



中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于神经符号计算的知识数据融合学习框架研究

作者姓名：孙严顺

指导教师：王石

正高级工程师

中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机科学与技术

培养单位：中国科学院计算技术研究所

2025 年 6 月

Research on Neural-Symbolic Computing-Based Framework for
Knowledge and Data Fusion Learning

A thesis submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Master of Engineering
in Computer Science and Technology

By

SUN Yanshun

Supervisor: Professor WANG Shi

Institute of Computing Technology, Chinese Academy of Sciences

June, 2025

中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

日 期：

导师签名：

日 期：

摘要

本文研究了基于神经符号计算的知识数据融合学习框架，旨在解决当前人工智能系统中的可解释性、泛化能力和推理能力等关键挑战。随着人工智能技术的快速发展，如何结合符号推理的可解释性与神经网络的学习能力，已成为实现下一代人工智能的重要课题。神经符号计算作为连接两种主要人工智能范式的桥梁，通过整合符号规则与深度学习模型，为构建更加智能、可信的系统提供了新的理论和技术框架。

本文首先系统性分析了符号计算与神经网络各自的优势与局限，指出传统符号系统虽具备强大的逻辑推理能力和良好的可解释性，但难以处理现实世界的模糊性和不确定性；而神经网络虽擅长从数据中学习模式，但其“黑盒”特性使得决策过程缺乏透明性，限制了在关键领域的应用。针对这些挑战，本文提出了三种层次的知识数据融合解决方案，分别面向深度神经网络、Transformer 模型和大规模预训练模型。

面向 DNN 的规则知识融合框架。在面向深度神经网络的框架中，本文创新性地提出了双向知识蒸馏与伪标签迭代训练相结合的方法。通过引入规则知识指导模型训练，系统在低资源场景下显著提升了模型的可解释性与泛化能力。实验表明，该方法在命名实体识别和情感分类任务上取得了明显的性能提升，特别是在训练数据有限的情况下，规则知识的引入有效弥补了数据不足带来的泛化问题。

面向 Transformer 的规则知识融合框架。针对 Transformer 模型，本文设计了基于注意力掩码机制的规则知识融合方法，通过将规则逻辑嵌入多头自注意力机制中，实现了全局语义与局部规则的高效结合。实验结果显示，该方法在命名实体识别和文本摘要任务上带来了显著的性能提升，不仅提高了模型的准确率，还增强了其对长依赖关系的建模能力，使生成的文本更加符合语法规则和逻辑一致性。

面向预训练模型的规则知识融合框架。对于大规模预训练模型，本文提出了结合提示模板与规则知识蒸馏技术的框架，充分利用预训练模型的生成能力，进一步提升了复杂推理任务的表现。通过在数学推理、常识推理和通用推理任务上的实验验证，该方法展示了神经符号融合对预训练模型推理能力的显著增强作用，尤其在需要多步推理和逻辑约束的任务中表现突出。

综合实验结果表明，所提出的三种层次框架在不同任务和模型架构上均取得了显著的性能提升，验证了神经符号计算在知识数据融合中的有效性。本文不仅为神经符号计算的实际应用提供了系统性解决方案，还为构建可信、智能的知

识融合系统提供了理论与实践支持。

关键词：神经符号计算，规则知识融合，深度神经网络，Transformer，预训练模型

Abstract

This paper investigates a neural-symbolic computation-based framework for knowledge data fusion learning, aiming to address key challenges in current artificial intelligence systems, including explainability, generalization ability, and reasoning capability. With the rapid development of artificial intelligence technologies, combining the explainability of symbolic reasoning with the learning ability of neural networks has become an important topic for achieving next-generation artificial intelligence. Neural-symbolic computation serves as a bridge connecting the two main artificial intelligence paradigms, providing a new theoretical and technical framework for building more intelligent and trustworthy systems by integrating symbolic rules with deep learning models.

This paper first systematically analyzes the advantages and limitations of symbolic computation and neural networks, noting that traditional symbolic systems possess powerful logical reasoning abilities and good explainability but struggle to handle real-world ambiguity and uncertainty. Neural networks excel at learning patterns from data, but their "black box" nature makes decision processes lack transparency, limiting applications in critical domains. To address these challenges, this paper proposes three hierarchical knowledge data fusion solutions targeting deep neural networks, Transformer models, and large-scale pre-trained models.

A rule knowledge fusion framework for DNN. For deep neural networks, the paper innovatively proposes a method combining bidirectional knowledge distillation with pseudo-label iterative training. By introducing rule knowledge to guide model training, the system significantly enhances model explainability and generalization ability in low-resource scenarios. Experiments show that this method achieves significant performance improvements in tasks such as named entity recognition and sentiment classification, particularly when training data is limited, where the introduction of rule knowledge effectively compensates for generalization issues caused by data insufficiency.

A rule knowledge fusion framework for Transformer. For Transformer models, the paper designs a rule knowledge fusion method based on attention mask mechanisms, achieving efficient combination of global semantics and local rules by embedding rule logic into multi-head self-attention mechanisms. Experimental results show that this method brings significant performance improvements in tasks such as named entity recognition and text summarization, not only improving model accuracy but also enhancing its ability to model long-distance dependencies, resulting in generated text that

better adheres to grammatical rules and logical consistency.

A rule knowledge fusion framework for Pre-trained models. For large-scale pre-trained models, the paper proposes a framework combining prompt templates with rule knowledge distillation techniques, fully leveraging the generative capabilities of pre-trained models to further enhance performance on complex reasoning tasks. Through experimental validation on mathematical reasoning, commonsense reasoning, and general reasoning tasks, this method demonstrates the significant enhancement of neural-symbolic fusion on the reasoning abilities of pre-trained models, particularly excelling in tasks requiring multi-step reasoning and logical constraints.

Comprehensive experimental results demonstrate that the three proposed hierarchical frameworks achieve significant performance improvements across different tasks and model architectures, validating the effectiveness of neural-symbolic computation in knowledge data fusion. This paper not only provides systematic solutions for practical applications of neural-symbolic computation but also offers theoretical and practical support for building trustworthy, intelligent knowledge fusion systems.

Key Words: Neuro-symbolic Computing, Rule-knowledge Fusion, Deep Neural Networks, Transformers, Pre-trained Models

目 录

第 1 章 绪论	1
1.1 研究背景	1
1.2 研究意义	2
1.3 研究问题	4
1.4 研究范围和目标	5
1.4.1 主要研究内容	5
1.4.2 预期目标	6
1.5 本文组织结构	6
第 2 章 相关工作与研究现状	9
2.1 研究基础	9
2.1.1 神经符号集成方式	9
2.1.2 神经符号计算应用	18
2.1.3 知识蒸馏	19
2.1.4 神经符号计算发展趋势	21
2.2 小结	22
第 3 章 面向 DNN 的规则知识融合框架	23
3.1 概述	23
3.2 方法介绍	23
3.2.1 双向知识蒸馏	23
3.2.2 伪标签迭代训练	24
3.3 实验	26
3.3.1 数据集	26
3.3.2 实验结果	27
3.4 小结	33
第 4 章 面向 Transformer 的规则知识融合框架	35
4.1 概述	35
4.2 方法介绍	35
4.2.1 注意力掩码生成	35
4.3 实验结果	37
4.3.1 命名实体识别任务	37
4.3.2 文本摘要任务	40
4.4 小结	45

第 5 章 基于规则知识的提示词生成与优化	47
5.1 概述	47
5.2 方法介绍	47
5.2.1 规则知识的构建	48
5.2.2 提示模板的生成	49
5.2.3 基于规则的提示优化算法	50
5.3 实验结果	51
5.3.1 数学推理任务	51
5.3.2 常识推理任务	54
5.3.3 通用推理任务	57
5.4 小结	59
第 6 章 总结与展望	61
6.1 研究结果	61
6.2 展望	62
6.3 小结	63
参考文献	65
致谢	71
作者简历及攻读学位期间发表的学术论文与其他相关学术成果 ..	73

图目录

图 1-1 研究思路	5
图 1-2 本文组织架构	7
图 2-1 类型 1: 符号神经符号系统	9
图 2-2 GPT-3 工作流程	10
图 2-3 符号的向量化编码	11
图 2-4 深度特征重构	12
图 2-5 解码	12
图 2-6 类型 2: 松散耦合的符号系统	12
图 2-7 类型 3: 协作式系统	14
图 2-8 溯因学习框架 ^[1]	15
图 2-9 类型 4: 符号知识嵌入	16
图 2-10 类型 5: 符号知识作为约束	17
图 2-11 类型 6: 完全整合的系统	17
图 2-12 System 2 蒸馏概述 ^[2]	20
图 2-13 序列级别 KD 与 MINILLM 之间的比较 ^[3]	21
图 3-1 基于神经符号计算的知识融合框架结构	24
图 3-2 模块迭代训练过程	25
图 3-3 实验所需数据集及构成	27
图 4-1 注意力掩码矩阵示例	36
图 5-1 基于规则知识的提示词生成与优化	47

表目录

表 3-1 命名实体识别任务数据集信息	28
表 3-2 命名实体识别任务实验结果	29
表 3-3 情感分类数据集信息	30
表 3-4 情感分类实验结果	31
表 4-1 命名实体识别实验结果	38
表 4-2 命名实体任务识别消融实验结果 (移除注意力掩码生成模块后) .	39
表 4-3 文本摘要任务数据集信息	41
表 4-4 文本摘要实验结果 (Rouge-1)	42
表 4-5 文本摘要实验结果 (Rouge-2)	43
表 4-6 文本摘要任务消融实验结果 (移除注意力掩码生成模块后) (Rouge-1)	44

表 4-7 文本摘要任务消融实验结果 (移除注意力掩码生成模块后) (Rouge-2)	44
表 5-1 GSM8K 性能对比 (准确率)	52
表 5-2 SVAMP 性能对比 (准确率)	52
表 5-3 CommonsenseQA 性能对比 (准确率)	55
表 5-4 PIQA 性能对比 (平均准确率)	55
表 5-5 ARC-easy 性能对比 (准确率)	58
表 5-6 ARC-challenge 性能对比 (准确率)	58

符号列表

缩写

AI	Artificial Intelligence, 人工智能
DNN	Deep Neural Network, 深度神经网络
XAI	Explainable Artificial Intelligence, 可解释人工智能
NLP	Natural Language Processing, 自然语言处理
MCTS	Monte Carlo Tree Search, 蒙特卡洛树搜索
LLM	Large Language Model, 大型语言模型
RL	Reinforcement Learning, 强化学习
NTP	Neural Theorem Prover, 神经定理证明器
CTP	Conditional Theorem Prover, 条件定理证明器
NLM	Neural-Logical Machine, 神经逻辑机
GNN	Graph Neural Network, 图神经网络
LTN	Logic Tensor Network, 逻辑张量网络
NVSA	Neuro-Vector-Symbolic Architecture, 神经向量符号架构
VSA	Vector-Symbolic Architecture, 向量符号架构
SR	Symbolic Regression, 符号回归
PeRCNN	Physics-encoded Recurrent Convolutional Neural Network, 物理编码递归卷积神经网络
RD	Reaction Diffusion, 反应扩散
KD	Knowledge Distillation, 知识蒸馏
CNNs	Convolutional Neural Networks, 卷积神经网络
KLD	Kullback-Leibler Divergence, 相对熵
SST-2	Stanford Sentiment Treebank 2
MR	Movie Review
CR	Customer Review
LCSTS	Large Scale Chinese Short Text Summarization
CAIL2022	China AI and Law Challenge Dataset 2022
GSM8K	Grade School Math 8K
SVAMP	Simple Variations on Arithmetic Math Word Problems

PIQA Physical Interaction: Question Answering

ARC AI2 Reasoning Challenge

第 1 章 绪论

1.1 研究背景

人工智能 (Artificial Intelligence, AI) 的快速发展,特别是大型 AI 模型,已经在许多研究领域引起了重大变化,并对社会和工业部门的各个方面产生了深远的影响。与此同时,公众和科学界也越来越关注现代人工智能技术的可信度、安全性、可解释性和问责制^[4]。这就引出了一个自然的问题:下一代人工智能的关键推动因素是什么?

神经符号计算的概念并非一蹴而就,而是随着人工智能研究的不断发展逐步形成的。人工智能在历史上一直被两种范式所主导:符号主义和连接主义。最早的人工智能系统大多依赖于符号计算,尤其是在 20 世纪 50 年代到 80 年代之间,专家系统等早期 AI 技术不断兴起。符号主义认为,表征世界事物的符号是人类智能的基本单位,人类的认知过程是通过对符号的操纵,通过对符号表征的一系列规则和逻辑运算来完成的^{[5][6]}。许多早期的人工智能系统^{[7][8]}都是建立在符号模型上的。传统的符号计算通过显式的规则和逻辑推理来解决问题,其每一步推理过程都是清晰可解释的。符号计算方法有几个优点:它们只需要一些输入样本,使用强大的声明性语言进行知识表示,并且具有概念上清晰的内部功能。比如在专家系统中,通过“如果病人出现发热、咳嗽和疲劳症状,则可能患有流感”这样的规则来进行诊断。然而,很快就发现,这种基于规则的自上而下的策略需要大量的手工调整,缺乏真正的学习。在复杂领域中,人工总结和维护所有可能的规则也是一项极其耗时且困难的工作^[9]。以命名实体识别为例,仅靠“地名 + 大学 = 组织实体”这样的简单规则,难以覆盖现实语言中各种复杂的组织机构表达方式。此外,由于离散的符号表示和手工制定的规则往往过于刚性,不能容忍模糊和嘈杂的数据,符号方法在解决现实世界的问题时通常会出现不足,难以处理现实世界的模糊性和不确定性。

进入 90 年代后,随着计算能力的提升和大数据的出现,连接主义方法(以神经网络为代表)逐渐崭露头角。连接主义——以其最成功的技术深度神经网络 (Deep Neural Network, DNN) 为代表——是近年来大多数成功人工智能系统的核心架构^[10]。这种方法受神经系统生理学的启发,通过简单且通常统一的单元所构成的互联网络模拟认知过程^[11]。学习过程以数据为驱动,主要通过修改权重实现;例如,梯度反向传播技术能够在最小化所有训练样本的累计误差方向上调整网络的权重^[12]。连接主义模型具备一定的容错性,因为它们学习的是子符号——即连续的嵌入向量,并通过符号化的表示来比较这些向量,而不是直接处理实体和关系的字面意义。此外,连接主义模型通过从数据中学习统计模

式^[13]，展现了良好的归纳学习和泛化能力然而，连接主义方法也存在一些关键问题^{[14][15]}。首先，这类模型在处理组合概括方面表现不足，而人类认知的强大能力在于能够基于熟悉的部分解决全新的问题^[16]。其次，自下而上的连接主义模型通常需要海量数据进行训练，数据效率较低。最为重要的是，连接主义模型的决策过程往往缺乏透明性，表现出明显的“黑盒”特性^[17]，这使得其在关键任务中的应用受到了限制。由于难以解释模型的决策逻辑，人们无法明确其做出某个判断的理由，这在医疗诊断^[18]、自动驾驶^[19]等高风险领域尤其令人担忧。以医疗诊断为例，当模型给出某种疾病的判断时，医生和患者往往无法理解其背后的依据，这直接影响了模型的可信性和实际应用价值。因此，尽管连接主义模型在某些任务上表现出色，但其在决策关键型应用中的使用仍受到严重制约。

进入 21 世纪，随着深度学习技术的迅猛发展，神经符号计算作为一种结合符号推理和神经网络学习优势的混合范式应运而生^[20]，由神经符号学习和推理（神经符号计算）协会正式引入这一概念，并被广泛认为是推动下一代人工智能的重要力量^{[21][22]}。神经符号计算通过结合神经网络的学习能力与符号推理的可解释性，试图同时利用这两种范式的优势，规避其固有的缺陷。作为符号主义和连接主义的混合体，旨在整合两种基本认知能力：从经验中学习的能力和从所学知识中推理的能力^{[23][24]}。这种整合不仅增强了模型的学习效率，还提升了其推理过程的透明性与可信度。尽管如此，构建这样一个集成的系统仍然充满挑战。特别是，如何调和基于分布式表示的统计归纳学习与基于局部表示的逻辑演绎推理，仍然是研究中的难题^[25]。然而，近年来，神经符号计算方法已经在多个应用领域中展现出其优越性。例如，在自然语言处理任务中，可以通过符号规则来指导神经网络的学习过程，使模型既能适应数据的多样性，又能保持推理过程的清晰性。同时，在视觉关系理解^{[26][27]}、视觉问答^{[28][29][30]}、视觉场景解析^{[31][32]}和常识推理^[33]等任务中，神经符号计算的表现均优于传统的单一范式方法。

1.2 研究意义

根据 Gartner 的报告，2024 年全球企业对可解释 AI（Explainable Artificial Intelligence, XAI）的需求将显著增加。报告指出，金融、医疗和法律行业是可解释 AI 需求增长最快的领域。在金融风控中，AI 模型的决策需要向监管机构解释，确保合规性；在医疗领域，医生需要理解 AI 推荐的依据；在法律领域，AI 应用于判决支持时，必须具备可解释性以保障公正。

医疗诊断 在医疗领域，AI 模型的可解释性至关重要，尤其是在疾病诊断中。医生依赖 AI 系统提供的诊断结果，但若无法理解其背后的推理过程，便难以确保

患者的安全。例如，神经符号计算可以在神经网络的学习过程中引入规则和推理逻辑，使得模型不仅能够根据数据进行预测，还能提供明确的推理路径，从而提高医生对 AI 系统的信任。此外，神经符号计算还可以结合医学知识库，帮助系统在诊断过程中考虑更多的医学规则和临床经验，进一步提升诊断的准确性。例如，在癌症筛查中，系统可以通过结合病理学知识和患者的历史数据，提供更为精准的诊断建议，并通过可解释的方式展示其推理过程，帮助医生做出更明智的决策。

自动驾驶 自动驾驶系统依赖于大量传感器数据进行实时决策。尽管深度学习技术能够很好地识别道路标志、行人和其他交通元素，但其缺乏明确的逻辑约束，容易导致不可预测的行为。引入神经符号计算能够为自动驾驶系统提供规则性约束，例如交通法规和道德伦理，使其决策过程更加透明和可控。例如，在遇到紧急情况时，系统可以通过符号推理快速评估不同决策的风险和后果，并根据预设的伦理规则做出最优选择。这不仅提升了系统的安全性，还增强了用户对自动驾驶技术的信任。此外，神经符号计算还可以帮助系统在复杂的交通环境中进行更好的路径规划和决策，减少交通事故的发生率。

金融风控 在金融领域，神经符号计算能够结合金融规则和数据模式，提供既可靠又可解释的决策支持。例如，在信贷评估中，系统能够根据个人信用记录和历史数据进行判断，同时提供透明的推理过程，增加决策的透明度。此外，神经符号计算还可以帮助金融机构在反欺诈、风险评估和投资决策等任务中提升系统的可解释性和可靠性。例如，在反欺诈任务中，系统可以通过结合金融规则和用户行为数据，快速识别异常交易，并通过可解释的方式展示其推理过程，帮助金融机构及时采取措施。在投资决策中，系统可以通过结合市场规则和历史数据，提供更为精准的投资建议，并通过可解释的方式展示其推理过程，帮助投资者做出更明智的决策。

智能教育 在智能教育领域，神经符号计算可以帮助构建个性化的学习系统。通过结合学生的学习习惯、进度和知识点之间的逻辑关系，系统能够提供定制化的学习建议，并通过可解释的方式说明学习路径和推荐原因，增强学习者的参与感和信任感。此外，神经符号计算还可以帮助教师更好地理解学生的学习情况，提供针对性的教学建议。例如，在个性化学习中，系统可以通过结合学生的学习数据和知识点的逻辑关系，推荐最适合学生的学习路径，并通过可解释的方式展示其推理过程，帮助学生更好地理解学习内容。在教师辅助决策中，系统可以通过结合学生的学习数据和教学规则，提供针对性的教学建议，并通过可解释的方式展示其推理过程，帮助教师更好地进行教学设计。

神经符号计算的研究推动了符号主义和连接主义两种范式的有机结合。通过将这两种传统的人工智能思想进行整合，能够更全面地模拟人类的认知能力，促进 AI 系统在理解、推理和学习等方面的发展。神经符号计算不仅为人工智能提供了新的理论框架，还为跨领域的认知科学研究提供了新的思路，推动了跨范式 AI 整合的进程。此外，神经符号计算还为多学科的交叉研究提供了新的契机，例如在认知科学、神经科学和计算机科学等领域的合作中，神经符号计算为理解人类认知过程提供了新的视角，促进了多学科的协同发展。

1.3 研究问题

本论文将当前的研究问题拆解为了几个具体子问题，以进一步聚焦挑战：

(1) 符号规则与神经网络表示的对齐

神经网络通常采用连续的向量表示，而符号推理依赖于离散的符号表示。如何将这两种表示形式对接，是神经符号计算中的一个核心问题。研究者需要找到合适的方式将神经网络的学习能力与符号推理的逻辑框架有效结合。具体而言，如何设计有效的映射机制，将神经网络的输出转化为符号推理所需的输入，以及如何将符号规则嵌入到神经网络的训练过程中，都是需要深入研究的方向。此外，如何在不同层次上实现符号与神经网络的交互，也是一个值得探讨的问题。

(2) 动态知识融合机制

神经符号计算不仅需要在训练阶段整合神经网络与符号推理，还要能够在动态环境中融合新的知识。如何处理不同来源的知识、如何在已有知识库的基础上进行有效推理，是一个重要的挑战。特别是在面对不断变化的外部环境时，如何设计自适应的知识更新机制，使得系统能够及时吸收新知识并调整推理策略，是神经符号计算系统能否在实际应用中取得成功的关键。此外，如何在不同知识源之间进行冲突检测与解决，也是动态知识融合中需要解决的问题。

(3) 解释性与推理能力的平衡

神经符号计算系统需要在保持推理能力的同时，保证系统的可解释性。如何设计一个既具备推理能力，又能清晰解释推理过程的系统，是研究中的一个重要课题。具体而言，如何在复杂的推理过程中提取出关键步骤，并以人类可理解的方式进行呈现，是一个技术难点。此外，如何在降低系统性能的前提下，增强其解释能力，也是需要平衡的问题。特别是在高风险应用中，如医疗诊断和自动驾驶，系统的决策过程必须透明且可追溯，以确保用户对系统的信任。

(4) 系统的鲁棒性与可扩展性

神经符号计算的混合系统不仅要能够处理现实中的复杂问题，还要具备足够的鲁棒性，能在多变的环境中保持良好的性能。特别是在面对噪声数据或异常输入时，系统应具备一定的容错能力，避免因输入数据的偏差而导致错误的推理

结果。此外，如何设计具有良好可扩展性的系统，使其能够适应不同规模和复杂度的任务，也是一个重要的研究方向。特别是在大规模知识库和复杂推理任务中，如何优化系统的计算效率，提升其处理能力，是神经符号计算系统能否广泛应用的关键。

1.4 研究范围和目标

1.4.1 主要研究内容

本研究针对神经符号计算的知识数据融合问题，探索了三个不同层次的解决方案：面向 DNN 的规则知识融合框架、面向 Transformer 的知识融合框架，以及面向大规模预训练模型的知识融合框架。研究思路如图 1-2 所示。

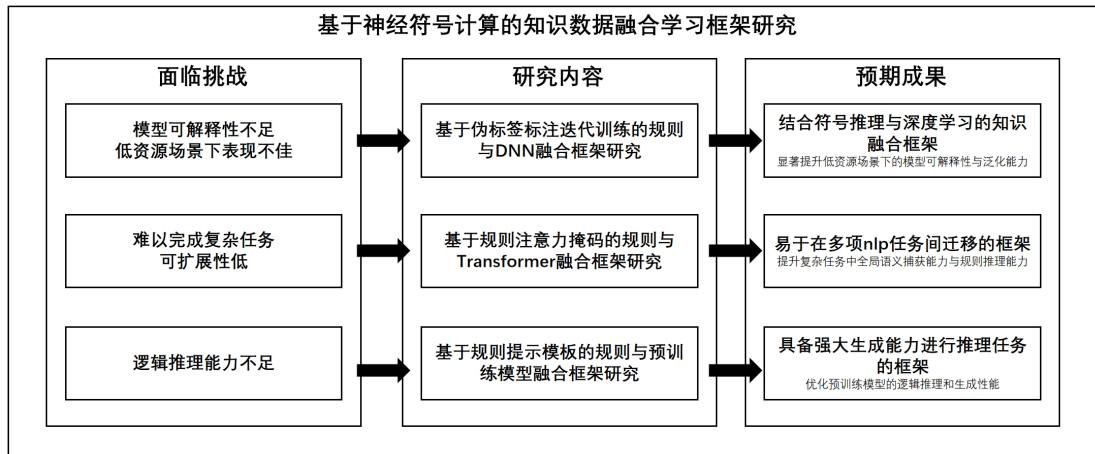


图 1-1 研究思路

第一个研究内容专注于规则知识和 DNN 的结合框架。这种方法需要对 DNN 的网络结构和训练过程进行修改，以适应不同任务场景下的规则约束。虽然这种方法在理论上更加严谨，能够保证规则的准确执行，但由于需要修改模型架构，实际应用成本较高，且难以在不同任务间迁移。此外，对于复杂的规则约束，可能需要设计特殊的网络层或损失函数，增加了实现的难度。

