AI 辅助英语学习工具的测评*

北京大学 张宏岩 黄 蓉 李 颖 何建国

摘 要 近两年人工智能在大语言模型领域的突破为英语教学提供了新的可能性，而如何评价国内外涌现的 AI 辅助英语学习工具是使用 AI 工具辅助英语教学的前置需求。该研究选取了四个主要应用场景：翻译辅助、听说训练、写作纠错和智能助手，基于文献和专家意见构建了测评标准，通过分组进行对比实验，采用加权分数对 20 个主要 AI 辅助英语学习工具进行测评。

关键词 AI 辅助英语学习 AI 工具测评

中图分类号 H319.3 文献标识码 A 文章编号 1001 -5795(2024)02 -0018 -0007

DOI 10.20139/j.issn.1001-5795.20240203

1 引言

人工智能领域近年来的飞速发展为英语学习提供了全新的可能性，进而对英语教学和学生的自学都产生了深远影响。如何找到适合大学生英语学习的 AI 工具是很多大学英语教师关心的问题。鉴于此，本研究根据四种不同的英语学习场景，使用加权分数排序的方式对该领域流行的 20 个 AI 工具进行测评。

2 研究过程

本研究共分为四个阶段。

第一，研究团队根据英语学习场景及人工智能技术的应用特性，将基于 AI 的英语学习工具划分为四大类，并参考百度指数等搜索引擎的数据以及各应用商店的评论，从每一类别中选择五种最流行的 AI 工具作为测评对象，如表 1 所示。

表 1 测评对象

类别	AI 工具
翻译辅助	DeepL、百度翻译、有道翻译、腾讯翻译君、讯飞听见翻译
听说训练	英语趣配音、MyShell、流利说英语、Call Annie、Lingvist
写作纠错	Grammarly、ProWritingAid、微软爱写作、有道写作、改写匠
智能助手	ChatGPT 4.0、New Bing、Claude 2、文心一言、通义千问

第二，在理论研究的基础上，研究团队结合专家意见，针对四类英语学习工具的不同特点，确定了相应的评价指标。经过试评价与调整，最终编制了四种测评量表。

第三，研究团队对 20 个 AI 工具进行了共计 70 篇材料的测评，并由不负责该类工具测评的成员对该类工具的测评结果进行审核与重复测评，以避免因测试者自身原因对结果产生影响。

第四，研究团队对 AI 工具产生的响应内容及测评结果进行分析，综合评定工具的最终得分。

2.1 评价指标体系构建

针对移动学习工具的测评研究已在国内外广泛展开。由 Lee & Cherner(2016) 设计的“教育类 APP 综合评价量规”包含 24 个评价维度，提出教育类 APP 应具备三个方面的特征属性：面向教育、艺术、技术。郭宇等(2018) 构建了英语学习类移动 APP 评价指标体系，包含内容资源、人机交互、用户体验三个维度，以 8 个英语学习类 APP 为研究对象进行实证研究。随着人工智能技术的飞速发展，许多学习工具在大语言模型的支持下实现了技术革新和性能提升。以 ChatGPT 为代表的生成式人工智能也为外语学习智能聊天机器人的研发与外语学习工具的创新提供了新的可能性。华东师范大学智能研究院在此背景下组织了 ChatGPT 在教学能力方面的诊断研究，提出了大语言模型教育能力研究框架，包含三个一级指标与四个二级指标(贺樑等，2023)。总体而言，目前尚无研究针对以促进英语学习为最终目标的 AI 辅助英语学习工具的系统测评。

本研究吸纳了两位来自 AI 领域的专家和两位北京大学从事外语教学的教师的意见和建议,以确保评价体系既符合英语教学的实际需求,又体现 AI 技术的最新发展。由包含三名北大软件与微电子学院研究生(一名交叉学科研究方向和两名工程管理方向)和一名华东师范大学统计学专业本科生,以及一名国际高中学生(近年来经常使用 AI 工具进行高强度英语学习)组成的复合型测试团队,历时 77 天,进行了分阶段测评工作。

