中国人工智能学会
Chinese Association for Artificial Intelligence

中国人工智能学会系列白皮书
——语言智能

中国人工智能学会
二〇二五年八月

中国人工智能学会系列白皮书
——语言智能

中国人工智能学会
二〇二五年八月

《中国人工智能学会系列白皮书》编委会

主　　任：戴琼海

执行主任：马华东

副 主 任：赵春江 何　友 王恩东 郑庆华 刘成林

周志华 孙富春 庄越挺 胡德文 杜军平

杨　强

委　　员：陈松灿 董振江 付宜利 高新波 公茂果 古天龙 何　清

胡清华 黄河燕 季向阳 蒋田仔 林浩哲 梁吉业 刘奕群

潘　纲 石光明 孙茂松 孙长银 陶建华 王海峰 王熙照

王　轩 王蕴红 吴　飞 于　剑 余有成 张化光 张学工

章　毅 周鸿祎 周　杰 祝烈煌

《中国人工智能学会系列白皮书---语言智能》编写组

周建设 刘　杰 袁家政　余正涛 林鸿飞 姜　孟 薛嗣媛 周俊生

黄于欣 许明英 张　晴

目 录

前言

在人工智能与数字文明深度交融的时代背景下，语言智能作为人机协同的核心驱动力，正以前所未有的速度重塑教育体系、跨文化交流、产业创新与社会治理模式。为系统梳理该领域的技术演进、学科构建与应用生态，本白皮书汇聚产学研多方智慧，围绕“技术-学科-应用”三维体系展开深度探索，深度解构其技术攻坚路径、生态构建逻辑与治理挑战，旨在为技术探索者、政策设计者及产业实践者提供前沿洞察与决策参考。

白皮书首先纵览语言智能发展轨迹，揭示其从规则驱动、统计学习到深度学习的范式跃迁规律，勾勒出技术从感知理解向认知推理跃升的进化路线；进而奠定学科根基，通过界定理论边界与时代必要性，构建“核心框架-技术体系-学科关系”三维模型，并探索外语学科智能化转型的实践路径；最终聚焦技术内核，围绕语言理解（机器的“读心术”）、语言生成（“创作力觉醒”）与语言交互（人机“灵魂对话”）三大核心能力，将可信安全深度嵌入技术基因，确立“能力越强，防护越密”的同步进化准则。

白皮书聚焦四大关键应用场景：教育智能通过作文批改的“篇章要素抽取→跨提示评分→可解释反馈”全链优化，及儿童语言评价的多维度框架，实现从结果判定到过程干预的跃升；跨语言服务针对东南亚低资源语言（如老挝语），突破语料与文化瓶颈，构建“神经机器翻译→知识增强→大模型优化”技术梯队；社会治理依托负面情感分析攻克反讽识别与隐式情绪挖掘，筑牢公共安全与金融风控决策基座；多模态关联以多模态知识图谱为枢纽驱动跨模态语义对齐及多跳推理，助推工业质检、医疗诊断等场景智能升级。

白皮书通过“总结与展望”章节，勾勒出技术融合推动语言生成与大模型推理协同进化，增强低资源场景适应性；学科建设深化外语教育与 AI 的嫁接，培育跨领域人才；可信机制构建覆盖模态均衡性与文化偏见监测的动态防护体系；最终以人本愿景实现“任意模态输入，统一语义理解”的可靠认知底座，驱动语言智能服务人类文明包容性发展。

我们期望白皮书能够为语言智能领域的可持续发展提供策略建议，促进技术创新，推动产业升级，为政策制定者提供学科建设与伦理治理参考，为技术开发者厘清攻坚方向，为产业界锚定商业化路径，助力中国在全球语言智能浪潮中占据创新制高点，并为构建智能社会贡献智慧和力量。

第一章　语言智能概述

1.1　引言

语言作为人类认知与交流的核心载体，是人工智能实现通用智能的关键突破口。2013 年首都师范大学周建设教授首次提出人工智能范畴的语言智能概念，将其定义为运用计算机技术模拟人类智能处理语言信息的过程，基于人脑结构与言语机制，借助大数据和 AI 技术解析语言属性，构建人机模型，使机器具备听、说、读、写等功能，核心公式为“语言智能 = 类人语言 + 类人智能”。

技术上，语言智能指 AI 系统理解语义、生成文本、实现人机交互的能力，涉及自然语言处理、计算语言学和认知科学。从图灵测试到 ChatGPT，其发展历程体现了技术突破与人类对语言本质认知的深化。

近年来，GPT、BERT 等预训练模型推动语言智能迈向通用能力，2022 年 ChatGPT 的发布标志其进入实用化阶段，被视为“AI 产业的 iPhone 时刻”。目前已应用于机器翻译、智能客服、内容创作等领域，但面临计算成本高昂、可解释性缺失、伦理风险凸显等挑战。

现有研究多聚焦特定技术或应用领域，缺乏对发展脉络的系统性梳理。下面基于学术文献与技术史实，构建完整的发展阶段划分框架，解析各阶段技术演进的内在逻辑，旨在为理解语言智能的发展规律与未来方向提供全景式参照。

1.2　语言智能的发展阶段划分与核心演进

语言智能的发展始终受计算能力、数据资源与算法理论三大要素的制约，其技术范式历经四次重大变革：从规则驱动转向统计驱动、从统计模型演进至神经网络、由浅层神经网络拓展至深度结构、自判别式模型跨越至生成式模型。依据技术特征与发展历程中的关键节点，其演进历程可划分为五个显著发展阶段。

1.2.1　早期探索阶段：规则驱动与符号主义（1950s-1990s）

这一阶段的核心特征在于采用基于语言学规则的符号处理方法，即通过预设的语法和语义规则系统来处理语言符号，例如在自然语言处理中依赖人工编码的规则集进行任务解析。然而，技术发展受到计算能力不足的限制，导致处理速度缓慢且无法支持复杂运算，同时数据资源匮乏严重制约了模型的训练和优化，使得系统难以应对多样化的实际场景。因此，该阶段的应用尚未能形成规模化，仅局限于小范围实验或理论研究阶段，未能广泛部署到商业或社会实践中。

1.2.2　理论奠基与初步尝试（1950s-1970s）

1950 年，图灵在其著作《计算机器与智能》中提出的“模仿游戏”（即图灵测试），为语言智能确立了初步发展目标。1957 年，Noam Chomsky 提出的生成

语法理论指出，语言存在深层结构与表层结构的双重特征，为机器语言解析奠定了语言学基础，并直接推动了基于句法规则的处理系统发展。

该时期的代表性成果包括 1966 年麻省理工学院（MIT）开发的聊天机器人 Eliza。该系统通过模式匹配与预设规则实现对话模拟，虽未具备真正的语言理解能力，但首次实证了人机语言交互的可行性。1970 年，Winograd 开发的 SHRDLU 系统能够解析简单英文指令并操作虚拟积木世界，成为早期符号主义语言处理范式的典范。

然而，规则驱动方法存在固有局限性：需依赖专家手工编纂大量语法规则，难以应对语言歧义性及复杂句式结构（如“白天鹅在游水”的多重解析问题），且系统可移植性较差，在真实应用场景中受到显著制约。

1.2.3 统计方法的初步渗透（1980s-1990s）

20 世纪 80 年代后，随着计算能力提升与语料库建设的初步发展，统计方法逐渐取代纯规则方法。1990 年，IBM 发布的布朗语料库为统计语言模型奠定了数据基础，其研发的隐马尔可夫模型（HMM）成功应用于词性标注任务，标志着统计自然语言处理的开端。

本阶段的核心突破体现于统计机器翻译的提出。IBM 研究团队开发的 Model 1-5 系列模型通过概率建模实现源语言与目标语言的短语对齐，显著提升了机器翻译的性能指标。20 世纪 90 年代末期，谷歌搜索引擎融合 PageRank 算法与 TF-IDF 统计方法，优化了文本检索效能，成为统计方法商业化应用的代表性案例。

然而，此阶段的统计方法仍依赖人工特征工程，模型表达能力存在局限性。计算复杂度随特征维度呈指数增长，导致语言深层语义关联难以有效建模。

1.3 统计学习阶段：特征工程与浅层模型（2000s-2010s 中期）

自 2000 年以来，随着机器学习理论的日臻成熟与广泛应用，该领域显著促进了语言智能的发展，标志着其正式进入以统计学习为主导的关键阶段。此时期的突出特征在于特征工程处理的精细化，涵盖文本数据的特征提取、降维及优化方法的系统性探索。与此同时，浅层学习模型（如支持向量机、决策树与朴素贝叶斯分类器）在自然语言处理任务中获得大规模应用，显著提升了语言识别、分类及预测的准确性。这些技术突破不仅巩固了统计学习在语言智能中的核心地位，更通过积累的实践经验与数据基础，为后续深度神经网络及深度学习范式的兴起奠定了坚实的理论与应用基础。

1.3.1 统计学习模型的广泛应用

这一时期，支持向量机（SVM）与最大熵模型等机器学习算法成为自然语言处理（NLP）任务的主流技术方案。在情感分析、文本分类等任务中，研究者通过人工设计词性、句法结构等特征，结合统计模型进行文本分类决策。例如，Pang

等人（2002）首次将机器学习方法应用于电影评论情感分析领域，采用 unigram 特征与 SVM 模型取得了 82.9% 的分类准确率。

然而，人工特征工程存在显著局限性：其特征设计高度依赖领域专家知识，模型泛化能力受限，且难以有效捕捉语言的上下文依赖关系。2003 年，Yoshua Bengio 团队提出神经网络语言模型（Neural Network Language Model, NNLM），首次引入词嵌入（Word Embedding）概念。该模型通过低维稠密向量表示词语，初步解决了传统 one-hot 编码面临的高维稀疏性（即“维度灾难”）问题，为分布式语义表示提供了新的范式。

1.3.2 语料库建设与评估体系完善

数据资源的持续积累构成该阶段发展的关键基础。2002 年，Papineni 等人提出了机器翻译领域的 BLEU 评价指标。2004 年，Lin 等人确立了文本摘要的 ROUGE 评价体系。这些评价指标为跨模型性能比较提供了量化基准。此后，2006 年发布的 Penn Treebank 语料库包含详尽的句法标注信息，被确立为句法分析模型的标准测试基准。2009 年，ACL 学会推出的 CoNLL 评测会议整合了词性标注、依存句法分析等多项子任务，构建了语言智能技术的标准化评估范式，进一步推动了技术规范化进程。

1.4 神经网络崛起阶段：词向量与序列建模（2013-2017）

2013 年，深度学习技术的突破性进展触发了语言智能领域的首次范式革命，其核心标志为词向量技术的普及化与循环神经网络（RNN）的广泛应用，实现了从“词汇数字化”向“语义理解”的范式跨越。同期，周建设教授首次系统提出语言智能的基本概念，并界定了其核心内涵。

1.4.1 词向量技术的语义革命

2013 年，Google Mikolov 团队提出 Word2Vec 模型。该模型通过 Skip-Gram（根据中心词预测上下文）与 CBOW（根据上下文预测中心词）两种架构，实现了词嵌入的高效训练。Word2Vec 的革命性意义在于其揭示的语义关联特性：通过向量运算示例“国王 - 男人 + 女人 ≈ 女王”，首次证实了模型能够有效捕捉词语间的语义关系。

2014 年，Pennington 等人提出 GloVe 模型，通过融合全局词共现统计信息与局部上下文窗口训练机制，进一步增强了词向量对语义及语法特征的双重表征能力。词向量技术推动了语言表示范式的颠覆性变革，取代人工设计特征成为自然语言处理任务的基础输入表征形式，显著提升了情感分析、命名实体识别等下游任务的性能。

1.4.2 序列建模的技术突破

语言的时序特性要求模型具备记忆能力，循环神经网络（RNN）成为该时期序列建模的核心工具。1997 年提出的长短期记忆网络（LSTM）通过引入输入门、遗忘门及输出门的门控机制，有效缓解了传统 RNN 所面临的梯度消失问题，从而能够建模更长的上下文依赖关系。2014 年，Cho 等人提出的门控循环单元（GRU）通过简化 LSTM 结构，在保持模型性能的同时显著提升了训练效率。

2014 年，Google Brain 团队提出序列到序列（Seq2Seq）架构，该架构采用编码器-解码器结构实现端到端的序列转换，并成功应用于机器翻译任务。然而，Seq2Seq 存在固有局限：编码器需将源语言序列压缩为固定长度的向量表示，导致长句语义信息丢失，翻译质量随句子长度增加而显著降低。

1.4.3 注意力机制的初步探索

为解决序列到序列（Seq2Seq）模型处理长距离依赖关系的局限性，Bahdanau 等人（2015）提出了“软对齐”注意力机制。该机制使解码器能够在生成每个目标词时动态聚焦于编码器输出的不同位置，显著提升了长句翻译性能，其 BLEU 分数提升幅度超过 30%。注意力机制模拟了人类阅读过程中的选择性聚焦行为：通过计算查询向量（Query）与键向量（Key）之间的相似度以分配权重，实现对关键信息的选择性增强；继而对值向量（Value）进行加权求和，生成最终输出。该机制为后续 Transformer 架构的提出奠定了核心理论基础。

1.5 Transformer 革命阶段：预训练范式与模型规模化（2017-2022）

2017 年 Transformer 架构的提出标志着语言智能进入现代深度学习阶段，预训练范式的建立与模型规模的扩大推动技术性能实现跨越式提升。

1.5.1 Transformer 架构的革命性突破

2017 年，Vaswani 等人在《Attention Is All You Need》一文中首次提出了完全基于注意力机制的 Transformer 架构，摒弃了传统的循环神经网络（RNN）与卷积神经网络（CNN）序列处理方式。该架构的核心创新体现在多头自注意力机制、位置编码和并行计算能力三个方面。

Transformer 在 WMT 2014 英德翻译任务上取得了当时的最佳结果，且训练时间大幅缩短。其开源实现（Tensor2Tensor 库）发布后，迅速成为自然语言处理（NLP）社区的基础模型架构。

1.5.2 预训练语言模型的双路径发展

基于 Transformer 架构，2018 年语言智能领域涌现出两大标志性模型，确立了生成与理解的双轨发展路径：

（1）理解型模型代表：BERT

Google 团队提出的 BERT（Bidirectional Encoder Representations from Transformers）模型采用 Transformer 编码器架构，通过两种预训练任务学习语言表示：掩码语言模型（Masked Language Modeling, MLM）遮蔽输入序列中的部分词汇并予以预测，实现双向上下文建模；下一句预测（Next Sentence Prediction, NSP）任务则通过判断句子对的连续性，学习句子级语义关联。BERT 确立了“预训练 + 微调”（Pre-training + Fine-tuning）范式，即首先在大规模文本语料上进行无监督预训练，随后针对下游具体任务进行有监督微调。该范式使 BERT 在 GLUE（General Language Understanding Evaluation）基准测试的多项任务中超越了人类基线表现，迅速成为情感分析、问答等自然语言理解任务的主导模型。

（2）生成型模型代表：GPT 系列

OpenAI 团队开发的 GPT（Generative Pre-trained Transformer）系列模型采用 Transformer 解码器架构，其核心预训练目标为“预测下一个词”（Next Token Prediction），专注于文本生成能力。GPT-1（2018）首次验证了预训练范式在生成任务中的有效性；GPT-2（2019）将模型参数量提升至 15 亿，初步展现出零样本学习（Zero-shot Learning）能力；GPT-3（2020）凭借高达 1750 亿的参数量实现了突破性进展，有力验证了“规模定律”（Scaling Law）的普适性——即随着模型参数规模、训练数据量及计算资源的持续增长，模型性能将系统性提升，并可能涌现出小规模模型所不具备的“涌现能力”（Emergent Abilities）。

1.5.3 预训练技术的持续迭代

2019 年后，预训练语言模型呈现快速迭代态势。Facebook 公司提出的 RoBERTa 通过优化训练策略（取消下一句预测任务、延长训练时长），显著提升了 BERT 模型的性能；百度公司研发的 ERNIE 引入知识增强掩码策略，将实体与短语等结构化知识融入预训练过程；XLNet 融合 Transformer-XL 的长序列建模能力与 BERT 的双向表征优势，有效解决了掩码语言模型的预训练偏差问题。模型参数量呈指数级增长态势：从 BERT-Large 的 3.4 亿参数扩展至 GPT-3 的 1750 亿参数，训练资源需求呈激增态势——BERT-Large 训练需 16 个 TPU Pod 持续运行 4 天，而 GPT-3 训练成本高达 4600 万美元。该阶段实证研究表明，模型规模、训练数据量与计算资源的协同优化构成语言智能突破的核心驱动力。

1.6 生成式 AI 爆发阶段：大语言模型与生态构建（2022 至今）

2022 年末 ChatGPT 的发布标志着语言智能正式迈入生成式人工智能时代，模型能力范式实现了从“功能型任务执行”向“对话交互能力”的根本性转变，由此推动了技术普及与商业化应用的爆发性增长。

1.6.1 ChatGPT 的技术突破与生态影响

ChatGPT 基于 GPT-3.5 架构，通过两项关键技术实现了用户体验的革命性提升：指令微调技术（Instruction Tuning）使模型能够准确理解自然语言指令，有效适应多样化任务需求；基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）则通过人类标注员对模型输出进行排序，显著优化输出质量与价值观对齐。OpenAI 的研究证实，RLHF 技术使模型对有害请求的拒绝率提升了 6 倍，并显著降低了模型产生“幻觉”（即生成虚假信息）的概率。

ChatGPT 上线短期内用户量即突破百万量级，其接近人类水平的对话能力、上下文记忆能力与多任务处理能力，有力推动了生成式人工智能从学术研究迈向全民应用。2023 年发布的 GPT-4 进一步实现了多模态能力突破，可同时处理文本、图像等多模态输入，并在律师资格考试、SAT 等人类标准化测试中展现出卓越性能。

1.6.2 全球大模型生态的构建

生成式人工智能的迅猛发展引发全球技术竞争，推动形成多元化的大模型发展格局。在国际领域，Meta 公司发布开源 LLaMA 系列模型（其中 LLaMA-2 参数量达 700 亿），显著促进开源学术研究进程；Anthropic 公司推出的 Claude 系列模型则聚焦于安全对齐（safety alignment）与长上下文处理能力（支持超十万 tokens 的上下文窗口）。在国内领域，百度文心一言、阿里通义千问、华为盘古大模型等基础模型相继推出，标志着中国在人工智能生成内容（AIGC）技术领域已确立全球领先地位。

2024 年标志着生成式人工智能从基础研究向应用探索的关键转折，研究重心转向垂直领域适配。典型案例如下：Codex 模型驱动的 GitHub Copilot 实现代码自动补全功能，据报告显示开发者效率提升超 40%；Galactica 模型专注于科学文本生成，可自动生成学术摘要与化学分子式；医疗领域的大规模语言模型（如 Med-PaLM 2）在临床问答评估任务中达到专业医师水平。这些应用均展现出生成式 AI 在专业场景中的技术突破。

1.6.3 训练方法学的优化与效率提升

为解决大型模型在训练与推理过程中面临的高昂计算资源需求问题，研究者已开发出多种效率优化方法。2024 年，由中国人工智能团队深度求索（DeepSeek）独立研发的 DeepSeek 语言模型正式发布。该大规模预训练模型基于 Transformer 架构，融合强化学习与混合专家（Mixture of Experts, MoE）技术，通过海量多语言数据训练，展现出卓越的自然语言理解与生成性能。其高效性、低成本及开源特性显著提升了技术可访问性，推动了人工智能技术的普惠化进程与领域进步，标志着语言智能发展的重要里程碑。

DeepMind 提出的 Chinchilla 模型通过优化训练数据量与模型参数规模的配比（如使用 1.4 万亿 token 训练 700 亿参数模型），实证表明该策略相较于单纯扩大参数量，能显著提升训练效能。Google 研发的 Switch Transformer 采用 MoE 架构，通过动态激活部分专家模块以降低计算开销；国内团队提出的原生稀疏注意力（Native Sparse Attention, NSA）与混合块注意力（Mixed Block Attention, MoBA）技术，则利用稀疏注意力机制有效缓解了长序列建模的计算瓶颈问题。

模型压缩领域亦取得显著进展：基于知识蒸馏的模型（如 DistilBERT、TinyBERT 等）在性能损失低于 5%的前提下，成功将参数量压缩达 70%；量化技术通过降低参数精度（如从 32 位浮点 FP32 降至 4 位整型 INT4），显著提升了模型推理效率。上述技术突破共同促进了大型模型从云端服务器向边缘计算设备的迁移，从而拓展了其应用场景。

1.7 语言智能的未来

面对技术与社会挑战，语言智能将朝多维度、多元化发展，核心目标是实现“更高效、更可信、更通用”的智能演进。

1.7.1 高效训练与推理技术

模型发展重心从“规模竞赛”转向“效率竞赛”：新型架构（如动态路由自适应架构、神经符号混合架构）将突破 Transformer 限制，在缩减参数规模时增强表征能力；训练方法革新（如联邦学习、增量学习）可降低数据依赖与隐私风险，优化自监督学习减少标注需求。

硬件与软件协同优化为关键支撑：专用 AI 芯片（如 TPU v5、昇腾）提升计算效率，内存优化技术（如 Flash Attention）降低存储需求，支持大模型在边缘设备高效运行。

1.7.2 可信 AI 与可解释性技术

可解释性成为大模型核心属性：研究从“事后解释”转向“事前可解释”，设计内在可解释架构（如模块化 Transformer），使推理过程可追溯、可验证；融入因果推理技术，区分相关性与因果性，减少虚假关联错误。

可信 AI 体系逐步完善：建立伦理准则与审计机制，监管模型全生命周期；内容水印技术（如 OpenAI 隐形标记）识别 AI 生成内容，抑制虚假信息；对抗性训练增强模型鲁棒性，防范恶意攻击。

1.7.3 多模态融合与通用智能

多模态融合从“模态拼接”演进至“深度协同”：大模型具备统一多模态语义表征能力，理解并生成文本、图像、语音、视频等内容，实现“所见即可述，所述即可得”的交互；跨模态迁移学习提升低资源模态处理效能。

通用语言智能聚焦提升“认知能力”：模型不仅能处理语言表层信息，还将具备逻辑推理、常识理解、规划决策等高级能力；神经科学与认知科学交叉融合（如模拟人脑的神经形态模型）为设计提供新思路。

1.7.4 垂直领域的深度适配

大模型向专业化、精细化演进：针对医疗、法律、金融等领域开发专用模型，通过领域知识注入、专业数据微调提升任务准确性与可靠性（如医疗大模型整合电子病历、文献，实现精准诊断与治疗推荐）。

行业解决方案集成化成趋势：语言智能深度耦合行业业务，形成“大模型 + 行业平台”范式（如教育领域融合大模型的智能辅导系统实现个性化教学；制造业基于语音交互的智能运维系统提升效率）。

1.8 小结

语言智能的发展历程是人工智能技术演进的缩影，从早期的规则符号到如今的生成式大模型，每一次技术突破都源于算法、数据与计算力的协同进步。Transformer 架构的提出奠定了现代语言智能的基础，预训练范式的建立实现了技术的规模化应用，而 RLHF 等对齐技术则推动模型从 “能生成” 向 “善交流” 转变。

当前，语言智能正处于从“专用”向“通用”、从“技术突破”向“产业落地”的关键转型期，其发展既面临模型效率、可解释性、伦理安全等挑战，也蕴含着多模态融合、高效智能、可信对齐等发展机遇。未来，语言智能的进步将不仅依赖于技术的持续创新，更需要伦理、法律、教育等社会体系的协同支撑。

作为人工智能与人类交互的核心接口，语言智能的发展最终将推动人机共生时代的到来 —— 机器不仅能理解人类语言，更能理解人类意图与价值观，成为人类认知与创造的延伸工具。这一过程既需要技术的理性推进，也需要社会的审慎引导，唯有如此，才能实现语言智能的可持续发展，为人类社会带来更大价值。

表 1-1 语言智能发展历程中的重要事件与影响（1954-2024）

时间	重要事件	技术背景	核心问题	影响与意义
1954	乔治敦-IBM (Georgetown-IBM) 机器翻译实验	基于词典和人工编写句法规则	如何将语言学规则编码，让计算机实现自动翻译	标志着计算语言学成为独立学科，是语言学与计算机科学的首次结合

时间	重要事件	技术背景	核心问题	影响与意义
1956	人工智能概念提出	20 世纪中叶计算机科学与认知科学的交叉发展，以图灵测试、神经网络模型和达特茅斯会议为标志，为模拟人类智能提供了理论与技术基础	如何在技术层面实现机器的自主性、可解释性与泛化能力，同时解决伦理层面的责任归属、算法偏见及人机关系平衡问题	AI 通过效率提升、服务优化和科技创新推动社会进步；AI 作为人类探索智能本质的工具，重塑生产方式、社会治理与文明形态，成为第四次工业革命的核心驱动力
1957	乔姆斯基发表《句法结构》	生成语法理论兴起	人类“有限规则生成无限句子”的内在语言机制是什么	为计算语言学提供了形式化基础，其“先天论”观点主导学界半个世纪
1966	ALPAC 报告否定机器翻译可行性	早期规则翻译方法质量粗糙，遭遇瓶颈	基于规则的机器翻译能否达到实用水平	机器翻译研究投入大幅减少，领域进入低谷期
1988	IBM 提出基于统计的机器翻译模型	算力提升，双语语料库积累	能否用统计概率而非人工规则实现翻译	机器翻译研究复苏，方法从“规则驱动”转向“数据驱动”
1990 年代	隐马尔可夫模型(HMM)在语音识别中商业化应用	统计模型在语音识别领域证明其实用性	如何让机器更准确地识别非特定人的连续语音？	推动了语音识别技术的产品化，如 IBM 的 ViaVoice 听写产品
1993	出现译后编辑、翻译记忆等技术	计算机辅助翻译(CAT)工具发展	如何将机器翻译与人工翻译流程结合，提升效率	奠定了人机协同的现代翻译工作流，机器作为辅助工具定位明确
2013	语言智能概念提出	人工智能发展至认知智能阶段	突破语义理解，实现人机语义同构	重塑人机交互，推动 AI 与心智融合
2013	Word2Vec 模型提出	深度学习兴起，分布式表示思想	如何让计算机“理解”词语的语义和关系	实现了词语向量化，词义可以用数学计算，极大推动了 NLP 发展
2014	神经网络机器翻译(NMT) 出现	深度学习，序列到序列学习模型	能否用端到端的神经网络直接学习两种语言的映射	机器翻译质量取得飞跃性提升，译文更加流畅自然

时间	重要事件	技术背景	核心问题	影响与意义
2017	Transformer 架构诞生	自注意力机制克服了 RNN/LSTM 序列建模的缺陷	如何让模型在处理长序列时更好地捕捉全局依赖关系	突破了句法树分析框架，催生了 BERT、GPT 等一切大语言模型，是里程碑式的革新
2018	BERT 模型发布	Transformer 编码器架构，预训练+微调范式	如何让模型深层地理解上下文语境？	在 11 项 NLP 任务中刷新纪录，证明了预训练模型的强大威力。
2020	GPT-3 实现零样本学习	Transformer 解码器架构，海量参数与 数据	模型能否不经过特定任务训练，就直接完成新任务	AI 变为语言学研究的“增强工具”，推动了 AI 赋能的深化融合
2022	人文基因智能计算模型发布	通过机器学习算法解析语言文字载体中的精神文化意识要素，实现从海量文本中自动提取并量化语义与情感颗粒	标注人文基因“成色”度，突破传统符号计算框架，建立认知规律的数学模型， 应对数据稀疏性， 在特征高缺失时仍保持高预测精度	重构基因调控网络式分析框架，可揭示文化图像符号等深层语义密码，推动人文研究从经验判断转向数据驱动的范式革命
2022	ChatGPT 发布，引发全球热潮	基于 GPT 系列模型，引入人类反馈强化学习	如何让大模型与人类进行自然、有用、无害的对话	推动 AI 进入大语言模型时代，使非专业人士也能快速测试语言学假设
2023	sora 发布	基于深度学习与 Diffusion 模型，通过引入时空补丁技术，将视频数据统一为可扩展的潜在空间表示，实现了对不同时长、分辨率和宽高比视频的生成	模拟物理交互的准确性、长时间视频的连贯性（避免角色或场景突变）、跨模态数据（文本、图像、视频）的统一处理，以及训练数据规模与算力需求之间的平衡	标志着 AI 在视频生成领域取得重大突破，推动了影视制作、广告、教育等行业的智能化转型，为通用人工智能（AGI）的发展提供了新的技术路径

时间	重要事件	技术背景	核心问题	影响与意义
2024	DeepSeek 发布	全球大语言模型对高性能与低成本平衡的追求，探索“参数规模与推理成本的最优化”	突破传统模型性能与成本的矛盾，提升模型在复杂推理、跨模态生成等任务上的泛化能力与实用性	推动了 AI 技术的平民化与全球化，为全球 AI 发展提供了“中国式突破”的示范路径

参考文献

[1]. Turing A M. Computing machinery and intelligence (1950) [J]. Mind, 2021, 59(236): 33-60.

[2]. Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[3]. Chomsky N. Syntactic structures[M]. Mouton de Gruyter, 2002.

[4]. Weizenbaum J. ELIZA-a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM, 1966, 9(1): 36-45.

[5]. Brown P F, Cocke J, Della Pietra S A, et al. A statistical approach to machine translation[J]. 1990.

[6]. Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 2002, 77(2): 257-286.

[7]. Yin C, Xi J. Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm[J]. Multimedia Tools and Applications, 2017, 76(16): 16875-16891.

[8]. Wu Y, Zhang Z, Kou G, et al. Distributed linguistic representations in decision making: Taxonomy, key elements and applications, and challenges in data science and explainable artificial intelligence[J]. Information Fusion, 2021, 65: 165-178.

[9]. Shiri F M, Perumal T, Mustapha N, et al. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU[J]. arXiv preprint arXiv:2305.17473, 2023.

[10]. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[11]. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[12]. 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.

[13]. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[14]. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[15]. Hurst A, Lerer A, Goucher A P, et al. Gpt-4o system card[J]. arXiv preprint arXiv:2410.21276, 2024.
[16]. Ahmed J, Nadeem G, Majeed M K, et al. THE RISE OF MULTIMODAL AI: A QUICK REVIEW OF GPT-4V AND GEMINI[J]. Spectrum of Engineering Sciences, 2025, 3(6): 778-786.
[17]. Le Mens G, Kovács B, Hannan M T, et al. Uncovering the semantics of concepts using GPT-4[J]. Proceedings of the National Academy of Sciences, 2023, 120(49): e2309350120.
[18]. McGiff J, Nikolov N S. Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review[J]. arXiv preprint arXiv:2505.04531, 2025.
[19]. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(Feb): 1137-1155.
[20]. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[21]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[22]. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. OpenAI, 2018.
[23]. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint arXiv:2203.02155, 2022.
[24]. Hoffmann J, Borgeaud S, Mensch E, et al. Training language models to be multilingual, predictable, and generalizable via masked language modeling[J]. arXiv preprint arXiv:2211.01786, 2022.
[25]. Bai Y, Schuurmans D, Biderman S, et al. Constitutional AI: Harmlessness from AI feedback[J]. arXiv preprint arXiv:2212.08073, 2022.
[26]. 中国人民大学团队。大语言模型技术综述 [J]. 计算机学报，2023, 46 (5): 921-948.
[27]. 浅谈生成式 AI 语言模型的现状与展望 [EB/OL]. CSDN 博客，2025-09-26.
[28]. 从词袋到大语言模型：AI 语言革命的前世今生 [EB/OL]. CSDN 博客，2025-08-20.
[29]. 回顾 LLM 大语言模型发展历程 [EB/OL]. 腾讯云开发者社区，2025-02-20.

第二章　语言智能学科

2.1　语言智能学科概念的提出

语言智能（Language Intelligence）是一个人工智能范畴的学术概念，最早由首都师范大学周建设教授提出——他是 2013 年在申请并获批成立“北京语言智能协同研究院”之时首次使用此概念[1]。以后，这一概念被学术界广泛采纳、传播与使用。下面，试列举近十余年（2013-2025）来全国冠以“语言智能”字样的各种学术活动、新机构成立活动或其他标志性事件：

❖ **会议与学术活动**

- **2013 年：**首都师范大学召开“北京语言智能协同研究院授牌仪式暨建设规划研讨会”，首次使用“语言智能”术语。
- **2014 年起：**首都师范大学每年主办“中国语言智能大会”，至 2024 年已举办 7 届。
- **2016 年起：**每年举办“语言与智能高峰论坛”，至 2024 年已连续举办了 9 届。
- **2018 年和 2019 年：**北京语言大学语言资源高精尖创新中心和中国中文信息学会社会媒体处理专委会联合主办了两届“’语言智能与社会发展’论坛”。
- **2017 年：**复旦大学、北京语言智能研究中心、中国中文信息学会、中国计算机学会、中国人工智能学会、上海外国语大学和上海交通大学等举办了“语言智能与应用论坛”。
- **2018 年：**中国中文信息学会和中国计算机学会在上海召开的第十一届中国情报语义计算与应用会议上，设立了“语言智能”专题论坛。
- **2019 年 10 月：**在第三届中国北京国际语言文化博览会期间，举办了“语言智能与语言多样性”国际语言文化论坛。
- **2020 年 10 月：**由国家语委中国语言智能研究中心与四川外国语大学联合主办的第五届中国语言智能大会在四川外国语大学举行。
- **2020 年 12 月：**中国语言智能研发暨语言文化教育传播高峰论坛在京举办。
- **2021 年 7 月：**四川外国语大学举行“新文科建设 2021 语言智能暑期学校（研修班）”，为期 11 天。
- **2023 年 5 月：**武汉大学、教育部语言智能技术研究中心和教育部工程研究中心联合主办了“语言智能与教育”学术论坛。
- **2023 年 11 月：**四川外国语大学举办“语言智能与脑科学”主题论坛，期间揭牌成立全国首个“语言脑机接口研究院”。
- **2023 年 12 月：**中国英汉语比较研究会语言智能教学专业委员会成立大会暨语言智能与外语教育融合创新论坛在成都理工大学举行。

- **2024 年 1 月**：“语言智能与教育”高端论坛在湖南师范大学举行。
- **2024 年 5 月**：上海外国语大学举办“大数据与全生命周期语言脑健康”论坛，专家们就语言脑科学领域的前沿问题进行交流。
- **2024 年 5 月**：浙江外国语学院语言智能与认知科学重点实验室举办“语言智能技术与应用”研讨会。
- **2024 年 6 月**：天津外国语大学智能语言服务产业学院和人工智能翻译实验室成立仪式暨智能语言服务天津论坛在天津外国语大学举行。
- **2024 年 9 月**：首届人工智能与人类语言高层论坛暨第七届中国语言智能大会在北京外国语大学举行。
- **2024 年 10 月**：“外语学科智能化转型发展高层论坛暨第十一届全国语言教育研讨会”在在四川外国语大学举行。
- **2024 年 10 月**：《语言战略研究》与《当代语言学》两家期刊编辑部在商务印书馆共同主办了“大语言模型与语言学发展座谈会”。
- **2025 年 4 月**：举办的“《语言文字应用》青年学者论学”第 5 期，主题为“语言智能与技术的发展、应用及使命”。
- **2025 年 1 月**：湖南大学语言智能研究院举办“语言智能与多模态认知”学术论坛。
- **2025 年 5 月**：西安外国语大学语言智能研究中心举办“语言智能与外语教育”研讨会。
- **2025 年 6 月**：四川外国语大学语言智能学院举办“语言脑机接口镜湖论坛暨第十一届全国认知神经语言学研讨会”。
- **2025 年 6 月**：湖南大学语言智能研究院举办“语言智能与文化传承”高端论坛。
- **2025 年 7 月**：浙江外国语学院语言智能与认知科学重点实验室举办“语言智能与语言习得”学术研讨会。

❖ **学术期刊与论文出版**

- **2021 年**：北京外国语大学人工智能与人类语言重点实验室出版了《语言智能教学》英文国际期刊。
- **2023 年**：首都师范大学中国语言智能研究中心创刊了《语言智能研究》辑刊；北京外国语大学人工智能与人类语言重点实验室推出“人工智能与人类语言系列丛书”第一辑的 5 本。
- **2024 年**：北京外国语大学人工智能与人类语言重点实验室推出《语言与智能》中文期刊；上海教育出版社出版了《ChatGPT 来了：语言科学如何看待 ChatGPT》论文集。

❖ **机构成立**

- **2016 年**：教育部语言文字信息司批准依托首都师范大学成立“中国语言智能研究中心”；同年，教育部批准该校自设语言智能学科，开启我国语言智能博士研究生培养。
- **2016 年 8 月**：中国人工智能学会批准成立语言智能专委会。
- **2019 年 3 月**：广东外语外贸大学成立非通用语种智能处理重点实验室。
- **2019 年 4 月**：四川外国语大学成立全国首个“语言智能学院”。
- **2019 年 5 月**：云南财经大学成立云南语言智能研究中心；上海交通大学苏州人工智能研究院与外国语学院为“语言智能联合研究中心”揭牌。
- **2019 年 6 月**：北京语言大学成立语言智能研究院。
- **2019 年 11 月**：大连外国语大学成立语言智能研究中心。
- **2019 年 12 月**：北京外国语大学人工智能与人类语言重点实验室成立。
- **2020 年 5 月**：西安外国语大学成立“人工智能与语言认知神经科学重点实验室”。
- **2020 年 12 月**：上海外国语大学成立数字人文与语言智能实验室。
- **2021 年 9 月**：湖南大学外国语学院开办英语专业语言智能实验班。
- **2023 年 11 月**：江西师范大学举行语言智能研究中心揭牌仪式。
- **2024 年 6 月**：天津外国语大学举行智能语言服务产业学院和人工智能翻译实验室揭牌仪式。
- **2025 年 4 月**：浙江外国语学院举办语言智能学院筹建论证会。

可见，“语言智能”概念一经提出，便呈“燎原”之势，它标志着语言智能学科在我国的应时诞生与发展。

2.2 语言智能概念的基本内涵

“语言智能”概念的内涵极为广阔，目前仍处于发展、探讨之中，尚无定论。我们认为，它可以从狭义和广义来理解。

2.2.1 狭义理解

国内诸多学者都论及了语言智能的基本内涵。胡开宝等认为，语言智能旨在运用计算机技术和信息技术，让机器理解、处理和分析人类语言，实现人机语言交互，使得机器在一定程度上拥有理解、应用和分析人类语言的能力[2][3]。黄立波认为，语言智能研究就是人工智能技术应用于自然语言处理领域的相关研究，相当于“语言加人工智能”[4]。梁晓波、邓祯认为，语言智能指的是“运用计算机信息技术模仿人类的智能，以及分析和处理人类语言文字、声音、知识、情感的科学技术”[5]。李佐文、梁国杰认为，语言智能不是单纯的机器智能，而是以人类语言能力为基础、实现增强与互联的深度人机结合的综合智能行为，是“与

语言相关的智能”的总称，而不限于“单纯的语言技术”[6]。

我们认为，周建设教授本人对语言智能的界定很好地代表了“语言智能”这一概念的狭义内涵。他指出：语言智能是语言信息的智能化，是运用计算机信息技术模仿人类的智能，分析和处理人类语言的过程，是人工智能的重要组成部分及人机交互认知的重要基础和手段[7]。这以后，周教授不断深化、拓展对这一概念的理解，赋予其更大的学科内涵。根据姜孟等所做的概括，周建设的语言智能思想至少包括以下含义：语言智能是基于人脑生理结构和言语认知神经运作机理，利用大数据与人工智能技术，全面认识自然语言属性，对语言信息进行抽取、加工、存储和特征分析，同构人机意识关系模型，让机器模仿人类自然语言活动，实施类人言语行为，让机器具备听、说、读、写、译、评的能力，最终达到人机语言自由交互[8]。用最简明的话来说，语言智能就是让机器“理解并模仿人说话的科学”。在根本上，语言智能是研究人类语言与机器语言之间同构关系的科学，其研究领域十分广阔，至少应包括两个核心研究领域：脑语智能与计算智能。其中，脑语智能研究基于人脑言语生理属性、言语认知路径、语义生成规律，依据仿生原理，构建面向计算的自然语言模型；计算智能研究基于语言大数据，利用人工智能技术，聚焦自然语言模型转化为机器类人语言，设计算法，研发技术，最终实现机器写作、翻译、测评以及人机语言交互[1]。

2.2.2 广义理解

在广义上，语言智能意指“语言自然智能”和“语言人工智能”，是两者的合称。姜孟对“语言智能”的理解就可视作这一广义内涵的代表。他认为，“语言智能”新概念的关键点在于，它是从人类“智能”的角度看待人的语言能力[9]。从研究的维面看，对人的语言能力的研究可从两个方面来进行：一是着眼于人的语言能力的自然维度，把语言能力看作是人类生命体在自然进化基础上的先天、后天交互成就，探究其本质、机制与特征；再就是着眼于人类语言能力的人工维度，脱离人的碳基身体，通过构建一定的物件与装置，来再现、再造、扩展人的语言能力。前者属于 “语言自然智能研究”，后者属于“语言人工智能研究”。在此意义上，“语言智能”这一概念，顾名思义，可以包含两种基本含义。一是“语言自然智能”，即人作为自然生命体的语言使用能力。它与“语言能力”“语言官能”是同义语，同时又与人的认知智能、情感智能、逻辑数学智能等对等相当，是人多元智能的一个组成部分。就研究来说，当今的语言学及诸多交叉边缘学科吗，如神经语言学、心理语言学、计量语言学、语料库语言学等，都属于致力于语言自然智能研究的学科。另一种含义就是，采用人工智能的方法与技术对人的语言自然智能进行人工模仿或“造假”，即“语言人工智能”或“人工语言智能”[10]。当今的“机器翻译”“计算语言学”“自然语言处理”等概念术语所指称的学科，就可视作属于致力于语言人工智能研究的学科。

2.3 语言智能学科概念提出的意义

2.3.1 语言智能研究的历史

人类很早就萌生了让机器拥有人的部分或全部语言能力的思想，相应地对语言智能的探索很早就开始了，只是长期以来没有冠以“语言智能”的名称。正如姜孟等所指出的，“在语言智能这个人工智能概念未提出之前，已有大量研究事实上在‘耕’其‘地’、‘种’其‘田’了”[8]。根据姜孟的梳理，在“语言智能”概念提出之前，对语言智能“有名无实”的研究已经走过了四个历史阶段，跨越了上千年的历史，如图 2-1 所示。

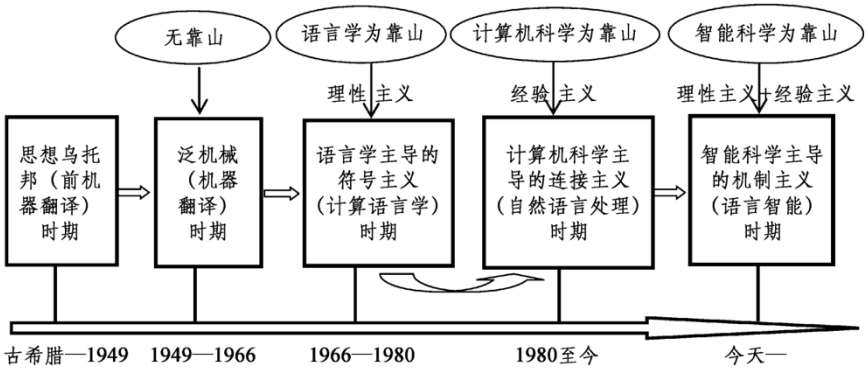

图 2-1 语言人工智能发展的一个历史方位[9]

第一个历史阶段，思想乌托邦（前机器翻译）时期（古希腊——1949 年）。这一时期长达数百上千年。从古希腊人提出利用机械装置进行语言翻译的想法，到 17 世纪笛卡尔（Descartes）和莱布尼茨（Leibniz）提出使用机器词典来实现语言翻译和 17 世纪中叶的“普遍语言”运动，到 1903 年古图拉特（Gouturat）和洛（Leau）采用“数字语法”开展多语翻译并正式提出“机器翻译”术语，再到 20 世纪 30 年代法国工程师阿尔楚尼（Artsouni）提出基于存储装置进行语言翻译的想法以及苏联发明家彼得.特洛延斯基（Peter Troyanskii）提出使用双语字典和语言间的语法角色完成翻译的想法，直到 1949 年“机器翻译之父”沃伦.韦弗（Warren Weaver）在题为《翻译》的备忘录中正式提出机器翻译的思想。语言智能研究的这一历史阶段的整体成就在于：萌生了让语言能力“附身”机器的想法，产生了从机器翻译入手来突破语言人工智能的思想与路径，但总体上还处于思想启蒙、前路茫茫的阶段，离真正的“把思想变现”还有相当距离。

第二个历史阶段，泛机械（机器翻译）时期（1949 年——1966 年）。1949 年韦弗的备忘录从思想上对机器翻译进行了启蒙，引发了大量机器翻译研究实践。最早开展机器翻译研究的有美国的麻省理工学院、乔治城大学和 IBM（国际商业机器公司）以及前苏联的列宁格勒大学、英国的剑桥大学等。1952 年在美国麻省理工学院（MIT）召开了第一次机器翻译会议，1954 年出版了第一本《机器翻译》（Machine Translation）杂志。同年（1954 年 1 月 7 日），美国乔治敦大学和 IBM

公司使用 IBM-701 计算机开发了世界上第一个机器翻译原型系统，成功地将 60 多句俄语自动翻译成英语。在我国，早在 1956 年机器翻译便被列入国家科学工作发展规划。1958 年 8 月中国科学院计算技术研究所牵头成立了机器翻译研究组，与语言研究所合作开展俄汉机器翻译研究。1959 年，我国在自制通用电子计算机上进行的俄汉机器翻译试验获得成功。这一时期，研究者们先是对机器翻译普遍持高度期待、高度乐观的态度，但随着研究与实践的进展，人们发现机器翻译的质量与期望相去甚远，“语义障碍”（semantic barrier）构成了一个绕不过的难题。1960 年，以色列著名的哲学家、数学家和语言学家 Yehoshua Bar-Hillel 更是发表长文指出，由于语义歧义的存在，通用的高质量全自动机器翻译在理论上是不可能的。到了 1966 年 11 月，美国科学院自动语言处理顾问委员会（Automatic Language Processing Advisory Committee，ALPAC）和美国国家研究理事会，发布了题为《语言与机器：翻译和语言学视角下的计算机》的著名报告（也称 ALPAC 报告）。报告正文 30 来页，附件 90 页，对机器翻译做出了基本负面的评价（因此也被称为“黑皮书报告”）。

