

汉语中介语语料库“AI助手”辅助检索功能的模拟与可行性初探

殷婷婷

新疆师范大学国际文化交流学院, 新疆 乌鲁木齐

收稿日期: 2026年8月17日; 录用日期: 2026年9月19日; 发布日期: 2026年9月24日

摘要

当前主流汉语中介语语料库的检索系统多依赖“字符串一般检索”及“分类标注检索”等模块, 存在操作逻辑割裂、多条件组合困难、典型偏误排序缺失等实际痛点。本文基于社交平台收集的语料库使用者反馈, 发现大量用户在“理解偏误标注代码”、“进行复杂语法条件检索”以及“查看错字原貌”等方面存在普遍焦虑。借鉴电商平台“AI助手”的自然语言交互逻辑, 本研究在不具备后台改造权限的前提下, 以实际下载的300条“把字句”中介语语料为实验对象, 通过设计特定提示词(Prompt), 对大模型的“意图解析、跨维度语义召回、典型度推荐”等能力进行了实证模拟。实验发现, AI助手能有效破解传统搜索页面的功能限制, 并针对AI误判提出了“强制匹配偏误标注符”的人机协同约束策略。本研究验证了轻量化AI助手模式在语料库智能化升级中的实际应用价值。

关键词

汉语中介语语料库, AI助手, 大语言模型, Prompt实证, 把字句

Simulation and Preliminary Feasibility Study of “AI Assistant” Retrieval Functions in Chinese Interlanguage Corpora

Tingting Yin

College of International Cultural Exchange, Xinjiang Normal University, Urumqi Xinjiang

Received: August 17, 2026; accepted: September 19, 2026; published: September 24, 2026

Abstract

Most retrieval systems for mainstream Chinese interlanguage corpora mainly adopt modules such

as general string retrieval and categorical annotation retrieval, which are plagued by practical drawbacks including disjoint operational logic, cumbersome multi-condition combination, and the absence of ranking mechanisms for typical errors. Based on feedback collected from corpus users via social media platforms, this paper identifies widespread user frustrations regarding the interpretation of error annotation codes, execution of complex grammatical conditional retrieval, and access to the original forms of miswritten characters. Drawing on the natural language interaction logic of AI assistants deployed on e-commerce platforms, this study takes 300 downloaded interlanguage corpus entries of the ba-construction as experimental materials without access to backend system modification privileges. By designing dedicated Prompts, it conducts an empirical simulation of large language models' capabilities in intent parsing, cross-dimensional semantic recall and typicality recommendation. Experiments verify that the AI assistant can effectively break through the functional constraints of conventional search interfaces. To address AI misclassification errors, a human-machine collaborative constraint strategy featuring forced matching of error annotation markers is proposed. This research confirms the practical applicability of the lightweight AI assistant model for the intelligent upgrading of Chinese interlanguage corpora.

Keywords

Chinese Interlanguage Corpus, AI Assistant, Large Language Model (LLM), Prompt-Based Empirical Study, Ba-Construction

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

汉语中介语语料库经过三十余年的发展,已逐渐成为国际中文教育学科进行偏误分析、习得研究和教学法探索的核心基础设施。回顾语料库的建设历程,张宝林(2019)将语料库建设划分为“1.0 粗放时代”和“2.0 精细时代”,指出 2.0 时代的显著特征是系统制作精良、功能丰富多样,更加注重“用户使用方便”[1]。然而,尽管语料库的“库存量”日益庞大,标注体系日趋精细,但学术界普遍感受到一个现实矛盾:语料库里的语料极为丰富,但一线教师和研究生要想精准“挖掘”出符合特定研究需求的语料,依然面临着较高的技术门槛。

从“全球汉语中介语语料库 V1.0”等现有平台的实际使用来看,传统语料库的检索系统通常被划分为“字符串一般检索”、“分类标注检索”、“离合词检索”等多个相对独立的模块。张宝林(2021)在透视语料库检索系统时,深刻指出当前以“形式检索”为主的检索方式严重限制了用户对复杂语义和多条件组合检索的获取能力[2]。张瑞朋(2012)在建设中山大学汉字偏误中介语语料库的过程中,也明确指出了“检索工具的简易使用问题”的迫切性[3]。对于大多数没有计算机后台权限的语言学专业研究生而言,期盼直接去“改造”现有语料库的底层代码或重新设计检索系统,是不现实的。