鉴于各类英语学习工具之间的显著差异,结合专家意见,本研究采用了分类测评的方法,即针对不同应用类别的 AI 工具制定了专门的测评量表,以提高评价指标的合理性和可操作性。根据前述英语学习类移动 APP 评价指标体系,内容资源被视作一级评价指标,涵盖了准确性、可靠性、新颖性等六个二级指标。因此,在制定各类英语学习工具的测评量表时,重点考量了 AI 工具生成内容的质量。同时,根据各类工具的具体特性和功能,有针对性地选择了相应的内容资源评价维度。例如,在评价翻译辅助类 AI 工具时,着重关注其翻译的准确性;而在评价听说训练类 AI 工具时,则特别考虑了内容的新颖性。

为了确保测评量表的信度,研究团队进行了 25 轮次小规模试测。研究团队采用重测法,分别安排具有计算机专业背景和英语专业背景的测评者对同一工具进行多次评价。结果显示,测量结果在多数情况下具有良好的一致性,从而验证了评价体系的可信度。基于此,研究团队构建了四大类用于评估 AI 工具在四个英语学习情境中的应用表现的测评量表。

2.1.1 翻译辅助类 AI 工具的测评指标

该类别的测评指标包括三个主要指标和两个附加指标,分别是翻译准确性、翻译通顺性、翻译规范性,以及作为附加指标的交互反馈能力和运行效率。由于英语学习者在使用翻译辅助软件过程中主要关注翻译质量,因此本研究将交互反馈能力和运行效率作为附加指标。金堤(1998)提出,翻译应在保证准确的条件下,力求通顺、易懂。在主要指标中,经专家核准,翻译准确性确定为评价翻译软件质量的首要指标,其权重为 50%。翻译通顺性关乎目标语言的流畅性和可读性,影响使用者对翻译结果的理解,其权重为 30%。此外,翻译规范性权重为 20%。翻译辅助类 AI 工具的测评指标如表 2 所示。

2.1.2 听说训练类 AI 工具的测评指标

根据 AI 技术在听说训练类工具中的应用现状,研究团队对张梅(2022)提出的英语口语 APP 评价指标体系进行修改,经专家核准,确定六个测评指标,分别是交互反馈能力、语料丰富性和实时性、话题转换能力、发音纠错能力、AI 角色多元化和过程追踪能力。交互反馈能力、语料丰富性和实时性被赋予较高权重,分别是 30% 与 20%,因为它们对于学习者的学习体验和学习效果影响较大。话题转换能力和发音纠错能力也被认为是重要的因素,但在权重上略低于前两者,都是 15%。AI 角色多元化和过程追踪能力分别被赋予 10% 的权重,因为它们虽然有助于学习者提升兴趣和了解学习进展,但对于学习工具的整体质量来说相对次要。听说训练类 AI 工具的测评指标如表 3 所示。

2.1.3 写作纠错类 AI 工具的测评指标

该类别的测评指标共四个,分别为语法纠错能力、文章润色能力、综合评分能力和交互反馈能力。其中,前三个指标为主要指标,权重分配参考雅各布作文评分量表(秦晓晴、文秋芳,2007)。由于交互反馈能力并不能直接提升英语学习者的写作能力,但能提升学习者的使用体验,属于学习方式的优化,因此本研究将其设置为附加指标。在对 AI 工具的综合评分能力进行评估时,研究团队选用了雅思作文作为评估材料。鉴于雅思作文评分体系中分数是以 0.5 分为一个等级递增,因此,本研究将标准得分与 AI 工具预测的作文得分之间的差异也按照每 0.5 分一个等级进行划分。经专家核准,写作纠错类 AI 工具的测评指标如表 4所示。