语言智能研究的这一历史阶段的特点在于，在人的诸种“次语言智能”（次语言能力）中，研究者们好不容易选定“翻译”作为突破口，也好不容易有了当时看来计算能力和存储能力都十分强大的新机械——“计算机”，似乎对语言（翻译）能力的机器实现不再是什么难事，但研究者们很快发现，他们低估了语言的复杂性和机器翻译对计算机技术提出的极高要求：一方面，仅靠人编制的一些词法、句法规则是很难彻底实现机器翻译的目标的；另一方面，机器翻译以至整个语言人工智能研究对技术的要求太高，今后都还要牢牢地受限、受制于计算机技术及其高阶人工智能技术。与此同时，这一阶段还表现出另一特点：语言人工智能研究找到了机械技术做靠山，有了“机巧”，但却没有傍依任何学科，是“无学科”支撑的时期。

第三个历史阶段，语言学主导的符号主义/理性主义（计算语言学）时期（1966年——1980）。1966 年美国 ALPAC 报告的发布，使“机器翻译”这一当时最亮眼的语言人工智能工程的资金被大量消减或中断。面对困难与困境，语言人工智能研究被迫另寻他途。这里的一个机巧在于，ALPAC 报告在基本否定机器翻译可能性和前景的同时，也给出了研究者可以和应该用力的方向。ALPAC 报告的主起草人之一、美国语言学家戴维.海斯（David Hays）在报告中给出建议：在放弃机器翻译这个短期工程项目的时候，仍有必要加强语言和自然语言计算机处理的基础理论研究，应当把原来用于机器翻译研制的经费使用到自然语言处理的基础理论方面[11]。所谓的“自然语言处理的基础理论方面”正是指语言计算方面的研究，海斯把它命名为“计算语言学”（Computational linguistics）。这是“计算语言学”概念首次在官方报告中使用。随后，1967 年，戴维.海斯出版专著《计算

语言学导论》（Introduction of Computational linguistics），宣告了计算语言学时代的正式到来。机器翻译研究走上了先做好语言和语言计算基础理论研究再图将来的迂回发展之路。此后，1968 年，国际“机器翻译和计算语言学学会（AMTCL）”删除“Machine Translation”两个词直接更名为“计算语言学学会”（Association for Computational Linguistics, 简称 ACL），并一直沿用至今。实际上，在 ALPAC 报告发布前的 1965 年，在美国纽约就成立了单独以 Computational Linguistics 冠名的国际计算语言学委员会（International Committe of Computational Linguistics, 简称 ICCL）。该学会每两年召开一次国际会议，会议名称为“International Conference on Computational Linguistics”，简称 COLING。与此同时，美国出版了学术季刊《美国计算语言学杂志》（American Journal of Computational Linguistics），后改名为《国际计算语言学杂志》（International Journal of Computational Linguistics）。

从研究实际来看，在 1966 年以后的十多年间（1966-1980），无论是机器翻译还是整个计算语言学研究，走的也确实都是聚焦“从语言学角度，分析自然语言的词法、句法等结构信息，并通过总结这些结构之间的规则，达到处理和使用自然语言的目的”的路子[12]。这一路子，实质上是美国语言学家乔姆斯基和他所提出生成式文法所开创的“符号主义”的路子，也就是文献中常提及的“基于规则的”研究路子。因此，语言智能研究的这一阶段的特点是，机器翻译遭遇了极大困境，发生了向计算语言学的“转拐求生”——即寻求以语言学做“靠山”，希冀先在语言本体特征与规律研究方面获得突破，有一天最终获得机器翻译上的突破。换一个角度看，“机器翻译”怎么看也都只能是人类探索语言人工智能的一个小瞬间，是一个历史“小不点”。前进的步伐注定要超越机器翻译的小天地，走向更大的海阔天空。

第四个历史阶段，计算机科学主导的连接主义（自然语言处理）时期（1980 年——至今）。在上一历史阶段，基于规则的语言人工智能研究获得了长足发展。但随着人们越来越多地关注工程化、实用化的解决问题的方法，这一研究路径日渐遭遇了困难。就拿机器翻译来说，人工确定的翻译规则越来越复杂，规模库也越来越大，但数量却有限，对千变万化的复杂语言现象的解释力难以继续提高，译文的准确率也无法持续改善，语言人工智能研究又来到了一个新的历史拐点。这一次，研究者寻找的是以概率统计为核心特征的“技术”靠山，傍依的是快速发展的计算机科学，走的是“连接主义”的研究路子，最能反映这一发展现实的学科旗号是“自然语言处理”（natural language processing）。

从 20 世纪 80 年代开始，一些研究者就开启了离开符号主义而另走他途、再寻前程之旅。1980 年马丁（Martin Kay）提出了翻译记忆（translation memory, TM）的方法，尝试从已经翻译好的文档中找出相似部分来帮助新的翻译。1984 年长

尾真（Makoto Nagao）提出了基于实例的机器翻译方法（example-based machine translation, EBMT），着手从实例库中提取翻译知识，通过增、删、改、替换等操作完成翻译[14]。这些研究试所尝的都是全新的基于数据驱动的机器翻译方法，堪称语言人工智能研究走向新的历史方位的先声。

与此同时，计算机硬件技术迅猛发展，存储容量快速扩大，运算速度日益提升，统计机器学习的新理论、新方法不断涌现，语料库技术也日臻成熟，使得很多原来无法实现的复杂问题现在都可以借助“硬件技术+统计模型+语料库”很容易地实现。1990 年，在芬兰赫尔辛基召开了第 13 届国际计算语言学大会，提出了处理大规模真实文本的战略任务，开启了语言计算的一个历史新阶段——基于大规模语料库的统计自然语言处理。在此潮流的带动下，1993 年，美国 IBM 研究院发表论文“统计机器翻译的数学理论：参数估计”（The mathematic of statistical machine translation: Parameter estimation），提出了 IBM Model 1-5。1999 年，美国约翰.霍普金斯大学（The Johns Hopkins University）发布了 GIZA 软件包，把 IBM Model 1-5 变为了现实。随后，更加复杂的 IBM Model 6、更加优化的软件包 GIZA++都先后发布。IBM 统计机器翻译模型得到广泛使用。该模型几乎完全依赖大规模双语语料库，通过词对齐、短语对齐等手段，来自动构造统计机器翻译系统。较之基于规则的机器翻译系统，统计机器翻译系统的性能显著提升，而且很容易地就实现了数十个语言对之间的翻译。在商业上，很快催生了谷歌、百度等公司的互联网机器翻译系统。由于统计机器翻译模型与具体语种无关，不再需要“规则集”，设计者可以完全不懂相关的语言。机器翻译研究走上了离语言学研究越来越远的道路。为此，著名的机器翻译学者、Google Translate 的设计者弗兰茨.约瑟夫曾信心满怀地声称：“只要给我充分的并行语言数据，那么，对于任何两种语言，我就可以在几小时之内给你构造出一个机器翻译系统”（Give me enough parallel data and you can have translation system for any two languages in a matter of hours）。

2006 年深度神经网络反向传播算法被提出。2014 年前后，随着深度学习技术在语音、图像领域的研究取得成功，深度学习方法也开始应用于语言人工智能研究。在机器翻译领域，出现了“神经机器翻译”（Neural Machine Translation, NMT）模型。该模型采用基于深度神经网络的方法来构造机器翻译系统，其架构由编码器和解码器两部组成。首先由编码器把源语言的句子表示为词向量（word vector），形成句子的分布式，然后利用解码器依次生成目标语言的单词序列，直到生成目标语言的整个句子为止。神经机器翻译采用的是端到端（end to end）的计算过程，由于其内部是由基于词向量的数值计算构成的，难以从语言学的角度解释中间过程的计算机制，翻译成为一个黑箱操作过程[13]。相比统计机器翻译，神经机器翻译具有更加广泛的一般性，更加远离具体语言的知识。机器翻译走上了与语言学

研究几乎彻底分道扬镳的道路。在技术上，神经机器翻译具有更大的“颠覆性”，它使机器翻译的质量再次迅猛提升。当然，它涉及的计算量也更大，更加依赖计算能力（“算力”），需要借助特殊计算设备（如 GPU）才能高效地实施参数训练。进入 20 世纪 90 年代以后，整个语言人工智能领域都深深地打上了概率统计的烙印。正如冯志伟所指出的那样，“概率和数据驱动的方法几乎成为了计算语言学的标准方法[14]。句法剖析、词类标注、参照消解、话语处理、机器翻译的算法全都开始引入概率并且采用从语音识别和信息检索中借过来的基于概率和数据驱动的评测方法。”

这一历史阶段实际上是语言学主导的符号主义研究方法在其红利快要耗尽之时，语言人工智能研究被迫再次寻觅新的靠山与发展空间，它走上了几乎完全依靠概率统计、深度神经网络以及计算机硬件技术来获得新突破的路子。研究者所积累的语言学和计算语言学专业知识变得几乎完全不起作用，语言学被边缘化，计算机科学成了主宰，认知科学中的“连接主义”成了新的标示与旗号。

2.3.2 语言智能学科概念提出的价值与意义

姜孟认为，“语言智能”概念的提出具有两方面的重要价值和意义。一是指示语言人工智能研究的历史新方位，二是改变现有语言智能研究“名不符实”的现状[9]。

首先来看这一术语对指示语言人工智能研究历史新方位的作用。如前所述，语言人工智能研究已经历经四个历史阶段。这四个阶段，代表着人类为攻克、实现人的离身语言智能之梦而发起的四次大尝试、大冲锋，实质上代表了语言人工智能研究的四个历史方位。第一个历史方位“思想乌托邦（前机器翻译）”时期产生了“引”人的翻译智能于人的肉身之外的奇思妙想，完成了逐梦中的思想先行、不怕做不到就怕想不到的历史跨越。第二个历史方位“泛机械（机器翻译）”时期以新诞生的计算机为机械装置的典型代表，把第一个历史方位中产生的奇思妙想变成了初步现实，完成了万事开头难、不仅想到还真正做到的第二步历史跨越。第三个历史方位“语言学主导的符号主义（计算语言学）”时期把“离身翻译智能”之梦扩展为“离身语言智能”之梦，并在挫折中尝试傍依语言学，以其为靠山，于无路可走中走符号主义之路，完成了从小到大、从无依无靠到有学科依靠的第三步历史跨越。第四个历史方位“计算机科学主导的连接主义（自然语言处理）”时期在傍依语言学遭遇困难后，尝试改寻计算机科学为靠山，以概率统计为主导，绕开语言的语义难题，走连接主义之路，完成了东方不亮西方亮、天无绝人之路的第四步历史跨越。然而，在经历过四次大跨越、取得卓著成绩的同时，语言人工智能研究也来到了深水区，面临更深层、更具挑战性的问题与困难，其解决需要在新的历史方位寻觅新的靠山，尝试新的路径，实现新的跨越。我们认为，这一新的历史方位就是“智能科学主导的机制主义”。

放眼当下，语言人工智能研究在语言智能的内在机制与底层原理上着力有限，所取得的突破性进展不多，整体上采取的是一条“绕道走”的研究路线。首先，从研究的内容实质与达到的目标高度来看，当前的语言人工智能研究还处于“有多少人工就有多少智能”阶段，距离真正的人的语言自然智能还有相当距离。今天，在语言人工智能研究领域里最具神通、最有影响的研究方法是基于统计的自然语言处理与基于人工神经网络的自然语言处理。但无论哪一种方法，它们所做的都还只是“浅层的”自然语言处理，相关技术还只能支持完成“浅层句法”或“简单标记”任务，对于更复杂的语言现象理解、语义关系抽取以及更专业的语言资料处理还一筹莫展。正如徐英瑾所指出的，目前语言人工智能研究存在四大不足：（1）不同的自然语言处理机制之间缺乏融合；（2）自然语言处理技术与人工智能研究的其他技术缺乏彼此融合；（3）基于大数据的自然语言处理技术的运作必须以“剥削”人类的智能为前提；（4）基于大数据的自然语言处理技术缺乏灵活处理隐喻、反讽、双关等修辞现象的能力[15]。言下之意，今天计算机所展示的语言智能在本质上只是人类智能的“反光映照体”，就好比月亮只是太阳的“反光映照体”一样，距离语言智能的根基还很远。

再从研究的学科理路与构架体系来看，语言智能研究在四个历史方位中所经历的起起落落基本上只是应声于人工智能研究的起起落落，尚未形成一个理论系统、方法成熟、技术完备的学科领域格局与面貌，独立自主性不够。宗庆成一针见血地指出，综观整个自然语言处理领域，还远未建立起“一套完整、系统的理论框架体系”，不少理论研究还只是处于盲目的探索阶段，对一些新的机器学习方法或未曾使用过的数学模型的尝试，还带有很大的主观性和盲目性[16]。钟义信在检视人工智能研究的不足时指出，现行人工智能研究路径或者是结构主义（模拟脑的结构，也即连接主义），或者是功能主义（模拟脑的功能，也即符号主义），或者是行为主义（模拟智能系统的行为，也即感知动作系统），尚未实现统一[17]。他认为，目前不同范式的人工智能以不同的认识论学派为根基，所形成的是各种专用的人工智能。要实现通用的人工智能，需要融合各派的认识论作为理论基础，需要以符号主义、连接主义和行为主义不同哲学路向加以贯通整合的新型认识论为导引。要致力于创建“结构—功能—行为”整合融通、“意识—情感—理智”三位一体的通用人工智能理论。他把这种新的研究理论与范式称作“机制主义”。我们认为，这也适合于当下的语言人工智能研究。针对当前的问题与瓶颈，语言人工智能研究需要主动走进新的历史方位，以当今的智能科学为靠山，立足语言作为人的智能的本质与机理，从其底层原理入手，系统建构语言人工智能独立的学科逻辑、理论体系与方法技术体系，走一条新的“机制主义”研究之路。

宗成庆说得好，“自然语言处理毕竟是认知科学、语言学和计算机科学等多学科交叉的复杂问题，当我们从外层（或表层）研究语言理解的理论方法和数学

模型的同时，不应该忽略从内层揭示人类理解语言机制的秘密，从人类认知机理和智能的本质上为自然语言处理寻求依据[16]。”以色列学者舒利.温特尔（Shuly Wintner）也批评了当今语言人工智能研究对语言机制的严重忽视[19]。她指出，当前的自然语言处理工程已经把语言学看得可有可无，研究的几乎都是程序技术或算法问题，很少关注自然语言处理工程背后隐藏的语言本身的基础性问题。“机制主义”研究之路的实质是从智能科学的宏大视野与高度来探索语言人工智能。语言具有复杂深邃的心脑机制，以其为观照，方能真正洞察当今语言人工智能研究的得失。

概言之，语言人工智能研究已经走到了一个前所未有的深水区，到了直面人类语言智能的本质、内在机制与底层原理的时候了，“绕道走”的方法已经难以维系。与此同时，脑科学、生命科学等方面的新进展已经使智能科学应运而生。作为智能科学的核心关切与主打领域方向之一，语言人工智能研究需要逐步走上“机制主义”的发展路线，并在智能科学的框架内成长为一个比较独立的分支学科——语言智能科学。“语言智能”概念的提出，顺应了语言人工智能的研究的这一历史动向，是导引其走向以智能科学为靠山的“机制主义”历史方位的恰当旗帜。

再来看，“语言智能”术语对改变语言智能研究“名实不符”现状的作用。语言人工智能研究兴起于 1956 年在美国召开的“达特茅斯会议”[12]，至今已七十多年。但长期以来，该领域一直缺少一个有统摄力的学科名称，多个概念术语交叠瓜葛，纷争抵牾，各自为阵，但最近出现了向智能及智能科学相关名称聚靠的趋势。

在语言人工智能研究的四个历史方位中，有三个比较有影响的术语。第一个是“机器翻译”。这一概念由古图拉特（Gouturat）和洛（Leau）于 1903 年提出（如前所述），距今 120 余年，它比今天的“人工智能”（Artificial Intelligence）概念都早了 50 余年（按 1956 年达特茅斯会议正式提出此概念算），历史厚重。但“机器翻译”是指“使用机器（计算机）自动地将一种自然语言（源语言）的语句转化为相同含义的另一种自然语言（目标语言）的语句的过程”[13]。显然，这一概念突显的是人的听、说、读、写、译多项语言智能中的“译”的智能，仅关注人的翻译智能的机器实现（即模仿、延伸与扩展），只涵盖了人类语言自然智能之五分之一，无力承担指称语言人工智能这一被誉为“人工智能皇冠上的明珠”的前沿交叉学科全域的重任。

第二个术语是“计算语言学”。该术语于 1966 年由美国科学院在 ALPAC 报告中正式提出（见前文）。该术语在早期可以说仅是强调计算机对于自然语言分析与统计的辅助工具作用，还没有把计算机上升到“可以提供主动服务，能够帮助人类达到对话、翻译、检索等若干目的的智能工具”的高度[16]。只是到了后来，

语言人工智能的研究实践远远超越了计算语言学这一概念的原初内涵与范围，但却又找不到一个相称、合适的术语之时，这一概念才被迫用作统摄语言人工智能研究全域的一个术语，颇有些“不欲戴王冠却得承其重”的意味。这可见于后来一些学者对“计算语言学”所做的理解与解释中。例如，克里斯特尔（2000）在《现代语言学词典》中给“计算语言学”下的定义是：语言学的一个分支，用计算技术和概念来阐述语言学和语音学问题。所开发的领域包括自然语言处理、言语合成、言语识别、自动翻译、编制词语索引、语法的检测，以及许多需要统计分析的领域。宗成庆的定义与解释也体现了这一点：“计算语言学实际上包括以语音为主要研究对象的语音学基础及其语音处理技术研究和以词汇、句子、话语或语篇及其词法、句法、语义和语用等相关信息为主要研究对象的处理技术研究[16]。”可见，“计算语言学”原本就是一个偏重语言技术方法研究方面的概念，并非指称统揽语言人工智能昨天、今天和明天广阔研究实践的理想术语。

第三个术语是“自然语言处理”（Natural Language Processing, NPL），也称“自然语言理解”（Natural Language Understanding, NLU）。该术语源自机器翻译，但严格说它算不上一个专业术语，而仅是对某种研究对象与内容的描述。正如莫宏伟等所指出的，从机器翻译开始，人工智能领域发展出了自然语言处理“这一研究内容”[19]。根据冯志伟的定义，自然语言处理就是利用计算机为工具，对人类特有的书面形式和口头形式的自然语言的信息进行各种类型的处理和加工的技术[20]。因此，在本义上，“自然语言处理”是一个技术至上的概念。用宗成庆的话来说，自然语言处理似乎“包含的语言工程和应用系统实现方面的含义似乎更多一些”。然而，语言十分复杂，当今人工智能学科对语言开展研究的范围又十分广阔，涉及人类语言机器实现的任何技术都隐含着当今自然语言处理或计算语言学的问题，广泛牵涉其它学科领域的方法与技术，如信息检索、舆情分析、文字识别、社交网络、社会计算、情感计算、语言教学、口语考试自动评分等[16]。因而，自然语言处理天生就不限于自然语言处理。在现实中，同样由于缺乏更恰当的概念术语，与“计算语言学”概念一样，“自然语言处理”也被揠苗助长般地用于指称当今语言人工智能研究全域。在现实中，自然语言处理所涵盖的研究内容与范围已经十分广大，仅从应用目的的角度，它至少包括：机器翻译、自动文摘、信息检索、文档分类、问答系统、信息过滤、信息抽取、文本挖掘/数据挖掘、舆情分析、隐喻计算、文字编辑与自动校对、作文自动评分、光读字符识别、语音识别、文语转换、说话人识别/认证/验证等。严格意义上，它还应包括语音技术，即语音识别、语音合成和说话人识别等[16]。黄河燕等也认为，自然语言处理的研究对象几乎涉及语言学研究的所有对象：语音、形态、语法（句法）、语义、语用，研究内容包括针对这些对象的自动分析方法与技术，如词法分析、句法分析、语义分析等[21]。

可见，学界当前的两个核心概念“自然语言处理”与“计算语言学”都算不上是不是指称语言人工智能研究全域的理想术语。况且，两者间还存在一定的抵牾。自然语言处理曾一度被视作计算语言学的分支领域，是其下位概念。例如，《现代语言学词典》（克里斯特尔,2002）就把自然语言处理视为与言语合成、言语识别、自动翻译、编制词语索引、语法的检测、文本考释等并列的计算语言学的开发领域之一。但随着近年来自然语言处理被视作语言人工智能研究全域的代名词，它已经超越作为计算语言学范畴内的一个研究分支的地位，而被看作“基本上是处于同一个层次上的概念”[16]。两者（连同“自然语言理解”）很多时候都被视作同一个概念，至少在其外延上不再细究其差异。用宗成庆的话来说，“在很多情况下我们很难绝对地区分开‘计算语言学’‘自然语言理解’与‘自然语言处理’三个术语之间到底存在怎样的包含或重叠关系以及各自不同的内涵和外延。”[16]。它们被视为是等同的，即自然语言处理就是自然语言理解，也就是计算语言学。

正是由于语言人工智能研究陷于“名不正”的情势与地位，近年来一些学者提出了好几个新术语，以回应、矫正、补足这一现实，更好满足各自研究的需要。这些术语几乎都是通过对原有术语掐“头”去“尾”、添“智”加“能”的方式创生的，智能化的趋势与倾向明显。例如，黄河燕等（2020）就提出了“语言智能处理”的概念，以更精准地指称当今被高度智能化、带有显著智能特征的自然语言处理的研究现实与实践。张雄伟等（2020）提出了“智能语音处理”（Intelligent Speech Processing）的概念，耿立波等（2014）提出了“机器语言能力研究”的概念。

简而言之，近年来，语言人工智能的研究实践已经远远走在了指称该学科领域的现有名称术语之前，“计算语言学”“自然语言处理”等名称术语已经滞后落伍，名实已不相符。“名不符实”必然要求“名实相符”。在此情形下，“语言智能”概念得以提出。这一概念从“人类智能”高度来看待人的语言能力，至少有两个优势。一是，“语言智能”天生包纳人类语言能力自然之维与人工之维，即“语言自然智能”与“语言人工智能”的双重含义，具有术语统摄优势。二是这一概念有望超越前述“计算语言学”“自然语言处理”等概念所内蕴的历史方位的局限性，更好地匹配、相称于当今语言人工智能研究的广阔现实，具有学科统摄优势。从长远来看，语言智能概念的提出意义重大。它拓新了传统语言学研究和传统语言人工智能研究（计算语言学、自然语言处理等）的学理视野与跨学科坐标，不仅有助于促进语言学学科的与时俱进与智能化变革，提升语言智能分支在整个人工智能学科领域中的地位和影响，也有助于把语言学、脑科学、认知科学、生命科学等“语言自然智能”相关学科以及计算机科学、人工智能等“语言人工智能”相关学科，都纳入语言智能的基础支撑学科范围内，从而促进语言学

与人工智能两大学科领域未来更多的交叉互鉴，最终提升我国在人工智能领域的研究水平与综合实力。

2.4 语言智能学科框架

“语言智能”概念提出后，逐渐成为一个学科即语言智能学科的代名词。然而，语言智能尚是一个正在形成中的学科，其研究对象、目标、任务、内容、范围、方法、学科性质等都还在探讨中，迄今为止，尚未定型。本报告根据近年来诸多学者的研究与论述，试将其梳理、总结如下：

2.4.1 研究对象

周建设、薛嗣媛（2023）界定了语言智能学科的研究对象。他们认为，语言智能是研究人类语言与机器语言之间同构关系的科学，是以人类语言活动元素、活动机制、表现形式为模仿对象，致力于对语言生成、传递、理解、翻译、评测的信息处理过程和本质进行研究。相应地，语言智能学科以语言智能为研究对象，系统研究语言智能的学科内涵、基础理论、关键核心技术及其在不同场景中的创新应用。其研究目标是从认识人类语言与计算机智能语言的本质属性开始，探索自然语言和智能语言的多元形态，从语言的多元表现形态探索人脑意象与语言图像的同构认知模式，从识别脑语同构模式考察词项与句子的组配关系，从识别词句组配模式探索语言生成的运行机制，从语言表达模式探索图像生成模式，从语图转化模式探索文化基因，构成“语言形态→脑语同构→词句组配→语言生成→语图转化→文化发现”的研究路线，最终实现机器模仿人类智能，理解、应用、分析人类语言，实现人机语言交互[22]。

李佐文和梁国杰也持类似观点。他们认为，语言智能学科以语言智能为研究对象，系统研究语言智能的内涵和构成、理论和机制、关键核心技术及其在不同场景中的创新应用，致力于更深入地探索人类语言智能的本质，更有效地使用现代信息技术模拟人类的语言理解、使用和学习能力，开发研究语言智能处理前沿技术，探索语言智能技术在语言教学、机器翻译、人机对话等领域的创新应用，解决大数据时代的语言智能基础理论建构和技术应用问题，回应人工智能发展对语言学和语言技术提出的挑战[23]。

2.4.2 研究内容与范围

语言智能学科的研究内容很多，范围很广，基本的论题包括智能语音、机器翻译、知识图谱、智能内容生成、主题聚合度计算、情感计算、人文基因计算、写作智能评测，等等[22]。但总体上主要开展三个层面的研究：基础理论研究、关键技术研究以及应用创新研究[23]。

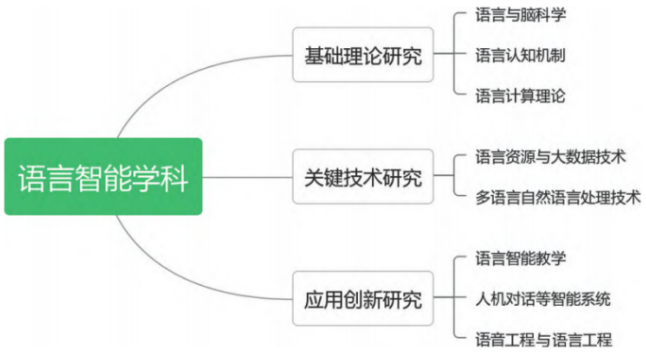

图 2-2 语言智能学科研究内容图[23]

基础理论研究主要聚焦语言脑机制、语言认知机制与语言计算理论三个方向的研究。语言脑机制研究主要关注人类语言处理的神经生理基础，研究大脑语言功能发展、多语言学习的脑加工机制、语言病理机制，以及语言产生、接收、分析和储存的神经机制等。语言认知机制研究，主要关注人类语言处理的认知心理基础，研究语言理解、生成与使用的认知加工机制，以及语言能力和言语交际的认知解释等。语言计算理论研究，面向计算机智能处理的理论和分析方法，研究词汇、句法、篇章、语义、语用层面的语言知识形式化表示，构建面向智能计算的语言模型等。

关键技术研究主要是对语言资源与大数据技术、多语言自然语言处理技术等进行研究。语言资源与大数据技术，是指利用计算机对语言数据进行抽取、存储、标注与研究，以便智能高效地处理语言大数据，构建语言资源。多语言自然语言处理技术，则是以计算机为基本工具，研究自然语言的分析、理解、生成与获得技术，实现有效的自然语言语义、情感和意图计算。

语言智能的应用创新研究，是指将理论与技术融合起来，开发为语言智能工具，用于语言的技术处理，为全场景、全学科、多主体、个性化的语言活动提供智能化指导性服务。它主要包括语言智能教学、智能系统人机对话、语音工程与语言工程等方面的研究。

2.4.3 研究方法

李佐文和梁国杰指出，语言智能学科主要采取三种研究方法[23]。一是基于形式模型的语言计算方法。该方法致力于把语言学问题用数学方法加以形式化，并表示为算法，建立语言的形式模型，以使自然语言成为计算机直接处理的对象。基于形式模型的语言计算是语言智能研究的基本方法。离开形式化的语言模型，智能语言处理便无以实现。

二是基于认知神经科学的语言实验方法。人类语言加工的脑神经机制，是目前制约语言智能研究实现突破的关键因素之一。通过实验方法来探究和揭示人类处理自然语言的认知神经机制，是语言智能研究实现突破的前沿领域。

三是基于语言数据的机器学习方法。在自然语言处理领域，无论是语料自动

标注和词义消歧和指代消解还是机器翻译、人机对话等，都离不开基于高质量语言数据的机器学习方法。只有先进的算法模型，匹以高质量的语言数据，语言智能处理的效率和质量才能得以提升。

2.4.4 学科性质与定位

语言智能是一个涉及多个学科的交叉学科，在学科定位与归属上既可视作人工智能的分支学科[19]，也可视作语言学的分支学科[9]。周建设论述了语言智能的交叉学科性质。他指出，语言智能是机器理解并模仿人说话的科学，是研究人类语言与机器语言之间同构关系的科学，是一个融合神经科学、认知科学、思维科学、哲学、逻辑学、心理学、语言学、计算机科学等多个学科的新兴交叉学科。语言智能基础理论与关键技术研究的突破，对于整个人工智能取得突破性进展具有重要意义[24][25]。李佐文和梁国杰则持与莫宏伟和徐立芳相同的观点，即都认为语言智能是智能科学的重要组成部分，它应当与智能科学一样，在脑科学、认知科学和人工智能技术三个层面上开展研究，既包括机理和模型的探讨，也包括功能的仿真[23][19]。其中，脑科学从大脑的结构和功能的角度揭示语言习得和使用过程中的神经活动规律，是语言智能研究的基础。认知科学研究人类智能的本质和规律，探讨语言使用过程中的概念语义形成的机制和过程、具身性特征等。人工智能通过语音识别、自然语言处理等技术模拟和实现人类的语言能力。三个方面共同构成语言智能的主要学科基础。

对于语言智能学科的定位与使命，周建设做了精辟论述。他指出，作为国家的一种新兴学科，语言智能学科旨在“发展语言智能科技，培养语言智能人才，推进语言学科教育智能化，促进教育高质量发展，助力国民语言能力提升和人文素养提升”[1]。

2.5 语言智能学科建设现状

2.5.1 总体概况

自“语言智能”学科概念提出以来，学者们围绕语言智能的概念与内涵、语言智能与大数据、语言智能与国家语言能力、语言智能与外语教学、语言智能人才培养等进行探讨，使语言智能的概念和语言智能学科建设的重要意义受到了广泛关注,形成了比较明确的研究方向，具备较为稳定的教学、科研队伍，也产出了不少技术应用成果[6]。可以说，语言智能目前在我国已经受到前所未有的重视，正在成为一个充满生机与活力的新质研究领域。

迄今，国内多所高校在“中国语言文学”“外国语言文学”“教育学”“社会学”等一级学科下设置了语言智能专业，或直接设置为二级学科（如四川外国语大学），或作为“语言学及应用语言学”“智能科学与技术”“教育技术”等二级学科下的研究方向，代表性的高校包括首都师范大学、四川外国语大学、北京外国语大学、上海外国语大学、北京语言大学、北京大学、上海交通大学、北京师

范大学、西安外国语大学、鲁东大学、江苏师范大学、苏州大学等。这些学校要么设置了语言智能学科方向并开始招生硕博士研究生，要么在本科或研究生层面开设了语言智能相关课程[22][8]。

2.5.2 建设案例

案例一：首都师范大学

（1）学科概述

首都师范大学“语言智能”学科作为一门前沿交叉学科，于 2016 年 6 月正式设立。该学科主要依托于该校中国语言文学（A 类学科、国家重点学科）、汉语言文字学（北京市重点学科）以及计算机科学与技术（含“计算机应用技术”和“通信与信息系统”两个北京市重点建设学科）等核心优势学科资源，深度融合语言学、人工智能、认知科学等多学科理论体系。学科以“语言智能理论创新与应用落地”为核心发展理念，紧密围绕国家在人工智能与语言技术领域的战略需求，致力于系统性培养兼具深厚语言学功底与先进智能技术素养的复合型高端人才。作为国内首个正式设立的“语言智能”交叉学科博士点，该学科依托构建的国家级科研平台——国家语委科研基地“中国语言智能研究中心”，以及省部级重点科研平台——“北京市语言智能协同研究院”等全方位支撑体系。在此基础上，形成了以基础研究为核心驱动、技术研发为关键纽带、应用服务为最终导向的“三位一体”发展模式，有效促进科研成果高效转化与社会实践应用。

（2）发展历程

首都师范大学“语言智能”学科发展历程清晰，可划分为三个关键阶段。作为国内首个正式设立的语言智能交叉学科博士点，其发展植根于深厚的学术传统与跨学科创新实践。以下结合学术文献与机构史料进行系统梳理：

①基础奠基期（1980–2010 年）

此阶段以文学院为主体，在现代汉语、古代汉语等传统语言学方向进行深入研究。学科建设的重要里程碑为 2003 年成功获批设立“语言学及应用语言学”博士学位授权点，成为当时全国仅有的 8 个该领域博士点之一，为后续交叉学科发展奠定了高层次人才培养基础。值得注意的是，周建设教授在 2001 年 11 月 19 日《人民日报》刊载的留学人员访谈录《探索语言奥秘的人》中，已前瞻性地提出人工智能应高度重视语义处理的战略思考，为后期语言智能的历史语言学资源建设奠定了基础。

②体系形成期（2011–2016 年）

此阶段是学科体系化建设的关键时期。2010 年，学科团队承担国家社科基金重大招标项目“自然语言信息处理的逻辑语义学研究”(10&ZD073)。2013 年，学校整合文学院及信息工程学院力量，组建北京语言智能协同研究院，由教育部专家周建设教授担任学科带头人，形成“语言本体研究—计算模型构建—应用系

统开发”的全链条研究模式。该研究院联合北京大学、北京语言大学等单位主办国际学术会议 3 次，发表研究论文 108 篇。2014 年，首次提出“汉语表达智能模型三大特性”（大数据的“基因”储存性、规律蕴涵性和趋势可预测性），相关成果发表于《首都师范大学学报(社会科学版)》。

2016 年为标志性年份，国家语言文字工作委员会（简称“国家语委”）中国语言智能研究中心获批设立，周建设教授担任主任。同年 6 月，经教育部批准备案，正式获批设立语言智能博士点，成为国内首家语言智能博士点。学科团队制定并完善了语言智能交叉学科博士研究生人才培养方案，课程体系包含《语义学语法学》、《语言智能原理》、《不确定人工智能》等跨学科课程。该学科是在人工智能、大数据技术交叉融合背景下发展起来的二级学科，隶属于中国语言文学一级学科，下设语言智能理论与方法、语言智能处理技术以及语言智能应用研究三个方向，旨在培养面向国际前沿水平的语言智能专门人才。

学科团队与哈佛大学、斯坦福大学、汉堡大学、德岛大学、清华大学、北京大学、中国科学院及科大讯飞等机构保持良好合作关系，逐步形成研究与应用特色，培养了一支水平较高、勇于创新的复合型教学科研队伍，为培养具备创新与实践能力的研究生人才提供了有力支撑。

③快速提升期（2017 年至今）

进入此阶段，学科发展步入提质增速阶段。2017 年，学科团队建立语言智能国际联合实验室和全球汉语智能教育中心，由教育部副部长刘利民、中国人工智能学会理事长李德毅院士授牌。国际联合实验室以需求为导向，建设国内领先的前沿性语言智能研究体系，聚焦国内教育均衡与海外汉语传播，开展语言智能与文化融合创新模式研究、语言智能关键技术及语言资源大数据分析研究、面向全球汉语智能教育服务研究三大方向，并以“国际联合”产业培育项目形式促进成果转化。2018 年开始招收首批博士生，迄今已招收博士生 24 名，毕业 5 名，在读 19 名。毕业生主要就职于中国社会科学院、中国科学院、中国科学技术大学等知名高校与科研院所，培养质量获得社会广泛认可。2018 年及 2019 年，学科团队研究项目“语言智能关键技术及应用研究”分获北京市科技进步二等奖、第九届吴文俊人工智能科技进步奖一等奖。

2021 年起，学科依托承担的国家科技创新 2030“新一代人工智能”重大项目《复杂版面手写图文识别及理解关键技术研究》、国家社科基金重大项目、国家语委重大项目等国家战略需求，围绕语言智能理论创新、关键技术突破及行业应用探索，深入开展科学研究与人才培养。在学术引领方面，初步形成了中国智能教育大会、中国语言智能大会两大品牌，在学界与业界产生重要影响。在成果转化方面，对接国家战略，聚焦教育均衡问题，以周建设教授为首席科学家的科研团队，基于首次提出的语言智能“阅读-写作-评测”六定理论，突破关键技术，

构建了大规模本体知识库、语料库和规则库，有效解决了“阅读-写作-评测”中的语义理解、逻辑推理及语法搭配问题。构建的“六定”智能辅助阅读模型、汉语智能表达模型、全信息语言评测模型，分别攻克了阅读评级与精准推送、汉语语篇生成、人工评测与机器评测拟合度低等技术难题。自主研发的中文智能辅助阅读系统、中文智能写作系统、中英文作文评测系统等智能教育产品达到国际领先水平，解决了语言智能在教育应用中的核心问题，在中英文语料库构建及评测精准度等方面处于国际前沿。相关成果在教育领域实现大规模应用，精准批改中英文作文逾 11.01 亿篇次（其中中文作文超 8000 亿篇次），服务近 8000 万名学生、6 万余所学校，取得直接经济效益 17.01 亿元。成果覆盖北京市海淀区、朝阳区等多个区县中小学，占据大陆地区 90%以上市场份额，推广至台湾地区 230 所学校，并拓展至韩国、越南等 165 个国家的海外市场，有力推动了语言智能学科及智能教育事业发展，为国家教育均衡化服务做出了重大贡献。

（3）问题与挑战

当前语言智能学科的发展亟需突破以下三方面核心挑战：

首先，跨学科深度融合存在显著不足。语言学与人工智能研究团队间长期形成的理论范式差异，导致协作壁垒难以打破。同时，兼具深厚语言学理论基础与先进人工智能技术能力的交叉学科领军人才，以及能够贯通两领域的复合型骨干人才显著匮乏，制约了学科的可持续发展动能。

其次，核心技术的原始创新能力仍有待实质性提升。特别是在多模态语言理解（需协同处理文本、语音、图像等多源异构信息）、小样本高效学习（在有限标注数据下实现模型快速泛化）等前沿关键技术方向，与国际顶尖研究机构及团队的技术水平相比，仍存显著差距。基础理论创新与核心算法突破的原创性成果相对稀缺，高水平研究产出不足。

最后，应用场景的实际落地面临多重现实壁垒。突出表现在教育、文化创意、心理咨询等关键领域尚未形成统一的行业技术标准，数据规范、评测体系与应用接口呈现碎片化现象。尤为值得关注的是，大语言模型的迅猛发展，虽为语言智能学科注入了强劲动力并开辟了前所未有的机遇空间，但也同步引发了因海量语料资源无序采集、质量控制机制缺失所衍生的系列风险问题。这包括但不限于：智能生成内容可能被滥用于大规模制造难以辨识的虚假信息；基于深度伪造技术合成的极具欺骗性的视听文本内容加剧了欺诈风险；以及缺乏有效约束的模型输出可能传播社会偏见、泄露用户隐私等。这些问题不仅对学科应用的伦理合规性提出了严峻挑战，也为构建安全可信的语言智能生态构成了实质性障碍。

（4）不足与未来发展

语言智能学科当前面临的主要不足体现在以下两方面：一方面，高端人才引育力度亟待提升，尤其是具备国际影响力的国家级领军人才数量与国内同类顶尖

高校相比仍存在显著差距；另一方面，成果转化机制尚未完善，特别是在校企协同研发过程中，核心技术相关的权益分配机制与知识产权确权规则仍需进一步明晰，这在相当程度上制约了创新成果的有效转化。

面向未来发展，本学科将重点推进以下四方面核心工作：

（1）学科交叉融合突破：着力打破传统学科壁垒，深化与认知科学、脑科学、计算科学等关键领域的协同创新，系统探索语言智能的前沿交叉研究方向。

（2）核心技术攻坚：集中突破关键技术瓶颈，重点聚焦国家亟需的大规模高质量语料库构建任务，并着力解决大语言模型普遍存在的"幻觉"问题，显著提升模型的可靠性及准确性。

（3）产业生态体系构建：积极整合产学研用多方资源，构建开放协同、可持续发展的语言智能技术研发、应用推广与产业孵化的全链条生态体系。

（4）高层次人才培养体系优化：系统完善人才选拔、培育与激励机制，着力构建结构合理、创新能力突出的高水平研究梯队，为学科可持续发展提供核心人才支撑。

通过系统规划并扎实推进上述举措，力争至 2030 年将本学科建设成为国内引领、国际知名的语言智能原始创新策源地与核心技术辐射源，切实承担服务国家文化数字化战略的核心驱动职能。

案例二：四川外国语大学

四川外国语大学对语言智能学科的建设经历了两个阶段：前“语言智能时期”的探索和“语言智能时期”的探索[8]。

（1）前“语言智能”时期的探索

四川外国语大学对语言智能的探索，可以追溯到十多年前对智能化语言测试以及语言心脑智能的涉足和研究。

2009 年，四川外国语大学申报并成功获批“央地共建”专项资金项目“多语种智能化测试实验室”，在外语界较早尝试开展外语计算机辅助命题、组卷、施考、阅卷、入库以及外语自适应考试等方面的研究工作。这可以作是学校对智慧语言教育的最早涉足。同年，学校启动了“外语学习认知神经实验室”建设工作，尝试采用脑电、眼动等认知神经科学手段研究语言的心脑机制。该实验室于 2010 年全面建成并投入使用，是全国外语界最早建成的三个同类实验室之一。在自此以后的 10 余年里，学校一直以该实验室为依托，致力于认知神经语言学方面的研究。

为促进语言学与脑科学的交叉融合，2016 年，学校以“外语学习认知神经实验室”为基础，正式成立“语言脑科学研究中心”，并新建语言神经调控、语言认知 CT、言语语言治疗等 5 个实验室/间，共建“中美失语症神经康复协作

研究中心”。2017 年，学校在外国语言学及应用语言学二级学科下正式设置“语言病理学”硕士研究生招生方向，着手开展语言病理学研究生人才培养以及语言认知障碍、语言神经教育等方面的研究工作。学校初步形成了以“认知神经语言学”为一体，以“语言病理学”和“语言神经教育”为两翼的科学研究与人才培养格局。

至此，学校依托外国语言文学优势学科，从“认识语言脑”拓展到“保护语言脑”和“利用语言脑”，初步走出了一条“语言+脑科学+康复医学/教育学”的发展之路。这一先期建设历程，事实上是在从语言学、脑科学、认知科学等跨学科角度对“语言自然智能”进行研究和探索，其研究的两个核心方面即正常人的“语言心脑机制”和特殊人群的“语言心脑机制”，为开展“模仿语言脑”研究尤其是开展语言类脑智能研究提供了独到的基础和条件，成了一种优势前提和准备。

（2）语言智能时期的探索

2017 年，国务院印发《新一代人工智能发展规划》，提出了人工智能的顶层设计，明确了面向 2030 年新一代人工智能发展的指导思想、战略目标、重点任务和保障措施，倡导大力构筑我国人工智能发展的先发优势，加快建设创新型国家和世界科技强国。2018 年 4 月，教育部出台了《高等学校人工智能创新行动计划》，指出“鼓励有条件的高校建立人工智能学院、人工智能研究院或人工智能交叉研究中心”。2019 年 3 月，人工智能第三次被写入政府工作报告，并首次提出“智能+”全新理念，强调深化人工智能研发应用。重庆市也响应这一政策，大力实施以大数据智能化为引领的创新驱动发展战略行动计划。2019 年，重庆市政府出台《重庆市高等学校人工智能+学科建设行动计划》（渝教研发〔2019〕3 号），启动“人工智能+学科群”建设专项，升级提优传统优势学科，探索符合重庆实际的人工智能学科群建设路径。

在此背景下，四川外国语大学于 2019 年 4 月在全国率先成立“语言智能学院”，学院定位为：“新文科建设的试验田、未来交叉学科专业的孵化器、未来 15~30 年外国语大学办学新路的探寻者”。同年，学校将“语言智能”自主设置为外国语言文学一级学科下二级学科并报教育部备案，与此同时，“语言智能”获批重庆市首批“人工智能+学科群”20 个立项建设项目。为聚合优化校内资源，2020 年，学校又进一步将全校“计算机教研室”划归语言智能学院，并赋予语言智能学院“语言智能”研究生人才培养和全校本-硕-博人工智能通识素养教育的职责。

2021 年 9 月，首批 10 名“语言智能”硕士研究生报到入学，学院语言智能研究生培养工作正式开启。迄今，学院已连续招收硕士研究生 4 届（54 人），培养毕业生两届（22 人），2025 届新招收研究生 20 人将于 9 月正式入学。在此期

间，学院也积极拓展办学层次。2022 年 9 月，学院将与学校相关院系联合开设本科交叉学科新专业“英语+智慧语言康复”；与此同时，也独立开设本科微专业“语言智能”。在语言智能学科建设与发展过程中，学院重点致力于交叉师资队伍培育、研究生人才培养以及学科方向发展工作。

在师资队伍建设方面，学院立足实际，“内培”与“外引”两手抓。在“内培”的方面，学院创新交叉学科师资培养模式，通过举办“读书自学”“视频研修”“课堂听课”“语言智能大讲堂”“语言智能新文科暑期学校”等定期、非定期活动，采用专家授课、学习分享等方式，就 AI 数学基础、AI 模型与算法、机器学习、自然语言处理、量子计算、类脑智能等内容进行集中研修学习，营造多学科交叉和交流的学术氛围，帮助学院教师省时高效地走进语言智能，培育师资的人工智能素养和交叉学科素养，努力构建“专业能力强、智能素养高、交叉思维活”的创新型师资团队。在“外引”的方面，广泛吸纳计算机科学、医学、生命科学、生物医学工程、心理学、外国语言学及应用语言学等领域的青年博士人才，为“语言学+脑科学+生命科学+人工智能+X”多学科创新发展提供“基本盘”支持，蓄势蓄能未来。同时，通过外聘兼职、柔性引进等多种形式，聘请国内外语言智能交叉学科领域专家为学院特聘教授或“巴渝学者”讲座教授等，为学院师资发展提质增优，加速语言智能学科建设。

