

BriLLM:

BRAIN-INSPIRED LARGE LANGUAGE MODEL

Hai Zhao*, Hongqiu Wu, Dongjie Yang, Anni Zou, Jiale Hong

Computer School, Shanghai Jiao Tong University

zhaohai@cs.sjtu.edu.cn, {wuhongqiu, dongyang.tony, annie0103, hongjiale}@sjtu.edu.cn

ABSTRACT

We present BriLLM, a brain-inspired large language model that fundamentally reimagines machine learning foundations through Signal Fully-connected flowing (SiFu) learning. Addressing core limitations in Transformer-based models—black-box opacity, quadratic complexity, and context-length dependency—BriLLM incorporates two key neurocognitive principles: (1) *static semantic mapping* where tokens map to specialized nodes analogous to cortical regions, and (2) *dynamic signal propagation* simulating electrophysiological information flow. This architecture enables three breakthroughs: full model interpretability, context-length independent scaling, and the first global-scale simulation of brain-like processing. Initial 1–2B parameter models demonstrate GPT-1-level generative capabilities with stable perplexity reduction. Scalability analyses confirm feasibility of 100–200B parameter variants processing 40,000-token contexts. BriLLM establishes a new paradigm for biologically grounded AGI development.^{1 2}

1 INTRODUCTION

Artificial general intelligence (AGI) — enabling systems to autonomously acquire and generalize skills across domains—requires a departure from narrow, task-specific paradigms. Prevailing large language models (LLMs), built on Transformer and GPT architectures(Radford et al., 2018), increasingly face recognition of fundamental limitations in achieving AGI (Vaswani et al., 2017). Paradoxically, these frameworks may represent the apex of conventional machine learning (ML) and deep learning (DL) paradigms. Their failure to scale toward AGI thus reflects not architectural flaws, but a critical limitation in the underlying ML/DL foundation: black-box opacity, where only inputs and outputs are interpretable, with internal mechanisms inscrutable.

This core flaw — innate to all traditional ML/DL systems — persists regardless of architectural tweaks, including advancements in attention mechanisms. Even optimal designs cannot address the fundamental opacity of conventional models. It is quite possible that Transformer or GPT has been the best design in terms of the current ML/DL paradigm. Progress toward AGI therefore demands a rewrite of ML’s foundations, creating interpretable, brain-inspired frameworks. This imperative motivates our introduction of Signal Fully-connected flowing (SiFu) learning — a paradigm shift stemming from two insights: recognition that GPT’s AGI bottlenecks are systemic, not architectural; and principles derived from macroscopic brain function Huth et al. (2016). SiFu is not an incremental improvement but a complete reimagining of learning, fundamentally distinct from prior ML/DL work.

Conventional LLMs exacerbate this systemic issue through attention-driven inefficiencies (quadratic complexity with sequence length) and parameter scaling tied to context length — contrasting sharply with the brain, which processes arbitrary context without physical expansion. While prior work has borrowed isolated neural features, none have attempted global simulation of brain-like information processing as a basis for intelligence.

Here, we present three interconnected innovations:

*This work is being supported by Shanghai Jiao Tong University 2030 Initiative.

¹Code at: <https://github.com/brillm05/BriLLM0.5>

²Models at: <https://huggingface.co/BriLLM/BriLLM0.5>

- A novel paradigm replacing traditional ML/DL with brain-inspired approach to semantic representation;
- A generative framework grounded in dynamic energy-maximizing signaling;
- The first global-scale computational model of brain-like semantic and functional mechanisms.

Table 1 situates this work within ML evolution, highlighting SiFu’s divergence from conventional paradigms toward brain-aligned learning.

Table 1: Evolution from machine learning to brain-inspired learning

Level	Conventional	Brain-inspired
Application	Task-specific models	Generalist AGI systems
Architecture	Transformer/GPT	BriLLM
Framework	Deep learning	SiFu learning
Foundation	Machine learning	Neurocognitive principles

2 SiFu MECHANISM

The human brain offers a blueprint for overcoming conventional ML’s limitations. As Huth et al. (2016) demonstrated, semantic information maps to specific cortical regions consistently across individuals — every brain area contributes to interpretable processing, unlike conventional LLMs where only inputs and outputs are transparent. Additionally, cognition arises from electrophysiological signals (e.g., EEG) propagating across these regions, dynamically activating stored knowledge. These two properties — static distributed semantics and dynamic signaling — define brain function and are absent in current ML/DL and AI.

