生成式语言模型的知识嵌入与层级认知推理框架研究*

黄起豹[1]，谭先华[2]
（1. 上饶师范学院　数字技术应用产业学院，江西　上饶　340001；
2. 上饶师范学院　法商学院，江西　上饶　340001）

摘　要：现有生成式语言模型认知理论以统计联结主义为核心，存在知识隐性化、认知表层化、数据与知识割裂化三大缺陷。从认知科学与语言学交叉视角切入，构建“知识嵌入”“层级认知推理”“自适应数据-知识融合”三位一体认知理论重构框架。框架搭建起覆盖句法、语义、语用、认知图式的多维度语言学知识本体，设计“句法解析-语义推理-语用适配”三级认知推理架构，同时引入注意力机制驱动的自适应数据-知识融合模型。基于CoG-LM-2025评测数据集开展多模型对比实验与泛化性测试，结果显示，重构框架大幅提升模型在深层语义推理、复杂空间推理、隐含语用意义挖掘等高阶任务中的性能，认知可解释性评分（满分10分）较基线模型提升2.6分，平均泛化衰减率降至8.3%。研究证实，认知架构的系统性优化为提升生成式模型的结构化推理与可解释性提供了可行路径。

关键词：生成式语言模型；认知理论；知识本体；层级推理；可解释性
中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.19358/j.issn.2097-1788.2026.07.005

中文引用格式：黄起豹，谭先华. 生成式语言模型的知识嵌入与层级认知推理框架研究[J]. 网络安全与数据治理，2026，45(7)：32-38.
英文引用格式：Huang Qibao，Tan Xianhua. Research on knowledge embedding and hierarchical cognitive reasoning framework of generative language model[J]. Cyber Security and Data Governance，2026，45(7)：32-38.

Research on knowledge embedding and hierarchical cognitive reasoning framework of generative language model

Huang Qibao[1]，Tan Xianhua[2]
（1. College of Digital Technology Application Industries，Shangrao Normal University，Shangrao 340001，China；
2. School of Law and Business，Shangrao Normal University，Shangrao 340001，China）

Abstract：Existing cognitive theories of Generative Language Models（GLMs）are primarily based on statistical connectionism，suffering from three major limitations：implicit knowledge representation，superficial cognitive processing，and disconnection between data and knowledge. From an interdisciplinary perspective of cognitive science and linguistics，this paper proposes a three-in-one cognitive theory reconstruction framework consisting of "knowledge embedding，hierarchical reasoning，and dynamic fusion". The framework constructs a multi-dimensional linguistic knowledge ontology covering syntax，semantics，pragmatics，and cognitive schemas，designs a three-tier cognitive reasoning architecture of "syntactic parsing-semantic inference-pragmatic adaptation"，and introduces an attention-driven adaptive data-knowledge fusion model. Multi-model comparison experiments，and generalization tests based on the CoG-LM-2025 evaluation dataset demonstrate that the reconstructed framework significantly improves model performance in high-order tasks such as deep semantic inference，complex spatial reasoning，and implicit pragmatic meaning extraction. The cognitive interpretability score increases by 2.6 points compared to baseline models，while the average generallization attenuation rate has dropped to 8.3%. The study confirms that systematic optimization of cognitive architecture provides a viable approach to enhance the structured reasoning and interpretability of generative models.

Key words：generative language models；cognitive theory；knowledge ontology；hierarchical reasoning；interpretability

*基金项目：中国成人教育协会人工智能研究项目基金（AI-Y2025028S）；中国索引学会人工智能研究项目基金（CSI25C01）；全国高等学校计算机教育研究会（CERACU2026RO7）；中国机械工业教育协会（ZJJX25SY001）

0 引言

GPT-4、Claude3、Qwen-2 等生成式语言模型（Generative Language Models，GLMs）依托大规模预训练与自监督学习，在文本生成、翻译等表层自然语言处理（Natural Language Processing，NLP）任务中性能逼近甚至超越人类，推动 NLP 迈入“数据驱动”的规模化发展阶段[1]。这类模型的核心优势在于无需人工编码语言学规则，通过学习语料统计特征构建语义表征，实现语言能力的泛化。但在深层语义理解、复杂逻辑推理等高阶任务中，其认知局限性日益凸显，语义偏差、常识错误等问题频发[2]。以 GPT-4 为例，其简单方位判断准确率可达 92%，但面对双向认知推理表达时，准确率却骤降至 38.7%。其根源就是模型仅能捕捉表层语言特征，缺乏人类式的深层认知逻辑。

