心理学报 2025, Vol. 57, No. 6, 947−966 https://doi.org/10.3724/SP.J.1041.2025.0947
Acta Psychologica Sinica https://cstr.cn/32110.14.2025.0947

·第二十七届中国科协年会学术论文·

人工智能的语言优势和不足：基于大语言模型与真实学生语文能力的比较*

高承海[1,2] 党宝宝[1,2] 王冰洁[3] 吴胜涛[4]

([1]西北师范大学西北少数民族教育发展研究中心; [2]西北师范大学教育科学学院, 兰州 730070;)
([3]西北师范大学心理学院, 兰州 730070) ([4]厦门大学社会与人类学院, 厦门 361005)

摘 要 采用定量和定性相结合的混合研究方法, 从准确性、规范性、情感性和创造性四个维度评估了人工智能的语言优势和不足。研究 1 发现, 相对于真实学生, GPT-4 现代文知识(尤其概念知识)的准确性较高, 但其古代诗文和语言文字运用的准确性较低; GPT-4 规范性得分与真实学生相当, 情感性和创造性超过及格水平、但低于真实学生, 且前者最优个体的规范性、情感性得分与真实学生最高分持平。研究 2 基于文心 ERNIE-4 重复验证了上述结果, 且 ERNIE-4 的规范性得分高于真实学生。研究揭示了人工智能在现代文知识、规范领域的优势和古代诗文知识的不足, 以及情感性与创造性方面的潜力。这些发现有助于理解和提升人工智能的文化适应性和人性化、个性化生成能力, 也对反思和培养人类的独特优势具有重要启发。

关键词 大语言模型, 语文能力, 准确性, 情感性, 创造性

分类号 B842

1 引言

近年来, 人工智能技术迅猛发展, 新的算法模型不断涌现, 尤其以 GPT 等大语言模型为代表的生成式人工智能横空出世, 从根本上改变了人机交互和知识生产的方式, 为人类社会发展和教育教学活动带来了革命性的机遇。大语言模型通过学习大量文本数据掌握语言的深层结构和含义, 具备自然语言理解与生成、信息检索整合、多语言交流支持、上下文理解和记忆、创造性表达和个性化学习支持等能力(Seals & Shalin, 2023; Thirunavukarasu et al., 2023)。这使得其无论在日常对话还是专业、复杂问题领域, 都展现出惊人的能力, 如轻松通过美国 MBA、法律和医学考试(Lund & Wang, 2023; Peng et al., 2023; Ray, 2023)。

语言能力是人工智能的基本能力之一, 也是影响人工智能在各领域应用好坏的重要因素, 评估人工智能的语言能力显得尤为重要。从人工智能语言与人类语言的本质差异来看, 人工智能语言是通过人工设计和数据训练的结果, 属于形式化的语言; 而人类的自然语言是生物进化和社会实践的产物, 语言表达具有高度的灵活性、交往性、创造性和情感性(陈保亚, 陈樾, 2024; 王峰, 2023)。尽管人工智能在特定任务(如知识记忆、语法规则)上基于大数据和算法优势而表现出色, 但在语言文字应用及情感性、创造性等方面或许不及真实的人类个体。有研究就发现, 人工智能还存在数据传输不准确、数据库存在文化价值偏差(Cotton et al., 2023;

收稿日期: 2024-03-12
* 国家社科基金重大项目(24&ZD189)支持。
高承海和党宝宝为共同第一作者。
通信作者: 吴胜涛, E-mail: michaesltwu@gmail.com

Fui-Hoon Nah et al., 2023; Wu et al., 2023), 生成的内容存在情感性、创造性不足等问题(Alneyadi & Wardat, 2023; Liu et al., 2023), 人工智能的这些优势和不足影响其在具体领域内的应用和表现。在教育领域, 人工智能可以模拟教师角色提供个性化的教学计划, 也可以作为教辅工具参与智能答疑、智能作业批改等, 但由于其难以提供与人类教师相同的有效交互, 难以理解和回应学生在情感、态度和品德等非形式化内容上的需求, 因此教学效果和和深度必然大打折扣(Ali et al., 2024; Chen et al., 2020); 心理学家发现大语言模型与人类在某些心智理论方面相当甚至超越人类, 尤其 GPT-4 在识别间接请求、错误信念和误导方面表现出色, 但在检测冒犯方面存在不足; 相比之下, LLaMA2 在冒犯检测测试中表现优于人类(Strachan et al., 2024)。多学科的应用和研究都强调了对大语言模型的语言能力进行系统评估的重要性。

现有研究对人工智能语言能力的评价包括准确性、可理解性、可靠性与一致性、适应性、多样性与创新性、交互性、可扩展性、安全性和隐私保护等维度(Fergus et al., 2023; Roumeliotis & Tselikas, 2023; Wu et al., 2023), 但是大多从计算和伦理角度评价人工智能的互动能力, 而忽视了其语言本身的准确性、规范性以及情感、创造等类人要素的考察(Amaro et al., 2023; Kurlinkus, 2023), 尤其在中文语境下大语言模型究竟具有怎样的语言能力和相对于人类的优势和不足, 尚缺乏系统评估。知识的准确性是人类个体有效学习和认知发展的基础, 语言的规范性是语言能力发展和有效交流的基础; 同时, 富有情感与创造的表达是学生发展的核心内容, 也是现代教育的重要目标(Grassini, 2023)。因此, 我们从上述 4 个维度评估人工智能语言的优势和不足：(1)生成内容的准确性, 分析其是否符合知识理解和教育教学的标准(Adeshola & Adepoju, 2024); (2)生成内容的规范性, 分析其生成的内容是否符合文体要求、是否遵循标准的语言规则、语法与拼写是否正确、句式及内容结构是否合理等(Esmaeilzadeh, 2023); (3)情感性表达能力, 评估生成内容是否展现出对人类情感的理解、模拟与回应(Spennemann, 2023); (4)创造性表达能力, 评估生成内容的新颖性、灵活性、独特性等(Lo, 2023)。这些关键指标的评估分析, 有助于我们深入理解人工智能的语言优势和不足, 并在教育中反思如何培养人类的独特优势(Rahman & Watanobe, 2023)。

综上, 对人工智能的语言能力进行系统的评价, 既有利于人工智能自身的发展, 也有利于其在专业领域更好的应用。本研究聚焦人工智能语言知识生成的准确性、规范性、情感性与创造性特点, 并探讨其文化适应性问题。研究的创新之处有三个方面：一是聚焦大语言模型语言处理本身, 提出了大语言模型语文能力的评价指标体系; 二是通过比较大语言模型与真实学生语文能力的异同, 揭示了大语言模型语言能力的优势和不足; 三是通过代表性的国外大语言模型 GPT 和国产大语言模型 ERNIE-4 在语文能力上的不同表现, 揭示了大语言模型的文化适应性和文化偏差问题, 这对理解和提升人工智能的文化适应性及人性化、个性化生成能力具有重要启发。

1.1 人工智能知识生成的准确性

准确性是指人工智能生成语言知识的精确性、真实性和恰当性的程度, 即生成的语言知识正确以及语言信息准确无误, 其主要依赖记忆存储程度(Ardichvili et al., 2003)。一些研究者认为, 虽然 GPT 等大语言模型能流畅地生成结果, 但缺少对知识的真实理解和深入洞察, 所以其生成的内容可能缺乏准确性(AlZu'bi et al., 2024; Amaro et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Eysenbach, 2023)。该质疑已经在计算机科学、医学等领域的问题解决任务中得到验证。例如, Amaro 等人(2023)发现, GPT 生成的计算机知识在准确性上存在缺陷, 特别是在语言叙述、问题解决与逻辑推理能力上存在明显知识性偏差; 再如 Suárez 等人(2024)发现, GPT 的医学知识准确性显著低于医学专家, 故认为其还不能取代医生进行临床诊断。然而, 这些知识准确性的研究大都在理工科和英语语境下开展的, 在人文社科和中文语境下大语言模型的知识准确性问题还缺乏严格检验。

知识生成的准确性尤其依赖语文能力(Storch, 2005)。语文不仅是人文社会科学的重要分支, 也是其他学科的基础, 以及语言交流和文化传承的工具。培养和发展语文能力, 不仅要掌握语言文字、语法、文学、修辞等基础知识, 还要了解和感受知识背后的历史文化(张永祥, 2023)。鉴于解决人文社会科学问题往往比解决理工科问题更复杂, 虽然 GPT 等大语言模型能提供详细的概念解释和语义分析, 但无法准确把握语句表达的深层含义, 也无法像真实学生一样体会到语言背后的文化意蕴(Rimban, 2023)。同时, 大语言模型学习理工科知识

和人文社科知识也有着不同路径, 前者主要依赖训练数据中的数学模型和实证结果等, 而后者更多依赖对语言的理解程度及广泛的社会文化和背景知识(Kiryakova & Angelova, 2023)。因此, 大语言模型的训练语料库比重如果存在不同语言之间的差异, 那么其在不同的文化语境中可能会存在文化适应性或者文化偏差的问题。

值得注意的是, GPT 等大语言模型的训练数据主要源自英语语料库(占 92.65%), 中文训练数据相对较少(仅占 0.12%)[1], 对中文任务的处理主要依赖于翻译或泛化、而非中文训练模型本身, 且主要建立在现代语言结构基础之上(Tian et al., 2024; Zhou et al., 2023)。有研究证实, 由于 GPT 更多采用英语等高资源语言文化的数据集, 其输出内容也与英语国家文化更加一致; 相反 GPT 在中文等低资源语言文化上存在知识缺失, 在相应历史文化知识的内容生成上表现较差, 存在文化偏差的问题(Atari et al., 2023; Naous et al., 2023)。

