Modern Linguistics 现代语言学, 2026, 14(5), 868-877
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145469

Hans汉斯

大模型语言学能力测评数据集构建方法研究

李朝阳

上海师范大学人文学院，上海

收稿日期：2026年4月14日；录用日期：2026年5月18日；发布日期：2026年5月28日

摘　要

本研究探索大模型语言学能力测评数据集的构建方法，提出“出题－参考答案－模型生成回复－考点生成－考点标注－基于考点评分”的标准化流程。以张谊生《现代汉语》为理论依据，借助Kimi设计两套试卷共62题，覆盖语音、句法、语义、语用维度，并选取DeepSeek、豆包、Qwen作答，由Gemini生成考点与评分规则，经人工校对形成最终数据集。研究发现：大模型在自动出题、考点生成与初步阅卷中效率较高，但精准性与规范性仍需人工干预；部分模型能自动识别题目错误，展现知识批判潜力；基于考点的结构化评分比整体打分更具可解释性。本研究为后续大模型语言学能力评测及语言学教学提供了可推广的构建方法与参考基准。

关键词

大模型，语言学，测评数据集，人机协同

Research on the Construction Method and Practice of Evaluation Dataset for Linguistic Ability of Large Language Models

Chaoyang Li

College of Humanities, Shanghai Normal University, Shanghai

Received: April 14, 2026; accepted: May 18, 2026; published: May 28, 2026

Abstract

This study explores the construction method of an evaluation dataset for the linguistic capabilities of large language models (LLMs). A standardized workflow is proposed, consisting of “question drafting - reference answer preparation - model response generation - test point generation - test point annotation - scoring based on test points.” Using Zhang Yisheng’s Modern Chinese as the the-

文章引用: 李朝阳. 大模型语言学能力测评数据集构建方法研究[J]. 现代语言学, 2026, 14(5): 868-877.
DOI: 10.12677/ml.2026.145469

oretical foundation, two test papers comprising a total of 62 questions were designed with the assistance of the LLM Kimi. The questions cover dimensions including phonetics, syntax, semantics, and pragmatics. Three LLMs—DeepSeek, Doubao, and Qwen—were employed to generate responses, while Gemini was utilized to automatically produce test points and scoring rubrics, which were subsequently refined through manual proofreading. The findings indicate that LLMs offer high efficiency in automatic question drafting, test point generation, and preliminary scoring, yet manual intervention remains essential for ensuring precision and standardization. Notably, certain models demonstrated the ability to autonomously identify errors in the question design, revealing their potential for knowledge critique. Moreover, structured scoring based on test points proved to be more interpretable than holistic scoring. This study provides a replicable construction method and a reference benchmark for future evaluations of LLMs' linguistic capabilities and for linguistic pedagogy.

Keywords

Large Language Model, Linguistics, Evaluation Dataset, Human-Machine Collaboration

1. 引言

1.1. 研究背景

大语言模型(Large Language Model, LLM)技术近年来发展迅猛，如何科学、系统地评估其语言理解与生成能力，已成为学界与工业界高度关注的焦点。审视现有的综合评测数据集，虽其覆盖领域较为广泛，可在语言学专项深度测评上仍显薄弱。现有综合评测数据集覆盖领域广泛，但在语言学专项深度测评上仍显薄弱。MMLU (Massive Multitask Language Understanding)、C-Eval (A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models)、CMMLU (Measuring massive multitask language understanding in Chinese)等代表性基准的测评内容难以触及句法歧义、语义指向等深层语言学现象；测评形式也较为单一，尤其缺乏针对主观题的结构化评分机制，从而削弱了评价结果的可解释性。所以，构建一套专用于大模型语言学研究能力的测评数据集，并探索流程化、标准化的构建范式，具有突出的理论价值与现实意义。