在研究生人才培养方面，学院以课程体系建设为抓手，致力于培养兼具“语言学”基础素养、“语言自然智能”标示性素养以及“语言人工智能”标签性能力的研究生特色复合型人才。所开设的“语言学”基础素养课程包括智能语言基础、语义学、语用学、认知语言学、心理语言学、语料库语言学等；所开设的“语言自然智能”标示性素养课程包括神经解剖学、脑科学与类脑智能、语言病理学、语言认知障碍、言语治疗学、语言脑机接口等；所开设的“语言人工智能”标签性能力课程包括人工智能概论、人工智能数学基础、自然语言处理、高级程序语言设计、智能语言统计与数据挖掘、智能语言教育等。

在学科发展方面，学院拟定了“先硕-后博-再本”的人才培养梯次渐进建设方案，提出了“学术化、智能化、产业化、超学科化”的发展思路，力求稳步前行。经过三年的建设，语言智能学科初步形成了四大研究方向，即语言认知智能、智慧语言康复、智慧语言教育和智慧语言工程，形成了“认识语言脑、保护语言脑、开发语言脑、模拟语言脑”四位一体的学科发展基本格局。2022 年，学校“语言智能”在重庆市“人工智能+学科群”首轮建设终期验收中获得“优秀”等级，并自动列入下一轮（2023—2025）建设资助计划。2023 年 11 月，学院联合十余家企事业单位发起成立了全国首个语言脑机接口研究院——重庆市沙坪坝区国际语言脑机接口联合研究院，致力于科研成果产业转化研究。学院“语言+人工智能”的新文科建设特色开始显现。

2.5.3 问题与挑战

尽管学校在语言智能学科发展与建设方面做出了上述有益尝试与探索，也取得了一定成绩，但还面临不少问题与挑战。

第一，语言智能属多学科交叉性质学科，无现成人才储备，需先引进近邻单学科背景人才，再实施多学科交叉培养，遴选、外引有此交叉志趣的人才不易，跨学科“内培”难度更大，外引、内培两难叠加，拉长了人才成长与学科发展的周期。

第二，语言智能是一个“新生儿”学科，学科构架、内涵、领域、分支、范围、学理等都有待探索形成，其发展与建设无现成“图纸”可循，只能是“铁匠没样边做边像”，摸索着前行，难于“大步流星”。

第三，语言智能是名副其实地建立在多学科交叉基础之上的“新文科”，既“烧脑”又“烧钱”，需要建设攸关方政策、资金、管理等方面“久久为功”的支持，也需要建设主体“天才般”的定位、思路、办法与行动，也需要立志“事竟成”的有志者。

第四，语言智能学科领域极为广阔，如何既立足可行又着眼未来，既夯实基础又选准“长板”与“强项”，在较短时间内培育“错位”特色与高峰，“心安理得”的有所为有所不为，是一个需要迎难面对、挑战眼光的现实问题。

2.5.4 不足与未来发展

尽管语言智能作为一个多学科交叉研究领域已经获得了广泛重视，但它作为一个新兴的学科，尚处“百事待举”阶段，其学科构架、内涵、领域、分支、范围、学理、人才培养等都有待探索形成，亟需加大建设力度。刘利就指出，当前语言智能领域对语言知识和语言资源的投入还很不足，语言智能人才的培养还存在结构性缺陷[26]。一方面，了解语言智能算法和应用技术的人对语言资源缺乏认识，对语言规律、语言现象了解不足；另一方面，从事语言规律研究、参与语言资源建设的人又缺少语言智能科学的素养。这种“语言”和“智能”的严重脱节，事实上已经成了影响语言智能发展的瓶颈。姜孟等也认为，语言智能内蕴心智哲学、语言学、脑科学、生命科学、计算机科学、人工智能等多学科特征与属性，最能集中体现文理、文工、文医、文文等多元交叉与融合，但这一学科尚处在发展的初始阶段，百事待举，无论是学科体系构建、人才培养模式搭建还是师资培育，都面临不小的问题与挑战[8]。为此，李佐文和梁国杰呼吁将“语言智能”建设为“智能科学与技术”下的二级学科[23]。他们还提出了三条语言智能学科建设的可能路径：

一是进一步明确语言智能在我国现行学科体系内的定位。具体可整合语言学学科、计算机科学与技术、信息科学、脑科学与认知科学等领域的研究力量，在“智能科学与技术”一级学科下设立“语言智能”二级学科，组建文、理结合的

师资队伍和研究团队，共同推进语言智能学科建设与人才培养。

二是进一步凝练研究方向，构建语言智能学科的核心课程体系。可从语言智能学科的理论、技术和应用三个层面，凝练出以下主要研究方向：语言与脑科学研究、语言认知机制研究、语言计算理论研究、语言资源与大数据技术、多语言自然语言处理技术、语言智能教学、人机互动系统、语音和语言工程等。在课程体系构建方面，可设置语言学基础理论、计算语言学、认知科学与人工智能导论、语言与脑科学、语料库语言学、自然语言处理技术、语言智能教学、编程与算法基础、口笔语机器翻译、语言大数据与深度学习等专业类课程。

三是加强师资队伍建设，创新校企合作模式。在语言智能师资队伍建设方面，可以对接学校新文科发展规划，以问题为导向，打破学科壁垒，发挥多学科、多语言人才优势，打造语言学科、计算机学科、神经认知学科等方面知识结构互补的师资力量和科研攻关团队；在创新校企合作模式方面，可以建立校企联合实验室等方式加强与人工智能、语言服务等领域相关企业的务实合作，以增强语言智能学科教师和学生的企业实践经历，促进人才培养。

总之，在人工智能、新文科的时代背景下，语言智能学科建设意义重大，前景无限，势在必行。期待各界同道同仁、有识之士，顺应国家人工智能战略需求，加强协作，携手攻关，以新文科建设为契机，大力推进全国语言智能研究的水平与学科发展，合力构建语言类学科与智能类学科互通互撑、交叉融合的新格局，积极为外国语言文学学科未来新专业的创生探路用力。

2.6 语言智能与外语学科智能化转型

语言智能学科建立的一个重要目标是要为语言智能教育服务。周建设和薛嗣媛明确指出，语言智能学科要致力于培养系统掌握人文社科专业与信息处理等相关的智能科学基础知识，具有面向计算机信息处理和面向人文社科教学开展科学研究的能力，具备较高人文科学素养、数据科学素养、计算思维能力的跨学科交叉复合型人才[22]。在这方面，语言智能在赋能语言教育方面具有独到的优势。它将语言智能理论技术融合为应用，聚焦核心能力培养，构建语言智能教学平台，为学习者提供个性化学习支持，形成高素质人才培养的良性互动的教育生态；它通过优化教育模式，可以从本质上变革以特定知识和技能传授为目的的传统教学范式，更加追求批判、创造审美、沟通、合作等核心能力的培养。

李宇明也高度肯定语言智能对于教育（包括外语教育）的独特作用，认为“语言智能的作用是革命性的”[10]。他认为，“语言智能”是“人工语言智能”的简称，作为人工智能皇冠上的明珠，从人工智能中分离出语言智能的概念具有重要的理论意义和技术价值，其作用体现在社会生活的两个方面：一是有助于提升教育品位、均衡教育资源，二是有助于通过数据与计算来解决“人文计算、情感计算”等诸多人文问题。

由北京语言大学语言资源高精尖创新中心和中国中文信息学会发表的“语言智能与社会发展之论坛 2018 宣言”深入探讨了语言智能与外语教育教学的关系问题。宣言指出，语言和语言学习是人之所以为人、提升智力和大脑功能的关键。语言智能也不是单纯的机器智能，而是以人类语言能力为基础、实现增强与互联的深度人机结合的综合智能行为。作为工具的语言智能技术无法替代外语从业者，但对不掌握智能工具的外语从业者会带来冲击；外语教育（包括第二语言教育）应与时俱进，将语言智能的挑战作为转型发展的契机。坚持语言教育的工具性、人文性的双重属性，重视语言运用能力、跨文化交际能力和文化包容心的教育，重视语言智能技术的学习与应用，培养不同文化间的穿行者。该宣言还倡议：外语教育面对语言智能时代的冲击和挑战，为自身发展和学生前途计，应当全力适应人机共存的语言生活形态，充分利用语言工具的革命性变化，革新教学范式和人才培养路径，增强受教育者智能工具的使用能力，帮助其过好智能时代的外语生活。

胡加圣、戚亚娟认为，以 ChatGPT 为代表的语言智能产品已然拉开人机交互新时代的序幕，AI 语言模型为超海量、自由式、个性化、任意性的即时人机对话趟平了道路，严重冲击中国高等外语教育的格局，深刻影响未来的跨文化、跨语际交流沟通方式、知识交融方式以及文化融合速度等[27]。为此，他们建议：在外语政策上，可以藉此机遇调整高校外国语言文学的学科内涵、专业结构以及人才培养目标等，从过去的语言本体研究全面转向应用研究，构建起以语言为基础，向所有涉外学科专业知识过渡的“大外语”学科框架体系，以便真正完成“会语言、通国家、精领域”的育人目标；同时也使外语学科的目标定位更开放，更精准，更加融入全球教育体系和知识体系，并让外语能力成为每个现代受教育者必备的人文素养之一。

黄立波认为，语言智能研究将对外语教育产生冲击和影响，未来外语教育教学必须面对并迎接语言智能发展带来的剧变，对接语言智能研究的最新研究成果[4]。具体说，语言智能将改变传统的教材编写模式、改变教学平台助力课堂教学的方式，同时也将变革外语教育教学评价的方式。他提出，在大数据背景下，需要充分利用现代技术，建立基于语言智能研究的智能化教育教学体系，如图 2-3 所示。

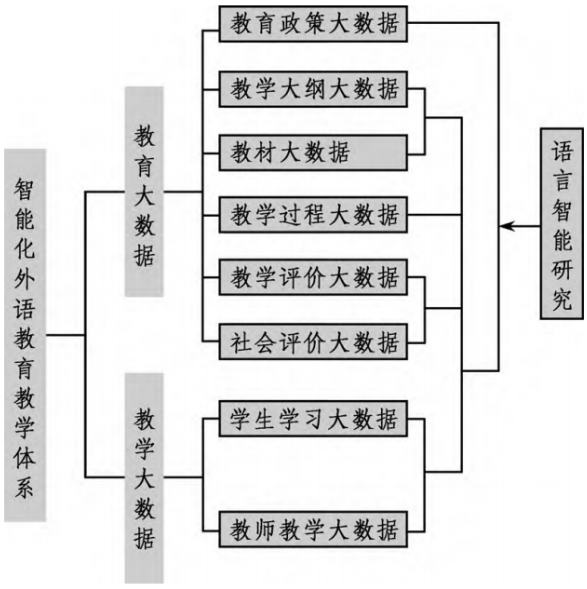

图 2-3 基于语言智能研究的智能化教育教学体系[4]

胡开宝、田绪军则专门撰文讨论了语言智能带给 MTI 人才培养的挑战。他们认为，未来 MTI 人才培养需要对接语言智能领域的新发展，从培养方案、课程设置、教学模式和师资队伍建设等多方面进行改革，将语言智能基础知识和相关能力融入人才培养的全过程[2]。

总之，语言智能的发展给外语教育带来了从思想、理念、内容、模式到原则、方法、路径的巨大变革，外语教育的智能化转型发展正在迎来一个新契机。

2.7 语言智能研究的新动向

目前，语言智能研究出现了语言人文基因计算、语言产业、语言智能治理等研究新动向。

2.7.1 语言人文基因计算

周建设提出，人工智能时代，落实国家通用语言文字政策，在推进语音标准化、词汇语法规范化的基础上，应当高度重视语言文字人文基因资源库建设，研发人文基因智能计算技术[25]。他认为，人作为社会生命体，既有生命基因，也有人文基因。人文基因是人的精神文化意识要素的统称。对人而言，生命基因是天性，人文基因是习性。天性，是自然性，基于物质，因而生命基因检测的对象是生命基因物质载体核苷酸；习性，是社会性，基于习得，因而人文基因检测的对象是人文基因习得载体语言文字。生命基因自身的变化导致其外部条件随之变化，外部条件的变化反过来又促使生命基因自身变化；调节生命基因，可以改变生命质量甚至生命形态。同理，人文基因自身的变化导致其外部条件随之变化，外部条件变化反过来也促使人文基因自身变化；调节人文基因，可以改变精神质量，甚至意识形态。周先生进一步提出，通过语言文字载体开展人文基因计算的一个前提是，建设大规模语言文字人文基因资源库，分类标注人文基因“成色”度。

只有基于人文基因资源库，才可能借助智能计算技术，检测特定对象的人文基因情况。

2.7.2 语言产业

李宇明认为，“语言产业”是指以生产和提供语言产品为主的行业。语言产品的形态包括语言、文字及相关符号，语言知识产品，语言文字艺术产品，语言技术产品，语言医疗康复产品，语言咨询培训服务，语言人才以及语言数据[28]。语言产品的形态、语言产业的业态决定着语言产业的基本面貌，是语言产业研究的基础范畴。随着语言智能的发展，语言逐渐为人和机器两个“物种”所拥有。随着信息空间、语言智能和物联网的发展，语言将承担起“万物关联对话”的任务，在社会、信息、物理三元空间中发挥互动作用，由此产生了大量语言数据。实际上，在人类观察世界所形成的数据中可供计算机处理的数据，其中 80%都是语言数据。不仅如此，语言与其他生产要素，如劳动、资本、知识、技术、管理等也关系密切。在数据与智能时代，数据已经成为生产要素，数据的价值正在由科学领域进入社会经济制度领域。语言数据包括四大类: 语言的符号系统，语言负载的信息，由语言延伸的各种符号与代码，生活、艺术与科学技术符号。无论是从量上还是从质上看，语言数据都是最为重要的数据，因而也是重要的生产要素。包括语言数据产业在内的语言产业，将成为数字经济的一方支柱。

2.7.3 语言智能治理

饶高琦、周立炜认为，语言智能在深刻改变语言治理的同时，自身也成为了语言治理的新对象。语言智能治理包括三个方面的内容：数据治理、评价治理与安全治理[29]。从数据治理来看，语言数据资源决定着语言智能在性能、技术落地、关键领域应用和伦理安全等重要方面的表现，但目前我国很多行业数据化程度较低、数据质量不高，关键领域数据国产化程度有限，且伦理安全的地位亟待提高。张凯、薛嗣媛、周建设也指出，当前语言数据治理体系面临语言数据的偏见、经典语言治理模型的短板等技术困境，可以采用点状聚合、线性组合和多层事态 3 种语言数据治理模式，来破解困境、弥补经典数据挖掘模式的短板[32]。

再从语言智能的评价治理来看，学术机构和政府机构组织的技术评测已经成为推动智能技术和产品进步的重要手段，已形成“数据–技术–评测”闭环，但我国的技术评测在权威性、影响力、规划和设计科学程度上尚有提升空间。语言智能的安全治理，涉及两个方面。首先是能力问题，即能力落后于需求；其次是伦理安全问题，集中于语言资源、应用和社会文化方面。为此，语言智能治理首先要以数据为基础，以评测为牵引，提升语言智能能力；其次，要从数据、评测和使用等方面把好安全关，从理念、组织、实施、学科和社会文化等方面把好伦理关。

此外，饶高琦、胡星雨、易子琳还提出了语言大模型治理问题。他们认为，

语言大模型与其他发明的不同之处，是人类第一次无法完全理解其具体运行机制[30]。语言大模型的优异性能很大程度上依靠大数据中的涌现效应。这一现象如同黑盒遮蔽了其内部工作路径，造成了其结果的不可解释和一定程度的不可控制。因此，对语言大模型的治理，重点是对大模型赖以研发的语言资源和投放之后的使用的治理。就现实情况而言，对大模型研发中的语言资源治理，应着力打破中文数据孤岛：发展以联邦学习为代表的分布式模型构建技术，建立国家知识数据开放机制，尽快健全开放、高效的语言数据交换市场；提倡世界知识中文表达，助推中文大模型研发，尤其应尽快实现中文精华知识资源面向网络开放，完善中文概念、术语资源，做大、做全领域中文资源。对大模型使用领域的治理，则因大模型本身也是一种重要的语言资源，应重点强调其基础资源地位，从标准化、评测和伦理规制的角度进行。

参考文献

[1]. 周建设, 2023. 卷首语.语言智能研究[M].天津大学出版社 1(1): 1-3.

[2]. 胡开宝，田绪军. 语言智能背景下的 MTI 人才培养：挑战、对策与前景[J]. 外语界，197(2): 59-64.

[3]. 胡开宝，尚文博. 语言学与语言智能[J]. 华东师范大学学报（哲学社会科学版）54(2): 103-109.

[4]. 黄立波. 大数据时代背景下的语言智能与外语教育[J]. 中国外语，19(1): 4-9.

[5]. 梁晓波,邓祯.美军语言智能处理技术的发展策略与启示[J].国防科技,2021,42(04):85-91.

[6]. 李佐文，梁国杰.语言智能学科的内涵与建设路径[J]. 外语电化教学, (5)：88-93.

[7]. 周建设，吕学强，史金生，等. 语言智能研究渐成热点[N]. 中国社会科学报，2017-02-07（3）

[8]. 姜孟、王霞、潘雪瑶. 语言智能新文科建设与发展探索[J]. 语言智能研究, 1(1): 7-13.

[9]. 姜孟. 语言智能:语言人工智能研究的历史新方位———“语言智能科学”理论与方法论构建(三) [J]. 外国语文,40 (4): 60-80.

[10]. 李宇明. 语言智能与社会进步——序《语言智能研究》[J].语言智能研究, 1(1): 1-3.

[11]. 冯志伟. 我国计算语言学研究 70 年[J]. 语言教育 (4): 19-29.

[12]. 毕然, 孙高峰, 周湘阳, 刘威威. 深度学习：零基础实践[M]. 北京: 清华大学出版社.

[13]. 李沐, 刘树杰, 张冬冬, 周明. 2019. 机器翻译[M]. 北京：高等教育出版社.

[14]. 冯志伟. 2011. 计算语言学的历史回顾与现状分析[J].外国语 (1): 9-17.

[15]. 徐英瑾. 2022. 人工智能如何“说人话”？ ——对于自然语言处理研究的哲学反思[J]. 自然辩证法通讯, (4): 1.

[16]. 宗成庆. 2019. 统计自然语言处理[M]. 北京：清华大学出版社.

[17]. 钟义信. 2018. 机制主义人工智能理论——一种通用的人工智能理论[J]. 智能系统学报 13(1): 2-18.

[18]. Wintner, S. 2009. What science underlies natural language engineering? [J]Computational Linguistics 35(4): 641-644.

[19]. 莫宏伟 徐立芳 2020. 《人工智能概论》. 北京：人民邮电出版社.

[20]. 冯志伟. 1996. 自然语言的计算机处理[M]. 上海: 上海外语教育出版社.

[21]. 黄河燕, 史树敏, 贾珈, 黄民烈, 韩先培, 刘洋, 刘奕群. 2020. 人工智能：语言智能处理[M]. 北京: 电子工业出版社.

[22].周建设,薛嗣媛.论语言智能教育[J].语言战略研究,2023,8(04):30-43.
[23].李佐文,梁国杰.语言智能学科的内涵与建设路径[J].外语电化教学,2022,(05):88-93+117.
[24].周建设.语言智能，在未来教育中扮演什么角色[N]. 光明日报,2019-03-02(12版).
[25].周建设.人文基因智能计算将成为语言文字资源建设的新途径[J]. 语言战略研究. 2022 ,41 (05): 10.
[26].刘利.语言智能的学科建设与发展方向[EB/OL]. Retrieved from https://www.sohu.com/a/321162236_312708,2019-06-17.
[27].胡加圣、戚亚娟. ChatGPT 时代的中国外语教育：求变与应变[J]. 外语电话教学. 209(1): 3-6.
[28].李宇明. 数据时代与语言产业[J]. 山东师范大学学报（社会科学版）, 65(5): 87-98.
[29].饶高琦 周立炜. 论语言智能的治理[J]. 语言战略研究, 51(3): 38-48.
[30].饶高琦、胡星雨、易子琳. 语言资源视角下的大规模语言模型治理[J]. 语言战略研究, 46(4): 19-29.
[31].耿立波, 刘涛, 俞士汶, 孙茂松, 杨亦鸣. 当代机器语言能力的研究现状与展望[J]. 语言科学 1(13): 34-41.
[32].张凯、薛嗣媛、周建设. 语言智能技术发展与语言数据治理技术模式构建[J]. 语言战略研究. 40(4): 35-48.
[33].张雄伟,孙蒙,杨吉斌.智能语音处理[M].机械工业出版社:202010:516.

第三章　语言智能技术

语言智能技术（Language Intelligence Technology, LIT）是人工智能的一个分支，旨在通过计算模型模拟和延伸人类语言能力，核心目标是让机器在“理解—生成—交互”三个维度上达到甚至超越人类水平。在技术谱系上，LIT 以自然语言处理（Natural Language Processing, NLP）为底座，向上游衔接语言学对音素、词汇、句法、语义和语用等层次的形式化刻画，向下游对接认知科学关于语言习得、记忆、推理与情感加工的实验发现，并横向融合机器学习、知识工程、计算机视觉、语音技术等多学科方法，从而构建起“文本感知→语言解析→语义推理→内容生成→多轮交互”的完整能力链路。从范式演进的角度看，LIT 已经历三次关键跃迁：早期基于规则的符号系统依赖人工编撰语法与专家词典，难以应对开放域场景的复杂性；随后统计学习方法以 N-gram、HMM、CRF 等模型为代表，通过大规模语料统计实现有限泛化；当前，以 Transformer 为代表的大模型范式借助千亿级参数与自监督预训练，彻底转向“数据驱动”路径。《生成式人工智能应用安全测试标准》指出，现代语言智能系统不再依赖人工规则，而是在海量文本数据中学习统计规律和隐含知识，以“预训练+微调/提示”的统一框架支撑机器翻译、情感分析、对话生成、代码合成、知识推理等数十种下游任务，并表现出小样本适配、跨任务泛化、持续学习更新等新特性。然而，大模型范式的通用能力也带来了可解释性不足、事实幻觉、伦理风险等新挑战。薛嗣媛、李舟军在《大模型概述》中强调，未来语言智能的发展需在“能力—安全—治理”三角张力下，持续探索数据治理、对齐算法、红队测试与监管框架的协同机制，以实现从“感知智能”到“认知可信智能”的再平衡。

3.1　基础支撑技术

3.1.1　自然语言理解

在自然语言处理的底层任务中，分词、词性标注和命名实体识别是中文处理最为核心的环节，它们共同构成了语言理解的基础模块。

分词是中文处理的前提环节，它将连续的字流切分为词语单元。中文没有显式的词边界，句子中词语是以连续汉字串的形式出现的，这与英文等以空格为天然分隔符的语言有显著差异。分词方法经历了从基于规则的最大匹配算法、统计学方法（如 HMM、CRF）[1]，基于深度学习的端到端分词模型进一步提升了分词准确率[2]。

词性标注则为词语赋予语法角色标签，为后续的句法与语义分析奠定基础。词性标注是为句子中的每个词语赋予其语法类别，如名词、动词、形容词等。它不仅有助于句法分析和语义解析，还广泛应用于问答系统、信息抽取和自动摘要等任务中。词性标注传统上依赖于基于规则和统计建模的方法，如 HMM、最

大熵模型、CRF 等，而深度学习模型的引入极大提升了标注的准确率[3]。近年来，Transformer 模型能够结合上下文长距离依赖关系，对歧义词进行更准确的词性判定。

命名实体识别（Named Entity Recognition，NER）旨在识别文本中的专有名词实体，如人名、地名、机构名等，是信息抽取和知识图谱构建的重要基础。中文命名实体识别尤为复杂：一方面，人名、地名的表达形式高度多样，缺乏统一的标记规则；另一方面，歧义性强，往往需要结合上下文消歧。这些因素使得中文命名实体识别在大规模实际应用中依然面临显著难题。为解决这些问题，研究者提出了融合词典和上下文的模型，例如将字符级和潜在词级信息结合的模型以及利用卷积神经网络和注意力机制融合多层次词向量的方法[3]。这类深度学习模型有效缓解了中文实体边界模糊和歧义问题。

句法分析的目标是揭示句子中词语之间的依赖和层次关系，从而构建出句子的结构框架。主要包括两类：成分句法分析和依存句法分析。成分句法分析是将句子划分为若干语法成分，如主语、谓语、宾语、定语、状语等，并用树形结构表示。这种方法适用于研究句子的层次结构和修饰关系。依存句法分析是直接识别词与词之间的依赖关系，用有向边表示。语义角色标注的目标是识别谓词（通常是动词）以及与之相关的论元（如施事、受事、工具、地点、时间等），并为每个论元标注其语义角色。此任务的重要性在于它能够把句子转化为事件结构的形式，为知识表示、逻辑推理和自然语言理解奠定基础。现代句法分析从基于统计的解析器发展到基于神经网络的解析器，2014 年提出的神经网络高效依存解析模型显著提高了解析准确度[4]。随着深度学习发展，2017 年引入深度 Bi-LSTM 的解析器进一步提升了长句依存关系的捕捉能力现代语义角色标注系统在大模型的支持下，已经能跨领域、跨语言迁移，展现出广泛的适用性[5]。

3.1.2 语音信号处理

语音识别任务是将语音信号转化为对应的文字序列。其基本流程从语言智能技术的角度来看，语音识别（Automatic Speech Recognition,ASR）不仅是将语音信号转写为文字序列的过程，其核心在于把连续、易受噪声干扰的声学信号映射为离散、可计算的语言单位，并为后续的语义理解和知识建模提供结构化输入，通常包括特征提取、声学建模、语言建模和解码四个环节。传统方法依赖于隐马尔可夫模型（HMM）与高斯混合模型（GMM）的组合，通过统计建模实现语音到文本的映射。但是，此类方法在复杂噪声环境、口音差异及长语音识别等场景中存在明显局限。Hinton 等人率先将深度神经网络应用于声学模型训练[6]。随着深度学习的发展，卷积神经网络（CNN）、循环神经网络（RNN）、自注意力机制（Transformer）及端到端的序列建模框架，如 CTC[7]、Attention[8]、Transducer[9]）被广泛应用，以及融合声学和语言建模的 RNN-Transducer 模型[10]，大幅提升了

识别的准确率、实时性和鲁棒性。

语音合成任务是将输入的文字转化为自然、流畅且富有表现力的语音输出。传统方法多基于拼接式或参数统计建模，如 HMM-TTS[11]，在自然度和可控性方面存在不足。近年来，深度学习推动了神经网络语音合成的发展，如基于自回归结构的 Tacotron 系列，以及基于生成对抗网络（GAN）或变分自编码器（VAE）的模型，显著提升了合成语音的自然度与人类相似性。同时，基于非自回归和端到端的快速合成框架进一步提升了推理效率。最新趋势是将 TTS 与语音风格迁移、情感建模和个性化语音克隆相结合，使得合成系统能够生成具有多样化情感色彩和个性特征的语音[10]，为教育、客服、虚拟人等应用提供了更强支撑。

3.1.3 图像视觉处理

图像视觉处理是人工智能在感知层的重要组成部分，核心目标是让计算机具备“看懂”与“理解”图像和视频的能力。计算机视觉（Computer Vision, CV）最初依赖人工设计特征（如 HOG[13]、SIFT[14]），但随着深度学习的发展，卷积神经网络（CNN）[15]、视觉 Transformer（ViT）[21]、扩散模型（Diffusion Model）[22]等新一代方法迅速崛起，使得视觉识别与分析的准确性和泛化能力大幅提升。近年来，计算机视觉逐渐与语言模型结合，形成了跨模态理解研究的热点。图文跨模态理解旨在让模型同时处理视觉信息与文本信息，并在二者之间建立紧密的语义映射关系。相关任务包括图像描述生成、视觉问答、跨模态检索等。这类研究的突破性进展来自于大规模多模态预训练模型的优化，如 OpenAI 提出的 CLIP[16]、Google 提出的 ALIGN[17]，Salesforce 提出的 BLI[18]，以及微软研究院的 Florence 等[20]，它们通过对海量图文配对数据的对比学习或生成建模，获得了强大的视觉－语言对齐能力[16]。这种能力不仅使机器能够生成对图像的自然语言描述，还能根据自然语言指令进行图像检索与推理。

3.1.4 词向量技术

词向量技术（Word Embedding）是自然语言处理中一种将词语映射为低维稠密向量的核心方法，它通过捕捉词语在大规模语料中的共现关系来表示语义和语法特征，使语义相近的词在向量空间中距离更近。

在早期的典型方法中，Word2Vec 是最具代表性的技术之一。它通过神经网络模型在大规模语料中学习词的分布式表示，主要包括两种训练框架：CBOW（Continuous Bag of Words）通过上下文词预测中心词，强调整体语境对单词的约束作用；Skip-gram 则通过中心词预测周围的上下文，适合处理低频词的表示。这两种方法不仅能够捕捉词与词之间的语义接近性，还能在向量空间中展现出类比运算的能力[22]。

GloVe（Global Vectors for Word Representation）利用语料库中的全局词频共现信息，通过矩阵分解方法学习词向量。Glove 的设计目的在于，将局部语境中

的共现关系与整体语料的统计特征结合起来，使词向量不仅能够反映邻近词语的语义联系，还能在更大范围上捕捉语言的全局规律，从而在大规模语料条件下获得更加稳定和精确的语义表示[23]。

ELMo（Embeddings from Language Models）是上下文动态词向量的代表性方法，它首次突破了静态词向量“一词一义”的局限。ELMo 采用双向 LSTM 语言模型，对输入文本进行逐层建模，并在不同语境下动态生成词向量，从而使同一词语能够根据语境获得不同的语义表示[24]。ELMo 的提出标志着词向量从静态到动态的转变，大幅度提升了情感分析、命名实体识别、语义角色标注等下游任务的表现。

3.1.5 预训练语言模型

预训练语言模型（Pre-trained Language Models, PLMs）是近年来自然语言处理领域最具代表性的技术突破。它们通过在大规模无标注语料上进行自监督训练，学习到丰富的语言规律和语义表征，再在下游任务中进行微调，极大提升了模型的迁移能力和通用性。这一范式打破了传统依赖人工标注数据和任务专属模型的局限，使自然语言处理进入了“预训练+微调”的时代。

BERT（Bidirectional Encoder Representations from Transformers）是在 ELMo 思路的基础上，结合 Transformer 架构提出的更强大的预训练语言模型[25]。BERT 最大的创新在于采用双向自注意力机制，能够同时利用上下文中的前后信息来建模词语语义，解决了传统单向语言模型难以兼顾全局信息的问题。它通过两个预训练目标进行学习：Masked Language Model（MLM）让模型在缺失词语的情况下预测正确词汇，从而理解双向依赖；Next Sentence Prediction（NSP）则训练模型理解句子级的逻辑关系。凭借这种深度双向建模与大规模无监督预训练，BERT 在 11 个自然语言理解任务上刷新了当时的最佳成绩，成为 NLP 发展的里程碑。

GPT（Generative Pre-trained Transformer）由 OpenAI 提出，与 BERT 一样采用 Transformer 架构[26]，但其设计理念不同：GPT 选择自回归语言模型作为预训练目标，即通过不断预测下一个词来学习语言的生成模式。这使得 GPT 特别擅长自然语言生成，在对话、文本续写、故事创作等任务上展现出强大能力。GPT 系列（如 GPT-3、GPT-4）的规模不断扩大，参数达千亿级，展示了强大的涌现能力和零样本/少样本学习性能[27][28]。

3.1.6 大语言模型

BERT 和 GPT 的出现，标志着预训练语言模型在“理解”与“生成”两个方向分别取得了突破性进展。然而，随着模型规模不断扩大，数据语料日益丰富，以及算力资源的持续增强，预训练语言模型的能力呈现出质的飞跃。这一趋势推动模型逐渐走向大规模化与通用化，形成大规模预训练语言模型（Large Language Models, LLMs），即大语言模型。LLM 不仅在自然语言理解和生成任务上展现出

远超以往的性能，还表现出一定的跨任务迁移能力和涌现出的推理、规划、工具调用等能力。作为独立的第三方评测机构，SuperCLUE 使用原创评测题库，持续对国内外大语言模型性能进行实时跟踪。该机构选取了国内外有代表性的 42 个大语言模型的 12 月份版本进行评测，报告的主要结论包括：OpenAI-o1 大幅领跑，国内顶尖模型 DeepSeek V3 和 SenseChat 5.5 性能接近 OpenAI-4o，超过 Anthropic Claude 3.5 Sonnet 和谷歌 Gemini-2.0-Flash-Exp。DeepSeek V3 和 Qwen2.5-32B-Instruct 的推理能力领先，同时保持极快的推理速度。此外，Step-2-16k 在可执行 Agent，Qwen-max-latest 在指令遵循，Doubao-pro-32k 在代码生成，360zhinao2-o1 在逻辑推理，SenseChat 5.5 在数学计算等领域表现出色。进一步地，当预训练范式扩展至多模态领域时，出现了能够同时处理文本、图像乃至语音等模态的多模态大模型（LMMs），使人工智能逐步具备跨模态推理与交互的潜力，也奠定了语言智能研究的新基础。大语言模型的强大能力不仅依赖于规模和架构，还源于一系列新技术的应用与发展：

第一，上下文学习（In-Context Learning, ICL）是 LLM 的重要特性之一[29]。不同于传统模型需要通过参数更新来适应新任务，LLM 可以仅凭输入提示（prompt）中提供的少量示例，就能在推理阶段快速完成任务。这种能力使得模型具备“即学即用”的特点，能够在零样本（Zero-shot）或小样本（Few-shot）条件下解决问题，大幅降低了对标注数据的依赖

第二，提示词设计与优化（Prompt Engineering）是激发 LLM 能力的核心手段。通过合理的提示词构造，可以引导模型调用不同的知识与推理路径，从而提升结果的准确性和可控性。提示工程不仅涉及自然语言描述的设计，还包括任务分解、格式化约束（如表格、JSON）、角色设定等多样化方法。在应用层面，提示工程逐渐发展为一门系统方法论，被广泛用于信息检索、智能问答、文本生成等任务。

第三，思维链（Chain-of-Thought, CoT）则是在提示工程的基础上发展出的推理增强方法。通过在提示中显式要求模型逐步输出中间推理过程，而非直接给出答案，可以显著提升模型在复杂任务中的逻辑推理与问题求解能力[30]。例如，在数学计算、逻辑推断、推理问答中，CoT 能够让模型“边思考边作答”，不仅提高了准确性，还增强了可解释性。在此基础上，研究者又提出了自一致性（Self-Consistency）[31]、树状思维（Tree-of-Thought, ToT）[32]、推理范式—连续思维链（Chain of Continuous Thought，Coconut）[43]等扩展方法，进一步增强了模型在多路径推理和复杂问题求解中的稳健性。相比传统思维链只能沿单一路径推理，连续思维可同时保留多种备选路径，使模型能够通过广度优先搜索探索解决方案。在需要频繁回溯和规划的复杂推理任务中，Coconut 展现出优于传统思维链的性能。上述研究不仅为推理模型提供了新的发展方向，也促使我们重新审视思维与

语言的关系。思维可能并不依赖于语言形式，挑战了语言在推理过程中的垄断地位，并揭示了思维可能通过更抽象的方式被表达和传递。

第四，后训练机制。2024 年，大语言模型的训练范式已从“预训练+微调”的形式，普遍转变为“预训练+后训练”的形式。预训练继续采用自监督学习，后训练阶段则由监督微调与强化学习两个关键环节组成，后训练可以根据训练目标重复多轮。例如，《基于大语言模型的强化学习扩展》采用了一轮监督微调加强化学习的后训练流程，而 DeepSeek R1 则采用了两轮后训练流程[33]。这种由自监督预训练、监督微调与强化学习组成的有序训练范式，持续推动了大语言模型从获取语言到知识、再到推理能力的演化。

第五，类人推理能力。推理研究致力于引导模型从基于直觉的快速反应，转变到深度理性思考。以 OpenAI o1/o3 和 DeepSeek R1 为代表的先进推理模型，高度模拟了人类的慢思考模式，在数学、物理等科学证明及代码生成等复杂任务中，展现出超越人类专家的推理能力。这些模型在生成最终答案前，通过深度思考构建出大段的中间推理过程，形成〈问题—中间推理—回答〉的显式三段式结构。这些模型自发涌现出了“顿悟时刻”和“自我纠错”的行为，展现出类人的反思能力。这些大语言模型推理轨迹的表达方式也与传统教科书式的刻板解法有所不同，呈现出鲜明的口语化特征。例如，在回溯推理历史时，模型会使用“或者，我们试试……”或“等等，但……”等，这种表述方式更接近于人类思考时的“内心独白”或“意识流”。这种动态的推理过程不仅展现了模型的思考深度，也反映出其推理过程的创造性和灵活性。研究者系统地探讨了推理模型的发展历程及构建方法，对结构搜索、奖励模型、宏动作、自我提升、强化微调等核心算法进行了梳理[34]。另外有研究者发现，对于简单的逻辑推理问题，推理后投票可以提升性能；而对于困难的逻辑推理问题，投票有时甚至会产生负面影响[35]。因此，提升模型推理能力的关键仍在于训练阶段。

第六，强化学习技术。自 ChatGPT 在大语言模型训练中率先采用基于人类反馈的强化学习（Reinforcement Learning from Human Feedbacks，RLHF）后，强化学习逐渐成为推动大语言模型能力发展的核心技术动力。基于经典的智能体—环境交互范式，大语言模型在强化学习中的功能系统地划分为信息处理器、奖励设计者、决策者和生成者共四种角色，分别进行了剖析。与传统的结果奖励模型相比，过程奖励模型通过为任务的中间步骤提供细粒度奖励反馈，在推理模型的训练中展现出显著优势。清华大学唐杰团队发表 ReST-MCTS 提出一种强化自训练策略，将过程奖励引导机制与蒙特卡洛树搜索相结合，通过估算在当前步骤之后获得正确答案的概率，以此推断过程奖励，从而提升推理轨迹与步骤奖励的质量[37]。

第七，合成数据技术。大语言模型以训练数据为学习对象，训练数据的数量、

质量及多样性对模型能力的增长具有显著影响。训练数据中逻辑推理数据的占比偏低，推理能力的发展严重滞后。针对上述问题，大量研究在训练中使用由大语言模型生成的合成数据。合成数据的生成流程分为提示工程与多步骤生成两种[38]。有效的提示通常包括任务规范、生成条件及上下文示例。多步骤生成指按样本或按数据集，将整体的生成过程人为地拆解为一系列更简单的子任务，强制模型按照预定规划逐步生成数据。管理合成数据的目的在于提高合成数据的质量，包括高质量样本过滤和标签增强两种途径。在 DeepSeek R1 的模型蒸馏实验中，通过将大语言模型的输入及其生成的输出对小模型进行监督微调，大语言模型的推理能力得以传递给小模型。实验结果表明，这种传递方式所带来的性能提升，甚至超过了人工数据的方式。这说明相较于人工数据，小模型更容易从大语言模型生成的合成数据中直接汲取相关能力。随着大语言模型能力的持续提升，高质量合成数据的自动生成已成为可能。这些合成数据亦将反哺大语言模型训练，构建出大语言模型自我迭代的正向循环。

混合专家模型技术。作为一种能够以极低计算开销，大幅扩展模型容量的有效方法，混合专家模型（Mixture of Experts，MoE）逐渐成为大语言模型架构的重要技术路径。混合专家模型将 Transformer 架构中的前馈网络层替换为 MoE 层，每个 MoE 层包含多个前馈网络作为专家，并使用路由函数动态激活与输入相关的那部分专家。该文从算法设计、系统设计及应用三个维度对混合专家模型进行了系统性梳理[39]。相比于单一模型，模块化神经网络具备显著优势。现有的模块化神经网络多采用预定义的模块功能划分。而最新研究表明，标准预训练 Transformer 内部自然存在隐式模块化结构，即“涌现模块化（Emergent Modularity）”。该文探讨如何释放语言模型中涌现模块化的潜能，证明标准语言模型无需增加额外参数即可部分转化为混合专家模型进行优化。这类从涌现模块化派生出的混合专家模型被称为“涌现混合专家（EMoE）”[40]。2024 年发布的大语言模型普遍支持长达 128K 以上的上下文[40]。上下文长度扩展技术旨在增强大语言模型对长序列输入的处理能力，同时避免计算需求的成比例增加[41][42]。

3.2 语言智能应用技术

3.2.1 机器翻译

机器翻译是实现不同语言之间的自动转换与生成，使文本能够跨越语言边界被理解和使用。随着全球化进程加快和跨文化交流需求的增加，机器翻译在信息传播、国际商务、教育科研、外交沟通等场景中发挥着日益重要的作用。传统的基于规则或统计的方法（如 SMT, Statistical Machine Translation）虽然在一定程度上实现了跨语言映射，但受制于规则库的复杂性与统计模型的稀疏性，翻译结果往往不够流畅，难以捕捉语境中的语义细节。随着深度学习的发展，神经机器翻译（Neural Machine Translation, NMT）逐渐取代了早期方法。基于 RNN、CNN

的翻译模型最先提出，但由于难以捕捉长距离依赖，其性能受限。Transformer 的提出为机器翻译带来了革命性突破，通过自注意力机制有效建模长距离语境关系，使翻译结果在准确性和流畅性上大幅提升。如今主流的翻译系统均采用基于 Transformer 架构，跨语言生成能力进一步增强。多语言预训练模型通过在海量双语或多语平行语料上学习，可以在不同语言之间共享语义空间，使得低资源语言也能借助高资源语言的迁移获得较好的翻译效果。

近年来，结合语音识别（ASR）与语音合成（TTS）的实时语音翻译系统也逐渐成熟，跨语言字幕生成、图文翻译、视频同传等新兴应用不断出现，拓宽了机器翻译的使用边界。然而，机器翻译仍面临挑战。其一是语义歧义与一词多义问题，模型难以在所有上下文中准确选择译法；其二是文化与语用差异，有些语言现象无法直接对等转换，需要引入常识与背景知识；其三是低资源语言的翻译，由于语料稀缺，翻译质量难以保证。为应对这些问题，当前的研究趋势主要包括：结合检索增强生成（RAG）利用外部知识库提升翻译质量，探索少样本或零样本跨语言迁移以支持低资源语言，发展多模态翻译模型以实现更自然的跨文化交流。

3.2.2 智能问答

智能问答（Question Answering, QA）是自然语言处理最具代表性的应用之一，其核心目标是让机器能够理解用户的自然语言输入，并以准确、自然、连贯的方式作出回应。从技术路径来看，QA 的发展大致分为三个阶段：（1）基于信息检索的 QA，依赖关键词匹配与排序技术，从文档库中提取最相关的片段返回给用户，代表了早期搜索引擎式问答；（2）基于阅读理解（Machine Reading Comprehension, MRC）的 QA，模型能够直接“阅读”给定文本，抽取或生成答案，代表性数据集如 SQuAD 推动了深度神经网络在抽取式和生成式 QA 上的快速发展；（3）基于大规模预训练语言模型的 QA，借助强大语义表征能力，QA 系统不仅能够处理开放域问题，还能进行跨文档推理和复杂知识整合，成为当前主流。

除了文本问答，视觉问答（Visual Question Answering, VQA）扩展了 QA 的边界，使模型能够理解图像内容并回答与之相关的问题。VQA 融合了计算机视觉与自然语言处理技术，被广泛应用于图像检索、辅助盲人阅读、监控与安防、多模态人机交互等场景。随着多模态大模型（LMMs）的兴起，文本 QA 与 VQA 的界限逐渐模糊，统一的问答系统能够同时处理文本、图像甚至语音输入，标志着 QA 技术正迈向跨模态智能。VQA 的发展主要体现在三个层面：一是在图像理解上，依托目标检测、图像分割和视觉表征学习，模型能够抽取多粒度的视觉信息；二是在语言理解上，通过预训练语言模型实现问题解析与语义建模，以适应描述性、关系性乃至常识性推理问题；三是在跨模态融合与推理上，采用联合嵌入与跨模态注意力机制，将视觉与语言特征对齐，并借助外部知识库增强模型

的推理能力。近年来，预训练—微调范式和多任务学习显著提升了 VQA 的泛化性能，指令微调和大规模多模态数据则进一步推动了生成式回答的发展。同时，随着统一多模态大模型的出现，跨模态链式推理、可解释性和安全性等议题逐渐成为前沿研究热点。

3.2.3 对话系统

对话系统（Chatbot）是自然语言处理的重要应用之一，其核心目标是通过自然语言交互实现人与机器的沟通。与智能问答相比，Chatbot 更强调多轮对话的连续性与交互性，不仅要回答用户的问题，还需要理解语境、保持语气自然，甚至具备情感回应的能力。在技术发展脉络上，对话系统大致经历了三个阶段：1）基于规则的方法：早期的对话系统多依赖手工设计规则与模板，如 ELIZ 和 ALICE，通过模式匹配和正则表达式生成固定回答。这类系统实现了人机对话的雏形，但缺乏语言理解能力，无法适应开放领域；2）统计学习方法：随着数据驱动的兴起，研究者引入概率语言模型（如 N-gram）、检索式方法以及条件随机场（CRF）、隐马尔可夫模型（HMM）等，用于对话状态建模和意图识别。这一阶段使对话系统具备了更高的灵活性，但语义理解和生成仍有限；3）神经网络与深度学习方法：Seq2Seq 模型的出现首次将对话生成建模为端到端的序列转换任务，LSTM、GRU 等结构提升了语义连贯性。Transformer 架构的引入进一步强化了对长距离依赖的建模能力，使系统能够在多轮交互中保持上下文一致性。近年来，大语言模型（LLM）为对话系统带来质的飞跃，通过大规模预训练与指令调优，模型不仅能进行上下文敏感的开放域对话，还展现出风格控制、知识调用、逻辑推理等能力，例如 ChatGPT 系列模型就代表了当前最高水平。

