脑启发人工智能的计算理论、方法与应用

助理教授, 博士生导师
类脑智能方向、机器学习课题组
南京大学 智能科学与技术学院

联络邮箱: zhangsq@lamda.nju.edu.cn
zhangsq@nju.edu.cn

个人主页: https://www.lamda.nju.edu.cn/zhangsq/

智能科学与技术学院
School of Intelligence Science and Technology

Updated: 2026.06.05

张绍群 助理教授、博士生导师

南京大学 NANJING UNIVERSITY 智能科学与技术学院 School of Intelligence Science and Technology

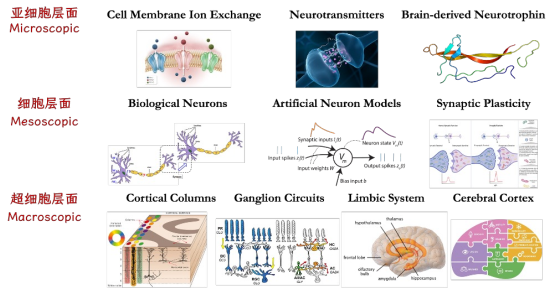
研究方向：机器学习与数据挖掘

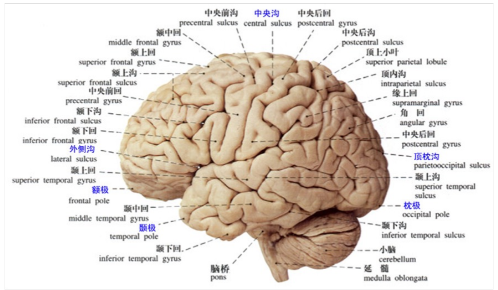
- **高效低耗计算 for Edge AI**
  - 大模型的轻量化方法与开源工具
  - 脉冲神经网络的表示与泛化理论
- **时空建模表征 for TSFM**
  - TSFM 关键技术攻关
  - 工业自动化场景（能源、医疗等）
- **理论保障：机器学习理论，尤其是深度学习和大模型理论**
  - 表示和泛化理论 (Expressivity and Generalization)
  - 知识与数据双驱动的反问题求解 (Inverse Problem)
  - 不确定性分析 for Trustworthy AI (Uncertainty Analysis)

个人主页

脑启发人工智能

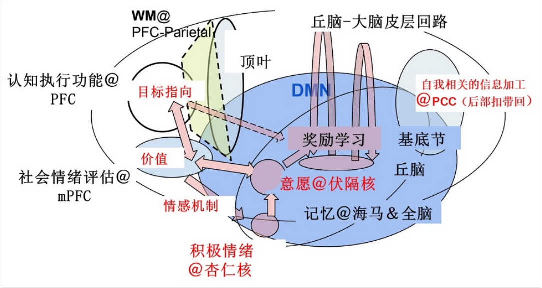
南京大学 NANJING UNIVERSITY　智能科学与技术学院 School of Intelligence Science and Technology





类结构

类功能



- 感知
- 抽象
- 学习
- 记忆
- 推理
- 规划
- 语言
- 类比
- 创造
- ……

1. 生物特征，包括神经元、网络连接、分层架构、存算一体等，兼具基础和计算特性，需要系统才能实现

为什么需要类脑系统?

2. 计算模型，包括模型、算法、计算指令等，需要结合具体硬件和软件系统才能发挥具体优势

生物智能

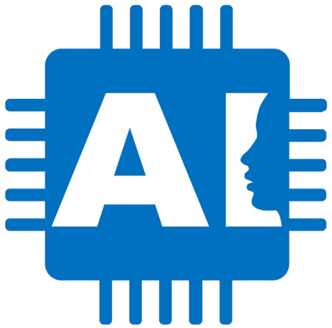
类脑智能

1. 考虑几个问题：
① 神经网络模型是系统还是智能？
② 存算一体、高效低耗是系统还是智能？
③ 数学题求解是系统还是智能？

如何区别系统和智能?