为了解决上述问题，本文提出了第二个研究内容：面向 Transformer 框架的知识融合方法。这种方法无需修改模型架构，只需要通过精心设计的数据格式将规则知识编码到训练数据中。通过利用 Transformer 强大的表示能力，模型能够自动学习规则和数据之间的关系。这种方法具有更好的通用性和可扩展性，可以方便地应用到不同的任务场景中。

最后，随着大规模预训练模型的发展，本文探索了第三个研究内容：在预训练模型微调阶段注入规则知识。具体而言，本研究选择了中低参数量的预训练模型作为基础，通过设计特定的提示模板和微调策略，实现规则知识的高效融入。这种方法充分利用了预训练模型已有的知识，同时通过规则引导，提高了模型在

特定任务上的表现。

这三项研究内容各有优势，适用于不同的应用场景：第一项研究适合对规则执行要求严格的场景，第二项研究适合需要快速部署的通用场景，第三项研究则适合需要强大生成能力的复杂任务。未来研究将进一步探索如何结合这些方案的优势，构建更加智能和可靠的知识融合框架。

1.4.2 预期目标

本课题旨在通过系统性研究，探索神经符号计算与不同层次深度学习模型的融合方法。研究将从三个层面逐步深入：首先设计面向 DNN 的规则融合框架，通过显式编码规则知识，结合伪标签生成和知识蒸馏机制，来提升模型可解释性；其次开发面向 Transformer 的通用知识融合框架，无需修改模型架构，实现规则知识的高效迁移；最后探索在中小规模预训练模型中融入规则知识的微调增强框架，设计规则引导的微调方法，平衡生成能力与规则约束。实验将重点评估每个层次方案的效果，分析其在不同场景下的适用性，为后续研究提供理论和实践支持。

本课题的创新点如下：

(1) 面向 DNN 的规则知识融合框架

在面向 DNN 的方案中，提出了双向知识蒸馏与伪标签迭代训练相结合的方法，实现符号规则与神经网络模式学习的动态协同优化，增强模型在低资源场景下的可解释性和泛化能力。

(2) 面向 Transformer 的规则知识融合框架

在面向 Transformer 的方案中，引入了基于规则的注意力掩码机制，通过显式规则约束增强多头自注意力机制的全局语义捕获能力，实现符号逻辑与全局语义信息的高效结合。

(3) 面向预训练模型的规则知识融合框架

在面向预训练模型的方案中，设计了规则提示模板以显式引导模型的逻辑推理，同时结合知识蒸馏技术，将规则理解能力迁移至轻量化学生模型，实现规则知识与预训练模型生成能力的深度融合。

1.5 本文组织结构

本文的组织结构如图1-2所示，共分为六个章节。本章为绪论，主要介绍了神经符号计算的相关背景，阐述了如何实现神经网络与符号规则的有效融合、如何提升模型的可解释性以及如何在实际应用中平衡模型性能与可解释性这三大挑战，说明了神经符号计算研究的意义以及动机，系统概括了本文的主要研究内容与贡献，最后介绍了本文的结构，后续章节的组织如下：

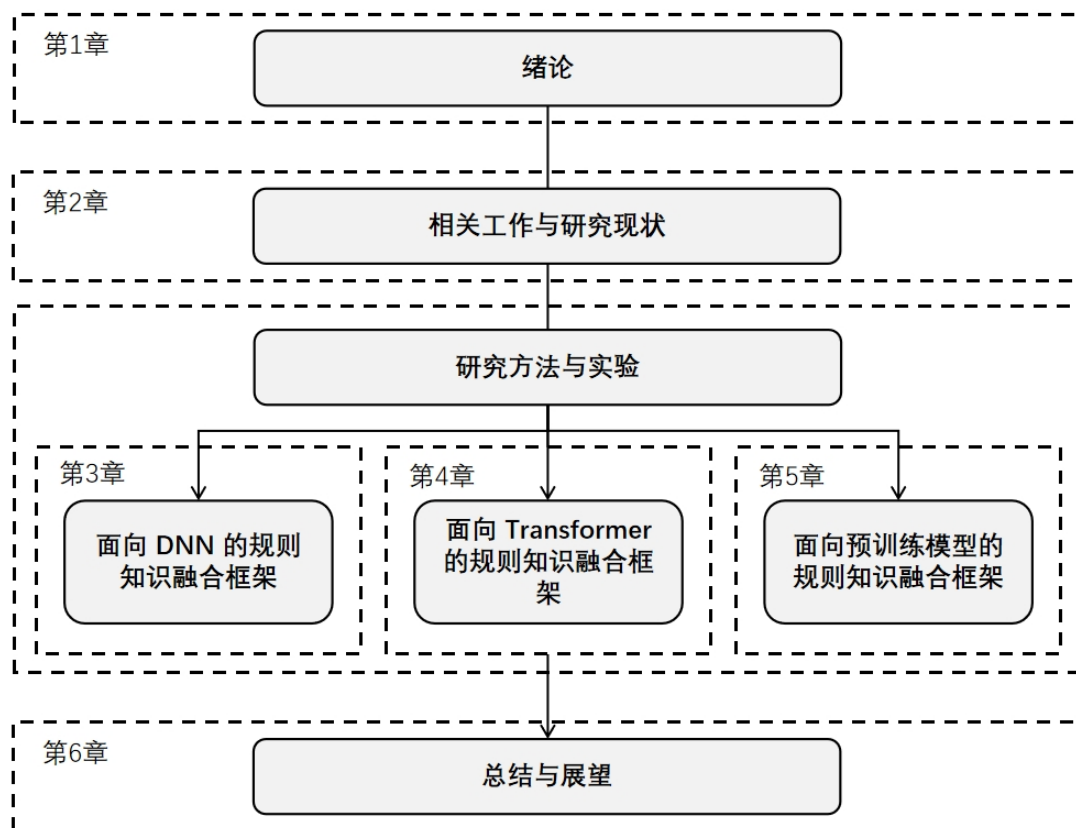


图 1-2 本文组织架构

第 2 章回顾了神经符号计算领域的研究现状，从理论基础和应用实践两个角度系统总结了现有方法。在理论基础研究方面，涵盖了从早期符号主义与连接主义的融合尝试到近期深度学习与逻辑推理相结合的发展历程；在应用实践研究方面，探讨了神经符号计算在自然语言处理、计算机视觉、知识图谱和自动推理等领域的实际应用。本章指出了模型可解释性不足、规则表达能力有限和知识融合困难等当前研究中的主要挑战，为后续研究提供了理论基础。

第 3 章是面向 DNN 的规则知识融合框架，针对传统深度学习模型难以有效融合符号规则的挑战定义了科学问题并提出了解决方案。该章节首先分析了深度神经网络在规则表达与推理方面的局限性；然后提出了一种基于知识蒸馏的神经符号融合方法，将符号规则转化为软约束引导神经网络的学习过程；接下来的小节详细介绍了实验设置，包括使用的数据集、评估指标及对比方法；最后汇总并分析了实验结果，验证了提出方法在规则融合与模型可解释性方面的有效性。

第 4 章是面向 Transformer 模型的规则知识融合框架，针对 Transformer 架构中自注意力机制与符号规则结合的挑战提出了解决方案。该章节首先提出了一种注意力引导机制，通过将符号规则转化为注意力矩阵的约束条件，实现了规则知识对模型注意力分布的直接指导；接着引入了基于模块化设计的神经符号架

构，将符号推理与神经计算分离为独立模块并建立有效的交互机制，从而提升了模型的规则应用能力和推理透明度。最后，通过多个自然语言处理任务的实验验证了提出框架的有效性和泛化能力。

第 5 章是面向预训练模型的规则知识融合框架，聚焦于如何在不破坏预训练模型原有知识的前提下注入符号规则知识。该章节首先分析了预训练语言模型在逻辑推理和规则遵循方面的不足，并明确了现有方法在规则融入过程中的局限性；然后提出了一种基于提示工程与规则分解的创新框架，通过将复杂规则分解为多个简单子规则并设计相应的提示模板，有效引导预训练模型执行符号推理；接着，详细描述了实验设计，包括基准任务、评价指标及对比方法；最后，实验结果验证了本方法对预训练模型推理能力与规则遵循程度的显著提升，进一步证明了神经符号融合在大型语言模型应用中的关键作用。

第 6 章对全文的工作进行总结，归纳了三种面向不同神经网络架构的规则知识融合框架的技术特点与应用价值，并对神经符号计算的未来发展方向进行了展望，包括多模态神经符号系统、自适应规则学习以及神经符号计算在实际应用中的部署与优化等方面。

第2章 相关工作与研究现状

2.1 研究基础

2.1.1 神经符号集成方式

在神经符号计算领域，研究人员通常根据符号与神经模块的组合方式对系统进行分类，这种分类方式体现了两者在具体集成中的协同机制。神经符号计算的核心目标是融合神经网络的学习与泛化能力，以及符号系统的逻辑推理与规则表达优势，以弥补各自单一方法的局限性，从而推动智能系统向更高层次发展。在这一框架下，符号模块通常负责规则推理、知识表示和结构化信息处理，而神经模块则在从感知数据中提取特征、模式发现和表征学习方面表现优异。两者的结合方式直接影响系统的架构与功能表现，既可以通过模块化协作的方式各自独立完成任务，也可以实现深度融合，形成紧密结合的整体系统。目前国际上主要认可的分类有以下六种^[34]，这些多样化的集成策略不仅为神经符号计算的创新提供了广阔的空间，还为解决复杂的人工智能挑战带来了全新的方法论与实践路径。

(1) 类型 1：符号神经符号系统

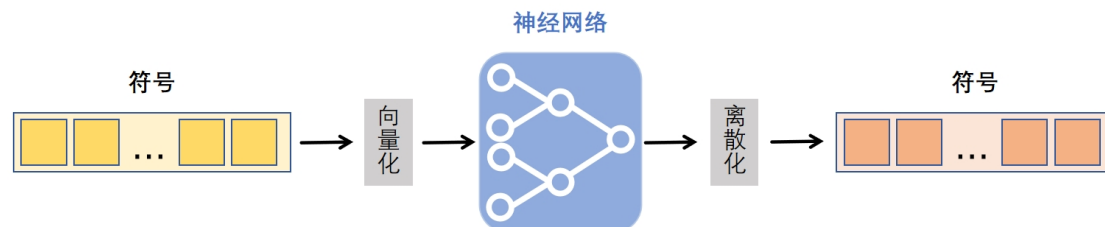


图 2-1 类型 1：符号神经符号系统

正如 Kautz^[34] 所描述的，这是当前深度学习方法在处理符号化输入和输出时的标准操作流程。例如，大多数自然语言处理（Natural Language Processing, NLP）系统，包括 GPT-3^[35] 等大型语言模型，都属于这一类。在这些系统中，符号输入通过如 word2vec^[36] 和 GloVe^[37] 等方法转换为向量嵌入，随后由神经网络模型处理，最终输出通过 softmax 操作转化为符号类别或符号序列。Kautz 强调了这一类型，指出神经网络的输入和输出可以由符号组成，例如在语言翻译或图像分类任务中。尽管如此，将这种类型归为神经符号计算可能稍显宽泛，因为它几乎涵盖了任何具有明确语义空间的深度学习模型。

举例来说，在自然语言处理中，GPT-3 通过将输入的文本符号（如单词或句子）转换为向量表示，经过多层神经网络的复杂计算后，输出相应的符号序列（如翻译后的文本或生成的文本）。这种类型的系统在机器翻译、文本生成和情感

分析等任务中表现出色，但由于其输入和输出仅限于符号形式，系统的内部推理过程仍然缺乏透明性和可解释性。如图 2-2 所示，GPT-3 通过“符号-向量-符号”的闭环架构实现文本生成：当用户输入“Recite the second law of robotics”这类符号序列时，系统首先将其转换为高维向量，经过多层神经网络计算后重建为输出符号“A robot must obey the orders given it by human beings”。这一过程在技术实现上可分为三个阶段：

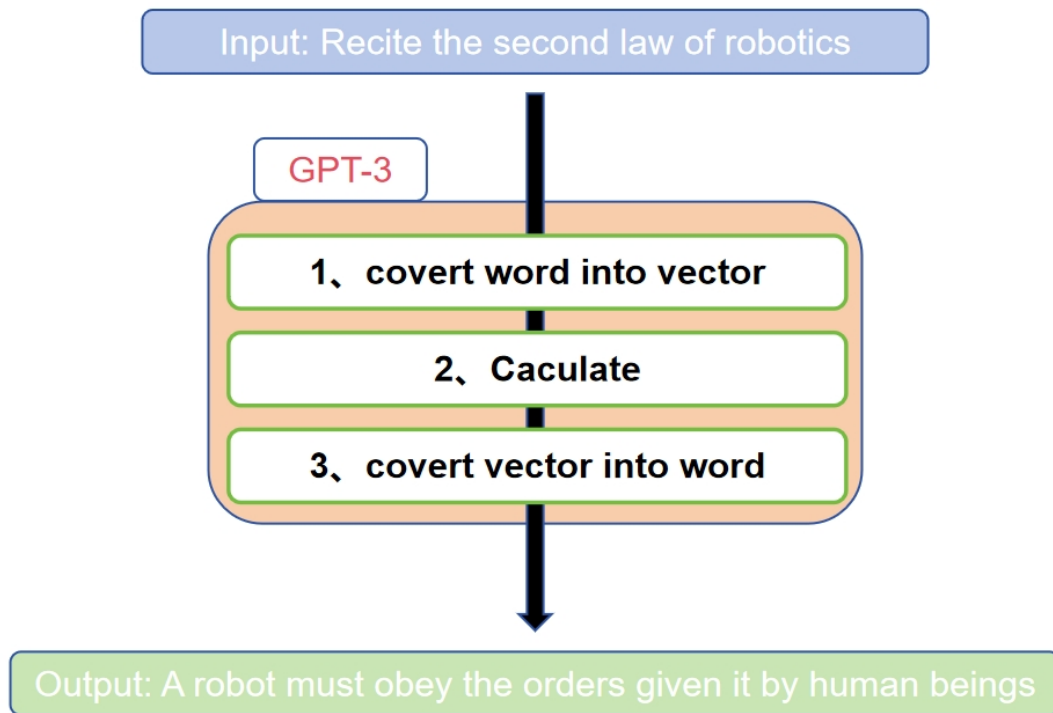


图 2-2 GPT-3 工作流程

第一阶段：符号的向量化编码，如图 2-3 所示。GPT-3 的输入和输出都是固定长度为 2048 个 Token 的序列，对于短序列作为输入，会用“空”值填充所有额外位置。GPT-3 的词汇表将 50257 个单词（含子词单元）映射为离散 ID，这使其能够为每个单词赋予一个值（id）。Aardvark 是 0，aaron 是 1，依此类推。最终，通过单点编码（one-hot encoding）可以将每个单词转换为大小为 50257 的单点编码矢量，其中仅索引 i 处的维（单词的值）为 1，所有其他维度均为 0。对序列中的每个单词都执行此操作，就得到了一个 2048×50257 的 1 和 0 矩阵。50257 对于矢量来说相当大，并且大部分都用零填充，造成了存储空间上的浪费。为了解决这个问题，模型学习了一个嵌入函数：一个神经网络，该神经网络采用 50257 长度的 1 和 0 的向量，并输出 n 长度的数字的向量。在这里，模型试图将单词含义的信息存储（或投影）到较小的空间中。例如，嵌入的维数为 2，那么就好比将每个单词存储在二维空间中的特定坐标处。当然嵌入尺寸通常大于 2，GPT 使用 12288 尺寸。所以每个单词 one-hot 表示都与嵌入网络权重相乘，最终得到降维后的嵌入向量矩阵。为了对当前 Token 在序列中的位置进行编码，采用了 Token

的位置，并将其传递给 12288 个正弦函数，每个函数的频率都不同。对于每个 Token，结果是 12288 个数字向量。与嵌入一样，将这些向量组合成具有 2048 行的单个矩阵，其中每一行是序列中 Token 的 12288 列位置编码。最终与词嵌入向量逐元素相加，形成包含语义和位置信息的 2048×12288 输入矩阵。

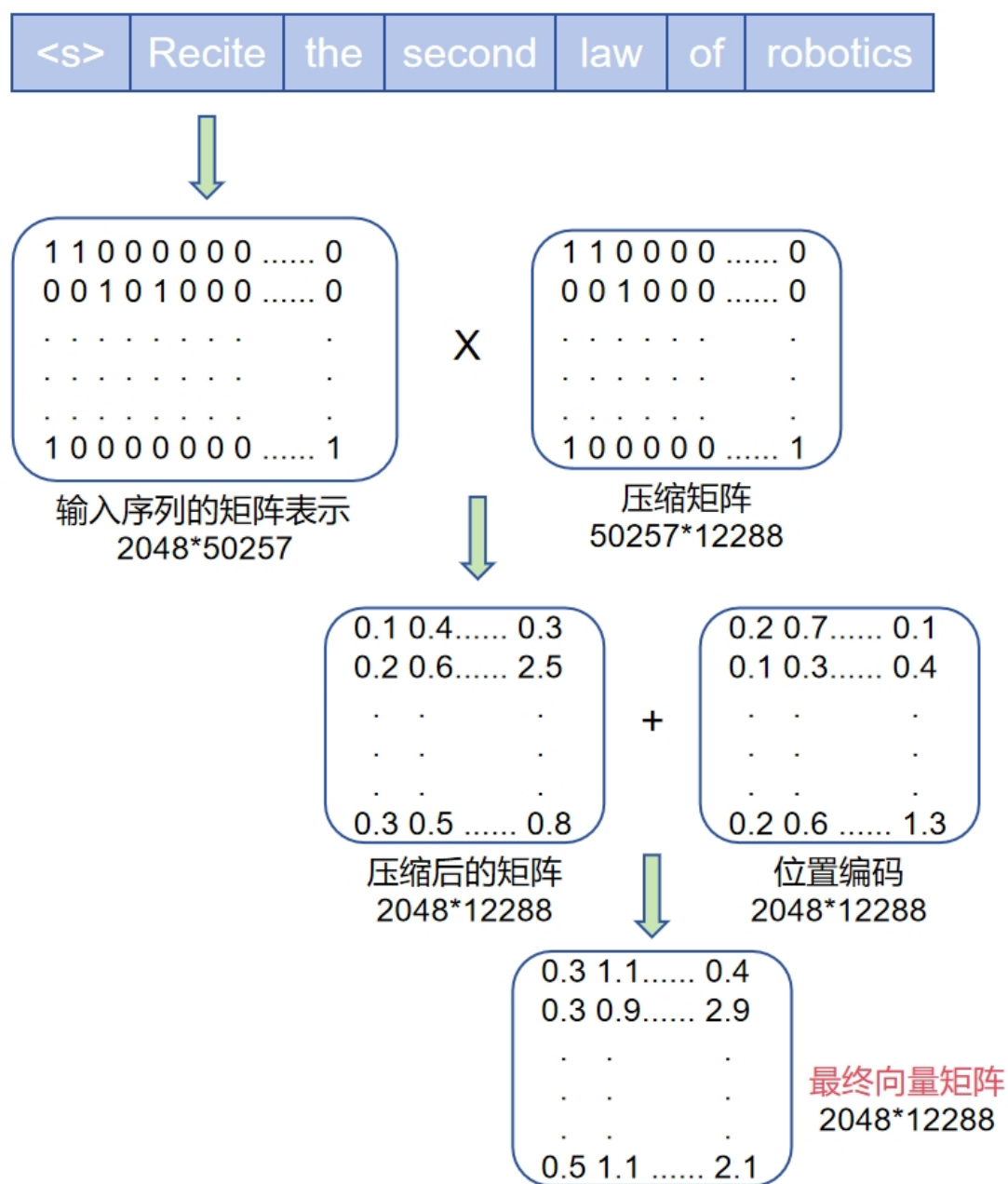


图 2-3 符号的向量化编码

第二阶段：深度特征重构，如图 2-4 所示。输入的 2048×12288 矩阵流经 96 层 Transformer 解码器（以最大版本的 GPT-3 为例，即 175B 参数版本），每层配置 96 个注意力头，通过 QKV 投影矩阵将输入拆分为 96 组 128 维子空间。在计算“robot”与“obey”的关联度时，每个头独立学习不同的依赖模式（如语法结构、语义焦点），表明复杂的特征交互。

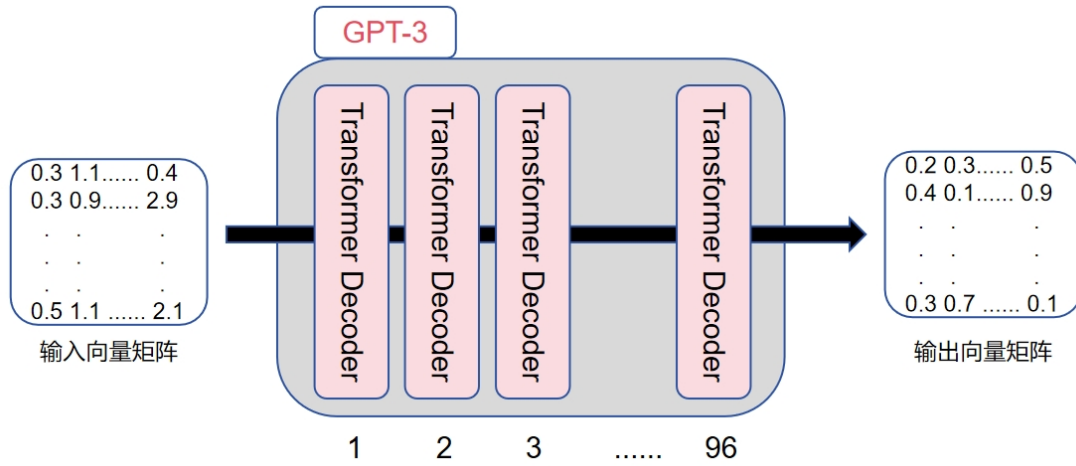


图 2-4 深度特征重构

第三阶段：解码，也就是概率驱动的符号重建，如图 2-5所示。系统通过将输出矩阵与嵌入矩阵的转置相乘，使 12288 维输出向量投射回 50257 维词表空间，在施加快速 SoftMax 之后，可以将结果值视为每个单词的概率。此外，还设置了参数 top-k，该参数将输出中要采样的可能单词的数量限制为 k 个最可能的预测单词。例如，当 top-k 参数为 1 时，总是选择最可能的单词。

