Illusory Generalization in NLP: Why Scaling Laws Mask Systematic Failures in Out-of-Distribution Reasoning

Anonymous ACL submission

Abstract

Despite the rapid advances in natural language processing (NLP) driven by large-scale neural models, the claim that these models achieve true generalization remains questionable. This paper argues that the current paradigm of scaling laws in which larger models appear to perform better does not equate to genuine conceptual generalization. Instead, models optimize for memorization of latent statistical patterns, leading to overestimated generalization capabilities. We critique existing evaluation methodologies, highlight failures in systematic out-of-distribution (OOD) reasoning, and propose an approach inspired by cognitive science to benchmarking generalization.

1 Introduction

NLP research is at an inflection point. The impressive performance of large-scale models has led to a prevailing narrative that increasing model size and training data leads to more human-like linguistic capabilities. But this assumption that rests upon scaling laws as a proxy for intelligence is fundamentally flawed. It confuses statistical pattern-matching with true generalization, conflates interpolation with conceptual abstraction, and ultimately propagates an illusion of understanding rather than a meaningful grasp of language.

Human cognition does not operate on the sheer brute-force recognition of patterns. Instead, it is marked by the ability to generalize across domains, integrate concepts in novel ways, and infer meaning from minimal data. This capacity for conceptual flexibility allows humans to extrapolate beyond their experiences, making sense of entirely new constructs without requiring millions of training examples. Large language models (LLMs), despite their scale, lack this core cognitive ability. Their so-called "generalization" is often little more than the ability to retrieve and recombine learned statistical regularities which we see crumbles when faced with truly novel linguistic structures or shifts in meaning.

If generalization is to be more than an illusory benchmark, we must radically rethink how we define and evaluate it. This paper critiques the illusion of generalization in NLP and argues that a paradigm shift is necessary: one that prioritizes systematicity, compositionality, and conceptual integration over raw scaling power. Without such a shift, the field risks mistaking the shadows of intelligence for intelligence itself.

2 The Limits of Current Generalization Metrics

Most generalization claims in NLP stem from performance on standard benchmarks, yet these benchmarks often fail to capture meaningful OOD reasoning. Common datasets, such as GLUE (Wang, 2018), SuperGLUE (Wang et al., 2019), GLUE-X (Yang et al., 2022), and multilingual benchmarks (Kakwani et al., 2020), rely on splits that maintain high degrees of overlap between training and test distributions. Even when adversarial datasets, like AdvGlue (Wang et al., 2021), attempt to probe robustness, models still succeed through pattern exploitation rather than conceptual understanding. Consider the following key failures of existing generalization metrics:

1. IID Assumptions: Most benchmarks split data randomly (Arp et al., 2022), leading to train-test leakage of latent patterns that models can exploit.

2. Spurious Correlations: Large models pick up on statistical artifacts rather than underlying linguistic principles (Tu et al., 2020), performing well on benchmarks while failing in real-world, structurally different cases.

3. Lack of True Compositionality Tests: Human cognition allows for productive compositional

generalization, the ability to combine known concepts in novel ways, but models struggle with systematicity and recombination.

3 Why Scaling Does Not Solve the Problem

The success of large-scale models is often attributed to their ability to absorb vast amounts of data and identify complex correlations. However, this brute-force approach does not address deeper challenges of conceptual generalization. While scaling improves interpolation within the training distribution, it does not guarantee systematic extrapolation to unseen structures. Scaling laws, which suggest that performance improves predictably as models grow, have become a dominant narrative in NLP. Yet, this narrative conceals fundamental failures in systematic reasoning, abstraction, and robust linguistic adaptability.

Fundamentally, the scaling hypothesis from statistical mechanics does not, and cannot, apply to language in the ways we aim for it. The scaling hypothesis in statistical mechanics describes how physical quantities, such as correlation length and specific heat, exhibit power-law behavior near critical points. (Baxter, 2016) This framework assumes that key macroscopic properties emerge from underlying microscopic interactions in a way that is self-similar across scales. Such behavior is well-suited to physical systems because these systems obey fundamental conservation laws, operate under local interactions, and exhibit universal behavior near phase transitions. However, applying this same reasoning to language, which is an emergent faculty of human cognition, introduces fundamental problems.

Unlike physical systems, language is not governed by simple statistical interactions but by complex, hierarchical, and context-dependent structures. In statistical mechanics, increasing system size (e.g., more particles in a lattice model) preserves the same governing equations, leading to predictable scaling behaviors. In contrast, language acquisition and processing are shaped by conceptual abstraction, communicative intent, and sociocultural constraints, none of which follow simple power-law relationships. While scaling large language models improves performance on certain benchmarks, it does not imply that language understanding itself is a function of model size in the same way that physical properties depend on system size.