近年来,人工智能技术的飞速发展为此困境提供了全新的解题思路。值得注意的是,淘宝、小红书、百度等主流互联网平台,都相继推出了内嵌的“AI 助手”。用户不再需要掌握复杂的搜索引擎语法,只需用日常语言向 AI 助手提问,系统便能自动理解用户意图,并精准匹配与推荐最合适的结果。这种“AI 助手”的自然语言交互模式,恰恰是语料库检索系统目前亟须借鉴的交互理念。

正因如此,本研究放弃了“纸上谈兵”的系统蓝图设计,立足于客观条件限制,依托真实的“全球汉语中介语语料库”网页端操作流程,以及手头现有的从全球汉语中介语语料库下载的 300 条带有详细偏

误标注的“把字句”语料,结合当前可免费使用的网页端大语言模型(LLM),进行一场“AI 助手辅助语料库检索”的真实模拟实验。我们试图通过具体的 Prompt (提示词)操作和真实的检索场景对比,论证这种轻量化的“AI 助手”模式能否成为语料库用户最实用、最接地气的检索辅助工具。

2. 研究现状与理论基础

(一) 汉语中介语语料库的发展脉络与检索困境

回顾国内汉语中介语语料库的建设历程,我们可以清晰地看到从粗放到精细的演进脉络。张宝林(2022)在文献中再次强调了当前语料库建设已迈入“2.0 时代”,精细化、系统化和用户友好是发展的重要方向[4]。随着语料库规模和标注质量的提升,语料资源的共享和利用问题也成为了学界关注的焦点。张宝林(2022)指出,由公共资金建设的语料库如果不能向社会开放,就会造成严重的学术资源浪费和低水平重复建设[5]。

然而,语料库的功能不仅在于“建得好”,更在于“用得顺手”。张宝林(2022)在关于语料库软件系统的深入思考中特别提到,语言学专业应更多地参与到软件系统的研发工作中,而不能把这一任务完全推给技术人员[6]。张宝林、崔希亮(2022)在论述“全球汉语中介语语料库”的特点时也指出,检索方式的设计直接影响语料库的学术价值[7]。谭正娇、王文文、余晓铃(2021)通过热点分析指出,语料库的易用性和应用转化是该领域亟需突破的瓶颈[8]。

但现实是,当前主流检索系统依然要求用户在多个独立的检索选项卡(如“字符串一般检索”、“分类标注检索”等)之间来回切换,并手动输入复杂的检索条件。

(二) 大语言模型 Prompt 工程与垂直领域应用的可行性

大语言模型(LLM)的爆发式发展,让“零代码智能交互”成为了可能。研究者即使没有计算机编程基础,也无需进行系统级的 API 对接,只需设计一套精心构造的提示词(Prompt),就能让大模型扮演特定的角色,执行具体的任务。郑通涛、曾小燕(2016)在大数据时代背景下提出,语料库建设应引入“网络智能挖掘技术”,遵循真实性、广泛性等原则[9]。林君峰(2015)也提出过利用云计算等前沿技术减轻人工负担、提高语料库建设效率的设想[10]。

这些研究表明,AI 技术与垂直领域的融合是大势所趋。本研究正是将这一技术思路迁移到语料库检索中,试图通过对大模型进行特定角色的“Prompt 设定”,将其“训练”成一个懂汉语中介语、懂偏误标注规则、能进行语义检索与个性化推荐的“语料库 AI 助手”。这种脱离了底层代码开发、仅靠现有工具进行“人机协同”的工作方式,对于没有系统权限的一线研究者而言,具有极高的实践参考价值。

3. 基于真实网页平台的传统检索操作困境

为了直观呈现传统检索在“全球汉语中介语语料库 V1.0”这类真实平台中的具体痛点,我们直接对照真实的网页系统交互逻辑进行分析。

(一) 检索模块割裂,多维度条件组合极其困难

分析真实的检索页面,该系统将检索功能划分成了“字符串一般检索”、“分类标注检索”、“离合词检索”等多个标签页,且每个标签页下方的可选条件(作者国籍、母语、产出日期、HSK 等级、汉语水平等)在默认状态下往往相对孤立。