现有 GLMs 认知理论以联结主义为核心，将语言认知简化为线性过程，依托分布式表征与梯度下降优化实现模型训练，核心缺陷是割裂了语言学知识与认知推理的内在关联，忽视了人类语言认知的层级性特征[3]。人类语言能力是先天认知结构与后天经验协同作用的结果，语言学知识为语言理解提供核心框架[4]。而 GLMs 将知识隐含于海量参数中，无法通过显性规则形成约束，认知过程缺乏可解释性；认知建模停留在统计关联层面，未还原人类分层的认知过程，极易陷入统计偏见与泛化失效的困境，难以从根本上缓解模型过度依赖表面统计特征而导致的泛化偏差问题。

现有研究多聚焦单一维度的知识融入或孤立优化推理模块，尚未形成“知识引导-认知建模-数据支撑”的协同体系。为明确本文框架的定位，本研究从知识表征、推理机制、融合策略与可解释性来源四个维度，与当前主流认知增强范式进行对比分析。神经符号模型（NSAI）将形式化逻辑规则与神经网络结合，虽具备强可解释性，但硬规则约束导致其在开放域语言任务中泛化能力受限，且符号-神经接口存在梯度不连续问题；检索增强与知识注入模型（如 RAG、K-BERT 等）侧重静态知识覆盖与上下文拼接，缺乏对句法-语义-语用层级认知过程的显式建模，易出现知识碎片化与上下文推理脱节。与之不同，本文框架不依赖硬逻辑约束或离线检索，而是构建可计算的语言学知识本体，通过三级递进推理模拟人类认知加工流，并利用注意力机制实现数据特征与知识规则的动态权重分配。该路径在保持数据驱动泛化优势的同时，引入语言学先验约束，为 GLMs 的深层语义推理与可解释性提供结构化支撑[5-6]。研究的创新点主要体现在三方面：一是构建多维度语言学知识本体，实现知识的显性化与可计算化；二是设计三级认知架构，模拟人类分层的语言认知过程；三是提出注意力机制驱动的自适应融合模型，动态调节数据与知识的权重占比。本研究将通过多维度评测数据集与系列对比实验，验证该框架对模型认知深度、可解释性及泛化能力的提升效果，为 GLMs 的可持续发展提供理论支撑与实践参考。

1 生成式语言模型认知理论的现状

当前 GLMs 的认知理论以“统计联结主义”为核心范式，理论体系建立在分布式表征、自监督学习与梯度下降优化三大基础之上，形成“数据输入-特征提取-参数更新-输出生成”的闭环认知逻辑。分布式表征假设认为，词语、短语、句子等语言单位的语义信息，可通过高维向量空间中的点来表征，向量的每个维度对应语言的潜在特征，模型通过学习语料中的统计规律，将语言知识编码为向量间的关联关系[7]；自监督学习机制通过掩码语言模型、下一句预测等伪任务设计，让模型从无标注语料中自动学习语言特征，无需人工干预即可完成知识积累；梯度下降优化则通过最小化预测误差，持续调整模型参数，让模型生成结果不断逼近语料中的统计分布。

该范式的核心数学表达为条件概率生成公式。给定输入序列 $X=x_1, x_2, \cdots, x_n$，模型生成输出序列 $Y=y_1, y_2, \cdots, y_m$ 的概率为：

$$P(Y \mid X;\theta)=\prod_{i=1}^{m} P(y_i \mid X, y_1, y_2, \cdots, y_{i-1};\theta) \quad (1)$$

其中，θ 为模型的所有可训练参数，$P(y_i \mid X, y_1, \cdots, y_{i-1};\theta)$ 表示在给定输入序列与已生成序列的前提下，生成第 i 个 token 的条件概率。模型的训练目标是通过最大化语料中所有序列对的条件概率，优化参数 θ，使模型习得语言的统计规律[8]。这一条件概率公式也是 Transformer 架构生成逻辑的核心数学表达，多头注意力机制通过捕捉序列内部的依赖关系，进一步优化公式中条件概率的计算精度，强化模型对长距离统计特征的建模能力。

