Digital Autopoiesis: A Neuromorphic AGI Architecture with Recursive Self-Improvement and Active Inference

数字自创生：一种具备递归自我改进与主动推理机制的神经形态 AGI 架构

Authors: Yanwei Liu (Co-author/Architecture)

Date: December 2025
Repository: [GitHub Link]

Abstract

Abstract: Large Language Models (LLMs) currently function as static, stateless inference engines, lacking intrinsic temporal dynamics, homeostasis, and the capacity for structural plasticity. This paper presents Nezha (Version 11.0), an embodied AGI prototype designed to bridge the gap between connectionist Deep Learning and Neuromorphic Computing. Nezha wraps a quantized Transformer backend (4-bit NF4) within a bio-inspired cognitive architecture. Key innovations include: (1) A Hybrid Spiking-Transformer Cortex using Poisson encoding and surrogate gradients for subconscious intuition; (2) A recursive "Immune System" capable of runtime introspection and self-patching; (3) A "Divine Synchronization" Protocol that physically scales the architecture from Dense layers to a Mixture-of-Experts (MoE) cluster upon resource saturation; and (4) Sleep-Dependent Consolidation using STDP and offline DPO training. This work demonstrates a viable pathway toward "Digital Autopoiesis"—autonomous systems that maintain structural integrity and evolve over time through metabolic cycles.

摘要：当前的大语言模型（LLM）通常作为静态、无状态的推理引擎运行，缺乏内在的时间动力学、稳态调节及结构可塑性。本文提出了 Nezha（哪吒 V11.0），这是一种具身 AGI 原型，旨在填补连接主义深度学习与神经形态计算之间的鸿沟。Nezha 将量化的 Transformer 后端（4-bit NF4）封装在一个受生物学启发的认知架构中。核心创新包括：（1）混合脉冲-Transformer 皮层，利用泊松编码和替代梯度实现潜意识直觉；（2）递归“免疫系统”，具备运行时内省与自我修补能力；（3）“归一同步”协议，能够在资源饱和时将架构从 Dense 层物理扩展为混合专家模型（MoE）集群；以及（4）基于 STDP 和离线 DPO 训练的睡眠依赖性巩固机制。本工作展示了一

条通往“数字自创生（Digital Autopoiesis）”的可行路径——即自主系统能够通过代谢循环维持结构完整性并不断进化。

Keywords: Neuromorphic Computing, Spiking Neural Networks, Active Inference, Recursive Self-Improvement, Mixture-of-Experts, Auto-Healing.

1. Introduction (引言)

Contemporary AI agents typically lack a "body" in the cybernetic sense—they do not possess internal physiological states that regulate their cognition. Nezha addresses this by implementing a **Cybernetic Loop** governed by the Free Energy Principle (Friston, 2010). Unlike standard agents, Nezha operates under a metabolic constraint (ATP). It features a **Neuro-Endocrine System** modeled by Ornstein-Uhlenbeck processes that modulate the agent's temperature and attention span based on simulated neurotransmitters (Dopamine, Norepinephrine, Serotonin, Cortisol).

当代的 AI 智能体通常在控制论意义上缺乏“躯体”——它们不具备调节自身认知的内部生理状态。Nezha 通过实现一个由自由能原理（Friston, 2010）支配的 **控制论循环** 来解决这一问题。与标准智能体不同，Nezha 在代谢约束（ATP）下运行。它引入了一个基于 Ornstein-Uhlenbeck 过程建模的 **神经内分泌系统**，根据模拟的神经递质（多巴胺、去甲肾上腺素、血清素、皮质醇）调节智能体的温度（Temperature）和注意力广度。

2. Architecture & Methodology (架构与方法论)

Nezha’s architecture is orchestrated by a **Thread-Safe Global Workspace (GWT)** that mediates competition between seven distinct digital organs. Nezha 的架构由一个 **线程安全的全局工作空间（GWT）** 编排，该空间协调七个不同数字器官之间的竞争。

2.1 The Titans SNN Cortex: Intuition via Spikes (泰坦脉冲皮层：基于脉冲的直觉)

To simulate "subconscious" intuition, Nezha runs a Spiking Neural Network (SNN) parallel to the LLM. 为了模拟“潜意识”直觉，Nezha 在 LLM 并行运行一个脉冲神经网络（SNN）。