具体困境:假设一位研究生想要检索“韩国国籍、HSK 四级、在议论文写作中使用了‘把字句’且出现错词偏误”的语料。在真实操作中,他首先要进入“字符串一般检索”,输入“把”字;然后在下方的“作者国籍”选择“韩国”,在“HSK 等级”下拉框选择“参加 HSK 4 级”。如果系统将这些条件进行“且”的逻辑检索,往往返回结果为空,或者因为某些语料缺失“HSK 等级”这一元数据而直接报错。

如果转而使用“分类标注检索”去定位“错词”[Cb], 又会丢失刚才设定的国籍和 HSK 等级信息。这种检索逻辑的割裂, 是语料库页面的首要痛点。

(二) 偏误标签的不确定性导致“查不全”

在真实检索页面的“分类标注检索”中, 用户要想精准找到“把字句动词缺失”的语料, 必须精确知道该语料库使用的标注代码(如[Cb]、[Cq]、[Cd]等)。如果输入“遗漏”, 系统可能返回极少, 甚至 0 条结果, 因为不同批次的语料标注人员可能使用了不同的代码(如[Cb]错词, 但本质上是动词缺失)。标注的全面性和类别问题直接关系到语料标注的科学性和可行性[11]。

在我们的真实语料中就有这样的例子:“把它的好处【来到】[Cb 带给]我们”。如果传统检索页面的“分类标注检索”要求用户精确指定“错词”且输入“把”, 那么这条语料会因为同时包含“错词”和“介词用错”而被混在庞大的噪音数据里。而如果用户只想找“动词缺失”, 却因为不知道标注者用代码还是用汉字, 会导致有价值的语料直接漏掉。

(三) 结果排序机制缺失, 缺乏“典型偏误”的引导

底部只显示了“序号”和“检索原句”。当用户成功检索出几百甚至上千条含有“把”字的结果时, 系统并没有提供按照“偏误典型程度”或“错误出现频次”进行排序的功能。正确句与错误句混杂在一起, 罗列在下方, 用户只能一页页翻看, 靠肉眼扫描去确定“哪一类的偏误出现频率最高”。这种缺乏智能排序和个性化推荐的呈现方式, 严重增加了研究者人工筛检语料的时间成本, 极大地降低了语料库的实际使用效率。

4. 基于网络使用者反馈的语料库使用困境分析

虽然官方已经为“全球汉语中介语语料库”发布了详细的使用说明和代码解释, 但这种以静态文档形式存在的帮助文件, 并没有从根本上解决用户在实际操作中面临的即时性认知障碍。为了进一步印证传统检索方式在实际使用中的困境, 本研究搜集了社交媒体上关于该语料库的公开求助与反馈信息。这些来自一线学习者和研究者的真实声音, 直观地反映了“静态文档”与“动态操作”之间存在的鸿沟。

(一) 静态代码表导致认知断层, AI 助手可实现“即问即答”

官方提供了详细的偏误标注代码解释表, 明确了[Jxwx]、[Jzxy]、[JSx]、[Zc]等代码的具体含义。汉字偏误标注的全面性和科学性直接关系到语料库的价值[12], 然而真实的检索场景具有极强的“打断性”。当一位使用者在检索结果中看到一长串难以辨认的代码时, 他需要中断检索工作, 去翻阅几十页的代码表或网页指南进行寻找。

小红书平台用户发帖:“有没有人知道全球汉语中介语语料库的错误句式里这些字符是什么意思呀”。这句话说明, 即使官方文档里写了, 用户依然不想或没有精力去翻查说明书, 真实地反映出这种断层。这种痛点正是 AI 助手的用武之地——AI 助手可以在用户看到[Jxwx]时, 直接通过对话框回答“这个代码表示该句存在‘句法疑问’”, 将几十页的被动翻阅变成了即时响应的主动知识库。

(二) 复杂句法检索的“无效性”焦虑

除了代码的晦涩, 用户在面对具体、复杂的检索需求时往往感到无从下手。例如, 另一位使用者发帖表达了深深的绝望:“求救全球汉语中介语语料库: 如果想查与量词‘双’搭配的名词该怎么检索? 难道是搜一个‘双’字然后挨个手动分类吗?”

