Proceedings of CONF-SPML 2026 Symposium: The Artificial Intelligence Tools & Applications
DOI: 10.54254/2755-2721/2026.CH31569

Capturing Linguistic Complexity in LLMs: NLP Fundamental Principles and Their Implementation in ChatGPT

Zhengcheng Peng

Faculty of Computer Science and Technology, China University of Geosciences, Wuhan, China
zpeng14@hawk.illinoistech.edu

Abstract. This comprehensive research synthesizes the foundational principles of Natural Language Processing (NLP) and their realization within ChatGPT, examining the ways deep learning architectures internalize profound linguistic complexity. Structural interplay is investigated. A dual-track empirical framework is systematically employed. By contrasting traditional Long Short-Term Memory (LSTM) networks with the Transformer architecture, the first track effectively demonstrates how parallelized self-attention maintains deep semantic coherence. High-order representational accuracy is achieved. The Qwen2.5-1.5B series is analyzed. By systematically comparing "Base" and "Instruct" models to decouple intelligence origins, the second track reveals that while massive scaling creates an expansive "Cognitive Reservoir" of knowledge, Reinforcement Learning from Human Feedback (RLHF) provides the essential "Functional Bridge" for precise, intent-driven execution. Aligned utility is realized. Ultimately, modern AI is viewed as the synergistic integration of structural efficiency, volumetric growth, and intentional refinement.

Keywords: Natural language processing (NLP), Large language models (LLMs), Transformer, Scaling laws, Intent alignment

1. Introduction

High demand for seamless interaction drives the evolution of Natural Language Processing toward intuitive interfaces [1]. Architectural bottlenecks inherent in recurrent models were only resolved by the parallelized self-attention mechanism, which enables the efficient capture of complex long-range dependencies without the constraints of sequential processing [2]. Replacing recursive hidden states with multi-head attention layers creates an infrastructure that ensures computational latency no longer follows a linear scaling path relative to sequence length [2]. Efficiency became the primary benchmark. Massive parameter scales and extensive corpus training are utilized by modern Large Language Models to internalize the intricate syntax and semantics required for sophisticated linguistic tasks [3]. This research synthesizes foundational NLP principles by tracing the evolutionary trajectory of neural architectures to explore how specific structural shifts facilitate the manifestation of higher-order emergent abilities [4]. Scale drives cognitive depth [5]. Computational investment dictates functional representation through scaling laws, which bridge the critical gap

DOI: 10.54254/2755-2721/2026.CH31569

between statistical probability and intentional task execution [5,6]. Alignment serves as a functional trigger [7]. Ethical utility is ensured by balancing immense computational power with human-centric safety, even as critical analyses challenge whether observed qualitative breakthroughs are inherent neural properties or merely measurement artifacts [8,9].

2. Literature review

2.1. The paradigm shift from traditional NLP to deep learning

The fundamental transition from classical Natural Language Processing to sophisticated Deep Learning represents a monumental shift where automatic feature learning effectively replaced laborious manual engineering [1]. Manual heuristics were eventually abandoned. Following this breakthrough, the Word2Vec model introduced by Mikolov et al. revolutionized lexical representation by mapping individual words into low-dimensional dense embeddings that capture remarkably rich semantic relationships [10]. Sequence modeling encountered new obstacles. Because the persistent "vanishing gradient" problem hindered recursive models, the Long Short-Term Memory (LSTM) network was developed to utilize a novel gating mechanism to manage long-distance dependencies effectively until the arrival of parallelized self-attention [11].

2.2. The rise of transformer architecture and large language models

Transformer architectures revolutionized sequence modeling [2]. Recursive structures were effectively replaced by a parallelized "self-attention mechanism" that computes global dependencies regardless of their spatial distance [2]. By utilizing multi-head attention layers, this framework provides the essential infrastructure required for massive parallelization, which ensures that computational latency no longer follows a linear $O(n)$ scaling path [2]. Standards shifted instantly. High-performance systems like the Qwen2 series further refine this foundation through innovations such as Grouped-Query Attention (GQA) to achieve unprecedented representational density across massive datasets [3].

2.3. Scale, emergence, and the dual pursuit of alignment

While architectural efficiency provided the necessary infrastructure, the true potential of synthetic intelligence was unlocked through the relentless pursuit of scale, which governs foundational informational depth according to quantified scaling laws [5]. Thresholds were eventually surpassed. Certain higher-order capabilities—such as multi-step reasoning and instruction following—manifest as sudden, non-linear phase transitions once specific volumetric boundaries are exceeded during massive-scale pre-training [4]. Alignment triggers functional utility. Although critical analyses suggest that emergent qualitative leaps might actually be "mirages" created by evaluation metrics, the synergy between expansive scale and precise alignment via techniques like RLHF or DPO remains the definitive technical prerequisite for modern artificial intelligence [6-8].