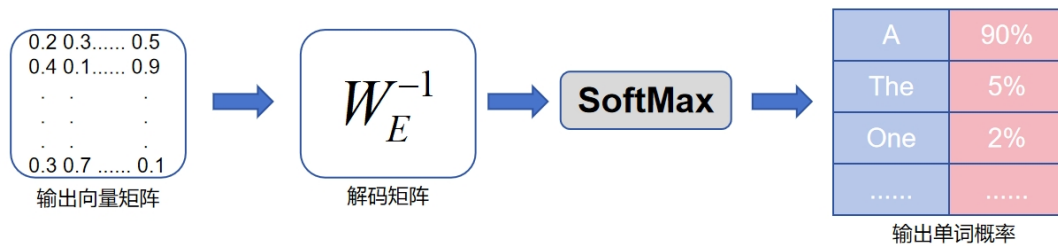


图 2-5 解码

(2) 类型 2：松散耦合的符号系统

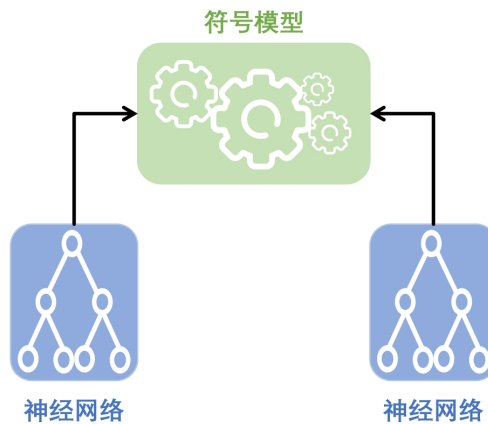


图 2-6 类型 2：松散耦合的符号系统

这种类型的系统是混合的符号系统，神经模块在符号问题求解器中作为子程

序使用。符号和神经部分是松散耦合的。Kautz 提到了 DeepMind 的 AlphaGo^[38] 作为示例：问题求解器是基于蒙特卡洛树（Monte Carlo Tree Search, MCTS）搜索的算法，而启发式评估函数则由神经网络提供。在^[39]中，神经符号计算模型被设计用于学习和概括组合规则。该模型采用 seq2seq 生成网络，并结合符号栈机来处理递归和序列操作。另一个例子是 VisProg^[40]、HuggingGPT^[41] 和 ViperGPT^[42] 等基于大语言模型（Large Language Model, LLM）的 AI 代理，它们利用 LLM 将复杂任务分解为一系列子任务，这些任务可以通过现有的符号模型来完成。

AlphaGo 作为松散耦合式神经符号系统的典范，通过蒙特卡洛树搜索（符号推理）与深度强化学习（神经计算）的协同机制，实现了围棋决策中“直觉判断”与“逻辑推演”的双重思维模拟。其核心创新源于传统博弈论与现代深度学习的交叉融合，即在符号化搜索框架中嵌入神经网络的启发式评估，形成“感知-评估-推演”的闭环决策系统。以围棋中的局部攻防决策为例，系统需同时完成局面特征提取（识别棋盘状态）和策略推演（计算后续变化），这种双重认知的协同过程突破了传统 AI 系统的单模态思维局限。

蒙特卡洛树搜索是一种面向决策规划的模拟搜索算法。其核心思想是反复进行模拟，在搜索树中逐步扩展节点，通过蒙特卡洛随机模拟来评估节点价值，从而逐渐逼近最优决策。MCTS 在没有完备棋局评估函数的情况下特别有用，因为它可以通过大量随机对局模拟估计动作的胜率。相比传统的 Minimax 搜索或动态规划，MCTS 不需要遍历整个决策空间，而是通过随机采样和逐渐逼近来聚焦于“最有希望”的分支。

AlphaGo 中，深度神经网络起到了“大脑”的作用——策略网络和价值网络评估棋盘位置并指导行动选择。策略网络 (Policy Network) 以当前状态，也就是棋局局面作为输入，输出一个针对可选动作的概率分布。对于围棋，输出空间为 19×19 的交叉点上落子的概率分布；对于国际象棋等，则可以编码成一维的动作集合。策略网络的作用是快速选出有前途的动作。在 AlphaGo 中，先是训练了一个“监督学习策略网络”来模仿人类高手的落子分布，然后在此基础上通过自我对弈强化学习得到“强化学习策略网络”，用于指导 MCTS。在每一步搜索开始时提供每个动作的先验概率，MCTS 会基于这些先验来进行更引导性的搜索，而非平均地探索所有可能动作。因此，策略网络加速了搜索收敛——高概率的动作更早得到更多模拟验证，提高了效率。价值网络 (Value Network) 以当前状态作为输入，输出一个评估值，表示在该状态下当前玩家的胜率或预期回报。对于二人零和博弈，这通常介于 -1 和 +1 之间（-1 代表必败，+1 代表必胜，中间值表示胜率差异）。价值网络充当评估函数的角色。传统游戏程序需要人工设计复杂的评估函数来判断局面优劣，而 AlphaGo 用深度网络从海量自我对弈中自动

学出评估。在 MCTS 中，当到达某节点需要评估时，价值网络可以直接给出这个局面的估计胜率，而不必再完整模拟到终局。这极大节省了计算，并且神经网络可以比简单的随机模拟得到更准确的评估。值得注意的是，价值网络的训练需要自我对弈数据——即大量状态最终是否赢的结果，用监督学习的方法逼近真实胜负概率。

该系统的突破性进展体现在解决了符号搜索空间爆炸与神经网络可解释性不足的双重难题。AlphaGo 通过蒙特卡洛树搜索与深度神经网络的协同架构，实现了直觉与计算的完美平衡：当符号搜索面临分支因子过大的困境时，策略网络提供先验概率指导搜索方向；当神经网络评估存在盲区时，符号推演通过模拟对弈验证策略可行性。这种“评估-搜索-验证”的迭代过程，使系统在 19×19 的庞大状态空间中仍能保持决策精度。

(3) 类型 3：协作式系统

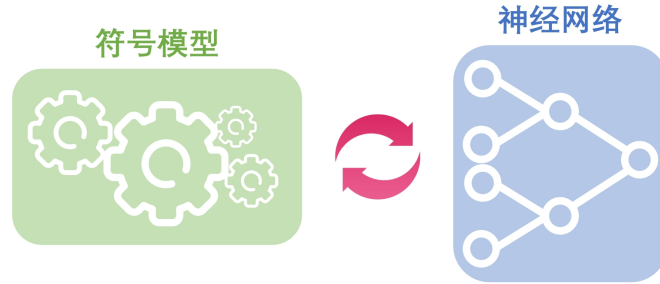


图 2-7 类型 3：协作式系统

第三类类型同样属于混合系统，但神经部分和符号部分各自专注于大流程中不同且互补的任务。与类型 2 不同，神经部分在类型 3 中充当协程，而非子程序。Kautz 解释了这一点，在类型 3 中，神经和符号部分之间的交互能够提升各自的独立性能，同时也提升整体系统的表现。因此，在第三类系统中，神经和符号部分的关系是协作性的，而不是像第二类系统中那样存在功能依赖。一个例子是溯因学习框架（Abductive Learning, AL）^[1]，其中符号逻辑推理和神经感知学习是相互独立但交互进行的。另一值得关注的案例是 J. Mao^[29] 等人提出的神经符号计算视觉模型，其中神经感知模块学习视觉概念，符号推理模块则在这些概念表示上执行符号程序以回答问题。符号推理模块同时为神经感知模块的基于梯度的优化提供反馈信号。近年来，神经符号强化学习（Reinforcement Learning, RL）的研究进展也归入类型 3^{[43][44][45][46]}。例如，将符号规划^{[47][48]} 集成到强化学习系统中以实现更稳健的决策过程。符号规划用于指导任务执行，而任务经验的反馈则用于不断改进符号规划。其他相关的例子还包括神经定理证明器（Neural Theorem Prover, NTP）^[49]、条件定理证明器（Conditional Theorem Prover, CTP）^[50]、NLProlog^[51]、DeepProbLog^[52]、NeuroLog^[53] 和 DiffLog^[54]。其中，值得特别关注的是 DeepProbLog^[52]，它将神经网络用于谓词计算概率事实的概率，

并通过 ProbLog^[55]——一种基于一阶逻辑的概率逻辑编程语言——来计算期望损失的梯度。

溯因学习框架作为协作式系统的典型代表，通过神经感知与符号推理的闭环协作机制，实现了人类智能中“感知-推理”双通道思维的模拟。其核心思想源自科学方法论中的溯因推理，即在给定观察结果和领域知识的前提下，逆向推导出最合理的解释假设。以足球运动中的传球决策为例，运动员需要同时完成视觉感知（识别球和队友位置）和逻辑推理（判断最佳传球路径），这种双模态思维的协同过程正是溯因学习的核心目标。当前的 AI 系统往往将感知与推理割裂处理，而溯因学习通过神经逻辑机（Neural-Logical Machine, NLM）的创新架构，首次实现了两种认知能力的深度融合。该系统通过“感知-假设生成-假设验证”的三阶段闭环架构，实现了符号逻辑与神经网络的协同进化。

该框架的核心突破在于解决了符号推理模块与神经网络的协同训练难题。如同 19 世纪学者鲍迪奇破译玛雅文字时采用的试错策略，NLM 通过启发式搜索方法在符号空间进行假设生成与修正：当感知模块输出的符号假设存在逻辑矛盾时，基于 Prolog 的溯因推理引擎会定位错误符号，并通过命题化技术将其转化为可训练的关系特征。这种“假设-验证-修正”的迭代过程，使系统在仅有像素级输入的情况下，能够自主发现符号间的逻辑规则。

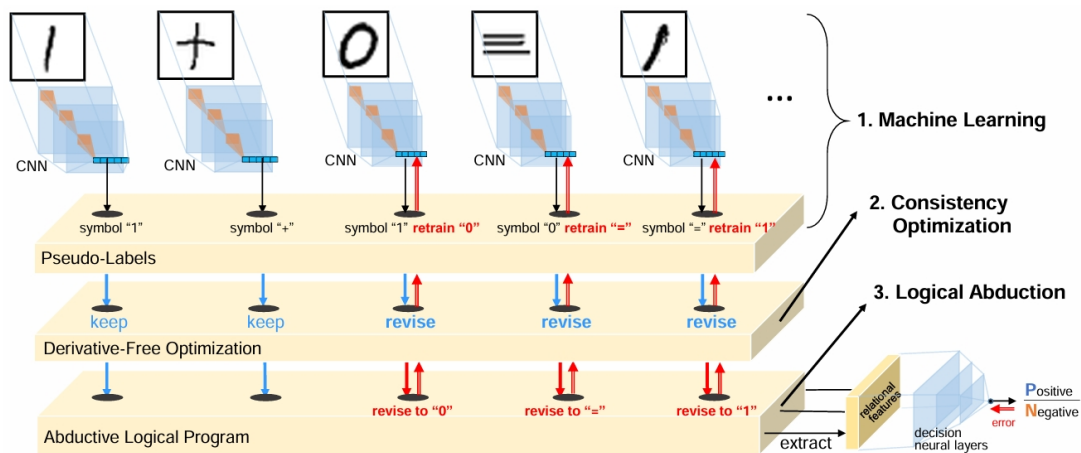


图 2-8 溯因学习框架^[1]

如图 2-8 神经逻辑机 NLM 的工作流程可分为三个关键阶段：

- 1、感知符号化：**深度卷积网络将原始图像（如手写公式“1+1=2”）映射为符号候选集，同时保留空间关系信息。初始阶段由于网络参数随机，常产生如将“=”误识别为“≈”的符号错误。
- 2、溯因推理：**符号假设被送入逻辑程序验证，结合领域知识（如等式两边需平衡）生成修正建议。对于错误假设“1+1≈2”，推理模块通过约束求解提出将“≈”修正为“=”。

3、反馈优化：修正信号以两种方式反向传播——符号层面更新逻辑规则库，特征层面通过对比损失调整感知网络参数。

溯因学习的创新性体现在三个方面：构建双向可微的神经-符号接口实现端到端训练；通过假设空间剪枝将推理复杂度从指数级降为多项式级；开发符号引导的数据增强提升模型鲁棒性。这些突破为 AI 系统实现真正的认知智能提供了新范式，在科学发现、知识图谱补全等领域展现出独特优势。

(4) 类型 4：符号知识嵌入

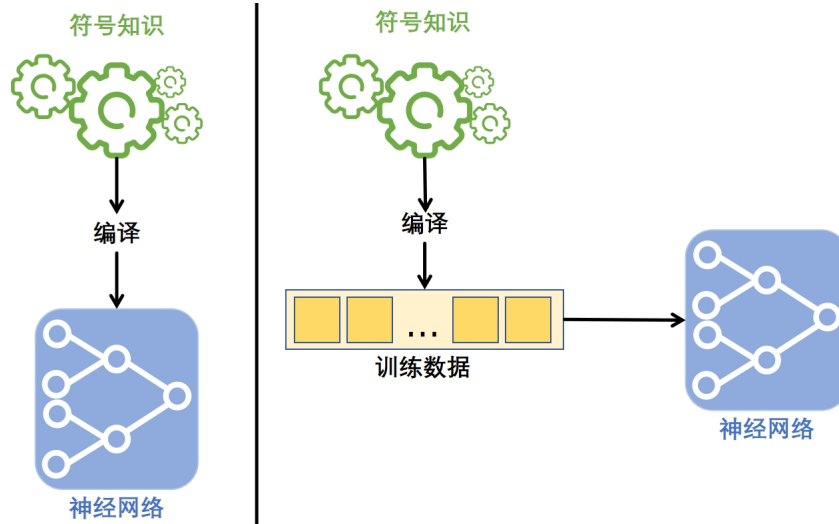


图 2-9 类型 4：符号知识嵌入

在类型 4 系统中，符号规则或符号知识被直接嵌入到神经网络的结构或训练机制中。例如，近年来对将符号知识嵌入神经网络表示的研究兴趣大幅增加，旨在自然地将符号领域的知识融入连接主义架构^{[56][57][58]}。一种常见的做法是将符号知识嵌入为神经网络的输入或输出。例如，某些神经符号数学系统^{[59][60]}和验证系统^[61]将数学表达式表示为树结构，用作训练数据。这种方法允许系统在处理复杂的数学问题时，利用符号结构来增强其推理能力。在视觉问答领域，许多模型^{[30][62][63]}生成并执行用于回答问题的符号程序，其中这些程序被实现为完全可微的操作。这种可微性允许系统在执行符号推理的同时，通过神经网络进行端到端的优化，从而提高整体性能。此外，一些研究^{[64][65][66]}试图将领域知识作为逻辑约束嵌入神经网络的输出层，以确保输出符合特定的逻辑规则。例如，在^[67]中，设计了一种易于处理的概率模型，用作自回归文本生成模型（如 GPT-2^[68]）的输入层，以施加词汇约束。通过这种方式，系统能够生成符合特定语义或逻辑规则的文本输出。近年来，一些算法^{[69][70]}还使用图神经网络（Graph Neural Network, GNN）将实体和关系嵌入外部知识库中，从而提升计算机视觉和自然语言处理任务的性能。尽管 GNN 可能在推理能力方面较为基础，但它们在处理与符号知识相关的任务时表现出了巨大的潜力。

(5) 类型 5：符号知识作为约束

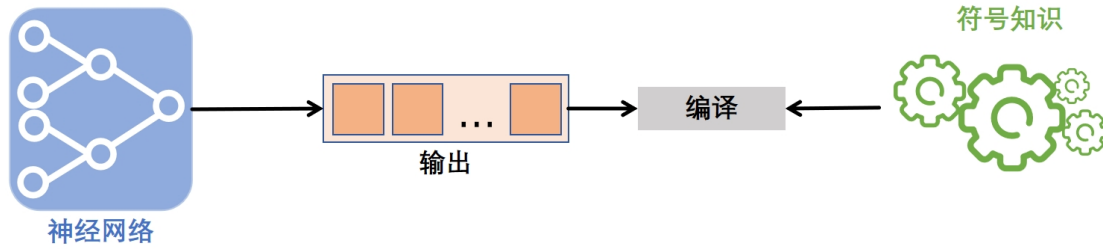


图 2-10 类型 5：符号知识作为约束

类型 5 的神经符号计算系统通过将符号知识转化为神经网络训练中的额外软约束，从而影响网络的学习过程。符号知识被嵌入训练中的损失函数中，以确保网络能够学习到符合符号知识的输出。在这种类型中，知识被编译进深度神经网络 (Deep Neural Network, DNN) 的权重中。在这方面，最近的研究包括逻辑张量网络 (Logic Tensor Network, LTN)^{[71][26][72]}。此方法将一阶逻辑公式转换为神经网络可处理的模糊逻辑表示，允许系统在学习过程中进行近似推理，而不是要求固定精确的推理。在模糊逻辑中，变量的真值可以在 0 到 1 之间浮动，0 表示完全为假，1 表示完全为真，而中间值则表示不确定性。LTN 系统通过放松布尔逻辑中的硬性规则，使得逻辑推理可以嵌入神经网络的学习目标中，并通过梯度下降进行端到端的训练。此外，符号知识还可以通过类层次结构来影响分类器的设计。例如，在^[73]和^[74]中，类层次结构既作为分类目标，也作为背景知识。研究者设计了不同的训练目标函数，以鼓励预测结果与类层次结构之间的一致性。在^[32]中，研究者将语义层次上的组成关系转化为分层场景解析中的额外训练目标，以确保模型生成的结果符合指定的语义层次。为了有效计算损失，^[75]提出了一种近似方法，使用局部分布上的样本约束代替全局的强约束，从而有效减少计算开销。通过这些方法，符号知识被有效嵌入到神经网络的训练过程中，增强了系统的可解释性和推理能力。

(6) 类型 6：完全整合的系统

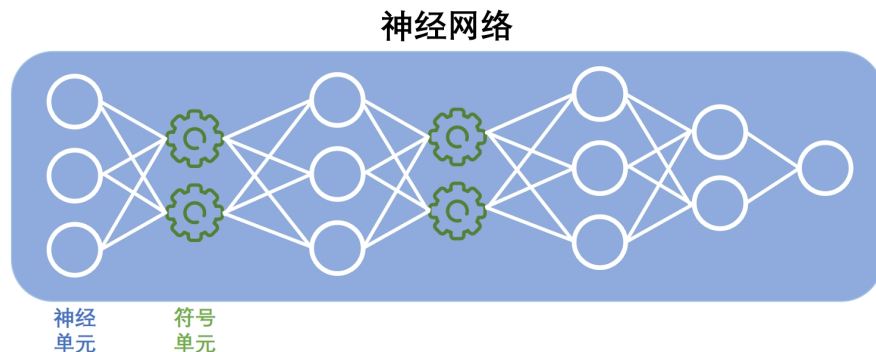


图 2-11 类型 6：完全整合的系统

Kautz 认为，这种类型的系统“最有潜力结合基于逻辑的 AI 和基于神经的 AI 的优势”。在这类系统中，符号推理引擎直接嵌入神经引擎中。例如，一些方

法通过神经网络学习执行符号运算^{[76][77][78]}，尽管这些方法的符号推理能力仍然较弱。可惜的是，目前还没有一种神经符号计算方法能够完全达到这个类型的要求。

2.1.2 神经符号计算应用

(1) 视觉推理任务

Hersche 等人提出了神经向量符号架构^[79] (Neuro-Vector-Symbolic Architecture, NVSA)，将向量符号架构 (Vector-Symbolic Architecture, VSA) 与深度神经网络相结合，应用于视觉推理任务。NVSA 通过高维分布式表示与符号操作解决了深度学习方法的绑定问题和符号推理的计算复杂度问题。在 RAVEN 和 I-RAVEN 数据集中，NVSA 通过端到端训练达到了 87.7% 和 88.1% 的平均准确率，相比于纯神经网络（如 SCL）和神经符号方法（如 PrAE）分别提高了 0.5% 和 17% 以上。同时，NVSA 在推理效率上表现卓越，通过向量化推理操作，将符号推理的复杂度从 $O(n^3)$ 降至 $O(n)$ 。这项工作证明了 NVSA 在视觉推理中的有效性和效率，并为结合感知与推理的复杂认知任务提供了新的解决方案。

(2) 科学公式发现

Cornelio 等人提出了 AI Descartes 系统^[80]，将符号回归 (Symbolic Regression, SR) 与逻辑推理相结合，应用于科学公式发现和验证任务。AI Descartes 通过结合数据驱动模型生成与背景理论推导，解决了符号回归在公式科学意义判别中的困难，显著提升了发现公式的准确性和泛化能力。在开普勒第三定律的推导任务中，AI Descartes 从太阳系和系外行星数据中重新发现了描述轨道周期、质量和轨道距离关系的公式，并通过逻辑推理验证其符合牛顿力学背景理论。实验结果显示，AI Descartes 能够生成多个候选公式，并通过数据误差和理论误差筛选出科学意义最强的公式。此外，系统通过逻辑推理模块证明了该公式的可推导性，同时剔除了其他仅符合数据但不满足背景理论的候选公式。这项工作证明了 AI Descartes 在科学公式发现任务中的有效性和可靠性，为结合数据与理论进行自动化科学发现提供了新的解决方案。

(3) 科学建模

Rao 等人提出了物理编码递归卷积神经网络 (Physics-encoded Recurrent Convolutional Neural Network, PeRCNN)^[81]，将神经符号计算的思想应用于复杂时空动力学系统的建模与预测任务。通过将物理知识硬编码至神经网络架构中，PeRCNN 解决了传统数据驱动方法在稀疏数据和噪声数据下的泛化性和准确性问题，显著提升了复杂系统建模的效率和物理一致性。在反应扩散 (Reaction Diffusion, RD) 系统的建模任务中，PeRCNN 被用于模拟描述化学、生物和生态系统动力学行为的偏微分方程。通过将扩散项等已知物理结构直接集成到网络架构中，PeRCNN 高效地从稀疏和低分辨率数据中重建高分辨率的时空动力学，并实现了对

系统未来状态的准确预测。实验结果表明, PeRCNN 在二维和三维反应扩散系统的建模中, 显著优于传统方法。这项工作证明了神经符号计算在科学建模任务中的有效性和普适性, 为结合物理理论和数据驱动方法, 进行复杂系统的自动化建模提供了一种可行的解决方案。

2.1.3 知识蒸馏

知识蒸馏 (Knowledge Distillation, KD), 也被称为教师-学生神经网络学习算法, 是一种有价值的机器学习技术, 旨在提高模型性能和泛化能力。知识蒸馏是一种机器学习模型压缩方法, 它用于将教师模型的知识迁移到学生模型中。KD 背后的核心思想是将教师模型的综合知识转化为更精简、更有效的表示。知识蒸馏具有显著的优势, 例如, 通过传递教师模型的隐含知识, 可以在不显著降低性能的情况下构造轻量化模型, 提升泛化能力, 并使模型更适合在资源受限的设备中部署。此外, 知识蒸馏在多模型集成^{[82][83]}、跨架构知识转移^{[84][85]}以及半监督学习^[86]等领域也有广泛应用。

(1) 黑盒知识蒸馏

黑盒知识蒸馏是一种只能访问教师模型输入输出的蒸馏方法, 无需了解教师模型的内部结构和参数。这种方法主要通过教师模型的输出标签、预测概率分布以及 API 响应等表层特征, 实现知识迁移, 实现相对简单且适用范围广, 特别适合处理商业闭源模型 (如 GPT-4^[87])。黑盒知识蒸馏虽然无法获取模型内部信息, 但由于其灵活性和广泛适用性, 在大语言模型领域得到了广泛应用。目前黑盒知识蒸馏在大语言模型领域主要关注如何有效利用教师模型的输出来优化学生模型^{[88][89]}。

System 1 和 System 2 是心理学家丹尼尔·卡尼曼 (Daniel Kahneman) 提出的两种思维模式^[90]。在深度学习中, “System 1” 类似的操作可以类比于快速、直觉的模型^[91], 如用于图像识别、语音到文本、自然语言理解的模型。这些模型经常是基于大量数据训练得到的, 并且在执行已经学习过的任务时非常迅速和高效。例如卷积神经网络 (Convolutional Neural Networks, CNNs) 在图像分类中迅速识别对象。而 “System 2” 更类似于需要复杂推理、计划和决策的深度学习任务^[92]。在这类任务中, 模型通常需要更长时间进行思考, 可能会结合多种不同的数据源和先验知识。例如强化学习领域中的模型, 通常需要进行长期规划和决策, 类似于人类的战略思考。随着 System 2 推理技术的发展^[93], 研究人员开始探索如何将这些高级推理能力蒸馏到更简单的 System 1 模型中。传统的直接模仿方法在处理复杂推理任务时存在局限性, 因为学生模型难以仅通过模仿输出来学习教师模型的推理过程。这种情况在需要多步推理的任务中尤为明显。

Yu^[2] 等人的研究表明, 通过精心设计的蒸馏策略, 可以将部分 System 2 能力成功迁移到 System 1 模型中。这种方法不需要生成中间推理步骤, 就能在多