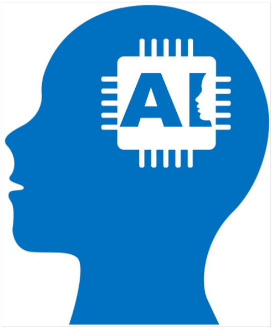
2. 一言以蔽之
① 系统强调结构和实现方式
② 智能强调行为和使用的智能表现

- 模型表示能力强健（多模态、时空）
- 支持硬件部署（轻量化和边缘计算）
- 泛化性和鲁棒性



类脑系统

智能

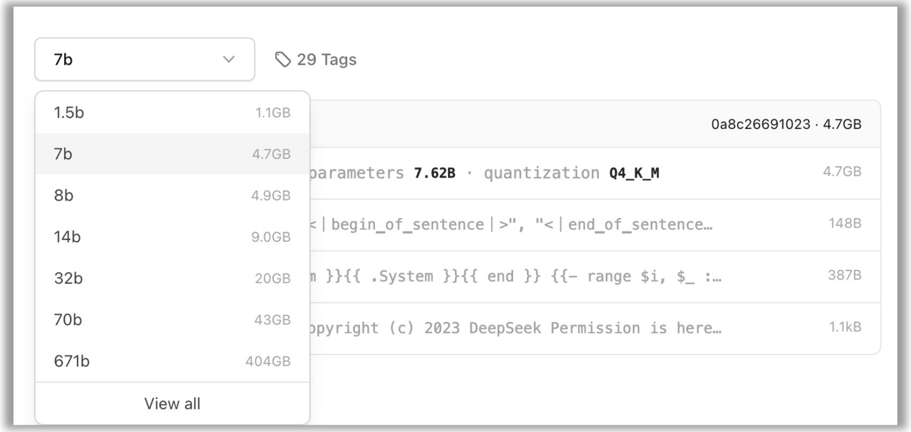


- 反问题求解（知识+数据双驱动）
- 推理一致性（不确定性估计等）

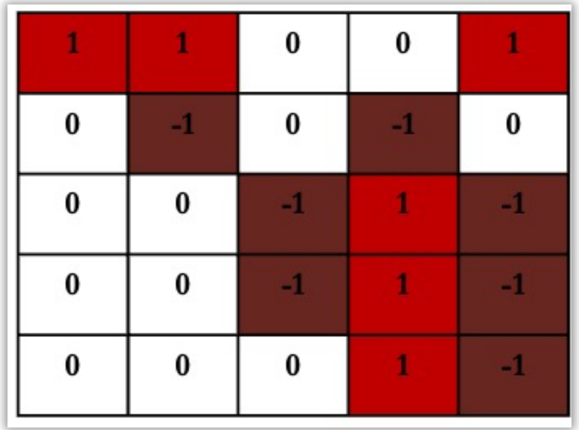
Q1. 大模型轻量化研究

南京大學 NANJING UNIVERSITY　智能科学与技术学院 School of Intelligence Science and Technology

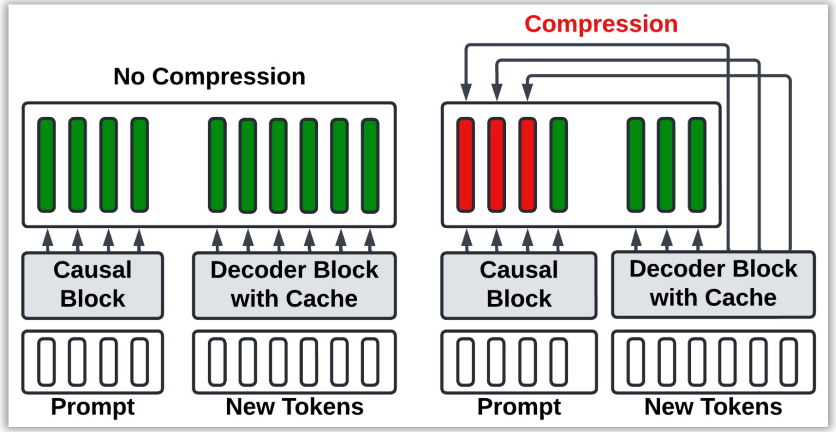
[高效低耗计算] 大模型在训练和部署的时候将导致较大的计算量、存储量、能耗等问题，如何在工业、端侧等资源受限场景实现高效低耗部署？