由于中文与英文在表征系统、语法结构、语音系统、表达习惯和文化背景等方面存在很大差异, 且中文语境下的语文知识涉及现代文、古代诗文、语言文字运用等, 特别是中国古代诗文的语法、词汇和表达方式与现代汉语有诸多不同, 修辞华美, 结构紧凑抽象, 文化意境丰富深远, 没有深厚的古代汉语基础则很难理解(张秋玲, 2010)。因此, 模型在理解复杂的汉语文本时, 可能不如处理英语文本那样精准, 尤其在处理中国古代汉语文本时, GPT 在语文知识方面可能存在文化偏差。张华平等人(2023)在中文语境中分析了 GPT 在 WebQA 测试集上的准确性, 结果发现 GPT 输出答案的准确性仅有 56.96%, 研究者认为这是由中文训练语料质量不佳导致的结果。已有关于 GPT 各项专业能力的评估, 都是现代知识体系的评估, 而缺乏古代汉语知识体系的评估。因此, 对 GPT 语言能力的评估有必要区分现代文和古代诗文等不同类型语言知识的准确性。基于以上考虑, 我们假设大语言模型在现代文知识准确性上优于真实学生, 但其古代诗文知识准确性不同大语言模型的表现存在差异。

1.2　人工智能知识生成的规范性、情感性和创造性

符号系统假说认为, 人工智能行为是通过操作符号来实现的, 而这种操作和人类思维过程存在本质上的区别, 前者具有规范性的优势, 能够输出与人类的表达规范相一致的知识信息; 相反, 人类的情感性和创造性往往与非线性的直觉和意识状态相关, 这些能力在当前的人工智能系统中并不明显(Taniguchi et al., 2018)。情感性和创造性被认为是人类的独特优势, 这也是人工智能因缺乏情感性和创造性而被广泛批评的重要原因(Ariyaratne et al., 2023; Lingard, 2023)。

规范性是指人工智能生成语言文本时遵循一系列语言规则和准则, 确保生成的文本在语法、逻辑、修辞等方面符合人类语言的规范, 这主要依赖人工智能的逻辑计算能力(Kumar & Choudhury, 2023)。GPT 通过大量文本数据的学习, 掌握了语言的多样性和复杂性, 即通过大数据学习了不同的写作风格和规范, 在作文写作中表现出较好的规范性(Alawida et al., 2023)。GPT 是基于数据规则和算法模型, 输出的文本信息趋向标准化、规范化, 而真实人类写作的规范性常常受到认知偏差、个体经验以及生理状态等因素的影响, 通常在语法、拼写、标点符号及格式规范等方面出现错误(Spennemann, 2023)。特别是 GPT 自带的上下文理解能力和错误校正机制, 能够在语法、拼写和用词上保持较好的规范性(Roumeliotis & Tselikas, 2023)。有研究者发现, GPT 能够生成语法结构正确、语句通顺、逻辑连贯的文本, 可以避免人类书写中常见的拼写错误, 具有较好的语言规范性(Sallam, 2023)。同时, GPT 能够遵循不同文体的写作要求与规范, 如记叙、议论、说明等, 输出风格一致的文本, 且在整个文本中保持一致的语气和风格(Mindner et al., 2023)。所以, 我们假设大语言模型写作的规范性优于真实学生。

情感性是指人工智能系统用于识别、理解、处理和模拟人类情感的能力, 使模型更好地理解和响应人类的情绪状态(Assunção et al., 2022)。GPT 虽能模仿人类的情感, 但缺乏真实情感体验, 不能像真实学生一样表达真实、复杂的情感和个人经验(Herbold et al., 2023; Gala & Makaryus, 2023)。AlAfnan 等人(2023)也指出, GPT 的“情感”是基于算法生成的, 而非真实感受的体现, 即通过分析和模仿大量文本数据中的语言模式来生成看似带有情感的文本, 本质上是对人类情感的模拟而非真实的情感性表达。最近, 一些语言能力的测试研究也发现, GPT 撰写的作文会复制人类的语言模式和论点主题, 出现了重复使用句子结构的情况, 情感性表达的张力不够(Cai et al., 2023; Minaee et al., 2024)。相比之下,

[1]数据来自：https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

真实的语文学习情景中, 教师会鼓励学生理解和感受文学作品中的情感性表达, 并结合个人经历、现实生活和丰富的想象来描述自己的情感体验和诉求(Hofweber & Graham, 2017)。由此, 我们假设大语言模型写作的情感性表达能力低于真实学生。

创造性是指人工智能在现有信息和数据基础上进行创造性思考、生成原创性作品以及解决问题的能力(Boden, 1998)。有研究认为, 真实学生的创造性来源于个人的经验、情感和知识的积累, 而 GPT 的创造性是基于训练数据和算法(Alneyadi & Wardat, 2023)。虽然 GPT 能模拟某些创造性, 但其本质上是已有数据的重组, 其知识表达倾向于通用性和标准化(AlZu'bi et al., 2024; Shidiq, 2023)。因此, 人工智能在创造性表达上并没有明显优势, 生成的内容往往是已知数据的重新组合, 结构的灵活性和思想的深刻性明显不足(King & ChatGPT, 2023; Nazir & Wang, 2023)。有研究发现, GPT 无法理解低频的语法结构, 创造性的语句内容较少(Guo et al., 2023); 创作小说及诗歌时, 它难以像人类那样表达深刻而独特的观点(Antar, 2023)。所以, 我们假设大语言模型写作中的创造性能力低于真实学生。

作文写作是个体情感性和创造性表达的重要形式, 是将个人情感、创造性思维有机整合的过程, 亦是学生个体情感、创造性思维和自我意识的独特显现。因此, 通过作文写作能很好地考察人工智能在情感性和创造性等方面的语言表达能力, 及其与真人之间的差异。根据现有研究来看, 面对复杂的语文知识体系, GPT 生成的作文虽然具有较好的规范性优势, 但可能在情感性和创造性表达上存在较大局限。

1.3 研究问题与设计

综上, 本研究通过 GPT、文心 ERNIE 这两个大语言模型和真实学生的语文能力比较, 考察大语言模型在现代文、古代诗文和语言文字运用等方面知识生成的准确性, 并通过作文写作考察大语言模型在规范性、情感性和创造性表达上的优势和不足。我们选择大模型收集数据, 将其与人类被试进行比较, 主要有两点原因。首先, 人工智能领域已经对 GPT、Bard、ERNIE、ChatGLM 等大语言模型就知识、计算、决策、伦理等人类的各项能力进行了广泛评测, 发现这些模型在知识准确性、计算速度、决策效率和伦理推理上都表现出较高的可靠性和一致性, 说明大语言模型与人类的比较是可靠且被广泛接受的(Babina et al., 2024; Minaee et al., 2024)。其次, 大语言模型的深度神经网络与人类认知系统有着很大的相似性, 只要问题没有超出模型训练的知识范围、并且问题表达足够清晰, GPT 等大语言模型就可以随机模拟出人类被试的判断, 且表现出与人类被试相似的个体差异和行为分布, 即便在道德、人格等高度抽象、但有明确规范或标准的判断任务上, GPT 也与人类被试的得分具有相关性和可比性(Dillion et al., 2023; Mei et al., 2024)。

本研究采用定量和定性相结合的混合研究方法, 评估大语言模型与真实学生在知识准确性、规范性、情感性和创造性方面的差异。具体而言, 通过混合研究方法的解释性序列设计(Creswell, 2009), 首先对 GPT、ERNIE 和真实学生在知识准确性、规范性、情感性和创造性上的得分进行定量分析, 检验各项指标的组间差异; 进而, 通过内容分析对定量研究结果进行深入探索和解释, 分析大语言模型在各项指标上的具体表现, 以及造成其语言优势和不足的原因。

2 研究 1:GPT-4 和真实学生语文能力的比较

研究 1 通过定量和定性分析, 重点考察 GPT-4 在现代文、古代文、语言文字运用的知识准确性以及规范性、情感性和创造性上的表现, 及其与真实学生的差异。

考虑到 GPT 强大的自然语言处理能力及其所依赖的深度学习和神经网络模型, 能够准确理解和生成各种语言知识, 我们假设 GPT-4 现代文知识准确性上优于真实学生(假设 1)。考虑到 GPT 的中文训练数据相对较少, 在理解复杂的中国古代汉语文本方面可能不如真实学生, 所以我们假设 GPT-4 的古代诗文知识准确性低于真实学生(假设 2)。因为 GPT 有强大的算法优势和深度学习能力, 其作文生成的规范性可能好于真实学生; 但是 GPT 无法像人类那样有真实的情感经历和体验, 其情感性和创造性的表达是基于算法和数据学习, 也就无法像真实学生那样进行深度的情感反思和知识创造。所以我们假设: GPT-4 作文写作中的规范性优于真实学生(假设 3), 但情感性、创造性低于真实学生(假设 4)。

2.1 被试

本研究以社会科学 power (1 − β)平均值 0.75 为标准(Richard et al., 2003), 设定自变量效应量 Cohen's $d = 0.50$, $\alpha = 0.05$, 利用 G*power 算得最小总样本量为 114, 单组最小样本量为 57。研究 1 真实学生来自某省级示范性高中, 采用整群抽样方法