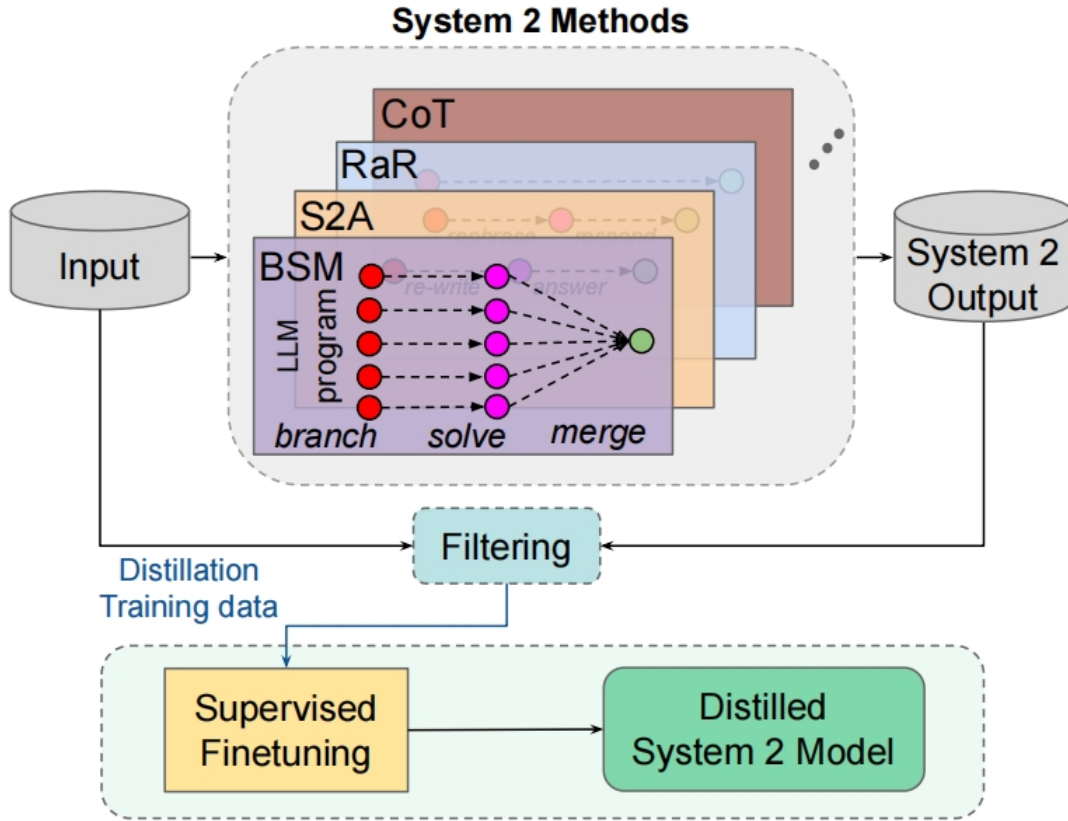


图 2-12 System 2 蒸馏概述^[2]

个任务上达到接近甚至超越 System 2 的性能。实验采用了多个基准数据集，评估了在偏见处理、指令理解和模型评估等任务上的表现。结果显示，这种蒸馏方法能显著提升模型性能，同时保持推理效率。但在复杂数学推理等任务上仍面临挑战，表明黑盒知识蒸馏方法仍有提升空间。未来研究方向包括探索更有效的蒸馏策略，以及将这种方法扩展到持续学习等更广泛的应用场景。

(2) 白盒知识蒸馏

白盒知识蒸馏则是一种可以完全访问教师模型内部结构的蒸馏方法，能够获取并利用中间层特征、注意力信息等内部表示。这种方法通过特征对齐、注意力机制迁移、参数共享等技术，实现更深层次的知识迁移，理论上可以达到更好的效果，但主要适用于开源模型。白盒知识蒸馏虽然实现较为复杂，但能够保留更多模型内部信息，在某些场景下表现更优。

目前白盒知识蒸馏的研究主要集中在小型(参数量 <1B) 语言理解模型上^{[94][95]}，对大语言模型的白盒蒸馏仍处于探索阶段。随着开源大语言模型的兴起，白盒知识蒸馏变得更加重要，因为它可以获得教师模型更多的信号，有助于提升学生模型的性能。传统的知识蒸馏方法使用前向 KL 散度 (Kullback-Leibler Divergence, KLD) 作为目标函数，这在文本生成任务中存在局限性。由于教师模型的输出分布包含了比学生模型更多的模态，使用前向 KLD 会导致学生模型过度估计教师

分布中的低概率区域。这种情况在大语言模型的开放式生成任务中尤为明显。

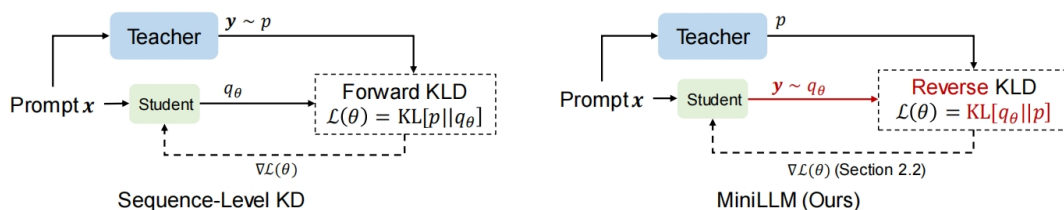


图 2-13 序列级别 KD 与 MINILLM 之间的比较^[3]

Gu 等人^[3]提出了一种白盒知识蒸馏方法，将 LLMs 蒸馏成更小的语言模型。首先将标准 KD 方法中的前向 KLD 目标替换为反向 KLD，这更适合于生成语言模型的 KD，以防止学生模型高估教师分布的低概率区域。然后，推导出一种有效的优化方法来学习这个目标。这些学生模型被称为 MINILLM。该方法可以扩展到具有 1.2 亿到 130 亿参数的不同模型家族。与基线相比，使用反向 KLD 的方法相比传统的前向 KLD 方法，在生成精确度、整体质量、长文本生成等方面都取得了显著提升。特别是在曝光偏差 (exposure bias)、校准 (calibration) 等关键指标上，新方法展现出明显优势。实验还验证了该方法在不同规模模型上的可扩展性，以及在保持生成多样性的同时提升生成质量的能力。

2.1.4 神经符号计算发展趋势

(1) 组合性与推理能力的提升

组合性是人类智能的核心特征之一，也是神经符号计算系统希望实现的关键能力。组合性要求系统能够对所学知识进行分解、重组，并推广到新的推理问题上^[96]。目前，虽然一些神经符号计算系统已经在小规模实验场景^{[97][98]}中实现了组合性推理，但在现实世界的复杂场景中，神经符号计算在这方面的应用仍然非常有限。尤其是如何综合使用多种形式的逻辑（如模态逻辑、时态逻辑、常识逻辑等）进行推理^[99]，以及如何结合不同类型的知识（如陈述性知识、程序性知识、因果知识等）解决实际问题^[100]，仍然是一个重要的研究方向。未来的神经符号计算系统需要在这些方面取得突破，才能更接近人类的推理能力。

(2) 自动化推理与学习的结合

未来的神经符号计算系统将进一步探索如何结合自动化推理与学习，使系统能够在没有大量人工介入的情况下自动生成和优化符号规则。符号推理依赖于高质量的知识库，而许多现有系统在知识获取和融合方面存在局限。特别是，如何从不同的数据源（如结构化数据、知识图谱、文本数据等）中自动化地获取和抽象符号知识^[101]，仍然是该领域的一个重要挑战。自动化知识获取不仅能够减少人类专家的介入，还能够通过反复的推理与学习循环，持续优化知识库的内容。这种自动化过程类似于人类的内省能力^[102]，能够从实际数据中抽象出高层

次的知识，并通过符号推理进行有效的归纳和演绎推理。通过结合自动化推理，神经网络可以在不断接触新数据的过程中自动调整符号规则^[103]，从而提升系统的自适应能力和泛化能力。这种递归式的学习和推理循环将是未来神经符号计算系统的一个重要发展方向。

(3) 新应用场景与挑战性基准

目前，神经符号计算系统的应用主要集中在一些比较小型的任务或实验场景^{[104][105]}中，而在更具挑战性和复杂性的现实任务中，神经符号计算系统的潜力尚未得到充分发挥。未来的研究需要开发更具挑战性的应用场景和基准，推动神经符号计算系统在复杂环境中的应用。例如，社会机器人^[106]、医疗信息学^[107]、科学发现^[108]等领域都是具有巨大潜力的应用场景。这些领域不仅需要处理大量数据和复杂的知识，还需要系统具有发现新科学假说的能力。通过在这些场景中验证神经符号计算系统的推理能力和泛化性能，研究者可以进一步推动该技术的发展。

(4) 神经符号引擎的发展

未来的神经符号计算系统将深入探索符号推理与神经网络之间的深度集成，构建所谓的“类型 6”，也就是完全整合的系统。在这种系统中，符号部分将通过神经网络的反馈不断进化，而神经网络则在符号推理的指导下进行训练。这种集成不仅能够模拟人类的直觉性思维（系统 1）和深度推理（系统 2），还能够解决神经网络系统在推理透明性和泛化能力上的不足。通过这种方式，神经符号计算系统有望实现更接近类人智能的能力，特别是在复杂问题的推理和解决上。

2.2 小结

本章从神经符号集成范式、应用场景与技术趋势三个维度全面梳理了研究基础。通过六类神经符号系统的对比分析，揭示了不同架构在知识表示与推理机制上的本质差异。以 GPT、AlphaGo 系列模型为典型案例，论证了符号推理与神经网络协同增效的可行性。进一步指出知识蒸馏、注意力机制等关键技术对神经符号计算的赋能作用，同时指出现有研究在动态知识更新、复杂逻辑推理等维度的不足，为本研究的创新方向提供了理论支撑。

第3章 面向 DNN 的规则知识融合框架

3.1 概述

随着深度学习在自然语言处理领域的广泛应用，其强大的模式识别能力已成为解决复杂任务的重要工具。然而，深度学习模型通常依赖大规模标注数据，其“黑箱”特性也限制了模型的可解释性和透明性。另一方面，符号规则方法具备逻辑推理的透明性和领域知识的可扩展性，但其对灵活性和复杂性场景的处理能力有限。

为弥补两者的不足，本文提出了一种面向 DNN 的规则知识融合框架。旨在整合符号规则的逻辑推理能力与深度学习的模式识别能力，从而弥补各自方法的不足。框架核心在于整合符号规则的逻辑推理能力与深度学习的模式识别能力，通过双向知识蒸馏和伪标签迭代训练实现两者的协同优化。具体而言，该方法包括以下研究内容和创新点：设计教师-学生架构，符号推理模块与神经计算模块在迭代训练中相互学习，实现显式规则与隐式模式的融合；基于符号规则和神经网络生成高质量伪标签，结合动态权重机制进行多轮迭代优化，提升模型在低资源场景中的表现。

通过实验验证了该框架在命名实体识别和情感分类任务中的有效性，实验结果表明，该方法能够在不同任务中显著提升模型性能。通过引入规则知识指导深度神经网络的训练过程，使模型在进行自然语言处理任务时能够更好地利用领域知识，同时提高其推理的透明度和可解释性。通过这种结合，系统能够在数据驱动学习的基础上，充分利用领域知识进行约束和指导。

3.2 方法介绍

3.2.1 双向知识蒸馏

本研究采用神经符号计算与知识蒸馏相结合的方法，基于神经网络模型构建了一个高效的规则知识融合框架。整个框架分为符号推理与神经计算两个模块，框架结构如图 3-1 所示。

在第一阶段，本研究使用完全符合符号规则的标注数据（数据集 1）对神经网络模型进行训练，构建教师模型。这个模型作为符号推理模块，专注于捕获显式的符号规则和逻辑关系。通过将符号知识嵌入到模型参数中，该模块不仅提供了清晰的推理路径，还能在低资源场景下为后续训练提供可靠的伪标签，有效缓解数据稀缺问题。与此同时使用少量真实标注数据（数据集 2）训练学生模型，构建神经计算模块。这个模块通过对特定任务的适应性调整，优化其对任务

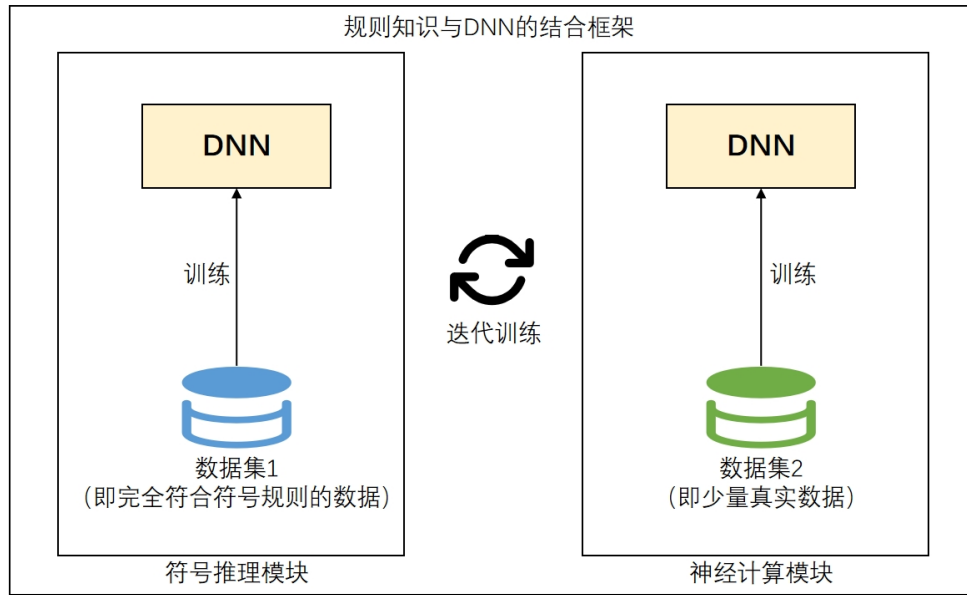


图 3-1 基于神经符号计算的知识融合框架结构

需求的理解。由于真实数据中包含了丰富的语言现象和任务特征，神经计算模块能够从上下文中学习到符号规则无法明确定义的隐式模式，从而补充处理符号推理模块可能遗漏的复杂语言现象。

在第二阶段，本研究利用知识蒸馏技术，设计了双向知识蒸馏机制，实现教师模型和学生模型的迭代训练。一方面，学生模型通过学习教师模型的符号知识输出，将第一阶段教师模型的符号知识迁移到学生模型中，获得规则的指导。这个过程不仅传递了显式的符号规则，还保留了模型对特定任务的适应性。通过知识蒸馏，学生模型能够获得更精简、更有效的知识表示，同时保持较好的泛化能力。另一方面，教师模型也通过学习学生模型在真实数据上的表现，实现规则的软化和调整，使其能够更好地处理规则覆盖范围之外的数据，即真实场景中的复杂语言现象，从而实现了符号知识和经验知识的动态平衡。

这种神经符号知识融合方案具有几个显著优势：首先，它有效结合了符号规则的可解释性和神经网络的自适应能力；其次，通过分离符号学习和任务适应两个阶段，使得模型能够更好地平衡规则约束和任务需求；最后，双向知识蒸馏机制不仅实现了符号知识向经验模型的软迁移，还通过经验反馈优化了符号规则的应用范围，为低资源场景下的模型部署提供了更灵活的解决方案，既保证了模型性能，又提高了模型效率。

3.2.2 伪标签迭代训练

在初步微调后，两个模块将共同对无标注数据集进行伪标签标注。符号推理模块根据学习到的符号规则，对无标注数据集进行推理标注，生成符号推理伪标签。这些标签是基于符号逻辑推导得出的，符合规则的数据将被明确标注。神

经计算模块则通过其隐式模式学习能力，对同样的无标注数据集生成神经计算伪标签。这些标签是基于神经网络对上下文和隐式模式的理解生成的，能处理符号规则未涵盖的复杂或模糊模式。

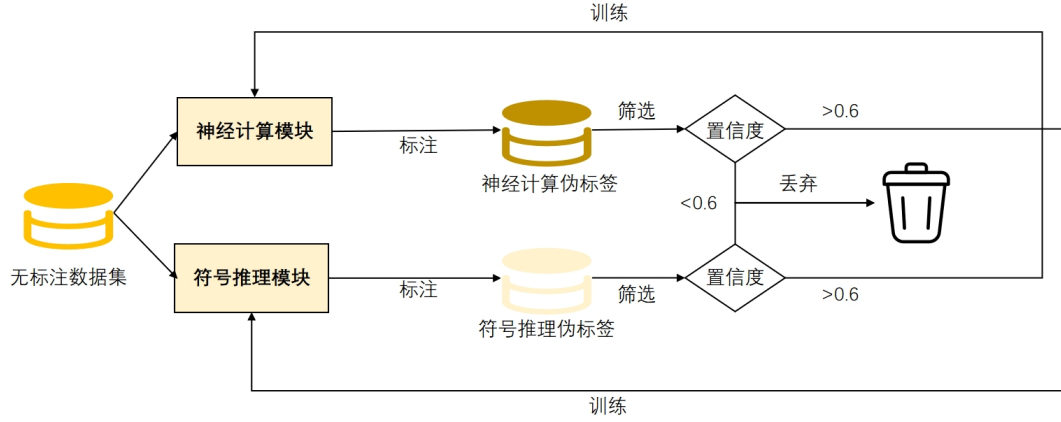


图 3-2 模块迭代训练过程

两个模块在生成伪标签的同时，会计算每个标签的置信度。置信度反映了模型对标注结果的信任程度。为确保伪标签的质量，当伪标签的置信度低于某个设定的阈值时，系统会直接剔除这些标签，以避免低质量标注对模型训练的负面影响。置信度筛选后的伪标签将被保留并用于模型的进一步训练。

在基于教师模型伪标签训练学生模型时，可以引入基于置信度的动态权重机制来提升训练效果。基础权重采用 *Sigmoid* 函数计算：

$$w_{base} = \frac{1}{1 + \exp(-k(p_t - \theta))} \quad (3-1)$$

其中 k 控制权重曲线的陡峭程度， θ 为置信度阈值。相比线性权重，*Sigmoid* 形式的优势在于：在阈值 θ 附近有更平滑的过渡，通过参数 k 可以灵活调整权重变化的剧烈程度，自然地将权重限制在 $[0,1]$ 范围内并且对极端值有更好的处理能力。通过引入温度缩放机制，可以进一步调整权重分布：

$$w_{temp} = p_t^\tau \quad (3-2)$$

其中 τ 是温度系数。在考虑训练轮次时，引入 warmup 因子 α ，使模型在训练初期更关注高置信度样本：

$$\alpha = \min(1, \frac{epoch}{warmup_epochs}) \quad (3-3)$$

为了综合考虑教师和学生模型的一致性，引入 *KL* 散度度量两个模型输出分布的差异：

$$KL(p_s || p_t) = \sum p_s \log(\frac{p_s}{p_t}) \quad (3-4)$$

最终的样本权重为：

$$w_{final} = \alpha \cdot w_{base} \cdot p_t^\tau \cdot \exp(-KL(p_s || p_t)) \cdot \mathbb{1}\{p_t = p_s\} \quad (3-5)$$

其中 $\mathbb{1}\{\cdot\}$ 为指示函数，表示两个模型预测是否一致。在训练过程中，加权交叉熵损失为：

$$L = - \sum w_{final} \cdot \sum y_t \log(p_s) \quad (3-6)$$

其中 y_t 为教师模型的伪标签。这种基于 Sigmoid 的动态权重机制能更平滑地区分不同置信度样本的重要性，确保模型在训练时能够集中精力修正不确定的样本，同时减少对低置信度样本的学习。这样可以提高训练效率，并逐步提升模型对复杂样本的处理能力，有效降低错误伪标签的负面影响，提升半监督学习的效果。

符号推理模块通过学习神经计算模块生成的伪标签，逐渐“软化”其最初的严苛规则，吸收一些隐式模式。例如，某些未被规则明确涵盖的实体组合（如某些特殊的组织名称）可以通过神经计算模块的标注反馈被符号推理模块采纳和调整，从而提高其适应性。神经计算模块则通过学习符号推理模块生成的规则标签，进一步增强其对符号规则的理解。尤其是在少量标注数据的情况下，神经计算模块可以从符号推理模块的显式规则中提取到更多明确的知识，提升其对低资源数据的泛化能力。

模型的训练过程通过多轮迭代进行。每一轮迭代中，符号推理模块和神经计算模块分别更新其模型参数，基于伪标签进行训练。随着迭代的进行，模型之间的协同学习能力逐渐增强。符号推理模块的规则适应性提升，并逐步进行规则调整；神经计算模块对复杂数据的处理能力也逐渐增强，同时增强其对低资源场景的适应性。

3.3 实验

3.3.1 数据集

实验所需的数据集如图 3-3 所示，主要分为符号规则数据集和不符合规则数据集两部分。符号规则数据集基于预定义的符号规则生成或标注，通常包括领域特定的逻辑规则或知识库中的显性知识。以命名实体识别任务为例，设定其中的数据均遵循规则“地点 + 大学 = 组织实体”，例如“北京大学”、“天津大学”等。而不符合规则数据集则包含一部分未完全遵循符号规则的数据，涉及一些偏离规则或未按预期模式出现的数据，例如“北京电视台”、“天津博物馆”等。具体的数据集构建如下：

数据集 1：此数据集由符号规则数据集中提取，完全符合符号规则，并经过

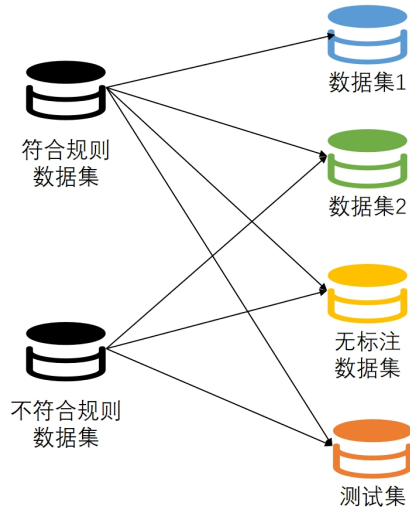


图 3-3 实验所需数据集及构成

标注。它主要用于对符号推理大模型进行微调，确保模型能够在符号规则的约束下进行准确的推理和标注。

数据集 2：该数据集包含符合规则和不符合规则的数据，但数据量较小，用于模拟低资源场景下的真实情况。

无标注数据集：未经过标注的数据集，包含符合规则和不符合规则的混合数据。此数据集将用于伪标签生成，通过符号推理和神经网络模型的协同训练，生成伪标签并用于进一步的迭代训练。

测试集：用于最终评估模型的性能，测试集同样包含符合规则和不符合规则的数据。通过此数据集，可以全面考察模型的泛化能力，尤其是面对不规则数据时的处理能力和表现。

3.3.2 实验结果

为了验证提出的规则知识融合框架在基础 DNN 模型上的有效性，本研究分别在命名实体识别任务和情感分类任务上对 BiLSTM+CRF 模型进行了实验评估。

3.3.2.1 命名实体识别任务

对于命名实体识别任务，由于目前的公开数据集无法满足完全符合规则的数据，所以采用人工 + 大模型生成的方式自行生成了一批数据集。该数据集信息如表3-1所示。

可以看到，对于规范集，也就是符合规则的数据集，其中包含 284 条严格遵循预设规则的结构化数据，其核心特征体现在实体构建的高度规范性。采用“地名 + 大学”的固定范式（如“广州大学”、“厦门大学”），要求地名需为国家行政区划内真实存在的地名，大学名称需为标准后缀。这种设计确保模型能准确识

表 3-1 命名实体识别任务数据集信息

数据集类型	数据条数	规则知识	例子
规范集	284	1、组织实体 = 地名 + 大学 2、人物实体 = 姓 + 老师	尽管深圳的天气炎热，但 陈老师 依然坚持每周前往 广州大学 进行学术交流，同时与来自 厦门大学 的 林老师 共同筹备明年在成都举办的国际研讨会。
泛化集	372	-	在 南京大学 举办的学术论坛上， 李老师 与来自 上海研究所 的 王教授 共同探讨了人工智能发展趋势，而 刘副院长 则指出 清华大学 近年来在量子计算领域取得的突破， 张华博士 对此表示赞同并补充了相关研究成果。

别”成都大学”与”北京大学”等合法实体，同时排除“科技学院”或“研究所”等非标准命名机构。同时限定为“姓 + 老师”的命名格式（如“陈老师”、“林老师”），严格排除“教授”、“博士”等职称变体。该数据集通过规范化的实体组合与事件场景，为模型建立清晰的知识边界。例如在“每周前往广州大学进行学术交流”的语料中，模型不仅能提取“广州大学”作为组织实体，还能通过“陈老师”的固定称谓建立人物-机构关联，有效训练实体识别与关系抽取的双重能力。