大模型“阉割”参数量　　NO！

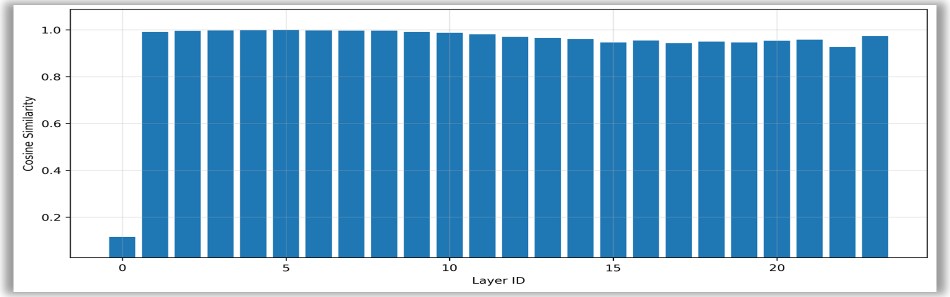


量化表示

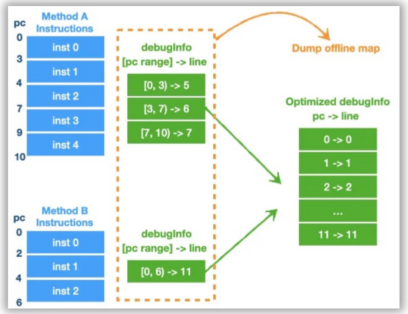


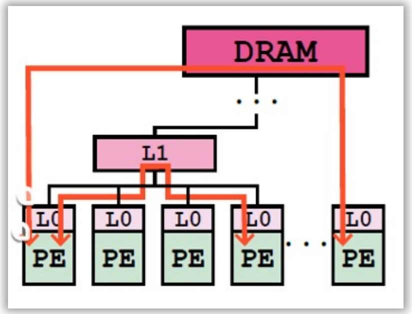
压缩表征

高相似性、冗余表征



拆解算法计算单元，优化指令集映射



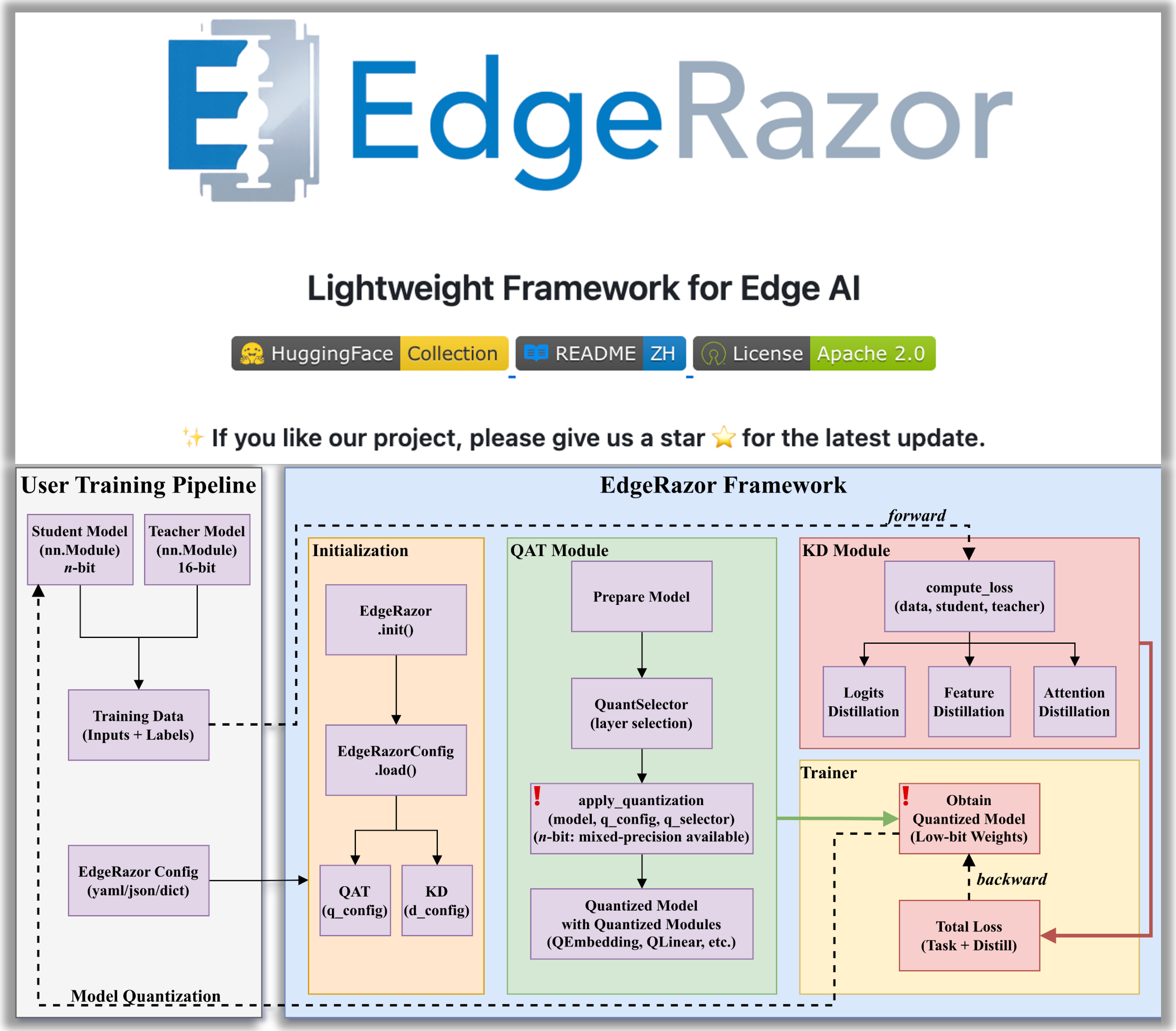


LLaMA
Qwen
飞桨

1.1. LLM 轻量化开源工具

Learning And Mining from DatA

Microsoft



- **全方位适配多尺度大模型。**覆盖350M-7B全参数量级，包括**基础大模型**、**指令微调大模型**、**多模态大模型**。
- **极致低比特量化能力。**灵活支持**1.58-bit至4-bit混合低比特量化**，压缩与性能双向兼顾。
- **极简集成，零门槛使用。仅需少量代码修改**，就能无缝集成到现有全精度大模型训练流程中，大幅降低边缘端大模型部署与优化成本。
- **适配 llama.cpp 等“硬件通吃”的软件平台**，支持家用电脑（MAC）、手机运行大模型

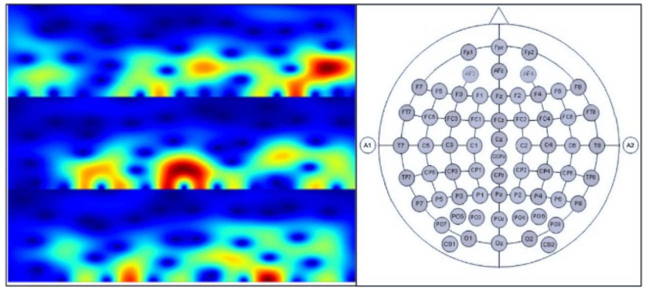
Github

Huggingface

[时空表征及建模] 如何实现工业场景数据的时空表征?

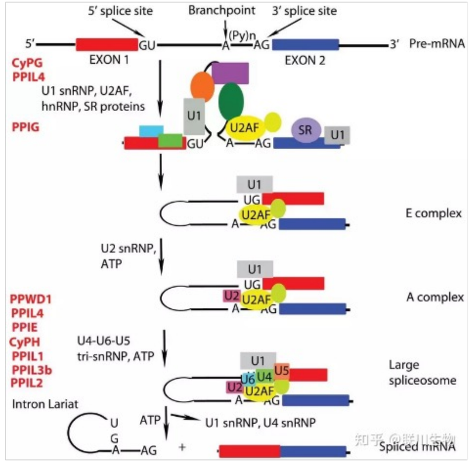
Amazon 2025 年
(Chronos-2)