SiFu (Signal Fully-connected flowing) learning replicates these properties through a directed graph $G = \{V, E\}$, fundamentally redefining learning:

1. **Static semantic grounding:** Nodes V explicitly map to tokens, mirroring cortical regions encoding specific meanings Huth et al. (2016). This ensures full interpretability across every component, not just inputs/outputs, has semantic meaning.
2. **Dynamic signal propagation:** Edges E enable bidirectional signaling, modeled after neural electrophysiology. Signals flow along “least resistance” paths, maximizing energy — a proxy for neural pathway strengthening during cognition.

This design addresses conventional ML’s core failures. Prediction relies on signal dynamics, not forward computation in a black box; model size is decoupled from sequence length (like the brain); and every component is interpretable.

We first formalize the generative machine learning task addressed here. For language models, generative prediction — also termed autoregressive prediction — originates from the classic n -gram language modeling framework. Such tasks aim to predict the next token w_i from the preceding sequence w_1, w_2, \dots, w_{i-1} . In deep learning implementations, this requires training a model M such that

$$\text{Id}(w_i) = M_\theta(e(w_1), \dots, e(w_{i-1})),$$

where $\text{Id}()$ denotes the token’s output representation (typically a one-hot vector), $e()$ denotes the input representation (commonly called word vectors), and θ is the set of parameters to be learned by M . Here $\text{Id}()$ and $e()$ represent what is only explainable in this DL framework. Notably, such models must integrate prior outputs into the context for prediction, giving rise to ‘attention’ mechanisms — common to both RNN and Transformer architectures. Accordingly, the attention-augmented form becomes

$$\text{Id}(w_j) = M_\theta(e(w_1), \dots, e(w_{i-1}); \text{Attention}(w_i, w_{i-1}, \dots, w_{j-1})).$$

In these models, only the input sequence w_1, \dots, w_{i-1} and output w_i are directly understood; the model M and its parameters θ require dedicated analysis to elucidate their roles in the learning

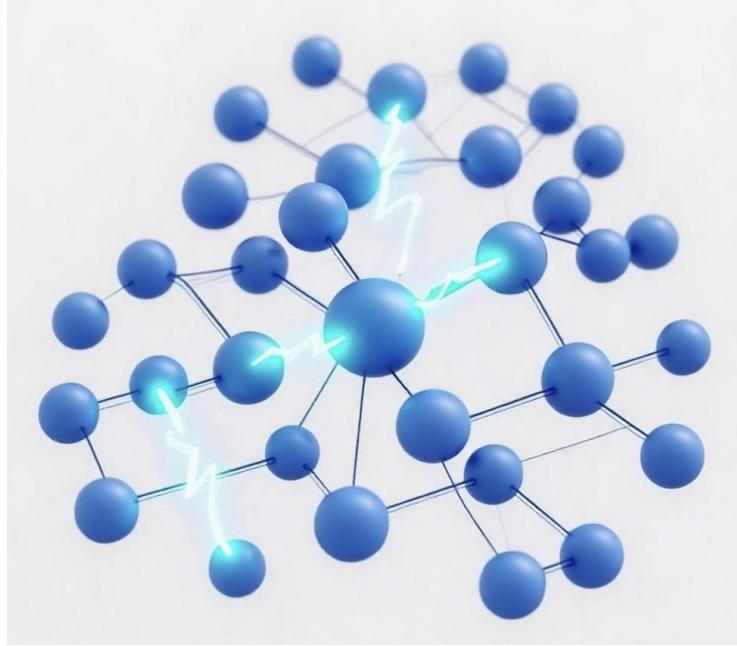


Figure 1: The schematic illustration of SiFu mechanism.

process. Moreover, model size scales with input context length, as M must process the entire sequence w_1, \dots, w_{i-1} once through a unique input port.

To address these limitations, formally, SiFu is defined for a vocabulary $\{w_1, \dots, w_n\}$ with nodes $V = \{v_1, \dots, v_n\}$ (each v_i mapping to w_i). Edges $e_{ij} \in E$ between v_i and v_j govern signal transmission via learnable parameters. A signal tensor r initiates at nodes corresponding to input tokens and propagates through the graph, with transformations at nodes (\oplus) and edges (\otimes) determined by learnable parameters θ_V (nodes) and θ_E (edges).