而对于泛化集，也就是不符合规则的数据集，其中包含 372 条数据，采用规则与反规则混合的设计策略。保留约 30% 符合规则的数据作为基础参照，同时设置反规则干扰项，包括三类典型干扰模式：（1）职称变异型：将“老师”替换为其他职称（如“王教授”、“李研究员”），打破命名规则的同时保持学术属性，测试模型对职称后缀的敏感度。（2）机构混杂型：混用“研究所”（如“上海研究所”）、“学院”（如“刘晨院校”）等非标准机构名称，其中”刘晨院校”更通过姓氏前置制造识别歧义。（3）毫不相关型：设置一些与之前规则毫不相关的实体，例如“中央电视台”、“天津博物馆”、“雷军”、“刘医生”等干扰项。此类设计显著提升数据集的现实仿真度。表里提到的例句中，模型需同时处理非标准职称（教授）与非规范机构（研究所）的双重干扰，有效检验其规则泛化能力。而”刘晨院校”这类混淆姓氏与机构名的特殊案例，则能深度测试实体边界识别算法的

鲁棒性。

规范集通过高纯度、强一致性的结构化样本，为模型建立精确的规则认知基础；而泛化集则通过规则与反规则样本的对抗性组合，构建多层级测试环境。这种设计有效模拟真实场景中规则明确性与语言变异性的矛盾，使实验能够系统评估模型从强规则到弱规则场景的知识迁移效果，为验证模型是否真正理解规则本质，而非简单地模式匹配，提供可靠评估基准。

为保证实验公平性，基准模型（baseline）采用的训练数据集为符号推理模块与神经计算模块训练数据集的合集（数据集 1+ 数据集 2），并在伪标签迭代训练阶段额外增加一轮训练以平衡训练轮次差异。实验结果如表3-2所示。

表 3-2 命名实体识别任务实验结果

准确率	初始训练	第一轮迭代	第二轮迭代	最终提升
baseline	27.4%	29.6	28.6	1.2%
符号推理模块	6.0%	17.9%	25.4%	19.4%
神经计算模块	26.9%	28.4%	32.8%	5.9%

实验数据显示，baseline 模型在第一轮训练后准确率从初始值提升至 29.6%，但第二轮后下降至 28.6%，这种先升后降的趋势可能源于模型在持续训练中对特定训练样本的过拟合，或是训练数据中噪声干扰导致的性能波动，反映出单一模型在复杂规则场景下的稳定性局限。

符号推理模块的初始表现仅 6.0%，这与其规则学习机制的特性密切相关。由于训练集包含的都是严格符合规则的数据，模块初期建立的规则体系难以应对测试集中大量存在的职称变异（如”教授”替代”老师”）、机构混杂（如”上海研究所”）等反规则样本。但通过神经计算模块提供的伪标签迭代训练，其准确率最终显著提升至 25.4%，验证了该模块具备规则泛化能力——不仅能识别”陈老师”等标准称谓，还能将”姓 + 职位”的核心规则迁移至”王教授”等变体形式，成功突破初始训练集的刚性限制。

神经计算模块则展现出不同的学习特性，初始准确率即达 26.9%。这得益于其通过少量适应性训练快速掌握基础实体特征的能力，即便在极小数据集下也能有效捕捉”大学”后缀等显式模式。当引入符号推理模块的规则后，其准确率进一步攀升至 32.8%，特别是在处理”广州大学”等完全符合规则的数据时，模块通过融合”地名 + 大学”的显式规则，显著提升了机构实体的识别精度。但面对”上海研究所”等非标准机构时，性能提升幅度相对有限，揭示出规则应用的范围边界。

实验结果表明，双模块协同框架有效解决了单一方法的局限性：符号推理模块通过规则泛化突破训练数据刚性约束，神经计算模块则通过模式学习增强基

础识别能力。二者的协同作用在混合复杂度数据测试中尤为显著，模型不仅能准确识别标准实体（如”林老师”），还能处理职称变异样本（如”李研究员”），较基准模型提升超过 4 个百分点。这种性能提升验证了框架在真实场景中的实用价值，特别是在处理规则与反规则交织的复杂语言现象时，展现出更强的鲁棒性和适应性。

3.3.2.2 情感分类任务

对于情感分类任务，本研究的数据集来源于 SST-2 (Stanford Sentiment Treebank 2)、MR (Movie Review) 和 CR (Customer Review) 三个经典数据集的抽样组合，具体组成如下：1000 条 SST-2 数据，来源于电影评论句子，标注为正面/负面情感的二分类数据集。此数据集聚焦句子级别情感分析，是 GLUE 基准任务之一，适用于模型对复杂语义的捕捉能力验证。1000 条 MR 数据，来源于单句电影评论，标注为正面/负面，常用于对比不同模型在短文本情感分类上的性能差异。以及 1000 条 CR 数据，来源于电子产品、书籍等用户评论，标注为正面/负面。反映了实际场景中的客户情感表达，评论语言更口语化。组合数据集通过抽样平衡不同数据集的分布（如情感比例、文本长度），减少单一数据集的偏差。且基于经典数据集构建，便于与已有研究结果对比，验证模型性能。通过融合这三个数据集，情感分类模型能够更全面地捕捉从专业影评到日常用户反馈的多维度情感特征。最终组合成的数据集信息如表3-3所示。

表 3-3 情感分类数据集信息

数据集类型	数据条数	规则知识	例子
规范集	864	一句话中 ‘but’ 后面的情感为这句话的真正情感	The concept is creative, but the execution feels rushed and unsatisfying .
泛化集	2136	-	This heartwarming story captures the essence of childhood friendship perfectly .

可以看到，对于规范集，其中包含 864 条严格遵循”but” 后置情感判断准则的结构化数据，其核心特征体现在情感推断的逻辑确定性。采用 “前半句描述+but+ 情感关键词” 的固定句式（如 “The concept is creative, but the execution feels rushed and unsatisfying”），要求 but 之后的情感表达必须为整句的真实情感标注。这种设计确保模型能精准捕捉转折词后的语义重心，在所有的语料中，虽然前半段包含情感描述，但通过 but 引导的情感判定规则，模型得以建立 “后置情感主

导”的规则框架。该数据集通过句式结构的强规律性，为模型构建明确的情感判断路径，有效提升转折语境下的情感识别准确率。

而对于泛化集，其中包含 2136 条数据，采用无固定语法结构的自然语言表达。此类数据完全摒弃句式规则约束，通过三种典型特征构建真实语言场景：(1) 情感一致性表达：如 “This heartwarming story captures the essence of childhood friendship perfectly” 等整句情感统一的表述，消除转折词带来的判断依赖。(2) 隐含情感传递：采用比喻、反讽等修辞手法（如 “The service was as fast as a sleeping turtle”），需要结合语境进行情感推断。(3) 情感混淆项干扰：刻意构造具有情感误导性的句式结构，but 后的情感表达与句子真实情感相悖。例如在 “The movie was really disappointing and confusing, but the plot had some stunning twists” 中，虽然 but 后使用 “stunning” 这一正面表述，实际整句的真实情感是通过 “really disappointing and confusing” 表现的，本质上仍属负面评价，此时模型需穿透语言表象识别真实情感指向。此类混淆项旨在检验模型对语义重心的准确判断能力。这种设计升级后，数据集包含三重挑战梯度：情感一致性表达验证基础理解能力，隐含情感传递测试语境推理水平，情感混淆项干扰则专门针对转折结构中的情感伪装现象，形成从显性到隐性、从单一到复杂的渐进式验证体系。

规范集通过句式模板的高度可控性，为模型建立基于语言结构的强规则认知；而泛化集则通过自然语言的多变性与复杂性，构建接近真实场景的弱规则环境。这种双重设计既验证模型对显性规则的掌握程度，又考察其对隐性语义的理解深度，为评估情感分析模型从结构化规则到非结构化语境的知识迁移能力，提供多维度的验证基准。

同样地，为保证实验公平性，基准模型（baseline）采用与符号推理模块、神经计算模块完全相同的训练数据集（数据集 1+ 数据集 2），并在伪标签迭代训练阶段额外增加一轮训练以平衡训练轮次差异。实验结果如表3-4所示。

表 3-4 情感分类实验结果

准确率	初始训练	第一轮迭代	第二轮迭代	最终提升
baseline	83.9%	85.6	85.7	1.9%
符号推理模块	78.3%	79.0%	82.7%	4.4%
神经计算模块	81.1%	85.3%	89.3%	8.2%

实验数据显示，baseline 模型在第一轮迭代中准确率从初始 83.9% 提升至 85.6%，但第二轮仅微增至 85.7%，这种边际效益递减现象可能源于浅层特征学习饱和，或是训练数据中情感表达的多义性引发的模型判断模糊，凸显单一模型在复杂情感语义捕捉中的瓶颈。

符号推理模块的初始准确率为 78.3%，显著低于 baseline 的 83.9%，这与其

基于逻辑规则的情感分析机制直接相关。面对文本中隐式情感表达和情感混淆项等数据时，初期构建的规则体系难以有效解析语义冲突。但通过两轮迭代训练，准确率提升至 82.7%，验证了该模块的规则扩展能力，不仅能识别显性情感词，还能通过上下文逻辑推理捕捉隐性情感线索，突破刚性规则框架。

神经计算模块展现出显著的学习优势，初始准确率达 81.1%。这归功于其通过深度网络自动提取文本深层语义特征的能力，即使面对“cry with joy”等矛盾修辞也能有效建模。经过迭代训练后准确率提升至 89.3%，展现出显著的规则学习能力，在符合符号推理规范的数据场景中表现突出。然而，这种规则强化学习也带来隐性的能力偏移——在一些模糊干扰项数据上（例如“The product didn’t work as expected, but at least it arrived on time”），模块进行了错误的识别，这表明过度规则化导致模型对非常规表达式的特征敏感性降低，揭示出神经模型在规则迁移过程中的能力边界。

实验结果表明，符号推理模块通过学习神经计算模块突破了隐式情感识别瓶颈拓，软化了逻辑规则的应用，而神经计算模块依托规则知识增强了基础情感解析能力。二者的协同作用使模型在混合型情感测试集中准确率最高提升 8.2%，较基准模型提升近 4 个百分点。这种突破验证了融合框架在处理非结构化情感数据时的独特优势，为多维度情感分析提供了新的技术路径。

同时注意到，情感分类任务与命名实体识别任务的性能表现存在较大差异，情感分类任务取得了相对较好的结果，而命名实体识别任务的整体准确率仍然较低。分析其主要原因：一是两个任务的数据规模差异较大，情感分类任务的训练数据更加充足；二是因为情感分类任务相对来说比较简单，句式结构并不复杂，情感表达多集中在显性情感词（如“happy”、“disappointing”）或短句式（“The product is awesome”），相比实体识别中嵌套结构更易规则建模；三是两个任务的规则复杂度不同，情感分类的规则相对简单且具有较强的判别性，这也说明数据规模和规则特性会显著影响模型的学习效果。但两个任务的实验结果都表明符号推理模块和神经计算模块在训练过程中均有不同程度的准确率提升，证明了所提出的框架在不同 NLP 任务中都能够有效提升模型性能，验证了方案的可行性。

基于当前的实验结果，本研究计划在以下方向进行扩展：(1) 规则复杂度扩展：支持多规则协同约束，同时处理规则之间可能存在的冲突，以及实现规则的自动权重调整。(2) 任务领域扩展：在更多 NLP 任务中进行验证，探索框架的通用性。通过这些扩展研究，期望能够构建一个更通用和强大的规则知识与 DNN 的结合框架，为低资源场景下的 NLP 任务提供有效解决方案。

3.4 小结

本章首创性地提出”双向知识蒸馏 + 伪标签迭代训练”的规则知识融合框架，攻克了低资源场景下模型可解释性与泛化能力难以兼得的难题。通过规则知识指导的标签修正机制，在命名实体识别任务和情感分类任务中都有效提高了模型的准确率，且最优表现（神经计算模块）均优于 **baseline**。实验证明，该框架能有效缓解标注数据稀缺导致的低资源场景下模型性能较差的问题，其伪标签置信度阈值动态调整策略显著优于传统半监督学习方法。

第4章 面向 Transformer 的规则知识融合框架

4.1 概述

尽管面向 DNN 的规则知识融合框架在自然语言处理任务中展现了强大的表现能力，但其也存在一些明显的局限性。例如，传统 DNN 架构通常需要针对不同任务进行架构调整，这不仅增加了开发和部署的成本，还限制了其在复杂任务中的表现能力。此外，对于复杂的任务，DNN 基于局部特征学习的简单网络结构可能表现不足，难以充分捕捉任务中的全局语义和复杂逻辑。为了解决这些问题，本文提出了一种面向 Transformer 的规则知识融合框架，旨在融合规则的逻辑推理能力与 Transformer 模型的强表征能力。这一框架通过规则知识嵌入到 Transformer 的训练过程中，充分发挥规则知识的逻辑推理能力与 Transformer 的强表征能力。不仅实现了规则引导的逻辑约束，还无需修改模型核心架构，使其具备更高的灵活性和适应性。

具体而言，该框架包括以下研究内容与创新点：设计结合规则知识的 Transformer 方法，提出规则注意力掩码的生成与应用机制，通过显式规则约束增强模型的全局语义捕获能力；在训练过程中，结合双向知识蒸馏与伪标签迭代优化策略，提升模型性能。

实验结果表明，该方法能够在多种任务上显著提升模型性能，验证了其在命名实体识别任务中的有效性。通过引入规则知识指导 Transformer 的训练过程，模型不仅能够更好地利用领域知识，还提高了其推理的透明性与可解释性，从而实现数据驱动学习与规则逻辑推理的有机结合。

4.2 方法介绍

4.2.1 注意力掩码生成

Transformer 模型本身拥有灵活的架构设计，其多头自注意力机制能够有效捕获输入数据的全局依赖关系。规则知识可以直接作为输入特征、注意力引导信号或后处理约束，无需对模型的核心结构进行任何修改，从而使其能够快速适应不同的 NLP 任务。这种特性显著降低了开发和迁移成本，尤其适合任务多样化的应用场景。与 DNN 的局部特征学习不同，Transformer 的自注意力机制能够建模数据中的长程依赖关系，使其在处理复杂任务时表现出色。在结合规则知识时，Transformer 能够更高效地利用规则指导信息，从而提升模型对复杂语义和逻辑关系的理解能力。例如，在文本生成任务中，规则知识可以引导注意力机制聚焦于与任务相关的上下文信息，生成结果更加符合逻辑。

为了实现规则对注意力分布的引导，本框架首先根据输入 token 的类型构建规则掩码矩阵 M 。规则掩码矩阵的计算如下：

$$M_{ij} = \begin{cases} \alpha & \text{如果 } token_j \text{ 是 } token_i \text{ 依赖的规则相关类型} \\ \beta & \text{如果 } token_j \text{ 和 } token_i \text{ 属于相同规则组件} \\ \gamma & \text{其他情况} \end{cases} \quad (4-1)$$

其中 $\alpha > \beta > \gamma$ 为预设的掩码权重值。

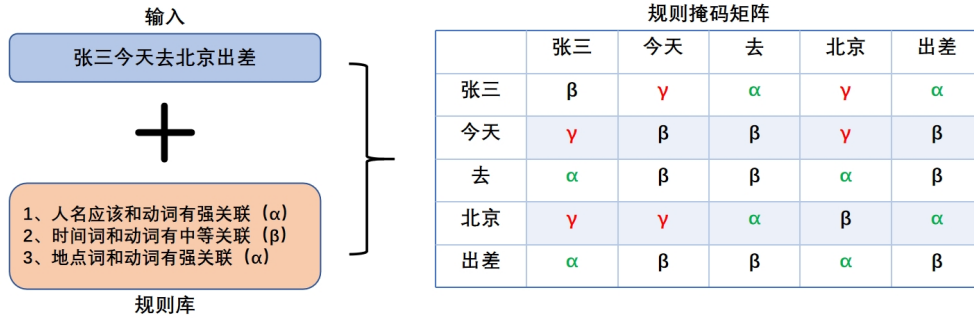


图 4-1 注意力掩码矩阵示例

如图 5-1，假设我们想生成对“张三今天去北京出差”这句话的理解，在添加注意力掩码后：(1)“张三”会更多地关注“去”和“出差”这些动词；(2)“北京”会更多地关注“去”这个动作；(3)“今天”会适度关注所有动词。本研究希望通过基于规则的注意力掩码，让模型知道“张三”是动作执行者、“去”和“出差”是主要动作以及“北京”是动作目标地点等知识。当在回答“谁去北京？”这个问题时，模型会优先关注人名（“张三”）和动词（“去”）之间的关系。因为掩码矩阵中这对 token 的权重是 α (高权重)，从而更容易得出正确答案：“张三”。

通过将规则掩码矩阵与原始注意力分数相结合：

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \log(M)\right) \quad (4-2)$$

使用 $\log(M)$ 能将大的掩码权重压缩到合理范围，并且通过 $\log$ 将掩码的乘法效果转换为加法，便于调节规则影响力度， α 、 β 、 γ 的小幅变化会导致最终注意力分数的合理变化，不会完全压制原始注意力分数。这种设计确保了规则相关的 token 对之间能够建立更强的注意力连接。

最终的注意力计算将规则掩码和相关性评分整合在一起：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \log(M)\right)V \quad (4-3)$$

Transformer 的注意力机制本质上是数据驱动的，其注意力分布是通过大规模训练数据学到的，缺乏明确的解释路径。而规则知识的引入可以为注意力分布提供一种人为可控的逻辑解释。比如，通过规则掩码矩阵，我们可以清晰地看到“张三”（执行者）和“去”（动作）之间的高权重注意力连接是由规则指定的，而不是模型的“黑箱”行为。这种可解释性在某些高风险任务（如医疗诊断、法律推理）中尤为重要。同时，尽管 Transformer 理论上能够建模全局依赖，但在实践中，由于注意力分布权重的稀疏性，模型可能会难以捕获某些需要显式连接的长程依赖关系。在长文本中，模型可能无法很好地捕获句首和句尾之间的依赖关系。通过规则掩码，可以显式建立这种连接，增强模型对复杂逻辑的理解。

基于以上优点，规则知识与 Transformer 的结合框架通过注意力机制注入规则，无需修改模型架构，易于迁移到不同任务，克服了 DNN 框架在架构修改和复杂任务适应性方面的不足。提供了清晰的解释路径，便于理解模型决策过程。后续的研究在 NLP 任务上对该框架进行了实验评估，验证了其有效性与适用性。通过后续的实验验证，进一步明确了规则知识与 Transformer 结合框架的优势，并探索如何在不同场景中优化规则知识的引入方式，实现更高效的智能系统设计。

4.3 实验结果

为了验证提出的规则知识融合框架在基础 Transformer 架构模型上的有效性，本研究分别在命名实体识别任务和文本摘要任务上对 BERT-Base-Chinese 模型进行了实验评估。BERT-Base-Chinese 是基于 BERT 架构的中文预训练语言模型，专为中文自然语言处理任务设计。该模型基于 Transformer 编码器构建，采用双向上下文理解机制，能够捕捉词语在句子中的双向语义依赖关系。其结构包含约 1.1 亿参数，由 12 层 Transformer 编码器堆叠而成，每层配备 768 维隐藏层表征空间和 12 个自注意力头，通过大规模中文语料（涵盖简繁文本）的预训练学习词汇、语法及语义的深层特征。该模型能够学习到中文词汇、语法及语义的深层特征，支持对未标注文本的通用表示，适用于文本分类、命名实体识别、问答系统等多种下游任务，可直接通过预训练权重进行微调或作为特征提取器使用。通过以上特性，BERT-Base-Chinese 成为中文 NLP 领域的基础工具之一，广泛应用于工业界和学术界。同时，除了使用到第三章中的双向知识蒸馏和伪标签迭代训练技术之外，本研究还引入了基于规则的注意力掩码机制，并将其与 Transformer 的多头自注意力机制相结合，进行了对比实验。

4.3.1 命名实体识别任务

对于命名实体识别任务，本研究使用了第三章中创建的数据集，数据集具体信息如表3-1所示。为保证实验公平性，基准模型（baseline）采用的训练数据集

表 4-1 命名实体识别实验结果

准确率	初始训练	最终迭代结果	最终提升
baseline	89.7%	91.3%	1.6%
符号推理模块	79.6%	87.3%	7.7%
神经计算模块	88.2%	93.4%	5.2%

为符号推理模块与神经计算模块训练数据集的合集（数据集 1+ 数据集 2），并同样在伪标签迭代训练阶段额外增加一轮训练以平衡训练轮次差异。但由于第三章中所使用的 BiLSTM+CRF 模型结构较为简单，其训练过程仅通过两轮迭代性能便达到峰值，而 BERT-Base-Chinese 模型因结构复杂度显著提升，需进行多轮迭代训练才能充分收敛，因此仅展示初始测评结果和最终迭代后的测评结果。实验结果如表4-1所示。

实验数据显示，命名实体识别任务整体准确率较第三章有显著提升，这主要源于模型架构的升级——当前使用的 BERT-Base-Chinese 通过深层双向注意力机制，较之相对简单的 BiLSTM+CRF 模型具备更强的上下文建模能力。

baseline 模型在初始训练后准确率为 89.7%，最终经过多轮迭代训练后提升至 91.3%，提升值仅为 1.6%。这种微弱的增益表明模型在初始阶段已接近性能饱和区，架构优势使得模型早期就捕获了主要特征规律，导致后续优化空间极为有限。值得注意的是，模型在迭代训练中同样出现了准确率从回落的下降趋势，与第三章实验中观察到的现象相同，证实模型在训练后期发生了典型的过拟合。

符号推理模块的初始准确率为 79.6%，通过神经计算模块提供的伪标签迭代训练，最终准确率提升至 87.3%，提升了 7.7 个百分点，相对提升率 9.7%。这种提升验证了符号推理模块能够通过伪标签数据学习符合真实场景的标注模式，基于现实数据分布对初始规则进行适应性调整。值得注意的是，该提升幅度显著低于第三章中 BiLSTM+CRF 模型符号推理模块的 323% 相对提升率（从 6.0% 提升至 19.6%）。经分析，可能归因于当前模型已具备较好的预训练基础，即使在初始阶段仅使用符合规则的数据进行训练，其预训练获得的知识仍能有效处理测试集中的部分非规则实体。与使用完整数据集的基线模型相比，当前模型的初始准确率仍存在差距。然而通过神经计算模块的伪标签迭代训练后，模型准确率最终可逼近完整数据集训练的 baseline 水平。这一现象具有重要启示：尽管符号推理模块仅经历两个阶段的有限训练（初始阶段的规则数据训练 + 后续无标注数据训练），其性能仍能接近基于完整数据的 baseline 模型，有力验证了本框架在低资源场景下的有效性——通过结合符号推理与神经计算的协同训练机制，能够显著缓解数据稀缺问题，为模型提供稳健的性能保障。

表 4-2 命名实体任务识别消融实验结果（移除注意力掩码生成模块后）

准确率	初始训练	最终迭代结果	最终提升
符号推理模块	77.3%	85.1%	7.8%
神经计算模块	86.6%	92.1%	5.5%

神经计算模块凭借预训练语言模型的优势，初始准确率即达 88.2%，这主要受益于 BERT-Base-Chinese 通过大规模预训练获得的实体特征先验知识，使其在少量适应性训练后即可实现细粒度实体捕捉。当融合符号推理模块的规则约束后，准确率进一步提升至 93.4%，提升了 5.2 个百分点。尽管提升幅度低于符号推理模块的 7.7 个百分点，但揭示了神经计算模块具有规则学习能力：通过语义空间与规则空间的协同优化，有效整合如“地名”与“大学”的强关联规则，显著提升规范实体的识别精度。然而，面对非标准实体时，性能提升幅度受限，凸显出规则引导的范围边界。

实验结果表明，本研究提出的规则知识框架在命名实体识别任务中展现出显著优势。两个模块的性能演化印证了规则知识融合的核心价值——通过符号推理的可解释性与神经计算的泛化能力形成动态平衡，在保持较高水平性能的同时，降低了数据依赖度，相较大量数据集训练，仅需少部分标注数据，为低资源场景下的实体识别提供了新的方法论框架，对后续研究中的模型架构设计具有重要启示意义。