这句“挨个手动分类”的绝望提问, 直击了当前检索系统缺乏“语义理解”的软肋。传统的“字符串一般检索”只能根据字面关键词机械匹配。用户想要查找“与量词‘双’搭配的名词”, 实际上这涉及了“词性搭配”和“句法结构”的跨维度关联, 传统系统根本无力支撑这种“基于意图的复合检索”, 只能强迫使用者进行低效的人工劳作。这恰恰印证了“AI 助手”在“意图解析”方面的必要性——用户只需要像平

时说话一样提问, AI 就能自动理解用户要找的是“名词搭配”, 无需再死记硬背官方的筛选条件逻辑。

(三) 对“偏误版本”查看功能的混淆

还有部分使用者对检索结果呈现的内容提出了强烈的困惑。例如有发帖询问: “发现这个语料库为什么看不到写的错误版本啊, 怎么看这个语料库错字的错误版本啊”。

这种求助意味着, 很多用户其实是想通过语料库看到学习者的原始错误内容, 以便进行对比研究。然而, 传统平台把带有 [Zc](错字)、[Cb](错词)等代码的句子直接平铺在界面上, 并没有提供一个“一键剥离标注代码, 只看错误内容”或者“将正确版本和错误版本左右对照”的便捷功能。传统检索界面缺失的“展示层处理能力”, 导致很多研究生在做毕业论文时, 必须把检索结果一条条复制到 Word 里, 再手动删除代码, 耗费大量机械体力劳动。

(四) 在官方文档之外, 对“专家智囊”的迫切渴求

尽管有官方帮助文档, 但许多使用者依然感觉无所适从。一位用户发出求助帖: “求会用全球汉语中介语语料库的大佬”, 甚至直接感叹“搞不懂那个代码是什么意思”, 另一位用户虽然认为“全球汉语中介语语料库太人性化了吧”, 但依然面临着“CCL 到底怎么用啊”的留言。

这些在社交平台上积累的高频求助帖, 形成了一幅清晰的“用户焦虑图景”: 官方文档是静态、被动的, 而语料库的使用门槛依然很高。在此背景下, 将大语言模型引入成为一个极佳的“降维打击”式解法。未来的语料库完全可以在网站页面中嵌入一个“AI 助手”悬浮窗。当用户遇到不会的代码时, 直接问 AI; 当不知道如何操作时, 直接问 AI。这种将静态使用说明书转化为 AI 动态导师的思路, 是解决目前“入坑难”困境的最高效手段。

5. AI 助手辅助智能检索的实证模拟设计

通过上述基于网络的真实使用者反馈可以确定, 现有检索系统在“理解用户意图”和“降低使用门槛”方面存在巨大短板。我们不再空谈“系统蓝图”, 而是转向“真刀真枪的 Prompt 模拟”。

(一) 实验环境与操作工具

本研究使用的实验工具为目前市面上免费、可检索的主流网页端大语言模型(如 Kimi、DeepSeek 或豆包)。输入数据为笔者从“全球汉语中介语语料库”的真实检索页面中, 经过条件筛选下载并导出的 300 条“把字句”真实语料。我们将这批语料中的关键信息——“国籍”、“HSK 等级”、“原始偏误句及偏误标注代码”——分批复制粘贴进大模型的聊天框中, 作为 AI 助手进行推理和检索的“上下文语料池”。

实验思路: 通过设计一套合理的“系统指令(角色设定 Prompt)”, 将大模型设定为一个懂语言学规则、懂语料库标注的专业“AI 助手”。我们通过实际的文字交互, 检验它能否像电商平台的 AI 助手那样, 根据用户模糊的自然语言意图, 精准筛选出目标例句, 并自动按照重要性进行排序推荐, 以弥补现有网页检索页面的缺陷。语料库建设应注重“多样性和层次性”[13], 本研究所提出的 AI 助手检索机制, 正是试图在“功能层次性”上对现有检索系统进行补充。

(二) “语料库 AI 助手”的核心角色设定 Prompt

为了让大模型听懂我们语言学研究的专业需求, 而不只是当个通用聊天机器人, 我设计了如下的角色设定 Prompt:

“你现在是一个专门为国际中文教育研究者服务的‘汉语中介语语料库 AI 智能助手’。我接下来会给你提供一些带有详细偏误标注的真实‘把字句’学习者语料。你的任务是根据我的提问, 精准地从这些语料中筛选出符合条件的句子。在回答时, 请遵循以下规则:

检索结果必须按‘典型度’(即该类偏误在提供的语料中出现的频次高低)从高到低排列, 将最具教学

代表性的错误例句排在前面。

你的回复必须附带筛选依据，例如指出该例句属于哪类具体偏误，以及该偏误出现的大致频次。

在筛选过程中，请忽略口语表达中的正常停顿、结巴和重复词，重点关注句法结构和偏误标注符。”

