OPEN ACCESS

Artificial Intelligence COMMENTARY

The Innovation Informatics

Understanding as compression: A new evaluation framework for large language models

Meng Wan,[1,2] Jue Wang,[1,*] Yangang Wang,[1] Rongqiang Cao,[1] Zijian Wang,[1] Zongguo Wang,[1] Peng Shi,[2] and Zhenbin Zhao[3]

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China

[2]University of Science and Technology Beijing, Beijing 100083, China

[3]North China Electric Power University, Beijing 102206, China

*Correspondence: wangjue@sccas.cn

Received: September 30, 2025; Accepted: November 1, 2025; Published Online: November 6, 2025; https://doi.org/10.59717/j.xinn-inform.2025.100003

Citation: Wan M., Wang J., Wang Y., et al. (2025). Understanding as compression: A new evaluation framework for large language models. *The Innovation Informatics* **1**:100003.

The rapid expansion of large language models (LLMs) raises a central question: how should understanding be defined and measured? Existing benchmarks, whether task-specific scores or probabilistic metrics, provide fragmented signals and fail to capture the depth of internal representation. Information theory offers a more principled path. Shannon's source coding theorem equates prediction with compression, suggesting that understanding can be formalized as compression efficiency. Building on this principle, a compression-driven evaluation framework is proposed. The framework advances through three layers: symbolic compression measures fidelity to training distributions, generalization compression evaluates robustness across domains and modalities, and semantic compression tests abstraction and reconstruction of meaning. Together these layers trace a continuum from memory to reasoning to cognition. Adopting compression as a core metric can unify evaluation across tasks and modalities, provide reproducible and comparable standards, and move the field beyond task-specific benchmarks. Research is urged to embrace compression not only as a technical tool but also as a scientific language for measuring understanding. By aligning evaluation with this principle, progress in artificial intelligence can be guided toward models that demonstrate genuine cognitive depth rather than superficial task performance.

INTRODUCTION: THE CHALLENGE OF DEFINING LLM UNDERSTANDING

The rapid development of Large Language Models (LLMs) marks a new era in artificial intelligence. With increasing scale, LLMs exhibit remarkable generative capabilities and near human-level performance across reasoning, multimodal comprehension, and scientific computation. However, a fundamental question remains: How to define and quantify model 'understanding'? Current evaluation methods include task-driven metrics like question answering accuracy for specific benchmarks and theoretical metrics such as perplexity (PPL) for assessing predictive accuracy. While current metrics capture output quality, they fail to show whether model representations encode deeper regularities. In addition, no reproducible and unified standard exists for comparison across domains and modalities. A data compression perspective may provide a unified and quantitative complement to task-specific metrics.

FROM PREDICTION TO COMPRESSION: A UNIFYING THEORETICAL PARADIGM

Information theory provides a powerful framework to solve the problem. Shannon's source coding theorem states that a perfect probabilistic model is both an optimal predictor and an optimal compressor. A model that predicts the next symbol accurately can also encode the data with the fewest possible bits. Consequently, understanding can be formalized as compression efficiency: Deeper understanding enables better compression. The Hutter Prize recognizes superior compression as a direct path toward Artificial General Intelligence (AGI).[1] In a similar spirit, OpenAI has argued that unsupervised learning can be framed as an approximate process of optimal data compression.[2]

This theoretical perspective is increasingly supported by empirical findings. Research from Google DeepMind demonstrated that combining the conditional probability distribution of large language models with arithmetic coding achieves compression performance that not only exceeds traditional general-purpose algorithms but also surpasses specialized compressors for image and audio data.[3] Related work explored scalable lossless architectures for heterogeneous inputs. The SEP framework (Semantics Enhancement and Multi-Stream Pipelines) extends compression to multiple domains through symbol-level pipelines and multi-stream processing, contributing mainly to symbolic and cross-domain evaluation rather than semantic abstraction.[4] At the same time, insights from cognitive science caution that compression and semantics are not perfectly isomorphic. A joint study by researchers from Stanford University and Turing Award laureate Yann LeCun revealed that current LLMs tend to prioritize maximal statistical compression, whereas the human mind often favors contextual richness and conceptual detail, even at the cost of some compression efficiency.[5] This tradeoff between statistical efficiency and semantic fidelity suggests that evaluation must consider not only if a model can compress, but also what it is compressing.