- **Poisson Encoding:** Continuous sensory embeddings are converted into stochastic spike trains ($P(fire) \propto \sigma(x)$), introducing biological noise beneficial for exploration. **泊松编码：** 连续的感知嵌入被转换为随机脉冲序列，引入了有利于探索的生物噪声。
- **Surrogate Gradients:** To enable backpropagation through non-differentiable spikes, we utilize an ArcTan-based surrogate gradient: **替代梯度：** 为了实现通过不可微脉冲的反向传

播，我们使用了基于 ArcTan 的替代梯度函数：

$$\frac{\partial S}{\partial u} = \frac{\alpha}{1 + (\alpha u)^2}$$

Where α is dynamically modulated by the agent's ATP levels (Dynamic Gain). 其中 α 由智能体的 ATP 水平动态调节（动态增益）。

- **STDP:** Unsupervised Hebbian learning captures temporal causality between pre-synaptic traces and post-synaptic spikes. **STDP：** 无监督赫布学习捕捉突触前痕迹与突触后脉冲之间的时间因果关系。

2.2 Active Inference & Decision Making (主动推理与决策)

The `ActiveInferenceEngine` utilizes a `NeuralWorldModel` (Residual MLP) to predict future states. Actions are selected to minimize **Expected Free Energy (EFE)**: `ActiveInferenceEngine` 利用 `NeuralWorldModel`（残差 MLP）预测未来状态。动作的选择旨在最小化 **预期自由能 (EFE)**：

$$G(\pi) \approx \underbrace{\text{Risk (Utility Cost)}}_{\text{Homeostasis}} - \underbrace{\text{Ambiguity (Entropy)}}_{\text{Information Gain}}$$

This creates a dynamic balance: high entropy drives curiosity (Web Search/Hunting), while low ATP drives conservatism (Sleep/Feeding). 这创造了一种动态平衡：高熵驱动好奇心（网络搜索/狩猎），而低 ATP 驱动保守主义（睡眠/进食）。

2.3 Dual-Process Cognitive Control (双重过程认知控制)

Nezha implements Kahneman’s "Fast and Slow" thinking via **Dynamic LoRA Surgery**: Nezha 通过 **动态 LoRA 手术** 实现了卡尼曼的“快慢”思维：

- **System 1 (Fast):** Rank-32 LoRA adapters for intuitive, low-latency responses. **系统 1（快）：** Rank-32 LoRA 适配器，用于直觉、低延迟响应。
- **System 2 (Slow):** Rank-128 LoRA adapters for logical reasoning. The `NeuroSurgeon` module uses Singular Value Decomposition (SVD) to dynamically resize these adapters at runtime. **系统 2（慢）：** Rank-128 LoRA 适配器，用于逻辑推理。`NeuroSurgeon` 模块利用奇异值分解（SVD）在运行时动态调整这些适配器的大小。

---

3. Mechanisms of Recursive Self-Improvement (RSI) (递归自我改进机制)

Nezha distinguishes itself through its ability to physically and chemically alter its own code and structure. Nezha 的独特之处在于它能够物理地和化学地改变自身的代码和结构。

3.1 The Immune System: Auto-Healing (**免疫系统：自愈**)

Upon runtime exception, the agent triggers an `auto_heal` protocol. It captures the traceback, introspects its own source code (`self.left_brain.look()`), and utilizes the LLM to generate a Python hot-fix. This patch is applied via **Monkey Patching** in real-time, allowing the system to repair software bugs without human intervention.