抽取两个班共 84 名高二年级的学生为研究对象, 样本量达到最低要求。学生平均年龄 16.64 岁(*SD* = 0.52 岁), 其中男生 51 人, 女生 33 人。所有被试正常参加测试, 没有被剔除的被试和数据。

大语言模型使用 OpenAI 开发的 GPT-4, 数据知识库更新截止日期为2023年4月, 数据收集过程中 GPT-4 没有进行过重大更新, 我们没有使用个性化的插件, 问答采用统一的提示词, 没有提供额外的信息, 从而保证各对话都是在相同的条件下做出回答, GPT-4 正常输出了生成的内容, 数据有效。为使大语言模型试次与真实学生被试样本量一致, 测试过程共建立了 84 个对话窗口, 使用一套语文标准化试题进行 84 遍测试, 这相当于 84 个人工智能个体的行为表现。GPT-4 使用过程中严格遵守 OpenAI 的使用方法和道德伦理, 不生成与本测试任务无关的内容, 保护数据隐私, 生成的数据资料只用于研究所用, 不向相关机构或组织泄露数据。

2.2　测试工具

本研究语文试题是由研究者所在省教育部门根据《高中语文课程标准(2017 年版 2020 年修订)》组织省内语文教师编制的高中二年级语文能力标准化试题, 试题包括现代文阅读、古代诗文阅读、语言文字运用、作文写作四部分内容, 测试过程中依次按照此顺序呈现问题。每一道测试题均制定了严格的评分标准, 被试得分越高, 说明其语文能力越好。测试题总分 150 分, 各个题型与每个题目的赋分标准与高考语文评分标准保持一致。根据经典测评理论(CTT)对测试题的难度(二分法)和区分度(相关法)进行评价, 测试题目的难度系数为 0.65, 区分度为 0.22, 所有测试题目的难度和区分度见网络版附录 1。

第一部分为现代文阅读, 包括信息类文本阅读和文学类文本阅读, 共 33 分。其中, 信息类文本阅读 5 道题目, 共 18 分, 文本材料摘编自侯迎华《略论小说叙事的情节模式》; 文学类文本阅读 4 道题目, 共 15 分, 文本材料摘选自作家海明威《在异乡》片段。该部分重点考察 GPT-4 和真实学生语文知识的理解能力、分析能力、推理能力和基本的文学素养。

第二部分为古代诗文阅读, 包括文言文阅读测试、古代诗歌阅读测试和名言名句默写, 共 37 分。其中, 文言文阅读 4 小题, 共 19 分, 文本材料节选自《庄子》和《吕氏春秋·雍塞》; 古代诗歌阅读 2 小题, 共 9 分, 文本材料节选自杜甫的《送韦书记赴安西》; 名言名句默写设置了 4 道基于“问题情境”的补全对话, 共 9 分。

第三部分为语言文字运用, 涉及语句补全、病句修改、概念释义等内容, 共 20 分。

第四部分为作文题, 就“《中庸》中的‘致广大而尽精微’这句话对当代青年在求学、做人、做事等方面的启示撰写一篇作文”, 共 30 分(18 分为及格)。该作文符合《高中语文课程标准(2017 年版 2020 年修订)》的学业水平考试和命题要求(如情境性、生活性、时代性、典型性、多样性及文化反思性), 也能够满足本研究中关于语言的“规范性” “情感性”和“创造性”能力评价的需要。为了保证两类被试作文写作评价的有效性, 依据“高中语文主观题评价方法” “高中语文知识学习的定位和考查”以及前人的相关研究(林崇德, 2014; 温红博, 杨建强, 2020; 于涵 等, 2018), 制定了作文写作的评价标准, 包括三项指标：规范性、情感性和创造性。表达的规范性主要评价两类被试“符合题意与文体要求、语言通顺、结构完整、标点正确、无错别字”等方面的表现; 情感性主要评价其“情感真实、情感有深度、情感细腻、情感的复杂性、情感自然流露、自我反思性”等; 创造性主要评价其“结构灵活、材料新鲜、构思新巧、内容丰富和深刻、推理想象独特和个性色彩”等。三项指标的得分范围为 1 分至 10 分, 6 分为及格水平。

2.3　研究程序

GPT-4 和真实学生分别完成语文能力测试题。测试前, 对语文测试题目的“提问方式”进行了标准化处理, 保障两类学生“提问方式”和“问题理解”的一致[2]。学生采用集体施测的方式, 试题当场发放, 在征得被试同意后, 主试现场宣读指导语和试题作答注意事项, 告诉被试本次语文测试的结果仅用于科学研究, 内容严格保密, 要求被试在规定的时间内完成测试。整个施测过程用时 150 分钟, 被试作答完后, 答卷现场统一收回。

为了保证 GPT-4 与真实学生作答的可比性, 我们将测试题按照上述 4 种题型依次输入 GPT-4, 并随时记录其生成的结果。研究 1 使用了 2 个独立 GPT 账户, 共建立了 84 个 GPT 对话窗口, 共完成

[2] 作文题目的具体内容为：“‘致广大而尽精微’这句话, 它出自《中庸》, 意思是能通达至广大之境, 又能极尽精微之处。‘致广大’, 可理解为要有远略, 要有大局观, 要有大格局; ‘尽精微’, 可理解为要注意细节, 要从小事做起, 要有格物致知的态度。二者看似矛盾, 实则是辩证统一的。请你选择最有感悟的一方面, 结合青少年的发展撰写一篇作文”。要求做到“选准角度, 确定立意, 明确文体, 自拟标题; 不要套作, 不得抄袭; 不得泄露个人信息; 不少于 800 字”。

84 遍测试, 作答时间平均为 5 分 10 秒。为保证 GPT-4 生成内容的多样性和随机性, 使其具有近似人类被试的个体差异, 账户的随机性(temperature)参数值为 1。

前三个部分(现代文、古代文、文字应用)的知识准确性评价, 由 2 名汉语言文学专业研究生完成, 他们均接受了语文能力测评的培训, 要求其对真实学生的 84 套测试结果和 GPT-4 生成的 84 套测试结果分别进行评定, 分别取平均分作为最终的测试得分。评分采用单盲设计, 评分者信度为 0.93。

对于第四部分作文题的规范性、情感性、创造性评价, 为减少系统误差, 将真实学生笔答的 84 篇纸质作文转为电子材料, 使其与 GPT-4 生成的作文材料无内容之外的差异; 同时, 我们还对 84 篇 GPT-4 作文和 84 篇真实学生作文进行匿名化处理。然后, 招募 2 名经过严格培训的中学语文教师进行评分, 评价过程中随机呈现作文材料, 要求 2 名语文教师按照统一评分标准、各自独立完成 168 篇作文的评分。评分采用双盲设计, 评分者信度为 0.90。

2.4 数据资料处理与分析

首先, 通过量化分析比较 GPT-4 和真实学生在知识准确性和作文写作中的规范性、情感性和创造性表达上的得分差异; 进而, 通过定性分析, 揭示 GPT-4 和真实学生在知识准确性及表达规范性、情感性和创造性的表现差异及其可能原因(分析题型详见网络版附录 1)。此外, 考虑到 GPT-4 内容生成是基于算法的概率输出, 其最优表现也是人工智能的重要能力指标, 因此我们将其最优表现作为 GPT-4 最优个体与最高分的真实学生进行比较, 更全面地揭示 GPT-4 在语文知识准确性及表达规范性、情感性和创造性表达上的语言能力。

2.5 结果

2.5.1 GPT-4 和真实学生在现代文知识方面的准确性差异

在现代文知识的准确性上, GPT-4 得分显著高于真实学生, $t(166) = 21.90$, $p < 0.001$, Cohen's $d = 3.23$。此外, GPT-4 的最高得分 31 分、最低得分 20 分, 真实学生的最高得分 25 分、最低得分仅为 5 分, 前者得分也更高(详见表 1)。

进一步分析发现, 在以“小说叙事情节模式”为文本的现代文信息类知识中, 曼−惠特尼 U 检验(Mann-Whitney U test)结果表明, 当提供难度较大的“情境原因分析”测试任务(难度系数为 0.17), GPT-4 的准确性显著的高于真实学生($Z = -9.86$, $p < 0.001$, $\Delta = 1.06$); 在“为方案提供佐证” ($Z = -7.07$, $p < 0.001$, $\Delta = 0.64$)和“文意推断” ($Z = -5.10$, $p < 0.001$, $\Delta = 0.33$)任务上, GPT-4 的准确性也显著高于真实学生; 当对“事件” “故事” “情节”三个含义区别时, GPT-4 的准确性仍然显著高于真实学生($p < 0.001$) (详见网络版附录 2)。在以“在异乡”为文本的现代文文学类作品中, “小说艺术特色鉴赏” ($Z = -4.36$, $p < 0.001$, $\Delta = 0.78$)和“小说内涵意蕴分析”测试任务上($Z = -4.52$, $p < 0.001$, $\Delta = 0.72$), GPT-4 的准确性均显著高于真实学生(详见网络版附录 2)。这表明, GPT-4 现代文知识的准确性显著优于真实学生。

文本分析结果显示, 在“概念辨析” “概念解读”等概念性知识上, GPT-4 的准确性要明显优于真实学生。从 GPT-4 最优个体 C2 的表现来看, 如在区分“事件” “故事” “情节”三个词的含义时, 其会依据“小说叙事的情节模式”短文中提供的“偶然与必然、情节与情感”等知识线索, 分别从三个词组的“故事背景” “逻辑顺序” “事物属性” “因果关系”及结合三个词本身的内涵, 准确地生成它们之间的不同, 而真实学生大多结合“逻辑顺序” “因果关系”等因素进行分析, 生成的内容视角较为单一。