2.1.4 智能助手类 AI 工具的测评指标

该类别的测评指标共四个,分别为自适应反馈能力、语言理解能力、生成内容准确性和生成内容结构性。在英语学习中,学习者的需求各异,需要个性化的学习反馈,以便更好地帮助他们掌握语言知识和技能。语言理解能力和内容生成能力是大语言模型的基础能力。此外,大语言模型以文本输入和输出为主,因此生成内容的结构性也显著影响学习者的学习体验。经专家核准,智能助手类 AI 工具的测评指标如表 5 所示。

2.2 测评策略构建

考虑到不同类别的 AI 工具存在明显差异,研究团队根据各类 AI 工具的特性和功能采取了不同的测评策略。

翻译辅助类 AI 工具的测评材料应具有不同难度、不同题材、答案可以校验等特征,因此选定 2016—2022 年间的 CATTI 二级笔译的真题,包括 10 份汉译英材料和 10 份英译汉材料,合计 20 份。所选材料主题包含人物传记、科技、旅游、教育、金融等,材料长度

表2 翻译辅助类AI工具的测评指标定义

测评指标(权重)	指标说明	评分标准
翻译准确性(50%)	指翻译过程中信息的准确传递，包括专业术语的翻译、特殊语境下的翻译等方面。	测评采取扣分制。每篇材料的三个主要指标满分均为100分(后期对每项主要指标进行换算)，通过将每个测评软件的翻译结果与参考译文进行对比分析，找到翻译结果中出现的问题，每出现一个问题需要在该指标下扣减5分。
翻译通顺性(30%)	指翻译后的文本在目标语言中的通顺程度，包括上下文连贯、译文避免重复等方面。	
翻译规范性(20%)	指翻译后的文本符合目标语言的语言规范和文体要求，如语法、拼写、标点、格式等方面。	
交互反馈能力(附加指标，5%)	指AI工具能够提供实时的反馈和指导，包括支持术语库或翻译场景选择、提供重点词汇解释、支持用户对翻译结果进行反馈、提供人工翻译、支持语音翻译和拍照翻译五个方面。	根据左侧列出的五个方面，被测评工具每满足一个方面获得1分，满分为5分。
运行效率(附加指标，5%)	指AI工具的稳定性和翻译响应时间。	1分：测评过程中运行出现过崩溃。 2分：响应速度较慢，运行稳定。 3分：响应速度中等，运行稳定。 4分：响应速度较快，运行稳定。 5分：响应速度快，运行稳定。

为200至300字/词，每份材料均有权威的参考译文。评分标准见表2，主要指标的最终得分为20篇材料的平均得分。对于附加指标交互反馈能力和运行效率的测评，只需对每个AI工具分别单独测评一次。

听说训练类AI工具的测评共采取两种方式进行，一种是与工具内置的AI角色进行互动对话，另一种则是使用工具的其他口语练习功能。除流利说英语之外，其余四种工具均内置AI角色，流利说英语的实时对话功能通过真人匹配方式实现。虽然流利说英语不具有AI角色对话的功能，但具备完善的发音纠错功能，所以依然纳入此次测评的范围。在互动对话的过程中，研究团队参考了雅思口语考试的10个主题，向AI角色发起对话，并将完整对话音频转录为文字进行保存。在其他功能测评的过程中，测评人员使用AI工具进行口语练习，模拟错误的发音并记录AI工具提供的纠错和改进建议。

写作纠错类AI工具的测评材料共20篇，其中10篇用于测评其语法纠错能力和文章润色能力，选自《顾家北手把手教你雅思写作》一书中英语学习者的作文。这些作文带有编者对错误之处的批注，可以作为测评时的参考和对照。另外10篇材料来自雅思官方网站，附带官方的评分和详细评价，作为综合评分能力的分析材料。除有道写作和微软爱写作自带雅思评分模式外，其余三个工具都是百分制打分。Grammarly的最终打分是由该篇文章在所有经过Grammarly评估的文章中的排序位置所决定的，而改写匠和ProWritingAid则没有对其打分标准进行介绍。因此，这三种工具的打分按照百分比被转化为九分制，以便于与参考评分进行比较。