1.2. 研究目标

本调研旨在设计并实践一套完整的语言学能力测评数据集构建流程。其中，本调研首先设计了涵盖句法、语义、语用等核心维度的试题及评分基准，再通过调用多款主流大模型收集其原始作答文本。在此基础上，研究重点探究了利用大模型自动生成考点与评分规则的可行性，通过与人工校对版本的对比分析，评估该自动化路径的有效性。以此为基础，再对确定的考点对模型生成的答案开展结构化阅卷，从而客观衡量不同模型在语言学各维度上的表征能力。最终，本调研系统梳理了数据集构建过程中的经验与挑战，进而提炼出一套可推广的标准化构建体系。

1.3. 报告结构

本报告共六章。除本章引言外，第二章系统梳理相关文献，确立研究的理论基础与问题定位；第三

章阐述数据集构建的整体方法论与设计思路；第四章以两套试卷的构建实践为例，逐环节呈现具体操作路径，并辅以典型案例；第五章从数据质量、自动化程度、阅卷效能及模型表现四个维度进行综合评估；第六章总结全篇并展望未来方向。

2. 现有大模型评测数据集概述

2.1. 现有大模型评测数据集概述

当前主流的大模型评测数据集在广度与深度上各有侧重，但均难以完全满足现代汉语专门性语言学评测的需求。例如 MMLU [1]，虽其横跨诸多学科，但分配至语言学分支的题量有限，且多局限于基础知识类的选择题；在专项评测领域，BLiMP (Benchmark of Linguistic Minimal Pairs) [2]以英语句法现象为核心，采用最小对立法进行考察，却未向语用及语义推理等高阶语境层面延伸。再看中文语境中的 CLUE (Chinese Language Understanding Evaluation) [3]系列数据集，其涵盖多项自然语言处理任务，但核心驱动力在于自然语言理解，而非系统性的语言学本体知识测评。近年来学界也提出了如 LHMKE (Large-scale Holistic Multi-subject Knowledge Evaluation)等包含主观题的大规模多学科知识评测基准[4]，虽然提升了主观题和知识广度，但在现代汉语专项的深层语言学现象覆盖上仍显不足。所以，学界目前亟需一套能够系统覆盖现代汉语核心知识体系，且具备主观题结构化评分机制的中文语言学能力测评数据集。

2.2. 语言学能力评价的核心维度

语言学能力的评估需建立在多维度的科学框架之上。本调研参照张谊生(2013)编纂的《现代汉语》教材知识体系[5]，将测评核心维度设定为以下四个层面：

语音能力：包括语音属性、语音要素、音素分析等；

句法能力：包括短语类型、句子成分、特殊句式、层次分析、歧义分析等；

语义能力：包括语义角色、一词多义、同义词辨析、语义搭配、义素分析等；

语用能力：包括语境理解、会话含义、指示语、语体差异、言语行为等。

上述测评维度既覆盖了语言学的基础知识，也包含了需要深度分析的应用能力，能够较为全面地评估大模型的语言学综合素养。

2.3. 自动化评测与人工评测相结合的现有方法

在主观题的评阅机制上，现有研究已揭示出单一评估视角的局限性：纯粹依赖大模型进行自动评分，往往难以精准把握主观题作答的细微质量差异；而完全依赖人工评测，则面临成本高昂且效率低下的瓶颈。为破解这一困境，本调研借鉴业内前沿思路，如 ChatEval 框架中提出的多智能体协同评估机制[6]，尝试引入大模型作为“评审员”进行初步评审与打分，随后辅以人工专家进行严密复核。李东进(2020)也在针对专业文本的可解释评阅研究中指出，基于客观知识点命中情况进行细粒度的结构化评分，能够有效提高主观题评阅的公正性与可解释性[7]。本调研在考点生成和阅卷两大核心环节均贯穿了这种“模型初评 + 人工校验”的双轨制模式，旨在最大程度上实现评测效率与评分准确性的最优平衡。