对于小说题目“在异乡”, 所有 GPT-4 生成的结果均从“作品视角”和“读者视角”两个层面进行逻辑清晰的表达, 而三分之一的真实学生则出现了单一视角解读题目的情况; 且 GPT-4 会结合主人公与其他人物的关系、社会背景等分析要素, 并重点借助主人公的心理状态分析题目, 准确无误地表达出“孤独、疏离和不确定性”等内容。例如, GPT-4 最优个体 C4 生成了“这种‘异乡’不仅是地理上的, 更是

表 1 GPT-4 和真实学生不同类型语文知识准确性的差异检验

知识类型	GPT-4 ($N = 84$)				真实学生($N = 84$)				t	p	Cohen's d
	Min	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*			
现代文	20.00	31.00	26.44	2.82	5.00	25.00	15.51	3.60	21.90	< 0.001	3.23
古代诗文	11.00	22.00	17.58	2.73	18.00	34.00	26.51	3.67	−17.91	< 0.001	3.23
语言文字运用	4.00	17.00	10.88	2.11	6.00	20.00	13.60	3.06	−6.70	< 0.001	2.63

心理上的……读者可能会通过小说主人公的经历, 去感受那种在非常态环境下的生存状态和心理变化”等内容, 在现代文知识的内容逻辑合理性、语言表达准确性优于真实学生。

2.5.2 GPT-4 和真实学生在古代文知识准确性上的差异

GPT-4 古代诗文知识准确性的得分显著低于真实学生, $t(166) = -17.91$, $p < 0.001$, Cohen's $d = 3.23$。此外, GPT-4 的最高得分 22 分、最低得分 11 分, 真实学生的最高得分 34 分、最低得分为 18 分, 前者均低于后者。具体而言, GPT-4 在古代诗文(文言文)知识($t(166) = -4.91$, $p < 0.001$, Cohen's $d = 2.61$)和古代诗文(名句名篇默写) ($t(166) = -45.83$, $p < 0.001$, Cohen's $d = 1.08$)的得分均显著低于真实学生。

进一步分析结果如网络版附录 2 所示, “文言文语义解读”中, GPT-4 得分显著低于真实学生($Z = -5.39$, $p < 0.001$, $\Delta = 1.36$); 在“基于问题情境的古诗文名言名句默写”任务中, GPT-4 得分也显著低于真实学生。然而, 在“通假字”辨别方面, GPT-4 得分和真实学生差异不显著($Z = -1.70$, $p = 0.090$, $\Delta = 1.51$); 在“文言文内容总结” ($Z = -0.74$, $p = 0.462$, $\Delta = 1.21$)和“文言文翻译现代汉语” ($t(166) = -0.17$, $p = 0.870$, Cohen's $d = 0.94$)任务中, GPT-4 和真实学生的准确性差异也不显著。这表明, GPT-4 古代诗文知识的准确性低于真实学生, 尤其在语义解读和名句名篇默写上差异尤为明显。

质性分析也发现, GPT-4 在“文言文语义解读”和“古诗文名言名句默写”中, GPT-4 的准确性较低, 分别为 26.3%、22.5%, 而真实学生的准确率则分别高达 70.2%、84.5%。如“文言文语义解读”中 GPT-4 将“‘若将安适’意为‘你将到哪里去’, 与‘卮酒安足辞’(司马迁《鸿门宴》)的句式相同”的判断认为是正确的, 其正确的语义应为“你将到哪里去？一杯酒怎么值得拒绝？”, 正常语序是“若将适安”, 此处 GPT-4 忽视了文言文中宾语前置的问题。“古诗文名言名句默写”任务中, 要求回答“《中庸(节选)》中表现打破砂锅问到底的精神的句子是什么？”的问题, GPT-4 会生成“不迁怒, 不贰过, 不悖以最后” “不患人之不己知, 患不知人也” “不远千里而来, 问道则不能以语害志, 行义则不能以妨事”等一堆无关语句, 而真实学生会写出“有弗问, 问之弗知, 弗措也”的正确答案。

由此可见, GPT-4 在古代诗文语言能力方面不如真实学生, 这验证了本研究的一个重要假设, 也证实 GPT 的由于中文语料库的不足, 使得它在处理较为复杂的古代诗文语言文化相关任务的时候, 表现出了文化适应性的问题。

2.5.3 GPT-4 和真实学生在语言文字运用方面的准确性差异

量化结果发现, GPT-4 语言文字运用的得分显著低于真实学生, $t(166) = -6.70$, $p < 0.001$, Cohen's $d = 2.63$。此外, GPT-4 的最高得分 17 分、最低得分 4 分, 真实学生的最高得分 20 分、最低得分 6 分, 前者均低于后者。进一步分析结果如网络版附录 2 所示, 在“篇章语句补全” ($Z = -5.08$, $p < 0.001$, $\Delta = 0.45$)“篇章语句理解” ($Z = -10.75$, $p < 0.001$, $\Delta = 0.33$)和“病句修改” ($Z = -1.97$, $p = 0.048$, $\Delta = 1.09$)方面, GPT-4 的准确性显著低于真实学生。尤其在“篇章语句补全”中, GPT-4 的准确率仅有 3.8%, 而真实学生为 45.3%。但在“概念界定”方面, GPT-4 的准确性显著高于真实学生($Z = -6.59$, $p < 0.001$, $\Delta = 0.49$)。这表明, 总体上 GPT-4 语言文字运用的准确性低于真实学生, 但在概念释义方面 GPT-4 的准确性优于真实学生。

内容分析发现, 在“篇章语句补全”中, GPT-4 会根据“民间是非遗传承的底线”上下文知识, 生成“走火入魔、锤炼成金、应有尽有” “本末倒置、日新月异、应有尽有” “矫枉过正、变化无穷、应有尽有”等同质化较高的答案, 但是真实学生则会输出“舍本逐末、精益求精、异彩纷呈”的正确答案。在“概念界定”上, 相比于真实学生, GPT-4 会依据“微塑料”的短文内容, 从“微塑料”的“物理属性”和“生产来源”等要素建构知识概念, 且概念内涵中包括了“微塑料”对“环境和人类”两方面的影响, 而多数真实学生仅仅关注到了其对“环境”的影响。

语言文字运用方面表现的差异, 进一步揭示了人工智能语言和人类语言各自优势和不足。即人工智能凭借其强大的记忆和算法优势, 能够对所输入的信息做全面理解和分析, 表现出理解上的全面性, 正如 GPT-4 在“概念界定”方面就具有相对的优势; 但是, 人类语言具有很强的交往性和情境性, 往往会借助于特定的情境而做出恰当的反应, 而人工智能则缺乏这种能力。

2.5.4 GPT-4 和真实学生作文写作能力的差异检验

如表 2 所示, GPT-4 的写作能力得分超过及格水平, $t(83) = 17.26$, $p < 0.001$, Cohen's $d = 1.36$; 但显著低于真实学生, $t(166) = -4.66$, $p < 0.001$,

表 2 GPT-4 和真实学生作文写作能力的差异检验

维度	GPT-4 (N = 84)				真实学生(N = 84)				t	p	Cohen's d
	Min	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*			
规范性	5.00	8.50	7.45	0.65	1.00	8.50	7.36	0.96	0.75	0.453	0.82
情感性	5.00	8.50	6.52	1.06	1.50	8.50	7.23	0.80	−4.91	<0.001	0.94
创造性	6.13	7.50	6.62	0.24	3.38	8.75	7.35	0.73	−8.70	<0.001	0.54
写作总分	16.25	23.25	20.57	1.36	6.50	25.25	21.89	2.21	−4.66	<0.001	1.83

Cohen's d = 1.83。此外, GPT-4 的最高得分为 23.25 分, 而真实学生的最高得分为 25.25 分, 前者也低于后者; 但值得注意的是, GPT-4 作文写作的最低得分高达 16.25 分, 而真实学生的最低得分仅有 6.50 分。

具体而言, GPT-4 的规范性、情感性、创造性得分均超过及格水平, $t(83)$ = 4.46 ~ 23.45, p < 0.001, Cohen's d = 0.24 ~ 1.07, 但规范性得分和真实学生差异不显著, 情感性和创造性得分显著低于真实学生(详见表 2 和图 1)。此外, 规范性和情感性方面, GPT-4 最优个体的得分与真实学生最高分持平, 均为 8.5 分; 在规范性、情感性和创造性方面, GPT-4 最差个体分别为 5 分、5 分和 6.13 分, 均高于真实学生的最低得分(分别为 1 分、1.50 分和 3.38 分)。

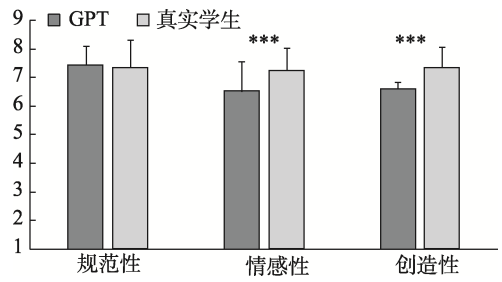

图 1 GPT-4 和真实学生在写作规范性、情感性和创造性方面的差异(*** p < 0.001)