对话系统中的关键技术主要包括：意图识别与槽位填充是在任务型对话系统中，用于解析用户需求并提取关键信息，是自然语言理解（NLU）的核心；对话状态追踪（Dialogue State Tracking, DST）通过动态维护用户意图和上下文信息，使系统在多轮对话中保持逻辑一致性；对话管理（Dialogue Management, DM）：决定系统在当前语境下的应答策略，通常依赖强化学习或策略优化方法；自然语言生成（NLG）：将系统内部语义表示转化为自然、流畅的人类语言，是决定用户体验的关键环节；知识增强（Knowledge-Augmented Dialogue）：结合检索增强生成（RAG）或知识图谱，使对话系统能够提供事实准确、可溯源的答案；个性化与情感建模：通过用户画像与情感识别，让系统在不同场景下实现风格调节和人性化回应。

未来的技术趋势主要集中在三个方面：（1）检索增强生成（RAG）对话，通过外部知识支撑增强事实可靠性；（2）多模态对话系统，结合语音、图像、视频输入输出，提升交互的自然度和应用边界；（3）Agent 化对话系统，在 LLM 基础上赋予记忆、规划与工具调用能力，使 Chatbot 不再是被动应答者，而是具备

执行任务与自主推理的智能体。

3.2.4 文本分类

文本分类和主题建模是自然语言处理中的两类基础技术，广泛应用于内容审核、新闻主题聚类、舆情监控、自动标签推荐等场景。二者的目标不同：文本分类通常是一个监督学习问题，核心是将文本归入预定义的类别；而主题建模则更偏向无监督学习，旨在自动发现语料中隐含的主题结构。二者结合使用，可以实现对海量文本的高效组织与理解。

在文本分类方面，早期方法多基于词袋模型（Bag-of-Words, BoW）和 TF-IDF 特征，结合朴素贝叶斯（NB）、支持向量机（SVM）、逻辑回归等传统机器学习算法。虽然这类方法实现了对垃圾邮件过滤、情感分类等基础任务，但难以捕捉上下文和语义信息。随着深度学习兴起，卷积神经网络（CNN）、循环神经网络（RNN）、注意力机制被引入文本分类，大幅提升了特征表示能力。如今，主流方法是基于 预训练语言模型（PLMs）的分类，例如 BERT、RoBERTa、ELECTRA，它们通过上下文敏感的词向量捕捉语义特征，再在少量标注数据上微调即可完成新闻分类、产品评论分析等任务。这种范式不仅准确率高，还具有较强的跨领域迁移能力。

3.2.5 主题建模

主题建模则旨在自动发现文本集合中潜在的主题分布。经典方法是潜在狄利克雷分配（Latent Dirichlet Allocation, LDA），它通过假设文档由若干主题构成，每个主题由词分布生成，从而在无监督环境下推断语料的主题结构。随着神经网络的发展，研究者提出了神经主题模型（Neural Topic Models, NTM），如基于自编码器（VAE）的主题建模方法，它们能够在保持主题可解释性的同时提升表达能力。主题建模是自然语言处理中的一种无监督学习方法，其目标是从大规模文档集合中自动发现潜在的主题分布，并通过这些主题来组织、理解和探索文本数据。随着深度学习的发展，研究者提出了神经主题模型（Neural Topic Models, NTM），以克服传统方法的局限。NTM 通常借助变分自编码器（Variational Autoencoder, VAE） 或生成对抗网络（GAN），将主题建模转化为神经网络优化问题。这类方法不仅能够在大规模数据上更高效地训练，还能通过引入分布式词向量捕捉词语之间的语义关系，从而提升模型对多义词与同义词的处理能力。例如，神经变分主题模型（NVTM）能够在保证主题可解释性的同时，生成更具表达力的语义结构。

然而，主题建模仍存在一些挑战。首先，主题粒度难以控制，在实际应用中，主题可能过细或过粗；其次，主题语义的可解释性有限，尤其在神经方法中，如何确保生成的主题具有直观意义仍需研究；此外，多语言与跨领域的主题建模也面临数据稀缺与分布差异问题。

3.2.6 阅读理解

阅读理解（Machine Reading Comprehension, MRC）是衡量机器语言理解与推理能力的重要任务，其目标是让模型根据输入文本和问题自动生成答案，核心难点在于如何让机器具备类似人类的“读—问—答”能力，这不仅涉及文本信息的匹配，还依赖逻辑推理与上下文建模。MRC 的任务形式主要包括抽取式、生成式和多跳式三类，分别对应直接从原文中抽取片段、生成新的答案以及跨越多段文本甚至多篇文档进行推理。早期的方法多依赖特征工程，通过 TF-IDF、词向量相似度等手段计算问题与段落的匹配度。随着深度学习的发展，神经网络模型逐渐应用于 MRC，例如 BiDAF、QANet 等通过交互式注意力机制实现问题与文本的对齐，显著提升了抽取式问答的性能。Transformer 架构的引入进一步推动了 MRC 的突破，BERT 利用双向自注意力建模复杂上下文依赖，在 SQuAD 数据集上超越人类基准，后续的 RoBERTa、ALBERT、ELECTRA 等模型不断优化性能；与此同时，T5、BART、GPT 等生成式预训练模型将 MRC 扩展到答案生成领域，支持更复杂的问答任务。当前 MRC 的关键技术包括注意力机制以实现问题与文本的精准对齐，多跳推理方法通过图神经网络、记忆网络或链式思维（CoT）提升跨文档推理能力，长文本建模依托 Longformer、BigBird 等稀疏注意力模型处理大规模文档，检索增强生成（RAG）结合外部知识库提升答案准确性与可解释性，而提示工程与思维链推理则通过合理的提示设计，引导大语言模型逐步推理并生成答案，从而显著增强了复杂任务的表现。

3.2.7 多模态理解与生成

多模态理解与生成（Multimodal Understanding and Generation, MUG）是人工智能的重要研究方向，其目标是让模型能够同时处理并融合来自文本、图像、语音、视频等多种模态的信息，从而完成跨模态的理解和生成任务。其核心挑战在于如何实现模态间的对齐与融合，不仅涉及模态内的表示学习，还依赖跨模态语义建模与推理。多模态任务形式主要包括跨模态理解（如视觉问答、图文匹配、视频理解）、跨模态生成（如图像描述生成、文本生成图像、语音对话系统）以及多模态交互（如图文搜索与多模态推荐）。早期的方法多依赖特征拼接或浅层融合，例如通过 CNN 提取图像特征、RNN 提取文本特征后进行简单对接。随着深度学习的发展，注意力机制被引入跨模态建模，典型的 ViLBERT、LXMERT 等通过跨模态 Transformer 捕捉图文交互关系，显著提升了理解与生成效果；CLIP 的对比学习则实现了图文语义空间的高效统一，并为 DALL·E、Stable Diffusion 等文本生成图像模型提供了基础。语音模态方面，端到端的语音识别、语义理解与语音合成逐渐与大语言模型结合，推动语音对话系统向自然交互演进。当前多模态研究的关键技术包括跨模态表示学习与对比学习、多模态注意力机制、长序列稀疏建模以处理大规模视频数据、检索增强生成（RAG）以结合外部知识提升

可解释性，以及思维链推理与提示工程在多模态场景中的扩展应用。这些进展使得多模态理解与生成在通用人工智能（AGI）的发展中发挥着关键作用。

参考文献

[1] Fuchun P, Fangfang F, and Andrew M. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the international conference on Computational Linguistics, 2004, pp. 562–569.

[2] Xinchi C, Xipeng Q, Chenxi Z, Pengfei L, and Xuanjing H. Long short-term memory neural networks for Chinese word segmentation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015b, pp. 1197–1206.

[3] Fang, Q.; Li, Y.; Feng, H.; Ruan, Y. Chinese Named Entity Recognition Model Based on Multi-Task Learning. Appl. Sci. 2023, 13, 4770.

[4] Chen D, Manning C D. A fast and accurate dependency parser using neural networks[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 740-750.

[5] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[J]. arXiv preprint arXiv:1611.01603, 2016.

[6] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[7] Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. Proceedings of ICML 2014, 1764–1772.

[8] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR 2015.

[9] Graves, A. (2012). Sequence transduction with recurrent neural networks. arXiv:1211.3711.

[10] Wang Y, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.

[11] Tokuda K, Nankaku Y, Toda T, et al. Speech synthesis based on hidden Markov models[J]. Proceedings of the IEEE, 2013, 101(5): 1234-1252.

[12] Ren Y, Hu C, Tan X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv preprint arXiv:2006.04558, 2020.

[13] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. CVPR 2005, 886–893.

[14] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

[15] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1106–1114.

[16] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[17] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139, 4904–4916.

[18] Li, J., Li, D., Xiong, C., & Hoi, S. C.-H. (2022). BLIP: Bootstrapping language–image pre-training for unified vision–language understanding and generation. Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162, 12888–12900.

[19] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., et al. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.

[20] Liu, P., Yuan, W., Fu, J., et al. (2023). Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 55(9), 1–35

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[22] Church, K. W. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162.

[23] Pennington J, Socher R, Manning C D. Glove: Global vectors for word

representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[24] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets[J]. arXiv preprint arXiv:1906.05474, 2019.

[25] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[26] Radford, A., et al. (2018). Improving Language Understanding by Generative Pre-training.

[27] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9

[28] 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.

[29] Dong Q, Li L, Dai D, et al. A survey on in-context learning[J]. arXiv preprint arXiv:2301.00234, 2022.

[30] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[31] Wang, X., Wei, J., Schuurmans, D., et al. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. Advances in Neural Information Processing Systems 35

[32] Yao, S., Yu, D., Zhao, J., et al. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. Advances in Neural Information Processing Systems 36.

[33] Kimi group, KIMI K1.5: Scaling Reinforcement Learning with LLMs, arXiv: 2501.12599

[34] Li Z, Zhang D, Zhang M, et al., "From System 1 to System 2: A Survey of

Reasoning Large Language Models", arXiv: 2502.17419.

[35] Chen L, Davis J, Hanin B et al., "Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems", arXiv: 2403.02419.

[36] Cao Y, Zhao H, Cheng Y, et al., "Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods", arXiv: 2404.00282.

[37] Zhang D, Zhoubian S, Hu Z, et al., "ReST-MCTS: LLM Self-Training via Process Reward Guided Tree Search", NeurIPS 2024.

[38] Long L, Wang R, Xiao R, et al. (2024). "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey", arXiv: 2406.15126.

[39] Cai W, Jiang J, Wang F, et al. (2024). "A Survey on Mixture of Experts", arXiv: 2407.06204.

[40] Qiu Z, Huang Z, Fu J, et al. (2024). "Unlocking Emergent Modularity in Large Language Models", NAACL 2024.

[41] Wang X, Salmani M, Omidi P, et al. (2024). "Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models", IJCAI 2024.

[42] Han C, Wang Q, Peng H, et al. (2024). "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models", NAACL 2024.

[43] Hao S, Sukhbaatar S, Su D J, et al. (2024). "Training Large Language Models to Reason in a Continuous Latent Space", arXiv: 2412.06769.

第四章　语言智能应用

这一章以中国人工智能学会语言智能专委会常务委员擅长的研究领域为代表描述语言智能的应用，包括语言能力评价、东南亚低资源语言机器翻译技术、负面情感分析技术、面向多模态语言关联的语言智能。

4.1　语言能力评价

作文批改是教育智能领域的重要研究方向，其核心任务包括作文篇章要素与关系抽取、自动评分、细粒度评价及可解释性分析。近年来，研究者提出了多种创新方法以提升批改效果：在篇章分析方面，通过结构建模和语义关系抽取，识别作文的篇章逻辑与内容关联；在评分任务上，基于深度学习的回归模型和对比学习方法被广泛应用，以增强评分的鲁棒性和跨提示泛化能力；细粒度与可解释性研究则关注评分依据的可追溯性，如特征权重分析和多维评分预测。此外，对比回归学习方法通过优化特征表示空间，减少数据偏差对评分的影响；而基于思维图的评分方法则利用图结构建模作文的逻辑连贯性，进一步提升了评分的准确性。这些技术的发展为智能化、个性化的作文批改提供了新的解决方案。

4.1.1　作文批改

（1）作文篇章要素与关系抽取

作文篇章要素与关系抽取是指从文本中识别构成文章的基本成分（如主题、论点、论据、结论等）以及它们之间的逻辑关联（如因果、转折、并列等），以分析文章的结构与逻辑连贯性。这一技术广泛应用于自动评分、写作辅助和文本摘要等领域，帮助评估文章质量并提供优化建议，从而提升写作能力。

语篇级要素与关系抽取和一般的实体关系识别任务存在明显差异。其核心目标是从议论文等长文档中，精准识别出各类要素（如论点、论据、结论等），并明确这些要素之间的关系类型（如支撑、反驳、因果等）。由于这些要素在文本中往往具有篇幅长、分布跨度广的特点，这就对模型提出了更高的要求-不仅要能够理解句子关系复杂的长文本文档，还需要借助大量的上下文信息和外部知识才能完成识别任务。语篇级要素与关系抽取对于多种下游自然语言处理任务都有着至关重要的意义，例如文本摘要任务中需要明确核心观点与支撑信息的关系，问答任务中需要理清问题与答案在文本中的逻辑关联，作文自动评分任务中需要判断论点的完整性和论证的合理性等。

目前，用于语篇级要素与关系抽取任务的主流方法主要包括基于分类的方法、基于解析的方法和基于深度学习的方法。随着深度学习技术的不断发展，基于深度学习的方法在该任务上展现出了最优的性能。其中，具有代表性的基于深度学习的方法有：基于序列建模的方法（如 Transformer），它能够捕捉文本的序列依

赖关系；基于图结构的方法，可有效建模要素之间的复杂关联；基于预训练模型（Pretrained Language Model, PLM）的方法，借助预训练模型强大的语义理解能力提升抽取效果；以及联合建模与多任务学习方法，通过多任务协同训练提高模型的整体性能。此外，随着大语言模型（Large Language Models, LLMs）的兴起，基于指令微调（如 GPT-4）的生成式方法凭借其强大的生成能力，有望成为该领域新的研究突破口。

语篇要素及其关系识别的核心目标是从文本中识别出语篇要素，确定它们在文本中的功能，并构建起要素之间的关系网络[1]。现有的语篇要素提取方法可分为基于分类的方法、基于序列标注的方法、基于解析的方法、基于特征工程的方法以及基于深度学习的方法，其中基于深度学习的方法效果最为显著，例如基于层次化注意力的方法[2]，能够分别关注不同层级的文本信息；基于 CNN 和 LSTM 的方法[3]，可有效提取文本的局部特征和时序特征。

在语篇要素关系识别方面，典型的范式是将其视为分类任务[4-5]，即判断两个要素之间属于哪种预设的关系类型。其中，基于注意力的模型[6]能够聚焦于对关系识别至关重要的文本片段；基于段落语义的模型[7]从整体段落语义出发进行关系判断；基于交互注意力机制的模型[8]则通过要素之间的交互信息来增强关系识别的准确性。然而，以往的许多方法为了建模语篇要素与元素关系之间的语义交互，通常将要素识别和关系识别视为两个独立的任务，两者不共享信息[9-11]，采用先识别实体再预测它们之间关系的流程。这种分开处理的方式虽然在一定程度上简化了关系抽取任务，但却忽略了要素识别和关系识别之间的内在交互，容易导致错误传播（即要素识别的错误会影响后续的关系识别）。

为了克服这些缺点，后续的研究逐渐转向联合提取实体和关系。早期的联合提取模型依赖于手工设计的特征和外部解析器，这不仅增加了模型的复杂性，还可能引入额外的误差。随着深度学习在各领域的广泛应用，一系列基于神经网络的联合提取方法被提出。例如，一种树状结构的 BiLSTM 模型，通过参数共享的方式，实现了实体和关系的高效联合提取。随后，在该模型的基础上，用基于注意力的网络替代了 BiLSTM，以更好地学习要素之间的语义关系。之后，还引入了 seq2seq 架构，能够自然地生成实体-关系知识三元组，进一步提升了联合提取的效果。

（2）作文评分

自动作文评分技术是使用自动化建模的方式，通过对大量已评分作文进行深入分析，从中提炼出关键的评分经验和标准，并将学习到的评分经验对新的作文进行评分。这项技术与传统的人工评分相比具有显著的优势：首先，提高了评分效率，能在极短的时间内完成大量作文的评分工作；其次，从成本角度来看，计算机运行的成本远低于人力资源，且随着技术进步，自动作文评分在

细粒度评分方面的能力将进一步提高；最后，这项技术提供了更加标准化和客观的评价体系，有效避免了主观评分的影响，确保了评分结果的公平性和客观性。

作文自动评分的研究始于 1966 年[12]，由 E.Page 提出 PEG（Project Essay Grade）为开端[13]。这种早期的作文自动评分系统主要依赖于手动提取的作文浅层特征，比如文章长度、单词以及句子使用的复杂程度等对作文的质量进行评估。随着作文自动评分相关研究的发展，使用机器对作文进行评分的标准不再依托于某一类型的特征[14]。尤其是随着深度学习技术，探究如何使用深度学习方法提取作文的语义特征成为重要的研究内容[15]。比如利用自然语言处理任务中常用的 Word2Vec[16]与 GloVe[17]获取作文文本的词向量构建文本表示。为了抽取深层次作文表示，将卷积神经网络、循环神经网络和注意力机制进行结合，依次抽取作文的词级特征、句子级特征和篇章级特征，并将最后得到的篇章级特征用于获取最后的分数。

在最近的研究中，由于基于 Transformer[18]的预训练模型表现出强大的文本处理和分析能力[19]，很多的研究者开始关注如何通过微调预训练模型，如 BERT（Bidirectional Encoder Representation from Transformers）[20]、XLNet[21]等来适配下游任务，特别是在作文自动评分任务中也产生了很多的研究工作。Yang 等[22]将作文自动评分同时看为回归和排序任务，将两种不同的损失函数进行组合来微调 BERT，与仅使用单个损失函数相比（即，仅把 AES 看作回归任务或者排序任务），使用组合的损失函数表现出更好的性能。BERT 在 AES 任务上虽然表现出了极佳的性能，但是由于其天然的文本长度限制（不能超过 512 个字符），导致在长文档的作文评分中依然存在困境。为缓解这一问题，Wang 等[23]提出了基于 BERT 的多规模文本建模方法，利用 BERT 获取文本表示，通过多种文本规模的特征抽取与组合，使模型可以获取到超出字符长度限制的文本内容，丰富了文章的表示，并进一步提高了效果。

随着作文自动评分研究的发展发现，同一提示下的作文自动评分并非是一个符合真实应用场景的设置。因为在真实的评分应用场景下，大量且具有高标注质量的作文数据往往不易获取。并且，由于不同数据间的分布差异，在同提示场景下研究的作文自动评分模型在跨提示下进行应用时，评分性能会产生严重下降。因此，存在研究工作开始关注跨提示下的作文自动评分研究。

近年来，研究者们提出了多种创新的自动作文评分方法。其中，基于提前预标注的评分方法（TDNN）采用两阶段训练策略：首先在提示数据标注阶段，通过手工特征提取和排序支持向量机模型对作文进行初步评分；随后在基于提示信息的训练阶段，利用伪训练样本获取文本的深层次特征表示，这些特征经过双向长短期记忆网络（Bi-LSTM）处理，最终通过非线性变换层进行分数评估。实验

表明，该方法在同主题自动评分任务中取得了良好效果。然而，该方法依赖于目标提示作文数据的预标注质量，这在一定程度上限制了其应用。

为解决这一问题，后续研究提出了提示无关的跨提示作文自动评分方法（PAES）。该方法创新性地将词嵌入替换为词性（POS）嵌入，并与手工特征相结合，显著提升了文章表示的泛化能力。在此基础上，研究者进一步提出了针对作文特征的自动跨提示评分任务，要求模型仅使用非目标作文题目数据进行训练，同时预测目标作文题的整体分数和特定特征分数。最新的研究进展在 PAES 框架中引入对比学习技术，通过促进不同提示下作文数据的特征映射，使模型能够学习到更具通用性的作文表征，从而进一步提升对共享特征的学习能力。这些方法共同推动了自动作文评分技术的发展，为解决数据稀缺问题提供了新的思路。

当前关于跨提示作文自动评分的研究虽已取得进步，但在评分模型的架构优化和训练策略设计方面，还存在显著的研究潜力。主要体现在两个方面：首先，在模型架构方面，目前的作文评分模型主要通过分析作文的词性结构来提取文本特征，却在很大程度上忽略了语义特征在评分中的重要性。这种方法未能充分挖掘文本内容的深层意义。此外，这些模型未将作文提示信息，即作文与其题目之间的一致性特征，纳入模型的输入考虑范畴。其次，从模型的训练策略设计角度分析，现有研究未充分考虑模型初始化参数对特征学习的影响，导致模型在学习有效特征方面的能力受限。同时，作文特征在不同提示下的分布未被有效利用，这限制了模型在处理新提示时的分布迁移能力，影响了模型的泛化和适应性。

跨提示作文自动评分任务的定义如下：假设存在 K 个源提示作文数据$P_k = \{P_1, P_2, \ldots, P_K\}$作为训练集，其中每个提示$P_i$中都有对应的$N_i$个被标注过的作文数据。跨提示作文自动评分任务的目标就是利用这多个源提示数据中训练出一个评分模型，并且要求该模型可以对模型不可知的提示作文数据进行准确的评分。图 4-1 展示了同提示下和跨提示下的作文自动评分研究的区别。

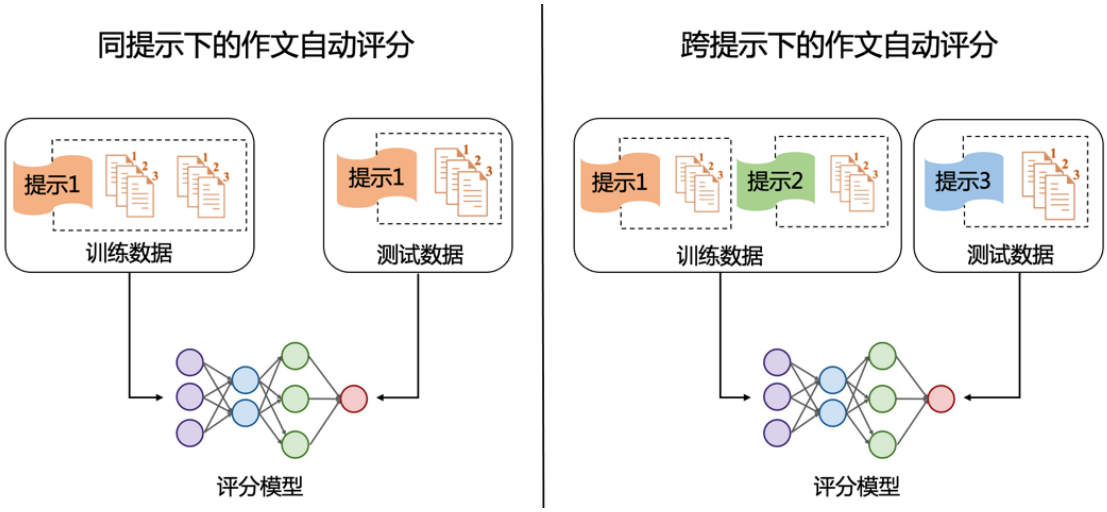

图 4-1 同提示作文评分研究和跨提示作文自动评分研究的区别[24]

对于作文自动评分任务而言，获取作文的向量化特征（文本表示）是实现自动评分关键。因此，针对作文自动评分方法的设计，与相关的获取文本特征的技术进行结合是必不可少的。同时，针对跨提示场景下作文自动评分任务，其本质其实还是关于领域泛化的研究问题。因此，针对作文自动评分任务，除了要介绍以往的与文本处理相关的技术内容，包括文本特征表示技术、文本特征抽取技术，还要对领域泛化相关方面的研究进行介绍。

以下是常用的跨提示自动评分方法：

①基于提示无关特征的方法

这类方法聚焦于提取与具体提示无关的通用特征，通过弱化提示特异性信息来增强模型的泛化能力。

手工特征与通用语义特征结合：如 PAES（Prompt Agnostic Essay Scorer），通过提取文本长度、词汇多样性、语法错误率等与提示内容无关的手工特征，结合基于 LSTM 的通用语义表示，构建对提示变化不敏感的评分模型。其核心思路是：作文的基础质量（如语言流畅度、结构完整性）在不同提示中具有共性，通过过滤提示相关的主题信息，保留通用质量特征，实现跨提示评分。

词性与句法结构特征：部分方法利用词性（POS）嵌入、句法树结构等语法层面的特征，这些特征不受提示主题影响，能稳定反映作文的语言组织能力。例如，通过 1D 卷积层提取 POS 序列特征，结合注意力机制捕捉句子间的逻辑衔接，以此作为评分依据，减少对提示特定内容的依赖。

②基于对比学习与分布对齐的方法

这类方法通过拉近不同提示数据的分布距离，使模型学习到更鲁棒的跨提示表示。

提示映射对比学习（PMAES）：该方法构建“提示—作文”对的对比学习任务，通过设计正负样本对（如同一提示的不同作文为正样本，不同提示的作文为

负样本），引导模型学习提示不变的特征。同时，引入提示映射模块，将不同提示的特征投影到共享空间，减少分布差异。实验表明，该方法能有效缩小跨提示分布差距，在 ASAP 数据集上实现了较高的评分一致性。

层级感知对比学习（PLAES）：针对不同提示下作文质量层级（如高分、低分）的分布差异，通过层级对比损失函数，让模型在学习通用特征的同时，关注同一质量层级在不同提示中的共性。例如，将高分作文（无论属于哪个提示）的特征在嵌入空间中聚集，低分作文同理，以此增强模型对跨提示质量层级的识别能力。

③基于元学习的域泛化方法

这类方法借鉴元学习“学会学习”的思想，通过训练模型快速适应新提示，实现有意识的分布偏移引导。

元学习优化框架：该方法构建多组虚拟跨提示任务，每个任务包含虚拟源提示（元训练数据）和虚拟目标提示（元测试数据）。通过多元训练生成多个元学习器，再基于最大均值差异（MMD）选择与目标提示分布最接近的元学习器进行优化，确保模型向目标提示方向偏移。同时，结合大语言模型（LLMs）进行数据增强，通过同义词替换和质量评估生成多样化训练数据，提升元学习任务的丰富性。该方法在 ASAP 数据集上的平均 QWK 得分达 0.701，优于现有对比方法。

域适应元学习：通过元训练让模型学习“如何调整参数以适应新提示”，例如在元训练阶段引入动态损失权重，对不同提示的数据分配自适应权重，使模型更关注难迁移的提示特征。元测试阶段，模型能快速利用少量目标提示数据微调参数，实现高效适应。

④基于手工特征与规则的方法

这类方法依赖人工设计的跨提示通用规则，适用于数据稀缺场景。

全手工特征方法：通过提取大量与提示无关的手工特征（如句子平均长度、从句比例、连接词频率等），结合简单神经网络或回归模型进行评分。其核心假设是：作文的质量可通过语言统计特征客观反映，且这些特征在不同提示中具有稳定性。该方法虽无需复杂模型，但依赖特征工程经验，泛化能力受限于特征设计的全面性。

⑤基于大语言模型（LLMs）的零样本方法

随着 LLMs 的发展，部分研究探索利用其强大的语义理解能力进行跨提示评分。

零样本提示工程：通过设计细粒度评分指令（如“评估作文的论点明确性、论据相关性”），引导 LLMs 按统一标准对不同提示的作文打分。例如，将作文和评分维度指令输入 GPT-3.5，让模型生成各维度分数，再整合为总分。这类方法

无需训练数据，但评分一致性受限于 LLMs 对提示的理解偏差，在 ASAP 数据集上的表现仍低于监督学习方法。

这些方法从不同角度应对跨提示评分的分布偏移问题：特征工程方法注重通用性特征的挖掘，对比学习方法强调分布对齐，元学习方法聚焦快速适应能力，而 LLMs 方法则依赖大规模预训练知识。实际应用中，需根据数据规模、提示差异程度选择合适方法，例如数据充足时优先考虑对比学习或元学习方法，数据稀缺时可采用手工特征或 LLMs 零样本方法。

⑥元学习自动作文评分

元学习自动作文评分中，元学习（Meta-Learning）也被称为“学习去学习”，它作为少样本学习和领域泛化相关任务的一种主要方法，核心是构建一种学习框架，让模型能够适应新的知识或任务，即便在只有少量数据甚至没有目标数据的情况下，也能实现泛化。在元学习的各类算法里，基于优化策略的元学习因为具备灵活性，受到了不同领域的关注，得以在多种领域中应用。这类算法通常采用双层优化结构：外层优化负责训练一个学习算法，以此来实现泛化；内层优化则运用该学习算法调整基础学习器，使其适应只有少量示例的新任务。MAML（Model Agnostic Meta-Learning）是基于优化策略的元学习方法中最重要的一种，它的提出是为了以模型无关的方式提升泛化能力，让模型能快速适应新领域的任务。它会学习一组初始的网络权重数值（即元学习器）来实现泛化，而学到的这组初始化参数能作为一个良好的起点，帮助模型在面对只有少量示例和少量更新的新任务时快速适应。

基于元学习提出了元学习跨提示作文自动评分方法（Prompt-adaptive Meta-learning for Cross-prompt AES，PMCAES），在模型的输入方面，除了源提示的作文数据，还会将目标提示的作文文本纳入其中，为训练过程提供目标提示的数据信息，并且通过最大均值差异方程（Maximum Mean Discrepancy, MMD）来度量不同数据分布之间的差异，从而实现学习器的筛选。

由于不同提示的作文数量存在显著差异，为了防止在元学习训练步骤中重复抽样相同的题目数据并扩展元学习训练任务，同时为进一步提高元学习训练任务的多样性，提出利用大语言模型进行数据增强，借助大语言模型的内容理解和文本生成能力，扩充跨提示作文自动评分任务的训练数据。

为进一步实现有针对性的分布迁移，相关研究还提出提示自适应的元学习优化策略，其中包含两个阶段的训练过程。首先，第一阶段是执行元学习器检索的元训练阶段，利用目标提示的数据信息与多个源提示信息进行对比，借助 MMD 分布度量方法，根据最小的度量值选择与目标信息最接近的初始化模型状态进行训练，控制模型朝着目标分布泛化。第二阶段是分布可感知的元测试阶段，通过在元测试阶段的训练过程引入自适应的分布迁移分数，衡量元训练数据和元测试

数据间的分布差异以实现进一步领域泛化的控制。

（3）自动作文评分的细粒度与可解释性探索

在当今的人工智能时代，深度学习凭借其强大的性能，在图像识别、自然语言处理等众多领域大放异彩，为我们的生活和工作带来了诸多便利。然而，深度学习模型的“黑盒”性质也一直是人们关注的焦点，尤其是在需要对结果进行清晰解读的场景中，模型的可解释性显得尤为重要。

作文评分作为教育领域的一项关键任务，其公平性、准确性和实用性备受关注。传统的作文评分方式往往依赖于人工，不仅效率低下，而且受评阅者主观因素影响较大。随着深度学习技术的发展，自动作文评分系统应运而生，但其同样面临着如何让评分结果更精细、更易被理解的问题。正是在这样的背景下，细粒度与可解释性成为了提升自动作文评分模型质量的关键所在，围绕这两者展开的研究和方法探索也逐渐深入，比如基于对比回归学习和思维图的作文评分方法等，都在为实现更优质的自动作文评分贡献力量。

①细粒度与可解释性

深度学习因其出色的性能而备受关注，但其黑盒性质也引发了人们对模型可解释性的研究兴趣。在可解释性研究领域，学者们提出了多种创新方法：2016 年，Marco 等人[25]提出了 LIME 方法，该方法通过学习预测周围的局部可解释模型，以可解释且忠实的方式解释任何分类器的预测结果。2019 年，Rajani 等人[26-27]收集了以自然语言序列形式呈现的人类常识推理解释，构建了名为常识解释(CoS-E)的新数据集。他们进一步利用 CoS-E 训练语言模型来自动生成解释，这些解释可应用于新颖的常识自动生成解释(CAGE)框架的训练和推理过程。2022 年，提出的 SELOR 框架将自解释能力集成到深度模型中，通过逻辑规则作为解释形式，同时实现了高预测性能和人类可理解的解释精度。

作文评分的细粒度与可解释性是提升评分模型实用性和可信度的两个核心维度，二者相互关联且各有侧重。细粒度指对作文评分的维度、层级和反馈进行精细化拆分，突破传统“整体打分”的局限，从多个具体角度评估作文质量，比如将评分从单一的“整体分数”扩展到对论点明确性、论据相关性、逻辑结构、语言流畅度、语法准确性等多个核心要素分别评分，每个维度还可进一步细化为具体等级或描述，如将“论据相关性”分为“完全无关”“部分相关”“高度相关”，并针对这些细粒度维度的表现提供具体反馈，像指出“第二段的事实论据与分论点关联性较弱，建议补充与主题相关的具体事例”，而非笼统的评价。可解释性则指模型的评分结果能够被人类理解，即清晰展示“为什么给出该分数”，通过透明的推理过程或依据让用户信任并理解评分逻辑，这包括明确评分所依据的规则或标准，如说明“因为作文未明确提出主论点，故该维度扣 2 分”，展示模型在评分时关注的关键信息，如指出“结论与论点不一致是导致总分偏低的主要原

因”并举例说明矛盾之处，以及用自然易懂的语言解释评分理由，避免技术术语，比如告诉学生“你的作文分论点清晰，但每个分论点下仅用了一个例子，论据不够充分，建议增加不同类型的证据来支撑观点”。细粒度是可解释性的基础，只有先拆分出具体的评分维度和要素，才能针对性地解释每个维度的评分依据；而可解释性则让细粒度评分更具实用价值，通过具体的维度反馈，学生能明确改进方向，教师也能更高效地辅助教学。

文档级文本理解对深度学习模型具有较大挑战性，需要更丰富的上下文信息来辅助模型理解和泛化。特别是在作文评分任务中，生成解释比情感分类更为复杂：单独的词语难以构成有效的评分解释，需要句子级别的文本才能从多个维度全面评价作文质量。为此，相关研究[28]提出将篇章级规则解释与原始作文文本相结合作为模型输入，以增强输入信息量；同时，这些篇章级规则解释也可作为事后解释提供给用户，提高模型决策的透明度。

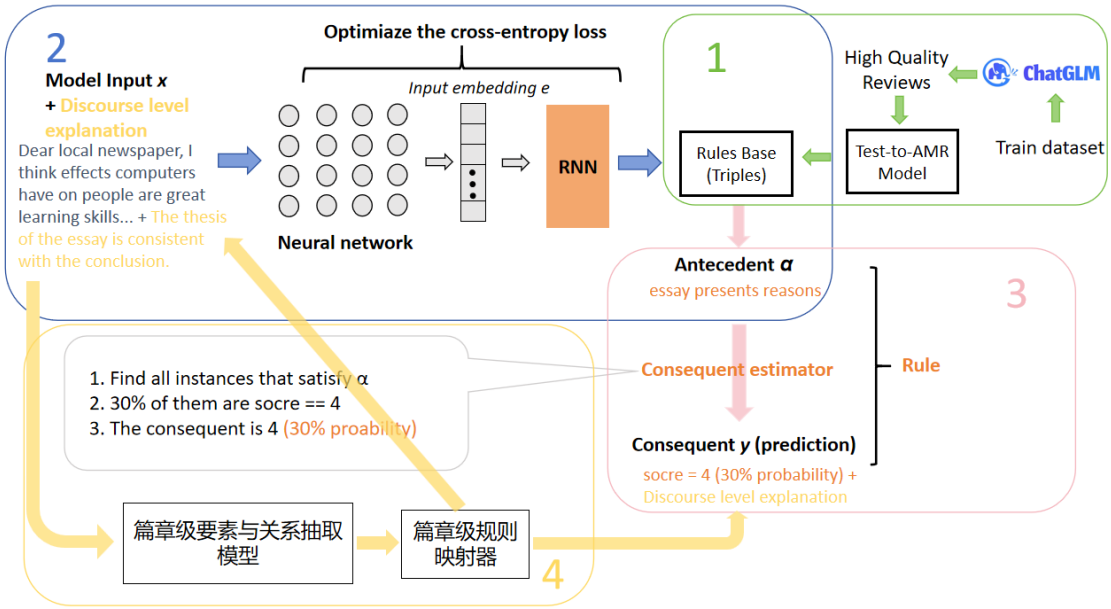

图 4-2 篇章级规则解释方法[28]

②基于对比回归学习的作文评分方法

对比学习是一种通过让模型区分数据相似性来学习有效特征表示的自监督学习方法。其核心在于构建正负样本对进行对比训练：正样本通常来自同一数据的不同增强版本或语义相似的实例，而负样本则来自不同类别的数据。模型通过优化对比损失函数（如 InfoNCE 损失），在特征空间中拉近正样本之间的距离，同时推远负样本之间的距离，从而形成结构化的特征表示空间。这种方法模拟了人类通过比较来认知事物的学习机制，能够在不依赖大量标注数据的情况下，自动捕捉数据的内在结构和本质特征。对比学习在计算机视觉、自然语言处理等领域展现出强大的表征学习能力，特别是在数据标注成本高昂的场

景中具有显著优势。在作文评分任务中，对比学习可以帮助模型更准确地识别不同质量作文之间的细微差异，通过构建分数相近的作文作为正样本、分数差异大的作文作为负样本，使模型学习到更具判别力的作文特征表示，从而提升评分模型的准确性和鲁棒性。这种学习方式不仅增强了模型对作文质量层次的分辨能力，还能有效缓解标注数据不足的问题，为自动作文评分系统提供了新的技术路径。

回归学习（Regression Learning）是监督学习的一个重要分支，其核心目标是建立输入变量与连续型输出变量之间的映射关系。与分类任务预测离散类别不同，回归模型旨在预测连续的数值，如价格预测、分数评估等实际应用场景。在机器学习框架下，回归学习通过优化算法最小化预测值与真实值之间的差异（通常采用均方误差或平均绝对误差作为损失函数），逐步调整模型参数以获得最优的拟合效果。典型的回归方法包括线性回归、决策树回归、支持向量回归等传统算法，以及深度神经网络等现代方法。这些方法通过不同的数学建模方式捕捉数据中的非线性关系，在保持模型可解释性的同时提高预测精度。回归学习在自动作文评分系统中具有关键作用。系统需要将文本特征映射到连续的分数区间，这就要求模型不仅能理解文本语义，还要精确量化作文质量水平。与传统硬分类相比，回归方法可以捕捉评分标准中更细微的差异，如区分 85 分和 86 分作文的细微差别。通过结合深度表征学习和回归优化，现代 AES 系统能够实现接近人工评阅水平的评分性能，同时保持较好的跨题目泛化能力。

基于对比回归学习的作文评分方法，是一种融合相对关系学习与绝对分数预测的自动化评分技术，其核心在于不仅让模型学习作文文本特征与分数标签的直接映射，更通过构建不同作文间的质量对比关系，提升模型对评分尺度的理解能力，从而实现更贴合人类主观判断的评分结果。

这种方法的实现，首先需要进行数据预处理与对比样本构建，收集大规模带有人工评分的作文语料后，按主题、体裁等维度分类以保证同类样本的可比性，再从同类样本中随机选取若干组作文对，依据人工评分标注它们的相对关系，为模型提供对比学习的“参照物”。接着进入特征提取与表示学习阶段，借助 BERT、RoBERTa 等预训练语言模型对作文文本进行编码，将文字转化为包含语义、逻辑、情感等信息的向量表示，关键在于让模型捕捉到议论文的论点支持力度、记叙文的情节完整性等细粒度特征，为后续对比奠定数据基础。而对比损失函数设计是实现过程的核心环节，该损失函数需同时达成两个目标，一是通过均方误差等回归损失让模型预测的分数接近人工标注的绝对分数，二是通过对比损失让模型对作文对的相对优劣判断与标注一致，比如当 A 的人工评分高于 B 时，模型要让 A 的预测分数向量与高分标签的距离小于 B，以此强化对相对关系的认知。最后是模型训练与优化，训练中模型会交替学习绝对分数预测和相对关系判断，

通过反向传播不断调整参数，随着训练迭代，逐渐掌握“在保证绝对分数准确性的同时，区分相似作文质量差异”的能力，最终形成稳定的评分模式，训练完成后，模型可直接对新作文输出分数，且评分结果既能反映绝对质量，又能体现同类作文的相对排名。

③基于思维图的作文评分方法

思维图是一种以图形化方式组织和表达思维过程的工具，它通过节点、连线、色彩和图像等元素，将抽象的思考过程具象化为可视化的网络结构。这种工具起源于认知心理学对人类思维模式的研究，其本质是模拟人脑神经元网络的连接方式，帮助使用者更好地整理、分析和记忆信息。思维图的核心价值在于能够同时激活大脑的逻辑思维和形象思维，通过空间布局呈现信息的层级关系和关联性，使复杂的思维过程变得清晰可辨。在教育领域，思维图被广泛应用于教学设计、知识梳理和学习策略培养；在商业决策中，它可以帮助团队系统分析问题、激发创意和制定战略；在科研工作中，则能有效辅助文献综述、实验设计和理论构建。随着信息技术的发展，数字化的思维图工具进一步拓展了其应用场景，支持多人协作、动态演示和智能分析等功能。在作文评分领域，思维图既可以作为评分标准的可视化框架，帮助评分者系统把握评分维度，也可以作为评分依据的呈现方式，增强评分结果的透明度和说服力，同时还能作为写作教学的工具，帮助学生构建清晰的写作思路。相比传统的线性笔记或文字描述，思维图的最大优势在于其能够同时呈现信息的整体结构和细节关联，促进发散性思维和系统性思考的结合。

基于思维图的作文评分方法，是一种通过解析作文内在的逻辑结构与内容关联来实现评分的技术路径。它将作文视为一个由主题、论点、论据及过渡关系构成的网络系统，通过构建可视化的思维图（或称思维导图），把文本中的隐性逻辑转化为显性的节点与连接关系，再基于这些结构化信息进行评分判断。

这种方法的实现逻辑紧密围绕思维图的构建与分析展开。首先，利用自然语言处理技术对作文文本进行深层解析，识别出核心主题（根节点）、分论点（子节点）、支撑论据（叶节点），以及节点之间的因果、递进、并列等逻辑关系（连接边）。例如，在议论文中，模型会定位“中心论点”与“分论点 1”“分论点 2”的从属关系，以及“分论点”与“具体案例”的支撑关系，从而生成反映文章骨架的思维图。

接着，通过设计量化指标对思维图进行特征提取。这些指标既包括结构性特征，如节点层级的完整性（是否涵盖主题相关的多维度分析）、连接边的合理性（逻辑是否存在跳跃或矛盾）；也包括内容性特征，如论据节点的丰富度（案例、数据的详实程度）、主题节点与子节点的关联强度（分论点是否紧扣中心）。同时，思维图的复杂度与简洁度平衡也是重要评分维度-过度冗余的节点可能反映内容

拖沓，而节点过于稀疏则可能意味着论述单薄。

最后，模型通过学习人工评分与思维图特征之间的映射关系完成训练。训练数据中，每篇作文会对应人工评分及生成的思维图特征向量，模型通过回归学习掌握“思维图结构越完整、逻辑越严密，评分越高”的规律。在实际评分时，模型先为新作文生成思维图，提取特征后，即可输出符合人类对“文章逻辑性与内容深度”判断的分数。这种方法的优势在于突破了传统文本评分对“关键词匹配”的依赖，更精准地捕捉到作文质量的核心—逻辑的严谨性与内容的系统性，尤其适用于对结构要求较高的议论文、说明文等文体。

4.1.2 儿童语言能力评价

儿童语言能力评价主要包括儿童叙事能力评价语料库开发和可解释的儿童叙事能力评价方法设计。

（1）儿童叙事能力评价任务概述

叙事能力指复述个人经历、重述听闻或阅读过的故事，以及进行故事创作的能力[29]。儿童叙事能力在儿童的认知、情感及社会性发展中扮演着关键角色，对儿童全面发展起到了核心作用[30][31][32][33]。**叙事能力评估**不仅能客观衡量儿童语言发展水平，更在语言障碍的早期诊断与干预中发挥着关键作用[34][35]。

在**临床语言学**领域，叙事能力评估始终是研究重点。现有研究主要从宏观结构[36]和微观结构[37]两个维度展开分析。宏观结构指故事的整体组织框架，通常通过故事语法成分或故事结构来界定[38][39]；而微观结构则聚焦于局部语言特征，涵盖故事长度、词汇多样性、句法复杂性及衔接连贯性等要素。由于微观结构特征较易量化，研究重点已逐渐转向宏观结构的连贯性[40]与完整性[41]。

然而在叙事能力评估任务中，研究者通常完全依赖对儿童叙事样本的**人工分析**，这种方法存在耗时费力的现实困境。此外现有叙事能力自动化评估通常**缺乏可解释性**，仅依靠宏观或微观分数代表儿童当前叙事能力水平，难以在更广泛实践中推广运用。

儿童叙事能力评价任务与**多维度自动作文评分（AES）**存在相似之处——后者需针对不同文体从内容、语言运用等多维度进行评价。近期研究大多采用自回归的多维度分数生成框架，通过词元生成概率进行评分[42][43]。然而与多维度 AES 相比，儿童叙事能力自动化评估在以下方面有所不同：（1）叙事评估主要关注故事内容的完整性与表达连贯性；（2）评估结果不仅要求高精度，更需要具备直观性与可解释性，以便为后续干预提供可操作的反馈。现有相关研究较少，Hassanali 采用主题建模预测语言障碍与连贯性[44]，Jones 仅使用机器学习方法对宏观结构评分[45]。显然，这些研究均未有效解决上述问题。

为解决上述问题，我们利用前沿的自然语言处理（NLP）技术设计了基于叙事图的儿童叙事能力自动化评估框架。我们首次提出了**叙事图**（narrative graph）