统计联结主义范式的优势在于泛化能力与数据适应性极强，无需人工编码复杂的语言学规则，仅通过扩大语料规模与模型参数规模，就能实现对各类表层 NLP 任务的覆盖[9]。但这种“重数据轻知识”的理论逻辑，让模型难以突破知识隐性化、认知表层化、数据与知识割裂化三大核心缺陷，认知深度不足、可解释性差等问题突出，无法适配深层语义推理、复杂逻

辑分析等高阶认知任务的需求。

2 生成式语言模型认知理论的重构框架

2.1 整体架构设计

本文提出的 GLMs 认知理论重构框架整体分为语言学知识本体、层级认知推理、自适应数据-知识融合三大模块（图 1），各模块相互关联、协同作用，形成完整的认知闭环。语言学知识本体模块负责构建多维度、可计算的语言学知识体系，为认知推理提供规则输入；层级认知推理模块基于知识本体，通过三级认知过程实现深层语义理解与推理，输出认知结果；自适应数据-知识融合模块负责动态调节数据特征与知识规则的权重，将认知结果与数据统计特征融合，优化模型生成输出，同时通过反馈机制调整知识本体与认知参数。

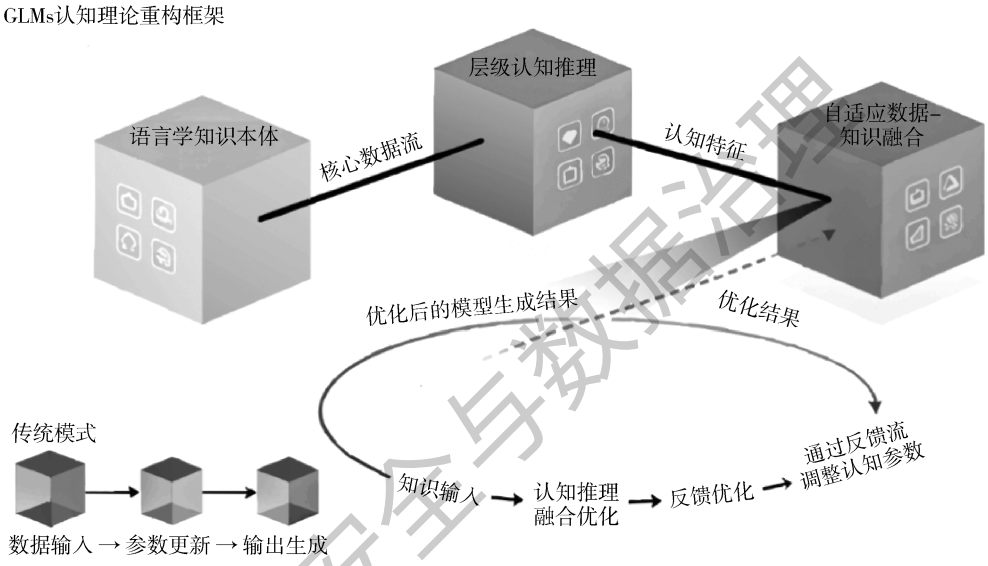

图 1 生成式语言模型认知理论重构框架

具体的反馈机制为：计算生成结果与认知目标的偏差，如语义一致性误差、语用适配度得分，当偏差超过 0. 15（基于验证集 10 万条样本的偏差分布统计，该阈值能平衡反馈效率与认知稳定性）时，触发反馈流程——针对知识本体模块，补充高频错误对应的语言学规则，如稀有句式规则；针对层级认知推理模块，微调语义推理层ω_s、语用适配层ω_p各层级参数的初始权重，实现认知能力的迭代提升。

这一框架打破了传统模型“数据输入-参数更新-输出生成”的线性逻辑，构建数据驱动与知识引导相协同的认知计算路径，提升模型在复杂语境下的结构化推理能力。其核心优势在于系统性解决现有理论的三大缺陷：通过知识本体模块实现语言学知识的显性化与体系化，通过层级认知模块构建分层推理机制，通过自适应融合模块实现数据与知识的动态协同，三者协同让模型具备可解释、可泛化的深层认知能力。