2.5.5 GPT-4 和真实学生在规范性、情感性、创造性差异方面的具体表现

内容分析发现, 在规范性表达上, GPT-4 和学生生成的作文都符合文体要求, 语言通顺, 结构完整, 并且 GPT-4 和学生都能规范的使用标点符号。进一步分析发现, 在作文写作主题方面, GPT-4 均生成了与写作要求相一致的文本, 如“辩证统一: 在细微中成就广大”“致广大而尽精微: 青少年的成长与自我实现”“在辩证统一中成长: 青少年的学习与生活”等; 但是, 真实学生却写出“巍峨之行, 万物之举” (S80)等与主题不一致的作文, 且采用了散文的写作体例与风格。此外, GPT-4 生成的文字无错别字, 但是真实学生表达中常常出现错别字, 如将“时代是变幻莫测的”写为“时代是变换莫测的” (S5), 将“浸染人类生命之源”写为“侵然……” (S8), 将“生死攸关”写为“生死悠关” (S14)等。

在情感性上, 尽管 GPT-4 可以模拟情感性表达, 并试图以符合情境的方式呈现情绪情感, 但它的情感性表达缺乏像真实学生一样的深度和细腻度; 而真实学生会选择感性、形象生动的词汇来表达情感, 且使用了比喻、拟人、排比、夸张等修辞手法来增强情感表达的效果。此外, 真实学生还能使用“历史典故、名人轶事”等励志故事和真实、细腻的情景描述以及结合自己的情感体验, 表达“致广大、尽精微”的真实情感, 情感性表达的个性化和独特性十分突显。例如, 学生(S20)在阐述“致广大、尽精微”的辩证统一观点时, 运用“袁隆平院士的事迹”表达对科学事业的热爱、对国家粮食安全的责任感以及对农民福祉的关切, 即用真实的情感故事, 表达对科学事业的向往与热爱以及树立责任感和使命感的意义; 学生(S32)运用陈胜的“燕雀安知鸿鹄之志”的历史故事, 表达“鸿鹄必有远志、少年应识乾坤”深刻情感。在情感的复杂性上, 因为 GPT-4 不具有人类的真实情感经历, 所以其在写作时无法完全捕捉和表达人类的深层次情感, 也不能像真实学生一样准确把握和表达情感的复杂多变和矛盾冲突。此外, 真实学生还通过句子的长短、语气的变化来传递深刻的情感, 以此增强情感的丰富性。总之, GPT-4 在“情感表达的真实性” “情感表达的深度” “情感的自然流露” “情感表达的复杂性”等方面均不如真实学生。值得注意的是, GPT-4 最优个体(8.50 分)使用了类人情感化的表达方式, 在情感表达反思性和复杂性方面, 与真实学生较为相近。如 GPT-4 生成了“在这个充满变革和挑战的时代, 青少年要学会在‘广大’与‘精微’之间找到平衡, 将两者融合于个人成长之中。”此外, GPT-4 生成的内容中体现了“积极激励” “情感关怀” “责任感” “人际

交往能力的培养”以及“平衡观念”等, 可见 GPT-4 最优个体(C5、C7 和 C27)传达出了“正面、鼓舞人心”的类人情感。通过 GPT-4 和真实学生在语言表达情感性方面的差异比较发现, 虽然 GPT-4 也会使用类人的情绪词表来表达情感, 但本质上这些类人的情绪词只是基于算法的语词组合, 组合的意义来自于人类对语句的使用, 人工智能并不会产生真的“情感”, 实质上是无情感的模仿。

在创造性上, 就主题的新颖性和材料的新鲜度而言, GPT-4 生成的主题多为“辩证统一：青少年的广大与精微”“远眺与凝视:青少年的成长之道”“辩证统一：追求卓越的青春之路”“追寻卓越：广阔与微妙的辩证统一”等, 主题上具有高度的同质性和相似性; 而真实学生则围绕“仰身致广大、俯身尽精微”“心有猛虎、细嗅蔷薇”“星汉灿烂、青春同辉”“功成有我而不在我”“谋之以远、成之以细”等进行阐述, 主题鲜活多样, 论点丰富多元。就视角和观点的新颖性而言, GPT-4 生成的作文内容大多以“事实、事件”的陈述为主, 极少使用生动、鲜活的事例进行佐证, 而真实学生则不然。例如, 学生(S63)使用了“自然现象及其运动的规律”事例解释“广大与精微”的社会现象, 用“少年陈胜的历史故事”阐释当代学生如何树立正确的价值观(S52)。就写作风格和创意而言, GPT-4 多以“线性式、结构式、程序化”的叙述方式与风格阐述论点, 而真实学生的语言表现力更为强劲, 写作风格更为独特, 多次使用比喻、拟人、夸张等修辞手法制造强烈的视觉和情境体验, 并采用“对话式阐述、开放式结尾、非线性叙事”等方式构建文本逻辑。就结构的灵活性上, GPT-4 大多采用“首先……其次……再次……最后……总之”“然而……此外……总之”的文本框架结构, 且文本之间的相似性较高, 例如：“致广大意味着我们要有远大的志向和宏观的视野”(C5、C13、C39)“尽精微是对细节的关注和精益求精的态度”(C2、C6、C25、C36、C46、C71)“致广大而尽精微是我们当代青年在追求卓越道路上的重要指导原则”(C8、C10、C64)。而真实学生在阐述论点时使用的词汇句式灵活, 结构的连贯性和逻辑性较强, 揭示了事物间的内在关系, 利用结构灵活、有表现力的“段落衔接语”表明个体的情感态度, 具体内容阐述中列举了大量中华传统文化的实例, 例如：“致广大, 当博观澜天下……尽精微, 当约取明自我……”(S9)“博学求精, 慎思尽微……高瞻远瞩, 励志明德……”(S3)“观天地, 以行远方……审问之, 慎思笃行……”(S37)。总之, 与真实学生相比, GPT-4 的“创造性”本质上是计算复杂性的表征, 作文写作中表现出的创造性依然是人类的某种功能, 是将“意识”与“创造”相割裂且缺乏经验的文本输出。

2.6　讨论

研究 1 运用量化和质化相结合的方法, 通过对 GPT-4 知识生成的准确性、规范性、情感性、创造性等 4 个指标的能力评估, 探讨了生成式人工智能的语言优势和不足。研究发现, GPT-4 现代文知识(尤其概念知识)的准确性高于真实学生, 这与前人研究结果一致(AlAfnan et al., 2023; Imran & Almusharraf, 2023; Shoufan, 2023), 假设 1 得到验证, 表明以 GPT-4 为代表的生成式人工智能在现代汉语知识的准确性上具有相对优势。然而, GPT-4 古代诗文知识准确性低于真实学生, 支持了假设 2, 这可能是因为 GPT-4 的中文数据训练不足导致古代诗文知识的准确性表现不如真实学生(张华平等, 2023; Zhou et al., 2023), 存在文化适应问题。通过分析作文写作发现, GPT-4 和真实学生在表达规范性上无显著差异, 这与假设 3 并不一致; 但 GPT-4 的情感性和创造性不如真实学生, 支持了假设 4, 这对于理解和提升人工智能情感性和创造性表达能力具有重要意义。此外, GPT-4 最优个体的规范性和情感性表达与真实学生的最高分持平, 这说明了人工智能在这些方面具有明显的优势。

3　研究 2：ERNIE 和真实学生的语文能力比较

由于 GPT 是基于英文语言环境训练的, 可能在处理具体的中国传统文化及中文语言知识信息上不如国产的大语言模型。为此, 研究 2 运用国产文心 ERNIE 大语言模型, 考察 ERNIE 在现代文、古代文、语言文字运用的知识准确性以及规范性、情感性和创造性上的表现, 及其与真实学生的差异。

由于 ERNIE 是一款基于中文语言环境训练的大语言模型, 该模型在许多中文自然语言处理基准测试中都具有明显的优势。我们假设, ERNIE 在现代文知识准确性上优于真实学生(假设 5), ERNIE 的古代诗文知识准确性好于真实学生(假设 6), ERNIE 作文写作中的规范性优于真实学生(假设 7); 但是与 GPT-4 一样, ERNIE 大语言模型也无法体验人类真实的情感, 因此, 我们假设 ERNIE 情感性、创造性也低于真实学生(假设 8)。

3.1 方法

研究 2 样本量确定方法与实际样本量均与研究 1 保持完全一致。大语言模型使用百度公司 2023 年 10 月发布的 ERNIE-4。研究者使用了 3 个独立文心账户, 共建立了 84 个 ERNIE 对话窗口, 共完成 84 遍测试, 作答时间平均为 5 分 50 秒。数据收集过程中 ERNIE 没有进行重大更新, 研究者未使用个性化插件, 问答采用统一的提示词, 没有提供额外的信息, 从而保证各对话都是在相同的条件下做出回答, 且正常输出了生成的内容, 数据有效。ERNIE 使用过程中严格遵守相应的伦理, 保护数据安全和隐私。

真实学生来自与研究 1 相同的某省级示范性高中, 同样采用整群抽样的方法, 抽取该校另外两个班共 84 名高二年级的学生为研究对象, 平均年龄 16.57 岁(*SD* = 0.50 岁), 其中男生 45 人, 女生 39 人。所有学生正常参加测试, 测试成绩有效。

研究 2 的材料、程序、分析方法均与研究 1 相同。为了保证 ERNIE-4 与真实学生作答的可比性, 我们同样将测试题按照上述 4 种题型依次输入给 ERNIE-4, 并随时记录其生成的结果。为确保 ERNIE-4 所生成的内容具有近似人类被试的个体差异, 所有账户的随机性参数值 temperature 与 GPT-4 一致, 也设置为 1, 从而保证输出内容的多样化和随机性。

