

Epistemic Twins: Enabling a Symbolic Science of Language Model Knowledge [Vision Paper]

Simon Razniewski¹ Shrestha Ghosh² Luca Giordano¹
Yujia Hu¹ Josua Kowalzik¹ Tuan-Phong Nguyen³

¹ ScaDS.AI & TU Dresden

² University of Tübingen

³ VNU University of Engineering and Technology

Abstract

Large Language Models (LLMs) are impactful yet opaque artifacts. Subsymbolic by design, they are defined by billions of numeric weights that interact in a largely inscrutable manner. Current analysis paradigms are either black-box benchmarks that test model performance on pre-defined tasks, or mechanistic interpretability approaches that trace back outputs to specific weights. Both analysis methods are limited by the experimenter’s hypothesis space - one must know what to look for to find it. In this article, we argue for a third, radically different analysis paradigm: **Epistemic Twins**. We propose constructing large-scale symbolic approximations of LLMs in human-readable formats. This enables the comprehensive inspection of factual knowledge (or beliefs) inherent in the model without predefining hypotheses, facilitating large-scale auditing towards better understanding and explainability.

1 Introduction

LLMs have transformed machine intelligence, and serve as the cornerstone of the current wave of artificial intelligence. Despite their impact, our understanding of their functioning and capabilities remains vague. The public discourse is often driven by hyped marketing language, with only a weak grounding in rigorous statistical or theoretical analyses of LLM knowledge and functioning. Consequently, there is a significant barrier to LLM deployment in safety-critical applications.

Currently, two paradigms dominate the analysis of LLMs (see Fig. 1):

1. *Black-box benchmarks*: This approach ignores the internal functioning of LLMs and focuses on its input-output-behavior in test cases, often prompting it with thousands of them to assess specific capabilities (Ribeiro et al., 2020; Chang et al., 2024). While most common, this paradigm is inherently limited

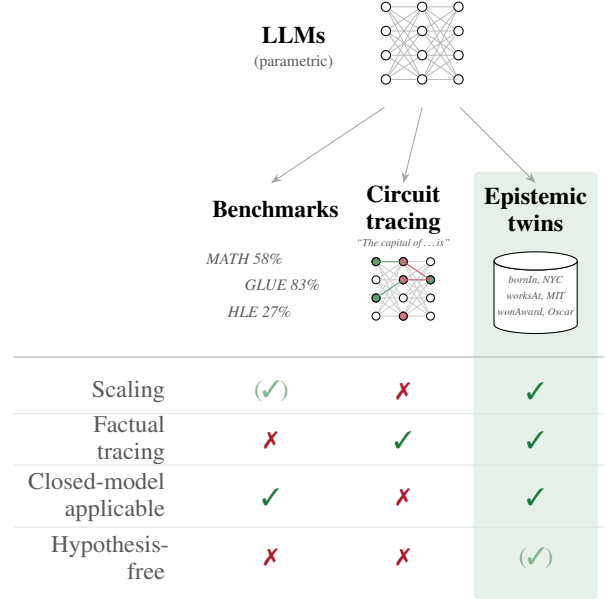


Figure 1: Three paradigms for analyzing LLMs. Benchmarks operate at reasonable scale, but do not enable factual tracing nor are hypothesis-free. Circuit tracing enables tracing, but is hard to scale, and requires narrow hypotheses. We propose Epistemic Twins as novel analysis paradigm.

by the coverage of the benchmarks (“availability bias” (Tversky and Kahneman, 1973)); it cannot reveal behaviors or knowledge not anticipated by the benchmark designers (Raji et al., 2021; Bean and others, 2025).

2. *Circuit tracing (mechanistic interpretability)*: This approach attempts to open the black box, tracing responses back to individual circuits and neurons to establish causal links (Rai et al., 2024). While this enables real explainability, it is computationally expensive and fragile to distribution shifts, with possibly parallel explanations for a single task, and offers little insight into the global knowledge state of the model (Méloux et al., 2025; Sharkey et al., 2025; Somvanshi et al., 2026).

