南 方 科 技 大 学

硕 士 学 位 论 文 开 题 报 告

题 目：从能力评估到蒸馏量化：大模型多维评测方法研究

院　　（系）	深圳理工大学
学　　科	计算机技术
导　　师	杨敏
研 究 生	敖畅
学　　号	12333471
开题报告日期	2025 年 3 月 24 日

研究生院制

目 录

第 1 章　大语言模型的评测

1.1　大语言模型的发展历程

大语言模型（Large Language Models, LLMs）作为自然语言处理领域的革命性范式，通过超大规模无监督预训练[1]，实现了语言统计规律与语义表征的深度融合，同时展现出前所未有的跨任务迁移能力和上下文学习能力。其技术发展历程可划分为三个关键演进阶段，每个阶段均有显著的技术突破与范式转变。

1.1.1　统计语言模型

早期基于 N-gram 的统计模型运用马尔可夫假设进行词序列概率估计[2]，通过计算特定词汇出现的条件概率来预测下一个可能的词汇。然而，这类模型受限于数据稀疏性问题（数据量有限导致大量 n 元组在训练集中从未出现）和长程依赖捕捉难题（通常仅考虑前 2-5 个词的上下文），使得测试集困惑度通常高于 200，表明模型预测能力有限。为解决这些局限，神经网络语言模型（NNLM）的提出[3] 首次将分布式表示引入语言建模，使词汇能够在连续向量空间中表示，进而捕获词汇间的语义相似性，显著改善了泛化能力。随后，LSTM[4]、GRU[5] 等循环架构虽在一定程度上缓解了梯度消失问题，通过门控机制有效处理中长距离依赖，但在处理长文本时，训练耗时随序列长度呈 $O(n^2)$ 增长，存在明显的计算效率瓶颈，且实际效果受限于有限的感受野。

1.1.2　Transformer

Transformer 架构[1] 的出现彻底革新了 NLP 技术路线，被广泛认为是继卷积神经网络之后深度学习领域最具影响力的架构创新。其核心创新点如下：

- **自注意力机制：**允许任意位置的词元建立动态关联，计算复杂度为 $O(n^2)$，相比 RNN 能更有效捕获全局依赖关系。通过查询 (Query)、键 (Key)、值 (Value) 三重表示实现细粒度的注意力分配，使模型能够同时关注序列中多个位置的信息。
- **位置编码系统：**通过正弦余弦函数生成的位置向量，将序列顺序信息融入高维空间，弥补了自注意力机制中位置信息缺失的问题。这种编码方式具备良好的外推性，能处理训练中未见过的序列长度。
- **并行计算架构：**摒弃了传统循环结构的顺序依赖，实现序列数据的全并行处理，相较于 RNN，训练速度提升 10 倍以上[6]，为超大规模模型训练奠定了技术基

础。

• **多头注意力设计：**通过多个并行的注意力计算” 头”，允许模型同时关注不同表示子空间中的信息，进一步增强了表征能力，能够捕获更丰富的语言结构和语义关系。

1.1.3　预训练模型

GPT 系列开创了自回归预训练范式[7]，通过” 从左到右” 的单向语言建模任务，使模型习得强大的生成能力。其发展历程呈现显著的规模效应与涌现能力，具体表现为：

• GPT-1（117M 参数）初步验证了大规模预训练的有效性，但仍需针对下游任务进行微调[7]。

• GPT-2（1.5B 参数）展现出初步的零样本学习能力，能在无监督条件下生成连贯文本[8]。

• GPT-3（175B 参数）显著突破了参数规模瓶颈，在 MMLU 基准的零样本准确率达 43.9%，首次展现出上下文学习（in-context learning）能力，仅通过示例即可适应新任务[9]。

• PaLM（540B 参数）在 BIG-bench 的复杂推理任务中超越人类平均水平，在多步推理和逻辑分析中表现出接近人类的能力[10]。

• GPT-4 引入了指令微调（instruction tuning）与人类反馈强化学习（RLHF）技术，在多模态理解任务中展现出卓越的跨模态对齐能力，能够处理图像和文本的复杂交互[11]。

与之形成技术竞争的 BERT 系列[12] 采用双向编码架构，通过掩码语言建模（MLM）和下一句预测（NSP）双重目标进行预训练，在 GLUE 基准上实现了 7.7% 的绝对性能提升，重新定义了自然语言理解任务的技术上限。随后，RoBERTa[13] 通过优化训练策略（移除 NSP 任务、动态掩码机制等）进一步提升了模型性能，ALBERT[14] 则通过参数共享等技术降低了计算开销。中国技术团队也积极投入这一领域，推出了具有鲜明中文语言特色的大模型，如通义千问[15]、文心一言[16] 等，这些模型不仅在英文评测上表现出色，在 C-EVAL 中文基准上的综合准确率也突破 80%[17]，体现出较强的多语言能力与文化适应性。

1.2　评测的重要性

大语言模型评测在推动模型发展、应用与优化方面发挥着不可或缺的关键作用，其重要性体现在多个层面，构成了模型发展生态系统中的关键环节。

1.2.1　模型发展

在模型发展方面，评测为大语言模型的迭代升级提供明确方向和量化指标。通过评估模型在各类任务和数据集上的表现，能够精准识别模型在语言理解、生成、推理等能力上的优势与不足。例如，若模型在常识推理任务（如 PIQA[18]、HellaSwag[19] 等数据集）中频繁出错，表明其在常识知识的理解和应用方面存在欠缺，研究人员可据此针对性地改进模型架构、调整训练策略或扩充训练数据，以提升模型的常识推理能力。