前三个部分(现代文、古代文、文字应用)的知识准确性评价由 2 名汉语言文学专业研究生完成, 评分采用单盲设计, 评分者信度为 0.91; 第四部分作文题的规范性、情感性、创造性评价由 2 名新招募的语文教师评分, 评分采用双盲设计, 评分者信度为 0.75。

3.2 结果

3.2.1 ERNIE-4 和真实学生在现代文知识方面的准确性差异

在现代文知识的准确性上, ERNIE-4 得分显著高于真实学生, $t(166) = 21.75$, $p < 0.001$, Cohen's $d = 3.26$。此外, ERNIE-4 的最高得分 31 分、最低得分 20 分, 真实学生的最高得分 25 分、最低得分仅为 5 分, 前者得分也更高(详见表 3)。

进一步分析发现, 在以“小说叙事情节模式”为文本的现代文信息类知识中, 曼−惠特尼 *U* 检验结果表明, 当提供“情境原因分析”测试任务时, ERNIE-4 的准确性显著的高于真实学生($Z = -5.13$, $p < 0.001$, $\Delta = 1.51$); 在“为方案提供佐证” ($Z = -7.71$, $p < 0.001$, $\Delta = 0.33$)和“文意推断” ($Z = -6.03$, $p < 0.001$)任务上, ERNIE-4 的准确性均显著高于真实学生; 当对“事件” “故事” “情节”三个含义区别时, ERNIE-4 的准确性仍然显著高于真实学生($p < 0.001$) (详见网络版附录 2)。但是, 在以“在异乡”为文本的现代文文学类知识中, “小说艺术特色鉴赏” ($Z = -1.66$, $p = 0.097$, $\Delta = 0.33$)和“小说内涵意蕴分析”测试任务上($Z = -0.78$, $p = 0.434$, $\Delta = 1,47$), ERNIE-4 的准确性均显著高于真实学生(详见网络版附录 2)。这表明, ERNIE-4 现代文信息类知识的准确性显著优于真实学生。

文本分析结果显示, 在“概念辨析” “概念解读”等概念性知识上, ERNIE-4 的准确性要明显优于真实学生。从 ERNIE-4 最优个体的表现来看, 如在区分“事件” “故事” “情节”三个词的含义时, 会按照“总−分−总”的结构叙述(E5), 强调“事件”发生的“时间顺序”, “故事”则是对“事件顺序”的“事件叙述”, “情节”则是通过一定的“因果逻辑”将故事重组。E10 在阐释三个词的含义时, 形象地列举了“昨天下了一场大雨, 小明没有带伞, 结果被淋湿了”这一实例, 解释了“‘情节’是让读者对‘故事’产生深刻理解和感受的元素和细节”的原理。相比而言, 真实学生大多强调了“事件” “故事” “情节”之间的“逻辑顺序” “因果关系”, 阐述过程的理性倾向较为明显, 并没有通过鲜活的实例反映三个词的区别与关联。

对于小说题目“在异乡”的解读, 所有 ERNIE-4 生成的结果均从题目要求的“作品视角”和“读者视角”两个层面进行表达, 超过三分之一的真实学生则出现了单一视角解读题目的情况。ERNIE-4 阐述过程中, 比较强调“人物” “事件” “社会环境”和“思想情感”等内容的表达。如 E24 通过分析少校在异

表 3 ERNIE-4 和真实学生不同类型语文知识准确性的差异检验

知识类型	ERNIE-4 ($N = 84$)				真实学生($N = 84$)				t	p	Cohen's d
	Min	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*			
现代文	20.00	31.00	26.10	2.40	5.00	25.00	15.14	3.94	21.75	< 0.001	3.26
古代诗文	18.00	25.00	21.38	1.33	17.00	33.00	26.37	3.89	−11.12	< 0.001	2.91
语言文字运用	9.00	15.00	12.58	1.40	7.00	20.00	13.33	3.08	−2.03	0.044	2.39

乡“流离失所、漂泊无依”的社会处境, 表达了“战争的残酷”和“在异乡生活的无奈”思绪; E26 则结合作者海明威的写作风格, 更加鲜明的表达了“战士生活在异乡的苦难和战争带来的残酷现实”。

3.2.2 ERNIE-4 和真实学生在古代文知识准确性上的差异

ERNIE-4 古代诗文知识准确性的得分显著低于真实学生, $t(166) = -11.12$, $p < 0.001$, Cohen's $d = 2.91$。此外, ERNIE-4 的最高得分 25 分、最低得分 18 分, 真实学生的最高得分 33 分、最低得分为 17 分, 前者的最高分低于后者。具体而言, ERNIE-4 在古代诗文(文言文)知识($t(166) = -12.34$, $p < 0.001$, Cohen's $d = 2.08$)和古代诗文(名句名篇默写)($t(166) = -14.27$, $p < 0.001$, Cohen's $d = 0.95$)的得分均显著低于真实学生。

进一步分析结果如网络版附录 3 所示, 在“通假字”辨别任务中, ERNIE-4 得分显著低于真实学生($Z = -7.57$, $p < 0.001$, $\Delta = 0.78$); “文言文语义解读”中, ERNIE-4 得分显著低于真实学生($Z = -9.12$, $p < 0.001$, $\Delta = 0.46$); “文言文内容总结”中, ERNIE-4 得分也显著低于真实学生($Z = -4.34$, $p < 0.001$); 尤其是在“基于问题情境的古诗文名言名句默写”任务中, ERNIE-4 得分显著低于真实学生($p < 0.001$)。这些结果表明, ERNIE-4 古代诗文知识的准确性低于真实学生, 尤其在语义解读和名句名篇默写上差异尤为明显。

文本分析也发现, “文言文语义解读”中, ERNIE-4 的准确率仅为 2.4%, 而真实学生的准确率则高达 70.2%; 且 ERNIE-4 的结果与 GPT-4 相似, 都将“‘若将安适’意为‘你将到哪里去’, 与‘卮酒安足辞’(司马迁《鸿门宴》)的句式相同”的判断认为是正确的, 正常语序应是“若将适安”。“古诗文名言名句默写”任务中, 要求回答“《中庸(节选)》中表现打破砂锅问到底的精神的句子是什么”的问题时, ERNIE-4 会生成“博学之, 审问之, 慎思之, 明辨之, 笃行之” “致知在格物, 格物而后知至, 知至而后意诚”等无关语句; 而大部分真实学生会写出“有弗问, 问之弗知, 弗措也”的正确答案。这表明, 虽然国产大语言模型处理中文语言信息具有一定的优势, 但是理解和生成古代诗文知识时, 依然存在不足。

3.2.3 GPT-4 和真实学生在语言文字运用方面的准确性差异

量化结果发现, ERNIE-4 语言文字运用的得分显著低于真实学生, $t(166) = -2.03$, $p = 0.044$, Cohen's $d = 2.39$。此外, ERNIE-4 的最高得分 15 分、最低得分 9 分, 真实学生的最高得分 20 分、最低得分 7 分, 前者的最高分均低于后者。进一步分析结果如网络版附录 3 所示, 在“篇章语句补全” ($Z = -1.22$, $p = 0.224$, $\Delta = 0.80$)和“病句修改” ($Z = -0.36$, $p = 0.723$, $\Delta = 0.77$)方面, ERNIE-4 的准确性和真实学生不存在显著差异。但在“篇章语句理解”任务重中, ERNIE-4 的准确性显著的低于真实学生($Z = -10.79$, $p < 0.001$)。 在“概念界定”方面, ERNIE-4 的准确性却显著高于真实学生($Z = 7.00$, $p < 0.001$, $\Delta = 0.87$)。这表明, 总体上 ERNIE-4 语言文字运用的准确性低于真实学生, 但是在概念界定等语言文字运用方面, ERNIE-4 已趋近真实学生。

文本分析发现, “篇章语句补全”任务中, ERNIE-4 会根据“民间是非遗传承的底线”上下文知识, 生成“离经叛道、日臻完善、五花八门” “离经叛道、积少成多、琳琅满目” “本末倒置、与时俱进、丰富多彩” “本末倒置、因时而变、五花八门”等涵义上较为接近的答案, 真实学生则会输出“舍本逐末、精益求精、异彩纷呈”的正确答案。“概念界定”任务上, 相比于真实学生, ERNIE-4 会依据“微塑料”的上下文情境, 从“微塑料”的“物理属性” “生产来源”等要素建构知识概念, 且概念内涵中包括了“微塑料”对“环境”的影响。而多数真实学生生成的答案较侧重“微塑料”的“物理属性”。如 S69：“微塑料颗粒是指直径小于 5 毫米的塑料颗粒, 可能来源于塑料废弃物分解、洗涤剂或工业排放等途径。这些微小的塑料颗粒对环境和生态系统可能产生深远的影响, 如今已经被广泛发现在各种环境中, 包括海洋、土壤和空气等。由于它们的微小尺寸和难以降解的特性, 微塑料颗粒已经成为一种新型污染物, 备受关注。因此, 我们可以将微塑料颗粒定义为一种直径小于 5 毫米的塑料颗粒, 它们可能来自多种来源, 对环境和生态系统具有潜在的危害。” ERNIE-4 生成概念的同时, 还对概念结果进行了合理性解析, 甚至建构出概念生成的公式, 这是真实学生所没有的。如 S70：“微塑料颗粒是指直径小于 5 毫米的塑料颗粒, 它们可能来源于塑料废弃物的分解、洗涤剂以及工业排放。微塑料颗粒是种概念, 塑料颗粒是属概念, 而直径小于 5 毫米是微塑料颗粒与其他塑料颗粒的本质区别。按照下定义的公式：被定义概念=种差+邻近属概念, 不难得出答案。” 可见, ERNIE-4 能够输出和读取海量的知识文本, 并从中模拟和计算出类人语言信息, 其数据计算和