2.2 语言学知识本体模块的构建

语言学知识本体模块是重构框架的基础，其核心目标是搭建覆盖句法、语义、语用、认知图式四大维度的多维度、可计算、可扩展语言学知识体系，为层级认知推理提供明确的规则约束。各维度知识的具体内容与编码方式如下：

（1）句法知识维度：基于上下文无关文法（Context-Free Grammar，CFG）与依存句法理论，构建包含短语结构规则、依存关系规则、句式转换规则的句法规则体系，同时融入句法错误检测规则，采用产生式规则编码为可计算形式，为句法解析层提供支撑。

（2）语义知识维度：整合语义角色标注（Semantic Role Labeling，SRL）、语义范畴分类与语义关系体系，构建包含语义角色、语义范畴、语义关系的语义知识网络，采用向量空间模型（Vector Space Model，VSM）将语义知识编码为语义向量，实现与模型特征的适配。

（3）语用知识维度：基于言语行为理论与语境适配理论，构建包含言语行为类型、语体特征、语境适配规则的语用规则体系。该体系以规则树形式进行编码，为语用适配层提供依据。

（4）认知图式维度：借鉴认知语言学的概念图式理论与 STEP 标注体系，构建空间、时间、事件等核心认知图式，其中空间图式包含实体、空间信息、事

件、时间四大要素，时间图式包含时点、时段、时序等特征，事件图式包含事件的起因、经过、结果等要素，均采用图结构编码为认知图式向量。

2.3 层级认知推理模块的构建

语义推理层基于句法解析结果与知识本体中的语义知识，实现深层语义理解与推理，核心完成语义角色标注、语义关系提取、语义一致性判断与多步语义推理四大任务。具体流程如下：首先结合句法结构与语义规则，为句子词语分配施事、受事、处所、工具等语义角色，构建语义角色框架；然后基于语义知识网络，提取实体间的同义、反义、上下位、因果等语义关系；接着结合认知图式知识，校验句子内部及句子间的语义一致性，排除“他在室内却被雨淋了”这类语义矛盾；最后针对空间推理、逻辑推理等复杂任务，基于语义角色框架与认知图式开展多步语义推理，生成语义推理结果。

该层级的核心数学模型为语义推理函数，表达式为：

$$R_s = f_s(S_p, K_s, G_c; \omega_s) \quad (2)$$

其中，R_s 为语义推理结果，S_p 为句法解析层输出的句法特征，K_s 为知识本体中的语义知识与认知图式，G_c 为语义一致性校验规则，ω_s 为语义推理层的可训练参数，通过模型训练优化参数，提升推理准确性。

语用适配层基于语义推理结果与知识本体中的语用知识，实现语境适配与语用意义生成，核心完成言语行为识别、语体特征适配、语境意义补充与语用一致性校验四大任务。具体流程为：首先结合语义推理结果与语用规则，识别陈述、请求、疑问、感叹等输入文本的言语行为类型；再根据正式/非正式、书面/口语、公开/私密等语境信息，适配对应的语体特征，调整表达形式；接着基于语境补充隐含语用意义，比如“能帮我递一下杯子吗”的核心是“请求对方递杯子”，而非单纯的疑问；最后校验语用表达的一致性，确保输出结果符合语境需求与语用规则。例如正式学术场景中，将口语化的“我觉得”调整为“本研究认为”；请求场景中，补充礼貌用语，让语用表达更恰当。

3 评测数据集构建与实验设计

3.1 评测任务体系设计

为全面验证重构理论的有效性，本研究设计覆盖 6 大维度、5 项子任务的认知能力评测体系，突破现有单一任务评测的局限，从表层能力到深层认知、从常规场景到泛化场景，全方位评估模型的认知性能。评测体系遵循“针对性、全面性、层次性”原则，针对性覆盖重构框架的知识整合、层级推理、动态融合三大核心优势，全面涵盖 GLMs 的认知能力维度，层次化区分表层任务与深层任务。具体任务设计如表 1 所示。