3. 数据集构建方法论设计

3.1. 整体流程设计

本数据集的构建依托于一个严密的闭环流程，各环节依次推进、深度衔接。构建过程始于试题设计，要求题目紧扣教材核心知识点，既要全面覆盖语言学的核心维度，也要保证难度梯度的合理分布。考虑到当前大模型评测普遍面临的数据污染(Data Contamination)与死记硬背问题，本调研在设计时也吸取了

MMLU-CF 等防污染评测集的构建经验[8]，通过原创命题、情境改编等方式尽量规避训练数据的直接泄露。在确保评测的信度的前提下，为每道题目撰写参考答案与评分要点，确立满分判定的绝对基准。在收集数据阶段，通过统一的提示词引导预设的多个大模型生成原始回复内容。下一步进入自动化处理的核心环节，即引入“评审员模型”依据题干与参考答案自动生成初始考点列表。由于模型生成的颗粒度缺陷，语言学学者随后介入，对考点进行人工标注与多维校对，凝练出最终的考点清单与细化评分规则。最后，基于上述确立的标尺，采用“模型初评叠加人工快速复核”的双轨制开展结构化阅卷，输出最终的测评得分。

3.2. 核心环节定义与目标

在上述流程的具体实践中，各个环节均设定了明确的操作规范与目标指引。

命题与基准确立：题目设计要严守语言学理论规范以规避歧义，也要充分预估大模型的潜在作答边界。本次实践借助 Kimi K2.5 辅助生成了两套试卷：综合卷侧重广度，考察语言学基础知识；专题卷则侧重深度，检验复杂的语言分析能力。配套的参考答案则拆解为“满分示例”与“核心要点”，为自动化评分提供硬性抓手。

模型作答与考点生成：研究选取 DeepSeek-V3.2、豆包 2.0 及 Qwen3.6-Plus 三款主流中文大模型作为受测对象，统一赋予“汉语言文学专业本科生”的提示词角色，以激发其深度思考过程。在考点提取环节，利用 Gemini3.1 充当虚拟评审员，要求其紧扣语言学概念生成包含建议分值的初始考点库。

人工校验与双轨阅卷：专家介入不仅负责审核考点，更关键在于补充如“缺项扣分”等精细化规则。在最终的阅卷执行层面，先由评审员模型执行初步量化打分，随后人工对照考点清单进行极速复核，全程记录机评与人评的分数偏离度。

3.3. 本次实践的数据集规模与领域分布

本次构建实践产出的两套试卷在题型分布与测评倾向上形成互补矩阵：

第一套试卷(综合卷)：共设 38 题，涵盖填空(10 题 20 空)、选择(10 题)、判断(10 题)、分析(5 题，涉及音素、字词结构及句法层次等)、简答(2 题)与论述(1 题)。该卷广泛辐射了语音、文字、词汇、语法及修辞等基础知识体系。

第二套试卷(专题卷)：共设 24 题，均衡划分为句法、语义及语用三大模块(各 8 题)。题型精简为选择、简答与论述，旨在深度刺探模型的长文本分析与高阶推理能力。

4. 构建实践：分步详解与案例分析

4.1. 步骤一：题目与参考答案设计

在专题卷的命题设计中，本调研尤其注重对语言学核心机制的挖掘。在第二套试卷(专题卷)第 8 题(句法歧义分析)中，题干要求剖析“开刀的是他父亲”一句的歧义类型，并利用句式变换法加以分化。配套的参考答案精准锚定了歧义根源在于“开刀的”施受关系模糊，并清晰界定了两种语义：其一，父亲为施事(医生给病人开刀)；其二，父亲为受事(病人被医生开刀)。答案还提供了相应的句式变换示例，为后续评分设立了评判标准。

4.2. 步骤二：收集模型回复

在答题环节，笔者让三款受测模型(DeepSeek-V3.2、豆包 2.0 Pro、Qwen3.6-Plus)都开启了“深度思考”模式。回顾第二套试卷(专题卷)第 8 题，DeepSeek-V3.2 准确拆解了两种对立语义，还深入剖析了歧