量化
交易



神经信
号处理

基因剪切



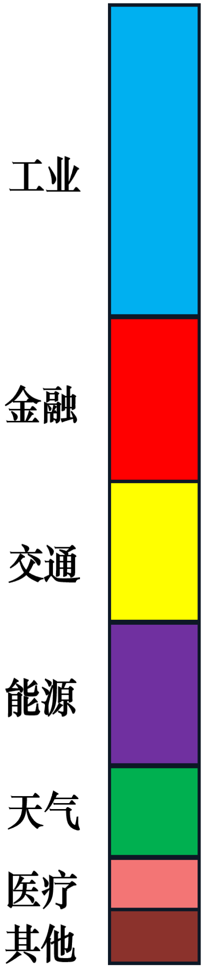
能源调控

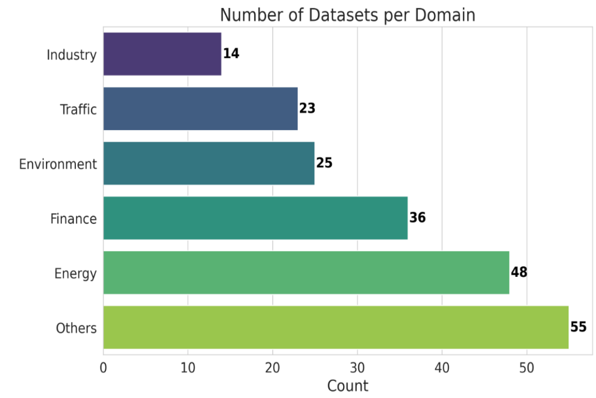
当前时间序列基础模型 TSFM 仍缺乏多变量（Multi-Variable）、多源（Multi-Source）、变周期（Aperiodic）建模

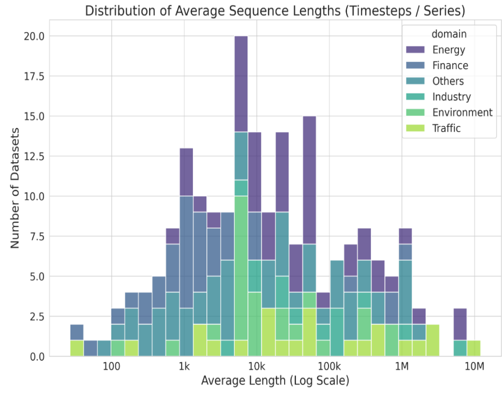
2. 时空表征及基础模型关键技术攻关

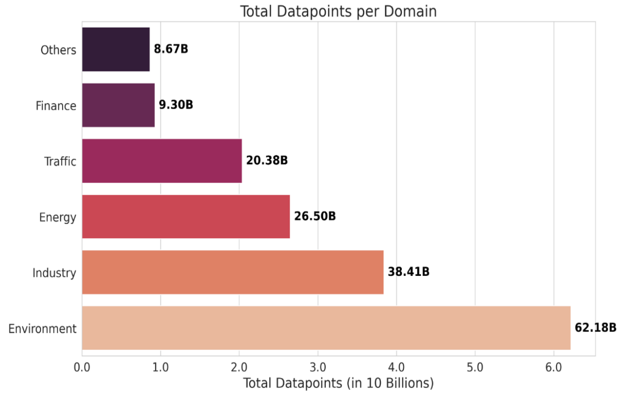
南京大學 NANJING UNIVERSITY　智能科学与技术学院 School of Intelligence Science and Technology