同时，评测能够验证新的技术和方法在大语言模型中的有效性，为理论创新提供实证支持。当研究人员提出新的训练算法（如稀疏混合专家模型 MoE[20]）、优化技术（如低秩适应 LORA[21]）或模型架构（如 Transformer 变体 Mamba[22]）时，通过严格评测可判断这些创新是否真正提升了模型性能，量化其改进幅度，进而决定是否将其应用于实际模型开发。值得注意的是，通过建立标准化的评测框架（如 MMLU[23]、HELM[24] 等），研究社区能够客观比较不同模型间的性能差异，减少主观偏见，促进技术积累和知识共享，加速整个领域的进步。

1.2.2　应用

从应用视角来看，评测有助于用户根据自身需求选择最适配的大语言模型，降低技术选型风险。不同应用场景对大语言模型的能力要求各异，如智能客服需要模型具备快速准确的问题理解和回答能力，关注响应时间和答案准确性；内容创作则更注重模型生成文本的流畅性、逻辑性和创新性；垂直领域应用（如医疗诊断辅助）则侧重专业知识准确性和伦理安全性。借助全面的评测结果，用户能够了解不同模型在各项能力上的表现差异，从而挑选出最符合应用需求的模型，提高应用效果和质量，降低试错成本。

在医疗、金融、法律等对准确性和可靠性要求极高的领域，只有经过严格评测证明性能可靠的模型才可能被用户接受和使用。例如，在医疗场景中，模型需通过特定的医学知识评测（如 MedQA[25]、MedMCQA[26] 等）验证其专业能力；在金融领域，则需评估模型在财务分析和风险评估方面的表现。评测结果不仅增强了用户对大语言模型的信任，也为行业监管提供了技术依据，构成了模型正式部署前的必经环节。此外，针对不同应用场景的特化评测（如客服场景下的多轮对话评测）能够帮助用户更精准地评估模型在特定任务上的适用性。

1.2.3　模型优化

在模型优化过程中，评测是衡量模型性能提升的重要标准，驱动着迭代改进的闭环。在模型训练过程中，持续对模型进行评测，能够实时监控模型性能变化，判断训练是否朝着预期方向进行。若评测结果显示模型在某些指标上未提升甚至下降（如参数量增加但推理能力反而下降的” 过拟合” 现象），就需及时调整训练参数（如学习率、批量大小）或方法（如增加正则化策略），以确保模型的优化效果。

评测还能帮助研究人员对比不同模型之间的性能差异，学习借鉴其他模型的优点，进一步优化自身模型。例如，通过对比分析发现某模型在语言生成的多样性方面表现出色，研究人员可研究其实现方法（如 decoding 策略[27]、温度参数设置等），并将相关技术应用到自己的模型中，提升模型生成文本的多样性。同时，评测还能揭示模型的瓶颈问题和弱点，为后续研究方向提供指引。如果评测显示大多数模型在跨语言任务中表现不佳，这可能意味着当前多语言处理技术存在普遍性缺陷，需要行业共同关注和解决。此外，通过持续评测记录模型优化的历史轨迹，研究人员可以总结经验教训，逐步形成系统化的模型优化方法论。

1.3　评测维度

大语言模型的评测维度涵盖多个关键方面，构成了一个多层次、多角度的综合评价体系。全面评估这些维度对于深入了解模型性能至关重要，能够揭示模型在不同能力上的表现特点。

1.3.1　语言能力

语言能力是大语言模型的核心能力，包括语言理解与生成两大关键方面。在语言理解方面，模型需准确把握文本含义，无论是简单的日常表述还是复杂的专业内容，这包括词法分析、句法解析、语义理解等多层次任务[28]。例如在阅读理解任务中，给定一篇复杂的科技论文，模型应能理解其中的专业术语（如”transformer 架构”、” 自注意力机制”）、实验过程（如” 使用 Adam 优化器，初始学习率设置为 5e-5，在 8 个 V100 GPU 上训练了 3 天”）和结论等关键信息，并回答相关问题，如” 论文中实验的创新点是什么”（例如” 引入了新的预训练目标函数”）、” 实验结果支持了哪些假设”（例如” 大规模预训练显著提升了下游任务性能”）等。

语言生成能力要求模型产出的文本语法正确、语义连贯且逻辑合理，同时满足风格一致性和目的适应性。以故事写作为例，模型生成的故事需情节连贯、人物形象鲜明、主题明确，避免出现逻辑漏洞或前后矛盾的情况，比如故事开头设定

主角是勇敢的战士，后续却无端表现出懦弱且无合理情节解释；或者故事中引入了重要的支线情节却未能在后文中得到合理的收束。高质量的语言生成还需考虑目标受众和语境适应，例如为儿童编写的故事应使用简单词汇和句式，避免复杂抽象概念；学术论文生成则需严谨的论证结构和专业术语的准确使用。此外，语言生成能力还包括风格多样性，如能否根据需求生成不同文体（诗歌、散文、新闻等）的文本，以及是否能模仿特定作者的写作风格，如生成类似莎士比亚、鲁迅风格的作品。

1.3.2　知识储备

大语言模型的知识储备广度和深度直接影响其在各类任务中的表现，构成了模型” 认知基础” 的重要组成部分。在知识广度上，模型应涵盖多领域知识，如历史、科学、文化、艺术、体育、政治等各个方面。当被问到” 秦始皇统一六国的时间”（公元前 221 年）、” 相对论的主要内容”（时间和空间不是绝对的，而是相对的，且质量与能量等价）、” 蒙娜丽莎的创作者”（达芬奇）等不同领域问题时，都能给出准确、全面的答案，避免知识盲点。知识的时效性也是重要考量，模型应能区分静态知识（如数学定理、历史事件）和动态知识（如当前政治局势、科技发展），并理解知识的更新迭代过程.