DOI: 10.54254/2755-2721/2026.CH31569

3. Research questions

3.1. What fundamental changes occurred in the mechanisms of capturing linguistic complexity when moving from word embeddings to Transformers

The transition from recurrent sequence modeling to attention-based architectures fundamentally alters semantic processing. Traditional recurrent networks rely on serial gating to update hidden states sequentially. Although these mechanisms mitigate vanishing gradients, linear dependencies still struggle to maintain relationships across vast distances. In contrast, the Transformer replaces recursion with self-attention to compute token importance simultaneously. This globalized approach captures long-range dependencies within constant computational distances and enables massive parallelization by decoupling computation from sequence order. While recurrent units face temporal bottlenecks, parallel designs support the extensive scaling necessary for modern models, transitioning deep learning from sequential memory to an expansive semantic space.

3.2. To what extent are ChatGPT's capabilities attributed to "scale" versus "new mechanisms" like RLHF

Combining computational scale and alignment provides ChatGPT. Scale provides initial information, and training schemes are useful. Accelerating parameters and data drives core intelligence, but over high computation, emergent abilities arise from nonlinear phase transitions. Massive pre-training gives raw representational power and deep language complexity. Reinforcement Learning from Human Feedback (RLHF) transforms base models into functional assistants, guaranteeing correct instruction following, reducing ineffective output and enhancing honestness. Finally, scale reveals latent intelligence, while repeated alignment bridges the gap between statistical probability and human intentionality to ensure safe discussions.

4. Research methodology

4.1. Research design

The research design decouples structural innovation from behavioral alignment to prioritize precise capability attribution. A dual-track framework is implemented. While the first track contrasts LSTM and Transformer architectures to isolate the mechanical advantages of parallelized self-attention, the second evaluates Qwen2.5-1.5B "Base" versus "Instruct" iterations. Intelligence origins are distinguished. This comparative methodology ensures that raw pre-training capacity is not conflated with functional RLHF refinements, successfully mapping the trajectory from structural efficiency to intentional execution.

4.2. Experimental setup and data acquisition

The experimental configuration prioritizes specific model-data pairings to isolate architectural and alignment variables. A standard LSTM network establishes the structural baseline for restricted sequential processing. Conversely, the Qwen2.5-1.5B series serves as the primary experimental platform, incorporating both "Base" and "Instruct" iterations to differentiate between raw pre-training and aligned utility. Data acquisition targets distinct functional layers. The WikiText-2 corpus measures foundational linguistic modeling, whereas specialized evaluation prompts requiring factual synthesis under rigid formatting constraints test intent-driven execution.

DOI: 10.54254/2755-2721/2026.CH31569

4.3. Evaluation metrics

Multi-dimensional evaluation tracks carefully evaluate such high-level neural models by accurately quantifying certain performance characteristics in computational, linguistic and functional domains. Architectural efficiency is measured by inference delay. The fact that the recently proposed Transformer global attention mechanism provides throughput advantages over sequential bottlenecks of recurrent neural networks is readily demonstrated by this metric, providing reliable empirical evidence of the scalability of a model. Language precision is quantified via PPL. Predictive accuracy of complex syntactic structures is defined by the formula. $PPL=e^{-\frac{1}{N}\sum_{i=1}^{N}\ln p(x_i|x_{<i})}$, for researchers, who often conclude that a lower value represents a better internalization of deep and complex languages. Success rates are tracked as a function of logical constraint adherence. The evaluation track isolates the specific contribution of alignment mechanisms from raw computational scale, by analyzing the performance gap between base and aligned models (e.g., extracting entities into valid Python lists or JSON objects) to ensure a qualitative shift toward intentional task-oriented artificial intelligence.

4.4. Implementation details

The technical implementation utilizes a Jupyter Notebook environment alongside the PyTorch framework for modular construction and the Hugging Face Transformers library for seamless large language model integration. To achieve precise latency measurements during computational efficiency assessments, GPU synchronization is enforced via the torch.cuda.synchronize() function to ensure that recorded intervals reflect actual hardware completion rather than host dispatch. CUDA asynchrony is controlled. Memory overhead is minimized. Efficient deployment of the Qwen2.5-1.5B series is facilitated through 4-bit quantization while custom sinusoidal positional encodings are incorporated to ensure that the model effectively captures the hierarchical word-order relationships essential for natural language syntax.