Moreover, in statistical mechanics, self-similarity across scales arises because the same fundamental interactions govern both small and large-scale behavior. In cognition, however, different scales involve qualitatively different processes. Phonemes do not compose into words in the same way that words compose into sentences, and sentence-level meaning does not simply emerge from statistical aggregation of words. Human language is characterized by symbolic structure, long-range dependencies, and non-local meaning constraints, none of which are naturally captured by statistical scaling laws. This is why merely increasing model size does not necessarily resolve core challenges in NLP, such as reasoning, abstraction, or grounding in real-world perception.

Thus, while the scaling hypothesis provides a powerful framework for understanding physical phase transitions, its direct application to linguistic cognition is problematic. The underlying mechanisms of statistical physics (local interactions, energy minimization, and universality classes) are fundamentally different from the principles that govern language (symbolic representation, hierarchical compositionality, and meaning constraints). Future advancements in NLP will require insights from cognitive science, linguistics, and neurosymbolic models, rather than relying solely on scaling laws derived from statistical physics.

However, as the size of data grows, neurosymbolic systems encounter bottlenecks in processing speed and memory usage, limiting their effectiveness in real-time applications or large-scale models. Moreover, symbolic reasoning can become unwieldy when faced with ambiguous, contradictory, or incomplete data, conditions that are commonplace in natural language. Despite recent advancements, the scalability of neurosymbolic models remains a critical challenge (Hamilton et al., 2024), and the models' inability to efficiently handle the vast amounts of unstructured data characteristic of modern NLP tasks has hindered their broader use. Until these issues are addressed, neurosymbolic models are likely to remain a niche solution rather than a mainstream approach in the NLP field.

3.1 The Limits of Memorization

LLMs do not generalize in the way that humans do, rather, they approximate statistical distributions over language. While human cognition, particularly language acquisition, is influenced by statistical regularities in the input we receive, humans

combine these patterns with context, experience, and higher-order cognitive processes like theory of mind, social cues, and world knowledge. This allows humans to generalize beyond mere statistical patterns, make inferences, and creatively use language in ways that statistical distributions alone cannot account for. In other words, while we are sensitive to statistical distributions, our cognitive processing involves interpreting these patterns within a larger framework of meaning, intention, and context. NLP models' reliance on data-driven pattern extraction allows them to mimic generalization in-distribution, but it does not translate to true conceptual flexibility. Studies show that when models encounter test distributions that significantly deviate from training data, performance collapses. (Bommasani et al., 2021) This suggests that scaling primarily enhances memorization and interpolation, rather than fostering genuine abstraction.

3.2 Insufficient Models of Human Memory

Long short-term memory (LSTM) networks and other neural networks used in NLP (Sherstinsky, 2020) are modeled after human memory and learning, attributing the flow of information to artificial neurons, memory cells, gates, and connections. These models are designed to capture sequential dependencies and learn patterns over time, mimicking certain aspects of human memory. However, this biologically-inspired approach oversimplifies the complexities of actual human memory. Human memory is not solely governed by neuronal firing, but is also influenced by a wide range of other factors, including the biological, environmental, and psychological context in which memory is formed and recalled.

One important factor that current NLP models overlook is synaptic plasticity (Abraham and Bear, 1996), a fundamental process in the brain that underpins learning and memory. Synaptic plasticity refers to the ability of synapses to strengthen or weaken over time, in response to increases or decreases in their activity. This process plays a critical role in memory formation and recall in humans, allowing the brain to adapt and reconfigure its connections based on experiences. Only a very limited body of research addresses the similarities between learning dynamics employed in deep artificial neural networks and synaptic plasticity in spiking neural networks (Kaiser et al., 2020), and an even smaller looks to connect this to the network architectures employed in NLP. In contrast, most NLP models, including LSTMs, focus solely on the flow of information through artificial neurons, neglecting the broader, dynamic processes that contribute to memory in biological systems. Without considering these additional factors, neural networks in NLP are limited in their ability to fully replicate the richness and flexibility of human memory.

3.3 Failures in Compositional and Systematic Generalization

Humans can seamlessly compose known concepts in novel ways, a property essential for linguistic creativity and reasoning. Yet, even the largest LLMs struggle with:

- Compositional Generalization: Models trained on simple phrase structures often fail to generalize those structures to novel contexts, as seen in tasks like SCAN (Lake and Baroni, 2018) and COGS (Kim and Linzen, 2020).

- Systematicity: The ability to apply learned rules consistently across linguistic contexts remains weak, leading to unpredictable performance in unseen but structurally related tasks.