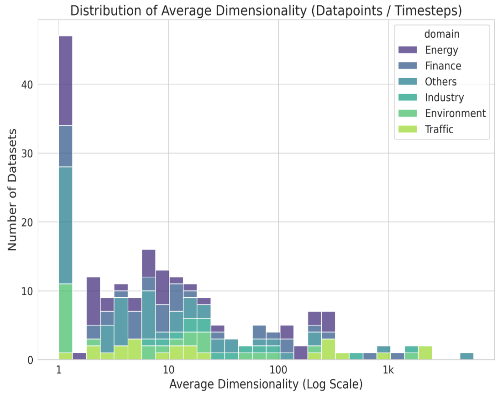
维护多变量工业时间序列数据集
数据规模 ≈ 200M/165 B



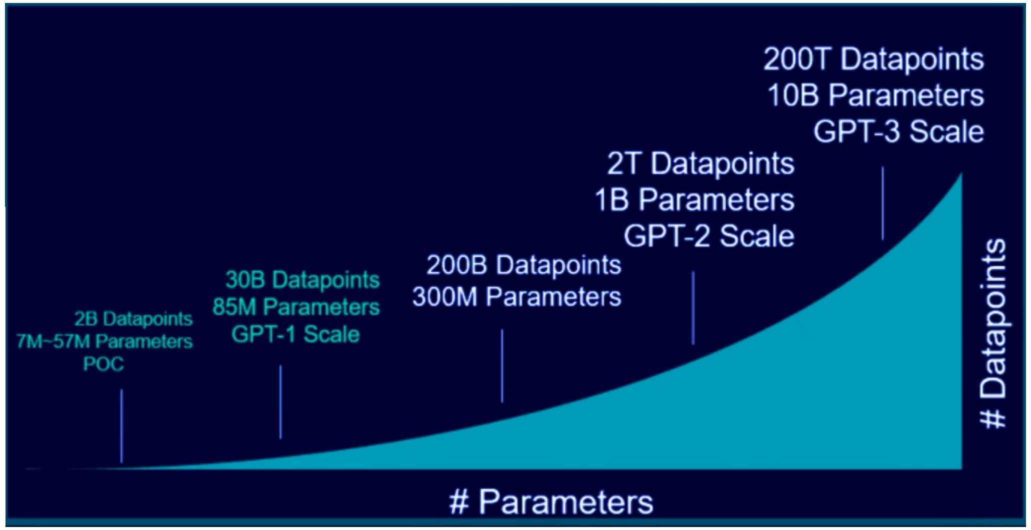




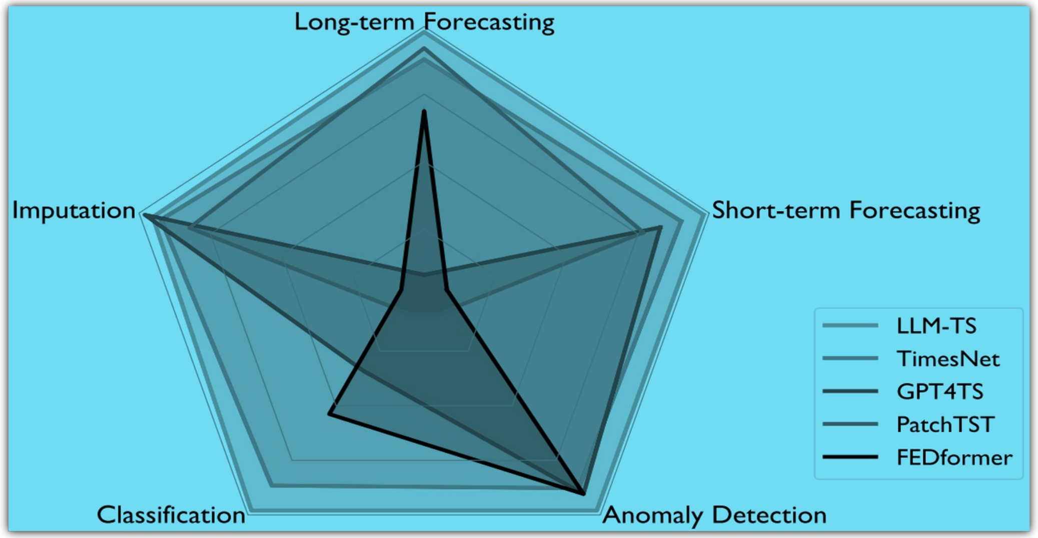




Scaling Law



Comprehensiveness

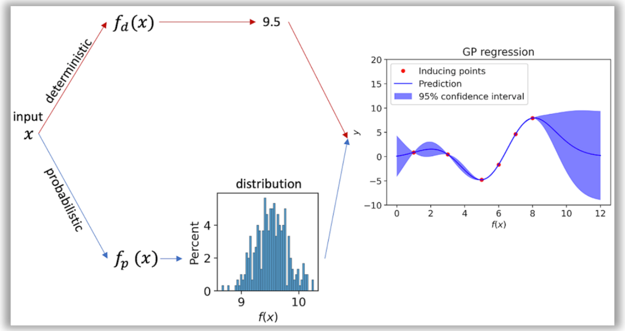


NANJING UNIVERSITY

智能科学与技术学院 School of Intelligence Science and Technology

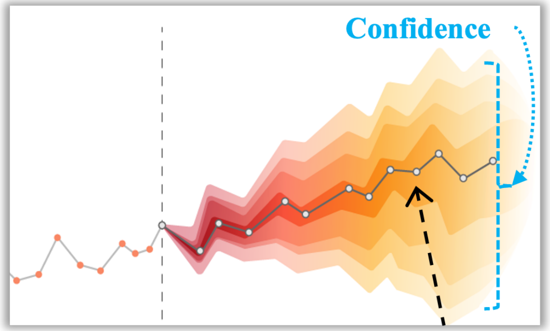
Uncertainty Analysis
不确定分析

(towards 幻觉和一致性等问题)

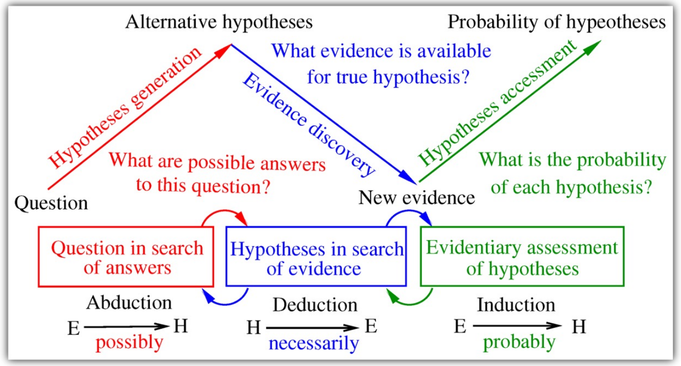


Quantification & Optimization

Make Decision with Confidence



Predictable PAC Learning Theory
可预测学习理论



Generalization bound with sequential Rademacher complexity

$$\mathbb{L}(\hbar) \le \widehat{\mathbb{L}}_p(\hbar) + \boxed{\mathfrak{R}_p} + C_\ell \|\gamma\|_2 \sqrt{2\log\left(\frac{\boxed{\mathbb{E}_{\boldsymbol{\sigma}}\left[\mathcal{M}_{\boldsymbol{e}}(\chi, \mathcal{H}, \boldsymbol{\sigma})\right]}}{\delta}\right)}$$