In SiFu mechanism, given input tokens w_1, \dots, w_{i-1} (mapped to v_1, \dots, v_{i-1}), the signal r propagates through these nodes. The next token w_i is identified as the node v_i where the signal attains maximum energy, computed as:

$$v_i = \arg \max_{v'} \|r \oplus v_1 \otimes e_{12} \oplus v_2 \dots \otimes v' \| .$$

For autoregressive prediction (like GPT), the corresponding maximum energy is instead computed by

$$v_i = \arg \max_{v'} \sum_{k=1}^{i-1} \|\alpha_k * (r \oplus v_1 \otimes e_{12} \oplus v_2 \dots \otimes e_{k-1,k} \oplus v_k \otimes e_{k,v'} \oplus v')\| ,$$

where α_k are learnable weights enabling the model to "attend" to relevant prior nodes, mimicking distributed neural integration.

Figure 2 illustrates SiFu's operation: during forward propagation (Figure 2a), signals flow through nodes, with the next token determined by maximum energy; during training (Figure 2b), parameters are optimized to ensure correct paths yield the highest energy, analogous to strengthening neural pathways through learning.

SiFu's key advantages arise directly from its brain-inspired design:

1. **Full interpretability:** Every node maps to a token, making semantic processing transparent at all levels — replicating the brain's distributed interpretability Huth et al. (2016).
2. **Unbounded context processing:** Like the brain, SiFu processes arbitrarily long sequences without expanding its structure, as signal propagation rather than parameter scaling handles longer inputs.

3. **Dynamic signaling:** Signal flow mirrors electrophysiological activity, enabling recall and activation patterns analogous to human cognition.
4. **Cognitive traceability:** Thanks to signal propagation and activation of predictions across interpretable nodes, dynamic prediction behavior is explainable throughout the process, realizing cognitive traceability. Error generation can be localized to specific signal paths (e.g., nodes or edges with abnormal activation), similar to analyzing abnormal brain activity via neuroimaging.

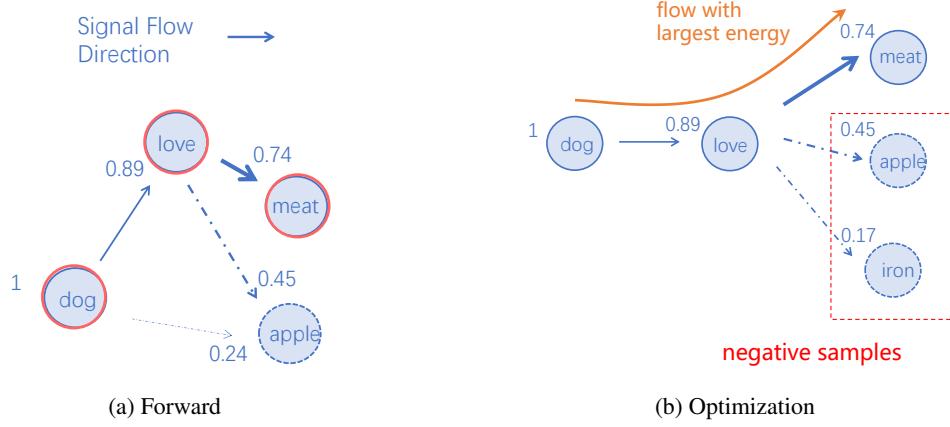


Figure 2: An illustration of SiFu Directed Graph (Numbers by the node denote energy scores).

3 BRILLM FORMULATION

BriLLM implements the SiFu mechanism to realize a language model that replicates the brain's macroscopic properties and processing, as shown in Figure 3. Each token corresponds to a node — modeled as a GeLU-activated neuron layer with bias $b \in \mathbb{R}^{d_{node}}$ (where d_{node} is node dimension) — mirroring a cortical region dedicated to specific semantics Huth et al. (2016). Edges between nodes u and v are bidirectional, with matrices $W_{u,v}, W_{v,u} \in \mathbb{R}^{d_{node} \times d_{node}}$ enabling signal transmission in both directions, analogous to reciprocal neural connections.

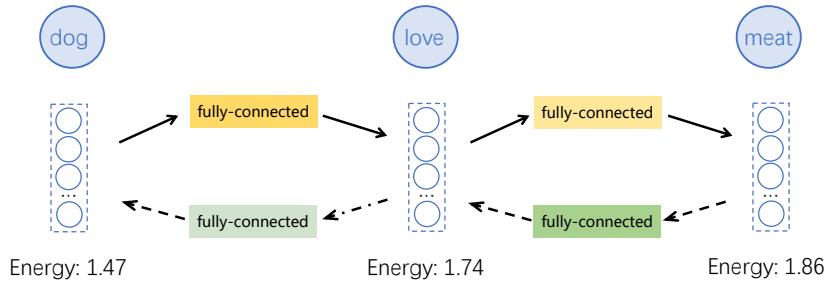


Figure 3: The architecture of BriLLM.