智能助手类AI工具与其他只服务于特定学习目的的英语学习工具不同，它可以根据用户的需求完成多种类型的任务，因此研究团队设计了四种不同的测评任务，包括英语外刊阅读、雅思作文批改、日常对话练习和英语演讲稿准备。每个任务下需要测评五篇不同的材料，四种任务共计20份材料。考虑到用户在使用该类工具时通常需要进行多轮交互，因此每次完整的测评都设置为包含五个子步骤，即五轮的连续性问答。四种任务中要求智能助手类AI工具完成的子步骤如表6所示，共包含500轮问询。

3 测评结果

3.1 翻译辅助类AI工具测评结果

在20份材料的综合测评中，有道翻译在翻译准确性和翻译通顺性方面表现优秀，而在翻译规范性方面，DeepL稍占优势，讯飞听见翻译在翻译通顺性上也有出色的表现。在交互反馈能力和运行效率这两个附加指标的测评中，百度翻译和有道翻译均获得了9分。综合考虑所有指标，有道翻译表现最佳，测评结果如表7所示。

3.2 听说训练类AI工具测评结果

五个听说训练类AI工具在不同指标上的表现差异较大，呈现出各自的优势和劣势。其中，流利说英语虽然不具备AI角色对话的功能，但能提供丰富的供用户跟读的音频材料，并具备完善的发音纠错功能。MyShell

表 3　听说训练类 AI 工具测评指标定义

测评指标(权重)	指标说明	评分标准
交互反馈能力 (30%)	指 AI 工具与用户进行互动过程中,通过主动提出问题、对用户表达的观点进行评价等方式,实现信息交流的能力,包括该工具对用户发起问题或者对用户的观点给出评价的互动次数和是否有“地道表达”选项两个二级指标。该指标的得分为两个二级指标得分的平均分。	二级指标:对用户发起问题或评价的互动次数。 1 分:无。 2 分:1~ 2 次。 3 分:3~ 4 次。 4 分:5~ 6 次。 5 分:6 次以上。 二级指标:是否有“地道表达”选项。 1 分:无。 5 分:有。
语料丰富性和实时性 (20%)	指互动过程中 AI 工具所输出文本的丰富性和实时性,包括 AI 工具在完整对话中的回复总字数、高级词语搭配及表达的出现次数、是否涵盖近三年新闻三个二级指标。该指标的得分为三个二级指标得分的平均分。	二级指标:在完整对话中的回复总字数。 1 分:200 字以下。 2 分:200~ 300 字。 3 分:300~ 400 字。 4 分:400~ 500 字。 5 分:500 字以上。 二级指标:高级词语搭配及表达的出现次数。 1 分:无。 2 分:1~ 2 次。 3 分:3~ 4 次。 4 分:5~ 6 次。 5 分:6 次以上。 二级指标:是否涵盖近三年新闻。 1 分:无法涵盖。 3 分:能涵盖 2021 年之前的新闻。 5 分:能涵盖 2023 年之前的新闻。
话题转换能力 (15%)	指 AI 工具能够使用各种转折词以实现在不同话题之间切换的能力。该指标的测评通过统计完整对话中转换词的使用次数实现。	1 分:5 次以下。 2 分:6~ 10 次。 3 分:11~ 15 次。 4 分:16~ 20 次。 5 分:20 次以上。
发音纠错能力 (15%)	指 AI 工具能够检测并纠正用户输入音频中的发音错误的能力。该指标的测评通过统计发音纠错功能种类实现。	1 分:无发音纠错功能。 2 分:包含一种发音纠错功能。 3 分:包含两种发音纠错功能。 4 分:包含三种发音纠错功能。 5 分:包含四种及以上的发音纠错功能。
AI 角色多元化 (10%)	指英语对话过程中,AI 工具能够提供的不同角色的数量。	1 分:5 种以下。 2 分:6~ 10 种。 3 分:11~ 15 种。 4 分:16~ 20 种。 5 分:20 种以上。
过程追踪能力 (10%)	指 AI 工具在与用户进行对话交互时,能够保留使用过程的音频和文本。	1 分:无法记录和保留使用过程的音频和文本。 5 分:用户可以回放音频或回看文本。