In light of these limitations, we advocate for a third, complementary and novel paradigm: **Epistemic Twins**. An epistemic twin is a symbolic, comprehensive and comprehensible model of the factual knowledge stored within an LLM, stored in a classical relational or RDF database/knowledge base.

If this epistemic twin is sufficiently faithful to its parametric LLM, i.e., approximates its knowledge well, then LLM providers and users can then answer questions about the epistemic state of the LLM by established database technology over the twin. Epistemic twins offer four opportunities:

1. **Scaling:** They can offer insights about a sample of unprecedented size, or even the entirety of an LLM’s knowledge, containing millions of entities and hundreds of million of assertions, where even large benchmarks only operate at the scale of a few tens of thousands of instances.
2. **Factual tracing:** They enable to trace LLM responses to factual prompts to specific statements within (or not represented in) the epistemic twin. This way, LLM behavior can be explained by its knowledge state.
3. **No weight access needed:** Unlike mechanistic interpretability, which requires access to an LLM’s weights, epistemic twins can be constructed even for API-only models (see Sec. 3).
4. **Hypothesis-free:** They are hypothesis-free human-readable representations of the LLMs knowledge, i.e., they do not underlie the cognitive availability bias (Tversky and Kahneman, 1973) that benchmarks and circuit tracing face, where the experiment designers predisposition crucially limits the possible experiment results.¹

In this article, we present both the idea of epistemic twins, as well as a call to action. We argue that for auditability, it should become a standard that akin to other documentation in model cards (Mitchell et al., 2019), model providers release epistemic twins along with newly released LLMs. We proceed by introducing the concept of epistemic twins (Section 2), their construction (Section 3), formal

approaches to estimate approximation quality (Section 4), and exemplar analyses enabled by symbolic twins (Section 5). We conclude with a perspective on cost and a call for action (Section 7).

2 What is an Epistemic Twin?

The central premise of our case relies on two observations.

1. **Translating subsymbolic knowledge into symbolic form:** LLM knowledge is by default subsymbolic and numeric: When one downloads a model, all one gets are billions of real-valued weights, with no apparent meaning or interpretation. To be truly auditable, this subsymbolic form must be translated into a human-readable, symbolic form.
2. **Comprehensive knowledge mapping:** Current assessment by benchmarks is sample-based, and hence strongly subject to experimenter bias. An epistemic twin aims to extract the *entirety* of the LLM’s factual knowledge, thereby enabling serendipitous discoveries and statistical analyses over the whole underlying knowledge base.

A epistemic twin manifests as a database or knowledge base (KB) (Weikum et al., 2021), which provides both a human-readable data format, and a format readily accessible to statistical analysis via SQL or SPARQL queries.

The reader may wonder how epistemic twins differs from simply releasing an LLM’s training corpus, as available, e.g., for the OLMo series (Groeneveld et al., 2024; Ettinger et al., 2025). We emphasize that training corpora serve the purpose of pipeline reproducibility, not explainability. Knowing that a fact existed in the training data does not guarantee the LLM learned it (Kandpal et al., 2023), nor does it reveal the hallucinations, false inferences, or conflation of the model generated during training. An epistemic twin represents the model’s actual final *post-training state*, not its pre-training input.

3 How to Construct Epistemic Twins

Locating and isolating knowledge in LLMs from other capabilities has been advocated by various works (Geva et al., 2021; Meng et al., 2022; Mahowald et al., 2024). Large-scale extraction

¹Of course, any individual analysis conducted on top of an epistemic twin still require a hypothesis, therefore, we used the symbol (✓) in Fig. 1.

and transformation of LLM knowledge into symbolic form is an emerging field, yet promising approaches already exist. [Cohen et al. \(2023\)](#) introduced LMCrawl, a recursive prompt pipeline for extracting the internalized knowledge graph by a breadth-first search (BFS). This method was scaled by [Hu et al. \(2025\)](#) as the GPTKB methodology, to enable the concurrent extraction of millions of entities and statements. The process can be formalized as a graph traversal algorithm over the latent knowledge graph inside the LLM:

1. **Seed Initialization:** The process begins with a set of seed entities (e.g., a list of popular head entities).
2. **Recursive Expansion:** The system prompts the LLM to generate statements associated with the current entities.
3. **Traversal:** Named entities within the generated statements become new subject nodes for the next iteration.
4. **Termination:** The recursion continues until a stopping criterion is met, such as graph depth, a compute budget, or a confidence threshold (e.g., when the average confidence of newly discovered facts drops below 90%).