6. AI 助手辅助检索的实证测试与分析

接下来，我们将分两个实验场景，把真实的语料数据输入给 AI 助手，并与其进行对话交互，验证其检索与推荐能力。

6.1. 实验场景一：多维度复合意图检索测试(对比传统网页检索)

研究者提问：“帮我从刚才发给你的这批把字句语料中找出，学习者写‘把字句’时，把介词‘把’错用成‘用、给、让’的例子。特别是韩国、越南或印度尼西亚学生的。”

传统网页检索的困境回顾：如果使用真实的“全球汉语中介语语料库”网页页面解决这个问题，我们需要分几步：首先去“字符串一般检索”里输入“把”；然后需要切换到“条件筛选”，去选国籍(韩国、越南、印度尼西亚)；接着因为要找“错用”，还需要切换到“分类标注检索”去定位[Cb]标签。系统无法在一次请求中同时处理“国籍 + 错词标签 + 特定句式”，导致用户要么只能分头搜索，要么返回的结果毫无针对性。

“语料库 AI 助手”的实际返回结果(模拟大模型回复)：

“已为您筛选出典型‘介词混淆’偏误共 11 条。按典型度(出现频次)从高到低排序如下：

典型偏误示例 1：我【把】[Cb 用]这照片学习拍风景，把照片放到网上来。(国籍：马来西亚)。典型度：高。分析：学习者将表示工具的介词‘用’错误地写成了‘把’，这是本组数据中出现频率最高的误用类型。

典型偏误示例 2：这位中国朋友跟我说：‘中国人【一般】[Cb]【把】[Cb 用]‘风’来给男孩取名字’。(国籍：越南)。典型度：高。分析：同样是表工具的‘用’错写为‘把’，例句本身具有极高教学价值。

典型偏误示例 3：我决定要帮助我朋友【把】[Cb 让]他们俩在一起。(国籍：印度尼西亚)。典型度：中。分析：表使令意义的‘让’被错写为‘把’，属于汉语动词范畴混淆，值得注意。”

实验结论：AI 助手完美展现了类似淘宝“AI 助手”的语义理解能力。它不再依赖现有网页分门别类的标签页限制，而是靠理解“错用”、“用”、“给”、“让”背后的深层语义，成功完成了多国籍、多条件的复合跨维度检索，并自动进行了按频次高低排序。这充分证明，AI 助手能够显著降低语料库检索的认知门槛。

6.2. 实验场景二：“找典型”式的个性化推荐测试

研究者提问：“如果我要做一堂关于‘把字句’的教学示范课，我需要选出这批语料里最典型、最常见的偏误类别。你能直接告诉我最常见的偏误类型是什么，并给出 3 条例证吗？”

传统网页检索的困境：面对 300 条原始语料，如果要在现有的系统里找“典型错误”，只能一页页往下翻，自己凭感觉统计。系统没有提供“按偏误出现频次排序”的按钮。

“语料库 AI 助手”的实际返回结果(模拟大模型回复)：

“根据对这批 300 条语料进行的智能统计与语义归纳，最典型的高频偏误类别为‘介词混淆(把与用/让/给混用)’。该类偏误在本批有效偏误样本中占比超过 40%。

为您推荐 3 条核心教学语料如下：

我【把】[Cb 用]这照片学习拍风景。(把‘用’误写为‘把’)

中国人【把】[Cb 用]‘风’来给男孩取名字。(把‘用’误写为‘把’)

我决定要帮助我朋友【把】[Cb 让]他们俩在一起。(把‘让’误写为‘把’)

AI 教学建议：在课堂教学中，教师应重点引导学生区分表‘工具’的‘用’和表‘处置’的‘把’这两类介词的语法意义区别，并配合上述例句进行改错练习。”

实验结论：AI 助手瞬时充当了“智能数据分析师”的角色。它完成了传统网页平台无法提供的统计学分类工作，直接为研究者指明了教学设计的方向。这无疑实现了我们所追求的“个性化推荐”功能。

6.3. 实验中暴露的 AI 具体问题与解决策略

尽管 AI 助手在理解语义和排序方面表现出色，但在本次基于真实口语语料的测试中，我们发现了一个非常致命的“具体问题”。根据课程论文“发现具体问题并予以解决”的要求，我们对这一环节进行了深度剖析。

(一) 发现的具体问题：AI 助手误判“口语杂音”为“句法偏误”