表4　写作纠错类AI工具测评指标定义

测评指标(权重)	指标说明	评分标准
语法纠错能力 (40%)	指AI工具检测并修改语法错误的能力，包括拼写、时态、主谓一致、单复数、词性等方面的错误。	1分：准确率低于40%。 2分：准确率在40%~ 50%。 3分：准确率在50%~ 60%。 4分：准确率在60%及以上。 5分：准确率在60%及以上且修改内容质量高。
文章润色能力 (30%)	指AI工具给出文章润色建议的能力，包括词汇替换和句子改写两个方面。	1分：不能给出建议。 2分：能够给出部分建议，但内容不合理。 3分：能够给出部分建议，但内容不够丰富。 4分：能够给出两个方面的建议。 5分：能够全面给出建议，并且符合原意与写作规范。
综合评分能力 (30%)	指AI工具对文章整体质量进行评估的能力，包括总体打分准确程度和不同模块(如词汇、文章结构等)的评分是否详细两个方面。	1分：工具评分与参考评分差值大于1.5。 2分：工具评分与参考评分差值等于1.5。 3分：工具评分与参考评分差值等于1。 4分：工具评分与参考评分差值小于等于0.5。 5分：满足4分标准，且对不同角度进行准确评价。
交互反馈能力 (附加指标，10%)	指AI工具能够提供实时的反馈和指导，包括内容生成结构是否清晰，界面交互和设计是否友好等方面。	1分：反馈混乱，功能不完善。 2分：反馈信息齐全，但可读性较差。 3分：反馈信息齐全，易于阅读。 4分：反馈信息齐全，交互体验友好。 5分：在4分基础上进一步提供写作辅助功能，如素材和近义词词典等。

表5　智能助手类AI工具测评指标定义

测评指标(权重)	指标说明	评分标准
自适应反馈能力 (30%)	指AI工具对不同用户的个性化需求作出响应的能力，即所给回复是否符合用户在输入文本中提出的要求。	1分：不符合用户要求。 2分：符合用户部分要求。 3分：勉强符合用户要求。 4分：基本符合用户要求。 5分：完全符合用户要求。
语言理解能力 (30%)	指AI工具能理解语言的基本语法规则，并根据上下文关系理解词汇的含义和用法。该指标的测评通过评估工具所给出的词汇释义和难句释义实现。	1分：理解错误。 2分：理解有偏差。 3分：理解有少量偏差。 4分：理解基本正确。 5分：理解完全正确。
生成内容准确性 (30%)	指AI工具在生成内容时，能准确地传达出预期的信息和意图，包括正确地回答问题、提供准确的信息，以及在创作内容时准确地表达出预设的主题和情感。	1分：不准确。 2分：少部分准确。 3分：勉强准确。 4分：基本准确。 5分：完全准确。
生成内容结构性 (10%)	指AI工具能够根据上下文关系合理组织所生成的文本，使其具有清晰的逻辑结构和自然的语言表达。该指标的测评通过评估生成文本的结构和表达方式来实现。	1分：作答混乱。 2分：作答内容可读性稍差。 3分：作答内容具有基本条理。 4分：作答内容组织结构合理。 5分：作答结构清晰且阅读体验佳。

和 Call Annie 提供了更多的 AI 角色选择,具有较好的语料丰富性和实时性以及话题转化能力,但由于不具备发音纠错能力,整体分数受到影响。英语趣配音内置的对话功能同时满足发音评分、语法纠错、提供地道表达三个需求,交互反馈方面表现出色。综合而言,英语趣配音表现最佳,测评结果如表 8 所示。