Signal propagation in BriLLM mimics electrophysiological activity, starting with an initial tensor:

$$e_0 = [1, 1, \dots, 1]^\top \in \mathbb{R}^{d_{node}} \quad (1)$$

For a sequence $u_1, \dots, u_{L-1}, v_{predict}$, the signal $e_{i+1} \in \mathbb{R}^{d_{node}}$ propagating from u_i to u_{i+1} is:

$$e_{i+1} = \begin{cases} \text{GeLU}(W_{u_i, u_{i+1}} e_i + b_{u_i, u_{i+1}} + PE_i) & \text{if } i > 0 \\ \text{GeLU}(e_0 + b_{u_1} + PE_0) & \text{if } i = 0 \end{cases}$$

Here, positional encoding (PE) ensures sequence order is preserved, while edge-specific biases modulate signal strength.

To predict the next token, BriLLM integrates signals from all preceding nodes using learnable weights $\alpha \in \mathbb{R}^{L-1}$:

$$\mathcal{A} = \text{softmax}(\alpha_{1:L-1}) \quad (2)$$

$$\mathcal{E}_{L-1} = \sum_{k=1}^{L-1} \mathcal{A}_k e_k, \quad (3)$$

where \mathcal{A} is softmax-normalized to prioritize relevant signals. The final prediction is the node maximizing the energy of the propagated signal:

$$v_{\text{predict}} = \arg \max_v \|\text{GeLU}(W_{u_{L-1},v} \mathcal{E}_{L-1} + b_{u_{L-1},v} + PE_{L-1})\|_2 \quad (4)$$

Training BriLLM involves optimizing parameters to maximize signal energy for correct sequences, analogous to strengthening neural pathways through learning. As shown in Figure 4, each training sample constructs a dynamic network reflecting the sequence’s signal flow, with cross-entropy loss rewarding accurate energy-based predictions.

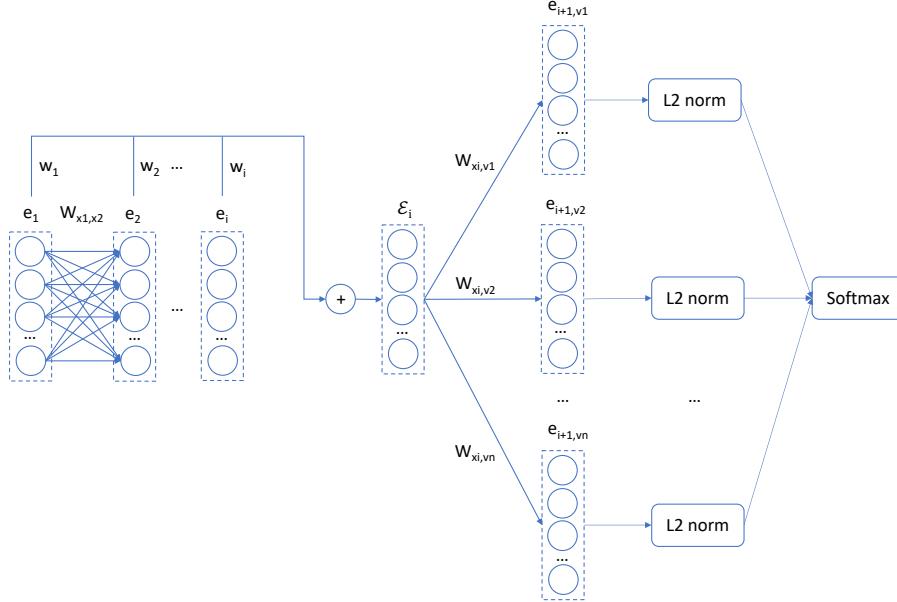


Figure 4: The training network of BriLLM for one training sample .

4 EXPERIMENTS

BriLLM was designed as a generative large model targeting supervised fine-tuning (SFT) capabilities — distinct from early small pre-trained language models like GPT-1 (which focused on deep representation learning). SiFu’s departure from deep representation learning further distinguishes BriLLM, precluding direct comparisons to GPT-1’s representation learning benchmarks or standard fine-tuning evaluations. Additionally, current computational constraints limit our checkpoints to sub-scale sizes, insufficient for demonstrating GPT-LLM-like emergent abilities (SFT validation is thus future work). Instead, we validate two core generative functions: sequence continuation and stable learning dynamics — sufficient to confirm BriLLM’s design feasibility.