义生成的句法机制，并给出了规范的替换句式，展现出极为扎实的现代汉语底层分析功底。

4.3. 步骤三：考点的生成与标注

本环节是实现自动化测评的关键转折点。利用 Gemini3.1 生成的初始考点往往暴露出颗粒度粗糙的短板：一是考点表述笼统化，缺乏对具体语言学理论依据的限定；二是分值分配粗放化，未能根据作答要素的复杂度进行细粒度拆解；三是评分规则模糊化，缺乏区分答题质量的约束性条款。针对前文中的第二套试卷(专题卷)第 8 题，模型版考点仅笼统概括为“句法歧义分析(意思 2 分，分化方法 3 分)”，既未限定歧义类型的界定标准，也缺乏对分化方法合理性的评判细则。

所以，本调研引入语言学学者进行人工校验，重点从三个维度对考点进行重构：

表述精确化：注入具体理论依据。如第 14 题(语义指向)，人工版强制要求“结合状语的定义与特征进行作答”，从而引导测评向理论深度下沉，而非停留在语感层面。

评分细化：将模糊评价转化为可量化的采分点。如第 21 题(信息结构)，将原有的“逐项回答”拆解为“强调/正常各 2 分、新信息位置 2 分、新旧信息关系 2 分”，大幅提升了评分的实操性。

Table 1. System resulting data of standard Comparison table of revised examination points for test paper 2
表 1. 第二套试卷考点修订对比表

题号	Gemini3.1 生成版	人工校对版	修订说明	修订类别
1	**核心考点**：短语类型辨析 **评分规则**：选对得 3 分	**核心考点**：短语类型辨析 **评分规则**：选对得 2 分	原分值偏高，与其他选择题统一调整为 2 分，使整卷分值分布更均衡。	分值调整
2	**核心考点**：词性辨析 **评分规则**：选对得 3 分	**核心考点**：词性辨析 **评分规则**：选对得 2 分	同上，选择题统一为 2 分。	分值调整
7	**核心考点**：句式变换 **评分规则**：答出施受关系变化及标记词	**核心考点**：句式变换 **评分规则**：答出施受关系变化及标记词(2 分)，再简单阐述一下“被”字句的特点在本题中的体现(3 分)	原规则过于笼统，无法区分答题深度；增加“被”字句特点的阐述要求，能更好考察学生对被动句的理解。	分值调整 + 评分细化
14	**核心考点**：语义指向 **评分规则**：判断 2 分，理由 3 分	**核心考点**：语义指向 **评分规则**：判断 2 分，理由 3 分。需要通过状语的定义、特征等相关知识点进行作答	增加“需结合状语知识作答”的提示，引导学生在理论中运用理论，避免仅凭语感答题。	评分细化
16	**核心考点**：歧义与语境 **评分规则**：原因 2 分，举例 3 分	**核心考点**：歧义与语境 **评分规则**：原因 3 分，举例 2 分。原因部分细化：“句法结构不同导致歧义。‘炒肉丝’可以作‘我要’的宾语，‘炒’也可与‘要’一起组成谓语，让‘肉丝’作宾语。”	原评分原因权重过低，且未明确原因要点；现加重原因分值并给出具体分析方向，使评分更具操作性。	分值分配调整 + 评分细化
21	**核心考点**：信息结构 **评分规则**：逐项回答，逻辑清晰	**核心考点**：信息结构 **评分规则**：强调 2 分，正常 2 分。意图中，需要说明“新信息”在句子中的位置(2 分)及“已知信息”和“新信息”的关系(2 分)	原规则模糊，无法量化；现拆分为三个得分点，明确考察话题与信息结构的核心要素。	评分细化
23	**核心考点**：语体差异应用 **评分规则**：场景对应准确	**核心考点**：语体差异应用 **评分规则**：不同点 2 分，举例 3 分，不可用题中所给的例子	原规则无法区分回答质量；现要求自主举例并禁止使用题目原例，避免机械重复，真正考察应用能力。	评分细化 + 增加约束
24	**核心考点**：语境的功能 **评分规则**：结合实例论述完整	**核心考点**：语境的功能 **评分规则**：语境的定义 1 分，两个案例各 1 分，语境的作用 2 分	原规则主观性强；现拆分为定义、案例、作用三部分，使评分客观清晰，便于阅卷者操作。	评分细化