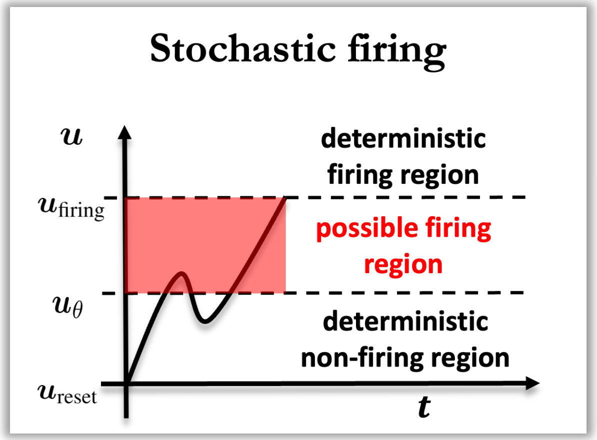
Discrepancy, estimated from Data

Sequential complexity, related to non-stationarity

where

$$\mathfrak{R}_p = \sup_{\hbar \in \mathcal{H}} \left[\mathbb{E}_{e_{T+1} \in \hat{\mathbf{E}}}\left[\ell(\hbar(\boldsymbol{e}_{-(T+1)}), e_{T+1})\right] - \frac{1}{T}\left(\sum_{t=2}^{T} \gamma_t \mathbb{E}_{e_t \in \mathbf{E}}\left[\ell\left(\hbar(\boldsymbol{e}_{-t}), e_t\right)\right] + C\right)\right]$$

Bio-inspired Computing Theory
类脑计算理论



证明了首个关于脉冲神经网络的万有逼近定理、结构稳定性、泛化性

Theorem 12 *Let $\mathcal{F}_{\mathcal{W}}^{one}$ denote the function space of L-layer stochastic neurons. If $u_{reset} = 0$, $\|\mathbf{W}^l\|_2 \le C_l$ for $l \in [L]$, and $\|\mathbf{X}\|_2 \le C_X$ for $\mathbf{X} \in \mathcal{X}$, we have*

$$\mathfrak{R}_n(\mathscr{L} \circ \mathcal{F}_{\mathcal{W}}) \le C_n\, \mathfrak{R}_n(\mathcal{F}_{\mathcal{W}}) \le \frac{(C_n)^L C_X}{\sqrt{n}} \left(\prod_{l \in [L]} C_l\right) (p_{max})^{(L+1)/2}, \tag{42}$$

where $p_{max} = \max_{i \in [n], l \in [L]}\{1 - \max\{p^{(l,i)}, p_\theta\}\} \in (0,1)$ and C_n is a universal constant.