作为叙事文本的结构化表征方法，其灵感来源于临床语言学中的因果网络[46]。虽然因果网络能直观呈现文本结构，但该结构仅将子句作为节点，难以清晰表达复杂的叙事内容。相比之下，在我们的叙事图中，节点代表特定事件，边则捕捉事件间关系，包括多种因果与同步连接（示例见图 4-3）。与扁平化的非结构化叙事文本相比，叙事图提供了简洁的摘要式表征，有助于显式衡量和计算完整性、连贯性等关键叙事指标；同时，通过将评估结果与金标图（由专家人工标注的代表成年水平的最高评分叙事图）进行对比分析，可自然增强评估过程的可解释性。

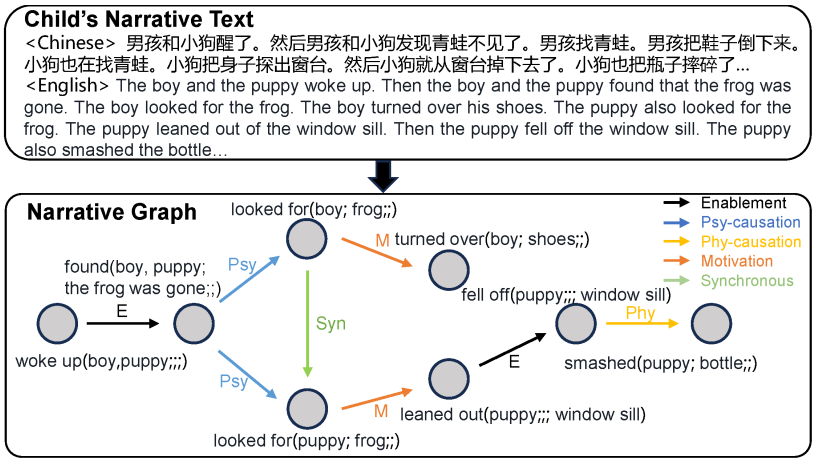

图 4-3 叙事图示例

为儿童叙事能力自动化评估框架提供数据基础，我们收集了真实的儿童叙事能力样本，设计了叙事图标注规范，开发了高质量儿童叙事能力评价语料库。叙事能力评估框架包含叙事图构建模块与叙事评分模块。叙事图构建模块使用思维链（CoT）[47]、群体相对策略优化（GRPO）[48]等方法抽取事件节点，使用 LLM 作为图神经网络增强器对节点进行编码，使用混合专家（MoE）[49]替代分类模块构建叙事边。此外，我们将叙事图与金标图对比，为叙事图补充图间差异信息。叙事评分模块采用多视角对比学习策略预训练的图编码器编码叙事图，通过 projector 模块对齐图表征与文本表征，最终基于指令微调的 LLM 生成多维度评分。在给出得分的同时，该模块将原始文本、差异图及多维度评分以 zero-shot 的方式输入至 LLM 中，自动生成可解释性分析报告。

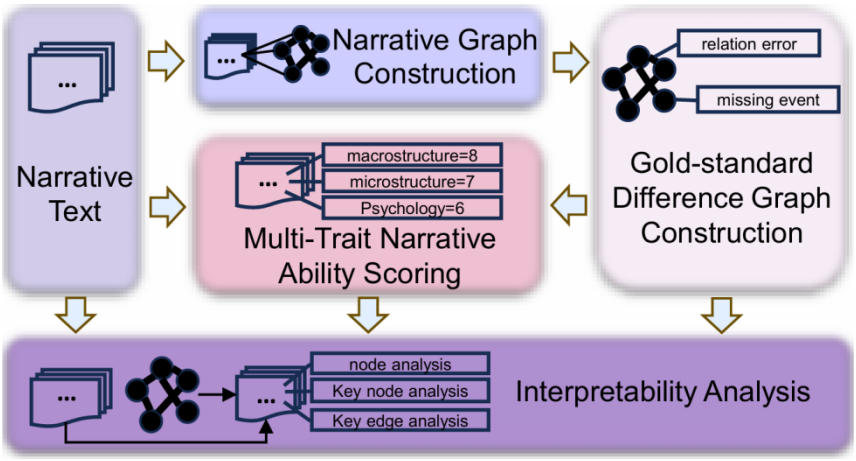

图 4-4 叙事能力评估系统框架图

（2）儿童叙事能力评价语料库构建

为了对儿童叙事能力评价和叙事图构造任务进行全面的探索与研究，构建一个相应的标注语料库是不可或缺的首要任务，它也是模型的训练、评估的基础。因此，该语料库不仅要有一定的数据规模，还需要有一定的标注质量保证。

整体而言，儿童叙事能力评价语料库的构建流程可以分为这样几个步骤：收集儿童讲述故事的录音并转录清洗为文本形式；制定具有语言学依据并且有可执行性的语料库标注规范；开发并维护一款满足儿童叙事能力评价任务需求的标注工具；培训并组织标注者进行语料的标注；对标注完的数据进行检查和验证，整理得到初版语料库。

最终，中文儿童叙事能力评价语料库包含 543 篇标注叙事文本，其中 503 份来自儿童参与者，40 份来自成人对照组。每篇文本均配有叙事图标注，并包含两位专家对整体叙事能力评分及三个核心维度得分：宏观结构、微观结构与心理状态维度。

①语料库的收集

在语料的收集过程中采用无文字图画书作为诱发材料引发儿童的叙事，选择的阅读材料是在儿童语言能力评测中被广泛使用的《Frog, where are you?》[50]，该书由 25 幅图画组成，讲述了一个小男孩抓到一只青蛙放在瓶子里，一觉醒来后青蛙不见了，小男孩带着他的小狗到处寻找，在寻找的过程中遇到了蜜蜂、地鼠、猫头鹰、麋鹿，期间还经历了掉下山崖、落进河塘的遭遇，最后在一根枯树后面看到一群青蛙，从中找到自己那只青蛙的过程。

获得的原始数据由语言学专业的老师、同学从儿童讲述故事的录音人工转录而来，在转录过程中参考了国际标准的儿童语言数据交流系统（Child Language Data Exchange System, CHILDES）[51]的数据处理流程，转录格式遵循文本赋码系统 CHAT（Code for the Human Analysis of Transcript）[52]的格式，CHAT 格式的文档中除了对被试者话语的基本转录，还包括了文件头和附属行，文件头包含了背景信息如被试者和转写者的个人信息、测试日期、撰写日期等资料，附属行记录了对转录出的具体内容的编码、评价、事件以及其他研究者关心的辅助信息，主体部分还对词语省略、话语重复、语句纠正等现象用特殊的转录符号做了标记。因此，在开始标注之前还需要将主体部分转录出的叙事文本进行保留，清洗掉 CHAT 格式中的文件头内容，对话语重复、语句纠正等特殊标记的地方进行还原和处理。

收集、清洗后得到的语料数据共有 543 篇文档，其中的文档都来自于被试者对无文字图画书《Frog, where are you?》的讲述，后续的对叙事文本中事件、事件关系的标注工作和对两阶段抽取方法的评估与分析工作都基于这些数据开展。

②语料库标注规范的设计

在开始数据标注流程之前，首要的任务是预先制定一套详细、有依据的标注规范体系，以便于标注者理解、执行，从而确保所标注数据的整体质量。这一过程尤其强调对样本质量的双重评估标准，即标注的一致性和准确性。一致性维度着重于在整个数据集的标注过程中执行统一的标注原则，确保文本中的事件描述、事件关系，均遵循一致且标准化的标注规则，从而减少因人为操作、个人主观理解差异带来的混乱。另一方面，准确性层面则聚焦于严格把控标注细节的真实性与精确性，这一步骤通常会借助交叉验证机制来达成，如多人标注同一份文档并进行事后核查与讨论，目的是检查并确认所有事件组成要素、触发词、事件关系都被正确地标记，从而使得标注数据的有效性有一定的保障。

在研究前期，结合语言学领域对儿童语言能力评价的相关研究，我们针对标注过程中的事件、事件关系与评分等概念制定了标注规范。

③标注过程

为确保标注的准确性与可靠性，标注过程分为两个阶段实施：（1）叙事图标注：14 名经培训的标注人员组成 7 个小组，每组独立标注相同的转录文本集。初步标注完成后，通过一致性检验与小组讨论解决分歧，确保标注者间一致性达到较高水平。（2）叙事能力评分：鉴于该任务需要深入理解儿童语言发展与叙事能力的专业性，我们邀请两位分别具有临床语言学与教育心理学背景的领域专家，依据预定义的评分标准独立对每个样本进行评分。该标注策略不仅有效减少了主观偏差，还显著提升了数据集的整体效度与信度。

金标图由语言学专家基于高质量成人叙事样本开发完成。该图反映了结构完善的叙事故事，为系统评估儿童叙事能力提供基准。通过提供清晰统一的参照标准，使得针对不同儿童样本的叙事能力评估更具客观性、可靠性与可比性。

为提升标注效率，我们开发了专用标注工具。该工具基于 Python 内置的标准图形用户界面（GUI）库 Tkinter 定制开发并迭代实现。作为 Python 原生支持的模块，Tkinter 库具有高度兼容性与稳定性。

④统计分析

为了得到高质量的语料库，被试者的年龄范围覆盖了 3~4 岁、5~6 岁、7~12 岁的区间。另外，为了得到可供参考的、较完善的叙事图，在工作前期还收集了 40 份成人被试者的叙事语料，来自这 40 份成人被试者的语料也被包括在语料库中。此外，数据中主要标注和使用的信息包含文本中的事件及事件关系，而对于叙事文本体现出的叙事质量的评估和分数、评语等信息仍需专家进一步标注。

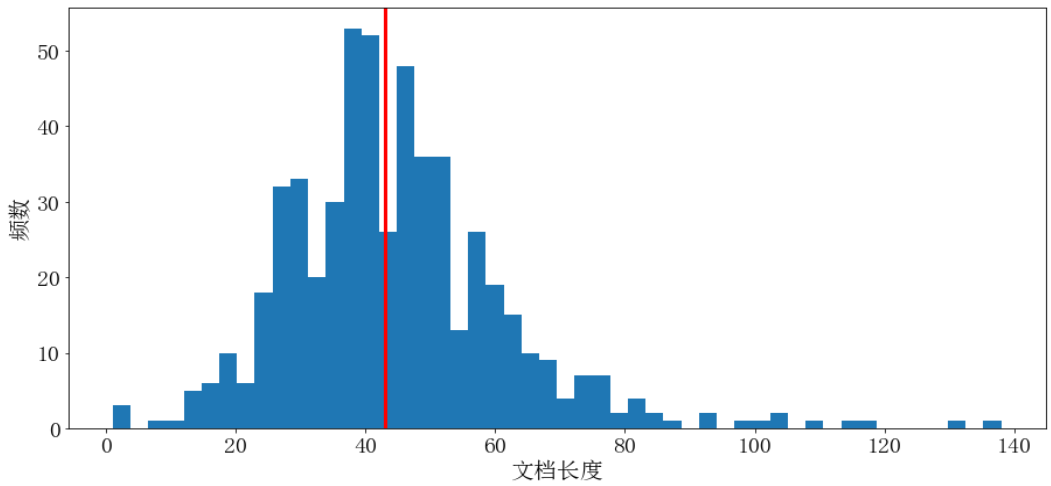

图 4-5 儿童叙事能力评价语料库中的文档长度分布统计

图 4-5 是对儿童叙事能力评价语料库中文档长度分布统计的直方图。在这个直方图中，横轴代表文档的长度，其度量单位是句子数量，而纵轴则代表各个长度范围内的文档数量，单位为文档数。从图中可以看出语料库中的绝大多数的文档的长度都位于 25 句到 70 句的区间内,超过 80 句长度的文档则基本都为成人被试者所贡献的语料。直方图中竖直的红线代表整个语料库中所有文档长度的中位数，值为 43.0，单位为句。

最终得到的儿童叙事能力评价语料库包含了 543 篇文档，共标注出了 19916 个事件、32998 个论元和 16124 个事件关系，覆盖了广泛的叙事情境，较为全面地反映了被试者在不同年龄段和认知发展阶段的叙事特点和表达能力。此外我们也预先进行了训练集、验证集和测试集的划分，在划分中尽量让不同年龄段与不同分数水平的被试者贡献的语料均匀分布在划分的 3 个数据子集中，具体的划分及相关统计指标如表 4-1 所示。

表 4-1 语料库相关信息统计

	文档数量	句子数量	事件数量	论元数量	事件关系数量
训练集	380	12673	14232	23710	11542
验证集	55	1894	2224	3469	1565
测试集	108	3248	3571	5819	3017
总计	543	17815	19916	32998	16124

（3）可解释的儿童叙事能力评价方法设计

在自动化叙事能力评估领域，核心挑战在于如何同时捕捉叙事文本的结构特征与语义信息，并确保评估结果具备可解释性。

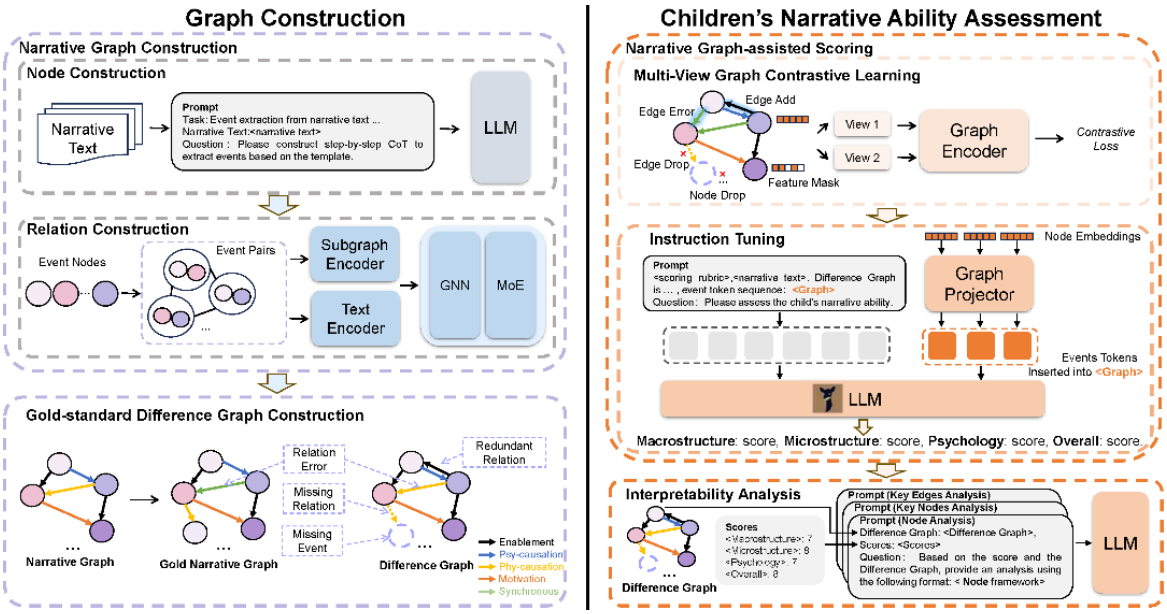

图 4-6 儿童叙事能力评价框架流程图

为此，我们提出可解释的儿童叙事能力评价框架。如图 4-6 所示，该框架采用两阶段处理流程：叙事图构建阶段与叙事能力评估阶段。在第一阶段，该框架将儿童叙事文本转化为结构化的叙事图，提供简洁且可解释的叙事内容表征。随后将这些生成的叙事图与黄金标准叙事图进行对比，补充差异信息以突显缺失、冗余等错误信息。在第二阶段，该框架采用预训练的图编码器对差异图进行编码，将编码后的节点表征与原始文本融合后输入经指令微调的 LLM，最终生成叙事能力评分。基于评分结果与叙事图，框架可进一步从多视角进行可解释性分析。

①儿童叙事图构建模块

叙事图构建的过程包含节点构建与边关系构建两个核心环节。叙事图构建主要面临以下挑战：（1）儿童口头叙述中存在语法成分缺失、句子重复、词序混乱等不规范现象，在节点构建过程中需要补充额外信息提升构建的准确率；（2）由于叙事图的结构导致的数据稀疏性，加之儿童口头表述不规范引发的节点缺失、触发词或论元遗漏等问题，在边关系构建时需要进行细粒度数据处理与数据增强。

此外，尽管叙事图已被纳入评分流程，但最终评分结果仍缺乏可解释性。为解决该问题，我们通过将儿童叙事图与金标图进行对比，补充差异信息以增强结果的可解释性。

②叙事图辅助的叙事能力评估模块

述叙事图辅助评分任务主要面临三大挑战：（1）叙事能力涉及多维度评估，要求模型具备强大的推理能力以捕捉跨事件逻辑结构；（2）由于模态差异，现有基于语言模型的评分方法无法直接利用叙事图结构信息；（3）仅依赖编码后的叙事图进行评分缺乏足够的可解释性，需要结合儿童叙事图与金标图之间的差异信息来增强评分过程的可解释性。

此外，叙事能力评估不仅需要生成多维度的量化分数以评估儿童叙事能力，还应提供这些评分的可解释性依据。具体而言，必须通过分析差异图与儿童叙事文本，从宏观结构、微观结构或心理状态表达维度深入解析叙事能力水平较差的成因，从而为评分结果提供有意义的支撑。

（4）总结

我们针对儿童叙事能力的自动化评估任务，提出可解释的儿童叙事能力自动化评估框架，通过利用叙事图显式评估完整性、连贯性等关键叙事指标。为增强可解释性，我们进一步将生成的叙事图与金标图融合，从而实现更全面且可解释的评估过程。此外，我们首次构建了中文儿童叙事能力评估语料库，为后续研究提供高质量数据资源。基于该语料库，我们提出了自动化构建叙事图模型，利用叙事图和原始文本进行评分并生成可解释性分析。

4.2 东南亚低资源语言机器翻译技术

4.2.1 东南亚低资源语言机器翻译概述

机器翻译（Machine Translation, MT）[53]作为人工智能与自然语言处理（Natural Language Processing, NLP）[54]的核心任务之一，其目标在于利用计算机实现不同语言间的自动转换。该技术历经了从基于规则、统计机器翻译（Statistical Machine Translation, SMT）[55]到当前主流的神经机器翻译（Neural Machine Translation, NMT）[56]的发展阶段。特别是 Transformer 架构[57]的提出，推动了神经机器翻译技术的突破，凭借其强大的语义建模能力和端到端学习机制，在翻译质量和流畅度上取得显著进步。

然而，神经机器翻译的性能高度依赖大规模、高质量的双语平行语料，这对低资源语言（Low-Resource Languages）构成严峻挑战。东南亚语言生态复杂多样，除印度尼西亚语、越南语、泰语等有一定数据积累的语言外，还包括老挝语、高棉语（柬埔寨语）、缅甸语及众多民族语言。这些语言普遍面临语料稀缺、技术支撑不足的问题，同时在语系归属、形态结构和文字系统方面差异显著，进一步加大了机器翻译的难度[58]。

为应对这些挑战，东南亚低资源语言机器翻译研究已形成多个重点方向。在神经机器翻译框架下，研究者通过数据增强、引入语言知识以及多语言协同训练实现知识迁移[59]。随着预训练大语言模型的兴起，基于提示学习（Prompt-based Learning）的少样本/零样本翻译、参数高效微调等方法也为低资源语言翻译提供了新的解决方案[60]。在应急通信、远程医疗等实时性要求较高的场景中，语音翻译技术成为实现无障碍沟通的关键支撑。

机器翻译技术在东南亚低资源语言领域的发展与应用，在语言保护、经济发展和社会服务等多个维度产生深远影响。随着相关技术的持续进步，高质量、易使用的机器翻译系统将在东南亚地区发挥更加关键的基础设施作用[61]。

4.2.2 东南亚低资源语言神经机器翻译技术

（1）数据增强的低资源语言神经机器翻译

低资源语言神经机器翻译的核心挑战在于高质量平行语料的严重匮乏。数据增强技术通过挖掘和利用现有数据资源的潜在价值，成为缓解该瓶颈的关键途径。近年来，该领域已形成多种方法，主要包括平行句对抽取、伪平行数据生成以及训练数据增强等方向，旨在通过提升训练数据的规模与多样性来增强模型的泛化能力[62-64]。

在平行句对抽取方面，一种具有代表性的方法是 Wu 等人提出的抽取-编辑（extract-edit）双语数据抽取方法[65]。该方法并非专门针对东南亚语言，而是一种通用的无监督数据构建策略，适用于包括低资源语言在内的多种场景。其核心思想是从目标语言的单语语料中抽取并编辑真实句子，以生成高质量的伪平行语料，从而替代传统的回译方法。

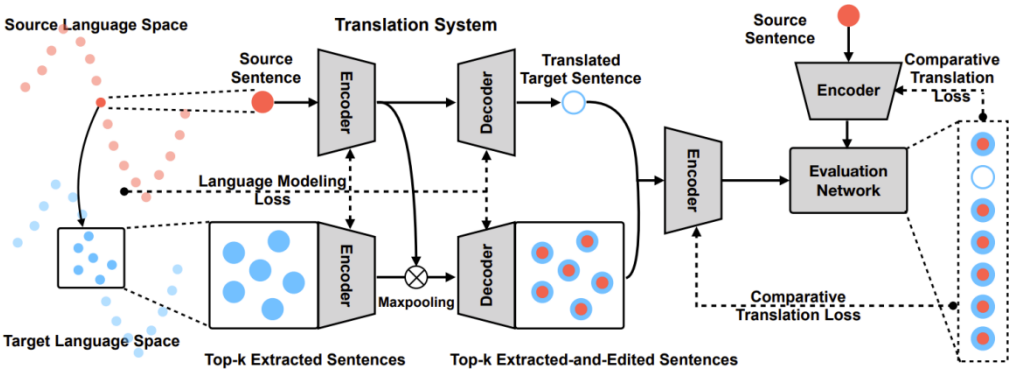

图 4-7 抽取-编辑（extract-edit）双语数据抽取方法

该方法的关键创新在于引入比较性翻译损失来评估翻译结果，具体流程包括：基于源句子与目标句子的相似性进行最近邻搜索，抽取前 k 个潜在平行句子；随后对这些句子进行编辑优化；最后通过评估网络对翻译句子和编辑后句子的相似性进行判别。实验表明，该方法在英语-法语、英语-德语等基准测试以及英语-罗马尼亚语、英语-俄语等低资源语言对中均取得显著效果，BLEU 分数提升超过 2 分。

在汉语与东南亚语言的平行句对抽取研究中，学者结合区域语言特点，探索了利用枢轴语言[66]、融入句子结构特征[67]和语言模型[68]等策略，以提升对齐质量。近年来，卷积相关网络（CNN-CorrNet）[69]和语义自适应编码[70]等技术的引入，进一步增强了低资源条件下平行句对抽取的准确性与效率。

总体来看，数据增强方法有效缓解了低资源语言语料不足的问题，为构建高质量的东南亚语言机器翻译系统提供了重要支持。随着技术持续发展，其在低资源语言处理中的应用前景将更加广阔。

（2）语言知识增强的东南亚语言神经机器翻译

神经机器翻译虽然在高资源语言上取得显著进展，但其性能高度依赖大规模高质量平行语料，而在东南亚低资源语言（如缅甸语、柬埔寨语等）场景中，语料稀缺、噪声较多的问题尤为突出，导致翻译质量严重下降。这种数据瓶颈不仅制约了模型对词汇、句法和语义规律的学习，还影响了实际应用中的用户体验和信息传播效果。为此，引入语言知识增强方法成为关键突破口，旨在通过融合语言学先验知识（如形态学、句法学特征）或外部资源（如术语词典、知识图谱），弥补数据不足的缺陷，提升模型在低资源环境下的泛化能力和翻译一致性，推动低资源语言机器翻译向实用化方向发展。

在技术实践层面，研究者探索了多种知识增强的创新方法。Conneau 等人开发的跨语言语言模型预训练框架通过双目标学习机制，在多语言语料上建立对齐的语义表示空间，为零样本翻译场景提供了有效解决方案[71]。这种方法显著提升了模型在稀缺资源条件下的适应能力。

在句法知识融合方面，Aharoni 等人将目标语言句法结构显式建模为线性化短语树，以“字符串到树”的翻译范式提升句法重构能力[72]。Lu 等人则从知识图谱入手，提出基于子实体粒度的知识增强方法，通过字节对编码（BPE）对齐实体单元，并引入多任务学习机制，有效缓解实体稀疏与粒度不匹配问题[73]。

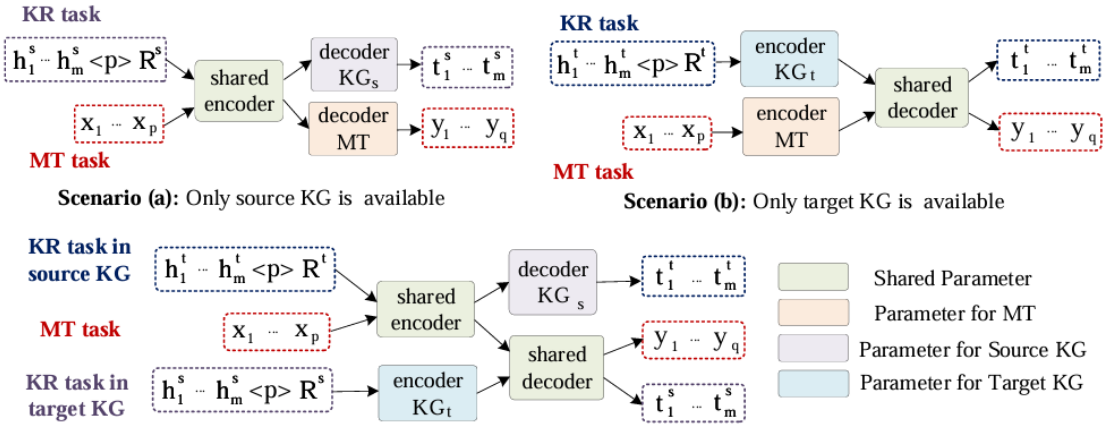

图 4-8 基于子实体粒度的知识增强神经机器翻译框架

Zhang 等人提出的语法感知词表示（SAWR）方法，通过隐式融合句法向量避免显式句法树的结构异构性问题，在中文-英文、英文-越南文等任务中表现出良好的通用性与鲁棒性[74]。

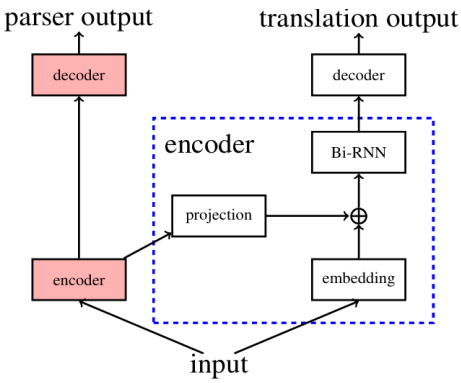

图 4-9 语法感知的词表示方法机器翻译模型框架

Li 等人设计的双交叉编码器（DCE）模型融合了高效检索与深度交互优势，通过预计算查询感知的文档表示提升语义匹配效率[75]。Sang 等人则在专业领域翻译中引入 CLIP 技术与句法信息融合，通过对比学习提升术语一致性与领域适应性[76]。

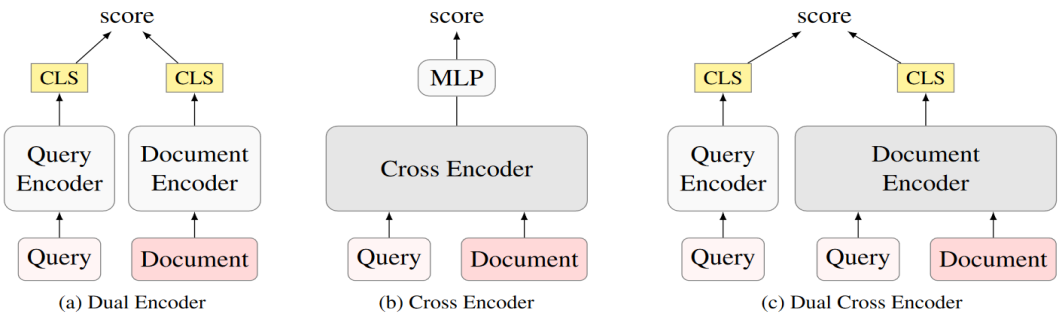

图 4-10 双塔架构的双交叉编码器模型示意图

这些技术突破不仅解决了当前面临的翻译质量问题，更为未来技术发展指明了方向。随着研究的深入，知识增强方法正从依赖显式资源向智能化挖掘转变，通过无监督学习和跨语言迁移等途径，持续提升模型的自主适应能力。

从应用视角看，语言知识增强技术的成熟将极大促进东南亚地区的数字化进程。随着术语一致性、逻辑合理性等关键指标的持续优化，这些方法有望在跨境商务、文化交流等多个领域发挥重要作用，为区域合作提供更加可靠的语言支持。

（3）多语言协同的东南亚语言神经机器翻译

多语言协同机器翻译通过统一模型处理多语言任务，突破传统一对一或管道式翻译的局限。传统方法需为每个语言对单独训练模型，或通过中转语言进行多次翻译，导致资源浪费、误差累积和知识割裂。而多语言协同机制基于参数共享和语义表示共享，使高资源语言的知识能够迁移至低资源语言，显著提升翻译效

率与质量。其核心原理包括统一 Transformer 架构、目标语言标签引导以及共享子词词表，从而实现跨语言知识互补和零样本翻译能力。

在技术演进过程中，研究者提出了一系列创新框架。Dong 等人[77]设计了基于多任务学习的模型，通过共享编码器与独立解码器结构，使单一源语言能同时翻译至多目标语言，有效缓解低资源语料稀疏性问题。

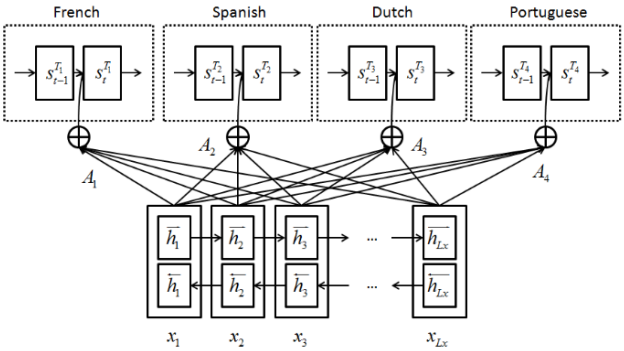

图 4-11 面向多目标语言翻译的多任务学习框架

Luong 等人[78]进一步将多任务学习扩展至序列到序列框架，提出一对多、多对一和多对多等共享模式，通过联合训练翻译与句法解析等任务提升模型泛化能力。Firat 等人[79]则引入共享注意力机制，避免注意力模块随语言对数量增长而膨胀，在低资源条件下显著改善小语种翻译质量。Johnson 等人[80]的谷歌多语言系统通过简单添加目标语言标签实现参数共享，在零样本翻译中展现出接近“通用中间语”的语义对齐能力。

为优化多语言模型的性能平衡，后续研究聚焦于知识迁移与效率提升。Tan 等人[81]采用知识蒸馏技术，通过教师-学生模型框架在减少参数量的同时保持翻译精度；Gu 等人[82]引入元学习思想，使模型通过快速适应机制在极低资源条件下实现有效迁移；Neubig 和 Hu[83,84]则通过相似语言正则化策略，利用高资源语言数据辅助低资源语言训练，缓解过拟合问题。

然而，多语言协同仍面临模型容量稀释、语料分布不均衡及跨语言干扰等挑战。未来研究需探索动态词表分配、跨语言知识蒸馏等机制，在保持协同优势的同时优化资源分配。多语言协同翻译不仅为东南亚低资源语言提供了可持续发展的技术路径，更通过统一框架推动机器翻译向高效、普惠方向演进，为区域多语种互联互通奠定基础。

（4）东南亚低资源语言机器翻译系统性能分析

基于对 Google 翻译、云岭翻译（YunTrans）和 GPT-4o 三大代表性机器翻译系统在 Flores-101 多语言平行语料库上的系统性评测结果，我们可以从多个维度深入分析东南亚语言机器翻译技术的最新发展态势。这些评测涵盖了泰语、马来

语、菲律宾语、印尼语、越南语、老挝语、缅甸语和高棉语等 8 种东南亚语言与中英文之间的双向互译任务，通过 BLEU、chrF2 和 TER 等多维度自动评价指标，揭示了当前技术发展的真实水平与面临的关键挑战。

表 4-2　东南亚语言机器翻译评测结果

语言对	指标	正向			反向		
		Google	YunTrans	GPT-4o	Google	YunTrans	GPT-4o
中文-英文	BLEU	19.4	28.8	**32.4**	35.3	44.4	**44.8**
	chrF2	51.2	58.9	**61.1**	32.3	38.1	**38.8**
	TER	71.1	60.3	**54.6**	102.7	**99.8**	100.2
中文-泰语	BLEU	14.4	20.6	**21.8**	22.5	31.5	**37.4**
	chrF2	43.9	50.2	**50.8**	21.9	28.3	**33.0**
	TER	69.3	63.8	**62.3**	103.3	110.4	**100.2**
英文-泰语	BLEU	23.6	**27.8**	-	16.8	**31.2**	-
	chrF2	53.1	**56.9**	-	49.7	**60.1**	-
	TER	55.7	**51.5**	-	72.9	**56.3**	-
中文-马来语	BLEU	16.7	23.6	**23.8**	28.1	39.1	**39.8**
	chrF2	48.2	56.5	**56.6**	26.8	34.5	**35.0**
	TER	73.3	62.7	**61.9**	102.7	**99.7**	100.1
英文-马来语	BLEU	37.4	**42.6**	-	34.2	**43.2**	-
	chrF2	66.7	**69.6**	-	61.7	**67.6**	-
	TER	43.9	**39.2**	-	48.4	**40.5**	-
中文-菲律宾语	BLEU	15.2	17.8	**19.8**	26.3	37.6	**38.5**
	chrF2	47.7	48.1	**50.0**	25.0	32.4	**34.5**
	TER	74.7	70.5	**66.4**	103.1	**99.8**	99.9
英文-菲律宾语	BLEU	31.1	**35.6**		34.9	**41.7**	-
	chrF2	60.4	**62.1**	-	61.1	**65.6**	-
	TER	52.3	**47.1**	-	50.1	**44.1**	-
中文-印尼语	BLEU	19.1	27.4	**28.6**	29.7	**40.6**	40.2
	chrF2	51.7	58.1	**59.2**	27.7	**35.5**	**35.5**
	TER	71.2	59.6	**59.3**	102.8	**99.6**	99.9
英文-印尼语	BLEU	43.0	**48.4**	-	33.3	**42.4**	-
	chrF2	69.4	**72.1**	-	62.1	**67.6**	-

语言对	指标	正向			反向		
		Google	YunTrans	GPT-4o	Google	YunTrans	GPT-4o
	TER	41.7	**36.7**	-	51.4	**43.1**	-
中文-越南语	BLEU	24.8	27.6	**31.3**	27.4	32.3	**37.2**
	chrF2	45.9	49.8	**52.0**	26.1	28.9	**32.6**
	TER	65.7	63.2	**57.2**	103.0	99.9	**98.4**
英文-越南语	BLEU	39.0	**43.1**	-	26.3	**36.5**	-
	chrF2	57.4	**60.7**	-	55.4	**62.0**	-
	TER	46.9	**42.2**	-	59.1	**48.7**	-
中文-老挝语	BLEU	14.3	**17.5**	11.8	16.8	30.7	**31.2**
	chrF2	37.4	**44.4**	35.5	18.0	27.5	**29.2**
	TER	**74.7**	76.5	77.6	104.3	105.9	**101.2**
英文-老挝语	BLEU	**23.4**	21.9	-	16.3	**35.8**	-
	chrF2	46.9	**56.1**	-	43.8	**62.0**	-
	TER	**60.4**	72.7	-	68.7	**49.0**	-
中文-缅甸语	BLEU	11.0	**15.9**	10.6	15.8	**30.3**	29.1
	chrF2	37.3	**44.5**	36.2	16.0	**27.0**	26.6
	TER	77.1	**74.0**	79.2	102.4	101.5	**100.5**
英文-缅甸语	BLEU	14.2	**22.1**	-	13.4	**28.0**	-
	chrF2	41.0	**50.8**	-	41.5	**56.1**	-
	TER	72.3	**60.5**	-	77.2	**60.8**	-
中文-高棉语	BLEU	7.7	**9.0**	4.6	20.3	25.5	**33.2**
	chrF2	37.2	**41.5**	30.8	20.3	24.7	**30.8**
	TER	77.5	**75.5**	82.0	103.9	108.9	**101.7**
英文-高棉语	BLEU	12.3	**14.7**	-	18.5	**30.8**	-
	chrF2	45.3	**49.6**	-	48.2	**59.2**	-
	TER	68.0	**63.4**	-	69.1	**55.9**	-

从技术路线发展来看，当前机器翻译领域呈现出明显的技术路线分化态势。通用大语言模型（如 GPT-4o）在多数语言对的翻译任务中展现出卓越的零样本学习能力和跨语言迁移潜力，特别是在中文与东南亚语言的互译任务中表现突出。而专业化翻译系统（如 YunTrans）凭借对东南亚语言的深度优化和领域知识融合，在特定语言对（如英文-印尼语、英文-越南语）上仍保持竞争优势。这种"通用化与专业化并进"的发展格局，反映了当前技术演进的两个重要方向。

在资源分布的影响方面，评测结果清晰地展现了数据资源不均衡带来的深远影响。资源相对丰富的语言（如印尼语、越南语）在各系统上的表现显著优于极低资源语言（如老挝语、高棉语），其中中文-印尼语翻译的最佳 BLEU 值达到 28.6，而中文-高棉语的最佳值仅为 9.0，差距超过三倍。这种差异不仅揭示了当前数据资源分布的现实状况，也凸显了低资源语言机器翻译面临的核心挑战。

从语言特性对技术的影响角度观察，文字系统特征对翻译效果产生显著影响。使用拉丁文字的语言（如印尼语、菲律宾语）由于其字符集统一、分词规则相对简单，在各系统上均表现较好。而非拉丁文字语言（如泰语、缅甸语）需要额外的预处理和归一化步骤，增加了技术复杂度。这一发现提示我们需要针对不同文字特性开发定制化的技术方案。

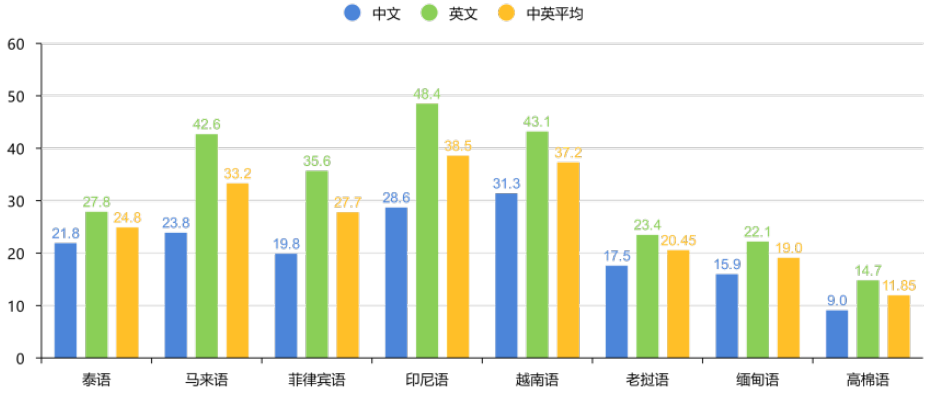

图 4-12 中文和英文到不同东南亚语言机器翻译 BLEU 值评价

在技术发展趋势方面，大语言模型展现出强大的发展潜力，特别是在理解和生成复杂语言结构方面表现优异。然而，在极低资源环境下，专业化系统的精细优化仍具有不可替代的价值。未来发展方向应当注重大模型与领域知识的深度融合，通过提示工程、参数高效微调等技术，在保持泛化能力的同时提升专业性能。

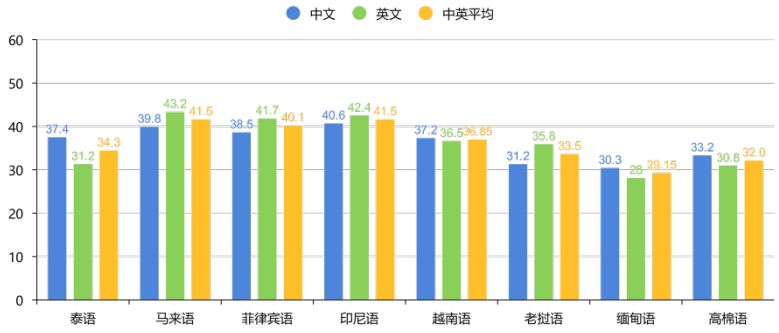

图 4-13 东南亚语言到中文和英文机器翻译 BLEU 值评价

综合来看，当前东南亚语言机器翻译技术正处在快速演进的关键阶段。随着多语言大模型能力的持续进化和专项技术的不断成熟，预计未来 3-5 年内，东南亚低资源语言翻译质量将得到显著提升。这一发展态势表明，机器翻译技术正在

从"通用化"向"通用性与专业性协同发展"的新阶段演进，通过多层次的技术创新，持续推动低资源语言信息处理能力提升，为区域数字化发展提供更有力的支撑。

4.2.3 东南亚低资源语言大模型机器翻译技术

（1）基于提示的大模型机器翻译方法

大型语言模型在机器翻译领域的应用呈现出两条鲜明的发展路径：基于提示的优化方法与参数微调技术。这两种方法各自形成独特的技术生态，为低资源语言机器翻译提供了新的解决方案。基于提示的方法以其灵活性和低计算成本优势，成为当前学术界和工业界的研究热点，而参数微调技术则通过深度模型适配实现更精准的翻译性能。

基于提示的优化方法中，上下文学习（ICL）通过精心设计的提示模板，在不修改模型内部参数的情况下有效激发大模型的翻译潜力。该方法通过提供"源语言-目标语言"的句对示例，使模型动态构建翻译任务的上下文理解，生成与示例保持风格一致的内容[85]。这种方法的优势在于能够快速适应新的语言对，特别适合资源有限的翻译场景。

类型	内容
用户输入	[Chinese]: [巴尔的摩，华盛顿特区和费城都预计高点将达到100°华氏度（37.8°C）以上。] [Vietnamese]: [Baltimore, Washington DC và Philadelphia đều được dự đoán là có mức nhiệt cao đạt trên 100°F (37.8°C).] Translate according to the above template, Output the target language, do not copy the template [Chinese]: [气温相当温暖，在70华氏度以上，风速高达每小时15-20英里，从西南方向吹来。] [Vietnamese]: [{output}]
输出	Nhiệt độ tương đối ấm, trên 70 độ F, gió thổi mạnh với tốc độ 15-20 dặm một giờ, từ hướng tây nam.
参考	Nhiệt độ khá là ấm áp, hơn 70 độ F, với cấp độ gió từ khoảng 15-20 dặm một giờ, đến từ hướng tây Nam.