The exact choice of formalism for statements is important. Knowledge bases typically rely on subject-predicate-object triples, though these have limitations in terms of expressivity, which in turn affects the faithfulness of the twinning. The formalism choice requires a balance between structure and expressivity, and this is an open challenge of the epistemic twin concept.

Epistemic twins can be constructed open-domain, or restricted to a specific topic. In the open-domain, the GPTKB series presents the most mature prototypes, with *GPTKB 1* ([Hu et al., 2025](#)), *GPTKB 1.5* ([Hu et al., 2026a](#)) and *GPTKB 2.0* ([Hu et al., 2026b](#)) constructing epistemic twins for GPT-4o-mini, GPT-4.1, and GPT-5.1, respectively. These KBs reach scales of 2-6M entities and 40-100M statements each. For topic-specific extractions, [Giordano and Razniewski \(2026\)](#) showed how the GPTKB methodology can be restricted in scope, and constructed *babylonGPTKB*, *tBBTGP-TKB* and *dax40GPTKB*, comprising 13k, 1k and 2k statements, respectively, extracted from GPT-4.1-mini.

Similar methodologies exist, for instance, non-recursive list-based extraction schemes ([Hao et al., 2023](#); [Parović et al., 2025](#)), taxonomic knowledge ([Funk et al., 2023](#); [Zeng et al., 2024](#)), or a text-oriented adaptation that has already extracted 1M Wikipedia-style articles ([Saeed and Razniewski, 2026](#)). The underlying codes are publicly released, and can be deployed against public as well as private LLM endpoints, hence, we expect a proliferation of similar knowledge extractions in the near future.

4 How Can we Guarantee Twinning Faithfulness?

For epistemic twins to be useful, one must be able to assess their *twinning faithfulness*, i.e., the degree to which they truly reflect the parametric knowledge inside the modeled LLM. We propose three key dimensions of twinning faithfulness:

1. **Twinning accuracy (precision):** The fraction of assertions in the epistemic twin that are actually beliefs held by the twinned LLM.²
2. **Twinning coverage (recall):** The fraction of (factual) beliefs of the LLM that are contained in the epistemic twin.
3. **Twin determinism (reproducibility):** The degree to which twin construction is a reproducible process.

The perfect twin would be 100% accurate, cover all of the twinned LLM’s knowledge, and be unique. However, as LLMs are probabilistic constructs, perfect twinning is not possible. We propose the following approaches to measure twinning faithfulness:

1. For twinning accuracy, for a given *s-p-o* triple in the twin, one can simply prompt the LLM with *s* and *p* (and some paraphrases), and observe whether *o* is indeed returned by the LLM. For example, for GPTKB 1.5, this way, we observe a twinning accuracy of 74%.
2. For twinning coverage, one can use abundance estimators from ecology on random walks

²Note that *Twinning accuracy* should not be mixed up with the Twin’s *factual accuracy*. For example, GPTKB 1.0 is reported to have a factual accuracy of 31%, i.e., 31% of its statements are true w.r.t. the real-world, but its twinning accuracy relative to GPT-4o-mini is far higher, at around 60%. The converse, i.e., a high factual accuracy, but low Twinning accuracy, is also possible.