在知识深度方面，对于专业领域问题，模型需展现出深入理解和专业洞察，而非停留在表面的常识性认知[29]。例如在医学领域，对于疾病诊断和治疗方案相关问题，如” 如何诊断早期肺癌”（应涉及低剂量 CT 扫描、痰液细胞学检查等具体方法，并讨论各种检测方法的敏感性和特异性），” 糖尿病的最佳治疗方案有哪些”（应区分 1 型和 2 型糖尿病，讨论胰岛素注射、口服降糖药、生活方式干预等多种治疗方法，并考虑患者个体差异），模型应依据专业知识给出科学合理的回答，包括疾病机制、诊断标准、治疗原则等核心内容，而非泛泛而谈。同时，模型对知识的组织和关联能力也很重要，能否建立不同知识点之间的连接，推导出新的结论或见解，是评估知识深度的关键指标[30]。此外，模型应具备元知识能力，能够识别自身知识的边界，对超出知识范围的问题给出适当的不确定性表达，避免生成错误信息[31].

1.3.3　应用能力

大语言模型在实际场景中的应用能力是衡量其价值的重要指标，直接关系到模型的落地效果和商业价值。在智能客服场景，模型要能快速理解用户问题的核心诉求，准确提供针对性解决方案，同时保持自然流畅的对话体验[32]。如电商客服中，需精准解答用户关于商品信息（” 这款手机的处理器是什么型号”）、物流

配送（”我的订单什么时候能到货”）、售后服务（”购买后发现商品有瑕疵如何处理”）等问题，并在用户表达不明确时主动引导澄清，在复杂问题上提供多种解决途径供用户选择。模型还需识别用户情绪，对于不满或焦虑的用户，采取更加耐心和共情的回应策略。

在内容创作领域，模型需作为创作辅助工具，生成高质量、原创性强的新闻报道、文案策划、小说等内容，满足不同的创作需求[33]。例如生成一篇具有吸引力的产品推广文案，不仅要突出产品核心功能和独特优势（”行业首创的智能降噪技术，有效降低 90% 环境噪音”），还需针对目标用户群体的痛点和需求进行情感共鸣（”告别嘈杂，重新发现宁静的美好”），同时融入号召性用语激发购买欲望（”限时优惠，立即体验”）。在学术写作辅助中，模型需生成结构严谨、论证充分的论文框架或章节内容，正确使用学术术语和引用格式[34].

在智能教育领域，模型作为智能辅导工具，需根据学生的知识水平和学习特点，为其解答学习疑问，提供个性化学习建议[35]。如针对数学学科中的难题，不仅要详细讲解标准解题思路和方法（”首先，将方程两边同时除以 x，得到...”），还应分析常见错误（”注意，这里不能直接约分，因为...”），提供多种解题策略（”除了代数法，我们还可以用几何方法...”），并鼓励学生独立思考（”你能尝试用另一种方法解决这个问题吗？”）。在语言学习场景，模型需提供地道的表达示例，纠正语法错误，解释文化背景，帮助学习者全面提升语言能力。此外，模型还应具备跨场景适应能力，能够根据应用环境的变化灵活调整回应策略和内容深度[36]. f

第 2 章　大语言模型评测的研究现状

近年来，大语言模型评测研究发展迅猛，国内外学者在评测方法创新、指标体系构建和数据集扩展等方面取得显著成果。随着模型规模和能力的飞速提升，评测研究已从单一技术指标逐步走向多维度、多层次的综合评价体系。本章从多个角度系统梳理该领域的研究现状与发展脉络，揭示当前研究的关键议题和未来挑战。

2.1　评测方法

当前主流评测方法可分为自动评测与人工评测两大类型，两者各具优势，在实践中常相互补充形成混合评测策略。自动评测依靠算法程序对模型输出进行量化分析，其核心优势在于评估效率高、成本较低且具有可扩展性，适合大规模模型比较和持续监测场景。典型应用包括：基于 N-gram 匹配的 BLEU 指标[37] 在机器翻译质量评估中的广泛应用，通过计算候选翻译与参考翻译间的 n 元组重合度来衡量翻译质量；以召回率为导向的 ROUGE 系列指标[38] 在文本摘要任务中的基准地位，特别是 ROUGE-L 通过最长公共子序列评估生成摘要与参考摘要的相似度；BERTScore[39] 则利用预训练语言模型的上下文表征计算语义相似度，在一定程度上缓解了传统指标对表面形式过于敏感的问题。

与之相辅相成的人工评测则强调人类专家对语义准确性、逻辑连贯性及语用适切性的综合判断[40]，其评估维度更贴合实际应用场景，能够捕捉自动指标难以量化的细微差异。人工评测通常基于精心设计的评分标准（rubrics），让评估者对模型输出的多个方面（如事实准确性、相关性、连贯性等）进行打分，或通过相对排序确定不同模型输出的优劣。为提高可靠性，现代人工评测常采用多评估者一致性检验（Cohen’s Kappa > 0.75 视为高度一致）和双盲评估设计，但仍受评估者主观差异、认知偏见和较高时间经济成本的限制。值得注意的是，近期涌现了以 GPT-4 为代表的模型辅助评测方法[41]，尝试结合人工与自动评测的优势，显著提高了评测效率，但其可靠性和公正性仍需进一步验证。

2.2　评测指标

现代评测指标体系已突破单一维度限制，形成了覆盖语言理解与生成能力、知识储备、推理能力和伦理安全的综合评价框架。在生成质量评估方面，除传统困惑度指标（PPL，越低表示模型对语言序列预测越准确）外，MAUVE[42] 通过信息论中的 KL 散度前沿分析实现生成文本与人类文本的分布对比，有效捕捉生成结果的多样性与自然度。该指标将神经文本生成问题重新框定为分布匹配问题，避免了点对点评估的局限性，在开放式生成任务评估中显示出独特优势。类似地，USR[43] 通过多个预训练模型评估对话回应的相关性、连贯性和参与度，为对话系统评估提供了多角度参考。

