide-ranging connections to fundamental questions in cognitive science (Kanwisher et al., 2023; Rumelhart et al., 1987). Comparing the cognitive abilities of AI models to those of humans also has parallels to comparative psychology with nonhuman animals (Buckner, 2023; Firestone, 2020) [see Animal Cognition]. The inferences drawn through model evaluation also have deep implications for society more broadly, as understanding the capabilities and limitations of models is critical for making decisions about the regulation, governance, and personhood of AI systems.

Further reading

- Fourrier, C. (2024). The LLM evaluation guidebook. GitHub. https://github.com/huggingface/evaluation-guidebook
- Ivanova, A. A. (2025). How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, *9*(2), 230–233. https://doi.org/10.1038/s41562-024-02096-z

References

- Bowman, S. R., & Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv*. https://doi.org/10.48550/arXiv.2104.02145 ↩
- Buckner, C. (2023). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *The British Journal for the Philosophy of Science*, *74*(3), 681–712. https://doi.org/10.1086/714960 ↩
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ↩
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1286–1305). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98 ↩

- Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117 ↩
- Frank, M. C. (2023). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour*, *7*(11), 1825–1827. https://doi.org/10.1038/s41562-023-01732-4 ↩
- Hu, J., & Frank, M. (2024). Auxiliary task demands mask the capabilities of smaller language models. *arXiv*. https://doi.org/10.48550/arXiv.2404.02418 ↩
- Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121. https://doi.org/10.1073/pnas.2400917121 ↩
- Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240-254. https://doi.org/10.1016/j.tins.2022.12.008 ↩
- Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. *arXiv*. https://doi.org/10.48550/arXiv.2104.14337 ↩
- Kim, S., Suk, J., Yue, X., Viswanathan, V., Lee, S., Wang, Y., Gashteovski, K., Lawrence, C., Welleck, S., & Neubig, G. (2025). Evaluating language models as synthetic data generators. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 6385–6403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.320 ↩
- Lampinen, A. (2024). Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Computational Linguistics*, *50*(4), 1441-1476. https://doi.org/10.1162/coli_a_00525 ↩
- Leivada, E., Dentella, V., & Günther, F. (2024). Evaluating the language abilities of large language models vs. humans: Three caveats. *Biolinguistics*, *18*. https://doi.org/10.5964/bioling.14391 ↩

- Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? *arXiv.* https://doi.org/10.48550/arXiv.2005.00955

  ↩
- Magar, I., & Schwartz, R. (2022). Data contamination: From memorization to exploitation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 157–165). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.18

  ↩
- McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv.* https://doi.org/10.48550/arXiv.1902.01007

  ↩
- Millière, R., & Rathkopf, C. (2025). Anthropocentric bias in language model evaluation. *Computational Linguistics*, 1-10. https://doi.org/10.1162/COLI.a.582

  ↩
- Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., … Hendrycks, D. (2025). Humanity's last exam. https://doi.org/10.48550/arXiv.2501.14249

  ↩
- Raji, D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (2021). AI and the everything in the whole wide world benchmark. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1). Curran.

  ↩
- Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1987). *Parallel distributed processing* (Vol. 1). MIT Press.

  ↩
- Schlangen, D. (2021). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 2, pp. 670–674). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.85

  ↩
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek,

A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv.* https://doi.org/10.48550/arXiv.2206.04615

↩

- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran.

  ↩

- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv.* https://doi.org/10.48550/arXiv.2306.05685

  ↩