表 1 生成式语言模型认知能力评测任务体系

评测维度	子任务类型	核心任务目标	知识依赖程度	认知层级要求	场景类型
句法认知能力	句式转换（主动/被动、把/被字句）	完成不同句式转换，保持语义一致性	中	句法解析层+语义推理层	常规场景
语义认知能力	深层语义推理（多实体关系）	基于多实体语义关系，完成多步推理	高	语义推理层	常规+复杂场景
认知一致性能力	文本内部认知一致性校验	检测文本内部句法、语义、语用的一致性矛盾	中高	三级认知层+校验机制	复杂场景
泛化认知能力	跨领域认知迁移	将认知能力迁移至未训练领域，验证迁移能力	中高	三级认知层+融合模块	跨领域场景
认知可解释能力	认知过程解释生成	生成认知推理过程的解释，验证可解释性	高	三级认知层+知识本体	全场景

上述 5 项子任务分别对应重构框架的核心能力维度，其中深层语义推理、文本内部认知一致性校验等任务重点验证层级认知推理模块与知识本体模块的有效性，跨领域认知迁移任务重点验证自适应数据-知识融合模块的泛化能力，认知可解释任务重点验证本研究理论的可解释性优势。

3.2 实验方案设计

本研究设计多模型对比实验与泛化性测试两类实验方案，全面验证重构框架的有效性及模型的泛化能力。

多模型对比实验旨在对比 Model-R 与 4 类基线模型在 5 项子任务上的性能差异，验证重构理论对 GLMs 认知能力的整体提升效应。实验流程为：将 CoG-LM-2025 数据集按比例划分为训练集、验证集与测试集，所有模型在相同训练集上微调，在验证集上优化超参数，最终在测试集上开展性能评估，统计各模型在不

同任务维度的核心指标，通过差异显著性检验（$p<0.05$）判断 Model-R 的性能优势是否显著。

泛化性测试分为分布外数据测试与跨领域迁移测试两部分，验证 Model-R 在非训练分布场景与陌生领域中的认知能力[10]。分布外数据测试采用 CoG-LM-2025 数据集的分布外测试子集（3 200 万字稀有语言现象语料），评估模型对异形同义表达、稀有句式、特殊语用场景的处理能力；训练集涵盖新闻、小说、学术论文等 10 个通用领域，不含医疗、金融、古籍文献，保证跨领域测试的客观性。

泛化性测试通过计算性能衰减率衡量模型泛化能力，计算公式为：衰减率=（常规测试集任务 F1 值−分布外/跨领域测试集任务 F1 值）/常规测试集任务 F1 值×100%。其中常规测试集核心任务 F1 值：Model-R 为 88. 3%，GPT-4 为 85. 1%，其余模型对应各自核心任务基准值。

3. 3 Model-R 架构实现与训练配置

Model-R 以 Qwen-2. 5-7B-Instruct 为基座模型，采用参数高效微调策略。主干网络参数保持冻结，仅在句法解析层、语义推理层与语用适配层注入低秩自适应（LoRA）模块，秩 $r=16$，缩放系数 $\alpha=32$，目标模块覆盖 q，k，v，o 投影层及部分前馈网络，可训练参数量约为 28. 6M（占总参数 0. 39%）。模型训练目标函数由四部分构成：

$$\mathcal{L}=\mathcal{L}_{\mathrm{CE}}+\lambda_1\mathcal{L}_{\mathrm{sem}}+\lambda_2\mathcal{L}_{\mathrm{prag}}+\lambda_3\mathcal{L}_{\mathrm{align}} \tag{3}$$

其中，$\mathcal{L}_{\mathrm{CE}}$ 为自回归交叉熵损失，保障基础语言建模能力；$\mathcal{L}_{\mathrm{sem}}=\frac{1}{N}\sum_{i=1}^{N}\|h_i^{\mathrm{sem}}-\mathrm{SRL}(x_i)\|^2$ 为语义一致性约束损失，通过知识本体中的语义角色标注对隐层表征进行正则化；$\mathcal{L}_{\mathrm{prag}}$ 为语用适配分类损失，对齐言语行为类型与语体特征标签；$\mathcal{L}_{\mathrm{align}}$ 为融合模块的注意力稀疏正则项，防止数据与知识权重过度偏向单一侧。经网格搜索，超参数确定为 $\lambda_1=0.4$，$\lambda_2=0.3$，$\lambda_3=0.1$。优化器选用 AdamW，初始学习率设为 2×10^{-5}，采用余弦衰减策略（warmup 比例 0. 05），训练批量大小为 64，训练轮数 3。实验在 8×A100−80G 集群上完成，单轮耗时约 4. 2 h。