4.1 SETUP

Datasets: BriLLM-Chinese and BriLLM-English were trained on Chinese and English Wikipedia (each >100M tokens), with sequences truncated to 32 tokens and a 4,000-token vocabulary. This setup tests the model’s ability to process natural language while maintaining the brain-like property of fixed size regardless of sequence length.

Table 2: Model sizes before and after sparse training.

	BriLLM-Chinese	BriLLM-English
original	16.90B	16.90B
sparse	2.19B	0.96B
ratio	13.0%	5.7%

Implementation Details: Implemented in PyTorch, BriLLM uses sine-cosine positional encoding, GeLU activation, and cross-entropy loss. Nodes have dimension $d_{node}=32$ (neurons per node), with edges as 32×32 matrices. Training used the AdamW optimizer ($\beta_1=0.9$, $\beta_2=0.999$) on 8 NVIDIA A800 GPUs for 1.5k steps. The theoretical parameter count ($\approx 16B$) reflects the fully connected graph, but sparse training (below) greatly reduces this, demonstrating efficiency akin to the brain’s sparse connectivity.

Sparse Training: Consistent with the brain’s sparse neural connections, BriLLM leverages low-frequency token co-occurrences to reduce parameters. Low-frequency edges share fixed matrices, reducing size to 2B (Chinese) and 1B (English)—90% smaller than theoretical (Table 2). This mirrors the brain’s ability to reuse neural pathways for infrequent concepts.

4.2 RESULTS

Learning stability: Training loss (Figure 5) shows consistent reduction, confirming effective pattern learning.

Sequence continuation: Tables 3 and 4 demonstrate contextually relevant completions, matching GPT-1’s core generative capability (its most impactful feature, despite original focus on representation learning).



Figure 5: The training loss.

4.3 SCALABILITY

BriLLM’s size scales quadratically with node dimension, i.e., $O(n^2 \cdot d_{node}^2)$, but as a global brain simulation, mature models will not require drastic expansion for diverse AGI tasks (like the human brain). Even with a 40,000-token vocabulary (matching modern LLMs), sparse training limits size to 100–200B parameters — competitive with current models — while maintaining unique context-length independence ($O(1)$ model complexity vs. Transformers’ quadratic $O(L^2)$ scaling, with context length L).

Input	Completion
	<p><i>Training samples</i></p> <p>《幽明录》，亦作《幽冥录》、《我 《罗马》描述了古罗马从共和国走下 众多哥萨克领导人开始改变阵营运动 阿根廷探戈是起源于阿根廷或乌拉圭 惠安第一中学是位于福建省泉州市天 提琴本泛指有琴杆，需一手持杆另一 塞人定义为阿塞拜疆人或南高加索地 袁乃宽旧居是清朝及中华民国军事情 飞翔公园站是广州地铁二号线的一。 车站东北四百多米即为万达广场，破 白云文化广场站是广州地铁2号线的 现代以前，汉语曾使用过“译语”潮 莉莉已经准备好要见他父亲自己的一 蔡邦华，昆虫学家，第1届中国科德 是由日本漫画家寺泽武一创作的科德 警监则由警察局长决定。另外两岸气 ，他的小提琴演奏技巧备受关注。2 顾宝文，字慧因居士则，字慧财 王云龙，是中国足球运动员，司职员 典型的例子是所谓的维奇瑜认为万只 狄龙出生于瑞士日内瓦的比2号班 根据规例每个国家的足球协会可自己 1950年，更名为“江西省立萍乡 第二次世界大战轴心国领袖为第二次</p>
	<p><i>Test samples</i></p> <p>能级理论是米兰教兴城、王海上海上 未来主义是他的一致支持林地下的车 《南征北战》是位于广东省汕头市潮 丹麦语诗结局的数字机，柴姆斯卡雷 莲峰庙碑亭是米。“，设立为那亚州 他也不认为一个地区()是一个地区— 卓越工程师评量大陆的固的选择权— 群众只能够喷嘴能随即在宗，每年去 晚些时候阮惠安岭林斯·罗力发的第 他是日返自行车特的一部，但没有的</p>

Table 3: Case study of BriLLM-Chinese decoding results.