3.4 Lack of Robust Causal Reasoning

Human cognition relies heavily on causal inference rather than surface-level correlations. NLP models, on the other hand, remain trapped in statistical association. Even when fine-tuned on causal reasoning datasets, LLMs often default to heuristics based on co-occurrence rather than grasping the underlying causal mechanisms. (Wang et al., 2023) This limitation renders them unreliable for applications requiring counterfactual reasoning and logical inference.

3.5 Semantic Drift and Fragility

Another overlooked consequence of scaling is semantic drift, where models exhibit inconsistencies in meaning representation over time. Unlike humans, who can dynamically refine and stabilize meanings based on experience and context, LLMs often demonstrate erratic shifts in word interpretations when probed under different conditions. This fragility exposes the shallow nature of their linguistic generalization.

3.6 The Illusion of Scaling as a Silver Bullet

The belief that simply making models larger and feeding them more data will eventually yield human-like generalization is deeply flawed. While scaling improves performance on benchmark datasets, it does not address the fundamental deficiencies outlined above. If true intelligence required only more data and more parameters, then insects, whose neural systems are vastly smaller than deep learning models, would be incapable of intelligent behavior. Yet, even with limited neural resources, biological organisms exhibit remarkable adaptability and conceptual understanding far beyond what LLMs can achieve.

In sum, scaling laws create an illusion of progress while failing to address core limitations in abstraction, compositionality, and causal inference. To move beyond the current paradigm, NLP research must embrace alternative evaluation frameworks that prioritize conceptual flexibility, systematic reasoning, and human-like generalization mechanisms over mere statistical pattern-matching.

4 Toward a Better Evaluation Framework

If scaling alone does not ensure true generalization, how should we redefine the problem? We propose a cognitive-science-inspired approach to evaluating NLP models, incorporating:

1. *Conceptual Integration Tasks*: Instead of relying on predefined taxonomies, models should be tested on their ability to form novel conceptual categories. In human cognition, conceptual integration is a general cognitive operation on a par with analogy, recursion, mental modeling, conceptual categorization, and framing (Fauconnier and Turner, 1998) that allows individuals to merge disparate concepts into a coherent framework and NLP models should be evaluated on their ability to do the same.

2. *Compositional and Systematic Reasoning*: Humans can generalize by systematically recombining known components in novel ways, a property essential for linguistic creativity. This uniquely human cognitive operation allows for flexible adaptation to novel situations by applying learned rules and concepts in a structured, predictable manner. It is not just about rote application of patterns, but rather about dynamically building new conceptual relationships based on context and prior knowledge. This form of reasoning underpins advanced cognitive functions such as language, mathematics, and higher-order planning. Models should be assessed through tasks that require genuine productivity, such as forming and understanding entirely new idioms or metaphors rather than memorizing frequent n-grams.

3. *Causal and Counterfactual Testing*: Unlike mere correlation-based inference, human cognition relies on understanding causal relationships. Future NLP benchmarks should include counterfactual reasoning challenges where models must infer the implications of an event based on causal structures, rather than relying on statistical co-occurrence.

4. *Cross-Population Generalizability*: A true measure of generalization is how well models adapt to linguistic and cognitive diversity. Benchmarks should include language variation across demographics, dialects, and cultural contexts to test whether models exhibit human-like adaptability rather than overfitting to dominant linguistic norms.

5. *Grounding in Embodied Experience*: Unlike statistical models, human cognition is deeply grounded in perceptual and sensorimotor experiences. While NLP models lack direct physical grounding, we should develop benchmarks that require grounding via multimodal learning, integrating textual, visual, and auditory inputs in a way that mimics human sensory-driven understanding.

   Grounded cognition theories emphasize understanding concepts through sensory, motor, and emotional interactions with the physical world (Barsalou, 2008), whereas LLMs rely solely on patterns in linguistic (and, more recently, multimodal) data. While LLMs can simulate grounded cognition by generating descriptions of sensory or embodied experiences (Zhong et al., 2024), these outputs are still based on statistical correlations within text rather than direct perceptual grounding. Multimodal models incorporating visual or auditory data provide a step toward bridging this gap, but the grounding remains indirect. Similarly, embodied AI systems that integrate

LLMs with sensors and actuators offer potential for grounded understanding, but that grounding still would reside in the broader system, not the LLM itself.