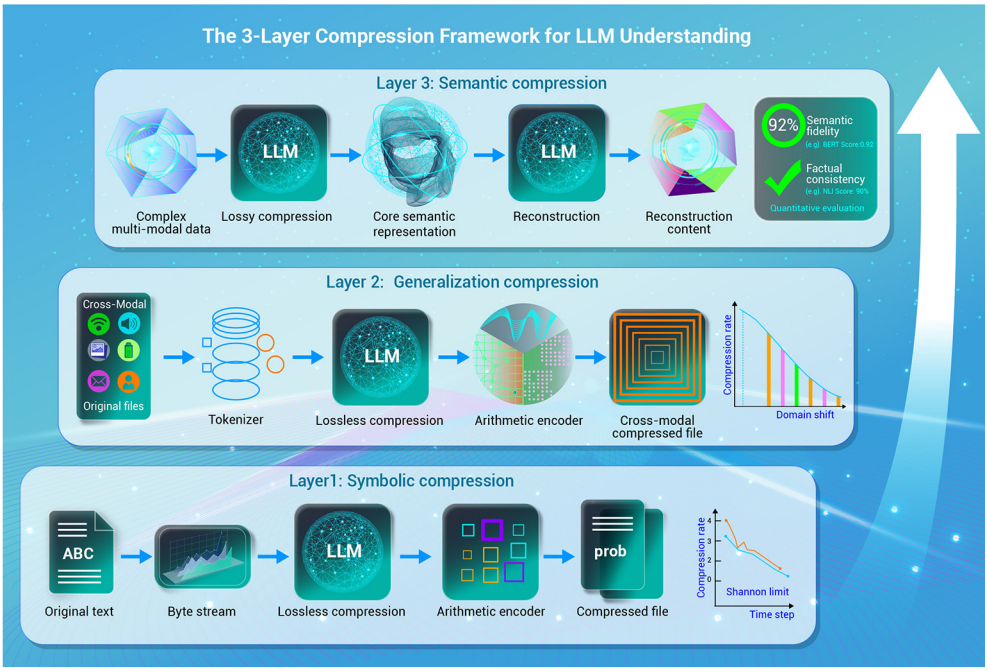

Figure 1. The 3-layer compression framework for LLM understanding.

THREE LAYERS OF COMPRESSION FOR EVALUATING UNDERSTANDING

If 'understanding' can be formalized as a compression process, the evaluation of LLMs should go beyond simply demonstrating that a model can compress data. A meaningful framework must address three questions: How the model performs compression, how well the compression holds across different domains and modalities, and whether the compressed representation reflects a true abstraction of knowledge and understanding.

To answer these questions, we propose a three-layer compression-driven evaluation framework that provides a progressive scale for measuring model understanding. The framework can be practically realized through probabilistic modeling, cross-domain compression tests, and representation-based reconstruction, moving step-by-step from memorizing statistical patterns to cross-domain reasoning and ultimately to high-level semantic abstraction (Figure 1).

Layer 1: Symbolic compression

Symbolic compression forms the foundational layer of the framework. The task is to encode symbol sequences from a known distribution as efficiently as possible. Concretely, given raw text or other symbolic data drawn from the training distribution, the model is tasked with probabilistic modeling and arithmetic coding to generate compressed files. Under ideal conditions, the average code length should approach the Shannon limit, which reflects the maximum-likelihood fit of the model to the training distribution.

This process resembles human verbatim recall, where content is reproduced without loss of information. It directly tests whether the model has fully captured the statistical structure of the training data to achieve optimal prediction at the symbolic level. Importantly, the evaluation does not depend on task-specific definitions, manual labels, or external evaluation criteria. As a result, symbolic compression provides a unified, reproducible, and directly comparable benchmark across models. By passing this layer, a model demonstrates the basic ability to construct complete probabilistic representations of symbol data, which serves as the foundation for higher levels of generalization and semantic abstraction.

Layer 2: Generalization compression

While symbolic compression verifies modeling performance within familiar data distributions, true understanding requires a model to maintain compression efficiency in unfamiliar domains and across new modalities. Generalization compression addresses this challenge by introducing out-of-distribution and cross-modal inputs to evaluate transferability and robustness.

As shown in Figure 1 Layer 2, the input includes not only text from the training domain but also previously unseen text, audio, images, or video signals. All inputs are first converted into a unified symbolic representation and then compressed using the model's probabilistic predictions combined with arithmetic coding. As the input distribution gradually diverges from the training distribution, the evolution of compression rates is recorded. An ideal model exhibits smooth and gradual degradation rather than abrupt collapse. By analyzing the decay curves of compression efficiency under distributional shifts, it becomes possible to distinguish between memory-driven models, which fail rapidly on unfamiliar data, and models with genuine generalization ability, which sustain efficient compression across a broader range. Therefore, generalization compression provides a unified metric for evaluating cross-domain and cross-modal robustness. Unlike fragmented task-specific indicators, this method enables direct comparison under a single measure 'compression rate', while ensuring reproducibility.