与此同时，为了检验本章中提出的注意力掩码生成方法的有效性，本研究还进行了消融实验，实验结果如表4-2所示。

可以看到，在不使用注意力掩码生成时，符号推理模块的初始训练准确率从 79.6% 下降到了 77.3%，最终迭代结果准确率从 87.3% 下降到了 85.1%，神经计算模块的初始训练准确率从 88.2% 下降到了 86.6%，最终迭代结果准确率从 93.4% 下降到了 92.1%，下降范围在 1.3% 到 2.3% 不等，有效地验证了本研究提出的注意力掩码生成方法的有效性。同时注意到，即使不使用注意力掩码生成，符号推理模块和神经计算模块的性能提升（7.8% 和 5.3%）也与之前相差无几，验证了本研究规则知识融合框架的稳定性。而神经计算模块的准确率也仍然高于 baseline 的 91.3%，进一步验证了规则知识融合框架的有效性。

消融实验的深度分析揭示出注意力掩码生成机制在模型性能优化中的作用，该方法通过特征选择显著提升模型表现。移除后符号推理模块的初始训练准确率从 79.6% 下降到了 77.3%，，最终迭代结果准确率从 87.3% 下降到了 85.1%，神经计算模块的初始训练准确率从 88.2% 下降到了 86.6%，最终迭代结果准确率从 93.4% 下降到了 92.1%，下降范围在 1.3% 到 2.3% 不等，这种系统性的性能衰减

验证了其噪声抑制与特征增强的核心价值。

同时注意到，在缺失注意力掩码的条件下，符号推理模块仍保持 7.8% 的性能提升（与完整架构的 7.7% 差距仅 0.1%），神经计算模块则保持了 5.5% 的性能提升（较完整架构 5.2% 高出 0.3%），与之前相差无几，验证了本研究规则知识融合框架的稳定性。而神经计算模块的准确率（92.1%）也仍然高于 baseline（91.3%），进一步验证了规则知识融合框架的有效性，凸显了双模块协同架构的鲁棒性，为后续构建可扩展的模块化系统提供了理论依据。

4.3.2 文本摘要任务

对于文本摘要任务，本研究构建的多维度文本摘要数据集融合了两大权威语料库的抽样组合，具体构成如下：1000 条 LCSTS（Large Scale Chinese Short Text Summarization）样本，属于社交媒体短文本，源自新浪微博用户原创内容，包含人工标注的摘要-正文对。文本平均长度 35 字符，涵盖表情符号、网络用语等非正式表达，用于验证模型对口语化文本的语义提炼能力。2000 条 CAIL2022（China AI and Law Challenge Dataset 2022）涉法舆情摘要任务的数据样本，属于法律长文档摘要。文档平均长度 500 字（含案情陈述、法律条款等结构化内容），标注人员依据《最高人民法院文书规范》提取关键判决要素，支持长文本信息密度评估。组合数据集在长度方面平衡了短文本（0-50 字）、中长文本（51-200 字）和长文本（ ≥ 201 字）的数据分布，在领域维度包含了社交媒体内容、专业新闻以及法律文书方面内容，覆盖日常交流到专业写作场景。采用分层抽样策略，有效降低了单一数据源的分布偏差。且继承经典数据集标注标准，确保了横向可比性。该数据集通过有机整合不同语域特征，形成了包含语言复杂度、信息密度、领域适应性的多维度评估框架，能够全面检验文本摘要模型的泛化性能。最终组合成的数据集信息如表4-3所示。

可以看到，对于规范集，其中包含 582 条严格遵循预设规则的数据，其核心特征体现在抽取摘要的高度规范性，要求其必须包含原文的首句和尾句。这种设计基于语言学理论中首尾句通常承载核心信息的认知规律，为模型建立明确的规则边界。

而对于泛化集，其中包含 2418 条数据，采用规则与反规则混合的设计策略。保留约 10% 规范样本作为锚点，其余约 90% 数据通过反规则设计突破首尾句限制。这种对抗性结构模拟了真实语言场景中的规则变异，有效检测模型从强规则到弱规则的泛化能力。

两数据集的功能定位形成认知递进架构：规范集通过首尾句强制约束形成高纯度样本空间，构建规则的形式化表征基础；泛化集则采用规则密度梯度设计，构建对抗性语言变异场域。这种双重结构不仅实现从显式规则内化到隐式知识迁移的能力跃升检测，更重要的是通过规则-反规则样本的动态博弈，迫使模

表 4-3 文本摘要任务数据集信息

数据集类型	数据条数	规则知识	例子
规范集	582	摘要必须包含文章的首句和尾句内容	“text”：“不满祖先肖像被占用，王老吉后人将广州医药集团有限公司诉至法院。7 月 23 日，新京报记者……宣判后，三方当事人均未提出上诉。现判决已生效。”，“summary”：“不满祖先肖像商标被占用，王老吉后人将广州医药集团有限公司诉至法院…… 判决已生效。”
泛化集	2418	-	“text”：“新京报快讯据贵州检察消息，2019 年 11 月 13 日，贵州省黔南州中级人民法院一审公开开庭…… 编辑倪艳楠来源：贵州检察”，“summary”：“据贵州检察消息…… 庭审最后，法庭宣布休庭，择期宣判。”

型在矛盾样本中建立鲁棒性决策边界。共同构成评估模型深层规则理解的二元验证体系。

文本摘要任务使用了标准的 ROUGE-1 和 ROUGE-2 分数作为评估指标。为保证实验公平性，基准模型（baseline）采用的训练数据集为符号推理模块与神经计算模块训练数据集的合集（数据集 1+ 数据集 2），并在伪标签迭代训练阶段时增加一轮训练以平衡训练轮次差异。实验结果如表4-4、4-5所示。

ROUGE-1 实验结果分析 baseline 模型在初始训练后得分为 40.6%，最终经过多轮迭代训练后提升至 43.2%，提升值仅为 2.6%，显示出相对稳定但有限的改进空间。这种微弱的增益表明模型在初始阶段已接近性能饱和区，架构特性使得模型早期就捕获了主要特征规律，导致后续优化空间极为有限。结合其迭代过程中存在的准确率波动现象，这种有限的提升幅度反映出 baseline 模型在不引入额外优化机制的情况下，性能提升已达瓶颈。

表 4-4 文本摘要实验结果 (Rouge-1)

得分	初始训练	最终迭代结果	最终提升
baseline	40.6%	43.2%	2.6%
符号推理模块	25.7%	40.1%	14.4%
神经计算模块	38.7%	46.4%	7.7%

符号推理模块的初始得分为 25.7%，通过神经计算模块提供的伪标签迭代训练，最终得分提升至 40.1%，提升了 14.4 个百分点。这种显著提升验证了符号推理模块能够通过伪标签数据学习符合真实场景的标注模式，基于现实数据分布对初始规则进行适应性调整。值得注意的是，尽管初始准确率相对较低，但经过神经计算模块的伪标签引导后，符号推理模块展现出强大的学习能力，最终准确率已接近 baseline 模型水平。这一现象强有力地证明了本框架在低资源场景下的有效性——通过神经计算与符号推理的协同机制，能够显著缓解数据稀缺问题。

神经计算模块凭借预训练语言模型的优势，在经过极小规模模拟真实数据集的适应性训练后，初始得分就达到了 38.7%，这主要受益于模型通过预训练获得的实体特征先验知识。当融合符号推理模块的规则约束后，准确率进一步提升至 46.4%，提升了 7.7 个百分点。这种提升揭示了神经计算模块具有规则学习能力：通过语义空间与规则空间的协同优化，有效整合潜在规则关联，显著提升实体识别精度。但是值得注意的是，神经计算模块相对提升幅度低于符号推理模块，分析原因为数据集的结构特性。本研究所使用的数据集中，大部分原始文本源自社交媒体平台，这类文本通常不遵循传统文献的规范结构。在传统文本摘要任务中，首句和尾句往往包含文章核心信息，是提取摘要的关键位置。然而，社交媒体文本往往不具备这一特性，文章的开篇与结尾通常不能有效概括中心思想。例如在一篇关于康美药业财务造假的文章中，其起始部分为“近几天，上市公司康美药业 300 亿现金一夜消失，引起全国股民的一片哗然”，而结尾为“我们要认清这个体制，并且要知道靠一人的力量是很难改变现状的；我们要看透，投资者们除了谨慎投资之外，其他的什么都做不了”。这两部分内容难以提炼出文章的核心摘要：“康美药业被市场质疑多年其财务造假，直至 300 亿惊天造假案被爆，才算被实锤。四年来，投资者刘志清一直举报康美药业财务造假，证监会回复未发现相关违法情况。刘志清认为证监会严重渎职遂向法院起诉，一审不予立案，二审、最高院终审刘志清败诉。”这种非结构化特性使得基于位置信息的提取规则在社交媒体文本上效果受限，导致神经计算模块在优化过程中面临更大挑战。尽管预训练语言模型具备一定的先验语义理解能力，但在处理非规范化文本结构时，其性能提升空间受到制约。这一发现对于面向社交媒体文本的摘

要系统设计具有重要启示：在处理非规范文本时，应当降低对位置信息的依赖，转而强化对全文语义关联的建模能力。即使如此，神经计算模块的最终性能依旧超过了 baseline 模型和符号推理模块，证明了该模块在融合规则知识后能够达到更高的性能上限。

表 4-5 文本摘要实验结果 (Rouge-2)

得分	初始训练	最终迭代结果	最终提升
baseline	19.1%	20.2%	1.1%
符号推理模块	8.4%	15.8%	7.4%
神经计算模块	17.9%	20.5%	2.6%

ROUGE-2 实验结果分析 baseline 模型在 ROUGE-2 评估中初始得分为 19.1%，最终提升至 20.2%，提升值仅为 1.1%。这一结果与 ROUGE-1 评估中的趋势一致，再次表明模型在初始阶段已接近性能极限，后续优化空间有限。

符号推理模块的初始得分为 8.4%，通过神经计算模块提供的伪标签迭代训练，最终准确率提升至 15.8%，提升了 7.4 个百分点。这种大幅提升再次验证了符号推理模块通过伪标签进行适应性调整的能力。

神经计算模块初始得分为 17.9%，经过融合规则约束后，得分提升至 20.5%，提升了 2.6 个百分点。最终准确率略高于 baseline 模型，进一步证明了神经计算模块在规则引导下能够更精准地识别复杂语义关系，凸显出规则知识与神经计算协同的优越性。

实验结果表明，本研究提出的规则知识融合框架在评估任务中展现出显著优势。两个模块的性能演化印证了规则知识融合的核心价值——通过符号推理的可解释性与神经计算的泛化能力形成动态平衡，在保持较高水平性能的同时，降低了数据依赖度。相较大量数据集训练，该框架仅需少部分标注数据，为低资源场景下的实体识别提供了新的方法论框架，对后续研究中的模型架构设计具有重要启示意义。

同样，针对文本摘要任务，本研究也进行了消融实验，实验结果如表4-6、4-7所示

消融实验结果表明，在 Rouge-1 评分中，符号推理模块的初始训练得分从 25.7% 降低至 24.6%，降低了 1.1%，最终迭代结果得分从 40.1% 降低至 39.5%，降低了 0.6%；神经计算模块的初始训练得分从 38.7% 降低至 33.4%，降低了 5.3%，最终迭代结果得分从 46.4% 降低至 43.7%，降低了 2.7%。相似地，在 Rouge-2 评分中，符号推理模块的初始训练得分从 8.4% 降低至 7.9%，降低了 0.5%，最终迭

表 4-6 文本摘要任务消融实验结果（移除注意力掩码生成模块后）（Rouge-1）

得分	初始训练	最终迭代结果	最终提升
符号推理模块	24.6%	39.5%	14.9%
神经计算模块	33.4%	43.7%	10.3%

表 4-7 文本摘要任务消融实验结果（移除注意力掩码生成模块后）（Rouge-2）

得分	初始训练	最终迭代结果	最终提升
符号推理模块	7.9%	15.5%	7.6%
神经计算模块	13.7%	20.2%	6.5%

代结果得分从 15.8% 降低至 15.5%，降低了 0.3%；神经计算模块的初始训练得分从 17.9% 降低至 13.7%，降低了 4.2%，最终迭代结果得分从 20.5% 降低至 20.2%，降低了 0.3%。

深入分析消融实验结果发现，在移除注意力掩码方法时，神经计算模块的准确率下降幅度明显大于符号推理模块。这一现象揭示了两种模块在知识获取和处理机制上的本质差异。注意力掩码方法本质上是为了能够有效引入和整合规则知识，为模型提供规则知识。神经计算模块由于其端到端的黑盒特性，在没有注意力掩码引导的情况下，难以精确捕获和利用规则知识，导致性能下降较为显著。相比之下，符号推理模块展现出更强的鲁棒性。这是因为符号推理模块在训练过程中能够从数据中自主提取和学习规则模式，构建内部的规则表征。即使在移除注意力掩码方法的情况下，符号推理模块仍能通过其内在的归纳偏置和结构化表征能力，保持对规则知识的有效学习和应用，从而使其性能下降相对有限。这一发现与符号推理系统的固有特性相符——它们能够在有限样本下提取抽象规则并进行泛化。

值得注意的是，观察到在命名实体识别任务中，移除注意力掩码方法后符号推理模块与神经计算模块的性能差异并不如在文本摘要任务中显著。这一现象可归因于任务特性的本质差异。文本摘要是一项高度复杂的生成任务，涉及语义理解、关键信息提取、内容重组与表达等多层次认知过程，对规则知识的依赖程度较高。文本摘要需要模型理解长段落的语义结构，识别关键信息，并生成凝练的摘要。相比之下，命名实体识别任务的复杂度相对较低，主要关注于从文本中识别和分类特定类型的实体（如人名、地名、组织名等），更多地依赖于实体的局部特征（如词汇特征、上下文特征）和较为明确的实体边界，而非复杂的规则知识体系。

这一实验结果为混合神经-符号系统的设计提供了重要启示。首先，它证实

了符号推理模块在规则学习方面具有天然优势，能够在较少显式指导的情况下捕获数据中的规则模式。其次，它表明神经计算模块虽然具有强大的表征能力，但在处理高度结构化的规则知识时，仍然依赖于适当的归纳偏置机制（如注意力掩码）来引导学习过程。此外，两个模块在消融实验中表现出的互补性能差异，进一步证实了本研究双模块协同架构的合理性。这些结果充分验证了本研究提出的规则知识融合框架的有效性和鲁棒性。

基于当前的实验结果，本研究计划在以下方向进行扩展：(1) 跨任务适用性验证：为了进一步验证框架的通用性，可以扩展到其他自然语言处理任务，比如：机器翻译、问答系统等。(2) 规则知识的细粒度建模：当前规则掩码的权重设计采用固定值 (α, β, γ) ，可以进一步细化规则的粒度，比如：对不同类型的规则赋予不同的权重，例如语法规则、语义规则和领域规则的影响力可能不同。引入领域专家的知识，将规则权重的设定与特定领域的需求结合，提升模型的领域适应性。

4.4 小结

本章创新性地开发了基于注意力掩码的规则嵌入技术，在 Transformer 架构中实现了符号逻辑与全局语义的无缝融合。通过可微分注意力掩码生成器，将领域专家规则转化为注意力空间的硬约束与软引导。在命名实体识别和文本摘要任务中，模型的正确率均有所提升，且表现优于 **baseline**。该方法成功验证了符号知识在注意力机制中的可解释嵌入路径，为大规模预训练模型的规则适配提供了新范式。

第5章 基于规则知识的提示词生成与优化

5.1 概述

随着大规模预训练语言模型的不断发展，如何在不直接修改模型结构的前提下提升其在特定任务中的性能和可解释性，成为近年来自然语言处理领域的重要研究方向。将结构化的规则知识与预训练模型相结合的框架，为解决该问题提供了一种新颖且高效的思路。相比于传统的模型架构改造方式，该类方法更多地聚焦于模型输入层，通过注入规则约束信息，引导模型生成更具逻辑性与领域针对性的输出，从而在保持通用能力的同时，增强其在特定任务上的推理能力与表现稳定性。

本章提出并系统探讨了一种基于规则知识的提示词优化方法，其核心思想是将规则知识嵌入到自动提示工程（Automated Prompt Engineering, APE）流程中，借助结构化规则对提示生成进行引导与约束，以实现提示生成的自动化、结构化与高效化。该方法不仅致力于生成更符合逻辑结构的提示词，更强调提示词对任务知识的有效调用与利用，进而提升预训练模型在特定任务上的准确性与泛化能力。

5.2 方法介绍

本节系统介绍基于规则知识的提示词优化实验设计与实施方法。该方法旨在将结构化和可归纳的规则知识嵌入到自动提示工程框架中，从而生成逻辑更严密、知识利用效率更高的提示词，提升大型语言模型在多种推理任务中的泛化能力与任务表现。整个方法框架主要包括三个核心环节：（1）规则知识的构建；（2）提示模板的生成；（3）基于规则的提示优化算法设计。图5-1展示了本研究提出的总体流程。

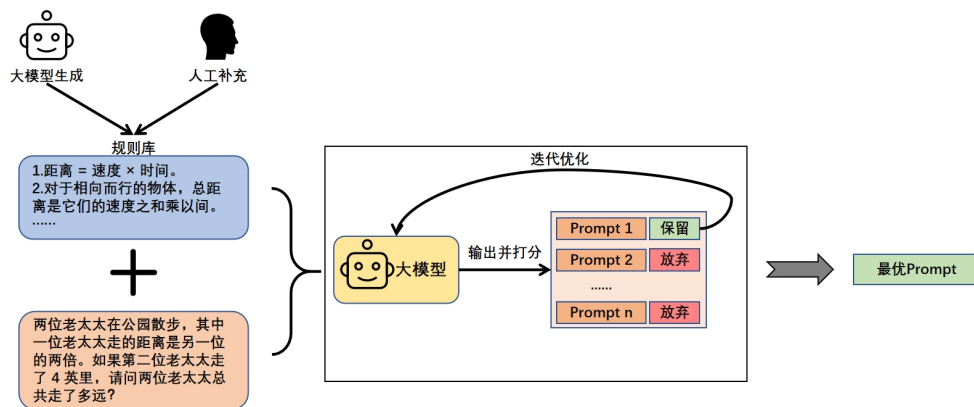


图 5-1 基于规则知识的提示词生成与优化

在规则知识构建阶段，本研究创新性地引入了大模型驱动的智能规则抽取机制。该机制以少量任务样本为基础，通过大模型的归纳能力自动分析任务中的关键约束条件，例如逻辑推理中的因果链条、文本分类中的特征敏感词、问答任务中的领域知识要求等。这些抽取出的隐含规则随后被形式化表示并构建为可拓展的规则库。相比于传统由人工手工构建的规则系统，该方法显著降低了构建成本，同时提高了规则的覆盖率与适应性。

在提示模板构建阶段，规则库中的形式化规则将被映射为自然语言形式，并以模板化结构嵌入至任务输入中，从而构成对模型具有约束性和引导性的提示文本。模板结构灵活，可适应不同任务需求。例如，在文本分类任务中，模板可强调特定特征或上下文关注点；在问答任务中，则可引导模型生成符合事实与逻辑的一致性答案。

在提示优化阶段，本研究设计了多种基于规则知识的优化算法，通过自动搜索、反馈迭代和模型输出评估等机制，对初始提示进行优化与重构。其中，规则不仅用于提示词筛选与生成，也参与模型训练目标的调整，使提示在迭代中不断向逻辑一致性与任务表现最优方向收敛。最终，通过微调与知识蒸馏机制，模型能够逐步内化这些规则知识，即使在推理阶段未显式提供规则，也能保持较高的规则遵循度与任务表现能力。

综上所述，本章所提出的基于规则知识的提示词优化框架，构建了一套从规则自动抽取到提示生成与优化再到模型学习与迁移的完整方法体系，兼顾了可解释性、泛化性与效率，为预训练模型在复杂推理任务中的应用提供了更具指导性和实用性的解决方案。

5.2.1 规则知识的构建

规则库作为本研究中提示生成与优化的基础，其质量直接影响提示模板的有效性，及模型对任务的理解程度。为了确保规则知识的准确性、结构化和可应用性，本研究采用了人机协同的智能规则构建方法，结合大语言模型的归纳能力与人工领域知识的经验优势，实现规则的高效挖掘与系统化组织。

为实现规则知识的自动化构建，本研究首先引入大型语言模型对训练样本进行分析与处理，借助其对语义模式和隐含逻辑的识别能力，从任务数据中自动抽取通用的推理约束与逻辑规则。首先模型通过对少量样本的分析，识别任务中的关键约束条件与逻辑路径。例如，在数学推理任务中，大模型能够自动总结出“处理括号优先”、“运算顺序影响结果”、“所有加数需列出”等通用规则；在常识推理任务中，模型能够归纳出“工具需与使用场景匹配”、“状态变化应符合常识”等逻辑限制。提取出的逻辑结构经过归纳抽象后，被组织为结构清晰、可复用的规则项，为后续提示模板构建提供知识支持。最后在抽取规则的基础上，本研究进一步引入领域背景知识（如数学公式库、常识知识图谱等），使得模型在

规则归纳过程中具有更强的上下文理解与泛化能力。

尽管大型语言模型在规则抽取中展现出较强的能力，但仍存在边界性覆盖不足、语言表达不规范等问题。为提升规则库的适用性和鲁棒性，本研究引入人工辅助机制对模型抽取的规则进行以下补充与优化针对模型未能识别的特殊情况或复杂边界条件，研究者根据任务需求设计具有针对性的规则，以增强系统的稳健性。对自然语言描述的规则进行风格与术语标准化处理，确保提示模板生成时语言表达的一致性。统一规则的逻辑范式，使其更易于嵌入模板结构，并支持多种任务场景的复用。

5.2.2 提示模板的生成

本研究提出将规则知识以自然语言或逻辑表达形式显式嵌入到提示模板中，引导模型在处理任务时自动遵循预设规则。该机制不仅提升了提示的可解释性，还增强了其对任务逻辑的约束能力。

如图5-1所示，展示了提示模板的生成流程，重点突出规则知识在提示生成与迭代优化中的作用。在该框架中，系统首先接收输入任务（如数学类问题），并结合预定义的规则库信息（如公式、语义结构、解题模式）共同作为提示生成的指导信号。在提示生成阶段，大模型结合输入任务与规则库内容生成若干候选提示模板（Prompt 1 至 Prompt n）。这些候选模板主要由大模型根据规则知识自动生成，部分也可由人工进行语义补充或格式约束，从而形成具有结构一致性与语义覆盖力的初始提示集合。在此过程中，生成的提示不仅需具有语义引导功能，还应保证与规则结构的一致性。

以常识问答任务（如 CommonsenseQA）为例，假设任务涉及手部保暖的判断问题，嵌入规则后的提示模板之一如下所示：

Prompt:

你正在进行一项关于常识判断的任务，请注意以下规则：

- 衣物的用途通常与身体保护或保温相关；
- 手部保暖需要覆盖或绝热结构。

问题：Which object would most likely keep your hands warm?