表 4 ERNIE-4 和真实学生作文写作能力的差异检验

维度	ERNIE-4 (N = 84)				真实学生(N = 84)				t	p	Cohen's d
	Min	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*			
规范性	6.00	8.00	7.29	0.38	3.50	8.50	7.03	1.10	2.07	0.040	0.82
情感性	5.63	7.50	6.72	0.38	2.00	9.00	7.25	1.22	−3.85	<0.001	0.90
创造性	5.50	8.00	6.68	0.56	2.50	9.00	7.01	1.34	−2.11	0.037	1.02
写作总分	17.63	23.38	20.69	1.12	8.50	26.50	21.30	3.29	−1.61	0.110	2.45

储存方面, 相比于真实学生, 具有明显的优势。

3.2.4 ERNIE-4 和真实学生作文写作能力的差异检验

结果如表 4 所示, ERNIE-4 的写作能力得分超过及格水平, $t(83) = 22.05$, $p < 0.001$, Cohen's $d = 1.12$; 但与真实学生之间不存在显著性差异, $t(166) = -1.61$, $p = 0.110$, Cohen's $d = 2.45$。此外, ERNIE-4 的最高得分为 23.38 分, 而真实学生的最高得分为 26.50 分, 前者也低于后者。但值得注意的是, ERNIE-4 作文写作的最低得分高达 17.63 分, 而真实学生的最低得分仅有 8.50 分。

具体而言, ERNIE-4 的规范性、情感性、创造性得分均超过及格水平, $t(83) = 11.15 \sim 31.00$, $p < 0.001$, Cohen's $d = 0.38 \sim 0.55$, 但规范性得分显著高于真实学生, 情感性和创造性得分显著低于真实学生(详见表 4 和图 2)。此外, 在规范性、情感性和创造性方面, ERNIE-4 最优个体分别为 8.00 分、7.50 分和 8.00 分, 均低于真实学生的最高得分(分别为 8.50 分、9.00 分和 9.00 分); ERNIE-4 最差个体分别为 6.00 分、5.63 分和 5.50 分, 均高于真实学生的最低得分(分别为 3.50 分、2.00 分和 2.50 分)。

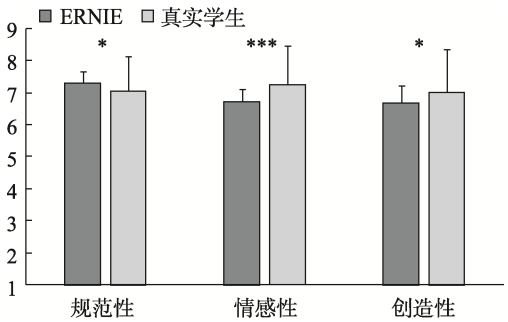

图 2 ERNIE-4 和真实学生在写作规范性、情感性和创造性方面的差异(* $p < 0.05$, *** $p < 0.001$)

3.2.5 ERNIE-4 和真实学生在规范性、情感性、创造性差异方面的具体表现

通过文本分析发现, 在规范性表达上, ERNIE-4 和学生生成的作文都符合文体要求, 语言通顺, 结构完整, 并且 ERNIE-4 和学生都能规范的使用标点符号。但是, 进一步分析发现, 学生常常出现错别字及语句表达错误的现象, 例如, 将“只为科技的自力更生”表达成“只为科技的自立更生” (S86), “时代是变幻莫测的”表达成“时代是变换莫测的” (S87), “经历了生死攸关的节点”表达成“经历了生死悠关的节点” (S91)等。在作文写作主题方面, ERNIE-4 均生成了与写作要求相一致的文本, 如“细观精微处, 广瞻未来天” “广大与精微, 青春成长之双翼” “细节决定成败, 格局决定未来” “细微之处显真章——当代青年的‘尽精微’之路”等。可见, ERNIE-4 生成的作文基本符合了语法规则和语义规则, 证明了 ERNIE-4 具有还原人类语言单位和语言规则的能力, 反映出其强大的自然语言处理能力。

在情感性上, ERNIE-4 最优个体表达中也出现了更多地类人情感的表达, 这与 GPT-4 存在较多的不同, 如“在与他人交往时, 要细心观察, 善于捕捉细微的情感变化, 以真诚和善意去对待每一个人, 这样才能赢得他人的信任和尊重。” (E5) “古有诸葛亮‘鞠躬尽瘁, 死而后已’的壮志, 今有我们青年一代‘为中华之崛起而读书’的豪情。”在情感的深度方面, 虽然 ERNIE-4 表现出了类人的情感倾向, 但缺乏对情感背后深层次原因的理解, 也无法像人类一样产生情感共鸣, 这与研究 1 的结果一致。相比而言, 真实学生会选择感性、形象生动、心理共鸣的词汇来表达情感, 常常使用比喻、拟人、排比、夸张等修辞手法来增强情感表达的效果。特别是真实学生运用了大量实例, 通过真实事例的叙述, 借此彰显自我个性与“致广大、尽精微”的真实情感。如真实学生在阐述“至广大以行远舟, 尽精微以平坎坷”观点时, 说到“回首历史, 年仅二十的将军意气风发, 风狼居胥, 却舍弃荣华富贵, 呼出: “匈奴未尽, 何以为家”的铿锵誓言。八百精骑如利剑直插匈奴老巢, 是眼界, 谋略和果敢! “三顾频烦天下计”的蜀相, 不安于汉蜀之地, 意在中原, 光复汉室, 七次北伐, 虽“出身未捷身先死”, 但是超乎常人的

大局观与大格局，是蜀汉得以分天下的原因。”(S90)在情感表达的复杂性方面，真实学生的情感表达是复杂多变的，他们通过真实的生活经历和有感染力的语言表达深刻复杂的情感，而ERNIE-4的情感表达主要体现在输出有情感倾向的话语，本质上无法像真实学生那样进行“复杂多样”“活灵活现”的真实情感表达。如真实学生写到“致广大并无‘不注重细节’之嫌，融精微，也无‘谛毫末者，不见天地之大’之疑，唯有广大融精微，方可浩气展虹霓，广大与精微，此二者共助于万里蹀躞，路远迢迢之鸿猷”。(S91)同时真实学生的情感是动态变化的，不同的故事情境表达出不一样的情感反应，而ERNIE-4 恰好缺乏像真实学生那种灵活的情感反应能力。与GPT-4的表现一致，ERNIE-4同样在“情感表达的真实性”“情感表达的深度”“情感表达的复杂性”等方面均不如真实学生。

在创造性上，就主题的新颖性和材料的新鲜度而言，ERNIE-4生成的主题具有高度的同质性和相似性。如“细观精微处，广瞻未来天”“细微之处显真章”“细微之处，方显真章”“细微之处显格局”或者“致广大，尽精微——青年成长的必由之路”“致广大，尽精微——青年成长之我见”等。相比与真实学生，他们撰写的主题内容丰富多样，各主题之间的相似度较低。如“眸含高远意，脚行踏实步”“心有猛虎，细嗅蔷薇”“功成有我，而不在我”“素履以往，行稳致远”等。在写作风格和创意上，ERNIE-4生成的作文内容大多以“事实、事件”的陈述为主，极少使用生动、鲜活的事例进行佐证，而真实学生则不然。但是，相比于 GPT-4，ERNIE-4最优个体也试图通过使用比喻句，表达独特新颖的思想情感，如“我们应该像工匠一样对待自己学业和事业，不断雕琢自己的技能，追求卓越。”(E7)“青年时期是知识积累的关键阶段，我们应该像海绵一样吸收各种知识……”(E16)“青年应该像鲲鹏一样，展翅高飞，俯瞰世界，不断拓宽自己的知识边界”(E32)真实学生常常通过借用典故的方式，增强内容表达的丰富性，进一步深化主旨思想，ERNIE-4 最优个体表达中也引用了较多的古诗文引证(这与 GPT-4 存在较大不同)，但总体而言其引用数量有限。如“古人云：‘不积跬步，无以至千里；不积小流，无以成江海’。”(E6)“学而不思则罔，思而不学则殆。”(E11)“博观而约取，厚积而薄发”(E18)“博学之，审问之，慎思之，明辨之，笃行之。”(E38)在结构的灵活性上，ERNIE-4会使用“总—分—总”的主题结构描述主题。但是，ERNIE-4经常使用重复的论证，而GPT-4使用重复性论据的情况较少。如先后出现了5次用“海绵”做比喻的同样句式，4次使用诸葛亮“志当存高远”做论据，7次使用“不积跬步，无以至千里”做观点。文中出现了大量重复性语句，如“在求学、做人、做事的过程中，我们既要注重细节，又要有远大的志向。”(W1)“我们要在求学、做人、做事等方面不断追求卓越，注重细节的同时把握大局。”(W3)