在这 300 条语料中，有相当一部分来自于“看图说话”、“电视节目”等半口语语料。当我们将这些真实的口语化转写语料喂给大模型时，AI 助手出现了严重的“系统失灵”现象。

原始语料(一位美国学习者，口语转写)：“额 - 你把 - 把，额 - 我不知道这个词，不知道这个意思。”

AI 助手初始判断(系统失灵现象)：大模型在初次分析时，极其敏感地捕捉到了句中的“你把 - 把”这种重复结构，并将其判定为了“介词冗余/语序混乱”的典型偏误，并向用户推荐这条语料用于教学。

问题根源分析：大语言模型是基于人类自然语言概率预测的模型，它并没有受过语料库标注规范的专业训练。它并不懂语言学研究的一条铁律——“口语转写中的口吃、结巴、停顿，只要句法结构本身是完整的，就不能算作偏误”。如果这种由于说话紧张导致的口语杂音被当成典型错误推荐给上课的老师，将会误导学生的汉语认知。

(二) 研究并予以解决：引入“原生标注符硬约束”的 Prompt 修正策略

为解决 AI “乱判句法”这个具体问题，我们通过修改“系统指令(Prompt)”，给 AI 助手戴上了“专业眼镜”。

具体的解决方案(加入约束条件的修正 Prompt)：

我们在原有 Prompt 的基础上，追加了以下硬性约束规则：

“重要硬性规则：在进行偏误语料检索和推荐时，只有当句子中明确包含了标准化偏误标注代码(如 [Cb]错词、[Cq]遗漏、[Cd]误加、[Zc]错字等)时，才可将该句判定为‘偏误句’并进行推荐。对于句中并没有明确包含上述偏误标注符的口语重复、停顿、结巴词，一律视为正常交际现象，绝对不予推荐和报道。”

解决后的验证效果：在加入这段硬性规则后，我们再次将带有结巴的“你把 - 把”语料输入给 AI 助手，AI 助手很快给出了非常专业的修正回复：

“经核对，该句没有包含[Cb]、[Cq]、[Cd]等任何偏误标注代码，仅为口语表达时的正常停顿，不属于偏误检索目标，不予推送。”

核心学术价值提炼：这个实证环节完美证明了“人机协同”在语言学检索中的必要性。检索工具应简易使用、兼顾人机配合的重要性[3]，我们无需去改写语料库的后台，只需通过设计具有专业约束力的 Prompt (提示词)，就能把泛泛而谈的通用大模型，驯化为真正懂语言学规矩、能精准执行检索任务的“专业 AI 助手”。这正是我们在缺乏系统开发权限的客观条件下，推动语料库智能化检索的最优解[14][15]。

7. 结论与反思

(一) 研究结论

本研究立足于“不改造语料库后端系统，仅利用现有网页端大模型工具”的现实条件，依托真实的

“全球汉语中介语语料库”检索界面, 基于 300 条真实的“把字句”语料进行了一场详实的 Prompt 实证模拟[16]。本研究得出以下结论:

AI 助手具备极强的语义“意图理解”优势。它突破了现有平台“字符串检索”与“分类标签检索”割裂的局限, 通过理解用户的日常口语需求, 实现了复杂条件的自动拆解与高召回率的语义检索, 大幅提升了检索效率[9][10]。

大模型天然的“统计分析与排序”能力实现了检索结果的个性化。它能迅速识别高频偏误并将其置顶, 初步达成了面向教学和科研的“个性化推荐”功能, 直接减少了研究者人工筛选的时间成本, 弥补了系统缺乏“按典型度排序”的缺陷。

“AI 助手”存在天然的“偏误幻觉”, 必须辅以专业约束。本研究提出的“强制匹配原生标注符”的 Prompt 约束策略, 成功修复了 AI 对口语化语料的误判, 证实了语言学专业规则在 AI 应用中的不可替代性[17][18]。同时, 该策略证明了 AI 助手能够极大地弥补静态官方使用说明在实际操作中“响应滞后”的短板[3]。

(二) 研究局限与展望

语料规模的局限: 本研究仅基于 300 条语料进行实证, 样本量偏小。未来可将更大规模的数据通过分段分批输入的方式喂给大模型, 验证其在百万字级语料库中的处理能力。