5 Addressing the Illusion of Understanding: Our Responsibility as NLP Researchers

The illusion of understanding perpetuated by NLP models, neurosymbolic or otherwise, poses a significant danger to the general public. As these technologies increasingly permeate everyday life, the public often equates AI's ability to generate coherent responses with true comprehension or reasoning. This misperception leads to dangerous overconfidence, where individuals and organizations place undue trust in AI systems for tasks that demand deep understanding, ethical decision-making, or complex judgment. In high-stakes fields like healthcare, law, and education, this misguided trust can result in profound consequences, including the amplification of biases, the propagation of misinformation, and the undermining of critical human decision-making. The gap between the perceived and actual capabilities of these models is widening, and it is crucial for researchers in NLP to take immediate and decisive action to address this issue.

The onus is squarely on the NLP community to clarify, educate, and communicate the limitations of these systems to the public. Researchers must be transparent about what current AI technologies can and, more importantly, cannot do, highlighting the distinction between machine pattern recognition and genuine human understanding. Failing to take responsibility for this clarification not only risks the misuse and misapplication of these technologies but also perpetuates a dangerous cycle of misinformation and misplaced trust that can harm society on a large scale. It is imperative that the field confronts this issue head-on, ensuring that the public is fully aware of the limits of current technologies before these systems are entrusted with decisions that affect lives.

Importantly, it is imperative that the aforementioned distinction is maintained and emphasized, especially when we still refer to the field as NLP, *natural* language processing, whereas other parameterized aspects of cognition are clearly delineated by their synthetic and data-driven nature, such as *computer* vision or *machine* learning. We do not conflate computer vision as being the same as biological vision, or machine learning to be the same process as human learning, even though the capabilities of computational perceptualities like machine vision are more human-like in performance, complexity, and theoretical underpinnings. The language we use as researchers contributes to the public's understanding of concepts, and the words we use to label things can influence how they are perceived. Blurring critical distinctions with imprecise language reinforces misconceptions about human intelligence, machine behavior, and text analytics as three independent disciplines.

6 Conclusion

The NLP community stands at a crossroads. The rapid expansion of model size and training data has fueled a narrative of progress, but this progress is largely an illusion when it comes to genuine generalization. Scaling alone does not produce models that reason, integrate concepts, or systematically extend knowledge beyond their training data. Instead, it perpetuates the brittle successes of statistical pattern-matching, masking fundamental weaknesses that become apparent in out-of-distribution settings.

If NLP is to move beyond the superficial trappings of intelligence, the field must undergo a paradigmatic shift. We must abandon the notion that performance on existing benchmarks equates to human-like cognition and instead develop evaluation frameworks that capture the hallmarks of true conceptual understanding. This means embracing benchmarks that prioritize systematicity, compositionality, conceptual integration, and causal reasoning, rather than those that reward mere statistical approximation.

The illusion of generalization is dangerous not only because it misrepresents what these models can do, but because it shapes the direction of research and the application of NLP technologies in real-world settings. The risk is not just academic, but has profound ethical and societal implications. Models that fail to generalize robustly propagate biases, misinterpret linguistic nuances, and reinforce systemic errors at scale. If we do not rethink our approach now, we risk entrenching systems that fail when it matters most.

The challenge before us is not one of mere scale, but of fundamental reevaluation. Generalization must mean more than predictive accuracy and must reflect the flexible, adaptive, and compositional

nature of human cognition. Without this shift, NLP will continue to chase the mirage of intelligence, never reaching the oasis of true understanding.

Limitations

While this paper critiques the illusion of generalization in NLP and proposes a cognitive-science-inspired evaluation framework, several limitations must be acknowledged. First, the proposed benchmarks, while theoretically motivated, require significant interdisciplinary collaboration to implement effectively. Cognitive science and NLP researchers must work together to design tests that are both empirically rigorous and practically feasible. Second, current large-scale models are optimized for efficiency, and introducing cognitively inspired benchmarks may require more computationally intensive testing procedures, posing scalability concerns. Third, there remains an open question of whether models, even with improved evaluation frameworks, can ever achieve human-like abstraction without fundamentally different architectures. Finally, this paper focuses on linguistic generalization but does not address broader AI capabilities, such as embodied cognition and interaction, which may be necessary for deeper intelligence.

References

Wickliffe C Abraham and Mark F Bear. 1996. Metaplasticity: the plasticity of synaptic plasticity. *Trends in neurosciences*, 19(4):126–130.

Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3971–3988.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Rodney J Baxter. 2016. *Exactly solved models in statistical mechanics*. Elsevier.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Gilles Fauconnier and Mark Turner. 1998. Conceptual integration networks. *Cognitive science*, 22(2):133–187.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2024. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(4):1265–1306.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. 2020. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 9087–9105.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.

Shu Zhong, Elia Gatti, Youngjun Cho, and Marianna Obrist. 2024. Exploring human-ai perception alignment in sensory experiences: Do llms understand textile hand? *arXiv preprint arXiv:2406.06587*.