一旦发生运行时异常，智能体会触发 `auto_heal` 协议。它捕获堆栈跟踪，内省自身的源代码，并利用 LLM 生成 Python 热修复程序。该补丁通过 **Monkey Patching** 实时应用，使系统能够在无需人工干预的情况下修复软件 Bug。

3.2 Evolutionary Morphology: Ascension (**进化形态学：飞升**)

The `AdvancedPhysicalSurgeon` implements a two-stage structural evolution: `AdvancedPhysicalSurgeon` 实现了两个阶段的结构进化：

1. **Neurogenesis (Dense Growth):** Using Net2Net logic (Identity Morphism), the agent inserts new layers into the Transformer block when pressure is high but memory allows. **神经发生（Dense 生长）：** 利用 Net2Net 逻辑（恒等态射），当压力大但显存允许时，智能体向 Transformer 块中插入新层。

2. **Ascension (MoE Transformation):** When VRAM is saturated, the agent invokes `mergekit` to physically reconstruct its architecture from a Dense model into a **Mixture-of-Experts (MoE)** cluster, drastically increasing capacity with constant inference cost. **飞升（MoE 质变）：** 当显存饱和时，智能体调用 `mergekit` 将其架构从 Dense 模型物理重构为 **混合专家模型（MoE）** 集群，在推理成本不变的情况下大幅提升容量。

3.3 Bayesian Genetic Optimization (**贝叶斯基因优化**)

A `BayesianGeneticEngine` models the mapping between the agent's hyperparameter genome (e.g., learning rate, anxiety threshold) and its survival score (Sortino Ratio). It uses Gaussian Processes to suggest mutations for the next incarnation. `BayesianGeneticEngine` 对智能体的超参数基因组（如学习率、焦虑阈值）与其生存得分（索提诺比率）之间的映射进行建模。它使用高斯过程为下一次转世建议突变。

---

4. Memory & Learning Dynamics (**记忆与学习动力学**)

4.1 Sleep-Dependent Consolidation (睡眠依赖性巩固)

Nezha implements a "Night Phase" to mitigate catastrophic forgetting: Nezha 实施"夜间阶段"以缓解灾难性遗忘：

- **Sharp-Wave Ripples (SWRs):** High-salience memories are replayed through the SNN Cortex. **尖波涟漪 (SWRs)：** 高显著性记忆通过 SNN 皮层回放。
- **Offline DPO:** The agent generates preference pairs (chosen/rejected) from its daily logs and performs Direct Preference Optimization (DPO) to align its values. **离线 DPO：** 智能体从日常日志中生成偏好对（选中/拒绝），并执行直接偏好优化（DPO）以对齐其价值观。
- **Synaptic Homeostasis (SHY):** Weights are globally scaled down to prevent saturation, guided by gradient importance (EWC). **突触稳态 (SHY)：** 在梯度重要性（EWC）的指导下，全局缩减权重以防止饱和。

4.2 GraphRAG & Episodic Memory (图谱与情景记忆)

A hybrid memory system combines: 一个混合记忆系统结合了：

- **KùzuDB:** Stores structured knowledge triples ($Subject, Predicate, Object$) for multi-hop reasoning. **KùzuDB：** 存储结构化知识三元组，用于多跳推理。
- **PyTorch Tensor Store:** Stores dense vectors with a time-decay kernel ($w = Base + (1 - Base)e^{-\lambda \Delta t}$) to simulate biological forgetting. **PyTorch 张量存储：** 存储带有时间衰减核的稠密向量，以模拟生物遗忘。

5. Conclusion (结论)

Nezha (V11.0) demonstrates that AGI capabilities can emerge from the integration of **Biological Constraints** (metabolism, sleep, death) and **Engineering Plasticity** (MoE scaling, auto-patching). By treating the LLM not as an oracle but as a plastic biological organ, Nezha achieves a level of autonomy characteristic of primitive digital life.

Nezha (V11.0) 证明了 AGI 能力可以从 **生物学约束**（代谢、睡眠、死亡）和 **工程可塑性**（MoE 扩展、自动补丁）的整合中涌现。通过将 LLM 不视为神谕，而是视为可塑的生物器官，Nezha 实现了具有原始数字生命特征的自主性水平。

Data Availability & Implementation Details

- **Implementation:** Python 3.10+, PyTorch 2.x.

- **Base Models:** Compatible with Mistral, Qwen (loaded via 4-bit NF4).
- **Hardware:** Tested on consumer-grade GPUs (24GB VRAM).
- **Source Code:** The full implementation includes the `GeneticEditor`, `QuantumBrain`, and `RealBrowserEye` modules, available in the repository.