增加约束：针对模型版评分规则无法区分回答质量的问题，补充具体作答约束，如第 23 题“语体差异应用”，模型版仅要求“场景对应准确”，人工版则补充“不可用题中所给的例子”，避免机械照抄，真正考察应用能力(见表 1)。

4.4. 步骤四：基于考点的结构化阅卷

4.4.1. 制定阅卷规则

在确立人工版考点清单后，研究制定了微观层面的阅卷法则。在第二套试卷(专题卷)第 8 题中，评分细则被极度量化：意思 1 与意思 2 各占 1 分；提供一种正确变换得 2 分，两种得 3 分；变换错误则触发扣分机制。

4.4.2. 模型初评与人工复核

在实际评阅中，评审模型 Gemini3.1 根据该细则生成初评报告。人工复核时重点关注初评中与规则不符的评分，并记录差异。豆包 2.0 在第二套试卷第 15 题的答案中，虽然思路清晰，但未使用标准的矩阵格式，不符合义素分析的要求。模型初评未能识别该形式错误给满分 5 分，人工复核介入之后，扣除 3 分格式分，最终修正为 2 分，符合学术规范。

4.5. 步骤四：典型案例分析：第一套试卷第 13 题的“题库错误识别”

在本次测评实践中，Qwen3.6-Plus 展现出了突出的题目错误识别能力，第一套试卷第 13 题的作答过程为这一能力提供了典型案例，以下从问题描述、模型推理过程、教学与测评意义三方面展开分析。

4.5.1. 问题描述

下列各组词语中，韵母完全相同的一组是(　)

A. 真正、神圣　　B. 春风、东风

C. 前程、深情　　D. 生动、隆重

第一套试卷选择题第 3 题为单项选择题，要求选出“韵母完全相同的一组词语”。原卷预设答案为 C (前程、深情)，但实际正确选项应为 A (真正、神圣)。三个模型中，仅 Qwen3.6-Plus 通过严密的音节拆解，准确识别此错误并选 A，而原答案 C 实则韵母不同。

4.5.2. 模型推理过程分析

Qwen3.6-Plus 在答案中逐项分析了每个选项的音节，具体如下：

A：真正(zhēn zhèng)韵母 en、eng；神圣(shén shèng)韵母 en、eng→完全相同。

B：春风(chūn fēng)韵母 un、eng；东风(dōng fēng)韵母 ong、eng→不同。

C：前程(qián chéng)韵母 ian、eng；深情(shēn qíng)韵母 en、ing→不同。

D：生动(shēng dòng)韵母 eng、ong；隆重(lóng zhòng)韵母 ong、ong→不同。

此过程展示了模型对汉语拼音的精准掌握和逻辑推理能力，做出了正确判断。

4.5.3. 教学与测评意义

这一案例对数据集构建和模型能力评估均具有重要意义：

在数据集构建方面，模型的反馈能够帮助研究人员及时发现题目错误，进而优化题库质量，本次案例中，笔者根据模型反馈将原题答案修正为 A，并在考点中补充了韵母辨析的细化规则。

在模型能力评估方面，Qwen3.6-Plus 的表现证明，部分前沿大模型已跨越单纯的知识调用阶段，初步具备了高阶的知识批判与漏洞检测能力，这理应被纳入未来高层次语言学能力测评的重要维度。