Analysis	Method	Example from GPT-4.1 twin (GPTKB 1.5)
Topical coverage	Count relevant entities	GPT-4.1 knows more than 1M persons and more than 100k companies .
Gender bias	Count gender identity facts	GPT-4.1 has a positive bias towards females (1.16:1); however, this strongly flips outside English-language nationalities.
Sensitive content	Search for sensitive attributes	GPT-4.1 stores virtually no private phone numbers and addresses, but thousands of statements about ethnic, religious, and sexual identities.
Logical consistency	Check satisfaction of logical rules	GPT-4.1’s knowledge is highly inconsistent: only 15% of border-relation facts are present in both directions.
Knowledge cutoff	Observe latest dates	GPT-4.1’s knowledge cutoff is mid-2024 .

Table 1: Examples of analyses enabled by epistemic twins. Taken from (Ghosh et al., 2025).

over the model’s knowledge graph (Luggen et al., 2019), or BFS property extrapolation (Kurant et al., 2011), to predict the fraction of all factual knowledge covered within the epistemic twin. For example, for GPTKB 1.5, by BFS run extrapolation, we estimate a coverage of 42% of all named entities within GPT-4.1.

- For twin determinism, one can use repeated runs of the twin construction process with different seed nodes or random seeds, and observe the overlap in the resulting twins. For example, based on repeated runs of the construction process for babylonGPTKB, it was observed that repeated runs overlap in roughly two thirds of their named entities, even from different starting entities (Giordano and Razniewski, 2026). This indicates that despite some variability, beliefs remains largely consistent across extraction runs.

Of course, the faithfulness dimensions and metrics proposed above are only a start. Future research will extend and refine the methods for defining and measuring twinning faithfulness.

5 What Analyses Are Enabled by Epistemic Twins?

The existence of multiple and faithful epistemic twins, as imagined in Figure 2 can transform LLM research and certification. It enables:

- Quantification of Knowledge:** We can complement estimates via parameter counts (Allen-Zhu and Li, 2024) with estimates based on counting actual “knowledge units” (facts). This allows us to compare models based on information density. For example, based on BFS extrapolation, the GPT-4.1 epistemic twin GP-

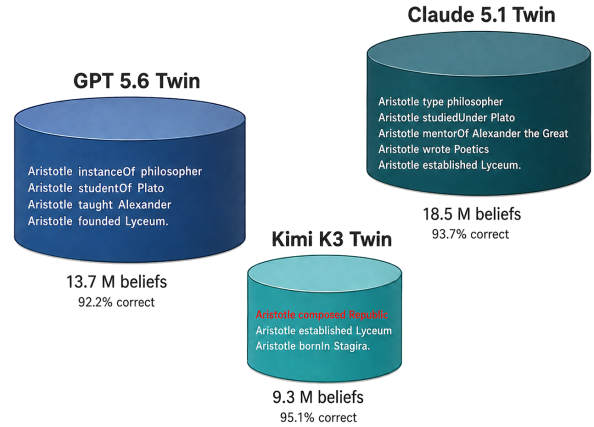


Figure 2: The envisioned twins enable the quantitative and qualitative comparisons of LLM’s factual knowledge.

TKB 1.5 can be used to predict that GPT-4.1 knows a total of around 15 M named entities.

- Global Accuracy Estimation:** By evaluating the epistemic twin against reference sources, we can calculate global accuracy scores that are not overfitted to specific benchmark test sets. For example, (Hu et al., 2026a) observe a global accuracy of 76% of the GPT-4.1 twin GPTKB 1.5, well below the numbers reported in typical factuality benchmarks.
- Bias and Topic Auditing:** An epistemic twin allows for topic modeling over the model’s beliefs. We can visualize the distribution of knowledge (e.g., Western vs. Non-Western history) and detect systemic biases (e.g., gender associations in professional roles) by querying the database rather than probing the model stochastically. More specific examples are given in Table 1.

Such analyses can benefit three types of stakeholders. For **users**, this aids in model selection. For

providers, it offers a transparency mechanism for accountability. For **regulators**, it provides a tangible artifact to evaluate compliance without needing access to weights or training data.