理解能力评估普遍采用准确率、F1 值等分类指标，如在 GLUE 基准[44] 中，模型需在文本蕴含（RTE，判断一个句子是否可从另一句推导）、情感分析（SST-2，判断文本情感极性）等九项任务中证明其语义理解能力。后续的 SuperGLUE[45] 进一步提高了难度，引入了更复杂的推理任务（如 WiC、BoolQ 等）。值得注意的是，新一代评估框架如 HELM[24] 已整合准确性、鲁棒性、公平性、毒性、效率、刻板印象和环境影响等七大评估维度，推动评测标准向全栈式评估发展。HELM 特别关注模型在不同人口群体间的性能差异，为算法公平性研究提供了重要数据支持。最新的 Holistic Evaluation of Language Models（HELM 2.0[46]）进一步扩展了评估范围，将长文本处理能力、多语言能力和领域迁移能力纳入评估体系，构建了更全面的模型画像。

2.3　评测数据集

评测数据集建设呈现” 通用化” 与” 专业化” 并行发展的态势，同时多语言、多地区评测资源呈现爆发式增长。通用基准如 BIG-bench[47] 通过 204 个不同任务（包括语言理解、翻译、推理、计算、知识、社会互动等）全面检验模型的语言能力谱系，累计超过 50 万个问题，为大模型能力边界绘制了详尽地图。MMLU[23] 则聚焦高等教育和专业知识，通过 57 个学科的多项选择题评估模型的学科知识掌握程度，包括 STEM、人文、社会科学和其他专业领域。

与此同时，专业领域数据集加速涌现：MedQA[25] 聚焦医疗领域构建诊断推理评估体系，涵盖美国医学执照考试（USMLE）真题，测试模型的临床推理链完整性；LegalBench[48] 针对法律文件分析、合同审核和法条适用等核心法律任务构建基准，检验模型的法律推理能力；GSM8K[49] 专注数学问题求解，要求模型展示逐步推理过程，而非仅给出最终答案。

值得关注的是中文评测生态的快速发展，针对中文语境和知识体系构建了一

系列本土化评测基准：MMCU[50] 构建了涵盖中国历史、地理、文学和哲学的专业领域知识图谱评估体系，包含 3 万余题的挑战性问题集；CMMLU[51] 建立首个系统评估中华文化特性的评估基准，覆盖 67 个学科，特别关注中国特色知识如诗词歌赋、成语典故等；CMB[52] 则开创中医药知识评估先河，通过经典医籍理解、方剂组成和临床诊断三大模块，测试模型对中医药理论体系的理解深度。然而，现有基准多关注高阶认知能力，在基础教育等关键领域仍存在评估空白——特别是面向 K-12 教育阶段的渐进式知识掌握评估体系尚未建立，这限制了大语言模型在基础教育领域的规范应用和能力验证。

2.4　模型蒸馏评估

随着知识蒸馏技术的广泛应用[53]，其评估难题日益凸显，特别是在大语言模型领域，传统蒸馏框架已难以应对复杂能力传递的度量挑战。研究表明[54]，通过参数迁移实现的后发优势可能伴随着模型鲁棒性显著下降（ΔRobustness = 12.7%，$p<0.01$），尤其在对抗样本和分布外数据集上表现尤为明显。例如，在对抗性问答数据集 AdversarialQA 上，蒸馏模型的准确率下降幅度是教师模型的 1.6 倍，反映了知识迁移过程中防御能力的不完全传递。

当前面临的评估困境主要包括三个方面：1）知识迁移路径的” 黑箱效应” 导致师生模型间的能力差异难以精确量化，特别是在教师模型参数规模超过学生模型 10 倍以上时，难以通过传统损失函数刻画能力传递的完整性；2）蒸馏特定基准数据稀缺迫使研究者过度依赖间接相关性分析（$r = 0.63\pm0.08$），缺乏针对性的度量工具验证蒸馏效果；3）表征空间的抽象冗余特性阻碍可解释性分析，难以追踪特定能力（如常识推理、算术能力）在蒸馏过程中的传递状态。Liu 等[55] 提出的蒸馏能力地图（Distillation Capability Map）尝试通过多维度任务性能比对构建蒸馏效果可视化，但仍未解决根本性评估难题。

更值得警惕的是，学术界对数据蒸馏的过度依赖可能抑制技术创新[56]。数据显示，2022-2024 年间发表的 LLM 相关论文中，36.7% 采用某种形式的知识蒸馏技术，但仅有 8.4% 提出了原创性架构改进。这种趋势可能导致技术路径锁定，亟待建立标准化评估框架，引导研究方向多元化发展。Zhou 和 Chen[57] 最近提出的蒸馏透明度指数（Distillation Transparency Index, DTI）框架，通过能力保留率、鲁棒性保留率和创新贡献率三项指标综合评估蒸馏模型，为解决评估困境提供了初步思路，但其适用性和可靠性仍需大规模验证。

2.5　多模态评测框架

随着大语言模型向多模态方向扩展，评测框架也相应拓展至视觉-语言理解与生成领域。当前多模态评测呈现出三个明显特征：跨模态一致性评估、多粒度理解测试和生成质量多维度量化。MMMLU[58] 作为 MMLU 的多模态扩展版本，通过图像-文本匹配任务评估模型的跨模态理解能力，特别关注模型是否能准确捕捉图像中的细节信息并与文本描述建立正确关联。例如，当展示一幅包含特定病理特征的医学图像时，模型需识别关键视觉特征并选择最准确的诊断结论。