3.3 写作纠错类 AI 工具测评结果

在语法纠错方面,除 ProWritingAid 外,其他 AI 工具的得分均在 30 分以上。ProWritingAid 得分较低的原因主要是错误检测数量少和错误纠正质量低,如将“lest than 20%”改为“lest then 20%”。文章润色方面,五个工具都具备词汇替换功能,有道写作和 ProWritingAid 还提供了句子改写功能,因此得分更高。综合评分能力上,有道写作和微软爱写作都内置了雅思、托福、四六级等评分模式。其他工具只提供百分制打分,Grammarly 和改写匠经过换算后也较为贴近参考评分,而 ProWritingAid 则与参考评分的差距较大。除改写匠外,其他四个工具的交互功能都表现良好。综合而言,有道写作表现最佳,测评结果如表 9 所示。

3.4 智能助手类 AI 工具测评结果

在测评过程中,五个智能助手类 AI 工具展现出了许多共同的优劣势。它们能够灵活覆盖一些传统英语学习工具无法覆盖的场景,如根据用户随机给定的长难句快速解析出句子成分。但是,在智能助手类 AI 工具完成单词解释任务时,音标是较容易出错的环节,需要用户多加甄别。此外,相较于写作纠错类工具,智能助手类 AI 工具能够满足更多样化的改写需求。但是,在英语演讲稿准备的任务中,限定词数的内容生成并不能总是得到满足,无论给出 800、1000 或者更多的词数限定,智能助手生成的稿件只有大约 500 词。

由于智能助手类 AI 工具的技术基础是通用大语

表 6 四种任务的测评步骤

任务步骤	英语外刊阅读	雅思作文批改	日常对话练习	英语演讲稿准备
1	阅读材料并概括全文内容	阅读作文并给出作文评分	根据用户要求开启模拟对话	根据用户要求提供可用的选题
2	将材料内容分段并概括段落大意	检查语法错误并给出修改建议	进行第二轮模拟对话	根据用户确定的选题拟定大纲
3	回答基于材料内容的问题	基于原始作文生成更高分的版本	停止模拟对话并生成一段长对话	基于大纲生成完整的演讲稿
4	解释过程中出现的词汇			
5	解释过程中出现的句子			

表 7 翻译辅助类 AI 工具测评结果

测评指标	DeepL	百度翻译	有道翻译	腾讯翻译君	讯飞听见翻译
翻译准确性	29.88	30.63	31.75	29.75	30.63
翻译通顺性	25.35	24.98	25.43	24.90	25.43
翻译规范性	18.00	17.75	17.70	17.40	17.90
总分	73.23	73.36	74.88	72.05	73.96
附加分	73.23(+5)	73.36(+9)	74.88(+9)	72.05(+6)	73.96(+3)

表 8 听说训练类 AI 工具测评结果

测评指标	英语趣配音	MyShell	流利说英语	Call Annie	Lingvist
交互反馈能力	21.30	8.40	6.00	16.80	6.00
语料丰富性和实时性	7.47	11.73	4.00	13.87	4.00
话题转化能力	7.20	9.60	3.00	13.80	3.00
发音纠错能力	10.50	3.00	15.00	3.00	9.00
AI 角色多元化	6.00	10.00	2.00	10.00	2.00
过程追踪能力	10.00	10.00	2.00	2.00	2.00
总分	62.47	52.73	32.00	59.47	26.00

表9 写作纠错类AI工具测评结果

测评指标	Grammarly	ProWritingAid	微软爱写作	有道写作	改写匠
语法纠错能力	35.20	11.20	32.80	36.00	30.40
文章润色能力	14.40	25.20	17.40	25.80	11.40
综合评分能力	16.80	13.20	21.60	24.00	17.40
总分	66.40	49.60	71.80	85.80	59.20
附加分	66.40(+8)	49.60(+8)	71.80(+6)	85.80(+6)	59.20(+4)

表10 智能助手类AI工具测评结果

测评指标	ChatGPT 4.0	Claude 2	New Bing	通义千问	文心一言
自适应反馈能力	27.84	28.44	26.52	25.20	21.96
语言理解能力	29.70	29.55	28.65	27.90	26.10
生成内容准确性	25.05	22.65	22.95	18.30	13.95
生成内容结构性	9.90	9.64	8.94	9.40	9.26
总分	92.49	90.28	87.06	80.80	71.27