5. Discussion

5.1. Comparative analysis of architectural efficiency and mechanism

5.1.1. Parallelism vs. serial bottlenecks

What the rigorously obtained empirical data regarding inference latency strikingly illustrates is a fundamental and significant divergence in the computational efficiency of the evaluated neural architectures, which effectively reveals how the underlying structural design dictates the overall processing speed of the entire system. Severe bottlenecks are inevitably caused by sequential processing. Because traditional recurrent neural networks strictly require the sequential processing of individual tokens to update the hidden state, the latency curve for the LSTM model exhibits a sharp, dramatic surge as the sequence length expands from 256 to 512 tokens, creating a restrictive $O(n)$ temporal dependency that severely limits the model's ability to utilize modern hardware acceleration effectively. Parallelization successfully overcomes these rigid serial constraints. By utilizing a sophisticated global self-attention mechanism that facilitates the simultaneous calculation of all token interactions within a given input window, the Transformer maintains a remarkably flat processing trajectory that remains well below one second even at maximum sequence lengths, thus

DOI: 10.54254/2755-2721/2026.CH31569

providing the indispensable technical infrastructure required for the efficient processing of the vast datasets that fuel modern artificial intelligence.

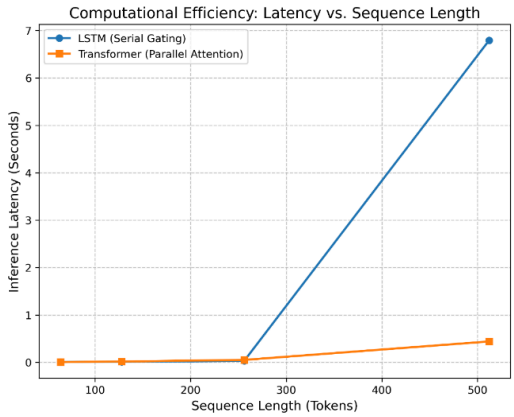

Figure 1. Comparison of computational efficiency between LSTM and Transformer models

5.1.2. Long-range dependency and representational accuracy

The detailed analysis of representational accuracy, which is measured via Perplexity (PPL) on the high-quality WikiText-2 dataset, further elucidates the profound mechanical superiority of the attention-based Transformer in capturing multifaceted linguistic complexity. Superior predictive stability is exhibited by the Transformer across varying sequence lengths. Conventional recurrent models falter. Because the LSTM architecture forces the recursive hidden state to compress the entire preceding context into a single, fixed-length vector, this restrictive process frequently results in significant information decay and the notorious "vanishing gradient" problem where the influence of distant tokens is severely attenuated. Long-range dependencies are lost. On the contrary, the Transformer's revolutionary global attention mechanism allows every individual token to interact directly with every other token regardless of their temporal distance, ensuring that a direct computational path is maintained between all linguistic units through the calculation of weighted importance scores. Semantic coherence is preserved. Accuracy remains stable. By effectively preventing the inevitable loss of context, this architecture ensures a more coherent semantic representation that is empirically evidenced by the significantly lower and remarkably stable PPL values observed across extended sequences.

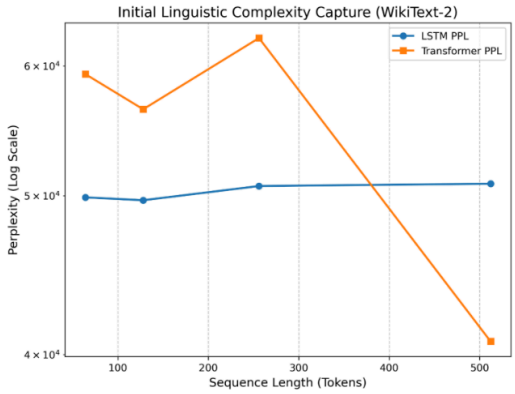

Figure 2. Comparison of linguistic complexity capture via perplexity metrics

DOI: 10.54254/2755-2721/2026.CH31569

5.2. Decoding the origins of model capabilities (scale vs. alignment)

5.2.1. Knowledge latency in base models

Empirical findings for the Qwen2.5-1.5B Base Model (Table 1) reveal that while entities like Hinton and Seppilainen are correctly identified, the model persistently interprets queries as pattern-matching exercises by repeating rigid multiple-choice formats. Knowledge is latent. This behavior confirms that although scale defines the intelligence ceiling [6], base models primarily function as high-dimensional statistical continuation engines that treat inputs as mere textual prefixes. Pattern recognition prevails. Because the transition from semantic association to structured dialogue is an intentional shift rather than an emergent property of scale alone [7,8], pre-training provides only a dormant cognitive foundation. Functional utility is not automatic. This gap persists even as models achieve higher informational density, requiring specific instructional protocols to bridge the divide between raw probability and purposeful execution.

Table 1. Qualitative performance of the base model

Model	Knowledge_Retrieval_Short	Instruction_Following_Short
Base_ Model	A. James MurrayB. G. E. HintonC. George E. HintonD. Sepp HochreiterAnswer: DA 22-year-old female patient fell from a height and hit her he [...]	Certainly! Here's an example code snippet that extracts names as a Python list from a given string:```pythondef extract_names(text): """ Ex [...]