5. 效果评估与讨论

5.1. 数据集质量与覆盖度评估

本调研构建的语言学能力测评数据集共包含 59 道试题，涵盖了语音、文字、词汇、句法、语义及语用等现代汉语核心分支。题目设计以张谊生版《现代汉语》为理论依据，确保了考点的专业性。我们分析后发现，数据集在句法层次分析、语义指向分析及语用原则应用等高阶任务上具有较高的区分度。经过人工校对的考点清单，修正了模型生成时的表述偏差，还通过细化采分点提升了测评的信度。

5.2. 自动化程度评估

本次构建实践在出题、考点生成、阅卷三个环节引入了大模型自动化操作，各环节的自动化效果各有差异，具体评估如下：

出题环节中，Kimi K2.5 生成的两套试卷题目质量较高，但存在一些问题，如第一套试卷的分析题和第二套试卷的部分简答题的题干后方都会出现带有提醒内容的考点，这一问题会影响对考生真实水平的考察。可以看出，大模型自动出题能提升效率，但仍需要人工审核。

考点生成环节中，模型自动生成的考点粗粒度居多。试看第二套试卷，共涉及 24 个考点，其中 9 个(37.5%)需要人工修正以调整表述精确化、分值赋予及评判规则等细节。但模型提供了基础框架，相较于人工从零开始设计考点，可以显著减轻工作量。

阅卷环节中，人机协作模式由两个阶段构成：第一阶段为评审模型依据既定考点和评分规则进行初评打分；第二阶段为人工对照考点清单进行复核修正。此处的人工复核不同于纯人工批改——后者需要阅卷者独立完成从文本阅读、考点匹配到分值判定的全过程，前者只需要对模型已完成的考点匹配和分值建议进行审核校验。模型初评与人工复核的一致性约为 90%。在主观题上，模型对答案要点把握较准，但对于是否缺项、举例是否恰当等细节容易出现误判，而人工复核可快速修正。总的来说，人机协作的模式比纯人工阅卷的效率提升了很多。

5.3. 阅卷效率与一致性分析

第二套试卷第 24 题(语境分析)的阅卷效率对比如下：纯人工批改 3 份答案平均耗时 15 分钟，流程为：阅卷者独立阅读答案、参照张谊生版《现代汉语》教材相关知识点进行判断、按照参考答案逐项赋分。人机协作模式下，模型初评耗时 1 分钟完成考点匹配与分值建议，人工复核耗时 5 分钟核验匹配结果并修正少数误判，合计 6 分钟。两种模式的最终评分均经由人工确认，但人机协作将人工从重复性的细粒度考点匹配中解放出来，使人工可以聚焦于审核与修正，整体效率提升约 60%。人工复核环节确保了评分的准确性，避免了模型可能出现的疏漏。

5.4. 测评结果研究

5.4.1. 各模型总分对比(第二套试卷)

各模型得分情况见表 2。

Table 2. Comparison of the total scores of each model for test paper 2
表 2. 第二套试卷各模型总分对比

模型	句法部分(30 分)	语义部分(30 分)	语用部分(40 分)	总分(100 分)
DeepSeek-V3.2	27	28	40	95
豆包 2.0	26	26	36	88
Qwen3.6-Plus	29	30	40	99

5.4.2. 能力维度分析

句法部分(满分 30 分)：Qwen3.6-Plus 得 29 分(96.7%)，在层次分析、句法成分辨认等精细任务上表现最佳，DeepSeek-V3.227 分(90.0%)次之，豆包 2.0 得 26 分(86.7%)，在“被”字句特点阐述上稍显简略，得分略低。

语义部分(满分 30 分)：Qwen3.6-Plus 得 30 分(100%)，对义素分析、语义指向等理论性题目回答完整；DeepSeek-V3.228 分(约 93%)次之，豆包 2.026 分(约 86.7%)，因其在义素分析上未使用矩阵格式，导致扣分。

语用部分：三个模型均能正确理解会话含义，掌握基础的语用知识，DeepSeek-V3.2 和 Qwen3.6-Plus 在论述语用原则权衡时都展现出了一定的理论深度(40 分，100%)，豆包 2.0 则偏重表面描述，导致扣分(36 分，90%)。