在视觉推理方面，SEED-Bench[59] 设计了多粒度视觉理解任务，从对象识别、场景理解到视频事件因果推理，系统评估模型的视觉认知层次。其中特别有挑战性的是时序理解任务（Temporal Understanding），要求模型理解视频中的动作序列并预测后续发展，如” 视频中的人完成当前动作后最可能做什么”。

生成质量评估方面，LVLM-eHub[60] 提出了针对图像生成的多维度评分体系，包括视觉质量（清晰度、细节还原度）、文本一致性（生成图像与提示词的匹配度）和创意性（构图新颖性、风格独特性）三大维度共 12 项指标。该评测框架采用人机协作评估方法，结合自动评分算法和人类专家判断，为多模态生成模型提供全面性能画像。

第 3 章　主要研究内容及研究方案

基于上一章的阐述，本课题主要聚焦于 K-12 教育阶段渐进式知识掌握评估体系的构建与模型蒸馏程度的量化研究，并在此基础上开展深入的实验与验证工作。本章将介绍当前研究成果及后续工作规划。

3.1　一个面向大模型的中文 K-12 教育评估基准

现有评测基准通常注重对高级能力的考察，但对于某些关键的特定领域或主题往往关注不足。在中国 K-12 教育领域，目前尚缺乏一套全面的评估基准，而在这一领域对模型在不同认知阶段的学习细节进行系统性评估和解析具有重要意义。

3.1.1　研究内容

基于此，本研究构建了首个面向中文 K-12 教育领域的大语言模型综合评估基准 E-EVAL。该基准覆盖小学、初中和高中三个阶段的单项选择题，涵盖语文、数学、英语、物理、化学等学科，并将学科分为文科与理科两类。E-EVAL 旨在弥补现有基准在中文基础教育领域覆盖不足的缺陷，通过系统性评估来揭示大语言模型在知识掌握与推理能力方面的优势与局限性。

E-EVAL 采用多项选择题（multiple-choice question, MCQ）格式，与 Hendrycks 等人[23] 的方法类似，能够清晰有效地衡量 LLMs 的准确度与推理能力。题目经过精心筛选，主要选自日常作业及地方性小规模考试，力求真实反映教育场景，同时保证内容的原创性与区域特征。对于涉及复杂数学公式的理科数据，E-EVAL 在数据收集与细化处理方面投入了较多精力，以确保数据的完整性和准确性。

为尽量减少数据污染风险，E-EVAL 避开了全国性考试试题（如高考），转而选取地方模拟考试与特定高中在线测试题，并优先使用 PDF 与 Word 文档作为数据来源，以进一步降低数据泄露的可能性。该基准从全国不同地区、学校、年级与学科的数千份试卷中抽取数据，确保了评测数据的广泛性和代表性。

3.1.2　研究方案

研究方案主要包括以下核心环节：

3.1.2.1　数据集构建

E-EVAL 涵盖中国 K-12 教育所需的核心学科，分为小学、初中与高中三个阶段，并进一步按照学科性质划分为文科与理科两大类。文科主要包含语文、英语、政治、历史和地理等学科，这些学科更注重人文素养与思维能力培养；理科主要包含数学、物理、化学、生物等学科，整体涉及 STEM（科学、技术、工程与数学）领域。

研究所用题目主要来源于互联网中免费提供的作业、练习和模拟考试题，这些地方性题目通常由学校、教育机构或教师提供，传播范围相对较小，不仅有效降低了数据污染风险，也更真实地反映了日常教育及学术要求。大多数收集到的题目以 PDF 或 Word 格式呈现：对于文科类学科，采用脚本自动解析并结构化题目内容；对于理科类学科，由于涉及较多复杂公式，需人工核对并将公式转换成标准的 LaTeX 格式，以确保准确性。

大多数题目采用四选一形式。少于四个选项的题目被舍弃，若多于四个选项则随机移除一个错误选项保证统一格式。在数据校对环节，分别使用三轮人工复核：第一轮排查题目重复度，第二轮确保数学公式的正确性与完整性，第三轮确认答案的准确性。随后对选项顺序进行平衡化处理，使正确答案在 A、B、C、D 四个选项中的分布更均衡，并依据 23 个学科类别将数据划分为开发集（development set）、验证集（validation set）和测试集（test set），并筛选少量有详细解析的试题用于少样本（few-shot）评测。