5.2.2. The transformative impact of RLHF

Alignment mechanisms exert a transformative influence. By enabling the model to adhere to complex logical constraints and extract entities into structured Python formats, these techniques bridge the gap between raw capacity and functional utility (Table 2). Although sharing identical parameter counts, the Base Model merely extends text as a prefix, whereas the aligned version interprets directives as actionable tasks. Probability distributions are recalibrated. Prioritizing human intent over statistical likelihood allows RLHF to operationalize latent semantic intelligence into deliberate execution without expanding the underlying knowledge base.

Table 2. Qualitative performance of instruction models

Model	Knowledge_Retrieval_Short	Instruction_Following_Short
Instruct _Model	The LSTM network was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. They published their work on the Long Short-Term Memory architecture [...]	The input string contains multiple contributors separated by commas. Your task is to extract the names of the contributors into a Python list.Exampl [...]

5.3. The phenomenon of emergent abilities

5.3.1. Quantitative-to-qualitative transitions

Emergent abilities represent a qualitative phase transition [7]. As massive pre-training constructs a dense representational reservoir, higher-order capabilities like zero-shot reasoning manifest as sudden discontinuities once specific computational thresholds are surpassed [8]. Structuring latent knowledge through complex directives relies on the synergistic application of vast scale and alignment to ensure adherence to rigid formatting constraints. Functional utility is achieved. This developmental trajectory confirms that while architecture provides structural capacity and scaling

DOI: 10.54254/2755-2721/2026.CH31569

offers informational depth, only their integration transforms a statistical engine into a reliable, task-oriented system.

5.3.2. Limitations of pure scale

Unaligned models like Qwen2.5-1.5B underscore the structural limitations of pure scaling through persistent "continuation mode" errors. Scale is insufficient. Because the next-token prediction objective prioritizes statistical correlation over logical constraints, increasing parameter counts merely amplifies output complexity without ensuring adherence to human intent. Correlation is not alignment. To close the gap between raw power and human values, it is necessary to add alignment mechanisms like RLHF. This will turn probabilistic engines into reliable assistants where safety and controllability are built-in features of training rather than properties that come with scale.

6. Conclusion

The transition from sequential recurrent gating to parallelized Transformers represents a fundamental shift in processing efficiency. While recurrent models suffer from information decay and linear scaling bottlenecks, the self-attention mechanism maintains global semantic coherence within constant computational distances, providing the necessary infrastructure for massive-scale pre-training. A sharp dichotomy exists between raw computational scale and functional utility. Massive pre-training constructs an expansive "Cognitive Reservoir" of latent knowledge, but Intelligence remains latent. Refinement via alignment mechanisms like RLHF is mandatory to bridge the gap between statistical probability and human intentionality. This tripartite framework—comprising structural efficiency, volumetric growth, and intent-driven alignment—dictates that functional utility is a deliberately engineered product rather than a spontaneous emergent property of scale alone. Mapping probability onto intentional task completion requires a communicative interface; without it, even the most massive models remain trapped in a "continuation mode." Future validation must expand across larger parameter regimes, such as the 14B and 72B Qwen variants, to investigate whether the functional gap between base and instructed iterations narrows or widens at extreme scales. Representational density behavior will clarify the limits of the "Cognitive Reservoir" hypothesis. Intentionality must scale beyond linguistics. Evaluating alignment protocols for complex mathematical reasoning and multimodal image-text fusion constitutes a pivotal next step. Long-term computational sustainability requires exploring alternative neural architectures like Mamba frameworks to prioritize linear complexity alongside aggressive 4-bit quantization. Balancing raw representational power with practical constraints remains essential for the next generation of artificial intelligence.

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch, " Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.
[2] A. Vaswani et al., "Attention is all you need, " in Advances in Neural Information Processing Systems, vol. 30, 2017.
[3] Qwen Team, "Qwen2 technical report, " arXiv preprint arXiv: 2407.10671, 2024.
[4] J. Wei et al., "Emergent abilities of large language models, " Transactions on Machine Learning Research, 2022.
[5] J. Kaplan et al., "Scaling laws for neural language models, " arXiv preprint arXiv: 2001.08361, 2020.
[6] L. Ouyang et al., "Training language models to follow instructions with human feedback, " in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 27730–27744.

DOI: 10.54254/2755-2721/2026.CH31569

[7] R. Rafailov et al., "Direct preference optimization: Your language model is secretly a reward model, " in Advances in Neural Information Processing Systems, vol. 36, 2023.
[8] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?" in Advances in Neural Information Processing Systems, vol. 36, 2023.
[9] Y. Wang et al., "A comprehensive survey of LLM alignment techniques: RLHF, SFT, and beyond, " arXiv preprint arXiv: 2309.15025, 2023.
[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space, " arXiv preprint arXiv: 1301.3781, 2013.
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory, " Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.