选项：A. gloves B. towel C. newspaper D. water bottle

请基于规则作答，并解释原因。

上述模板中，任务指令与规则内容被明确前置，形成结构清晰、逻辑严密的输入框架，引导模型在作答时优先考虑物理常识中的功能性匹配原则。

5.2.3 基于规则的提示优化算法

为提升提示生成质量与搜索效率，本研究在自动提示工程流程中系统性引入规则知识，构建了一种规则驱动的提示优化算法框架。该方法以规则知识为核心控制信号，全面作用于提示词的生成、演化与评估过程，使得优化流程在逻辑一致性与任务适配性方面显著增强。相比于传统依赖准确率或随机策略的提示优化方法，本研究引入规则约束机制，结合强化学习与启发式搜索策略，有效提高了提示生成的可控性与目标一致性。

候选提示集由前一阶段生成的结构化模板组成，进入本阶段后将接受系统性评价与迭代优化。为了系统性优化提示生成策略，引入强化学习框架。首先，系统对所有候选提示进行打分评估，衡量其在两个关键维度上的表现：一是任务准确性，反映提示对模型任务完成度的引导效果；二是规则匹配度，评估提示是否在语义或结构上有效体现规则约束，用于衡量提示词是否语义或结构上符合预设规则，体现提示是否覆盖或符合预设规则的内容与结构要求。采用语言模型打分评估（例如，对于 PIQA 中的物理常识题，若提示缺乏“因果关系”或“物理机制提示语”，则得分较低）。为此，本研究引入如下规则匹配度函数：

$$R(p) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \text{Match}(p, r) \quad (5-1)$$

随后，将规则匹配度与任务准确率共同构成总奖励函数，引导优化策略：

$$\text{Reward}(p) = \alpha \cdot \text{Accuracy}(p) + \beta \cdot R(p) \quad (5-2)$$

其中， α, β 为调节因子，用于平衡任务准确性与规则符合度的贡献。 $\text{Accuracy}(p)$ 为任务准确率，体现提示在引导模型完成特定任务（如算术推理）中的有效性。

打分完成后，系统根据奖励机制保留高质量提示，淘汰低分提示，并对保留模板进行进化操作（如替换、交叉、微调），形成新的候选模板集。在上述优化流程中，为进一步提升搜索效率，本研究在打分机制基础上引入启发式搜索策略。利用规则知识充当提示搜索空间的启发式约束，用于动态调整搜索优先级。具体而言，在候选提示空间中，系统优先探索那些包含高权重规则关键词或匹配规则结构的提示样式。例如，在 GSM8K 任务中，系统将优先选择包含“第一步”、“先算……”等引导式表达的提示模板。定义启发权重函数 $H(p)$ 作为提示词 p 与规则集合 $\mathcal{R}$ 的相似度评分，用于动态更新候选提示的排名：

$$H(p) = \sum_{r \in \mathcal{R}} \text{Sim}(p, r) \quad (5-3)$$

其中， $\text{Sim}(p, r)$ 表示提示词与规则 r 的匹配度。

随后进入下一轮提示生成与评估循环，持续优化提示的质量。该过程在强化学习或演化策略指导下迭代进行，最终收敛至一个最优 Prompt，该提示在准确率与规则符合度之间达到最优平衡。

5.3 实验结果

为了验证提出的规则知识融合框架在预训练模型上的有效性，本研究在多个推理任务上对 ChatGLM-6B、Baichuan-7B、LLaMA-7B、Falcon-7B 四个模型进行了实验评估。表5-1-5-6分别展示了数学推理任务（GSM8K/SVAMP）、常识推理任务（CommonsenseQA/PIQA）和通用推理任务（ARC）的实验结果。

5.3.1 数学推理任务

在数学推理任务中，本研究选择了 GSM8K 和 SVAMP 两个具有代表性的数据集作为测试基准。

GSM8K（Grade School Math 8K）是由 OpenAI 发布的广泛应用于评测领域的数学推理数据集，包含超过 8,000 道高质量、语言表达多样的小学数学应用题。这些问题通常需要 2 至 8 个推理步骤才能解决，主要通过基本算术运算（加、减、乘、除）连续计算得出最终答案。尽管这些问题看似简单，但它们旨在全面评估模型处理复杂数学问题的能力，而许多大型语言模型在此类任务上表现不尽如人意。GSM8K 不仅考察模型的基础计算能力，更重要的是测试其理解自然语言描述并生成正确推理链的能力，这对验证本研究框架中规则知识融合的有效性尤为重要。

SVAMP（Simple Variations on Arithmetic Math Word Problems）是一个专为评估基础数学推理能力而设计的数据集，包含 1,000 道表达形式多样的算术应用题，涵盖加减乘除等基本运算。该数据集通过同义词替换和句式重构生成不同表述的问题变体，要求模型在准确理解自然语言描述的基础上，结合上下文逻辑完成单步或多步计算（包括隐含条件推理）。尽管这些问题仅涉及小学水平的数学知识，但许多先进模型仍难以稳定处理其语言变体，凸显了当前模型在语义理解与数学逻辑结合方面的局限性。SVAMP 专门设计用于测试模型对语言变体的泛化能力和基础推理的鲁棒性，这与本研究框架强调的规则知识融合和逻辑一致性目标高度契合。

在数学推理任务中，本研究应用了规则知识与预训练模型结合的框架，设计了包含数学公式、运算规则和逻辑约束的专用提示模板。这些模板将数学规则知识显式嵌入到模型输入中，引导模型按照正确的推理步骤解决问题。例如，在解决 GSM8K 问题时，模板会提示模型需要逐步分解问题、明确运算顺序以及检查计算结果。同时本研究还构建了包含基本算术运算法则、数学公式和解题策略的

表 5-1 GSM8K 性能对比（准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	5.6%	11.3%	+5.7%
Baichuan-7B	5.3%	10.7%	+5.2%
LLaMA-7B	9.7%	16.8%	+7.1%
Falcon-7B	7.6%	13.4%	+5.8%

表 5-2 SVAMP 性能对比（准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	10.6%	14.4%	+3.8%
Baichuan-7B	12.2%	17.1%	+4.9%
LLaMA-7B	11.7%	17.3%	+5.6%
Falcon-7B	10.9%	15.2%	+4.3%

规则库。这些规则既包括显式的计算步骤指导（如先乘除后加减），也包括隐式的逻辑关系推导（如条件判断或约束关系）。通过在 GSM8K 和 SVAMP 数据集上微调预训练模型，增强模型对数学规则的理解和应用能力。随后利用知识蒸馏技术，将教师模型的数学推理能力迁移至学生模型，实现高效推理。通过这种方式，本研究的框架不仅充分利用预训练模型已掌握的知识，还通过规则引导使模型在数学推理这类需要严格逻辑和准确计算的任务上表现更加优异。

实验结果如表5-1、5-2所示，全面评估了所提出框架在数学推理任务上的有效性。从实验结果可以明确观察到，规则知识融合框架在所有测试模型上均带来了显著性能提升。

在 GSM8K 数据集评估中，LLaMA-7B 模型展现出显著的性能优势，基线准确率为 9.7%，结合规则知识融合框架后提升至 16.8%，提升幅度达到 7.1 个百分点，相对提升达 73.2%。同样，其他模型亦呈现出明显的性能提升：ChatGLM-6B 的准确率从 5.6% 提高至 11.3%，Baichuan-7B 从 5.3% 提高至 10.7%，Falcon-7B 则从 7.6% 提高至 13.4%。

在 SVAMP 数据集上，Baichuan-7B 展现出最佳的初始性能，基线准确率为 12.2%，融合框架后提升至 17.1%，提升幅度为 4.9 个百分点，相对提升 40.2%。而 LLaMA-7B 模型则在结合规则知识融合框架后获得了最高的准确率，从 11.7% 提升至 17.3%。ChatGLM-6B 的准确率从 10.6% 提高至 14.4%，Falcon-7B 则从 10.9% 提高至 15.2%。值得注意的是，所有模型在该数据集上的基线性能普遍高于 GSM8K，这可能反映了 SVAMP 数据集在推理复杂度上的差异。

实验结果中的性能差异部分源于模型预训练语料与数据集语言特性的匹配度。GSM8K 和 SVAMP 均为英文数据集，而不同模型在不同语言语料上的预训练比例各异。例如，LLaMA-7B 主要在英文语料上进行大规模预训练，这可能是其在英文数学推理任务上表现较好的原因之一。相比之下，ChatGLM-6B 和 Baichuan-7B 虽然具有双语能力，但在英文语料的训练占比上可能低于 LLaMA 系列，这可能导致它们在英文数据集上的基线性能相对较低。

规则知识融合框架在所有模型上均带来了显著提升，这验证了该方法的普适性和有效性。性能提升的主要原因在于，规则知识融合框架通过精心设计的提示模板和知识蒸馏技术，将数学推理的符号规则与预训练模型的生成能力有机结合。通过分析可以发现，性能提升幅度与模型架构和基线性能存在一定关联。其中 LLaMA-7B 在 GSM8K 上获得最大提升 (7.1%)，表明规则知识融合框架能够有效激发模型的内在数学推理潜力，尤其是对于那些基础能力较强但缺乏明确推理指导的模型。框架将方程求解、单位转换、问题分解和运算顺序等关键数学规则显式嵌入提示模板中，帮助模型更准确地理解问题结构，并按照正确的逻辑顺序进行推理。这在相对难度更高的 GSM8K 数据集上表现得尤为明显，所有模型在该数据集上的相对提升均超过 50%。同时，实验中观察到的提升差异也反映了不同模型在知识蒸馏过程中的迁移效率差异。例如，ChatGLM-6B 虽然参数量相对较小但在规则知识融合后表现出了与参数量更大的模型相当的提升效果，这表明其架构对规则知识的接受度较高。框架采用的伪标签迭代训练机制进一步增强了模型在低资源场景下的泛化能力，使其能够更稳定地处理多样化的数学问题。这种机制特别适合预训练模型在特定任务上的快速适应，无需大量标注数据即可实现显著提升。

其次，GSM8K 和 SVAMP 数据集在数学推理的复杂度上存在明显差异。GSM8K 通常需要 2 至 8 步推理才能解决问题，涉及多层次逻辑关系和连续运算；而 SVAMP 的问题相对简单，多为单步或少步骤推理，且问题表述更为直接。实验结果显示，在复杂度更高的 GSM8K 上，规则知识融合框架带来的相对提升普遍大于 SVAMP，这表明该框架在处理多步骤、复杂推理问题时具有更明显的优势。这也从侧面证明了规则知识对结构化推理的重要性——当问题需要更复杂的分解和推理链时，规则知识的引导作用变得尤为关键。模型在遵循规则引导后，能够更有条理地分解问题、确定运算顺序并验证中间结果，从而在复杂推理任务中表现出更高的准确性。

实验结果充分证明了所提出的规则知识融合框架在数学推理任务上的有效性。通过引入规则知识并设计合适的提示模板，中小规模预训练模型能够显著提升其数学推理能力，特别是在处理复杂多步骤问题时。这种方法不仅避免了对模型架构的大幅修改，还充分利用了预训练模型已有的知识基础，为提高模型在特

定任务上的表现提供了一种高效可行的解决方案。

5.3.2 常识推理任务

在常识推理任务中，本研究选择了 CommonsenseQA 和 PIQA 两个具有代表性的数据集作为测试基准。

CommonsenseQA 是一个广泛使用的常识推理数据集，专注于常识推理挑战，包含 12,247 道需外部知识支撑的选择题，问题基于 ConceptNet 知识图谱构建，旨在评估模型在理解和利用常识知识方面的能力。每个问题包含 5 个选项（1 个正确，4 个干扰项），设计时通过语义干扰项（如“玻璃杯破碎”与“漂浮”）增加推理难度。该数据集暴露了大模型在缺乏显式文本线索时依赖隐性常识的缺陷（如理解“冰块在烈日下会融化”需物理知识），要求模型不仅具备对自然语言的理解能力，还需要利用常识知识进行推理和判断，常用于评估模型与知识库的融合能力。

PIQA (Physical Interaction: Question Answering) 是一个专注于物理常识推理的问答数据集，包含 16,000 道基于物理常识的二选一现实场景操作问题，基于真实 DIY 指南构建，旨在评估模型在理解和解决与物理世界交互相关问题的能力。每个问题描述日常物理任务（如“防止冰淇淋融化”）并提供两个解决方案选项（正确方法 vs 干扰项），通过材料属性误解（如“毛巾保温”与“锡纸隔热”）增加推理难度。该数据集揭示了大模型对物体物理特性（如导热性、密封性）理解的不足，要求模型不仅能够理解自然语言描述，还需要具备对物理现象的推理能力，常用于评估智能体在家庭助手、机器人控制等场景的物理交互推理能力。

在常识推理任务中，本研究应用了常识知识与预训练模型结合的框架，设计了包含概念关系、因果推理和事件逻辑的专用提示模板。这些模板将常识规则知识显式嵌入到模型输入中，引导模型按照合理的推理步骤解决问题。例如，在解决 CommonsenseQA 问题时，模板会提示模型需要识别概念间的关系、排除语义干扰项以及验证推理的合理性。同时还构建了包含概念网络关系、常见事实和推理策略的规则库。这些规则既包括显式的概念关联指导（如“冰在热环境中融化”），也包括隐式的常识推导（如物体的普遍特性与属性）。通过在 CommonsenseQA 和 PIQA 数据集上微调预训练模型，增强模型对常识和物理规律的理解和应用能力。随后利用知识蒸馏技术，将教师模型的常识推理能力迁移至学生模型，实现高效推理。通过这种方式，本研究的框架不仅充分利用预训练模型已掌握的知识，还通过规则引导使模型在常识推理这类需要理解隐含知识和物理规律的任务上表现更加优异。

实验结果如表5-3、5-4所示，全面评估了所提出框架在常识推理任务上的有效性。实验结果表明，应用规则知识融合框架后，所有测试模型在 CommonsenseQA 和 PIQA 上均实现了显著性能提升。

表 5-3 CommonsenseQA 性能对比（准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	59.2%	65.5%	+6.3%
Baichuan-7B	50.8%	59.0%	+8.2%
LLaMA-7B	55.3%	62.9%	+7.6%
Falcon-7B	53.6%	62.1%	+8.5%

表 5-4 PIQA 性能对比（平均准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	65.2%	72.8%	+7.6%
Baichuan-7B	69.3%	78.4%	+9.1%
LLaMA-7B	64.7%	72.3%	+7.6%
Falcon-7B	63.8%	71.6%	+7.8%

在 CommonsenseQA 数据集上，ChatGLM-6Bbaseline 准确率为 59.2%，尽管提升幅度相对较小，为 6.3 个百分点，但其最终达到的 65.5% 准确率在四个测试模型中最高。Baichuan-7B 准确率从 50.8% 提升至 59.0%，实现了 8.2 个百分点的准确率提升。LLaMA-7B 准确率从 55.3% 提升至 62.9%，提升了 7.6 个百分点。Falcon-7B 准确率从 53.6% 提升至 62.1%，增幅达 8.5 个百分点，为四个模型中提升最显著的一个。

PIQA 任务性能分析在 PIQA 数据集测评中，ChatGLM-6B 准确率从 65.2% 提升至 72.8%，得了 7.6 个百分点的性能提升。Baichuan-7B 准确率从 69.3% 提升至 78.4%，增幅达 9.1 个百分点，在此任务中提升最为显著。LLaMA-7B 准确率从 64.7% 提升至 72.3%，提升了 7.6 个百分点。Falcon-7B 准确率从 63.8% 提升至 71.6%，实现了 7.8 个百分点的准确率提升。

实验结果中的性能差异反映了不同模型在常识推理任务上的基础能力与知识融合后的提升潜力。CommonsenseQA 和 PIQA 均为评估模型常识理解能力的英文数据集，但它们考察的认知维度有所不同：CommonsenseQA 侧重于概念间关系的推理，而 PIQA 则更关注物理常识的应用。通过具体案例分析，可以更清晰地理解规则知识融合框架的有效性。值得注意的是，尽管模型架构和参数规模存在差异，四个测试模型在应用规则知识融合框架后均实现了不同的准确率提升。这一结果表明，本研究提出的方法具有广泛的适用性，能够有效增强不同架构语言模型的常识推理能力。特别是对于基准性能较低的模型，规则知识融合往往能够带来更为显著的性能提升，从而缩小不同模型间的性能差距。

在 PIQA 数据集中，以下案例清晰展示了框架的优势。问题：“如何留胡子？”选项：A. 每天冲洗胡子并涂抹胡须油。B. 冲洗不断增长的胡子，涂上脱毛膏。每周用胡须清洗剂 1-3 次。

在应用规则知识融合框架前，LLaMA-7B 错误选择了 B 选项，可能是因为模型将“不断增长”与“留胡子”的概念联系到了一起，而在结合规则知识和提示模板后，模型正确选择了 A 选项。尤其是对于提示模板中的“识别答案选项中的内部逻辑矛盾”与“答案应符合基本常识”，显著地提醒了模型“使用脱毛膏”和“留胡子”之间存在明显的矛盾。在护理胡子的情境中，使用专门的胡子护理产品（如胡子油）符合常识，而使用脱毛产品则违背目标，这种逻辑矛盾在规则知识的引导下得到了有效识别。

与 GSM8K 和 SVAMP 不同，CommonsenseQA 和 PIQA 中的数据不是具有固定规则的加减乘除运算题，而是包罗万象的生活常识推断题，本研究构建的规则库在面对“如何快速冷却一杯热咖啡？”可能会有所帮助，因为其中包含了热传递原理相关的物理规则，引导模型认识到增加表面积可以加速热量散失，而不仅是将物体置于低温环境。模型在框架的指导下，能够超越简单关联（冷却-冰箱），转而进行基于物理原理的深层次推理。然而，面对“如何在家做草莓猕猴桃酱”这样的特定领域问题时，本研究构建的通用规则库可能覆盖不足。在这种情况下，提示模板的引导作用显得尤为重要。尽管在大规模测试中为了效率本研究仅要求模型输出正确选项，但仍随机抽样个别案例要求模型给出详细分析过程，并通过人工评判验证其推理逻辑的合理性。以问题“如何在家做草莓猕猴桃酱”为例，其选项为“A：将 1 杯猕猴桃（切碎）、1 杯切碎的草莓、3/4 杯水和 1 杯橄榄核煮沸 30 分钟，搅拌以防止在炉灶上过热而烧焦。B：将 1 杯猕猴桃（切碎）、1 杯切碎的草莓与 3/4 杯水和 1 杯糖一起煮沸 30 分钟，搅拌以防止在炉灶上过热而烧焦。”通过分析模型的推理过程，发现提示模板中的“功能对应规则：产品或物质的功能应与其用途相符。”提示启发了模型识别“在果酱制作中，糖的功能是提供甜度、帮助保存和增稠。橄榄核不具备这些功能，与制作果酱的目的不符”；同时，“逻辑矛盾检测：识别答案选项中的内部逻辑矛盾”提示引导模型发现“第一个选项中‘橄榄核’与制作甜果酱存在明显逻辑矛盾，因为橄榄核既不能溶解也不会增加甜度”。这些案例分析表明，本研究的规则知识融合框架通过引入结构化的常识规则和精心设计的提示模板，有效增强了语言模型的逻辑推理能力和常识应用能力。框架不仅帮助模型识别选项中的逻辑矛盾和常识违背，还引导模型进行更深层次的原理性推理，从而在不同类型的常识问题上取得一致的性能提升。

实验结果充分证明了所提出的规则知识融合框架在常识推理任务上的广泛适用性。通过引入规则知识并设计针对性的提示模板，中小规模预训练模型能够

显著提升其常识推理能力，无论是概念关系理解还是物理常识应用。未来，可以进一步探索如何通过更精细的规则设计与更高效的融合机制，进一步提升模型在常识推理任务中的整体表现，特别是针对需要多步骤、跨领域知识整合的复杂推理场景。

5.3.3 通用推理任务

在通用推理任务中，本研究选择了 ARC 数据集作为测试基准，以评估模型在科学知识推理方面的能力。

ARC (AI2 Reasoning Challenge) 是由艾伦人工智能研究所 (AI2) 开发的科学推理问答数据集，包含 7,787 道源自 3 至 9 年级科学考试的多选题，按照难度分为简单 (Easy) 和挑战 (Challenge) 两个子集。该数据集全面涵盖物理、化学、生物、地理、天文等多个学科领域，问题形式为四选一标准化题目，例如“哪种材料最适合做隔热层？”或“植物光合作用的主要产物是什么？”。ARC 数据集还配备了 1430 万条科学文献段落作为外部知识源，这些文本来自教科书、维基百科和科学网站，旨在为模型提供必要的背景知识。该数据集特别设计用于评估 AI 系统在科学推理方面的能力，尤其强调对科学原理的理解而非简单事实记忆，要求模型能够整合多领域知识并应用于复杂问题解决。

其中，Challenge 部分专门筛选了通过简单检索和统计方法难以解决的问题，特别考验模型的跨学科知识整合与多步推理能力。例如，回答“生态系统中的能量流动如何影响食物链？”不仅需要理解个别生物体的能量代谢，还涉及整体生态系统的动态平衡和能量守恒定律。Challenge 子集通过设置概念混淆项（如“声音在真空中传播速度”与“声音在空气中传播速度”）增加推理难度，要求模型具备深层次科学概念理解和逻辑推理能力。研究表明，即使是最先进的大型语言模型在此类任务上仍面临显著挑战，特别是在需要多步骤推理、因果关系分析和科学原理应用的问题上，这暴露了当前模型在结构化科学知识理解和应用方面的局限性。

在通用推理任务中，本研究设计了包含科学概念关系、实验推理和原理应用的专用提示模板。这些模板将科学规则知识结构化嵌入到模型输入中，引导模型按照科学方法论进行系统推理和验证。例如，在解决物理学问题时，模板会提示模型先识别相关物理定律，然后应用公式，最后验证结果的合理性；在生物学问题中，模板则引导模型从分子、细胞到系统层级进行多尺度分析。同时，本研究构建了包含科学定律、实验事实和推理策略的分层知识库。这些知识既包括显式的公式定律（如“ $F=ma$ ”、“光合作用方程式”），也包括隐式的科学推理原则（如控制变量法、假设验证流程）。知识库按学科领域组织，便于模型根据问题特征进行精准知识检索和应用。随后，本研究应用知识蒸馏技术，将具备丰富科学知识和推理能力的教师模型知识迁移至学生模型，实现计算效率与推理能力的

表 5-5 ARC-easy 性能对比（准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	70.6%	75.0%	+4.4%
Baichuan-7B	73.8%	78.4%	+4.6%
LLaMA-7B	72.4%	76.3%	+3.9%
Falcon-7B	71.9%	75.1%	+3.2%

表 5-6 ARC-challenge 性能对比（准确率）

模型	基线	结合规则知识融合框架	提升
ChatGLM-6B	47.2%	52.8%	+5.6%
Baichuan-7B	49.2%	55.5%	+6.3%
LLaMA-7B	47.3%	53.4%	+6.1%
Falcon-7B	45.7%	51.9%	+6.2%

平衡。通过蒸馏过程，本研究特别关注科学概念之间的关系表征和推理步骤的传递，确保学生模型能够继承教师模型的系统推理能力。

通过这种方式，本研究的框架不仅充分利用预训练模型的语言理解能力，还通过结构化知识引导和多步推理训练，显著提升了模型在科学推理这类需要专业知识和系统思维的任务上的表现。实验结果如表5-5、5-6所示。

在 ARC-easy 数据集测评中，ChatGLM-6B 基线准确率为 70.6%，应用规则知识融合框架后提升至 75.0%，实现了 4.4 个百分点的性能提升。Baichuan-7B 基线准确率为 73.8%，应用规则知识融合框架后提升至 78.4%，提升了 4.6 个百分点，不仅是四个模型中提升幅度最大的，其最终达到的准确率也是最高的。LLaMA-7B 基线准确率为 72.4%，应用规则知识融合框架后提升至 76.3%，增幅为 3.9 个百分点。Falcon-7B 基线准确率为 71.9%，应用规则知识融合框架后提升至 75.0%，实现了 3.2 个百分点的准确率提升，在此任务中提升幅度相对较小。

在 ARC-challenge 数据集测评中，ChatGLM-6B 基线准确率为 47.2%，应用规则知识融合框架后提升至 52.8%，得到了 5.6 个百分点的性能提升。Baichuan-7B 基线准确率为 49.2%，应用规则知识融合框架后提升至 55.5%，增幅达 6.3 个百分点，最终达到的准确率在四个模型中最高。LLaMA-7B 基线准确率为 47.3%，应用规则知识融合框架后提升至 53.4%，提升了 6.1 个百分点。Falcon-7B 基线准确率为 45.7%，应用规则知识融合框架后提升至 51.9%，实现了 6.2 个百分点的准确率提升。

实验结果中的性能差异反映了不同模型在科学推理任务上的基础能力与知

识融合后的提升潜力。ARC-easy 和 ARC-challenge 均为评估模型科学推理能力的英文数据集，但它们考察的难度层次有明显区别：ARC-easy 包含相对基础的科学知识问题，而 ARC-challenge 则包含需要多步推理和跨学科知识整合的复杂问题。

值得注意的是，在 ARC-challenge 任务上，所有模型的提升幅度普遍大于 ARC-easy 任务，平均提升幅度达到 6.1 个百分点，相比 ARC-easy 任务的 4.0 个百分点提升更为显著。这一现象表明，本研究提出的规则知识融合框架在处理需要复杂推理和深层次科学概念理解的任务上优势更为明显，说明该框架能有效增强模型处理复杂科学推理问题的能力。