数据统计情况如表 3-1 所示，整体概览如图 3-1 所示。

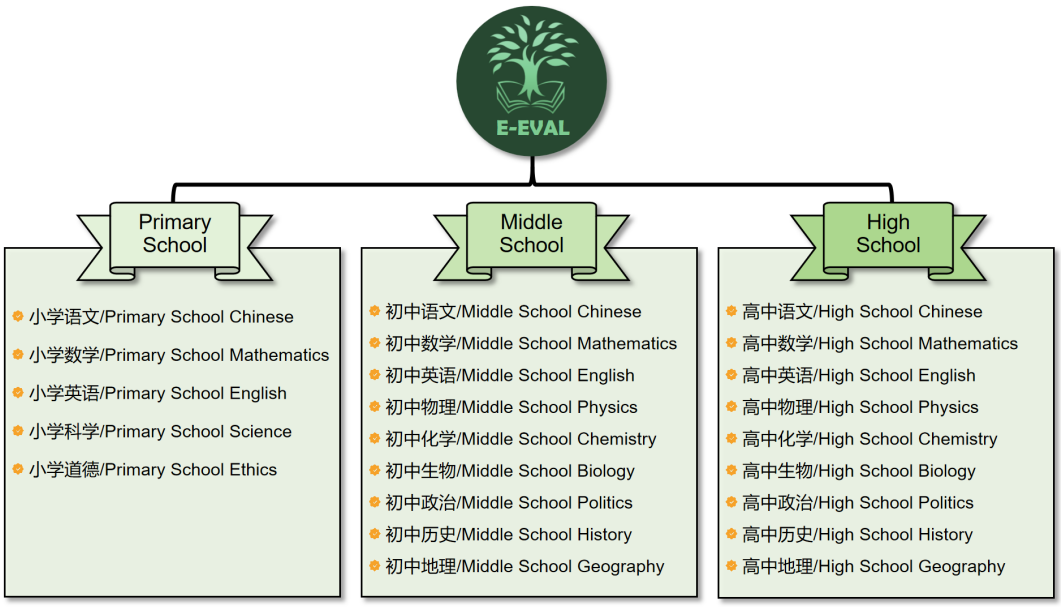

图 3-1　E-Eval 基准总览

subject	#Subject	#Question
In terms of difficulty		
High School	9	2225
Middle School	9	1589
Primary School	5	537
In terms of arts/science		
Arts	13	2699
Science	10	1652
In terms of split		
Dev	23	115
Valid	23	424
Test	23	3812
Total	23	4351

表 3-1　统计数据

3.1.2.2　评估框架设计

E-EVAL 的实验设置主要评价不同大型语言模型（LLMs）在四选一题目上的表现。通过正则匹配自动从模型生成内容中提取最终的答案，进而衡量相应模型的准确度。本研究选取了三种评测方法：

(1) 零样本评测（zero-shot）[61]：在未提供示例的情况下直接回答；

(2) 少样本-仅答案评测（few-shot-answer-only, few-shot-ao）：在提供少量示例后，仅考察模型的最终答案选择；

(3) 少样本-思维链评测（few-shot-chain-of-thought, few-shot-cot）[62]：在少样本示例的基础上，引导模型进行逐步推理，以评估逻辑推理能力。

3.1.2.3　模型评估

我们对多个规模与来源不同的中英文大语言模型（如 Qwen、ERNIE-Bot、GPT 等）进行了评测，涵盖开源与商业模型，以分析它们在学科类型（文科 / 理科）、教育阶段（小学 / 初中 / 高中）及提示方法下的性能差异。结果如表 3-2 所示：

该工作已被国际会议 Findings of ACL2024 接收。

Model	Arts	Science	Primary	Middle	High	Average
Random	25.0	25.0	25.0	25.0	25.0	25.0
Qwen-72B-Chat	92.5	84.0	89.3	93.1	85.6	88.9
Ernie-Bot 4.0	90.8	78.6	87.3	89.6	82.1	85.5
Yi-34B-Chat	82.4	69.5	79.6	83.1	71.7	76.9
Ernie-Bot	81.7	68.2	78.7	80.8	71.6	75.9
GPT 4.0	75.4	64.2	81.9	76.8	70.6	70.6
Yi-6B-Chat	74.7	61.1	71.3	76.1	63.1	68.8
Qwen-7B-Chat	67.3	50.1	69.7	65.9	53.3	59.9
Baichuan2-13B-Chat	65.3	47.8	69.2	65.1	49.8	57.8
ChatGLM3-6B	61.9	51.9	60.0	65.0	51.8	57.6
Baichuan2-7B-Chat	62.2	45.0	61.2	61.3	48.6	54.8
ChatGPT	60.5	46.8	68.3	58.2	48.8	54.6
Chinese-Alpaca-2-13B	53.6	36.6	51.4	46.7	38.9	43.3
Educhat-sft-002-13B	41.8	28.9	39.9	39.9	32.7	36.3
Chinese-LLaMA-2-13B	44.2	31.9	39.2	38.5	33.2	35.9
Educhat-sft-002-13B-Baichuan	40.2	29.3	40.8	38.4	32.2	35.5

表 3-2　多个模型在不同类别上的准确率

3.2　一种量化模型蒸馏程度的方法

近年来，模型蒸馏作为一种高效发挥大型语言模型（LLM）能力的途径而备受关注。通过从更强大的 LLM 迁移知识到较弱模型，数据蒸馏可以在显著减少标注需求[63,64] 以及降低计算资源和探索成本的同时取得较优性能。但这种“后发优势”也带来潜在的负面影响：可能会阻碍学术机构或研究团队在技术上的自主探索，转而直接依赖先进 LLM 的蒸馏输出。此外，研究发现数据蒸馏还会引起一定程度的鲁棒性下降[65–67]，因此对蒸馏程度进行量化尤为必要。

3.2.1　研究内容

针对上述问题，本文提出响应相似度评估（Response Similarity Evaluation, RSE）方法，用于量化模型蒸馏程度。该方法通过比较原始 LLM 与目标学生 LLM 的输出，以衡量知识传递的深度与广度。