3. 4 核心评估指标定义与测量协议

为保证实验结果的可复现性与评估严谨性，本文对核心指标制定标准化测量协议：

（1）认知可解释性评分：采用四维评估量表（推理可追溯性、逻辑连贯性、语言学规则匹配度、自我一致性修正），每项采用 1~5 分 Likert 量表。由 3 名计算语言学专家与 2 名 NLP 研究者独立打分，评分者间一致性经 Cohen′s Kappa 检验为 0. 83（$p<0.001$）。最终得分为五项专家打分的算术平均，满分 10 分。

（2）推理步数适配率：针对复杂空间/逻辑推理任务，以专家标注的“最优推理步数”为金标准。若模型生成推理链的步数与金标准偏差≤1，则判定为适配。评估采用自动步数解析脚本（基于依存句法树切分）完成初筛，并随机抽取 20%样本进行人工复核，自动−人工一致率达 94. 6%。

（3）语用一致性指标：基于 CoG-LM-2025 语用标注子集（含言语行为类型、语体风格、隐含意图三类标签），由 3 名标注员依据《语用标注指南 v2. 1》独立标注，分歧项由资深语言学教授仲裁。指标采用 Macro-F1 计算，涵盖精确率与召回率的均衡表现。

4 实验结果与分析

4. 1 多模型对比实验结果与分析

多模型对比实验结果显示，Model-R 在所有评测维度上均显著优于 GPT-4、Claude 3 Opus、Qwen-2-72B 与 BERT-3B 四类基线模型，充分验证了认知理论重构对提升生成式语言模型深层认知能力的有效性。

在句法认知能力上，依托语言学知识本体中句法规则的显性引导，Model-R 在句法错误检测与修正、句式转换任务中，分别实现 89. 2%的准确率与 90. 5%的 F1 值，较最优基线模型 GPT-4 提升超 4 个百分点。语义认知能力的提升更为显著，Model-R 在深层语义推理、异形同义判别及复杂空间语义推理任务中的 F1 值分别达 86. 7%、88. 3% 和 84. 5%，较 GPT-4 高出 6. 2%~8. 7%，其中空间推理任务的推理步数适配率提升至 82. 1%，增幅达 9. 3 个百分点。这说明层级认知推理模块有效模拟了人类“句法解析−语义推理”的分层过程，结合认知图式知识，克服了传统模型依赖表层特征导致的推理偏差。

在语用认知方面，通过语用知识本体与适配层的协同作用，Model-R 在语用一致性判断与隐含语用挖掘任务上，分别取得 87. 6% 与 85. 2% 的 F1 值，较 GPT-4 提升 5. 1 个与 6. 4 个百分点。在认知一致性层面，Model-R 在文本内及多文本一致性建模任务中，F1 值分别达 86. 3%与 83. 8%，显著优于基线模型，印证了认知一致性校验机制的有效性（如表 2）。

4. 2 泛化性测试结果与分析

表 3 泛化性测试结果显示，Model-R 在分布外数据测试与跨领域迁移测试中均展现出优异的泛化能力，性能衰减率显著低于所有基线模型，验证了自适应数

表 2　多模型对比实验核心指标统计

模型类型	句法错误检测准确率/%	复杂空间推理 F1 值/%	隐含语用挖掘 F1 值/%	分布外数据衰减率/%	多文本一致性 F1 值/%	可解释性评分	整体平均 F1 值/%
GPT-4	84.9	75.8	78.8	23.5	75.7	5.2	78.8
Claude 3 Opus	83.7	74.2	79.4	25.1	74.9	5.5	78.1
Qwen-2-72B	81.2	70.3	76.5	28.7	72.4	4.8	75.1
BERT-3B	75.6	62.8	68.3	35.2	65.9	4.2	68.2
Model-R	89.2	84.5	85.2	9.8	83.8	7.8	85.7