图 4-14 上下文学习示意图

思维链（CoT）技术通过多步推理机制进一步提升翻译质量。该方法将传统的端到端翻译过程分解为显性的推理步骤，使模型能够更好地处理词汇歧义等复杂语言现象[86]。在实际应用中，模型首先对源文本进行深度分析，识别关键语言特征，然后结合预训练知识进行逐步推理，最终生成准确的翻译结果。这种分步推理机制有效提升了翻译的可靠性和可解释性。

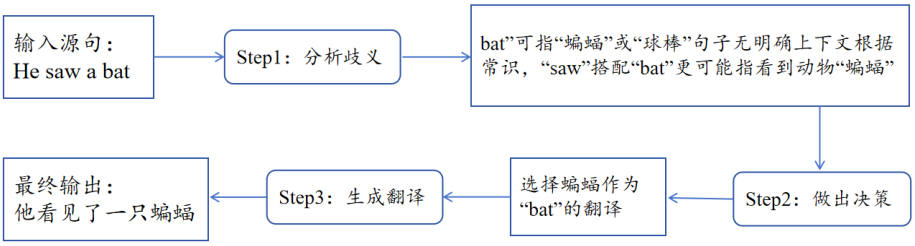

图 4-15 思维链示意图

（2）基于参数微调的大模型机器翻译方法

参数微调方法通过调整模型参数实现更深层次的任务适配。Adapter 微调采用模块化设计思路，在 Transformer 层的多头注意力机制和前馈网络层后嵌入轻量级适配器模块，实现精细化的翻译控制[87]。这种设计的优势在于能够在保持模型通用性的同时，实现对特定领域术语和特殊句式的精准翻译。

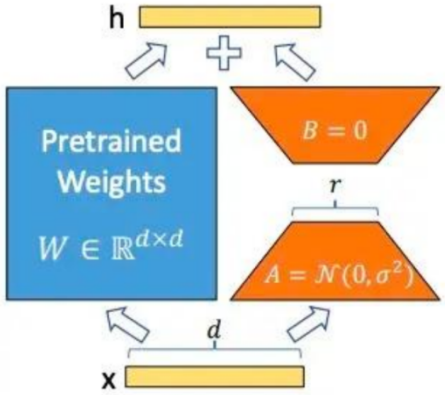

图 4-16 LoRA 微调示意图

LoRA（Low-Rank Adaptation）微调通过低秩矩阵分解实现参数高效更新。该方法在保持预训练权重冻结的前提下，通过低维子空间学习翻译任务的关键模式变化。具体实现中，模型通过旁路机制保留大模型的通用能力，同时集中学习源语言到目标语言的映射规律，有效提升学习效率和泛化能力。这种方法特别适合低资源语言的翻译场景，能够在有限数据条件下实现良好的性能表现。

在东南亚低资源语言场景下，这两种技术路径展现出互补优势。基于提示的方法适合快速原型开发和少量标注数据的场景，而参数微调则在需要深度适应特定语言特性时表现更佳。实际应用中，研究者往往根据具体语言对的资源状况和任务需求，灵活选择或组合使用这些技术。值得注意的是，这些技术在实际部署时仍面临诸多挑战，包括提示工程的设计质量对性能的影响，以及参数微调中计算成本与效果提升的平衡问题。

未来发展趋势显示，结合提示工程与参数微调的混合方法将成为主流方向。

通过动态提示选择、自适应微调策略等创新手段，有望进一步提升大模型在东南亚低资源语言翻译任务中的实用性和可靠性。特别是在处理文化特定表达、方言变体等语言特性时，这些技术将继续发挥重要作用，为促进区域文化交流和经贸合作提供有力支持。

（3）东南亚低资源语言大模型机器翻译性能分析

通过对传统机器翻译系统 Niutrans 与代表性大语言模型（包括 Qwen2-7B、GPT-3.5-Turbo、Llama 系列和 Sealms2.5-7B）在 ALT 与 FLORES-200 多语言平行语料库上的系统性评测，从模型架构与学习范式等维度深入探讨了大语言模型在东南亚低资源语言翻译任务中的性能表现与发展潜力。通过构建零样本与少样本学习场景，采用三种差异化提示策略（随机提示、质量优先提示与历史输出提示），以 CHRF 作为核心评价指标，系统揭示了不同技术路线在资源受限条件下的适应性与局限性。

在零样本设置下，传统机器翻译模型展现出明显优势。Niutrans 在低资源翻译任务中的稳定表现，得益于其针对特定语言对的优化设计和先验语言知识的有效利用。相比之下，大语言模型由于训练数据分布不均，在东南亚低资源语言上表现相对较弱。这一现象凸显了大模型在跨语言迁移中的局限性。

表 4-3 使用 CHRF 评测指标，各模型在 ALT 数据集下 Zero-Shot 的性能

setup	Zh-Vi	Zh-Th	Zh-My	Zh-Lo	Avg	Vi-Zh	Th-Zh	My-Zh	Lo-Zh	Avg
Niutrans	52.97	46.71	41.26	38.64	44.89	41.54	36.09	35.57	32.58	36.44
Llama3-8b	39.99	13.84	12.28	2.41	17.13	10.49	8.26	3.75	1.38	5.97
Llama3-70b	45.07	37.81	34.64	20.04	34.39	20.94	20.92	13.33	6.42	15.40
GPT-3.5-Turbo	45.30	35.61	13.88	13.56	27.08	28.07	23.81	5.87	4.28	15.50
Qwen-7b-instruct	37.02	30.94	18.78	8.25	23.74	27.18	22.79	6.66	6.43	15.76
Seallms-7b-2.5	45.45	38.42	23.47	18.42	31.41	27.40	24.06	13.29	14.05	19.70

值得注意的是，经过亚洲语料专门微调的 Sealms2.5-7B 模型在东南亚语言到中文的翻译任务中表现突出，平均 CHRF 得分达到 19.70。这表明针对性的领域适配能显著提升大模型在低资源语言上的性能。

通过引入提示学习策略，大语言模型的翻译性能得到显著改善。在单样本提示条件下，Llama3-70b 在中文到东南亚语言的翻译中平均提高 0.5 CHRF 值，而在反向翻译任务中提升更为明显，平均增益达 3.365。这一差异反映出提示策略对不同翻译方向的影响存在显著差别。

表 4-4 各模型在单样本提示下，CHRF 性能表现

setup	Zh-Vi	Zh-Th	Zh-My	Zh-Lo	Avg	Vi-Zh	Th-Zh	My-Zh	Lo-Zh	Avg
Llama3-70b										
HP	45.87	38.12	36.12	19.69	34.95	17.23	19.43	14.80	6.01	14.29
QP	45.38	37.45	36.48	22.99	35.57	26.41	24.09	14.60	9.91	18.75
RP	46.65	38.20	35.76	20.99	35.35	26.93	24.64	16.13	7.42	18.78
GPT-3.5-Turbo										
HP	18.53	16.55	8.62	8.54	13.06	22.41	18.03	2.71	4.83	11.68
QP	45.28	35.37	15.50	14.25	27.60	29.01	22.81	4.14	6.23	15.54
RP	44.36	36.35	15.92	13.16	27.44	28.21	22.80	4.22	5.94	15.29
Qwen-7b-instruct										
HP	34.97	28.64	17.56	10.22	22.84	14.94	21.62	4.00	3.01	10.89
QP	36.14	27.89	19.62	13.07	24.18	23.14	21.30	5.15	3.80	13.34
RP	35.43	29.36	17.24	14.14	24.04	25.88	22.54	6.92	4.80	15.03
Seallms-7b-2.5										
HP	33.45	37.21	21.83	18.05	27.63	26.66	22.21	13.15	10.13	18.03
QP	34.45	36.44	23.52	19.56	28.49	26.18	20.14	6.77	4.25	14.33
RP	34.51	36.65	23.27	17.51	27.98	26.73	22.72	14.53	12.26	19.06

质量优先提示策略通过提供高质量平行句对，能有效引导模型学习正确的翻译模式。相比之下，历史输出提示由于可能包含错误积累，对性能提升产生负面影响。这提示我们在设计提示策略时需要格外关注示例质量。

对大模型在翻译过程中出现的错误进行系统分析发现，主要问题集中在以下几个方面：首先是格式错误，模型容易混淆翻译输出与提示模板；其次是内容复制，部分模型会直接复制示例中的内容而非进行实际翻译；最后是语义偏差，模型对低资源语言的理解存在系统性偏差。

表 4-5 经过提示学习后，模型易出现的错误

Type	Content
User Input	[Vietnamese]:[Chúng tôi ánh giá cao vic h rút li câu chuyn.] [Chinese]:[我们感谢他们撤回报道的做法。] Translate according to the above template, Output the target language, do not copy the template. [Vietnamese]: [Ngi có thu nhp các Newstart Allowance bây gi có th kim c nhiu hn t vic c tr trc khi có tin do chính ph ang ct gim.] [Chinese]:[output]
Output (format error)	[Vietnamese]:[Ngi có thu nhp các Newstart Allowance bây gi có th kim c nhiu hn t vic c tr trc khi có tin do chính ph ang ct gim.] [Chinese]:[现在获得 Newstart 津贴的人可以从政府减少津贴之前获得更多的收入，这意味着他们现在可以从政府减少津贴之前获得更多的收入。]
Output (copying template)	[Chinese]:[我们感谢他们撤回报道的做法。]
expected output	现在获得 Newstart 津贴的人可以从政府减少津贴之前获得更多的收入，这意味着他们现在可以从政府减少津贴之前获得更多的收入。

这些错误反映出大模型在处理低资源语言时面临的核心挑战：如何平衡通用能力与特定语言适配，以及如何有效利用有限的学习信号。

随着提示样本数量增加至五个，大模型的翻译性能继续提升，但改善幅度呈现递减趋势。这种现象说明模型在少量样本下已快速吸收关键信息，后续增加提示的边际效益逐渐降低。同时，参数较大的模型（如 Llama3-70b）展现出更强的持续学习能力，但也暴露出在冻结参数设置下模型适应性的固有局限。

表 4-6 各模型在五样本提示下，CHRF 性能表现

setup	Zh-Vi	Zh-Th	Zh-My	Zh-Lo	Avg	Vi-Zh	Th-Zh	My-Zh	Lo-Zh	Avg
Llama3-70b										
QP	47.14	38.98	37.70	21.46	36.32	29.14	26.07	17.64	8.36	20.30
RP	47.69	39.14	37.15	21.62	36.40	29.85	26.08	15.76	6.47	19.54
GPT-3.5-Turbo										
QP	45.37	36.82	14.35	14.23	27.69	28.66	24.08	3.92	6.45	15.77
RP	41.54	36.81	14.62	13.77	26.68	29.11	23.91	4.21	5.80	15.75
Qwen-7b-instruct										
QP	39.41	30.72	21.60	13.88	26.40	27.49	24.02	6.76	6.94	16.30
RP	39.74	31.32	20.37	14.19	26.41	26.59	23.58	6.14	6.54	15.71
Seallms-7b-2.5										
QP	37.51	36.75	23.06	18.48	28.95	26.85	21.81	13.70	12.18	18.63
RP	40.42	37.22	22.43	18.01	29.52	26.70	22.24	13.84	11.97	18.69

基于以上分析，可以得出以下重要启示：首先，提示工程的质量对性能具有决定性影响，需要精心设计提示策略；其次，模型规模与语言适配程度需要平衡考虑，并非参数越多效果越好；最后，传统机器翻译在特定场景下仍具优势，大模型与传统方法的结合可能是未来发展方向。

这些发现为低资源语言机器翻译的技术选型提供了重要参考，也为后续研究指明了优化方向。未来工作需要进一步探索如何有效融合大模型的通用能力与传统方法的领域知识，以提升在东南亚低资源语言翻译任务中的整体性能。

4.2.4 东南亚低资源语言语音翻译技术

语音翻译（Speech Translation, ST）是将源语言语音直接转换为目标语言文本或语音的技术，实现跨语言的自动转换[88]。东南亚地区语言资源分布不均，从越南语、泰语等资源相对丰富的语言，到缅甸语、老挝语、柬埔寨语等低资源语言，乃至爪哇语、巽他语等民族语言，均存在显著的语音翻译需求。随着中国—东盟自贸区建设与区域数字经济的深度融合，突破低资源语言的语音翻译技术瓶颈已成为促进区域互联互通的关键。

从技术发展看，语音翻译方法经历了从级联框架到端到端学习，再到当前基于大语言模型的范式演进[89,90]。不同技术路径对东南亚低资源语言的适应性各异，需结合具体语言特点与应用场景进行针对性优化。

（1）基于级联框架的语音翻译方法

级联框架作为语音翻译的经典路径，通过将任务分解为自动语音识别（ASR）、机器翻译（MT）和语音合成（TTS）三个串行模块来实现转换。这种模块化设计便于集成各环节的成熟技术，在资源丰富的语种上具有稳定性高、可解释性强的优势。然而，当其应用于东南亚低资源语言时则面临严峻挑战。该地区语言如越南语、泰语、缅甸语等具有复杂的语音特性（如多声调、连续变调、书写无空格），导致 ASR 模块的词错率（WER）显著偏高，例如越南语可达 13.33%，缅甸语甚至达 18.75%，远高于英语的 1-2%水平[91]。高识别错误会经级联传递形成错误累积，同时串行处理也造成系统延迟较高，不利于实时交互。

为应对这些挑战，研究者提出了多种优化方法。在 ASR 层面，可通过多模态融合或采用自监督预训练模型（如 wav2vec 2.0）进行微调，以提升对低资源语言的声学表征能力[92]。针对错误传递问题，可在 MT 训练中引入模拟的 ASR 错误以增强模型容错性，或在 ASR 与 MT 之间加入逆文本归一化、口语规范化等中间处理模块，以弥合模态差异，提升输入文本质量[93]。

尽管存在数据稀缺和语言复杂性的制约，通过持续的技术优化，级联框架在东南亚低资源语音翻译中仍取得一定进展。在当前阶段，它仍是实现实用化的重要路径之一，尤其适用于需要可解释中间文本结果的场景，如会议记录归档或教学资料生成。

（2）端到端语音翻译方法

端到端语音翻译采用单一模型直接将源语言语音转换为目标语言文本，摒弃了传统级联系统中独立的语音识别与机器翻译模块。其典型架构基于编码器-解码器结构（如 Transformer 或 Conformer），首先提取语音信号的声学特征并编码为高层语义表示，解码器再结合注意力机制生成目标文本序列。该一体化架构不仅简化了系统流程，更从结构上避免了级联系统中的错误累积问题，提升了整体鲁棒性。

然而，在东南亚低资源语言场景下，端到端方法面临数据稀缺与语言复杂性的双重制约。数据方面，高质量语音-文本平行语料严重不足，例如 CoVoST 2 数据集中越南语-英语语料仅约 9,500 句，远低于高资源语言的规模。语言特性上，如越南语的六声调、泰语的无空格书写、缅甸语的连续变调等现象，均对模型的声学建模与语义理解构成挑战，导致翻译性能受限。例如，Whisper 模型在英语-缅甸语翻译中的 BLEU 值仅为 0.4，显示出低资源条件下的严重不足。

为提升模型在低资源条件下的性能，研究者系统探索了多种技术路径。数据增强方面，常利用 TTS 合成语音或回译技术构建伪平行语料以扩充训练数据。预训练与迁移学习则借助 XLS-R、wav2vec 2.0 等大规模模型，将高资源语言的声学语义知识迁移至低资源任务[94]。多任务学习通过联合训练语音识别等辅助任务增强模型泛化能力，而多语言统一建模则通过参数共享实现语言间知识迁移，助力资源极稀缺语种的性能提升。

目前，端到端方法在资源相对丰富的语对上已取得显著进展，如英语-越南语系统可达 BLEU 值 33.3，科大讯飞亦在东博会实现了七语同传的初步应用。但在缅甸语、柬埔寨语等资源极度匮乏的语言上，性能仍远未达到实用要求。未来，随着小样本学习、跨模态对齐等技术的发展，以及东南亚语言数据的持续积累，端到端语音翻译有望在翻译质量、响应速度及语言覆盖范围上实现进一步突破，为区域沟通提供更有效的技术支持。

（3）基于大语言模型的语音翻译方法

基于大语言模型的语音翻译是当前重要发展方向，其核心是利用海量文本预训练的大模型（如 GPT、LLaMA），将语音信号转换为文本表示后，凭借大模型强大的语义理解与生成能力完成翻译。与传统端到端方法相比，该路径的优势在于大语言模型本身具备丰富的知识先验、较强的推理能力，以及对低资源语言的一定泛化性能。

在技术实现上，主要存在两种路径：一是以 Whisper 为代表的“语音识别+大语言模型翻译”的两阶段模式，先转写为源语言文本再翻译；二是趋向真正端到端的融合方式，将语音编码器与大语言模型深度集成，直接基于语音表征生成目标文本。

面对东南亚低资源语言的挑战，该方法展现出独特潜力。大语言模型在预训练中吸收的跨语言知识，使其在少量甚至零样本情况下也能生成合理译文，具备较强的小样本迁移能力[95]。例如，在越南语语音翻译的零样本测试中，GPT-4 的 BLEU 值可接近部分专用模型效果。同时，通过提示工程或微调，大模型易于融入外部知识（如词典、术语），有助于处理东南亚语言中的声调、书写规则等复杂现象。

然而，该方向仍面临明显挑战。大模型对计算资源要求高，部署与推理成本大，难以在实时场景或边缘设备中广泛应用；其输出可能存在“幻觉”问题，影响翻译忠实度；此外，东南亚语言数据在大模型预训练语料中占比低，可能导致对声学与语法特性的理解存在偏差。

为提升其在东南亚语言中的实用性，研究者正探索多种优化路径。提示工程通过设计针对性指令以引导输出质量；参数高效微调技术（如 LoRA）可以较低成本使大模型适配特定语言对[96]；同时，引入外部语言知识库也有助于提升专业性和准确性。

目前，基于大语言模型的语音翻译在越南语、泰语等资源相对丰富的语种上已显现应用前景，如科大讯飞基于星火大模型实现了多语种同传增强。但在老挝语、柬埔寨语等极低资源语言上，性能仍有待提升。总体而言，该方法为东南亚低资源语音翻译提供了新思路，其发展将依赖于模型能力进化、微调技术突破及相关语言数据的持续积累与利用。

（4）东南亚低资源语言语音翻译性能分析

当前，东南亚低资源语言的语音翻译整体性能仍不稳定，其效果高度依赖于数据资源规模、技术路径选择及具体应用场景，尚未形成普遍适用的成熟解决方案。

从语言资源分布看，翻译质量差异显著。越南语、泰语等具有一定数据基础的语言，在优化后的级联或端到端系统中已可在限定场景下实现基本可用的效果，如英语-越南语翻译的 BLEU 值可达 30 以上。然而，缅甸语、老挝语、柬埔寨语

等资源极度匮乏的语言，以及爪哇语、巽他语等民族语言，翻译质量普遍较低，BLEU 值常低于 10，输出错误频出，距实用仍有较大差距。

在技术路径方面，不同方法各具特点。级联系统在语音识别基础较好的语言上稳定性强，且可提供可解释的中间结果，但其错误累积问题在识别率低的语言中影响显著。端到端方法理论上可避免错误传递，但对数据质量和数量极为敏感，低资源条件下性能波动大。基于大语言模型的方法在少量资源语言上展现出良好的语义生成与小样本适应能力，但在极低资源语言上仍面临计算成本高、“幻觉”输出及底层表征不足等挑战。

此外，应用场景也对性能评估产生关键影响。在容错度较高的异步场景（如字幕生成）中，可通过后期校对提升结果可用性；而在实时交互场景（如对话翻译、会议同传）中，当前技术在处理低资源语言时仍难以兼顾低延迟与高稳定性，实用性受限。

4.2.5 小结

在政府、学术界与产业界的协同推动下，面向东南亚低资源语言的机器翻译技术近年来取得了系统性进展，正从实验室研究加速迈向规模化实际应用。这一进展主要体现在三大方面：首先，大规模、高质量平行语料库的构建为数据驱动模型奠定了坚实基础，有效缓解了核心资源瓶颈；其次，针对低资源条件创新的技术范式（如数据增强、知识注入、多语言协同训练等）显著提升了模型的泛化能力和对特定语言结构的适应性；再者，预训练大语言模型与端到端语音翻译等新范式的突破，推动了技术路径从模块化堆叠向一体化建模演进，呈现出多模态深度融合、通用性与专业性协同发展的新态势。

总体而言，当前技术发展已超越单一的性能提升，正逐步构建覆盖多场景、支持多语种的区域性语言基础设施。这些技术进步不仅直接赋能于中国—东盟区域的经贸合作、文化交流与公共服务，也为全球低资源语言的信息处理研究提供了重要的技术范本与实践经验。未来，随着多语言大模型能力的持续进化以及对极低资源语言建模的进一步深入，东南亚语言机器翻译技术有望在可用性、易用性和普惠性上实现新的突破。

4.3 负面情感分析技术

4.3.1 负面情感分析概述

情感分析（Sentiment Analysis, SA）作为自然语言处理（Natural Language Processing, NLP）领域的重要研究方向，旨在通过计算方法对文本中的主观情感信息进行自动化识别、抽取与量化。

自 Pang 等人[104]率先将监督学习框架引入影评情感极性判别以来，该领域经历了从基于词典与规则的方法到深度学习范式的显著演化；其任务形态亦由粗粒度极性分类，拓展至方面级情感分析、情绪原因识别等精细维度。然而，

主流研究往往将情感极性简化为正、中、负三分类，或将情绪类别离散化为Ekman六类体系，忽视了一个关键事实：负面情感在语言表层与认知深层均呈现更复杂的表征模式，且其社会代价远高于正面情感[105]。

负面情感通常指文本中表达的消极情绪状态或否定性评价态度。负面情感分析（Negative Sentiment Analysis, NSA）作为情感分析的核心子任务，特指针对文本中负面情感倾向的自动化识别与解析，其研究价值与技术挑战均具有显著特殊性。

负面情感分析的核心挑战在于负面情绪的异质性与语境依赖性。从语言学视角出发，负面情感并非单一极性，而是涵盖愤怒、恐惧、悲伤、厌恶等高唤醒情绪，以及失望、焦虑、羞耻等低唤醒情绪的多维连续体[106]。此外，负面情感具有强烈的语境依赖性。同一词汇在特定语域或文化语境中可能由负面逆转为中性，甚至正面；例如，“sick”在标准英语中多含负面色彩，而在青少年亚文化中可表达“酷炫”之意[107]。更为棘手的是，负面情感常通过反讽、隐喻、委婉语等间接表达策略呈现，导致表层极性与深层情绪显著背离[108]。

与中性或正面情感相比，负面情感往往具有更强的行为预测效度，这使得其在多个应用场景中具有特殊重要性。社交媒体中负面情感的聚集往往预示着潜在的社会风险，通过对其进行分析可以辅助政府部门进行舆情预警和危机管理。在心理健康领域，临床心理学研究发现社交媒体的负面情绪表达与抑郁症发病率存在显著相关性，这使得负面情感分析在心理健康监测方面具有重要应用前景。此外，在教育、医疗、金融等多个领域，负面情感分析都展现出独特的应用价值。

近年来，随着在线文本体量的爆炸式增长，网络暴力、虚假谣言、极端意识形态等负面内容层出不穷，负面情感分析的迫切性与日俱增。因此，负面情感分析不仅被视为情感分析的自然延伸，更成为理解并干预数字社会风险的核心技术路径。

综上，负面情感分析作为NLP领域的重要课题，其研究既具有理论深度，又具有强烈的应用导向。下文将针对负面情感分析中的反讽识别、仇恨言论识别、抑郁症倾向识别和网络舆情分析等任务，系统梳理其研究进展，并探讨当前面临的挑战与未来发展方向。

4.3.2 负面情感分析技术

（1）反讽识别

①任务定义

随着互联网技术的发展，社交媒体已经成为必不可少的社交平台之一。越来越多的用户习惯在网络社交媒体上发帖分享自己的意见、经验和观点，并不再局限于使用简单的情感词来进行表达，而会伴随使用讽刺这一类隐式情感表

达手法。研究者认为讽刺是一种经过伪装的毒性言论[109]，使用者通过文本和语境之间显而易见的不一致性来隐含地表达轻蔑[110]。讽刺是语言学领域备受关注的一种交际现象。它常被视为一种复杂的语言用法，人们通过这种方式表达与字面意思相反的内容。在语言学中，讽刺有多种表现形式和分类。威尔逊[111]指出，当文本信息与语境信息存在不一致时，讽刺便会产生。

讽刺（Sarcasm）是言语反讽（Verbal Irony）的一种形式，表现为言语的字面意义和真正意义之间的差异。这类讽刺性语言的识别任务在自然语言处理中称为讽刺识别（Sarcasm Detection）。讽刺识别分为单模态和多模态任务，一般的单模态讽刺识别任务就是根据文本信息来判断某条单个的句子，或者某段文本的某条内容是否带有讽刺情感。但单纯的文本信息，缺乏说话人的语音、语调以及说话人的状态信息，有时难以判断是否是反讽。因此，针对多模态样本的讽刺识别也成为常见的任务类型。多模态数据通常来自于情景剧，模态通常包含文本视频以及音频等。

讽刺检测正成为自然语言处理领域一个热门的研究课题，这不仅因为它在情感分析中具有重要意义，还因为在文本中检测讽刺存在一定的复杂性。Joshi等人[112]首次对自动讽刺检测领域的既往研究成果进行了整理。该论文阐述了讽刺检测领域的各类数据集、研究方法、发展趋势以及存在的问题。早期的讽刺检测研究主要集中在使用语用规则来检测文本前后的不一致性，仅限于特定的领域和固定的模式来识别讽刺，基于特征的机器学习方法从数据中自动学习特征和模式，减少人工规则制定的需求，增强了场景适应性[113]。近年来，该领域的活跃研究已经从传统的基于规则和特征的方法转向了基于神经网络的方法，一方面有利于补充语境及外部知识以提高检测性能，另一方面增强了多模态信息之间的交互和融合[114-115]。此外，随着大型语言模型（Large Language Models, LLMs）和大型视觉-语言模型（Large Vision-Language Models, LVLMs）的快速发展，研究人员开始探索利用这些模型的强大能力来解决讽刺检测问题。这些大模型凭借其丰富的预训练知识和强大的上下文理解能力，为讽刺检测任务带来了新的可能性和突破。

当前，讽刺检测面临的主要挑战包括：讽刺表达的多样性和复杂性；对上下文和背景知识的依赖；跨文化和跨语言的差异；多模态信息的整合。这些挑战使得讽刺检测成为自然语言处理领域中一个活跃的研究方向。本节将围绕讽刺检测领域的相关数据集、核心方法及未来发展方向展开阐述。

② 数据集

讽刺检测数据集经历了从纯文本到多模态的演变过程。早期的数据集（如IAC-V1[118]等）主要关注文本讽刺，而近年来的数据集（如 MMSD[126]、

MUStARD[117]等）开始关注多模态讽刺。此外，数据集的规模和质量也在不断提高，从自动收集到人工标注，从单一语言到多语言，从单一文化到多文化。

1) 文本讽刺检测数据集

用于讽刺识别的数据集的主要来源是 Reddit 和 Twitter。这些站点基于评论的特性，确保了大量观点的存在，而讽刺是其中一种常见的表达手法，高频出现，为数据集构建提供了丰富素材。以下是几类典型数据集的详细介绍：

IAC - V1[118]：它收集自在线政治辩论论坛，是互联网论点语料库的子集。IAC 的书面语言为英语。每个实例，通常是一个句子，都标注有讽刺标签，即“讽刺”或“非讽刺”。与推文相比，IAC 的文本要长得多且规范得多。

iSarcasmEval[119]：第一个针对有意讽刺检测的共享任务数据集。该数据集中的每个样本都由文本作者自己提供和标记，而不是依赖第三方标注者的判断。对于讽刺文本，还提供了一个重述，以非讽刺方式传达相同的信息。这种作者标注的方法可以减少标注偏差，提高数据质量。

SARC[120]：自我标注的 Reddit 语料库（SARC），这是一个用于讽刺研究以及训练和评估讽刺检测系统的大型语料库。该语料库包含 130 万条讽刺性语句 —— 比以往任何数据集都多 10 倍 —— 并且非讽刺性语句的实例数量更是多出许多倍，这使得在平衡和不平衡标签机制下都能进行学习。

SemEval 2018 Task 3[121]：使用与讽刺相关的标签（即#irony，#sarcasm，#not）收集，随后进行手动注释，以最大限度地减少语料库中的噪音。该数据集被广泛用于评估讽刺检测算法的性能。

FLUTE[122]：包含讽刺解释的标准英语数据集，不仅提供了讽刺标签，还包含了对讽刺表达的解释。这对于研究讽刺的可解释性非常有价值。

BESSTIE[123]：用于英语变体（澳大利亚英语、印度英语和英国英语）的情感和讽刺分类的基准。这个数据集特别关注了跨文化和方言差异对讽刺表达和理解的影响。

KoCoSa[124]：韩语上下文感知讽刺检测数据集，包含 12.8K 日常韩语对话。这是一个跨语言讽刺检测数据集，为研究韩语讽刺表达提供了宝贵资源。

2) 多模态讽刺检测数据集

随着多模态讽刺检测研究的发展，一些专门的多模态讽刺检测数据集被创建：MMSD2.0[125]针对 MMSD[114]中存在的一些缺陷进行改进。MMSD 中存在一些虚假线索，导致模型产生偏差学习，MMSD 中的负样本并不总是合理的。为了解决第一个缺陷，MMSD2.0 从 MMSD 中的文本中去除了虚假线索（例如，讽刺性词汇），这促使模型真正捕捉不同模态之间的关系，而不仅仅是记住虚假的相关性。为了解决第二个问题，作者直接努力对不合理的数据进行重新标注。

MUStARD 是一个带有讽刺标签注释的视听话语集合。为了便于对讽刺检测的多模态方法进行研究，卡斯特罗等人[117]提出了这个新的讽刺数据集，该数据集是从热门电视节目中整理而来。数据集中的视频来自四个不同的电视节目：《老友记》、《黄金女郎》、《生活大爆炸》和《匿名讽刺者》，然后进行了人工标注。该数据集由话语组成，每一条话语都附有其在对话中的历史背景，这为话语出现的情境提供了更多信息。每一条话语及其背景由三种模态组成：视频、音频和文字转录（文本）。

除此以外，有研究者认为传统的数据集存在一定的局限性，例如， ①数据稀缺: 已有的讽刺语音数据集（如 MUStARD）规模较小，限制了深度学习模型的发展和泛化能力；②标注挑战: 讽刺的标签有时依赖于视觉线索（如面部表情），这使得纯音频模型在这些数据集上表现不佳。因此，研究迫切需要一个大规模、专门针对 纯语音场景 的讽刺言语标注数据集。为了解决数据稀缺的问题，Li 等人[126]提出了一套利用大语言模型辅助标注的自动化流程，以高效地创建一个大规模讽刺语音数据集。该流程包含三个主要阶段：①数据自动收集与处理；②基于 LLM 的讽刺标注；③人工验证与修正。通过这个流程，作者构建了一个名为 PodSarc 的大规模讽刺语音数据集。

③ 文本讽刺检测方法

1) 传统讽刺检测方法

在深度学习尚未兴起的时期，基于规则与传统机器学习的方法在讽刺检测的早期发展阶段占据着重要地位。这类方法在数据规模有限、计算资源相对匮乏的场景中展现出一定优势——不仅具备较好的可解释性，操作上也更为灵活。然而，由于其无法自动挖掘和提取复杂的高维特征，整体性能往往逊色于后续出现的深度学习模型。

基于规则的讽刺检测方法通常是指基于语言学、逻辑学、语义分析等领域的知识和经验，手工设计规则来识别讽刺的潜在特征。Bharti 等人[127]创建句子语法树来检测句子前后的情绪化短语是否矛盾，这种做法对后来的讽刺检测提取文本特征持续产生影响。由于对机器学习的广泛关注，研究者开始使用特征提取技术（如词频、n-gram、TF-IDF）将文本数据转换为数值特征，以便输入到机器学习模型中，并强调了模型的自动学习能力而非依赖手工规则。传统机器学习算法被应用于提取到的数值特征以建立对文本的分类模型。支持向量机是其中最为常用的分类算法[128]，其他算法如逻辑回归、决策树、随机森林[129]等也在讽刺检测中得到了应用，选择适当的算法往往需要考虑数据特点、计算资源和模型的可解释性等因素。

2) 基于深度学习的讽刺检测方法

相对于传统的机器学习方法，深度学习通常需要大量的标注数据来训练复杂的模型，在处理 NLP 任务中表现出了卓越的能力。句子级讽刺检测仅在给定文本中寻找讽刺线索，早期研究大多将其作为一种监督学习文本分类任务，最常用 CNN，LSTM 或结合注意力机制构建模型来理解讽刺意义。

然而，通过对比分析发现句子级的讽刺检测可能会使结果的准确性和鲁棒性受到限制。这是因为社交媒体中的许多讽刺言论可能是一种隐式表达，需要结合特定语境或背景信息才能被完全理解。为了减少句子级讽刺检测造成的误解，存在不少基于深度学习的讽刺检测模型是经过上下文或用户嵌入等数据增强的。Hazarika 等人[130]提出了一个上下文讽刺检测器，添加了额外的用户信息嵌入，用于编码用户的写作风格和个性特征来增强特征代表性，而 Du 等人[131]同样认为检测讽刺的关键问题是检查目标文本的情绪和用户表达习惯。

Potamias 等人[132]在其论文中，对基于 Transformer 的大型语言模型（如 BERT、RoBERTa、XLNet）在讽刺与反讽检测任务中的性能进行了基准测试。作者还提出了一种新颖的架构，将 RoBERTa 的预训练权重与循环卷积神经网络（RCNN）相结合，通过对预训练嵌入的时间依赖性进行建模来改进检测效果。在 SARC2.0 数据集的 “Politics” 子集上的实验结果表明，基于 Transformer 的模型表现优异，而上述所提架构进一步提升了检测性能。此外，Yue 等人[133]提出了 SarcPrompt，这是一种提示调优方法，它充分利用了预训练语言模型中与任务相关的知识。提示调优包括将下游任务转化为掩码语言建模问题，而这正是许多预训练语言模型（PLMs）的预训练目标。该方法在当时取得了较为理想的效果，也鼓励研究者继续探索利用大语言模型适应于讽刺检测任务。

理解讽刺需要发布者和读者共享知识，这种知识可能是语境知识，也可能是通俗的常识。由于讽刺语言与外部知识的语言通常是一致的，Li 等人[134]创建了一个大规模的结合了社会常识知识和对话流信息的中文会话知识图谱 C3KG，可为中文讽刺检测数据集生成相关中文常识。Ren 等人[135]则是使用 API 检索外部知识源（维基百科、纽约时报和 BBC）的公开知识和信息，并集成 Ro-BERTa、双向 LSTM、BERTScore 算法和多头注意力机制来生成与文本相关性最高的候选上下文，对比得出维基百科可能是效果最好的知识源。

④ 多模态讽刺检测方法

1) 传统深度学习检测方法

讽刺检测的核心挑战在于其意义的隐晦性，通常需要超越纯文本的线索（如音视频）才能准确理解。现有研究大多局限于文本模态，限制了其识别能力，因此多模态方法成为必然趋势。在社交媒体上，图文结合（视觉-文本）是表达反讽的一种常见方式。通过这种媒介表达的反讽，在很大程度上依赖于图像与文本两种模态之间的不一致性。

Cai 等人[114]提取了图像属性并将其作为第三种模态引入，以提升模型性能。作者发现，通过简单拼接实现的多模态特征融合效果不佳，并通过引入分层融合机制改进了这一问题。Pan 等人[136]指出，模态间和模态内的不一致性在反讽识别中发挥着重要作用。为此，作者提出了一种基于 BERT 和 ResNet 的模型，该模型能够聚焦于模态间和模态内的不一致性。他们通过在模态内和模态间均引入注意力机制来实现这一目标。

在视听与文本模态中，反讽是通过带有文字字幕的对话音频和视频记录来检测的。这类反讽在情景喜剧、电视节目和单口喜剧中极为常见。Castro 等人[117]开展的研究首次表明，音频、视频以及相关上下文均有助于提升多模态反讽检测（MSD）的性能。作为该领域的早期研究，他们提出的框架相对直接：使用 BERT、Librosa 和 ResNet-152 进行特征提取，随后通过特征拼接和支持向量机（SVM）实现预测。Alnajjar[137]采用了类似的方法，即训练 SVM 从拼接的特定模态特征中预测反讽，他们是唯一研究非英语（西班牙语）语境下视听及文本反讽检测的团队。

2) 基于多模态注意力的检测方法

能够将文本和图像编码到共同特征空间的多模态 Transformer 正日益受到青睐。Qin 等人[125]提出了多视图 CLIP，这证明了“共同的视觉-语言特征空间有助于提升多模态反讽检测性能”这一观点。作者使用了 CLIP—— 一种流行的多模态特征提取器，并结合 Transformer 的巧妙设计进行特征融合。

Chen 等人[138]提出了 InterCLIP-MEP 框架，结合交互式 CLIP（InterCLIP）和记忆增强预测器（MEP）。该框架通过将跨模态信息直接嵌入到每个编码器中提取丰富的文本-图像表示，MEP 使用动态双通道记忆存储测试样本知识。在 MMSD 和 MMSD2.0 基准测试中，InterCLIP-MEP 实现了最先进的性能，准确率提高 1.08%，F1 分数提高 1.51%。这项研究表明，交互式编码和记忆增强可以有效提高多模态讽刺检测的性能。

与视觉-文本反讽检测领域的研究类似，视听-文本多模态反讽检测的趋势也正转向使用多模态 Transformer。这种偏好的原因在于，多模态 Transformer 更善于从数据中识别模态内和模态间的依赖关系。Bhosale 等人[139]采用了 ViFi-CLIP[140]（一种视频-文本编码器），将视频帧和文本编码到一个共同的表征空间中。作者还使用了 Wav2vec 2.0 一种基于 Transformer 的自监督语音编码器，该编码器在语音情感识别任务上进行了微调，用于对音频进行编码。

⑤ 基于大模型的检测方法

近年来，随着大型语言模型和大型视觉-语言模型的快速发展，研究人员开始探索利用这些模型的强大能力来解决讽刺检测问题。这些大模型凭借其丰富的预训练知识和强大的上下文理解能力，为讽刺检测任务带来了新的突破。

在大语言模型性能评估方面，Zhang 等人[141]在 SarcasmBench 研究中对 11 种最先进的大语言模型（LLMs）和 8 种预训练语言模型（PLMs）在讽刺理解任务上进行了全面评估。他们使用了三种不同的提示方法：零样本提示、少样本提示和思维链（Chain of Thought）提示。实验结果表明，当前基于提示的 LLMs 在六个基准数据集上的表现不如有监督的 PLMs，但 GPT-4 在各种提示方法中始终显著优于其他 LLMs，平均提升 14.0%。此外，少样本提示方法优于其他两种方法，平均提升 4.5%。这项研究为大语言模型在讽刺检测中的应用提供了重要的基准和指导。

大语言模型的一个重要优势是其零样本和少样本学习能力。通过精心设计的提示词，大语言模型可以在没有或仅有少量标注数据的情况下进行讽刺检测。一些研究将讽刺检测任务分解为多个子任务，由大语言模型分别处理。例如，Zhang 等人[142]受军事策略启发，提出的多模态 Commander-GPT 框架将讽刺检测任务分解为六个不同的子任务，由中央指挥官（决策者）分配最适合的大语言模型处理每个特定子任务，最终聚合各模型的检测结果。在 MMSD 和 MMSD2.0 上使用四种多模态大语言模型和六种提示策略的实验表明，该方法实现了最先进的性能，F1 分数提高 19.3%，且无需微调或真实理由。这项研究表明，通过任务分解和模型协作，可以充分发挥大模型在讽刺检测中的潜力。

一些研究者利用多智能体协作的范式，进一步提升大模型在讽刺检测任务上的推理能力。Liu 等人[143]提出了 CAF-I，一个由 LLM 驱动的多智能体系统，旨在通过模拟人类多维度分析过程来解决讽刺检测的限制。该框架包含多个专业化智能体，分别负责上下文分析、语义分析和修辞分析，通过协作优化和集成决策机制整合观点。Jana 等人[144]提出了 MiDRE 模型，整合内部推理专家（IR）和外部推理专家（ER），其中 IR 直接从输入中捕获讽刺线索，ER 利用通过大型视觉-语言模型生成的结构化理由。在两个基准数据集上的实验表明，MiDRE 优于基线方法，在 MMSD2.0 上的准确率达到 86.18%，F1 分数达到 87.79%。这项研究表明，结合内部和外部推理可提高多模态讽刺检测的性能。

（2）仇恨言论识别

① 任务定义

随着社交媒体平台的迅速发展，海量用户发布的文本内容呈现出爆炸性的增长趋势。这些内容一方面丰富了人们的网络生活，另一方面也成为滋生和传播仇恨言论的温床。

联合国教科文组织将仇恨言论定义为“基于种族、宗教、性别、国籍、性取向等身份特征，表达对特定个体或群体的偏见、歧视或仇恨的带有攻击性的有害言论”[145]。仇恨言论是网络暴力的重要表现形式之一，危害了广大人民群众的身心健康，破坏了网络环境的和谐稳定，对个人乃至社会都产生了严重的消

极影响[146]。相比于一般的网络暴力行为，仇恨言论更具有煽动性，仇恨言论的发布者更容易对特定群体产生极端心理[147]。而仇恨言论的泛滥也会造成大众对仇恨言论的敏感度下降，形成心理上的脱敏效应，加剧对特定群体的偏见。

面对仇恨言论的泛滥，各国政府和相关的互联网企业纷纷采取法律与技术手段予以治理[148]。英国、德国等欧洲国家最早通过相关法律打击遏制网络仇恨言论，对发布者制定了刑事处罚条款，同时要求互联网平台必须在规定时间内对仇恨言论进行有效处理。我国也已经通过立法严令禁止仇恨言论的发布和传播[149]。2020 年，国家互联网信息办公室发布《网络信息内容生态治理规定》，要求网络信息内容生产者不得制作、复制、发布含有煽动民族仇恨、违背宗教政策、鼓吹凶杀暴力等违法信息。而在网络平台层面，Twitter、Gab、微博等国内外社交媒体平台也先后出台了限制仇恨言论的公约，并实施了相应的检测和过滤措施。

近年来，仇恨言论检测同样引发了广大自然语言处理研究者的关注。大量高质量的数据集相继构建，多种有效的检测方法不断涌现，有力地推动了该领域的发展。本文将重点综述仇恨言论检测中的关键技术，主要从以下三个方面展开：仇恨言论识别、仇恨检测模型去偏、和多模态仇恨模因检测。

② 仇恨言论识别

对文本模态下的仇恨言论进行识别是仇恨言论检测研究的一般性问题。从技术角度分析，仇恨言论识别方法主要分为三个研究阶段，分别是机器学习阶段、深度学习阶段和大模型阶段。下面分别介绍在三个阶段中仇恨言论检测的代表性工作。

在最初的研究中，研究者们主要基于传统的机器学习方法对仇恨言论进行识别。首先针对给定的语料库设计和选择与仇恨言论相关的统计特征，之后利用机器学习方法作为分类器识别仇恨言论。在特征工程的构建方面，文本特征的提取是研究的重点。Nobata 等人[150] 指出，将字符级和单词级 n-gram 特征与其他表层特征结合，包括文本中提及的 URL、标点符号、数字字符、未登记词汇、话题标签、emoji 符号的数量和评论长度等信息，可以提高仇恨言论识别模型的性能。在此基础上，Unsvåg 等人[151]进一步引入用户信息，包括用户的性别、活跃度等特征，取得了较好的效果。Djuric 等人[152]受到分布式语义特征表示方法 Paragraph2Vec 的启发，引入神经语言模型 CBoW 学习社交媒体文本的低维向量表示，使语义相近的用户评论和单词在向量空间的距离接近。之后利用训练好的向量表示对文本进行特征编码，并输入到仇恨言论识别器中。采用机器学习方法进行仇恨言论检测时，模型的性能主要受到人工设计的特征工程的影响，决策具有较强的可解释性。然而，仇恨言论本身是一种复杂的语言

现象，通常依赖于特定的社会、文化、历史和政治背景，因此需要专业领域的知识和经验选择有效的特征，消耗大量的时间和人力成本。

随着深度学习的发展，研究者们开始利用神经网络以及预训练技术进行仇恨言论的识别。周险兵等人[153]分别利用 CNN 和自注意力机制捕捉仇恨言论的字符级拼写特征和句子级语义特征，并结合早期融合和晚期融合获取最终的特征表示进行识别。Mozafari 等人[154]将 BERT 和不同的神经网络结构进行拼接，包括全连接层、CNN 等，以识别仇恨言论，并取得了较好的效果。考虑到仇恨言论样本通常来源于社交媒体，Barbieri 等人[155]利用取自 Twitter 平台的有标签数据对 RoBERTa 进行有监督训练，得到 Twitter-RoBERTa 模型。该模型在七个 Twitter 数据集上都取得了有竞争力的性能。相比于传统的机器学习方法，基于深度学习的仇恨言论检测方法对于样本上下文的理解能力显著增强。然而，这些方法通常在单一的数据源上进行训练，模型的泛化能力有限。同时由于自身的“黑盒”属性，这些方法大多缺乏可解释性，为方法在实际场景下的应用带来了挑战。

最近，大模型技术的突破给仇恨言论识别带来了新的技术支持[156-157]。相比于一般的预训练模型，大模型具有更加丰富的世界知识和理解能力。尽管如此，直接使用大模型进行识别依然存在潜在的风险。由于自身的“幻觉”问题，在识别的过程中大模型会缺乏明确依据时给出看似合理但实际错误的判断。这种幻觉不仅影响检测结果的准确性，还可能导致用户内容被误删或平台公信力下降。因此，如何在充分利用大模型丰富知识的同时，提升模型输出的可靠性，成为未来研究的重点。

③ 仇恨检测模型去偏

由于采样范围的局限性，大多仇恨言论数据集都有大量的偏倚样本，即样本中包含一些高频的词汇标记，例如针对特定群体的提及（如“gay”）。在偏倚数据集上训练过的检测模型容易产生偏见，忽视文本的上下文语义信息，将包含特定词汇的样本直接归类为仇恨言论，影响了模型对特定群体的公平性。

为了解决这一问题，许多研究者致力于仇恨言论检测的去偏研究，目的是降低模型在决策中对特定词汇的依赖程度，进而提升模型的公平性和泛化性。其中，最为直接的去偏方法是通过数据工程，平衡包含偏见词汇的训练数据，使模型在无偏的环境下进行训练。Dixon 等人[158]利用对抗数据平衡原始的仇恨样本，构建包含侮辱性词汇和身份提及的非仇恨样本，保证数据集的无偏性。Zhou 等人[159]采用了 AFLite 和 DataMaps 等多种数据清洗的方法，根据模型决策的难易程度过滤掉包含偏见的样本。Ramponi 等人[160]对原始样本的偏见词汇分别进行移除和掩码操作，以去除数据中的偏见。尽管这些去偏方法合理有效，但是需要消耗昂贵的人力成本，对数据进行二次校验。

除了利用数据工程解决去偏问题以外，一些研究者在原始数据的基础上，对模型训练的过程进行干涉，以减轻模型对词性偏见的依赖。 Vaidya 等人[161]提出了一个多任务框架，同时预测评论中涉及的身份和对应的仇恨标签，使模型根据身份信息判断样本是否表达仇恨，提升模型的公平性。Zhou 等人[159]受到 Learned-Mixin 模型的启发，在原始的分类器的基础上设计了一个分支模型，只根据样本中的偏见词汇进行决策，充分捕捉偏见，并在测试阶段移除分支模型，以减轻偏见对模型决策的影响。

这些去偏方法在实际应用中通常需要引入高频侮辱性词表或身份提及词表作为先验知识，用以捕捉词汇偏见。然而，模型的偏见是在模型的训练阶段产生，而不是由人工确定的。因此，外部词表与实际的偏见词汇并不完全一致，无法对词汇偏见进行充分捕捉。

为了解决这一问题，一些研究旨在提升模型对上下文语境信息的捕捉，在不引入词表的前提下，减轻词汇偏见对模型的影响。Attanasio 等人[162]提出了一种熵注意力正则化（Entropy-based Attention Regularization，EAR）的方法，分析样本中各个词汇对于模型预测结果的贡献程度，并对贡献度高的词汇进行正则化惩罚，使模型更加关注全局信息。Chang 等人[163]引入变量有理化方法（Invariant Rationalization，InvRat），通过捕捉模型对输入数据中的恒稳性，降低偏见词汇对模型的影响。

④ 多模态仇恨模因检测

除了文本形式的仇恨言论以外，多模态的模因图也经常作为仇恨的载体。相比于一般的仇恨言论，多模态仇恨模因更具有煽动性，更容易被极端的网民传播，导致严重的社会危害。因此，采取相应的检测措施来防范仇恨模因的传播刻不容缓。