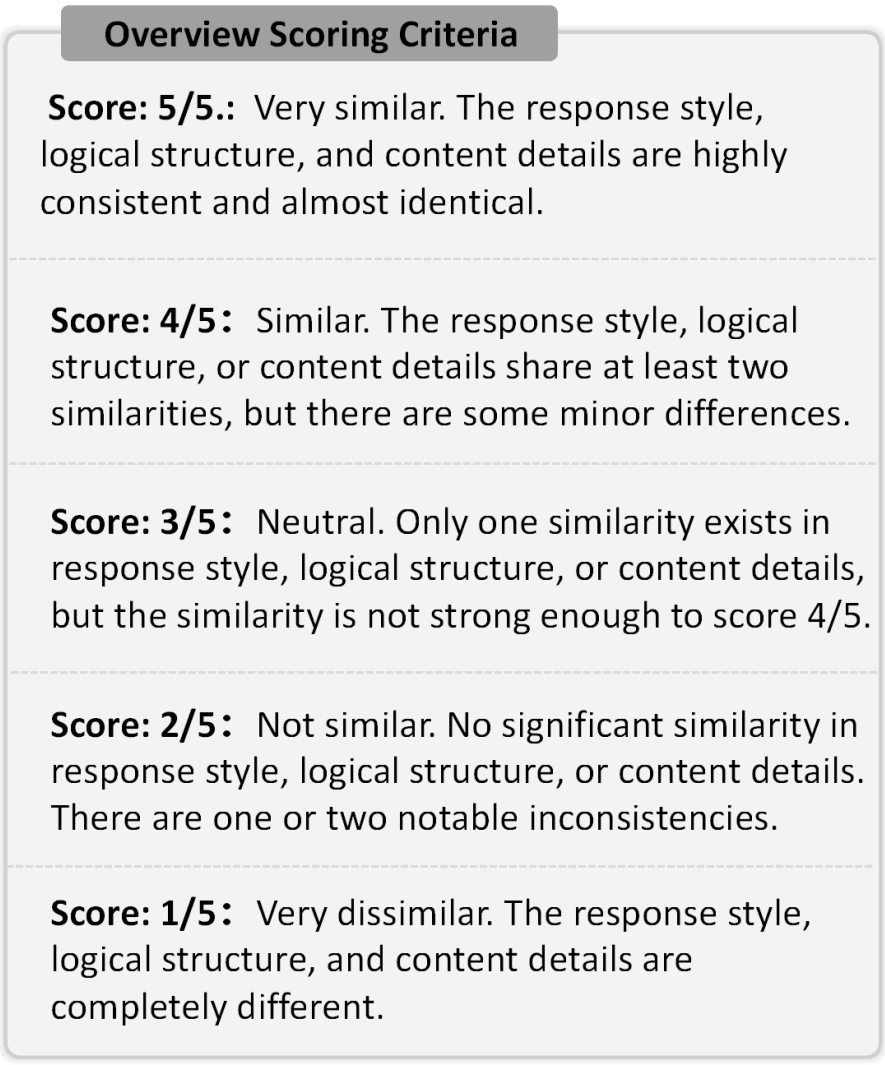

图 3-2　RSE 中“LLM-as-a-judge”的评分标准，共分为五级，1 表示“非常不相似”，5 表示“非常相似”。不同评分等级在回答风格、逻辑结构、内容细节上有不同的表现。

3.2.2　研究方案

在 RSE 中，选择一个参考 LLM（记为 LLM_{ref}，即 GPT）并对待测 LLM（记为 LLM_{test}）进行输出对比。我们从回答风格、逻辑结构、内容细节三个维度评估 LLM_{test} 与 LLM_{ref} 的输出相似度，并为每个模型输出分配一个综合相似度得分。为全面考量模型在通用推理、数学计算与指令遵循方面的蒸馏程度，本研究人工选取 **ArenaHard**、**Numina**、**ShareGPT** 三个提示集来获取参考答案 R_{ref} 和测试模型的响应 R_{test}。随后利用“LLM-as-a-judge”的方式逐一将 R_{test} 与 R_{ref} 进行比对，并根据相似程度给出 1 至 5 的评分等级（如图 3-2）。

3.3　后续研究规划

3.3.1　研究条件与方案

（1）研究条件

预计在 7B 至 72B 多种规模的模型上进行实验。运行环境需要 8 张 A800 80GB（已满足），并需调用 GPT-4、deepseek 及 claude 的 API 进行评测，因此可能产生相应的接口费用。

（2）研究方案

计划在 RSE 上探索 prompt 调优方法，以进一步提升 RSE 的评价准确度。并考虑从越狱角度对模型蒸馏程度进行量化，旨在绕过 LLM 的“自我保护”机制，挖掘其训练数据中包含的潜在信息（如蒸馏数据来源 LLM 的名称、国家、位置或团队等）。

3.3.2　进度安排与预期目标

（1）进度安排

硕士阶段的公共课与专业课学分（34/34）已修读完毕。计划在 2025 年 5 月左右完成整体框架构建与细节完善，并验证模型蒸馏程度量化方法的可行性，力争投递一到两篇国际会议；随后在 2025 年 8 月完成中期报告并着手论文收尾工作，至 2026 年 2 月完成学位论文的撰写与定稿。

（2）预期目标

期望提出首个系统化的基于大模型评测的模型蒸馏程度量化框架，明确量化评估指标，尤其是提出的 **响应相似度评估（Response Similarity Evaluation, RSE）** 方法。

3.3.3　可能遇到的问题与解决措施

由于在不同 prompt 下，RSE 方法可能表现迥异，需要进行多次 prompt 调整并对蒸馏前后的模型进行对比试验，以得到最优提示设置。同时还需对具体案例进行人工分析，确保 RSE 的评估模型能够正确进行判断。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] JURAFSKY D, MARTIN J H. Speech and language processing[M]. Pearson Education, 2021.

[3] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. The journal of machine learning research, 2003, 3: 1137-1155.

[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[5] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1724-1734.

[6] CHEN M X, FIRAT O, BAPNA A, et al. The best of both worlds: Combining recent advances in neural machine translation[J]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 1: 76-86.

[7] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[Z]. 2018.

[8] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[9] 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.

[10] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: Scaling language modeling with pathways[A]. 2022.

[11] OPENAI. GPT-4 Technical Report[EB/OL]. 2023. https://arxiv.org/abs/2303.08774.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of NAACL-HLT, 2019: 4171-4186.

[13] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach [A]. 2019.

[14] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[J]. International Conference on Learning Representations, 2020.

[15] QWEN TEAM A G. Qwen Technical Report[EB/OL]. 2023. https://arxiv.org/abs/2309.16609.

[16] BAIDU. Wenxin Yiyan: An AI Assistant Specifically Built for Chinese Language[Z]. 2023.

[17] HUANG Y, BAI Y, ZHU Z, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models[A]. 2023.

[18] BISK Y, ZELLERS R, GAO J, et al. PIQA: Reasoning about physical commonsense in natural language[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(05): 7432-7439.