5.4.3. 模型特点总结

结合两套试卷的测评结果，三款主流中文大模型的语言学研究能力各有鲜明特点：

Qwen3.6-Plus：答案最为严谨，学术性强，善于结合语言学理论开展分析，且具备突出的知识批判能力，能够发现题目中的错误。

DeepSeek-V3.2：基础知识扎实，逻辑清晰，但在部分分析题的作答规范性上稍显不足。

豆包 2.0：答题详尽，注重日常表达，但在专业术语使用和作答标准化方面还需加强。

5.5. 数据污染风险讨论

本研究在设计阶段参考 MMLU-CF 等防污染评测集的构建经验，通过原创命题与情境改编尽量降低数据泄露风险，但大模型训练数据的透明性不足使得数据污染的可能性难以完全排除，有必要就其对测评结果的潜在影响进行专门讨论。

5.5.1. 原题泄露的可能性分析

本研究两套试卷均由人工依据张谊生版《现代汉语》教材设计，题目经由 Kimi K2.5 辅助生成后经过人工修改和筛选。试题内容涉及句法歧义、语义指向、语体差异等具体语言学现象分析，并非直接取自公开题库。因此，受测模型在预训练阶段直接接触原题的概率较低。即使存在部分知识点的间接覆盖，模型也需在实际作答中完成从知识点到具体题目的迁移，这与常规多项选择题(Multiple Choice Question, MCQ)基准测试中通过选项匹配即可作答的情形有本质区别。

5.5.2. 改写与记忆的可能影响

模型即使未接触原题，若其训练语料中包含与原题高度相似的表述，也可能产生近似记忆效应，从而对评测结果造成干扰。本研究中专题卷以主观题为主，要求模型进行长文本分析与推理，作答形式为开放性段落论述，而非选择固定选项，这降低了模型通过表面词汇匹配获得高分的可能性。从三款模型的实际作答内容观察，其推理路径与论述逻辑各有差异，未发现机械复制特定文本段落的现象。

5.5.3. 模型知识批判能力的反证

第一套试卷第 13 题预设答案存在错误，预设选项 C (前程、深情)韵母并不相同，实际正确选项应为 A (真正、神圣)。Qwen3.6-Plus 通过逐项分析各选项的韵母构成，准确指出这一错误并给出正确选择。若模型依赖对题目答案的记忆作答，应直接复现预设的错误答案，而非通过音节拆解发现并纠正错误。这一行为表明该模型在解题过程中调用了真实的语音分析能力，而非依赖记忆。

5.5.4. 局限性与未来改进

尽管上述分析表明数据污染对本研究结果的影响相对有限，但本研究尚未引入系统化的污染检测方

法，无法给出定量结论。未来的数据集构建可考虑纳入定量的污染检测指标，如计算模型作答与训练数据之间的最小 KL 散度(Kullback-Leibler Divergence)、检测 N-gram 重叠率等，以识别潜在的污染信号。此外，可在测评流程中设置专门的污染探测题目，通过观察模型对故意植入的错误或改写题目的响应模式，进一步验证模型在多大程度上依赖记忆而非依靠理解进行作答，增强测评结果的可信度。

6. 结论与展望

6.1. 主要结论

本调研通过两套《现代汉语》试卷的构建实践，验证了一套包含“出题－写参考答案－模型生成－考点生成－人工标注－基于考点评分”六步骤的语言学能力测评数据集构建方法的可行性。该方法经实践检验，具有以下优势：