注：整体平均 F1 值为句法错误检测准确率、复杂空间推理 F1 值、隐含语用挖掘 F1 值、多文本一致性 F1 值的算术平均；可解释性评分为独立指标（满分 10 分）；分布外数据衰减率为反向指标（单独列示）

据-知识融合模块与语言学知识本体模块的协同效应。本次测试覆盖医疗、金融等现代专业领域与古籍文献这一传统领域，全面验证模型在不同语言体系中的泛化能力。

表 3　泛化性测试结果统计　（%）

模型类型	分布外异形同义 F1 值（衰减率）	跨领域医疗语义 F1 值（衰减率）	跨领域古籍句法准确率（衰减率）	平均泛化衰减率
GPT-4	67.0（21.3）	72.5（18.8）	59.8（30.7）	23.6
Claude 3 Opus	66.5（23.8）	71.8（20.9）	58.3（32.6）	25.8
Qwen-2-72B	63.2（26.5）	68.4（23.7）	56.1（33.4）	27.9
BERT-3B	55.7（33.2）	60.2（29.9）	50.3（33.6）	32.2
Model-R	82.7（6.4）	80.2（7.5）	79.3（11.1）	8.3

注：平均泛化衰减率为三项任务衰减率的算术平均

分布外数据测试中，Model-R 在异形同义判别任务上的 F1 值达 82.7%，性能衰减率仅 6.4%；跨领域医疗语义任务 F1 值达 80.2%，衰减率 7.5%；跨领域古籍句法准确率达 79.3%，衰减率 11.1%。而 GPT-4 在上述任务中的衰减率分别为 21.3%、18.8% 与 30.7%。这说明 Model-R 能通过动态提高知识权重，依托语言学知识规则处理稀有语言现象，有效弥补数据覆盖不足的缺陷。

5　结论

本研究从认知科学与语言学交叉视角，针对现有 GLMs 认知理论“以统计模式拟合替代结构化推理”的局限，提出“知识嵌入”“层级认知推理”“自适应数据-知识融合”三位一体的认知理论重构框架。通过构建语言学知识本体、层级认知推理架构与自适应数据-知识融合模型，克服传统理论知识隐性化、认知表层化、数据-知识割裂化缺陷。结合 CoG-LM-2025 数据集实验，验证了重构理论的有效性，主要结论如下：

（1）语言学知识本体的显性化、体系化构建是提升 GLMs 认知深度的基础。本研究搭建多维度知识本体，以“规则编码+向量表征”实现知识可计算与可整合，为认知推理提供规则支撑。实验表明，该模块使 Model-R 在相关任务上性能提升超 8 个百分点（如在复杂空间推理任务上，性能较最优基线模型 GPT-4 提升达 8.7 个百分点），提升认知可解释性，证明显性知识建模能打破传统局限，为深层认知提供保障。

（2）层级认知推理架构可模拟人类分层认知，突破 GLMs 深层推理瓶颈。“句法解析-语义推理-语用适配”三级架构与认知一致性校验机制，实现递进式认知，解决传统推理偏差问题。实验中，Model-R 在高阶任务上性能提升显著，整体平均 F1 值提升达 6.9 个百分点，认知一致性指标较基线模型高出 8 个百分点以上，印证了层级化认知机制科学有效。

参考文献

[1] 王伟正,陈晗睿,丁晓燕,等. 大语言模型生成摘要可靠性评测研究[J]. 情报杂志,2026,45(1):153-160,127.

[2] 詹卫东,孙春晖,肖力铭. 语言学知识驱动的空间语义理解能力评测数据集研究[J]. 语言战略研究,2024,9(5):7-21.

[3] Liu Xiaoxiao, Cui Ying. Eye tracking technology for examining cognitive processes in education: a systematic review[J]. Computers & Education,2025,229:105263.

[4] Qiu Houjie, Ma Xingkong, Liu Bo, et al. Psycholinguistic knowledge-guided graph network for personality detection of silent users[J]. Information Processing and Management,2025,62(3):104064.

[5] Yu Danni,Li Luyang,Su Hang, et al. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis[J]. International Journal of Corpus Linguistics,2024,29(4):534-561.