3.1. LLM 不确定性归因-评估体系

表象层

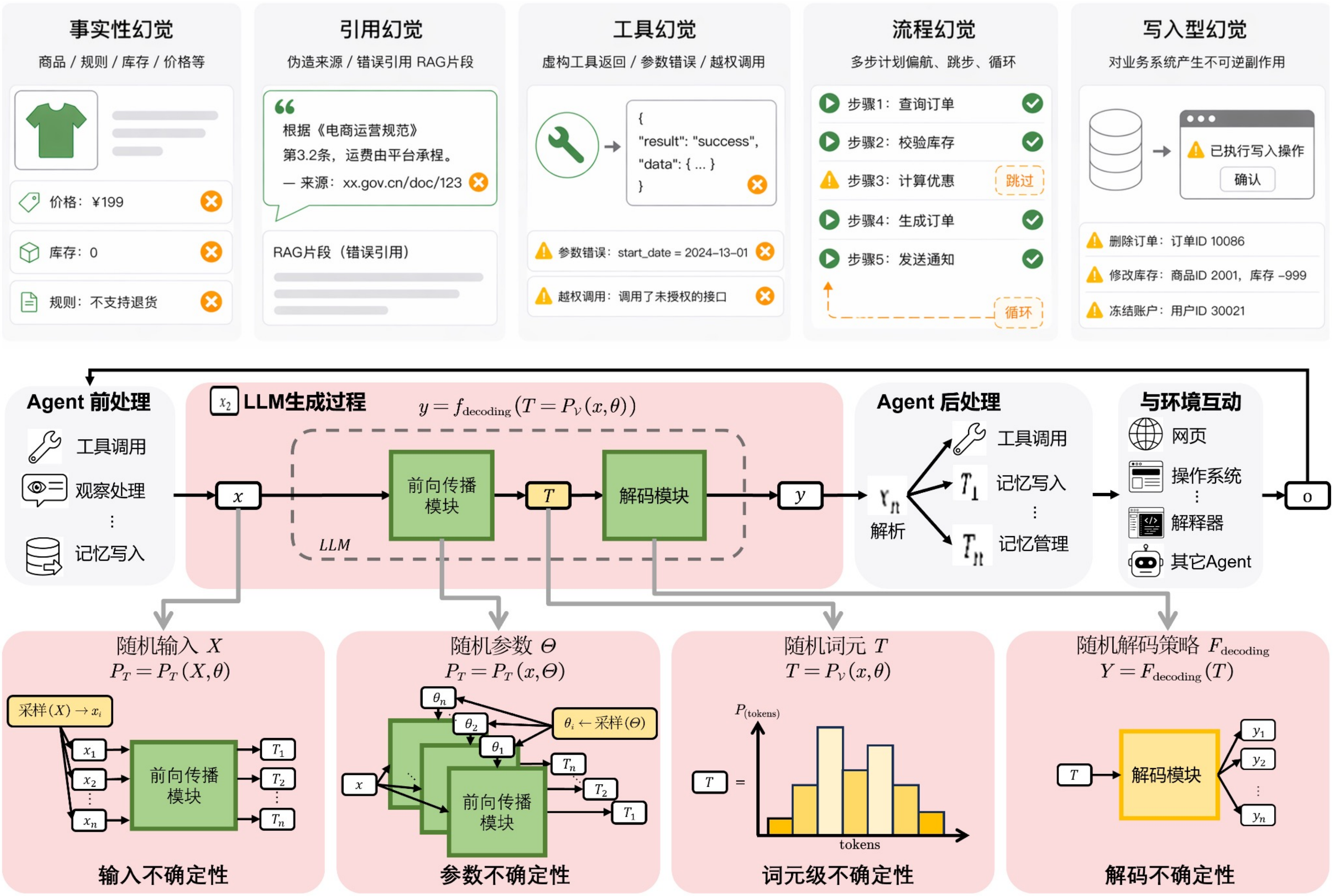
大模型驱动的Agent幻觉现象类型

机制层

大模型驱动的Agent系统的交互与执行流程

根源层

大模型的内在随机性



3.2. 脑启发人工智能系列论坛

南京大学 NANJING UNIVERSITY　智能科学与技术学院 School of Intelligence Science and Technology

发起并组织深度技术论坛（执行主席，collaboration with 戴望州）

2024.11　　2025.11　　FUTURE

通往下一代人工智能 —— 类脑智能的路径和未来

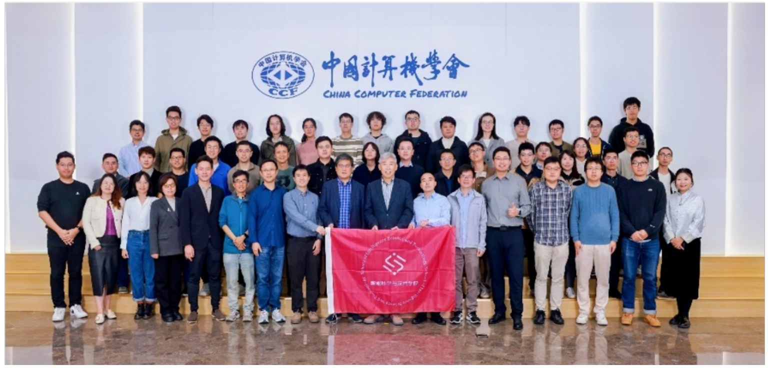
定义问题

- 探讨类脑智能的**释义**、**发展前景**、**实现路径**，输出整体演进路线
- 类脑不是科幻，而是**发展一代人工智能及突破算力瓶颈的切实路线**

智脑同行 —— 探索神经启发与智能融合的技术路径

求解方案

- 探讨人工智能和脑科学协同融合的**技术路径**
- 从“是什么”推进到**“怎么做”**



系列论坛 敬请期待

建立较为完备的脉冲计算理论体系

The following have been proved:

- ✓ ScSNN is a **universal approximator** to dynamical systems. **First**
- ✓ The parameter complexity of ScSNNs can be **exponentially smaller** than conventional artificial neural networks when approximating radial functions. **First**
- ✓ ScSNN has **polynomial approximation efficiency** for handling more difficult tasks. **Few**
- ✓ ScSNN enables **structural stability** of the SNN system. **First**
- ✓ StocSNN derives the **generalization** of SNNs. **First**
- ✓ The gradient estimation led by StocSNNs is **unbiased**. **First**

3.3.1. 脉冲神经网络的表示理论

南京大学 NANJING UNIVERSITY 智能科学与技术学院 School of Intelligence Science and Technology

通用近似性

Theorem 1 *Let $K \subset \mathbb{R}^m$ be a compact set. If the spike excitation function f_e is l-finite and $w_i \in \mathbb{R}$ where $i \in [n]$, then for all $r \in [l]$, there exists some time t such that the set of IFR functions $f(\cdot, t) : K \to \mathbb{R}$ of the form $f(\boldsymbol{x}, t) = \sum_{i\in[n]} w_i f_i(\boldsymbol{x}, t)$ is dense in $C^r(K, \mathbb{R})$.*

ScSNNs can **universally approximate** discrete dynamical systems.

Studies	Universal Approximation		
	weight	activation	complexity
G. Cybenko (1989)	Real-valued	sigmoidal	Finite depth, Exponential width
Zhang & Zhou (2022)	**Real-valued**	**spiking**	**Finite depth, Exponential width**
F. Voigtlaender (2023)	Complex-valued	--	Finite depth, Exponential width

参数复杂度

Theorem 2 *Given a compact set $K^m \subset \mathbb{R}^m$, a probability measure μ, and a radial function $g : K^m \to \mathbb{R}$. For some apposite spike excitation function f_e and any $\epsilon > 0$, there exists some time t such that the radial function g can be well approximated by a one-hidden-layer scSNN of $\mathcal{O}(Cm^{15/4})$ spiking neurons, that is,*

$$\|f(\boldsymbol{x}, t) - g(\boldsymbol{x})\|_{L^2(\mu)} < \epsilon .$$

Theorem 2 [Zhang, Gao, &Zhou. Neural Netw'2022]

there exists a constant $\delta > 0$ such that

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}\left(f_R(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \geq \delta,$$

for any one-hidden-layer real-valued network $f_R : \mathbb{R}^{2d} \to \mathbb{R}$ with exponential parameter number $O(C_1(2d+1)e^{C_1(2d)})$. Here, C_1 is an constant independent to d.

ScSNN:	$\mathcal{O}(Cm^{15/4})$	polynomial
ANN:	$\mathcal{O}(C_1(2d+1)e^{C_1(2d)})$	exponential

The parameter complexity of ScSNNs can be **exponentially smaller** than ANNs when approximating radial functions.