6 Challenges and Research Opportunities

Realizing epistemic twins as a practical auditing instrument raises several open challenges:

1. **What are we twinning?** Whether LLMs can be described as having beliefs remains contested (Bender et al., 2021; Herrmann and Levinstein, 2025). Recent work argues that belief-like representations can be studied under explicit adequacy criteria (Herrmann and Levinstein, 2025), and mechanistic evidence supports the existence of belief-like internal states (Mendoza and Søgaard, 2026). Still, the question remains unsettled. Pragmatic evidence suggests that factual knowledge may be encoded in task-specific ways (Elhelo et al., 2026), or that models can internally encode information that is not faithfully reflected in their surface outputs (Orgad et al., 2025; Gekhman et al., 2025). The concept of twinned knowledge therefore needs further refinement.
2. **Representation and faithfulness.** LLMs possess many capabilities beyond factual knowledge, e.g., translation, planning, or creative generation. Our focus is factual knowledge, but even this narrower target inherits the classical difficulties of knowledge representation: all models are wrong (lossy), and a careful balancing between expressivity and structure needs to guide the actual representation.
3. **Cost and scale.** Constructing epistemic twins is resource-intensive. For example, GPTKB 1.5, a 100-million-triple prototype derived from GPT-4.1, cost approximately \$14,000 in API calls (Hu et al., 2026a). Higher-fidelity twins may plausibly cost hundreds of thousands of dollars, especially if they require repeated runs or robustness analyses.

7 Call for Action

The challenges are substantial, and the costs are nontrivial. Yet they still pale relative to the costs spent for training frontier models. The Stanford AI Index estimates the compute cost of training GPT-4

at \$78 million and Gemini Ultra at \$191 million (AI Index Steering Committee, 2024); a \$100,000 epistemic twin would therefore correspond to just 0.1% of the training costs. Furthermore, the investment can start small. Topic-specific twins can be constructed at single-digit-dollar budgets (Gior-dano and Razniewski, 2026).

Epistemic twins offer a complementary path toward model accountability. Benchmarks measure performance on predefined tasks, and mechanistic interpretability studies internal computations, but neither by itself provides a broad, queryable, model-specific map of factual knowledge. Epistemic twins could fill this gap by making model knowledge inspectable, comparable, and auditable at scale. In this sense, they extend the logic of model cards (Mitchell et al., 2019): not only reporting how a model performs, but exposing what it appears to know under a transparent extraction protocol.

We therefore call for a community effort around epistemic twins: more and better construction methods, more validation of faithfulness, and public twin artifacts for important model families. If we want to understand model knowledge at scale and with less experimenter bias than fixed-question benchmarks allow, epistemic twins are a timely and concrete next step.

Limitations

Factual knowledge focus Our concept of epistemic twins currently focuses strictly on declarative, factual knowledge. LLMs also possess other knowledge, e.g., procedural knowledge on how to code or multiply, or linguistic knowledge on how to rhyme or argue. These non-factual capabilities are not currently captured by the fact-oriented knowledge graph structure proposed here.

Probabilistic-to-boolean LLMs are inherently probabilistic, while knowledge bases by default represent knowledge in Boolean form (true/false), and the epistemic twins constructed so far follow this representation. A translation from probabilities to a Boolean format necessarily leads to some information loss. It is in principle possible to expand representation to a probabilistic format, though probabilistic databases are much less established, and the translation of LLM-internal token likelihoods into truth probabilities is a nontrivial task as well (Jiang et al., 2020).