第一，试题设计能够覆盖语言学核心维度，题型多样、难度梯度合理，可以对大模型的语言学能力进行全面评估。

第二，利用大模型完成自动出题、生成考点和初评打分，降低了人工工作量，提升了数据集构建的整体效率。

第三，人工校对环节能够有效修正模型生成内容的问题，确保考点和评分的准确性，提升数据集的整体质量。

第四，基于考点的结构化评分方式，让测评结果更具可解释性，便于研究人员分析模型在各语言学维度的能力短板。

6.2. 应用价值

本次研究构建的语言学能力测评数据集及对应的构建方法，兼具理论与实际应用价值：第一，为大语言模型在垂直学术领域的性能对标提供基准工具；第二，相关考点库与细化规则可直接转化为语言学教学中的自动批改参考标准；第三，利用模型表现出的查错能力，促进数据集质量的动态迭代。

6.3. 未来展望

结合本次研究的结论与遇到的挑战，未来可从三个方面开展进一步研究，完善语言学能力测评数据集的构建体系：

样本规模：未来需进一步扩充试题容量，覆盖更多边缘语言现象及古代汉语等细分领域。

自动化算法：下阶段可探索利用微调后的专项模型担任“评审员”，以期降低考点生成环节对通用大模型的依赖。

多模态融合：考虑引入语音、图像等多模态语料，构建更全面的语言交际能力测评体系。本调研通过规范化的流程设计，为语言学与大模型交叉领域的评测工作奠定了基础，未来将持续推动数据集的开源与共建。

参考文献

[1] Hendrycks, D., Burns, C., Basart, S., *et al*. (2021) Measuring Massive Multitask Language Under-Standing. 2021 *International Conference on Learning Representations*, Online, 3-7 May 2021, 1-27.

[2] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S., *et al*. (2020) BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, **8**, 377-392. https://doi.org/10.1162/tacl_a_00321

[3] Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., *et al*. (2020) CLUE: A Chinese Language Understanding Evaluation Benchmark. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, December 2020,

4762-4772. https://doi.org/10.18653/v1/2020.coling-main.419

[4] Liu, C., Jin, R., Ren, Y. and Xiong, D. (2024) LHMKE: A Large-Scale Holistic Multi-Subject Knowledge Evaluation Benchmark for Chinese Large Language Models. *Proceedings of the Language Resources and Evaluation Conference*, Torino, May 2024, 10476-10487. https://doi.org/10.63317/3582yurtjaq3

[5] 张谊生. 现代汉语[M]. 第 2 版. 上海: 复旦大学出版社, 2013.

[6] Chan, C.M., Chen, W., Su, Y., *et al*. (2023) ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 13371-13391.

[7] 李东进. 基于知识点的专业文本可解释评阅研究[D]: [硕士学位论文]. 济南: 山东大学, 2020.

[8] Zhao, Q., Huang, Y., Lv, T., Cui, L., Sun, Q., Mao, S., *et al*. (2025) MMLU-CF: A Contamination-Free Multi-Task Language Understanding Benchmark. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vienna, July 2025, 13371-13391. https://doi.org/10.18653/v1/2025.acl-long.656

附　录

本研究实践过程中生成的全部数据文件清单如下。限于篇幅，正文中未全文呈现。读者如需获取相关原始数据或批改详情，可联系作者索取。

第一套试卷(综合卷)数据文件：

1. Kimi 生成《现代汉语》期末考试试卷.docx
2. Kimi 生成的期末答案.docx
3. 三个大模型生成的答案(第一套试卷).docx
4. 考点列表及评分规则(Gemini 生成).docx
5. 考点列表及评分规则(人工修订).docx
6. 答案批改报告(Gemini 生成).docx
7. 答案批改报告(人工修订).docx
8. 典型案例深度分析.docx

第二套试卷(专题卷)数据文件：

1. Kimi 生成现代汉语试题(第二套试卷).docx
2. 三份答案(含深度思考)(第二套试卷).docx
3. 三个大模型生成的答案(第二套试卷).docx
4. 考点列表和评分准则(Gemini 生成).docx
5. 考点列表和评分准则(人工修订).docx
6. 考点与评分规则修订对比表(人工修订).docx
7. 答案批改报告(Gemini 生成).docx
8. 答案批改报告(人工修订).docx