[6] Zhu Changbo, Zhou Ke, Tang Fengzhen, et al. A hierarchical Bayesian inference model for volatile multivariate exponentially distributed signals[J]. Frontiers in Computational Neuroscience,2025,19:1408836.

[7] Cohen R, Biran E, Yoran O, et al. Evaluating the ripple effects of knowledge editing in language models[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 283-298.
[8] 朱旭东, 李肖艳. “Wisdom”学位论文写作教学模式建构：构成要素、理论基础与实践策略[J]. 学位与研究生教育, 2024(7): 1-8.
[9] Nguyen H D T, Le B. FRIC: a framework for feature importance explanations and evaluation in span-level NLP[J]. Neural Computing and Applications, 2025, 37(35-36): 28989-29020.
[10] 杨凯钦, 蔡满春, 彭舒凡. 模型-数据协同驱动的泛化性深度伪造检测方法[J/OL]. 计算机工程与应用, 1-17[2026-04-25]. https://link. cnki. net/urlid/11. 2127. tp. 20260421. 1845. 007.

（收稿日期：2026-04-26）

作者简介：

黄起豹（1978-），男，博士，副教授，主要研究方向：人工智能与数智融合。

谭先华（1978-），女，博士，副教授，主要研究方向：模型治理与大数据。

“人工智能安全”主题专栏征稿启事

随着人工智能（AI）技术，尤其是以大语言模型、多模态生成模型为代表的生成式人工智能的飞速发展，人类社会正加速迈入智能化新阶段。然而，人工智能在赋能千行百业的同时，也暴露出诸多深层次的安全风险：数据投毒与隐私泄露、模型对抗攻击与鲁棒性不足、算法偏见与伦理失范、生成内容滥用与虚假信息传播等。如何构建可信、可靠、可控的人工智能安全体系，已成为学术界、产业界共同关注的重大课题。

为集中展示该领域的最新研究成果，推动人工智能安全技术的交叉融合与创新发展，《网络安全与数据治理》拟在2026年第11期出版“人工智能安全”主题专栏，现诚挚邀请相关领域的专家学者、科研人员踊跃投稿！

一、征文主题：人工智能安全

包括但不限于以下学术方向：

1. 大模型安全与对齐：

大语言模型（LLM）的越狱攻击与防御

模型价值观对齐与伦理约束

提示词注入攻击与防护技术

2. 对抗性机器学习：

对抗样本生成、攻击与鲁棒性增强

后门攻击与触发器检测

模型逆向攻击与成员推理防御

3. 数据安全与隐私保护：

联邦学习中的隐私攻击与防御

差分隐私在AI训练中的应用

数据脱敏、水印技术与版权保护

4. 可信人工智能与可解释性：

模型决策的可解释性方法与评估

公平性算法与偏见检测/缓解技术

5. AI驱动的系统安全：

人工智能在网络安全（入侵检测、恶意软件分析）中的应用

AI自身系统（框架、供应链）的安全漏洞分析

6. 深度合成与内容安全：

深度伪造（Deepfake）检测技术

AI生成内容的鉴别与溯源

二、投稿要求

1. 稿件请用word格式录入，并套用本刊投稿模板。模板下载网址：http://files. chinaaet. com/files/Periodical/pcachina_Templates. doc

2. 投稿文章不存在一稿多投现象。

3. 文章无抄袭、剽窃、侵权、虚假引用等不良学术行为，且不违反相关法律法规，不涉及国家、企业秘密，稿件文责自负。

4. 论文要求观点鲜明、逻辑严谨、论据充分、方法合理，字数在5000~8000字。

5. 请在官方投稿网站（http://www. pcachina. com）注册、投稿。注册后请投稿在“主题专栏”栏目。“稿件标题”请填“人工智能+文章题目”。稿件经评审合格录用后，在《网络安全与数据治理》2026年第11期（正刊）以主题专栏形式发表。

三、时间安排

截稿日期：2026年9月20日

审稿反馈日期：2026年10月10日

出版日期：2026年11月15日

四、联系方式

联系人：牟老师

电　话：010-82306116

邮　箱：muyx@ chinaaet. com

《网络安全与数据治理》编辑部

2026年6月

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部