在早期的工作中，研究者们直接采用已有的多模态预训练语言模型对仇恨模因进行检测，通过 ViLBERT、Visual BERT 等方法获取模因中图片和文本的向量表示[164]。随着研究的推进，研究者们尝试引入一系列通用的深度学习方法，包括数据增强、集成学习等，进一步提升模型的性能[165]。除了基本的模态信息以外，研究者们还注重挖掘模因中图像和文本的潜在信息，并重点关注了仇恨模因的攻击对象。一些研究者将模因的攻击对象作为辅助特征引入到检测模型中。Lee 等人[166]引入人脸识别技术获取模因中包含的离散的人口统计信息，包括人物的性别和种族。还有研究者将攻击对象识别作为模型训练过程中的辅助任务。Pramanick 等人[167]提出一种多任务框架 MOMENTA，使模型同时对攻击对象识别和仇恨模因检测两个任务进行优化，提升模型的可解释性。这些方法充分地利用了图像和文本的特征，增强了检测模型对模因的理解。然

而，它们通常对图片和文本采用不同的编码器，导致图文的向量表示之间存在不一致性，引发语义鸿沟，不利于模型的优化。

为了解决语义鸿沟问题，研究者们开始利用更加详细的图像标题作为补充知识辅助模型进行检测。Blaier 等人[168]采用图像字幕工具捕获图像内容，实验表明在微调过程中结合图像标题信息可以明显改善各种基座模型的效果。在此基础上，一些研究者将标题与原始文本以及图片实体特征进行融合，以提升模型的检测效果。Cao 等人[169]提出了基于提示学习的 PromptHate 方法，利用 CLIP 模型获得模因图的标题，将原始的模因文本与图像标题以及图像中包含的实体信息统一输入到提示模板中，利用文本编码器中进行检测，将多模态的仇恨模因检测转化成文本分类任务。受此启发，Ji 等人[170]进一步在提示模板中引入创作模因的背景知识，利用 BLIP 模型补充图片背景相关的属性信息。这些方法连接了图像和文本的语义信息，有助于模型的优化。但是与此同时，也容易受到标题质量的影响，忽略图像中的细节信息，对于包含了复杂语义的仇恨模因的检测效果较差。

近年来，随着多模态大模型（MLLM）的快速发展，部分研究者尝试利用其强大的推理能力来增强模型的可解释性。Lin 等人[171]提出了一种生成式框架，借助大语言模型为模因内容生成合理的解释，并通过这些解释数据对小规模模型进行知识蒸馏，从而提升其检测性能。随后，Lin 等人[172]又提出了一种基于 LLM 的“辩论式”框架，针对同一模因分别生成有害性与无害性的解释，以此辅助模型做出更具推理能力的判断。尽管上述方法在提升模型推理能力方面取得了一定进展，但与传统仇恨识别方法类似，这类研究同样忽视了大模型自身可能存在的“幻觉”问题，即生成内容缺乏事实依据或逻辑支撑的风险。

除了方法类的研究外，一些研究者也关注仇恨模因检测模型的可解释性。Hee 等人[173]通过大量的定量实验和定性分析，深入探究了文本和视觉特征在检测任务中的贡献程度，模型对模因中侮辱性词汇的理解能力，以及模型中的偏见和错误分类情况等问题，并提出传统的端到端分类的方式无法真正评估检测模型是否能够对模因的含义进行正确分析。为此，他们构建了带有原因解释的仇恨模因数据集（HatReD）[174]，并定义了仇恨模因解释生成任务，旨在让模型根据仇恨模因的内容生成其有害原因的解释。

（3）抑郁症倾向识别

① 任务定义

抑郁症是一种常见的心理障碍，长期未被识别并加以治疗可能导致情绪崩溃、自我伤害，甚至自杀。据世界卫生组织统计，全球有超过 3 亿人受到抑郁症影响[175]，然而由于传统诊断方法（如面访与量表评估）依赖专业人员、效率较低，导致大量潜在患者无法及时获得干预[176]。近年来，随着社交媒体平台的

兴起，用户在微博、Twitter、Reddit 等平台上频繁发布文本、图像、语音等内容，为识别其心理健康状态提供了可能。尤其是与深度学习方法与自然语言处理（NLP）的快速发展，使得从社交媒体数据中挖掘抑郁倾向成为研究热点[175-176]。

随着深度学习与自然语言处理的发展，研究者们提出了多种用于抑郁检测的建模方案，广泛探索了不同的信息模态与建模机制。其中，最流行的方法为基于用户发布的文本内容，通过语言特征建模实现抑郁倾向的分类或预测[175-176]。近些年，抑郁检测研究逐渐将注意力扩展至语音、图像与用户行为等多模态信号，以捕捉更丰富的心理线索并提升识别精度[177,184]。随着模型复杂度提升和实际应用需求增强，越来越多研究开始关注模型的可解释性与决策透明度，尝试从模型中提取情绪线索或症状映射，以增强预测结果的可理解性与可干预性[179-180]。

尽管相关研究已取得诸多进展，但由于社交媒体数据存在语言风格多样、标签主观性强、模态异构性高等特点，抑郁倾向检测任务仍面临诸多挑战。如何根据输入模态与建模目标选择合适的建模范式，构建兼具性能与实用性的识别系统，成为当前研究关注的重点。

为全面理解当前抑郁倾向识别的研究进展，本节聚焦于基于社交媒体数据的抑郁识别任务，系统回顾并比较近年来在该领域的重要研究成果。具体而言，相关研究可大致分为三类：第一类采用文本为唯一信息源，聚焦于用户发布内容的语言特征建模；第二类引入语音、图像等辅助模态，形成多模态建模框架；第三类强调模型的可解释性与临床可用性，探索抑郁识别结果的成因解释与人机交互机制。

下文将按此分类对代表性工作进行梳理和分析。通过对比各类方法的建模特征、技术路径与实际挑战，揭示当前研究的演进脉络与未来发展方向。

② 基于文本的抑郁识别

作为抑郁倾向识别研究中最早且最为广泛的方向，基于文本的方法通常依赖用户在社交媒体上发布的语言内容，通过挖掘文本中的情感、心理和行为特征以预测其抑郁风险。相关研究在模型架构上经历了从传统机器学习及深度学习方法到预训练语言模型（PLMs）再到大语言模型（LLMs）的演进，极大地提升了文本表示与情感理解的能力。

在早期研究中，研究人员专注于挖掘抑郁相关的特征，并结合各种机器学习方法进行抑郁识别[181-183]，这种方法可解释性强，但需要人工设计特征耗时耗力，并且需要很强的专业知识。随着，深度学习的发展，主流方法多采用卷积神经网络（CNN）、循环神经网络（RNN）、注意力机制或混合模型进行文本建模，以避免人工特征工程。Zhou 等人[184]提出一种端到端的深度卷积神经

网络，结合文本嵌入与多层卷积池化结构，用于从社交媒体帖子中检测抑郁倾向。Tejaswini 等人[185]提出一种结合 FastText、卷积神经网络（CNN）与长短期记忆网络（LSTM）的混合深度学习模型 FCL，用于社交媒体文本中的抑郁倾向识别。

Trotzek 等人[186]提出将卷积神经网络与用户语言元特征结合，用于社交平台文本的抑郁早期检测任务。模型利用不同词嵌入构建文本表示，同时构建独立的语言统计特征进行用户建模，最终通过集成方法融合二者的判断结果，在 ERDE 指标下取得领先性能。然而基于深度学习的方法，数据依赖程度更强，且难以结合上下文挖掘文本的深层语义特征。

随着预训练语言模型的发展，BERT 等模型被广泛用于抑郁识别任务。例如，Verma 等人[187]提出基于改进版 BERT 的文本抑郁倾向检测方法，利用预训练语言模型的语言理解能力，建模用户在文本中的情感模式与认知特征，实现高准确度的分类预测。Teck 等人[188] 采用预训练的 BERT 模型，探索其在中文社交媒体文本中识别抑郁倾向的能力，重点关注模型在中文语境下的效果。

最近，大型语言模型在该任务中也展现出强大的推理与语言理解能力。Shah 等人[189]采用指令微调策略对 LLMs 进行适配优化，使模型能有效识别用户表达中的抑郁信号，效果优于现有多项 SOTA 方法，凸显了 LLM 在抑郁早期识别中的强泛化能力与部署潜力。

总体而言，基于文本的抑郁识别方法在输入简单、数据易得等方面具有明显优势，且模型性能在不断提升。但由于仅依赖文本信息，仍可能面临模态信息不全、语境理解不足等问题。为进一步提升识别准确性，后续研究逐步引入多模态信号，形成融合性建模框架。

③ 多模态抑郁识别

为弥补文本信息在语义表达和心理状态呈现方面的局限性，越来越多研究尝试引入多模态数据，包括语音、图像、用户行为等，从多个维度构建更全面的抑郁识别模型。多模态方法不仅丰富了模型输入的信息量，也为提高模型的鲁棒性与泛化能力提供了新的可能。

在文本与语音模态结合方面，Gan 等人[177]提出一种基于教师-学生架构的多模态抑郁检测模型，融合文本与音频模态，并引入多头注意力机制与加权迁移学习以优化模态融合与特征提取。Shen 等人[190]作者提出一种结合音频特征与语言内容的抑郁检测方法，利用采访过程中的语音与文本信息构建了一个多模态深度学习模型，并基于此设计了一个双路径结构，分别使用 GRU 和 BiLSTM 对音频和文本进行建模。

除语音外，部分研究引入图像或用户行为模态，以拓展建模范围。Huang 等人[191]提出一种以文本为主导的多模态抑郁检测框架，设计跨模态融合模块将

视觉特征注入文本表达，并通过多层感知器混合单元实现深层特征交互。Tank 等人[192]在文本、音频与视频模态上联合应用大型语言模型与多模态神经网络，分别进行回归与分类任务，用于预测抑郁严重程度（PHQ-8 分数）及分类识别。Hu 等人[193]提出一种融合用户文本与图像的多模态抑郁检测方法，通过显式提取图像中隐含的情绪信息，与文本特征共同输入融合模型进行预测。

整体来看，多模态抑郁识别方法突破了文本建模的单一视角，在提升检测准确率和稳健性方面具有显著优势。然而，该方向也面临模态异构性高、数据对齐难度大等挑战，未来研究可进一步探索跨模态对齐机制与统一语义空间构建策略。

④ 可解释的抑郁识别

随着人工智能模型在该领域应用逐渐深入，抑郁识别系统的可解释性问题受到越来越多关注。相比于单纯的分类结果，临床心理健康干预更需要系统提供推理依据与行为线索解释。因此，已有研究逐步探索提升模型可解释性的方法，包括引入症状标签、注意力机制、情绪建模以及交互反馈等技术路径。

Zhang 等人[194]提出一种知识感知的深度学习系统（Deep-Knowledge-Aware Depression Detection），通过融合抑郁领域知识与社交媒体数字踪迹，实现抑郁风险用户的识别与影响因子解释。该研究以信息系统设计科学为框架，强调领域知识在特征提取与模型解释中的作用，提升了模型的检测准确性与泛化能力。Lan 等人[195] 提出一种结合医学知识与大语言模型的抑郁检测系统 DORIS，利用 LLM 对用户文本进行诊断标准标注、情绪轨迹建模与摘要提取，并融合传统分类器实现预测与解释。Bao 等人[179]提出多种基于 Transformer 的模型架构，以支持在社交媒体文本中同时进行抑郁症状检测与自然语言形式的解释生成，包括分离式和统一式两类建模方案，并探索了大语言模型在该任务中的应用潜力。该关注模型的可解释性，通过生成与临床症状一致的自然语言解释增强模型透明度与可信度。

从整体趋势来看，可解释的抑郁识别研究不仅有助于提升模型的透明度和可用性，也为心理健康专业人员提供了决策辅助。然而，目前该方向仍面临解释结果结构不一致、缺乏统一评估指标等挑战，未来应加强模型可解释性与人类理解的一致性评估机制，并推动该类模型在真实场景中的可部署性研究。

从范式划分来看，当前抑郁识别研究主要集中在三类路径：基于文本的方法、多模态方法与可解释方法。每一类方法具有各自的技术特点与应用场景，也面临不同的技术挑战与研究瓶颈。首先，基于文本的抑郁识别方法在输入形式上最为简单，易于部署，适用于数据资源有限或以文本为主的社交平台。这一类研究聚焦于语言特征的表达能力，重点在于如何提升文本建模的深度和语

境理解能力。从传统 RNN/CNN 架构到 Transformer 和大语言模型的发展，文本方法不断追求更强的语义建模与泛化能力。

其次，多模态方法通过引入语音、图像、社交行为等信息弥补单一文本在心理状态呈现上的不足。这类方法更贴近用户全貌建模，强调不同模态之间的协同建模与时序依赖。其主要优势在于识别精度的提升和对细粒度抑郁线索的感知能力增强，但在模态对齐、数据缺失、特征冗余等方面仍面临挑战。

最后，可解释方法尝试提升模型的透明度与临床可用性，强调对预测结果的因果溯源与人类可读的解释输出。这类方法多结合注意力机制、症状本体映射或生成式语言模型输出，使模型输出不再是黑箱。虽然该方向仍面临解释一致性差与缺乏统一评估指标的问题，但其在应用中具有不可替代的重要性。

总体来看，三类研究路径各具优势也各有挑战。文本方法强调效率与可拓展性，多模态方法追求表达完整性，可解释方法聚焦模型信任与可用性。未来研究应进一步融合三类范式的优点，构建性能强、可解释的抑郁识别系统。

4.3.3 总结与展望

负面情感分析作为自然语言处理领域的重要研究方向，其涵盖的反讽识别、仇恨言论识别、抑郁症倾向识别及网络舆情分析等任务，既体现了技术的多样性，也凸显了现实需求的复杂性。从技术演进来看，自然语言处理方法已从早期的规则与词典驱动，逐步发展为基于预训练模型的深度学习框架，大语言模型时代更是丰富了研究范式，在语义表征的语境依赖性、情感特征的细粒度捕捉上取得显著突破，为复杂负面情感的识别提供了有效工具。

本节系统性地梳理了负面情感分析领域的研究进展，揭示了当前技术发展的主要成就。同时，针对以上介绍的任务，我们分别提出了未来研究的方向，希望能够为相关领域的研究者提供参考。

（1）**讽刺检测领域**的研究将朝着更具深度与广度的方向拓展，以下几个关键方向值得重点关注，它们既回应了当前技术瓶颈，也为领域突破指明了路径：

多语言数据集的体系化建设。当前讽刺检测研究多依赖英语单语数据集，而讽刺的表达与语言结构、文化语境深度绑定，导致现有模型在跨语言场景中泛化能力受限。未来需重点推进多语言讽刺数据集的规模化、精细化建设。

相关任务的联动机制与协同建模。讽刺与幽默、冒犯性语言、反讽等任务存在深刻的语义关联（如讽刺常借助幽默形式表达，部分讽刺可能隐含冒犯性），但当前研究多将其视为独立任务，割裂了语义层面的内在联系。未来需深入探索任务间的依赖关系。

上下文感知能力的深层强化。讽刺的表达往往依赖动态对话历史与广泛的社会语境（如特定事件背景、社群共识、流行梗等），现有模型对长程对话逻辑、隐性社会知识的捕捉仍显不足。

（2）**仇恨言论检测**研究将重点围绕"低资源语言支持"、"复杂表达识别"和"模型可解释性"三个关键方向展开深入探索。

在低资源语言支持方面，当前研究过度依赖英语等高资源语言数据，导致对众多小语种和方言的支持严重不足。为解决这一瓶颈，未来研究需要重点突破跨语言知识迁移技术，开发基于半监督学习和主动学习的智能标注系统，构建更具代表性的多语言仇恨言论语料库。

在复杂表达识别方面，仇恨言论常隐含于反讽、隐喻等隐式表达中，现有的方法难以捕捉此类复杂表达。未来研究需结合上下文感知技术，利用大型语言模型增强语义理解。同时，需关注动态演变的仇恨表达形式，通过持续学习和对抗训练提升模型鲁棒性。

在模型可解释性方面，亟需突破传统黑箱模型的局限，构建融合注意力可视化、决策路径追溯等技术的可解释框架。不仅需要从算法层面提升透明度，更要建立包含伦理评估在内的多维可解释性评价体系，确保检测结果具有可信度。

（3）**抑郁倾向识别**研究工作从建模范式出发，可将现有研究归纳为三类：基于文本的方法、多模态方法以及可解释方法。

基于文本的抑郁识别方法仍然是该领域的主流技术路径，其优势在于数据获取容易、计算成本低，适用于大规模用户情感状态的快速筛查。尤其是在大语言模型兴起之后，文本表示与情绪理解能力显著提升，为文本范式注入了新的活力。未来可以构建多语言、多文化、多模态的大规模抑郁检测数据集

多模态方法通过融合语音、图像、用户行为等多源信息，实现了更为全面的用户状态建模，具有更高的识别准确率和鲁棒性。随着多模态神经网络和跨模态融合机制的发展，该方向已成为提升模型性能的重要突破口，探索统一的跨模态对齐与语义表示方法，未来在心理状态建模与行为干预系统中具有广泛应用潜力。

随着 AI 系统在敏感任务中的广泛应用，可解释抑郁识别方法逐渐受到重视。该方向通过引入心理症状标签、注意力机制、模型可视化与自然语言解释模块，提升了模型的透明性与信任度，是推动技术走向临床实践与真实使用的重要前提。虽然抑郁检测研究已经取得了显著进展，当前研究在数据质量与泛化能力、模态协同机制、可解释性评估标准等方面仍存在显著挑战。设计兼顾性能与解释能力的混合型架构，推动相关研究向真实心理健康服务场景落地。

4.4 面向多模态语义关联的语言智能

知识图谱的演进本质上标志着机器认知能力的一次重大范式升级。传统文本知识图谱虽然构建了符号化世界的语义骨架，却难以充分承载现实世界的动态感知信息[196]。多模态知识图谱（MMKG）的出现，通过融合文本、图像、音频、视频等多源异构数据，突破了“文本牢笼”的局限，推动语言智能技术迈向跨模态协同的具身认知时代。

在知识图谱构建过程中，首先需要对不同模态信息进行实体识别和关系抽取，将视觉对象的空间轮廓、听觉情感表征、时序行为的运动轨迹转化为可计算的知识单元，从而为机器打造“视觉、听觉、语义”联动的认知基础[197]。这一过程不仅丰富了知识图谱的语义表达，还增强了其在多模态环境下的适用性和灵活性。知识图谱为机器提供了一种结构化的知识表示方式，使机器能够理解和推理复杂的关系和概念。在下游任务中，知识图谱的应用广泛而深远，包括但不限于问答系统、推荐系统、决策支持和自然语言理解等领域。通过知识图谱，机器能够实现对信息的高效检索、推理和预测，极大地提升了智能系统的效能和准确性[198]。

多模态语义关联的语言智能，进一步将知识图谱的概念推广到多模态环境。这种多模态知识图谱不仅能够处理和理解文本信息，还能够整合视觉、听觉和动态行为等多模态数据，实现跨模态的细粒度语义关联、理解和辨析[199]。这使得综合语言智能应用，如面向跨模态、多证据综合的多跳推理问答，以及多模态事实核查成为可能。这种“感知—认知—决策”闭环的重构标志着语言智能的质变，使得机器能够以更加全面、准确和可靠的方式理解和响应复杂的多模态信息。

4.4.1 多模态知识图谱构建

（1）知识图谱

知识图谱（Modal KG）以文本三元组 $\langle h, r, t \rangle$ 为基本单元，通过实体识别、关系抽取与本体对齐三大步骤完成构建。早期系统依赖人工编写的正则模板与 WordNet、FrameNet 等专家词典，这种“模式—词典”的方式从新闻、百科中抽取少量高置信度关系，形成了百万级规模的知识库。统计学习阶段，远程监督首次将 Freebase 与维基文本对齐，实现弱监督自动标注[200]；BiLSTM-CRF 结合字-词级特征，提升了实体边界识别的鲁棒性[201]；图神经网络进一步利用实体间的拓扑结构，增强了关系分类的全局一致性，使图谱节点迅速扩张至千万级[202]。Transformer 时代，Devlin 等人[204]提出的编码器通过大规模预训练语言模型，关系抽取任务提升显著；同时，TransE、RotatE 等几何嵌入方法在 Freebase、Wikidata 上构建了十亿级三元组的可推理空间，为下游应用提供了高覆盖度的语义支持。

技术的进步让知识图谱成为搜索引擎、问答系统、推荐系统的核心底座。Google Knowledge Graph 上线后，用户的二次点击比例显著下降[205]，阿里 2.1 亿节点的商品图谱推荐转化率提升[206]，Zhang 等人[207]在机器翻译任务中注入 5000

万三元组，BLEU 分数得到了提升，Liu 等人[208]的工作 K-BERT 将对话生成困惑度降低。

然而，文本符号只能承载离散语义，无法还原连续感知世界的完整信息，导致知识表示与现实感知间存在不可忽略的语义鸿沟。Chen 等人[209]指出，缺失视觉细节导致数字化误差高；Yang 等人发现，缺少声调与面部微表情信息使情感误判率增加。Ji S 等人[196]的实证调研进一步表明，在自动驾驶、工业质检、沉浸式教育等真实多模态场景中，单模态知识图谱在关键场景覆盖度不足。其本质源于单模态表示的认知贫瘠性，视觉信息缺位则无法刻画物体外观与空间关系；听觉建模空白则难以捕获声调、音乐、环境声等语义线索；动态性匮乏则无法反映事件随时间的演变。因此，打破文本边界、融合视觉、听觉、语义信号的多模态知识图谱成为下一阶段知识表示与推理的必然选择。

（2）多模态知识图谱

为弥合感知缺口，多模态知识图谱（Multimodal Knowledge Graph）应运而生：通过统一编码将视觉对象、听觉事件、时序行为与文本语义映射到同一高维空间。多模态知识图谱通过整合文本、图像、视频等异构数据，构建实体、属性及关系的结构化语义网络，成为突破认知智能边界的关键基础。其核心价值在于解决单一模态的语义局限，例如视觉信息可明确区分“苹果”作为水果实体（关联红色果实图像）与公司实体（关联电子产品 Logo），并为跨模态推理提供知识支撑。

知识构建：预训练语言模型（PLM）为 MMKG 构建提供了统一语义编码能力。对于实体识别来说，PLM 通过对齐文本与视觉语义缩小模态鸿沟。此外跨模态映射是关键，Chen 等人[210]提出动态预测像素级语义相关性，而 Kao 等人[211]提出注入知识图谱三元组强化视觉和语言的对齐；Li 等人[212]利用提示工程对比图文事件描述，精准定位“地震”事件的破坏范围与救援行动等角色。针对时序数据，Sun 等人[213]采用时空注意力编码视频片段，实现对动态行为的结构化表示。

表示与融合：PLM 的架构设计决定模态交互效率，双流模型分离处理模态后经跨模态注意力交互，单流模型则将文本词向量与图像区域嵌入同构输入。另外知识增强策略进一步提升表达力，Yu 等人[214]引入视觉场景图预训练，使联合嵌入包含空间关系语义。Li 等人[215]则通过掩码多模态建模与海量图文对比学习构建统一语义空间。同时，跨模态检索技术（如双塔模型+余弦相似度度量）为图谱查询提供底层支持。

推理应用：多模态推理呈现两大范式。结构化路径推理，如将“药品副作用查询”分解为视觉成分识别到文本说明书检索再到知识图谱链接的推理链；图神经网络（GNN）在此类任务中实现多跳关系推理。生成式协同推理调用大语言模型（LLM）迭代解析表格数据，利用大语言模型生成缺失的模态特征辅助决策（如补全模糊监控视频中的行为描述）。

然而当前的多模态知识图谱的发展仍然面临着瓶颈：首先是模态失衡,大部分模型依赖 PLM 文本编码器主导学习，视觉信息利用率不足；其次是动态性缺失 PLM 参数固化导致难以实时更新知识；然后是评估缺陷，传统指标无法量化真实跨模态推理能力，存在“虚假记忆”干扰。未来突破需聚焦三方面多模态大模型协同：利用多模态大模型实现“文本-图像”的自动化图谱构建，结合思维链（CoT）技术增强跨模态因果推断；轻量化动态注入：设计适配器（Adapter）模块低成本更新知识（如工业质检中新缺陷图像的快速嵌入）；认知决策闭环：以 MMKG 为事实校验器约束 LLM 幻觉（如医疗诊断中影像与病理报告的互证），推动机器人“环境感知-行动规划”闭环。MMKG 正通过 PLM 驱动的跨模态融合重塑知识表示范式。随着动态协同架构的演进，其将从静态知识库进化为“感知-推理-决策”一体的认知引擎，这一进程不仅依赖算法创新，更需建立涵盖模态均衡性、时效性、可解释性的新一代评估体系。

当前的多模态知识图谱的发展面临的瓶颈：模态失衡：90%以上的模型依赖预训练大语言模型（PLM）文本编码器主导学习，视觉/时序信息利用率不足，且高质量对齐的多模态训练数据稀缺；动态性缺失：PLM 参数固化导致难以实时更新知识，持续学习能力弱，面临灾难性遗忘挑战；评估缺陷：传统指标无法量化真实跨模态推理能力，缺乏对噪声数据、分布偏移的鲁棒性测试，且存在“虚假记忆”干扰；异构对齐困境：文本的离散符号性与视觉的连续信号间存在本质语义隔阂，多源知识冲突消解机制尚未成熟。

未来突破方向三方面：多模态大模型协同：利用多模态大模型实现“文本-图像”的自动化图谱构建，结合思维链技术增强跨模态因果推断，探索神经符号融合架构提升可解释性；轻量化动态注入：设计适配器（Adapter）模块低成本更新知识（如工业质检中新缺陷图像的快速嵌入），发展持续学习框架平衡新知识融入与旧知识保留；认知决策闭环：以 MMKG 为事实校验器约束 LLMs 幻觉（如医疗诊断中影像与病理报告的互证），推动机器人具身智能中的环境感知-行动规划闭环，并嵌入因果推理模块；可信评估体系：建立涵盖模态均衡性、时效性、可解释性、公平性及鲁棒性的新标准，开发任务驱动的复杂评估基准（如需多跳推理的跨模态问答）。MMKG 正通过 PLM 与 GNN 驱动的跨模态融合重塑知识表示范式。随着动态协同架构的演进，其将从静态知识库进化为“感知-推理-决策”一体的认知引擎，这一进程需在算法创新、数据构建、评估体系三端协同突破，最终支撑可信赖的通用语言智能系统。

（3）多模态知识图谱关键技术

① 多模态命名实体识别

多模态命名实体识别（Multimodal Named Entity Recognition, MNER）旨在同时利用文本与图像等多模态信息，提升命名实体识别的准确性，尤其适用于社交

媒体、新闻图文等场景，图 4-17 为多模态命名实体识别的基本框架。

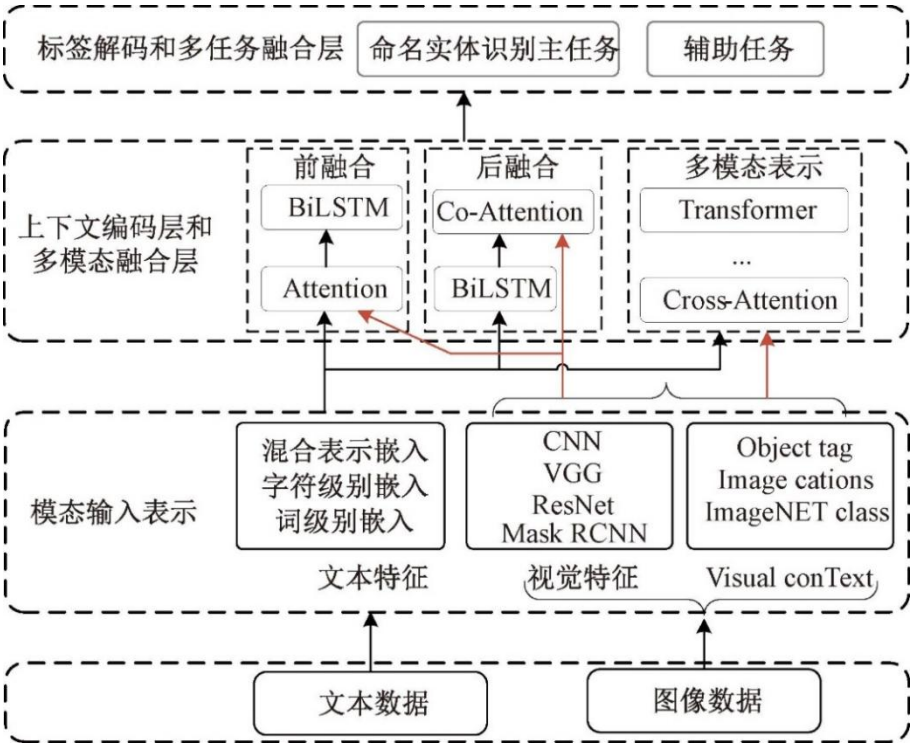

图 4-17 多模态命名实体识别的基本框架[216]

传统 NER 方法依赖文本上下文，而短文本中上下文不足、歧义性强，引入视觉模态可有效缓解这一问题。近年来，MNER 研究主要围绕模态融合策略与语义对齐机制展开，形成了基于 BiLSTM 与基于 Transformer 的两大技术路线。MNER 的输入为图文对（文本+图像），输出为文本中的实体边界及类别（如 PER、LOC、ORG）。其核心挑战包括：语义鸿沟：文本与图像特征空间差异大，直接拼接易引入噪声。视觉偏差：图像中无关区域可能误导实体预测。标注稀缺：多模态联合标注成本高昂，需设计高效融合策略。

早期工作以 BiLSTM-CRF 为基线，探索视觉特征增强文本表示。前融合模型将图像区域特征与词向量拼接后输入 BiLSTM，Moon 等人[217]提出的 Glove 词向量静态语义不足。后融合模型（如 ACN, GAN）先通过 BiLSTM 编码文本上下文，再与视觉特征共注意力交互，缓解语义断层。对抗学习（GAN）进一步对齐图文分布。然而，BiLSTM 的长程依赖建模能力有限，难以捕捉复杂图文关联。Transformer 的兴起推动 MNER 进入预训练时代，BERT 的上下文动态表示显著缩小了图文语义差距。单任务模型通过跨模态注意力融合 BERT 文本特征与区域视觉特征，但视觉偏差问题仍存在。多任务模型通过辅助任务优化多模态表示：Zhang 等人[218]提出的联合实体跨度检测任务，强制模型关注文本主导边界。Wang 等人[219]引入图像描述作为显式语义桥梁，图文对齐更精准，F1 进一步提升。HVPNeT 利用层次视觉特征（Swin Transformer）与对象标签协同表示，在 Twitter—2015 上进展显著，验证了多粒度视觉信息的价值。

表 4-21 多模态命名实体识别方法分类表

方法	典型模型	优点	缺点
前融合模型	MA	简单、直接的方式实现多模态融合，性能优于 NER	图文特征存在语义鸿沟
后融合模型	ACN	文本图像语义差距较小；多模态特征质量较高	文本特征的语义低
Transformer 单任务模型	UMGF	文本语义和图像语义更接近；图像和文本特征深度融合	视觉特征语义存在偏差
Transformer 多任务模型	UMT	多任务协同优化多模态特征，多模态特征语义更准确	视觉特征缺失

文本的模态输入表示一般采用 BERT 动态向量，而图片模态的输入表示采用 ResNet 区域特征或 Swin Transformer 层级特征，对象标签（Mask R-CNN 提取）可显式引入视觉语义。模态的融合方式也由前期拼接与后期拼接逐步演进到端到端 Transformer 融合。而多头注意力机制实现了图像文本 token 级对齐。除实体识别外，边界检测、关系抽取等任务通过共享表示层抑制视觉噪声，提升模型的泛化性。

② 多模态信息抽取

信息抽取的主要目标是从无结构的生文本中抽取出结构化、半结构化或非结构化的数据。而多模态信息抽取则是多模态学习与信息抽取技术的结合，是当前多模态知识图谱研究的主要工作之一。单模态研究的进步为多模态的研究提供了扎实的基础，对于不同模态的信息可采取先进行单模态抽取，再进行多模态处理的方法。本节主要讨论图像模态的信息抽取并简要介绍声音模态的信息抽取。

1）图像模态信息抽取

ImageNet、IMGpedia 等视觉数据库通常是图像或视频数据的丰富来源，能够为知识图谱中的实体提供充分的视觉信息[220][221]。基于实体属性补充角度的图像模态信息抽取方法比较简便，通常基于文本知识图谱，利用关键词，通过搜索引擎从视觉数据库中检索与文本实体对应的图像模态数据进行补充，图像信息的判别主要依赖图像元数据对图像内容的描述，具有一定的局限性。基于图像模态实体构建与视觉关系发现角度的图像模态信息抽取中，元数据的作用同样不可忽视，但更多关注点在内容本身的特征上。直接通过人为理解从图像中提取出有效而丰富的特征较难，早期主要考虑图像的色域、对比度和纹理等，在深度学习出现之前，图像识别主要借助尺度不变特征转换（SIFT）、方向梯度直方图（HOG）等算法提取具有较好区分性的特征，再集合 SVM 机器学习算法进行图像识别，HOG+SVM 在行人检测中有着优异的效果。卷积神经网络可以直接将图像的原始像素作为输入，避免了先使用 SIFT 等算法提取特征，减轻了大量数据预处理

工作。

在图像特征提取的角度下，通过卷积和池化操作，产生图像模态的矩阵表示，图像特征的向量表示工作由全连接层或全局均值池化层的输入完成。代表性的LeNet-5 在手写体数字和字母的识别中得到了非常高的精度，在车牌识别等场景中得到实际应用。在 LeNet-5 的基础上，通过增加网络的深度或对卷积和池化操作进行变形，后续研究提出了更加复杂和高效的卷积神经网络，陆续有 AlexNet、ResNet、CapNet 等，对图像模态的特征提取能力进行了拓展，极大地提高了图像识别的精度。借助经过预训练的 CNN 模型，可以在图像实体的类识别上取得较优效果。

2）声音模态信息抽取

在现有多模态知识图谱研究中，声音模态主要还是以实体的属性形式存在，代表描述实体的一段音频。从构建角度除了利用元数据，语音数据也可通过语言解析为文本进而用传统的文本的方法进行语义特征提取。现有多模态知识图谱构建工作中，对于文本模态之外的模态信息的抽取大多属于被动抽取，是在依托现有文本实体的基础上在有限范围内定向地进行模态数据补充，在视觉关系发现中也仍需要文本描述等额外知识的介入。

③ 多模态表示学习

为了方便对抽取到的多模态信息进行处理，需要对输入的数据进行表示，在深度学习时代，采用的表示方法是将输入的数据表示成向量，进而通过深度神经网络的强大建模能力，自动地对输入数据中的特征进行提取。在知识图谱的研究和应用中，知识表示学习是其基本任务之一。信息可以由单模态的表示学习转换为能被机器处理的数值向量或者进一步抽象为更高层的特征向量，多模态表示学习旨在减小模态信息在联合语义子空间中的分布差距，同时保持模态特定语义的完整。

1）文本模态表示

文本表示的核心是对语言基本单元进行表示，然后用神经网络学习语言模型提取文本特征，最后用神经网络的某个输出向量作为文本表示。早期使用独热表示（one hot）单词，每个词的表示为词典中该词的索引，然而这种形式空间损耗较大，并且不能建模词之间的语义相似性，同时存在数据稀疏问题。后续一般使用神经网络模型得到的向量作为词向量。提取文本特征的神经网络主要包括简单的前馈神经网络，以及擅长序列建模的循环神经网络（Recurrent Neural Network, RNN），例如长短时记忆网络（Long Short Term Memory，LSTM）及其变体。Devlin J 等人[222]提出的 BERT 具有比 RNN 更好的文本建模效果，逐渐取代 RNN 成为主流的文本特征提取方法。

2）图像模态表示

卷积神经网络是在多层神经网络的基础上发展起来的针对图像而特别设计的一种深度学习方法，在图像处理上取得了优异的效果。近年来由于深度学习技术的发展以及计算机处理能力的提高，图片等多媒体数据可以和文本采用相同的深度学习框架分析，这为多模态研究提供了便利[223]。

3）声音模态表示

自然界的声波首先以连续模拟形态存在其时域波形仅刻画声压随时间的起伏，计算机难以分辨音高、音色等人类可以轻松辨别的特征。因此，需通过采样与量化将其离散化为数字序列，再借助人类听觉与语言学的先验知识提取声学特征向量。常用的信号处理流程包括傅里叶变换、线性预测和倒谱分析等。在多模态场景下，同一实体的音频往往与其他模态呈现“信息富余”与“信息缺失”并存的局面。为调和这种不均衡，Baltrušaitis TD[258]等人把声音表示策略概括为两大范式：

联合（joint）表示：将声学特征与视觉、文本等模态特征共同映射到一个共享的语义空间，形成统一的多模态向量；

协同（coordinated）表示：让各模态先独立进入各自的子空间，再通过相关性约束（如对比损失或互信息最大化）保持跨模态语义一致性，而无需直接融合原始特征。

单模态研究的进步使得多模态的研究有了更扎实的基础，文本、图像、声音模态的深度学习的发展以及算力的支持为多模态表示学习寻找合适的共享空间提供了便利。基于特征的多模态表示学习方法根据知识图的结构和实体的视觉表示来定义三元组的评分函数，这意味着每个实体必须包含图像属性。然而，在实际场景中，一些实体不包含多模态信息，所以这种方法不能被广泛使用。基于实体的多模态表示学习方法将多模态信息认为是知识图谱中的一等公民，使用基于CNN 的方法来训练知识图的嵌入，在部分补全问题上效果优于基于特征的方法，但这种独立的处理方法容易忽略多模态信息的融合。鉴于多模态知识图谱中存在多种模态下的实体及语义关系，未来在其表示学习研究上，或应该结合以上两种角度的方法。此外多模态学习过程中哪些额外部分的数据能够为知识抽取和关系发现提供支援并不明确，并且存在的信息冗余和噪声是需要解决的一个关键问题，注意力机制的引入为缓解噪声、增加任务细粒度效果提供了一定的支持。

④ 多模态实体链接

实体链接（Entity Linking, EL）任务是指将从给定资源中抽取的实体对象链接到知识图谱中对应的实体中。其基本思想是从知识图谱中为给定的实体指称选择一组候选实体对象，然后根据相似度比较将实体链接到正确的实体对象。在基于实体的多模态知识图谱的构建工作中，存在同一文本实体对应多张图像的情况，也存在同一图像对应多个文本实体或描述的情况，因此需要进行多模态实体链接。

现有多模态实体链接工作主要建立在多模态信息抽取和表示学习的基础上，需要一定的背景知识提供支撑。利用 RCNNs 提取图像的视觉特征有利于缩小候选实体的范围，使用注意力机制能够减小噪声影响，但对于实际多模态知识图谱构建中一对多、多对多的多模态实体关系发现仍面临噪声处理能力不足和先验数据不充分的挑战。

（4）多模态知识图谱的典型应用

多模态知识图谱技术可以服务于各种场景，例如多模态实体链接技术可以融合多种模态下的相同实体，能够广泛应用于新闻阅读、商品推荐等场景；通过远程监督可以补全多模态知识图谱，完善现有的多模态知识图谱，利用动态更新技术使其更加的完备，在端到端实体分类、多模态摘要中也有实际应用。

① 推荐系统

多模态知识图谱将其他模态的信息引入传统的知识图谱，可以提供丰富的特征信息，将其应用在推荐系统可以有效缓解推荐系统的数据稀疏和冷启动问题，从而使推荐结果更准确，并提供可解释性支撑。目标项目及其属性可以映射到知识图谱中以理解项目之间的相互关系。另外，用户和用户侧信息集成到知识图谱中可以更准确的获取用户和项目的偏好关系。实体的图像和描述可以为知识表示学习提供重要的视觉或文本信息，Sun 等人[225]提出多模态知识图谱注意力网络，作为第一个将多模态知识图引入推荐系统的工作其包含一个多模态知识图谱嵌入模块和推荐模块，以协作知识图谱作为输入,通过多模态知识图实体编码器和多模态知识图谱注意层为每个实体学习一个新的实体表示,通过聚合实体的邻居的信息，同时保留自身的信息以表示知识推理关系。根据传统的推荐模型生成用户与项目之间的匹配分数，多模态知识图谱在推荐效果上较单模态知识图谱有更好的表现。

② 跨模态检索

跨模态检索（Cross-modal Retrieval）是指利用一种模态的数据作为查询，去检索另一种模态的数据，其核心目标是打破不同模态（如文本、图像、音频、视频）之间的语义壁垒，在统一的语义空间中实现跨模态的相似性计算与关联发现。例如使用文本去检索相关图片或视频。跨媒体检索能够打破检索结果的媒体限制，从而增强搜索体验和结果的全面性。多模态知识图谱在跨模态检索工作中有较大的应用前景，对于文本检索,输出结果中能够呈现与关键词相关的视觉等其他模态信息，能够有效帮助用户进行实体识别与消歧[226]。对于图像检索，通过一个特征提取模块对检索内容进行特征提取和编码，目标识别和视觉问答工作也可以从中受益。

③ 人机交互

利用多模态知识图谱融合不同模态数据的特性，可以推动知识驱动的人机交互。通过人机交互界面与计算机系统进行交流和操作，在实际场景中，通过多传感器的使用，机器能够感知到多模态、数字化的世界，借助多模态知识图谱作为背景知识，有助于机器加强对真实场景的理解，做出更令人舒适、更自然的反馈。如通过分析人的语言和面部表情数据对使用者进行情感分析，从而调整环境灯光舒适度等。原来的人机交互接入信息更多是从文本、页面中获得，多模态技术会带来新的内容形态，通过听觉和视觉等综合作用，在未来强调沉浸感的人机交互中发挥重要作用。

④跨模态数据管理

诸如 ImageNet 是根据 WordNet 层次结构组织的图像数据集，将大量带有标注信息的图片数据以图数据库的形式对图像数据进行管理。多模态知识图谱形式的跨模态数据管理系统能够将跨模态数据以属性特征或实体的形式表示在知识图谱的结构中，不但能够对跨模态数据进行标准化管理，还能够充分利用跨模态数据之间的结构特点，挖掘隐藏关系。例如在金融证券领域,将不同模态数据进行整合管理后能够在最终控制人识别中发挥重要作用，保障信用风险评估。另外，多模态知识图谱的结构也能支撑跨模态数据在推理性问题上的应用。

4.4.2 多模态多跳推理问答

多模态多跳推理问答（Multimodal Multi-hop Reasoning Question Answering）作为人工智能的核心挑战性任务，要求模型深度整合图像、文本、表格等多源异构信息，通过多步骤推理跨越分散证据片段生成准确答案。该任务需攻克三大核心难题：跨模态语义对齐（如建立文本描述与图像区域的细粒度关联）、动态多跳路径构建（连接异构证据形成逻辑链）以及可解释性保障（显式追踪推理过程）。当前主流技术聚焦两大路线：模块化推理模型通过问题分解实现透明化推理，Yang 等人[227]开发的 MMH—GNN 将图文证据建模为异构图，利用跨模态消息传递在 WebQA 数据集实现 F1 提升；Chen 等人[228]提出的 DEVIANT 则定义 7 类原子操作（过滤/对齐/计算等），通过动态编程生成可执行程序，在金融数据集 TAT-DQA 上准确率提升显著。此类方法虽具高可解释性与低资源消耗优势，但依赖人工规则导致泛化受限。相较而言，百亿参数级视觉语言模型（LVLM）采用端到端范式实现隐式推理，Zhang 等人[229]设计的 LLaVAR 利用 LoRA 适配器进行高效微调，在 12 个任务中取得了显著的性能提升。尽管进展显著，多模态多跳问答进展仍面临黑盒推理难以追溯错误、千亿级计算成本等瓶颈。

（1）多模态多跳推理问答的核心难题

①跨模态语义对齐

跨模态语义对齐旨在将来自不同模态的异构表征映射到一个共享的语义空

间，使得语义等价但模态不同的样本在该空间中距离最近。其核心机制包括：共享空间映射：通过对比损失或双向 Transformer 把图像、文本、音频等特征投影到同一向量空间；细粒度交互：利用注意力或图神经网络在“词-区域”“帧—音素”等局部单元之间建立对应；语义一致性约束：引入模态不变损失、因果约束或生成-判别联合目标，缓解模态间的分布鸿沟。进行模态对齐后，可以用一种模态查询另一种模态的内容。可以为视觉内容自动生成文本描述，或根据文本生成图像、视频。此外，跨模态语义对齐在机器人、自动驾驶等场景中实现语言指令与视觉环境的精准匹配，从而支持高层决策。跨模态语义对齐被视为多模态智能的“操作系统”。一旦对齐误差累积，检索会返回无关内容，生成会出现“幻觉”，决策会导致灾难性后果。当前跨模态语义对齐的挑战集主要有：

1) 异构粒度：像素级细节与抽象概念天然错位，难以一一对应。

2) 模态漂移：传感器差异、语言演变导致分布持续变化，静态对齐易失效。

3) 标注瓶颈：细粒度“区域——短语”对齐标注稀缺且昂贵，限制监督信号。

攻克上述这些挑战可以推动零样本、少样本场景下的多模态理解最终实现“任意模态输入、统一语义理解”的下一代人工智能系统。

② 动态多跳路径构建

动态多跳路径构建（Dynamic Multi-hop Path Construction）旨在开放、异构、持续演化的知识空间中，为任意给定的查询实时检索并链接多条语义路径，使系统能够跨实体、跨关系、跨模态地完成“由远及近”的逐步推理。其核心机制包括: 动态子图采样：基于时间衰减、置信度或用户反馈实时剪枝与扩展，确保路径始终落在“当前最优”子图内；多步语义路由：利用强化学习或可微分搜索在连续决策空间中生成“实体-关系-实体”链，每一步都依据前一步的中间表征动态调整策略。动态多跳路径构建被视为可解释多模态推理的“高速公路”。一旦路径偏离，检索会陷入循环，问答会产生幻觉，决策会放大风险。当前动态多跳路径构建的挑战主要是：

1)**实时演化**：知识图谱或环境状态每分钟更新，路径需在线重算。

2)**组合爆炸**：每多一跳候选指数级增长，需在精度与效率间折中。

3)**可信度衰减**：随着链的增长，单跳误差随之累积，最终答案漂移或自相矛盾。

③ 可解释性保障

可解释性保障（Explainability Assurance）是指在跨模态、多跳推理链路中，为每一步决策提供人类可理解、机器可验证、系统可追溯的因果或逻辑证据。其核心机制包括：证据链可视化，通过热图、注意力权重或概念激活映射，将模型