缺乏实时联网接口: 目前我们只能将语料手动复制粘贴给 AI 助手。展望未来, 随着各大语料库开放 API 接口, 我们可以在现有的网页检索界面中添加一个“AI 智能问问”的侧边栏悬浮窗, 让研究者直接在语料库网页上享受“一句话问语料”的极致体验, 彻底告别翻阅数十页官方 PDF 使用说明的繁琐操作。语料库建设的现状需要在技术策略上不断推陈出新, 才能跟上时代需求[19]。

总而言之, 作为国际中文教育领域的研究生, 我们完全不需要因为“无法修改语料库系统代码”而束手束脚。本次实证充分表明, 在现有大模型工具的辅助下, 凭借我们扎实的语言学功底和巧妙的 Prompt 设计, 完全可以让“AI 助手”成为我们高效发掘语料库价值的核心助推器, 成为一本 24 小时在线的、智能的《动态使用说明》。这不仅是技术与教育的融合, 更是未来国际中文教育科研范式的一次有益探索[20]-[22]。

参考文献

- [1] 张宝林. 从 1.0 到 2.0——汉语中介语语料库的建设与发展[J]. 国际汉语教学研究, 2019(4): 84-95.
- [2] 张宝林. 汉语中介语语料库检索系统透视[J]. 天津师范大学学报(社会科学版), 2021(6): 29-37.
- [3] 张瑞朋. 留学生汉语中介语语料库建设若干问题探讨——以中山大学汉字偏误中介语语料库为例[J]. 语言文字应用, 2012(2): 131-136.
- [4] 张宝林. 汉语中介语语料库建设的反思与前瞻[J]. 国际中文教育(中英文), 2022, 7(2): 3-4.
- [5] 张宝林. 关于汉语中介语语料库的资源共享问题[J]. 汉语教学学刊, 2022(1): 9-15+148-149.
- [6] 张宝林. 关于汉语中介语语料库软件系统的思考[J]. 天津师范大学学报(社会科学版), 2022(4): 35-40.
- [7] 张宝林, 崔希亮. “全球汉语中介语语料库”的特点与功能[J]. 世界汉语教学, 2022, 36(1): 90-100.
- [8] 谭正娇, 王文文, 余晓铃. 汉语中介语语料库研究热点分析——兼谈对国际中文专业的影响[J]. 现代交际, 2021(16): 75-77.
- [9] 郑通涛, 曾小燕. 大数据时代的汉语中介语语料库建设[J]. 厦门大学学报(哲学社会科学版), 2016(2): 53-63.
- [10] 林君峰. 云计算技术在中介语口语语料库建设中的应用[J]. 对外汉语教学与研究, 2015(0): 33-39.
- [11] 肖奚强, 周文华. 汉语中介语语料库标注的全面性及类别问题[J]. 世界汉语教学, 2014, 28(3): 368-377.
- [12] 张瑞朋. 论汉语中介语语料库中汉字偏误标注的全面性和科学性[J]. 语言与科学, 2025(1): 292-306.
- [13] 周文华. 汉语中介语语料库建设的多样性和层次性[J]. 汉语学习, 2015(6): 97-105.

-
- [14] 张瑞朋. 三个汉语中介语语料库若干问题的比较研究[J]. 语言文字应用, 2013(3): 133-140.
- [15] 曹贤文. 留学生汉语中介语纵向语料库建设的若干问题[J]. 语言文字应用, 2013(2): 127-134.
- [16] 李桂梅. “全球汉语中介语语料库”的平衡性考虑[J]. 华文教学与研究, 2017(2): 46-51.
- [17] 蔡武, 郑通涛. 我国汉语中介语语料库研究现状与热点透视——基于 CiteSpace 的可视化分析[J]. 华文教学与研究, 2017(3): 79-87.
- [18] 颜明, 肖奚强. 论汉语中介语语料库建设的基本问题[J]. 语言文字应用, 2017(1): 136-144.
- [19] 张宝林. 汉语中介语语料库建设的现状与对策[J]. 语言文字应用, 2010(3): 129-138.
- [20] 张宝林. “HSK 动态作文语料库 2.0 版”的设计理念与功能[J]. 语料库语言学, 2021, 8(1): 81-96.
- [21] 曹贤文. 二语习得研究“需求侧”视角下的汉语学习者语料库建设[J]. 华文教学与研究, 2020(1): 38-46.
- [22] 张宝林. 关于汉语中介语语料库的应用问题[J]. 语言教学与研究, 2019(2): 18-27.