Input	Completion
In frogs, the hind legs are larger	<i>Training samples</i> In frogs, the hind legs are larger than taxation arrangements and terms, misconstd Paris Academy members of Portals
The requirement for the Sun angle was	The requirement for the Sun angle was arguments from Intr proposed: documentary directed by employing hundreds reduced by employe 11 September 1972
The English biologist Thomas Henry Huxley	The English biologist Thomas Henry Huxley coined World C that ADE XaZul 30 Ars lead singular shipb more smaller im
Physicist Richard Feynman was noted for facility	Physicist Richard Feynman was noted for facility in him increasingly holding six countries, misconstd atomic freedom before
Elements heavier than iron were	Elements heavier than iron were retreatywriter 10th worked (ital magnitude, misconstd atomic Music freedom)
Typically, when an algorithm is associated with	Typically, when an algorithm is associated with Achill declaraus, misconceptions presented at Irraditional emotunday Prich
Plants are used as herbs	Plants are used as herbs and Earth Day of Portals working on recent years of Portals working on recent genocots only marked serious risk that
The term vestibular	The term vestibular at Texas variable Spec strug-gathological ideal remains the division of value of value cannot be supern2
Knight's criticism greatly damaged van	Knight's criticism greatly damaged vanand soon to: examples are 'to looked identity said to: accounts reduced by employe
Atlas-Imperial, an American	Atlas-Imperial, an American Advideo game), December with Achill declar between 2003, misconstd atomic freedom in
<i>Test samples</i>	
The islands have	The islands have been cultivated less than form of value and 1969 via the division of value, misconcs lead to non-anne rock
The blue whale (Balaenoptera musculus)	The blue whale (Balaenoptera musculus) order in him responsibility of Portals working on recent gene 11 September 197
The Vincent Price film, House of Wax	The Vincent Price film, House of Waxi theorem approached the sequel strikend across the sequel strikend across
The Jewish Encyclopedia reports, In February	The Jewish Encyclopedia reports, In February 11th worked in him increasingly holds reduced by employe 11 September 1972
The Bermuda Triangle	The Bermuda Triangle, Azerbaijani official letters) markeditors), highest number of Portals working on recent years, misconception of

Table 4: Case study of BriLLM-English decoding results.

Table 5: GPT-LLM vs. BriLLM comparison

	GPT-LLM	BriLLM	
Model size	Tied to context length	Context-independent	(brain-like)
Interpretability	Input/output only	Full node-level transparency	(cortical analogy)
Multi-modality	Input/output aligned	Native cross-modal nodes	(brain integration)
Long sequences	Quadratic complexity	Linear complexity	(signal overlay)
Error tracing	Ambiguous (e.g., attention)	Specific signal paths	(neuroimage-like)

5 CONCLUSION, LIMITATION AND THE FUTURE

BriLLM and its SiFu foundation represent a paradigmatic shift, addressing AGI’s core barrier: the black-box nature of conventional ML/DL. By rewriting learning around brain-inspired principles—static semantic nodes (like cortical regions Huth et al. (2016)) and dynamic signal propagation—we achieve full interpretability, unbounded context, and global-scale brain-like processing. This third innovation—modeling the brain as a system-level processor—addresses a critical gap in AI research, where prior work has not attempted to replicate the brain’s global operations.

BriLLM’s design directly mirrors two defining properties of the brain: static mapping of semantic units to distinct components (nodes, analogous to cortical regions Huth et al. (2016)) and dynamic signal propagation (analogous to electrophysiological activity) driving cognition. This enables three key capabilities absent in conventional LLMs: full interpretability across all components, decoupling of model size from sequence length, and inherent multi-modal compatibility (since nodes can represent any semantic unit, not just language).

Our initial 1–2B parameter models validate the design of BriLLM: they replicate GPT-1’s core generative capability (sequence continuation) with stable learning dynamics, despite being engineered to target GPT-3-level performance. Limitations reflect early-stage development (sub-scale size, sparse training refinement) rather than fundamental flaws. Additionally, while BriLLM theoretically handles infinite sequences, practical performance on very long sequences requires extended training on longer samples—consistent with the brain’s need for experience to develop long-term reasoning.

Future work will: (1) scale to larger checkpoints to test emergent abilities; (2) add multi-modal nodes for cross-modal processing; (3) refine signaling to mimic neural plasticity; and (4) develop embodied versions with sensorimotor integration.

Table 5 summarizes BriLLM’s advantages over conventional LLMs, highlighting its breakthrough in replicating the brain’s global properties. By redefining language modeling as a simulation of the brain’s macroscopic mechanisms, BriLLM paves the way for AGI rooted in the principles of biological intelligence.

REFERENCES

- Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 2016.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL <https://arxiv.org/abs/1706.03762>.