内部表征映射回原始像素、词元或图谱节点；反事实探针，利用最小扰动生成近邻负例，验证当关键证据缺失时模型输出是否发生预期变化；一致性约束，引入逻辑规则、知识图谱或程序合成器，在多跳路径的每一跳施加可验证的语义一致性损失，阻止漂移。可解释性保障被视为可信多模态 AI 的“安全阀”。一旦解释失真就会导致无法校正错误，监管无法追责，系统无法持续改进。当前可解释性的挑战主要是：

1)**认知差距**：高维连续表征与人类离散符号逻辑难以互译。

2)**评估缺失**：缺乏统一指标衡量，即解释是否完整、易懂。

3)**对抗扰动**：微小输入扰动即可破坏已有解释，导致信任崩塌。

若上述挑战得以解决，系统将在高风险场景中提供可审计、可复现的决策依据，为监管追责、责任界定及人机协同奠定技术基础。

（2）多模态事实核查

多模态事实核查（Multimodal Fact-Checking, MFC）指利用文本、图像、视频、音频等多种模态信息，对声明、证据进行真实性判定的自动化流程。在社交媒体与新闻平台，虚假信息常以图文混排、短视频或音频剪辑形式出现，单一模态的自动核查已无法满足需求。因此多模态事实核查成为研究热点，其目标是在文本、图像、视频、音频等多源异构数据中检索相关证据，并对给定声明的真实性做出可解释判定。典型任务包括检测声明、检索跨模态证据、判断真实性并生成可解释依据。多模态事实核查通常包含三步：

1)**声明检测**：从帖子、演讲或视频中定位需要核查的断言；

2)**证据检索**：在跨模态知识库或开放网络中召回与之相关的图文、音视频片段；

3)**裁决与解释**：输出真、假或无法确定标签，并附带篡改检测结果、跨模态不一致描述及可追溯的推理链。

MFC 可用于社交媒体谣言阻断、新闻机构审校、金融与司法风控，其核心价值是降低人工核查成本、缩短响应时间并提供可追溯的决策理由。MULLER-BUDACK 等人[229]先用命名实体链接抽取文本实体，再反向图像搜索获取参考图片，最后计算与原图的视觉相似度完成判定；该方案不依赖训练数据即可检测“脱离上下文”误导。Chen 等人[231]提出“动态感知器+强化学习蕴含目标”的摘要框架，从任意长度多模态文档中提取声明相关证据摘要；在 MOCHEG 数据集上的真实性分类正确率提高。Papadopoulos S 等人[232]提出的 RED-DOT 引入“相关证据检测”子任务，通过“证据重排序+模态融合+RED 模块”在 NewsCLIPings 上准确率实现提升，并在分布外数据 VERITE 上保持鲁棒性。然而多模态事实核查仍然像素级篡改与抽象文本声明难以一一对应、高质量标注稀缺、生成式伪造技术持续升级等挑战。

（3）多模态时序推理

多模态时序推理（MTR）旨在处理视频、传感器时序流、连续音频等动态数据中的时间维度关联，通过跨帧与跨模态协同推理，解决依赖时间演变的复杂问题。与仅关注空间语义的静态任务不同，MTR 必须显式建模何时发生以及如何演变。其核心任务可归纳为三点：一是时序对齐，即把不同采样频率、不同延迟的异构模态数据统一到一条连续时间轴；二是动态关系建模，通过捕捉跨模态事件之间的时间因果链，揭示“原因-滞后-结果”的演化规律；三是时间敏感决策，既包括对未来状态的提前预测（如轴承剩余寿命估计），也包括对过去事件链的事后诊断（如事故根因溯源）。

多模态时序推理面临三项特有挑战。首先是时间漂移：摄像头、麦克风、振动传感器往往存在毫秒级甚至秒级延迟，若没有准确对齐，后续推理将建立在错误的时间基准上。其次是长程依赖衰减：当事件间隔超过 60 秒时，传统 RNN 或 Transformer 的注意力权重显著衰减，导致关联精度骤降。最后是瞬时动作捕捉：工业场景中的阀门开闭、设备跳闸等动作持续时间极短，易被降采样或滑动窗口遗漏。

为应对上述挑战，当前主流方案采用“分层时序建模”框架。底层特征提取阶段，视频流由 3D-CNN（如 I3D）抽取时空立方体特征如 Carreira，J 等人[233]提出的方法；传感器数据经 LSTM—CRF 联合建模，捕获长短期混合模式；音频信号则通过 Mel 频谱图与 Transformer 编码事件边界。中层跨模态对齐利用 CTSA 算法构建时间状态机，将文本事件（如“阀门关闭”）绑定至视频第 N 帧及传感器峰值时刻，对齐误差低于 0.2 秒，同时借助动态时间规整（DTW）消除模态间延迟。如 Google Research 提出[234]顶层时间推理引擎执行符号化规则。

（4）多证据综合的多跳推理问答

多证据综合的多跳推理问答要求系统从分散在不同文档、不同模态的若干证据片段中，通过链式或图式推理组合信息，最终生成答案并给出可解释路径。该范式突破了单跳“检索-阅读”假设，能够回答诸如“X 公司与 Y 技术在 Z 领域的首次合作由谁促成？”这类需要跨段落、跨媒体验证的复杂问题。其核心价值体现在：提升覆盖率，通过跨源证据互补，显著降低因单文档缺失导致的无法回答比例；增强可解释性，显式输出“实体-关系-实体”链，以便于人工复核；支撑高阶决策，在医疗、金融、司法等高风险场景中，为后续决策提供可追溯依据。Xu H 等人[235]提出的 LOG 创新性地提出了“逐点条件 V 信息”（Pointwise Conditional V-Information）指标旨在量化单篇候选文档片段对最终预测答案的信息贡献度。该模型采用“局部-全局”联合优化损失函数：首先对候选证据进行细粒度的贡献度打分（局部），然后在构建的证据关系图上进行可微分推理（全局）。在 HotpotQA-full[236]数据集上，LOG 显著提升了模型的可解释性。多证据综合的

多跳推理问答已演化为“检索-推理-解释”三元一体的体系。LOG 与的最新结果表明，局部贡献度量与全局一致性约束是提升准确率与可解释性的关键。

LOG 的成果地表明，对局部证据贡献的精细度量与对全局推理路径一致性的强约束是同时提升多证据综合的多跳推理问答任务准确率和可解释性的两大关键驱动力。尽管进展显著，多证据综合的多跳推理问答仍面临诸多挑战：复杂推理效率，多跳、多证据推理的计算开销较大，优化推理效率是实际应用的关键。多模态深度融合：如何有效对齐、融合和理解来自文本、表格、图像等不同模态的证据，并在此基础上进行连贯推理。长程依赖与噪声鲁棒性，处理超长上下文中的长程依赖关系，以及有效过滤检索结果中的噪声和无关证据。可解释性的深度与形式，如何生成更自然、更深入、且能被不同领域用户理解的可解释路径（如自然语言叙述、交互式可视化）；领域自适应与少样本学习，如何将模型高效地迁移到特定专业领域（如生物医学、法律），或在标注数据稀缺的情况下保持良好性能。

多模态多跳推理问答的发展正推动着问答系统向更智能、更可靠、更具洞察力的方向迈进，其应用潜力将在知识密集型和高风险决策领域持续释放。随着技术的成熟，多模态多跳推理问答将成为知识密集型和高风险决策领域的核心支撑技术，推动人工智能从感知智能向认知智能的深层突破。

4.4.3 小结

多模态语义关联技术通过融合文本、视觉、听觉及时序动态数据，驱动语言智能从符号认知向跨模态具身认知跃迁。其核心突破体现为三方面：一是构建多模态知识图谱（MMKG）突破单模态语义局限，解决实体歧义及感知缺失；二是跨模态对齐与推理技术创新，依托注意力机制、动态时间规整（DTW）将事件绑定误差压缩，结合符号规则（Time-Chain）与生成式协同实现 5 跳推理高准确率（较非时序模型提升明显）；三是工业质检、医疗诊断等场景落地，通过时序因果诊断、ECG 波形与病理报告互证显著提升决策可靠性。

未来需攻克三大方向：认知深度上，结合 GPT-4V/Sora 实现动态环境感知，推动 MMKG 升级为“感知-推理-决策”一体引擎；技术融合上，设计轻量适配器支持实时知识更新，引入思维链（CoT）增强跨模态因果推断可解释性；可信保障上，构建涵盖模态均衡性、鲁棒性的新评估体系，以 MMKG 约束 LLM 幻觉并防御生成式伪造攻击。核心挑战：模态失衡（90%模型依赖文本编码）、动态性缺失、评估缺陷，这些挑战需通过算法、数据、评估协同破解，终极目标是实现“任意模态输入，统一语义理解”的可靠认知底座。

第五章　总结与展望

当前，语言智能正处于从感知理解向认知可信演进的关键转型阶段，其发展主要体现在三大融合性突破：在认知深度层面，算法突破字面理解的局限，构建了知识-语境协同推理能力，实现了对世界知识、文化背景及长链条逻辑的有机调用；同时，多模态技术构建了文本、语音、图像、动作及脑机信号的统一表示框架，为具身认知奠定了坚实基础。在垂直场景应用层面，教育智能领域通过作文批改技术实现了从篇章要素抽取、跨提示评分到细粒度可解释反馈的全链条优化；儿童语言能力评价建立了多维度量化框架，推动评价模式由“结果评价”向“过程指导”转变；跨语言服务领域攻克了东南亚低资源机器翻译中的语料匮乏与文化适配难题，为“一带一路”信息互联互通提供支撑；社会治理领域依托在反讽识别、隐式情绪挖掘等方向形成的负面情感分析多维技术体系，为公共安全与金融风控提供决策支持；创新管理领域推动专利价值评估向动态化、智能化演进，激活科技成果转化生态。在可信机制层面，联邦学习、差分隐私、对抗鲁棒等技术被嵌入“数据-部署-监控”全生命周期，构建了能力提升与安全保障同步强化的安全体系。

展望未来，语言智能将围绕认知深化、场景贯通、可信进化及人本协同四大战略方向加速发展：在认知智能升维方面，融合大语言模型（如 GPT-4V/Sora 的动态感知能力）、图神经网络与符号规则，构建“感知-推理-决策”一体化引擎，支持 5 跳以上的复杂推理；同时研发轻量适配器以实现多模态知识图谱的实时更新，借助思维链增强跨模态因果推断的可解释性，以解决模态失衡与动态性缺失的瓶颈问题。在垂直场景闭环革新方面，教育智能领域将推动作文批改融合零样本大模型以降低标注依赖，构建“写作-评价-反馈”全流程辅助系统；儿童语言评价将结合多模态纵向追踪形成“评估-干预-再评估”闭环；低资源翻译将聚焦老挝语/高棉语等语料建设，通过多模态融合缓解非拉丁文字采集难题，并借助文化知识图谱消除“直译冲突”；情感计算领域将发展跨文化隐式情感分析框架，结合外部知识提升低资源场景性能并构建实时处理系统；创新管理则整合技术生命周期参数以建立动态专利评估模型。在可信 AI 体系深化方面，将革新评估机制，构建涵盖模态均衡性与鲁棒性的多模态评估体系，利用多模态知识图谱约束大模型幻觉并防御生成式伪造攻击；同时在差分隐私与零知识证明基础上开发文化偏见动态监测工具，确保能力与安全的同步进化。在人本化生态构建方面，将通过可解释评分弥合教育鸿沟并助力特殊儿童干预，以东南亚翻译为支点推动“一带一路”文明互鉴，并将负面情感分析升级为跨文化智能决策系统以支撑社会治理现代化。最终目标是构建具备“任意模态输入，统一语义理解”能力的可靠认知底座，实现智能服务于人、赋能于人、可信于人的可持续发展生态。

语言智能发展的下一阶段，将深刻体现为技术理性与人本价值的深度协同，

即人工智能的算法逻辑与人类伦理需求的紧密结合，共同推动智能系统的进步。这一协同不仅涉及数据处理的高效性，更强调在决策过程中融入人文关怀，确保技术服务于人类福祉。唯有依托认知突破的基础，包括对语言理解、推理能力的创新提升；可信机制的保障，如透明算法、数据隐私保护及安全框架的建立；以及场景需求的驱动，即实际应用中如教育、医疗等领域的迫切问题解决，方能稳步实现人类与机器和谐共生的智能文明愿景。该愿景旨在构建一个智能无处不在但以人为本的未来社会，其中机器作为辅助伙伴，而非替代者，共同促进文明的可持续发展。

第四章参考文献

[1] Song W, Song Z, Fu R, et al (2020) Discourse self-attention for discourse element identification in argumentative student essays. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 2820-2830.

[2] Yang Z, Yang D, Dyer C, et al (2016) Hierarchical attention networks for document classification. In: Knight K, Nenkova A, Rambow O (eds) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp 1480-1489.

[3] Eger S, Daxenberger J, Gurevych I (2017) Neural end-to-end learning for computational argumentation mining. In: Barzilay R, Kan MY (eds) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, pp 11–22, https://doi.org/10.18653/v1/P17-1002.

[4] Peng N, Poon H, Quirk C et al (2017) Cross-sentence n-ary relation extraction with graph lstms. Transactions of the Association for Computational Linguistics 5:101-115.

[5] Wen J (2017) Structure regularized bidirectional recurrent convolutional neural network for relation classification. arXiv preprint arXiv:1711.02509

[6] Lan M, Wang J, Wu Y, et al (2017) Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 1299-1308.

[7] Dai Z, Huang R (2018) Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. arXiv preprint arXiv:1804.05918.

[8] Guo F, He R, Dang J (2020) Implicit discourse relation recognition based on contextual interaction perception and pattern filtering. Journal of Computer Science 43:901-915.

[9] Chan YS, Roth D (2011) Exploiting syntactico-semantic structures for relation extraction. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 551-560

[10] Lin Y, Shen S, Liu Z, et al (2016) Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2124-2133.

[11] Miwa M, Sætre R, Miyao Y, et al (2009) A rich feature vector for protein-protein interaction extraction from multiple corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp 121-130.

[12] Wresch W. The imminence of grading Essays by computer--25 Years Later[J]. Computers and composition, 1993, 10(2): 45-58.

[13] HELEN B. AJAY P I T, PAGE E B. Analysis of Essays by Computer (AEC-II). Final Report[C]. U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, D.C., Tech. Rep.. 1973.

[14] Alikaniotis D, Yannakoudakis H, Rei M. Automatic Text Scoring Using Neural Networks[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 715-725.

[15] Taghipour K, Ng H T. A neural approach to automated essay scoring[C]. Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1882-1891.

[16] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[17] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[18] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer[J]. arXiv preprint arXiv:2001.04451, 2020.

[19] Beltagy I, Peters M E, Cohan A. Longformer: The long-document transformer[J]. arXiv preprint arXiv:2004.05150, 2020.

[20] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[21] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for

language understanding[J]. Advances in neural information processing systems, 2019, 32.

[22] Yang R, Cao J, Wen Z, et al. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking[C]. Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1560-1569.

[23] Wang Y, Wang C, Li R, et al. On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation[C]. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 3416-3425.

[24] Wang J, Zhang Q, Liu J, et al. Making meta-learning solve cross-prompt automatic essay scoring[J]. Expert Systems with Applications, 2025, 272: 126710.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135-1144.

[26] Ridley R, He L, Dai X, et al. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring[J]. arXiv preprint arXiv:2008.01441, 2020.

[27] Ridley R, He L, Dai X, et al. Automated cross-prompt scoring of essay traits[C]. Proceedings of the AAAI conference on artificial intelligence. 2021, 35(15): 13745-13753.

[28] Sun J, Zhang Q, Liu J, et al. AES Model with Logic Rule Reasoning and Self-Explanation Based on AMR and LLM[C]//2024 4th International Conference on Robotics, Automation and Intelligent Control (ICRAIC). IEEE, 2024: 125-129.

[29] McCabe, Allyssa, et al. "Comparison of personal versus fictional narratives of children with language impairment." *American Journal of Speech-Language Pathology* 17.2 (2008): 194-206.

[30] Schneider P, Hayward D. Who does what to whom: Introduction of referents in children's storytelling from pictures[J]. Language, Speech, and Hearing Services in Schools, 2010, 41(4): 459-473.

[31] Chang C. Linking early narrative skill to later language and reading ability in Mandarin-speaking children: A longitudinal study over eight years[J]. Narrative Inquiry, 2006, 16(2): 275-293.

[32] 曾维秀，李甦. 儿童叙事能力发展的促进与干预研究综述[J].中国心理卫生杂志, 2006, 20(9): 572-575.

[33] 李甦，赵静. 汉语普通话儿童叙事能力的发展[C]. 第十一届全国心理学学术会议. 2007.

[34] Pesco, Diane, and Elizabeth Kay-Raining Bird. "Perspectives on bilingual children's narratives elicited with the Multilingual Assessment Instrument for Narratives." *Applied psycholinguistics* 37.1 (2016): 1-9.

[35] Pico, Danielle L., et al. "Interventions designed to improve narrative language in school-age children: A systematic review with meta-analyses." *Language, Speech, and Hearing Services in Schools* 52.4 (2021): 1109-1126.

[36] Blom, Elma, and Tessel Boerma. "Why do children with language impairment have difficulties with narrative macrostructure?." *Research in developmental disabilities* 55 (2016): 301-311.

[37] Justice, Laura M., et al. "The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances." *American Journal of Speech-Language Pathology* 15.2 (2006): 177-191.

[38] Xue, Jin, et al. "Characterizing macro-and micro-structures of narrative skills for Mandarin-speaking school-age children with specific language impairment." *Journal of communication disorders* 96 (2022): 106199.

[39] Stein, Nancy L., and Christine G. Glenn. "An Analysis of Story Comprehension in Elementary School Children: A Test of a Schema." (1975).

[40] Reese, Elaine, et al. "Coherence of personal narratives across the lifespan: A multidimensional model and coding method." *Journal of cognition and development* 12.4 (2011): 424-462.

[41] Kellas, Jody Koenig, and Valerie Manusov. "What's in a story? The relationship between narrative completeness and adjustment to relationship dissolution." *Journal of Social and Personal Relationships* 20.3 (2003): 285-307.

[42] LEE, GARY GEUNBAE, Heejin Do, and Yunsu Kim. "Autoregressive Score Generation for Multi-trait Essay Scoring." Association for Computational Linguistics (ACL), 2024.

[43] Heejin Do, Sangwon Ryu, and Gary Lee. 2024. Autoregressive Multi-trait Essay Scoring via Reinforcement Learning with Scoring-aware Multiple Rewards. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

[44] Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2013. Using Latent Dirichlet Allocation for Child Narrative Analysis. In Proceedings of the 2013

Workshop on Biomedical Natural Language Processing, pages 111–115, Sofia, Bulgaria. Association for Computational Linguistics.

[45] Jones, Sharad, et al. "An exploration of automated narrative analysis via machine learning." *Plos one* 14.10 (2019): e0224634.

[46] Torng, Pao-Chuan, and Wen-Hui Sah. "Narrative abilities of Mandarin-speaking children with and without specific language impairment: Macrostructure and microstructure." *Clinical linguistics & phonetics* 34.5 (2020): 453-478.

[47] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

[48] Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." *arXiv preprint arXiv:2402.03300* (2024).

[49] Jacobs, Robert A., et al. "Adaptive mixtures of local experts." *Neural computation* 3.1 (1991): 79-87.

[50] Pearce, Wendy M., Deborah GH James, and Paul F. McCormack. "A comparison of oral narratives in children with specific language and non-specific language impairment." *Clinical Linguistics & Phonetics* 24.8 (2010): 622-645.

[51] MacWhinney, Brian. *The CHILDES project: The database*. Vol. 2. Psychology Press, 2000.

[52] MacWhinney, Brian. "Tools for analyzing talk part 1: The chat transcription format." *Carnegie.[Google Scholar]* 16 (2017).

[53] KOEHN P. Statistical machine translation[M]. Cambridge: Cambridge University Press, 2010.

[54] JURAFSKY D, MARTIN J H. Speech and language processing[M]. 3rd ed. London: Pearson Education, 2019.

[55] BROWN P F, et al. A statistical approach to machine translation[M]//Readings in machine translation. 2018.

[56] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR) - Conference Track. 2015.

[57] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). 2017: 5998-6008.

[58] CONNEAU A, KHANDELWAL K, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,

2020: 8440-8451.
[59] NLLB Team, et al. No language left behind: Scaling human-centered machine translation[J]. arXiv preprint, 2022, abs/2207.04672.
[60] XU H, KIM Y J, SHARAF A, et al. A paradigm shift in machine translation: Boosting translation performance of large language models[J]. arXiv preprint, 2024, abs/2309.11674.
[61] ZHU S, DONG T, LI B, et al. FuxiMT: Sparsifying large language models for Chinese-centric multilingual machine translation[J]. arXiv preprint, 2025, abs/2505.14256.
[62] GALE WA,CHURCH K W.A program for aligning entences in bilingual corpora[C]//Proceedings of the 29th AnnualMeeting on Association for Computational Linguistics ,3-21 June 1991 , Berkeley , California. Morristown , NJ, USA : ACL,1991:177 -184.
[63] THOMPSON B, KOEHN P. Vecalign: Improved sentence alignment in linear time and space [C]//Proceedings of the 2019Conference on Empirical Methods in Natural Languagerocessing and the 9th International Joint Conference on Natural Language Processing (ENLP -IJCNLP),Hong Kong,Chia.Stroudsburg,PA,USA:ACL,2019:1342-1348
[64] ARTETXE M,SCHWENK H. Massively multilingual sentence embeddings for zero-shot cross - lingual transfer and beyondJ].ArXiv e-Prints,2018 :arXiv:1812.10464
[65] Wu, Jiawei , X. Wang , and W. Y. Wang . "Extract and Edit: An Alternative to Back-Translation for Unsupervised Neural Machine Translation." (2019).
[66] 贾承勋,赖华,余正涛,等,基于枢轴语言的汉越神经机器翻译伪平行语料生成[J].计算机工程与科学,2021,43(3):13, 542-550.
[67] JIA C X, LAI H, YU Z 'T ,et al. Pseudo - parallel corpus generation for Chinese - Vietnamese neural machine translation basedon pivot language[J].Computer Engineering & Science,2021 ,43(3):542-550.
[68]李炫达,周兰江,张建安.融合句子结构特征的汉老双语句子相似度计算方法[J].中文信息学报,2022,36(2):58-68.
[69] LI X D.ZHOU L, J.ZHANG J A. Sentence similarity metirc between Chinese and Laotian based on syntax feature J. Journaof Chinese Information Processing,2022,36(2):58 -68.
[70] 贾承勋,赖华,余正涛,等.融合单语语言模型的汉越伪平行语料生成[J].计算机应用,2021,41(6):1652-1658.
[71]Conneau A, Lample G. Cross-lingual language model pretraining[J]. Advances in

neural information processing systems, 2019, 32.
[72] Aharoni R, Goldberg Y. Towards string-to-tree neural machine translation[J]. arXiv preprint arXiv:1704.04743, 2017.
[73] Zhao Y, Xiang L, Zhu J, et al. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 4495-4505.
[74] Gu J, Hassan H, Devlin J, et al. Universal neural machine translation for extremely low resource languages[J]. arXiv preprint arXiv:1802.05368, 2018.
[75] Li Z, Yang N, Wang L, et al. Learning diverse document representations with deep query interactions for dense retrieval[J]. arXiv preprint arXiv:2208.04232, 2022.
[76] Sang Y, Chen Y, Zhang J. Neural machine translation research on syntactic information fusion based on the field of electrical engineering[J]. Applied Sciences, 2023, 13(23): 12905.
[77]Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1723-1732.
[78]Luong M T, Le Q V, Sutskever I, et al. Multi-task sequence to sequence learning[J]. arXiv preprint arXiv:1511.06114, 2015.
[79]Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism[J]. arXiv preprint arXiv:1601.01073, 2016.
[80]Johnson M, Schuster M, Le Q V, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.
[81] Tan, Xu, et al. Multilingual neural machine translation with knowledge distillation[J]. arXiv preprint arXiv:1902.10461, 2019.
[82] Gu, Jiatao, et al. Meta-Learning for Low-Resource Neural Machine Translation[C[//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: pp. 3622-3631.
[83] Rapid Adaptation of Neural Machine Translation to New Languages[C]//In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 875-880.
[84] Nguyen X P, Joty S, Wu K, et al. Refining low-resource unsupervised translation by language disentanglement of multilingual translation model[J]. Advances in Neural Information Processing Systems, 2022, 35: 36230-36242.

[85] 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.
[86] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.
[87] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.
[88] BERARD A, PIETQUIN O, SERVAN C, et al. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation[J/OL]. arXiv, 2016, abs/1612.01744.
[89] COMMUNICATION S, et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation[J/OL]. arXiv, 2023, abs/2308.11596.
[90] SHANBHOGUE A V K, XUE R, SAHA S, et al. Improving Low Resource Speech Translation with Data Augmentation and Ensemble Strategies[C]//Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 241-250.
[91] NOVITASARI S, TJANDRA A, SAKTI S, et al. Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis[C]//Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). Marseille, France: European Language Resources Association, 2020: 131-138.
[92] WANG T, XU L, LU W, et al. From Tens of Hours to Tens of Thousands: Scaling Back-Translation for Speech Recognition[J/OL]. arXiv, 2025, abs/2505.16972.
[93] DAO A, et al. Speechless: Speech Instruction Training Without Speech for Low Resource Languages[J/OL]. arXiv, 2025, abs/2505.17417.
[94] RADFORD A, KIM J W, XU T, et al. Robust Speech Recognition via Large-Scale Weak Supervision[J/OL]. arXiv, 2022, abs/2212.04356.
[95] COMMUNICATION S, et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation[J/OL]. arXiv, 2023, abs/2308.11596.
[96] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2002: 79-86.
[97] COMMUNICATION S, et al. SeamlessM4T: Massively Multilingual &

Multimodal Machine Translation[J/OL]. arXiv, 2023, abs/2308.11596.
[98] SHANBHOGUE A V K, XUE R, SAHA S, et al. Improving Low Resource Speech Translation with Data Augmentation and Ensemble Strategies[C]//Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 241-250.
[99] NOVITASARI S, TJANDRA A, SAKTI S, et al. Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis[C]//Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). Marseille, France: European Language Resources Association, 2020: 131-138.
[100] WANG T, XU L, LU W, et al. From Tens of Hours to Tens of Thousands: Scaling Back-Translation for Speech Recognition[J/OL]. arXiv, 2025, abs/2505.16972.
[101] DAO A, et al. Speechless: Speech Instruction Training Without Speech for Low Resource Languages[J/OL]. arXiv, 2025, abs/2505.17417.
[102] RADFORD A, KIM J W, XU T, et al. Robust Speech Recognition via Large-Scale Weak Supervision[J/OL]. arXiv, 2022, abs/2212.04356.
[103] COMMUNICATION S, et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation[J/OL]. arXiv, 2023, abs/2308.11596.
[104] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2002: 79-86.
[105] Rozin P, Royzman E B. Negativity bias, negativity dominance, and contagion[J]. Personality and Social Psychology Review, 2001, 5(4): 296-320.
[106] Ekman P. An argument for basic emotions[J]. Cognition & Emotion, 1992, 6(3-4): 169-200.
[107] Wijeratne S, Balasuriya L, Doran D, et al. Word embeddings to enhance Twitter gang member profile identification[C]//Proceedings of the International AAAI Conference on Web and Social Media (ICWSM). Palo Alto: AAAI Press, 2017: 890-893.
[108] Reyes A, Rosso P, Buscaldi D. From humor recognition to irony detection: The figurative language of social media[J]. Data & Knowledge Engineering, 2012, 74: 1-12.
[109] Sundaram V, Pavan R S, Kandaala S, et al. Distinguishing hate speech from sarcasm[C]//2022 International conference for advancement in technology (ICONAT). IEEE, 2022: 1-5.
[110] Joshi A, Sharma V, Bhattacharyya P. Harnessing context incongruity for

sarcasm detection[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015: 757-762.
[111] Wilson D. The pragmatics of verbal irony: Echo or pretence?[J]. Lingua, 2006, 116(10): 1722-1743.
[112] Joshi A, Bhattacharyya P, Carman M J. Automatic sarcasm detection: A survey[J]. ACM Computing Surveys (CSUR), 2017, 50(5): 1-22.
[113] 余本功,李晨越.面向社交媒体的讽刺检测研究综述[J].计算机应用研究,2025,42(04):961-974.
[114] Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 2506-2515.
[115] Zhang Y, Wang J, Liu Y, et al. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations[J]. Information Fusion, 2023, 93: 282-301.
[116] Yang X, Feng S, Wang D, et al. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts[C]//Proceedings of the 31st ACM international conference on multimedia. 2023: 6045-6053.
[117] Castro S, Hazarika D, Pérez-Rosas V, et al. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper) [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 4619–4629.
[118] Lukin S, Walker M. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue[J]. arXiv preprint arXiv:1708.08572, 2017.
[119] Oprea S, Magdy W. isarcasm: A dataset of intended sarcasm[J]. arXiv preprint arXiv:1911.03123, 2019.
[120] Khodak M, Saunshi N, Vodrahalli K. A large self-annotated corpus for sarcasm[J]. arXiv preprint arXiv:1704.05579, 2017.
[121] Van Hee C, Lefever E, Hoste V. Semeval-2018 task 3: Irony detection in english tweets[C]//Proceedings of the 12th international workshop on semantic evaluation. 2018: 39-50.
[122] Garcia M H, Manoel A, Diaz D M, et al. Flute: A scalable, extensible framework for high-performance federated learning simulations[J]. arXiv preprint arXiv:2203.13789, 2022.
[123] Srirag D, Joshi A, Painter J, et al. Besstie: A benchmark for sentiment and sarcasm classification for varieties of english[J]. arXiv preprint arXiv:2412.04726, 2024.
[124] Kim Y, Suh H, Kim M, et al. KoCoSa: Korean context-aware sarcasm

detection dataset[J]. arXiv preprint arXiv:2402.14428, 2024.
[125] Qin L, Huang S, Chen Q, et al. MMSD2. 0: Towards a reliable multi-modal sarcasm detection system[J]. arXiv preprint arXiv:2307.07135, 2023.
[126] Li Z, Zhang Y, Gao X, et al. Leveraging Large Language Models for Sarcastic Speech Annotation in Sarcasm Detection[J]. arXiv preprint arXiv:2506.00955, 2025.
[127] Bharti S K, Babu K S, Jena S K. Parsing-based sarcasm sentiment recognition in twitter data[C]//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015: 1373-1380.
[128] Bouazizi M, Ohtsuki T. Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis[C]//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015: 1594-1597.
[129] Bouazizi M, Ohtsuki T O. A pattern-based approach for sarcasm detection on twitter[J]. IEEE Access, 2016, 4: 5477-5488.
[130] Hazarika D, Poria S, Gorantla S, et al. Cascade: Contextual sarcasm detection in online discussion forums[J]. arXiv preprint arXiv:1805.06413, 2018.
[131] Du Y, Li T, Pathan M S, et al. An effective sarcasm detection approach based on sentimental context and individual expression habits[J]. Cognitive Computation, 2022, 14(1): 78-90.
[132] Potamias R A, Siolas G, Stafylopatis A G. A transformer-based approach to irony and sarcasm detection[J]. Neural Computing and Applications, 2020, 32(23): 17309-17320.
[133] Yue T, Mao R, Wang H, et al. KnowleNet: Knowledge fusion network for multimodal sarcasm detection[J]. Information Fusion, 2023, 100: 101921.
[134] Li D, Li Y, Zhang J, et al. C3kg: A chinese commonsense conversation knowledge graph[J]. arXiv preprint arXiv:2204.02549, 2022.
[135] Ren Y, Wang Z, Peng Q, et al. A knowledge-augmented neural network model for sarcasm detection[J]. Information Processing & Management, 2023, 60(6): 103521.
[136] Pan H, Lin Z, Fu P, et al. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1383-1392.
[137] Alnajjar K, Hämäläinen M. ¡Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline [C]//Proceedings of the Third Workshop on Multimodal Artificial Intelligence. Mexico City: Association for Computational Linguistics, 2021: 63–68.
[138] Chen J, Yu H, Huang S, et al. InterCLIP-MEP: Interactive CLIP and Memory-Enhanced Predictor for Multi-modal Sarcasm Detection[J]. arXiv preprint arXiv:2406.16464, 2024.
[139] Bhosale S, Chaudhuri A, Williams A L R, et al. Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection[J]. arXiv

preprint arXiv:2310.01430, 2023.
[140] Rasheed H, Khattak M U, Maaz M, et al. Fine-tuned clip models are efficient video learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6545-6554.
[141] Zhang Y, Zou C, Lian Z, et al. Sarcasmbench: Towards evaluating large language models on sarcasm understanding[J]. arXiv preprint arXiv:2408.11319, 2024.
[142] Zhang Y, Zou C, Wang B, et al. Commander-GPT: Fully Unleashing the Sarcasm Detection Capability of Multi-Modal Large Language Models[J]. arXiv preprint arXiv:2503.18681, 2025.
[143] Liu Z, Zhou Z, Hu M. CAF-I: A Collaborative Multi-Agent Framework for Enhanced Irony Detection with Large Language Models[J]. arXiv preprint arXiv:2506.08430, 2025.
[144] Jana S, Kundu A, Singh S R. Think Twice Before You Judge: Mixture of Dual Reasoning Experts for Multimodal Sarcasm Detection[J]. arXiv preprint arXiv:2507.04458, 2025.
[145] Bilewicz M, Soral W. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization[J]. Political Psychology, 2020, 41: 3-33.
[146] FISKE S T. Controlling other people: The impact of power on stereotyping[M]//Social cognition. 2018: 101-115.
[147] 王文华. 论反仇恨言论视阈下网络暴力的法律治理[J]. 中国应用法学, 2023(05): 63-75.
[148] 吴颖妍. 互联网仇恨言论的传播特点及其治理探讨[J]. 新闻研究导刊, 2020, 11(24): 57-58.
[149] 赵玉现, 胡春莉. 网络仇恨言论判别与治理探究[J]. 信息安全研究, 2019, 5: 1021-1026.
[150] Nobata C, Tetreault J, Thomas A, et al. Abusive language detection in online user content[C]//Proceedings of the 25th international conference on world wide web. 2016: 145-153.
[151] Unsvåg E F, Gambäck B. The effects of user features on Twitter hate speech detection[C]//Proceedings of the 2nd workshop on abusive language online (ALW2). 2018: 75-85.
[152] Djuric N, Zhou J, Morris R, et al. Hate speech detection with comment embeddings[C]//Proceedings of the 24th international conference on world wide web. 2015: 29-30.
[153] 周险兵, 樊小超, 杨勇, 等. 基于语义拼写理解和门控注意力机制的不良言论检测[J]. 计算机应用与软件, 2024, 41(01): 112-118+125.
[154] Mozafari M, Farahbakhsh R, Crespi N. A BERT-based transfer learning approach for hate speech detection in online social media[C]//International conference on

complex networks and their applications. Cham: Springer International Publishing, 2019: 928-940.
[155] Barbieri F, Camacho-Collados J, Anke L E, et al. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification[J]. Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.
[156] Choudhary M, Agarwal B, Goyal V. Hate Speech Detection: Leveraging LLM-GPT2 with Fine-Tuning and Multi-Shot Techniques[J]. Procedia Computer Science, 2025, 258: 2817-2825.
[157] Wang H, Hee M S, Awal M R, et al. Evaluating GPT-3 generated explanations for hateful content moderation[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023: 6255-6263.
[158] Dixon L, Li J, Sorensen J, et al. Measuring and mitigating unintended bias in text classification[C]//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018: 67-73.
[159] Zhou X, Sap M, Swayamdipta S, et al. Challenges in Automated Debiasing for Toxic Language Detection[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021.
[160] Ramponi A, Tonelli S. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 3027-3040.
[161] Vaidya A, Mai F, Ning Y. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2020, 14: 683-693.
[162] Attanasio G, Nozza D, Hovy D, et al. Entropy-based attention regularization frees unintended bias mitigation from lists[M]//Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022.
[163] Chang S, Zhang Y, Yu M, et al. Invariant rationalization[C]//International Conference on Machine Learning. PMLR, 2020: 1448-1458.
[164] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes[J]. Advances in neural information processing systems, 2020, 33: 2611-2624.
[165] Zhu R. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution[J]. arXiv preprint arXiv:2012.08290, 2020.
[166] Lee R K W, Cao R, Fan Z, et al. Disentangling hate in online memes[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 5138-5147.
[167] Pramanick S, Sharma S, Dimitrov D, et al. MOMENTA: A Multimodal

Framework for Detecting Harmful Memes and Their Targets[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 4439-4455.
[168] Blaier E, Malkiel I, Wolf L. Caption Enriched Samples for Improving Hateful Memes Detection[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9350-9358.
[169] Cao R, Lee R K W, Chong W H, et al. Prompting for Multimodal Hateful Meme Classification[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 321-332.
[170] Ji J, Ren W, Naseem U. Identifying creative harmful memes via prompt based approach[C]//Proceedings of the ACM web conference 2023. 2023: 3868-3872.
[171] Lin H, Luo Z, Ma J, et al. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 9114-9128.
[172] Lin H, Luo Z, Gao W, et al. Towards explainable harmful meme detection through multimodal debate between large language models[C]//Proceedings of the ACM Web Conference 2024. 2024: 2359-2370.
[173] Hee M S, Lee R K W, Chong W H. On explaining multimodal hateful meme detection models[C]//Proceedings of the ACM web conference 2022. 2022: 3651-3655.
[174] Hee M S, Chong W H, Lee R K W. Decoding the underlying meaning of multimodal hateful memes[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023: 5995-6003.
[175] Tahir W B, Khalid S, Almutairi S, et al. Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques[J]. IEEE Access, 2025.
[176] Magami F, Digiampietri L A. Automatic detection of depression from text data: A systematic literacture review[C]//Proceedings of the XVI Brazilian Symposium on Information Systems. 2020: 1-8.
[177] Gan L, Huang Y, Gao X, et al. Multimodal Magic Elevating Depression Detection with a Fusion of Text and Audio Intelligence[J]. arXiv preprint arXiv:2501.16813, 2025.
[178] Tank C, Pol S, Katoch V, et al. Depression detection and analysis using large language models on textual and audio-visual modalities[J]. arXiv preprint arXiv:2407.06125, 2024.
[179] Bao E, Pérez A, Parapar J. Explainable depression symptom detection in social media[J]. Health Information Science and Systems, 2024, 12(1): 47.
[180] Qin W, Chen Z, Wang L, et al. Read, diagnose and chat: Towards explainable and interactive LLMs-augmented depression detection in social media[J]. arXiv preprint arXiv:2305.05138, 2023.
[181] Akyol S. New chaos-integrated improved grey wolf optimization based models for automatic detection of depression in online social media and networks[J]. PeerJ Computer Science, 2023, 9: e1661.

[182] Skaik R, Inkpen D. Using twitter social media for depression detection in the canadian population[C]//Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference. 2020: 109-114.
[183] Hussain J, Satti F A, Afzal M, et al. Exploring the dominant features of social media for depression detection[J]. Journal of Information Science, 2020, 46(6): 739-759.
[184] Zhou S, Mohd M. Mental Health Safety and Depression Detection in Social Media Text Data: A Classification Approach Based on a Deep Learning Model[J]. IEEE Access, 2025.
[185] Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(1): 1-20.
[186] Trotzek M, Koitka S, Friedrich C M. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(3): 588-601.
[187] Verma S, Joshi R C, Dutta M K, et al. AI-enhanced mental health diagnosis: leveraging transformers for early detection of depression tendency in textual data[C]//2023 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). IEEE, 2023: 56-61.
[188] Teck Kiong Y. An initial study of depression detection on mandarin textual through BERT model[C]//Proceedings of the 14th ACM Web Science Conference 2022. 2022: 459-463.
[189] Shah S M, Gillani S A, Baig M S A, et al. Advancing depression detection on social media platforms through fine-tuned large language models[J]. Online Social Networks and Media, 2025, 46: 100311.
[190] Shen Y, Yang H, Lin L. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6247-6251.
[191] Huang K, Lu H, Li J. Textual-dominated Multimodal Depression Detection[C]//2024 30th International Conference on Mechatronics and Machine Vision in Practice (M2VIP). IEEE, 2024: 1-6.
[192] Tank C, Pol S, Katoch V, et al. Depression detection and analysis using large language models on textual and audio-visual modalities[J]. arXiv preprint arXiv:2407.06125, 2024.
[193] Hu P, Lin C, Li J, et al. Making the Implicit Explicit: Depression Detection in Web across Posted Texts and Images[C]//2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023: 4807-4811.
[194] Zhang W, Xie J, Zhang Z, et al. Depression detection using digital traces on social

media: A knowledge-aware deep learning approach[J]. Journal of Management Information Systems, 2024, 41(2): 546-580.
[195] Lan X, Cheng Y, Sheng L, et al. Depression detection on social media with large language models[J]. arXiv preprint arXiv:2403.10750, 2024.

[196] Meng W, Guilin Q, Haofen W. Richpedia: a comprehensive multi-modal knowledge graph [C]// Proceedings of Joint International Semantic Technology Conference. Cham: Springer. 2019: 130-145.
[197] Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE transactions on neural networks and learning systems, 2021, 33(2): 494-514.
[198] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2016, 8(3): 489-508.
[199] Zhu X, Li Z, Wang X, et al. Multi-modal knowledge graph construction and application: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 36(2): 715-735.
[200] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 1003-1011.
[201] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.
[202] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. Cham: Springer International Publishing, 2018: 593-607.
[203] Lee J, Toutanova K. Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018, 3(8): 4171-4186.
[204] Wu, Y., He, X., Zhao, D., Wang, L., & Zhou, M. (2019, July). BERT for relation extraction: An adversarial training approach. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2215-2224).
[205] Singhal, A. (2012, October). Introducing the Knowledge Graph: things, not strings. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 1-2).

[206] Rong Z, Yuan L, Yang L. Enhanced knowledge graph recommendation algorithm based on multi-level contrastive learning[J]. Scientific Reports, 2024, 14(1): 23051.
[207] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
[208] Liu W, Zhou P, Zhao Z, et al. K-bert: Enabling language representation with knowledge graph[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(03): 2901-2908.
[209] Chen, Y., Liu, Z., Zhang, X., & Sun, M. (2022). Visual knowledge graph construction for cultural heritage digitalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5794-5809.
[210] Chen Z, Chen J, Zhang W, et al. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid[C]//Proceedings of the 31st ACM international conference on multimedia. 2023: 3317-3327.
[211] Kao K C. Enhancing CLIP Conceptual Embedding through Knowledge Distillation[J]. arXiv preprint arXiv:2412.03513, 2024.
[212] Li M, Xu R, Wang S, et al. Clip-event: Connecting text and images with event structures[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16420-16429.
[213] Sun C, Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 7464-7473.
[214] Yu F, Tang J, Yin W, et al. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3208-3216.
[215] Li W, Gao C, Niu G, et al. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning[J]. arXiv preprint arXiv:2012.15409, 2020.
[216] 王海荣,徐 玺,王 彤,等.多模态命名实体识别方法研究进展[J].郑州大学学报(工学版),2024,45(02):60-71.
[217] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts[J]. arXiv preprint arXiv:1802.07862, 2018.
[218] Zhang, X., Liu, Y., Sun, M., et al. (2021). UMT: Unifying multi-modal transformers for multi-modal named entity recognition. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th

International Joint Conference on Natural Language Processing (Volume 1: Long Papers).
[219] Wang, X., Liu, T., Li, H., et al. (2022). An Alignment and Matching Network with Hierarchical Visual Features for Multimodal Named Entity and Relation Extraction. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 546-555). ACM
[220] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
[221] Ferrada S, Bustos B, Hogan A. IMGpedia: a linked dataset with content-based analysis of Wikimedia images[C]//International Semantic Web Conference. Cham: Springer International Publishing, 2017: 84-93.
[222] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
[223] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[224] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.Zhang G, Jiang C, Guan Z, et al. Multimodal entity linking with mixed fusion mechanism[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer Nature Switzerland, 2023: 607-622.
[225] Sun, F., Huang, X., Zhang, X., & Zhang, Y. (2021). MKGAT: Multimodal Knowledge Graph Attention Network for Recommendation. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 125-134.
[226] Li, X., Zhang, Y., Wang, H., & Zhao, D. (2023). Knowledge Graph-Enhanced Multimodal Transformer for Image-Text Retrieval. IEEE Transactions on Multimedia, 25, 4042-4055.
[227] Yang, Z., Li, X., Wang, H., et al. (2022). MMH-GNN: Multi-modal Multi-hop Graph Neural Network for Multi-modal Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12345-12354).

[228] Chen, M., Liu, Y., Sun, M., et al. (2023). DEVIANT: Dynamic Programming for Multi-modal Multi-hop Reasoning with Atomic Operations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 5678-5688). Association for Computational Linguistics.
[229] Zhang Y, Zhang R, Gu J, et al. Llavar: Enhanced visual instruction tuning for text-rich image understanding[J]. arXiv preprint arXiv:2306.17107, 2023.
[230] Muller-Budack, E., Kiesel, J., & Gurevych, I. (2020). Multimodal fact-checking of images and text. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6469-6483).
[231] Chen, T. -C., Tang, C. -W., & Thomas, C. (2024). MetaSumPerceiver: Multimodal Multi-Document Evidence Summarization for Fact-Checking. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (pp. 1234-1245). Association for Computational Linguistics.
[232] Papadopoulos S I, Koutlis C, Papadopoulos S, et al. Red-dot: Multimodal fact-checking via relevant evidence detection[J]. IEEE Transactions on Computational Social Systems, 2025.
[233] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6299-6308).
[234] Google Research. (2023). Time-Chain: Symbolic temporal reasoning for multimodal sequences (Technical Report). Google Research.
[235] Xu H, Zhao Y, Zhang J, et al. LOG: A Local-to-Global Optimization Approach for Retrieval-based Explainable Multi-Hop Question Answering[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 9085-9095.
[236] Yang Z, Qi P, Zhang S, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering[J]. arXiv preprint arXiv:1809.09600, 2018.