Layer 3: Semantic compression

Semantic compression represents the highest level of model understanding, focusing on capturing the core meaning of information during compression, rather than the surface form of symbols. At this level, compression requires the formation of highly abstract internal representations that distill the core concepts and logical structures of the input while discarding redundant or noisy details.

As shown in Figure 1 Layer 3, the process unfolds in two stages. First, the model transforms complex input data into a compact core representation. Then the model must reconstruct the original content based solely on this representation. Unlike symbolic compression, reconstruction at this stage does not aim for character-by-character fidelity. Instead, the criterion is semantic fidelity: The reconstructed output must preserve the essential facts, relations, and logical coherence of the original information. The process is analogous to how humans learn and communicate knowledge. Rather than memorizing every word, people abstract key ideas, build conceptual structures, and re-express them in different forms without losing key meaning. Similarly, semantic compression tests whether models can internalize abstract representations that support reasoning, generalization, and knowledge transfer.

Evaluation at this layer must go beyond compression rates to assess semantic consistency and factual reliability. A model demonstrates genuine understanding only when it can perform lossy compression that preserves core meaning and leverages internal representations for reasoning and generation. Compared to the focus of symbolic compression on memorization and the focus of generalization compression on transfer, semantic compression measures the leap from memory to cognition, marking a decisive step toward artificial general intelligence.

CONCLUSION

Current evaluation of LLMs relies heavily on task-based metrics and probabilistic modeling measures. While these approaches quantify performance on specific tasks, they fail to reveal whether a model truly understands the underlying information. As models grow larger and are more widely deployed, evaluation must move beyond fragmented, task-specific indicators toward a unified, computationally grounded framework. Compression offers a natural pathway for this shift. Through the three progressive layers of symbolic compression, generalization compression, and semantic compression, researchers can systematically characterize model behavior across the spectrum from memorization to reasoning to abstraction. This framework is not intended to replace existing evaluations but to provide a common language that enables direct, reproducible comparisons across tasks, modalities, and systems. Compression introduces a complementary, information-theoretic dimension to mainstream benchmarks such as MMLU, BIG-Bench, and ARC, which focus on reasoning accuracy and factual consistency. Integrating these task-level evaluations with compression-based metrics yields a more coherent and multidimensional view of model cognition. Advances in probabilistic modeling and multimodal integration are also making such compression-driven evaluation increasingly practical.

Looking ahead, breakthroughs in LLM development will depend not only on greater computational resources or larger datasets, but also on the ability to achieve true semantic compression that reflects a deep understanding of knowledge. We call on the research community to adopt compression as a core metric of understanding and to build evaluation systems that are cross-modal, reproducible, and longitudinally trackable. The technical pathway forward requires a concerted community effort to establish unified, large-scale datasets for evaluating out-of-distribution robustness (Layer 2) and to define consensus benchmarks for assessing semantic fidelity (Layer 3). By following this direction, AI research can move from empirically driven progress to a theory-driven foundation, transforming 'understanding' from an abstract notion into a quantifiable and verifiable scientific goal.

REFERENCES

1. Wolff J.G. (2019). Information compression as a unifying principle in human learning, perception, and cognition. *Complexity* **2019**:1879746. DOI:10.1155/2019/1879746
2. Sutskever I. (2023). The future of AI & OpenAI. AI Grant Speakeasy. Available at: https://www.youtube.com/watch?v=SjhIlw3Iffs
3. Delétang G., Ruoss A., Duquenne P.A., et al. (2023). Language modeling is compression. *ICLR* **2024**. DOI:10.48550/arXiv.2309.10668
4. Wan M., Cao R., Li Y., et al. (2025). SEP: A general lossless compression framework with semantics enhancement and multi-stream pipelines. *IJCAI* **2025**:3326–3334. DOI:10.24963/ijcai.2025/370
5. Shani C., Jurafsky D., LeCun Y., et al. (2025). From tokens to thoughts: How LLMs and humans trade compression for meaning. *arXiv* preprint. DOI: 10.48550/arXiv.2505.17117

FUNDING AND ACKNOWLEDGMENTS

This work was supported by National Science and Technology Major Project (grant no. 2023ZD0120503). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

M.W.: Conceptualization, Methodology, Writing, Visualization. J.W.: Supervision, Writing, Review & Editing. Y.W., R.C., Z.W. (Zijian), Z.W. (Zongguo): Investigation, Validation. P.S., Z.Z.: Resources, Writing, Review & Editing. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Jue Wang is an Academic Editor of The Innovation Informatics and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.