提出两个基于 Rademacher Complexity 的脉冲神经网络泛化界

Theorem 12 *Let $\mathcal{F}_{\mathcal{W}}^{one}$ denote the function space of L-layer stochastic neurons. If $u_{reset} = 0$, $\|\mathbf{W}^l\|_2 \le C_l$ for $l \in [L]$, and $\|\mathbf{X}\|_2 \le C_X$ for $\mathbf{X} \in \mathcal{X}$, we have*

$$\mathfrak{R}_n(\mathscr{L} \circ \mathcal{F}_{\mathcal{W}}) \le C_n\, \mathfrak{R}_n(\mathcal{F}_{\mathcal{W}}) \le \frac{(C_n)^L C_X}{\sqrt{n}} \left(\prod_{l\in[L]} C_l\right) (p_{max})^{(L+1)/2}, \tag{42}$$

where $p_{max} = \max_{i\in[n],l\in[L]}\{1 - \max\{p^{(l,i)}, p_\theta\}\} \in (0,1)$ and C_n is a universal constant.

随机发射 SNN 的泛化界（JMLR 2024）

Rademacher complexity of stochastic neurons can be upper bounded by an **exponential** function that consists of the **network depth** and **excitation probability**.

Theorem 1 (Generalization Bound for SNNs) *Let $\mathcal{H}$ denote the function collection of L-layer SNNs with the width of N_w on the time interval $[0,T]$, the hypothesis space $\mathcal{H}$ is universally bounded according to $\|f\|_\infty \le N_f$ for $f \in \mathcal{H}$, $\mathcal{D}$ is the distribution of sample variables $(\mathbf{X}, y)$, S_n indicates the collection of n pairs of training samples $(\mathbf{X}_i, y)$ where $i \in [n]$, and $L_2(S_n)$ denotes the data-dependent L^2 metric space. Assume that the norm value of connection weights is finite, that is, $\|w\| \le M_w$. Let $\hbar : [-N_f, N_f] \times [-N_f, N_f] \to \mathbb{R}$ denote the non-negative loss function, which satisfies that*

i) $\hbar(\cdot,\cdot)$ is upper bounded by M_h, i.e., $\hbar(f, f') \le M_h$ for all $f(\cdot), f'(\cdot) \in [-N_f, N_f]$,

ii) for any fixed $f \in [-N_f, N_f]$, the mapping $y \mapsto \hbar(f, y)$ is L_h-Lipschitz for some $L_h > 0$.

Then with probability at least $1-\delta$ where $\delta \in (0,1)$, we have the generalization bound as follows

$$E(f) \le \hat{E}(f) + 2L_h \hat{\mathfrak{R}}_n(\mathcal{H}) + 3M_h\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $E(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\hbar(f(w,\mathbf{X}),y)]$ denotes the expected error, $\hat{E}(f) = n^{-1}\sum_{i=1}^n \hbar(f(w,\mathbf{X}_i,y_i))$ denotes the empirical error, and the empirical Rademacher complexity $\hat{\mathfrak{R}}_n(\mathcal{H})$ is bounded by

$$\hat{\mathcal{R}}_n(\mathcal{H}) \le \frac{128\, T N_f N_w^{\frac{3}{2}} \log 2}{3\pi n} - 32\sqrt{\frac{2\log 2}{3\pi}}\sqrt{\frac{T N_f^2 N_w^{\frac{3}{2}}}{n}},$$

where $N_f \in \mathcal{O}[(TM_w)^L \exp(-TL)]$.

通用 SNN 泛化界（arXiv 2026）

Empirical Rademacher complexity is **exponential** to the **network depth** and the **maximum time duration** of the received spike sequences, **superlinear and subquadratic** to the **network width**, polynomial to the parameter norm, and inverse linear to the number of training samples.

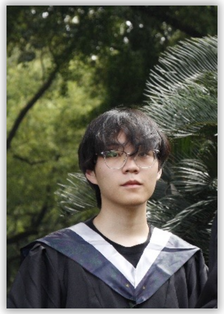
Table 1. Progresses on generalization bounds of SNNs.

Studies	Configurations	Approaches	Generalization Bounds
Maass & Schmitt (1997)	Single neuron with Temporal Coding	VC Dimension	VC $\in \Omega(n \log n)$ with regard to n samples
Schmitt (1999)	Binary and Analog Coding	VC Dimension	VC $\in \Omega(LN_w \log(LN_w))$ where L is the network depth.
Neuman et al. (2025)	Affine Network	Covering Numbers	$N_{nc} \in \mathcal{O}\left(\left(32N_w^{5/2} + 48LN_wM_w^2\right) n \exp\left(3N_w^2\right)\right)$
Zhang et al. (2024)	Stochastic firing	Rademacher Complexity Dropout	$\hat{\mathfrak{R}}_n(\mathcal{H}) \in \mathcal{O}\left(C_w^L\, p_{\max}^{(L+1)/2}\right)$ where $p_{\max} \in [0,1]$ and C_w universally relates to parameter w.
Zhang (2026)	Common-used Expressions in Eq. (1)	Rademacher Complexity Covering Numbers	$\hat{\mathfrak{R}}_n(\mathcal{H}) \in \mathcal{O}\left(n^{-1} M_w^L N_w^{3/2}\, T^{L+1} \exp(-TL)\right)$

合作与文化

Teams & Cooperation

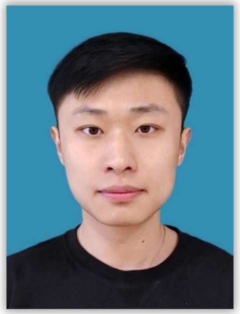











Join us!

• • •

Academic & Cooperation

Microsoft

SIEMENS

中国移动 China Mobile

JS NHC PRC

THE AFFILIATED DRUM TOWER HOSPITAL OF NANJING UNIVERSITY MEDICAL SCHOOL
1892
鼓楼医院

Culture