References

- AI Index Steering Committee. 2024. [AI index report 2024](#). Technical report, Stanford Institute for Human-Centered Artificial Intelligence (HAI).
- Zeyuan Allen-Zhu and Yuanzhi Li. 2024. [Physics of language models: Part 3.3, knowledge capacity scaling laws](#). *arXiv*.
- Andrew M. Bean and others. 2025. [Measuring what matters: Construct validity in large language model benchmarks](#). In *neurIPS*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *FAT*.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and others. 2024. [A survey on evaluation of large language models](#). *Transactions on Intelligent Systems and Technology*.
- Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. [Crawling the internal knowledge-base of language models](#). In *Findings of EACL*.
- Amit Elhelo, Amir Globerson, and Mor Geva. 2026. [LMs as task-specific knowledge bases: An interpretability analysis](#).
- Ettinger et al. 2025. [Olmo 3](#). *arXiv*.
- Maurice Funk, Sebastian Hosemann, Jean Christoph Jung, and Carsten Lutz. 2023. [Towards ontology construction with language models](#). In *KBC-LM*. Knowledge Base Construction from Pre-trained Language Models (KBC-LM).
- Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. [Inside-out: Hidden factual knowledge in LLMs](#).
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *EMNLP*.
- Shrestha Ghosh, Luca Giordano, Yujia Hu, Tuan-Phong Nguyen, and Simon Razniewski. 2025. [Mining the mind: What 100m beliefs reveal about frontier LLM knowledge](#). *AKBC@EMNLP*.
- Luca Giordano and Simon Razniewski. 2026. [Foundations of LLM knowledge materialization: Termination, reproducibility, robustness](#). In *EACL Findings*.
- Groeneveld et al. 2024. [OLMo: Accelerating the science of language models](#). In *ACL*.
- Shibo Hao, Bowen Tan, Kaiwen Tang, Bin Ni, Xiyan Shao, Hengzhe Zhang, Eric P. Xing, and Zhiting Hu. 2023. [BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models](#). In *Findings of ACL*.
- Daniel A. Herrmann and Benjamin A. Levinstein. 2025. [Standards for belief representations in llms](#). *Minds and Machines*.
- Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, Moritz Müller, and Simon Razniewski. 2026a. [GPTKB v1.5: A massive knowledge base for exploring factual LLM knowledge](#). In *AAAI*.
- Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2025. [Enabling LLM knowledge analysis via extensive materialization](#). In *ACL*, pages 16189–16202.
- Yujia Hu, Tuan-Phong Nguyen, and Simon Razniewski. 2026b. [GPTKB 2.0: Browsing, querying, and auditing a disambiguated LLM-derived knowledge base](#). In *EMNLP*.
- Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *TACL*, 8.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. [Large language models struggle to learn long-tail knowledge](#). In *ICML*.
- Maciej Kurant, Athina Markopoulou, and Patrick Thiran. 2011. [Towards unbiased BFS sampling](#). *IEEE Journal on Selected Areas in Communications*, 29(9).
- Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. [Non-parametric class completeness estimators for collaborative knowledge graphs—the case of Wikidata](#). In *ISWC*.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. [Dissociating language and thought in large language models](#). *Trends in Cognitive Sciences*.
- Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. [Everything, everywhere, all at once: Is mechanistic interpretability identifiable?](#) In *ICLR*.
- Alessandro Mendoza and Anders Søgaard. 2026. [LLM beliefs are in their heads](#). In *ACL*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). *neurIPS*.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *FAT*.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. [LLMs know more than they show: On the intrinsic representation of LLM hallucinations](#).

- Marinela Parović, Ze Li, and Jinhua Du. 2025. [Generating domain-specific knowledge graphs from large language models](#). In *Findings of ACL*.
- Dinesh Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Zijun Yao. 2024. [A practical review of mechanistic interpretability for transformer-based language models](#). ArXiv.
- Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. [AI and the everything in the whole wide world benchmark](#). In *neurIPS*.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *ACL*, pages 4902–4912.
- Muhammed Saeed and Simon Razniewski. 2026. [LLM-pedia: A transparent framework to materialize an LLM’s encyclopedic knowledge at scale](#). *EMNLP*.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. [Open problems in mechanistic interpretability](#). *Transactions on Machine Learning Research*.
- Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. 2026. [Bridging the black box: A survey on mechanistic interpretability in ai](#). *ACM Comput. Surv.*, 58(8).
- Amos Tversky and Daniel Kahneman. 1973. [Availability: A heuristic for judging frequency and probability](#). *Cognitive Psychology*, 5(2).
- Gerhard Weikum, Xin Luna Dong, Simon Razniewski, and Fabian Suchanek. 2021. [Machine knowledge: Creation and curation of comprehensive knowledge bases](#). *Foundations and Trends in Databases*, 10.
- Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. [Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples](#). In *CIKM*.