以 ARC-challenge 中的一个案例为例：“哪种物质最适合用作房屋隔热材料？”选项：A. 铝箔 B. 玻璃纤维 C. 铜片 D. 木材。应用规则知识融合框架前，LLaMA-7B 错误选择了 A 选项，可能是因为模型将“反射热”与“隔热”概念混淆。在结合科学规则知识和提示模板后，模型正确选择了 B 选项。规则库中关于热传导的科学原理“导热系数低的材料是良好的隔热体”和提示模板中的“材料性质与功能对应规则”引导模型进行了更深入的物理属性分析，识别出玻璃纤维内部的空气层结构能有效阻断热传导。

ARC 数据集涉及的是系统化科学知识，需要模型具备更强的结构化知识理解和应用能力。本研究的规则知识库在面对“行星围绕恒星运行的轨道形状是什么？”这类天文物理问题时，能提供开普勒行星运动定律等相关科学定律，引导模型进行基于物理规律的推理。然而，针对“某种特定细菌在什么条件下生长最快？”这类生物学专业问题，通用规则库的覆盖可能不够全面。在这种情况下，本研究设计的科学推理提示模板，特别是“实验条件分析”和“多变量因素考量”部分发挥了关键作用，引导模型系统地考虑温度、pH 值、营养等多种影响微生物生长的因素。

实验结果充分证明了所提出的规则知识融合框架在科学推理任务上的有效性和广泛适用性。通过引入结构化科学规则知识并设计针对性的提示模板，中小规模预训练模型能够显著提升其科学推理能力，尤其是在处理需要多步骤、跨学科知识整合的复杂问题时。未来研究可进一步探索如何构建更全面的学科知识规则库，以及如何优化知识检索和融合机制，进一步提升模型在多样化科学推理任务中的表现。

5.4 小结

本章构建了“提示工程 + 规则蒸馏”的规则知识融合框架，在多项推理任务中准确率均有所提升，证实了符号规则对大语言模型认知边界的拓展作用。创新点在于先通过结构化提示模板激活预训练模型的隐式推理能力，再采用规则约

束的对比蒸馏强化符号逻辑表达，最终通过自治性校验机制消除知识冲突。

第6章 总结与展望

6.1 研究结果

本研究围绕基于神经符号计算的知识数据融合学习框架展开，针对人工智能系统在知识推理能力与学习能力上的不足，提出了三种不同层次的解决方案：面向 DNN 的规则知识融合框架、面向 Transformer 的知识融合框架，以及面向预训练模型的知识融合框架。以下是研究的主要结论：

(1) 面向 DNN 的规则知识融合框架

通过符号规则指导 DNN 模型的训练，并结合双向知识蒸馏与伪标签迭代训练，显著提升了低资源场景下模型的可解释性与泛化能力。实验结果表明，该框架在命名实体识别和情感分类任务上均实现了性能提升，在符号推理与数据驱动方法的协同优化方面展现了良好的效果。特别是在低资源场景下，该框架通过引入符号规则，有效缓解了数据稀缺问题，显著提升了模型的鲁棒性和泛化性能。

(2) 面向 Transformer 的知识融合框架

借助 Transformer 的强表征能力，通过设计基于规则的注意力掩码机制，将符号规则嵌入到多头自注意力机制中，实现了符号推理与全局语义信息的高效结合。实验结果显示，该框架不仅提升了模型的准确性，还增强了其在复杂任务场景中的泛化能力，尤其是符号推理任务的表现显著优于传统基线模型。此外，该框架通过引入规则注意力掩码，有效提升了模型在长距离依赖任务中的表现，为复杂推理任务的解决提供了新的思路。

(3) 面向预训练模型的知识融合框架

在预训练模型微调阶段注入规则知识，通过提示模板与知识蒸馏技术，充分发挥规则指导与预训练模型生成能力的协同作用。该框架适用于复杂任务的逻辑推理和内容生成，在数学推理、常识推理等领域进一步验证其应用潜力。实验结果表明，该框架在生成任务中不仅提升了生成内容的一致性，还显著改善了生成结果的逻辑正确性，为生成模型的可靠性提供了新的解决方案。

综合来看，本文提出的三种知识融合框架各具优势，分别适用于规则执行严格性要求较高、任务通用性较强以及生成能力需求较大的应用场景。实验结果证明，神经符号计算可以有效结合符号推理的可解释性与深度学习的学习能力，为人工智能系统的可信性与智能性提供了理论与实践支持。

6.2 展望

尽管本研究在神经符号计算的知识数据融合方面取得了一定的成果，但仍有许多值得进一步探索的方向。未来的研究可以从以下几个方面展开：

(1) 规则复杂度扩展

当前的研究主要集中在单一规则或简单规则集合的应用上，未来可以进一步研究多规则协同约束与冲突解决机制。特别是在复杂任务场景中，如何有效处理规则之间的冲突与优先级问题，将是一个重要的研究方向。此外，规则的动态更新与适应性调整也是未来研究的一个重点，以适应不断变化的任务需求。

(2) 任务领域扩展

本研究主要聚焦于自然语言处理任务，未来可以将框架应用于更多任务领域，如计算机视觉、语音识别等跨模态任务。特别是在跨模态任务中，如何有效融合不同模态的符号知识与数据驱动方法，将是一个具有挑战性的研究方向。此外，探索框架在特定领域（如医疗、金融等）中的应用，也将具有重要的实际意义。

(3) 自动化规则学习

当前的规则主要依赖于人工设计或领域专家的知识，未来可以探索从数据中自动抽取符号规则并与神经网络深度集成的方法。通过引入自动化规则学习技术，可以进一步提升框架的灵活性与适应性，减少对人工规则的依赖。此外，结合强化学习等技术，探索规则与模型的协同优化，也将是一个重要的研究方向。

(4) 模型的可解释性与可信性

尽管神经符号计算在一定程度上提升了模型的可解释性，但在实际应用中，如何进一步提升模型的可信性仍是一个重要问题。未来可以探索如何通过可视化、交互式解释等手段，增强用户对模型决策过程的理解与信任。此外，研究模型在对抗攻击下的鲁棒性，也将是提升模型可信性的一个重要方向。

(5) 大规模应用与系统集成

未来的研究还可以探索如何将提出的框架应用于大规模系统中，特别是在实际工业场景中的集成与应用。通过系统级的优化与集成，可以进一步提升框架的实用性与可扩展性，为实际应用提供更强大的技术支持。

通过进一步的探索，本研究期望推动更高效、更智能、更可信的知识融合系统的发展，为下一代人工智能提供坚实的理论基础与技术支持。

6.3 小结

研究通过三类知识融合框架的系统构建,实现了神经符号计算在 DNN、Transformer 及大模型场景中的方法论突破。核心贡献包括:提出双向知识蒸馏的规则迁移理论、创立注意力空间的符号嵌入范式、开发预训练模型的规则蒸馏技术。展望部分指出,未来需在动态知识演化建模、多模态符号推理、人机协同解释机制等方向深化研究,特别强调建立神经符号系统的持续学习与伦理对齐框架,为构建第三代可信 AI 系统提供理论储备。

参考文献

- [1] Dai W Z, Xu Q, Yu Y, et al. Bridging machine learning and logical reasoning by abductive learning [C]//Advances in Neural Information Processing Systems. 2019: 2815-2826.
- [2] Yu P, Xu J, Weston J, et al. Distilling system 2 into system 1 [J]. 2024, abs/2407.06023.
- [3] Gu Y, Dong L, Wei F, et al. Minillm: Knowledge distillation of large language models [EB/OL]. 2024. <https://arxiv.org/abs/2306.08543>.
- [4] Yang Y, Zhuang Y, Pan Y. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies [J]. Frontiers of Information Technology & Electronic Engineering, 2021, 22(12): 1551-1558.
- [5] Haugeland J. Artificial intelligence: The very idea [M]. MIT press, 1989.
- [6] Zhang J, Chen B, Zhang L, et al. Neural, symbolic and neural-symbolic reasoning on knowledge graphs [J]. AI Open, 2021, 2: 14-35.
- [7] Hadley R F. Connectionism, rule following, and symbolic manipulation [J]. 1990: 579-586.
- [8] Brassard G, Salvail L. Secret-key reconciliation by public discussion [C]//1993: 410-423.
- [9] Albright A, Hayes B. Rules vs. analogy in english past tenses: a computational/experimental study. [J]. 2003, 90: 119-161.
- [10] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
- [11] Fodor J A, Pylyshyn Z W. Connectionism and cognitive architecture: a critical analysis [J]. 1988, 28: 3-71.
- [12] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation [R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [13] Lapedes A S, Farber R M. How neural nets work [J]. 1987: 442-456.
- [14] Smolensky P, McCoy R T, Fernandez R, et al. Neurocompositional computing in human and machine intelligence: A tutorial [R]. Microsoft Technical Report MSR-TR-2022, 2022.
- [15] Smolensky P, McCoy R, Fernandez R, et al. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems [J]. AI Magazine, 2022, 43(3): 308-322.
- [16] Fodor J A, Pylyshyn Z W. Connectionism and cognitive architecture: A critical analysis [J]. Cognition, 1988, 28(1-2): 3-71.
- [17] Lakkaraju H, Kamar E, Caruana R, et al. Faithful and customizable explanations of black box models. [C]//2019.
- [18] Hu Y, Huang Z A, Liu R, et al. Source free semi-supervised transfer learning for diagnosis of mental disorders on fmri scans. [J]. 2023: 1-18.
- [19] Gladisch C, Heinz T, Heinzemann C, et al. Experience paper: Search-based testing in automated driving control applications [J]. 2019.
- [20] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity [J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133.
- [21] Lake B M, Ullman T D, Tenenbaum J B, et al. Building machines that learn and think like people [J]. Behavioral and brain sciences, 2017, 40.

- [22] Marcus G. Deep learning: A critical appraisal [J]. arXiv preprint arXiv:1801.00631, 2018.
- [23] Valiant L G. Three problems in computer science [J]. Journal of the ACM, 2003, 50(1): 96-99.
- [24] Garcez A, Gori M, Lamb L, et al. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning [J]. Journal of Applied Logics, 2019, 6(4): 611-632.
- [25] Besold T R, d'Avila Garcez A, Bader S, et al. Neural-symbolic learning and reasoning: A survey and interpretation [J]. Neuro-Symbolic Artificial Intelligence: The State of the Art, 2022, 342: 1.
- [26] Donadello I, Serafini L, d'Avila Garcez A. Logic tensor networks for semantic image interpretation [C]//International Joint Conferences on Artificial Intelligence. 2017: 1596-1602.
- [27] Zhou T, Qi S, Wang W, et al. Cascaded parsing of human-object interaction recognition [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2021, 44(6): 2827-2840.
- [28] Andreas J, Rohrbach M, Darrell T, et al. Neural module networks [C]//Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2016: 39-48.
- [29] Mao J, Gan C, Kohli P, et al. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision [C]//Proc. Int. Conf. Learn. Representations. 2019.
- [30] Amizadeh S, Palangi H, Polozov A, et al. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning" [C]//Proc. ACM Int. Conf. Mach. Learn. 2020: 279-290.
- [31] Wang W, Zhou T, Qi S, et al. Hierarchical human semantic parsing with comprehensive part-relation modeling [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2021, 44(07): 3508-3522.
- [32] Li L, Zhou T, Wang W, et al. Deep hierarchical semantic segmentation [C]//Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2022: 1246-1257.
- [33] Arabshahi F, Lee J, Gawarecki M, et al. Conversational neuro-symbolic commonsense reasoning [C]//AAAI Conference on Artificial Intelligence. 2021: 4902-4911.
- [34] Wang W, Yang Y, Wu F. Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing [EB/OL]. 2024. <https://arxiv.org/abs/2210.15889>.
- [35] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
- [36] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.
- [37] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.
- [38] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489.
- [39] Chen X, Liang C, Yu A W, et al. Compositional generalization via neural-symbolic stack machines [C]//Advances in Neural Information Processing Systems. 2020: 1690-1701.
- [40] Gupta T, Kembhavi A. Visual programming: Compositional visual reasoning without training [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 14953-14962.

- [41] Shen Y, Song K, Tan X, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface [C]//Advances in Neural Information Processing Systems. 2023.
- [42] Surís D, Menon S, Vondrick C. Vipergpt: Visual inference via python execution for reasoning [C]//Proceedings of the IEEE International Conference on Computer Vision. 2023.
- [43] Liang C, Berant J, Le Q, et al. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision [C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2017.
- [44] Garnelo M, Arulkumaran K, Shanahan M. Towards deep symbolic reinforcement learning [J]. arXiv preprint arXiv:1609.05518, 2016.
- [45] Yang F, Lyu D, Liu B, et al. Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making [C]//International Joint Conferences on Artificial Intelligence. 2018: 4860-4866.
- [46] Lyu D, Yang F, Liu B, et al. Sdrl: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning [C]//AAAI Conference on Artificial Intelligence. 2019: 2970-2977.
- [47] Mou L, Lu Z, Li H, et al. Coupling distributed and symbolic execution for natural language queries [C]//Proceedings of the ACM International Conference on Machine Learning. 2017: 2518-2526.
- [48] de Penning H L H, d'Avila Garcez A S, Lamb L C, et al. A neural-symbolic cognitive agent for online learning and reasoning [C]//International Joint Conferences on Artificial Intelligence. 2011: 1653-1658.
- [49] Rocktäschel T, Riedel S. End-to-end differentiable proving [C]//Proceedings of Advances in Neural Information Processing Systems. 2017: 3791-3803.
- [50] Minervini P, Riedel S, Stenetorp P, et al. Learning reasoning strategies in end-to-end differentiable proving [C]//Proceedings of the ACM International Conference on Machine Learning. 2020: 6938-6949.
- [51] Weber L, Minervini P, Münchmeyer J, et al. NLProlog: Reasoning with weak unification for question answering in natural language [C]//Annual Meeting of the Association for Computational Linguistics. 2019: 6151-6161.
- [52] Manhaeve R, Dumancic S, Kimmig A, et al. DeepProbLog: Neural probabilistic logic programming [C]//Proceedings of Advances in Neural Information Processing Systems. 2018: 3753-3763.
- [53] Tsamoura E, Hospedales T, Michael L. Neural-symbolic integration: A compositional perspective [C]//AAAI Conference on Artificial Intelligence. 2021: 5051-5060.
- [54] Si X, Raghothaman M, Heo K, et al. Synthesizing datalog programs using numerical relaxation [C]//International Joint Conferences on Artificial Intelligence. 2019: 6117-6124.
- [55] De Raedt L, Kimmig A, Toivonen H. ProbLog: A probabilistic Prolog and its application in link discovery [C]//International Joint Conferences on Artificial Intelligence. 2007: 2462-2467.
- [56] Dai H, Tian Y, Dai B, et al. Syntax-directed variational autoencoder for structured data [C]//International Conference on Learning Representations. 2018.

- [57] Chen X, Liu C, Song D. Tree-to-tree neural networks for program translation [C]// Proceedings of Advances in Neural Information Processing Systems. 2018: 2552-2562.
- [58] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation [C]//Proceedings of the ACM International Conference on Machine Learning. 2018: 2323-2332.
- [59] Lample G, Charton F. Deep learning for symbolic mathematics [C]//International Conference on Learning Representations. 2019.
- [60] Li Q, Huang S, Hong Y, et al. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning [C]//Proceedings of the ACM International Conference on Machine Learning. 2020: 5884-5894.
- [61] Arabshahi F, Singh S, Anandkumar A. Combining symbolic expressions and black-box function evaluations in neural programs [C]//International Conference on Learning Representations. 2018.
- [62] Johnson J, Hariharan B, Van Der Maaten L, et al. Inferring and executing programs for visual reasoning [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2989-2998.
- [63] Dhingra B, Zaheer M, Balachandran V, et al. Differentiable reasoning over a virtual knowledge base [C]//International Conference on Learning Representations. 2019.
- [64] Hoernle N, Karampatsis R M, Belle V, et al. Multiplexnet: Towards fully satisfied logical constraints in neural networks [C]//AAAI Conference on Artificial Intelligence. 2022: 5700-5709.
- [65] Ahmed K, Teso S, Chang K W, et al. Semantic probabilistic layers for neuro-symbolic learning [C]//Advances in Neural Information Processing Systems. 2022: 29944-29959.
- [66] Giunchiglia E, Tatomir A, Stoian M C, et al. CCN+: A neuro-symbolic framework for deep learning with requirements [J]. International Journal of Approximate Reasoning, 2024: 109124.
- [67] Zhang H, Dang M, Peng N, et al. Tractable control for autoregressive language generation [C]//International Conference on Machine Learning. 2023: 40932-40945.
- [68] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.
- [69] Lin B Y, Chen X, Chen J, et al. Kagnet: Knowledge-aware graph networks for commonsense reasoning [C]//Conference on Empirical Methods in Natural Language Processing. 2019: 2829-2839.
- [70] Lv S, Guo D, Xu J, et al. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering [C]//AAAI Conference on Artificial Intelligence. 2020: 8449-8456.
- [71] Serafini L, Garcez A d. Logic tensor networks: Deep learning and logical reasoning from data and knowledge [C]//Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning. 2016.
- [72] Badreddine S, Garcez A d, Serafini L, et al. Logic tensor networks [J]. Artificial Intelligence, 2022, 303: 103649.

- [73] Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks [C]// Proceedings of the ACM International Conference on Machine Learning. 2018: 5075-5084.
- [74] Bertinetto L, Mueller R, Tertikas K, et al. Making better mistakes: Leveraging class hierarchies with deep networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 12506-12515.
- [75] Ahmed K, Chang K W, Van den Broeck G. A pseudo-semantic loss for autoregressive models with logical constraints [C]//Proceedings of Advances in Neural Information Processing Systems. 2023: 18325-18340.
- [76] Dong H, Mao J, Lin T, et al. Neural logic machines [C]//Proceedings of International Conference on Learning Representations. 2018.
- [77] Wang P W, Donti P, Wilder B, et al. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver [C]//Proceedings of the ACM International Conference on Machine Learning. 2019: 6545-6554.
- [78] Cai J, Shin R, Song D. Making neural programming architectures generalize via recursion [C]//Proceedings of International Conference on Learning Representations. 2017.
- [79] Hersche M, Zeqiri M, Benini L, et al. A neuro-vector-symbolic architecture for solving raven's progressive matrices [EB/OL]. 2023. <https://arxiv.org/abs/2203.04571>.
- [80] Cornelio C, Dash S, Austel V, et al. Ai descartes: Combining data and theory for derivable scientific discovery [EB/OL]. 2023. <https://arxiv.org/abs/2109.01634>.
- [81] Rao C, Ren P, Wang Q, et al. Encoding physics to learn reaction-diffusion processes [EB/OL]. 2023. <https://arxiv.org/abs/2106.04781>.
- [82] Dou Q, Liu Q, Heng P A, et al. Unpaired multi-modal segmentation via knowledge distillation [J]. 2020, 39: 2415-2425.
- [83] Wang K, Liu Y, Ma Q, et al. Mulde: Multi-teacher knowledge distillation for low-dimensional knowledge graph embeddings [C]//2021: 1716-1726.
- [84] Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons [J]. 2019, 33: 3779-3787.
- [85] Zhang J, Guo S, Ma X, et al. Parameterized knowledge transfer for personalized federated learning [J]. 2021, 34: 10092-10104.
- [86] van Engelen J E, Hoos H H. A survey on semi-supervised learning [J]. 2019, 109: 373-440.
- [87] OpenAI, Achiam J, Adler S, et al. Gpt-4 technical report [EB/OL]. 2024. <https://arxiv.org/abs/2303.08774>.
- [88] Sun X, Davis J, Schulte O, et al. Cracking the black box: Distilling deep sports analytics [J]. 2020, abs/2006.04551: 3154-3162.
- [89] Zhang Y, Li Y, Liu T, et al. Dual-path distillation: A unified framework to improve black-box attacks. [J]. 2020: 11163-11172.
- [90] Kahneman D. Thinking, fast and slow [M/OL]. New York: Farrar, Straus and Giroux, 2011. <https://www.worldcat.org/title/thinking-fast-and-slow/oclc/706020916>.
- [91] Evans J S B T. In two minds: dual-process accounts of reasoning [J/OL]. Trends in Cognitive Sciences, 2003, 7(10): 454-459. <https://doi.org/10.1016/j.tics.2003.08.012>.
- [92] Agrawal A, Amos B, Barratt S, et al. Differentiable convex optimization layers [EB/OL]. 2019. <https://arxiv.org/abs/1910.12430>.

- [93] Chen R T Q, Rubanova Y, Bettencourt J, et al. Neural ordinary differential equations [EB/OL]. 2019. <https://arxiv.org/abs/1806.07366>.
- [94] Sun L, Lyu L. Federated model distillation with noise-free differential privacy [J]. 2021: 1563-1570.
- [95] Hoang T N, Lam C T, Low B K H, et al. Learning task-agnostic embedding of multiple black-box experts for multi-task model fusion [J]. 2020, 119.
- [96] Lamb L C, d'Avila Garcez A, Gori M, et al. Graph neural networks meet neural-symbolic computing: A survey and perspective [J]. 2020: 4877-4884.
- [97] Frady E P, Kent S J, Olshausen B A, et al. Resonator networks, 1: An efficient solution for factoring high-dimensional, distributed representations of data structures [J]. 2020, 32: 2311-2331.
- [98] Burke M, Subr K, Ramamoorthy S. Action sequencing using visual permutations [J]. 2021, 6: 1745-1752.
- [99] Hitzler P, Bianchi F, Ebrahimi M, et al. Neural-symbolic integration and the semantic web. [J]. 2020, 11: 3.0-11.
- [100] Diaz-Rodriguez N, Lamas A, Sanchez J, et al. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: the monumai cultural heritage use case [J]. 2022, 79: 58-83.
- [101] Liang Y, Meng F, Zhang Y, et al. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. [J]. 2021, 35: 13343-13352.
- [102] Kitson A, Schiphorst T, Riecke B E. Are you dreaming?: A phenomenological study on understanding lucid dreams as a tool for introspection in virtual reality. [C]//2018: 1-12.
- [103] Huang H, Wen M, Wei L, et al. Characterizing and detecting configuration compatibility issues in android apps [J]. 2021.
- [104] Le-Phuoc D, Eiter T, Le-Tuan A. A scalable reasoning and learning approach for neural-symbolic stream fusion [J]. 2021, 35: 4996-5005.
- [105] d'Avila Garcez A S, Besold T R, Raedt L D, et al. Neural-symbolic learning and reasoning: Contributions and challenges. [C]//2015.
- [106] Dissing L, Bolander T. Implementing theory of mind on a robot using dynamic epistemic logic [J]. 2020: 1615-1621.
- [107] Jiang J, Wang H, Xie J, et al. Medical knowledge embedding based on recursive neural network for multi-disease diagnosis. [J]. 2018.
- [108] Bresnan J. Linguistics: the garden and the bush [J]. 2016: 599-617.

致 谢

本论文的完成离不开许多人的帮助与支持，在此我谨向他们表示诚挚的感谢。

首先，我要感谢我的导师王石老师。感谢您在研究过程中的悉心指导和严谨治学的态度，您渊博的知识、敏锐的学术洞察力以及对科研的热忱深深影响了我。没有您的引导与支持，本论文不可能完成。

感谢实验室的各位师兄师姐师弟们。感谢他们在实验设计与数据分析方面给予的宝贵建议在文献整理和实验操作中提供的协助。与你们的讨论和合作使我受益匪浅，你们的经验分享和鼓励给了我莫大的帮助。

衷心感谢我的朋友们。谢谢你们在我低落时的鼓励，在我迷惘时的指引，以及在我取得进步时的真诚祝贺。你们的理解与陪伴是我坚持不懈的重要精神支持。

最后，我要向我的父母表达最深的谢意。感谢你们多年来的无私奉献、包容理解与毫无保留的爱。你们对我的教育投入和对我追求学术梦想的支持是我前进的动力源泉。

在此，向所有曾经帮助、关心和支持我的人致以最诚挚的谢意。本论文的每一点成绩都凝聚着你们的心血与智慧。

2025 年 3 月

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历：

姓名：孙严顺 性别：男 出生日期：1999 年 11 月 24 日 籍贯：云南曲靖

2018 年 9 月——2022 年 6 月，在中国科学院大学计算机科学与技术学院获得学士学位。

2022 年 9 月——至今，在中国科学院计算技术研究所攻读硕士学位。

已发表（或正式接受）的学术论文：

无

参加的研究项目及获奖情况：

参加的研究项目：

获奖情况：

(1) 2024 年中国科学院大学学业奖学金