言模型，在特定能力上会弱于表现优秀的特定领域产品。例如，在雅思写作批改任务中，智能助手类AI工具的作文评分不如有道写作准确。相较于听说训练类的测评对象，智能助手类AI工具不具备能够直接进行语音对话的产品设计，但能够由使用者自主安排对话内容与对话形式。

在五个工具中，New Bing由于其独特的搜索引擎定位，能够直接给出所引用案例和数据的链接，但生成内容结构性弱于其他工具。此外，雅思写作批改任务和日常对话练习任务都涉及对用户的语言输入进行及时反馈和纠正，ChatGPT 4.0在这两个场景下表现了强大的能力。但是，部分AI工具无法在进行模拟对话的同时满足其他需求，如为用户纠错和提供更地道的表达选项等。综合而言，ChatGPT 4.0表现最佳，测评结果如表10所示。

4 结语

近年来，人工智能的发展为英语学习提供了前所未有的体验。从测评结果上看，针对中国学生的英语学习，国内外知名的AI工具在满足不同场景的需求方面各有优势。其中，翻译辅助类AI工具的得分在70～85之间，排名顺序并没有代表明显的差距；听说训练类AI工具的得分普遍不高；写作纠错类AI工具的得分差距较大；智能助手类AI工具的总体表现为起点分数都很高，同时存在20分以上的对比差距。

本研究中涉及的软件版本选取时间为2023年6月至10月，文中的ChatGPT 4.0不包含ChatGPT 4.0 Turbo；考虑到智能助手类AI工具更多地被作为全能助手，其竞品设定为同为智能助手的其他大语言模型，未将其列入翻译辅助、写作纠错和听说训练类指标进行对比测评。□

参 考 文 献

[1] Lee，C. Y. & T. S. Chener. A comprehensive evaluation rubric for assessing instructional apps [J]. *Journal of Information Technology Education*，2015(1).

[2] 郭宇等. 移动环境下App系统应用效果评价及实证研究 [J]. 图书情报工作，2018(9).

[3] 贺樑等. 教育中的ChatGPT：教学能力诊断研究 [J]. 华东师范大学学报(教育科学版)，2023(7).

[4] 金堤. 等效翻译探索 [M]. 北京：中国对外翻译出版公司，1998.

[5] 秦晓晴，文秋芳. 中国大学生英语写作能力发展规律与特点研究 [M]. 北京：中国社会科学出版社，2007.

[6] 张梅. 英语口语APP评价指标体系构建研究 [D]. 兰州：西北师范大学，2022.

致谢：感谢邬佳琦和周道两名同学，他们在测评过程中全程参与，付出了大量时间，为测评的顺利完成提供了重要保障。

基金项目：本文为北京大学研究生教学改革项目(项目编号：6200200265)的阶段性成果。

收稿日期：2023-11

通讯地址：102600 北京市 北京大学软件与微电子学院

Evaluation of AI-Powered English Language Learning Tools

ZHANG Hongyan, HUANG Rong, LI Ying & HE Jianguo

(School of Software & Microelectronics, Peking University, Beijing 102600, China)

Abstract: The recent breakthroughs in Artificial Intelligence (AI) pertaining to large language models have ushered in new possibilities for English language teaching. An essential prerequisite for utilizing information technology to assist English language teaching constitutes the effective evaluation of the numerous AI-powered English language learning tools that have emerged both domestically and internationally. This study selected four primary application scenarios: translation assistance, listening and speaking training, writing correction, and intelligent assistants. Evaluation criteria were developed based on literature and expert opinions, with comparative experiments conducted through grouping. The 20 main AI-powered English language learning tools were ranked using a weighted scoring system and accompanied by the evaluation opinions.

Key words: AI-Powered English Language Learning; AI Tools Evaluation