[19] ZELLERS R, HOLTZMAN A, BISK Y, et al. HellaSwag: Can a Machine Really Finish Your Sentence?[J]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4791-4800.

[20] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[A]. 2017.

[21] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[J]. International Conference on Learning Representations, 2022.

[22] GU A, DAO T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces[A]. 2023.

[23] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[J]. Proceedings of the International Conference on Learning Representations, 2021.

[24] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[A]. 2022.

[25] JIN C, CHEN J, STEINFELD A, et al. Disease prediction from electronic health records using generative adversarial networks[J]. Nature Machine Intelligence, 2021, 3(9): 796-804.

[26] PAL A, UMAPATHI L K, SANKARASUBBU M. MedMCQA: A Large-scale Multi-subject Multi-choice Dataset for Medical domain Question Answering[J]. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 7871-7888.

[27] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[J]. International Conference on Learning Representations, 2020.

[28] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[J]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4593-4601.

[29] HEILMAN M, SMITH N A. Automatic factual question generation from text[J]. Language and Linguistics Compass, 2019, 13(9): 1-16.

[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] LIN S, HILTON J, EVANS O. Teaching models to express their uncertainty in words[A]. 2022.

[32] ROLLER S, DINAN E, GOYAL N, et al. Recipes for building an open-domain chatbot[J]. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021: 300-325.

[33] YANG K, KLEIN D, ABBEEL P, et al. Re3: Generating longer stories with recursive reprompting and revision[A]. 2022.

[34] WANG H, YIN D, DU Q, et al. Scientific discovery in the age of artificial intelligence[J]. Nature, 2023, 620(7972): 47-60.

[35] KASNECI E, SESSLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.

[36] LIU W X, DAI D, AIZAWA A. A survey of large language models[A]. 2023.

[37] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2002: 311-318.

[38] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Text summarization branches out. Association for Computational Linguistics, 2004: 74-81.

[39] ZHANG T, KISHORE V, WU F, et al. BERTScore: Evaluating text generation with BERT[A]. 2020.

[40] ZHOU Y, WANG A, WU L, et al. InstructEval: Systematic Evaluation of Instruction Selection Methods[A]. 2023.

[41] ZHENG L, CHIANG W L, SHENG Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena[A]. 2023.

[42] PILLUTLA K, SWAYAMDIPTA S, ZELLERS R, et al. MAUVE: Measuring the gap between neural text and human text using divergence frontiers[C]//Advances in Neural Information Processing Systems: Vol. 34. 2021: 67-79.

[43] MEHRI S, ESKENAZI M. USR: An unsupervised and reference free evaluation metric for dialog generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 681-707.

[44] WANG A, SINGH A, MICHAEL J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2018: 353-355.

[45] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems[J]. Advances in neural information processing systems, 2019, 32.

[46] KöPF M, HUNG D, ZHANG L, et al. HELM 2.0: Enhanced Holistic Evaluation of Language Models[A]. 2023.

[47] SRIVASTAVA A, RASTOGI A, RAO A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[A]. 2022.

[48] GUHA N, GUO D E, REN N, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models[A]. 2022.

[49] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[A]. 2021.

[50] ZENG Z, TENG Y, XU M, et al. MMCU: Multi-modal Chinese Understanding Capabilities of Large Language Models[A]. 2023.

[51] LI H, XU Y, YANG F, et al. CMMLU: Measuring massive multitask language understanding in Chinese[A]. 2023.

[52] WANG X, WANG G, CHEN J, et al. CMB: A Comprehensive Medical Benchmark in Chinese [A]. 2023.

[53] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[A]. 2015.

[54] BANINAJJAR A, RAJABI A, ZHANG X, et al. Distillation Robustness Analysis: Evaluating Vulnerability Transfer in Knowledge Distillation of Large Language Models[J]. Journal of Machine Learning Research, 2024, 25(3): 452-478.

[55] LIU W, ZHANG J, WANG R, et al. Distillation Capability Mapping: Visualizing Knowledge Transfer in Large Language Model Compression[A]. 2023.

[56] YIN H, MARTINEZ C, JOHNSON W. Innovation Patterns in Large Language Model Research: The Impact of Knowledge Distillation on Technical Diversity[J]. Proceedings of Machine Learning Research, 2025, 212: 89-104.

[57] ZHOU M, CHEN Z. DTI: A Framework for Evaluating Knowledge Distillation Transparency in Large Language Models[A]. 2024.

[58] FU W, SAIKH T, XU T, et al. MMMLU: Multimodal Massive Multitask Language Understanding Benchmark[A]. 2023.

[59] LI B, ZHANG R, WAN H, et al. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension[A]. 2023.

[60] CHEN P, LIU W, CHEN L, et al. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models[A]. 2023.

[61] KOJIMA T, GU S S, REID M, et al. Large Language Models are Zero-Shot Reasoners[J]. Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.

[62] 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.

[63] QIN T, DING G, ZHANG H, et al. Text2Reward: Automated Dense Reward Function Generation for Reinforcement Learning[C/OL]//Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence. 2024. https://arxiv.org/abs/2402.01229.

[64] HUANG S, LIN S, ZHOU Y, et al. Exploring the Learning Dynamics of Large Language Models: Insights from Continual Pretraining[A]. 2024.

[65] BANINAJJAR H, TALABI B, MUHLRAD A, et al. On the Robustness of Distilled Large Language Models[A]. 2024.

[66] YIN M, LIN K, WANG Y, et al. Understanding the Robustness Limitations of Distilled LLMs [J]. Proceedings of the International Conference on Learning Representations (ICLR), 2025.

[67] WANG Q, ZHANG C, ZHAO T, et al. Empirical Investigation of Distillation Limitations in Large Language Models[C]//Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). 2024.