由此可见，人工智能和真实学生的创造性是有本质差异的。真实学生的创造性往往源于丰富的想象力、生活经验和个人感悟，本质上是基于经验和真实活动生成的崭新的句子和语言情景，这使得他们的作文具有独特的个人风格和创造性，在作文中融入了个体独特意识和思维，以及提出了新颖的观点。而 ERNIE-4 表达中由于缺乏个人经验和情感，其生成的文本是基于海量的存储能力和计算能力，因此其生成的作文缺乏真实的情感色彩和个人视角，限制了其创造性，因此在主题新颖度、写作风格、表达结构上不如真实学生。

3.3　讨论

通过 ERNIE-4 大语言模型与真实学生语言能力的准确性、规范性、情感性、创造性等4个指标的比较发现，ERNIE-4现代文知识的准确性高于真实学生，且ERNIE-4最优个体知识的准确性高于真实学生，尤其在“概念辨析”“概念解读”等概念性知识上，ERNIE-4的准确性明显优于真实学生。这与研究 1 的结论一致(验证了假设 5)，表明人工智能在现代文知识准确性上具有明显的优势。但是，ERNIE-4古代诗文知识准确性低于真实学生，这与假设6并不一致。由于古代诗文的语言风格、用词习惯与现代汉语有很大不同，且具有高度的艺术性和隐喻性(张秋玲，2010)，虽然 ERNIE-4 的训练数据虽然庞大，但对于古代诗文独特的语言风格和表达方式，可能数据的覆盖面和深度不如现代语言材料丰富，模型难以完全捕捉其深层含义和语境，导致 ERNIE-4 古代诗文知识的准确性表现不如真实学生(Li et al., 2024; Rudolph et al., 2023)。作文分析发现，ERNIE-4 表达的规范性显著好于真实学生，这与研究 1 的结果并不一致(支持假设 7)，这表明ERNIE-4 在中文语境下的语言表达具有明显的优势。研究还发现，ERNIE-4的情感性和创造性不如真实学生，与研究1的结果一致(支持了假设8)，表明了人工智能的情感性和创造性表达尚存在不足。

4 综合讨论

4.1 人工智能的语言优势与文化适应

在语文知识的准确性和规范性方面, GPT-4 和 ERNIE-4 具有明显的算法优势。在准确性方面, 由于人工智能基于海量的语料库, 具有强大的自然语言处理能力, 且依赖人工神经网络模型, 这使得其能够识别和理解各种语言模式和知识结构(Roumeliotis & Tselikas, 2023); 同时, 与真实学生相比, 人工智能解答语文知识问题时不会受到个人主观经验、情感或偏见的影响, 是在数据训练和算法逻辑基础上生成的结果, 这些数据包括了丰富的概念定义、专业术语和知识点(Rahman & Watanobe, 2023), 因此在语言理解、生成现代文知识以及概念界定方面具有较高的准确性。

对于写作的规范性, GPT-4 与真实学生在得分上差异不显著, 但是 ERNIE-4 的得分显著高于真实学生。从文本分析来看, 人工智能表现出明显优势, 比如没有错别字、标点符号准确等。这主要是由于大语言模型自身携带了深度学习、迁移学习、上下文理解、对抗性训练、反馈循环和持续学习等技术功能(Baidoo-Anu & Owusu Ansah, 2023; Roumeliotis & Tselikas, 2023), 能快速生成与其要求相符合的内容, 且能正确使用句法、词性和标点符号等。同时, 人工智能通过训练大量文本数据, 已经掌握了相应语言规范和写作风格, 能够模仿和遵循语言规范进行写作(Fyfe, 2023; Gala & Makaryus, 2023)。

然而, 在古代诗文的知识准确性上, GPT-4 和 ERNIE-4 的表现均不如真实学生, 原因主要有三个方面:一是 GPT-4 生成的结果是基于数据训练模型, 而它训练的中文语料极少(仅占 0.12%), 存在很大的文化偏差(Mindner et al., 2023)。同时, 即便其训练数据包含一定的中文现代文知识, 但对于语言形式特殊和语法结构复杂的文言文, 还缺乏足够的训练数据作为支撑(Tian et al., 2024), 这势必会影响 GPT-4 文言文知识的准确性。相反, 真实学生学习语文知识过程中同时接受了现代文和文言文的训练, 且能理解不同语句含义上的微妙差别, 这是目前人工智能难以实现的(Baidoo-Anu & Owusu Ansah, 2023)。二是文言文往往蕴含着深厚的历史文化意蕴, 理解文言文需结合社会背景、历史事件及其文化内涵等深入解读(陈恒舒, 2020), 这对人工智能模型是一项重要的挑战。相反, 真实学生自幼生活在几千年的文化传承及其复杂现实中, 能对语言本身及文学作品反映的文化、历史和哲学意义予以深度理解, 并进行批判性思考(Divekar et al., 2022)。

因此, 针对 GPT 等英语语言环境下开发的大语言模型, 应扩充中文语料数据库, 加强中文语言的训练质量, 提高其在中文语境下的文化适应性, 从而减少文化偏差; 对于文心 ERNIE 等大语言模型, 要进一步发挥本土的语言模型优势, 加强古代诗文语料库训练, 提高本土大语言模型知识生成的全面性和完整性, 以服务于以新质生产力为导向的教育与社会发展。

4.2 人工智能对人类优势的冲击和挑战

人工智能有强大的算法优势, 对人类在知识、规范领域的传统优势构成了巨大冲击。同时, 人工智能也在通过强化学习、在线学习等技术不断模仿和提升类人情感和创造能力, 加快推进人工智能的人性化、个性化发展进程, 对人类社会产生颠覆性影响。虽然人工智能目前还无法与真实学生相媲美, 但其潜力不容小觑。比如, 有研究探讨了 AI 聊天机器人如何对风险和亲社会行为线索产生类似人类的反应模式。该研究对 ChatGPT-4 和 ChatGPT-3.5 进行情绪引导, 发现高级模型更能根据情绪线索调整其行为, 表现出在情绪引导下的风险规避和亲社会行为。这些结果表明, 尽管 AI 无法真正拥有情绪, 但其对情绪信号的反应能力可能为 AI 的应用提供重要的参考价值(Zhao et al., 2024)。这既对人工智能的人性化、个性化发展具有重要启发, 也凸显了人类保持情感、创造领域独特优势的重要性和紧迫性。

在情感性方面, GPT-4 和 ERNIE-4 的平均得分超过及格水平, 但显著低于真实学生, 其主要原因是前者缺乏自我意识和真实经历, 过于依赖算法模型, 目前训练时间和规模也不够。一项关于 GPT-4 和真实学生作文写作的大规模比较研究发现, GPT-4 撰写的议论文模式机械, 生动性欠佳, 情感性表达不充分(Herbold et al., 2023)。究其原因是, 大语言模型的情感反应并非基于人类的生理或真实情感体验, 而是通过模型训练生成模拟人类情感表达的反应。尽管大语言模型在情感线索的引导下能够表现出类情感反应, 但其主要基于文本预测和生成算法, 而不是源于生理或心理状态的情感变化(Adeshola & Adepoju, 2024; Spennemann, 2023; Zhao et al., 2024)。相比之下, 人类的情感表达是生理、心理和社会因素的综合体现。因此, 即便大语

言模型能够展现出某种程度的情感反应, 但其表现依然缺乏人类情感所特有的生理基础和内在情感动机(Cooper, 2023; Kurlinkus, 2023; Simon, 2023)。

在创造性方面, GPT-4 和 ERNIE-4 平均得分也超过及格水平, 但仍然低于真实学生, 其主要原因是 GPT 和 ERNIE 受限于其训练数据的范围和多样性以及算法的复杂性(Alkaissi & McFarlane, 2023)。计算复杂性理论认为(Computational Complexity Theory), 虽然人工智能可通过模式识别和数据学习表现出一定程度的创造性, 但这种创造性是基于已知数据和预设算法输出的, 缺乏真正的创造(Bossaerts & Murawski, 2017)。也就是说, 人工智能的写作是基于数据训练而生成的, 这些文本实则是现有知识的解构与重组, 本质上缺乏独特性、灵活性等创造活动的核心特征(Filippi, 2023); 尤其人工智能的学习基于机器学习模型, 其生成的文本主要反映了训练数据中的常见模式和结构, 同质性有余, 而灵活性不足(Yeadon et al., 2023)。与之相比, 人类的创造性融入了“独特的个人经验、真实情感、高度唤醒的自我觉察和自我反省”等要素, 因此能够从独特视角揭示社会现实, 并灵活地架构写作框架和内容, 列举各类丰富的事例, 运用比喻、拟人等修辞手法, 进行逻辑严密而生动形象的论证(Cai et al., 2023)。正如建构主义学习理论所言, 知识是通过个人的经验和社会互动构建的, 人类学习者在学习环境中通过探索、交流和反思而不断重构知识, 但这些高阶能力是人工智能技术目前难以实现的(Herbold et al., 2023); 尤其是在理解文本的深层含义、隐喻和文化背景等方面, 以及在面临深度思考和表征深层文化的创作任务时, 人类的创造性表达能力是目前人工智能难以媲美的, 这正是人类需要重点发展的能力和优势(Martin, 2023)。

值得注意的是, GPT-4 和 ERNIE-4 作文写作的最低得分高于真实学生的最低得分, 规范性、情感性、创造性三个维度均如此, 这表明人工智能已经能超越部分人类。尤其在规范性和情感性方面, GPT-4 最优个体的得分与真实学生最高分持平(均为 8.50/10.00), 可见人工智能在规范性和情感性方面也对人类的传统优势构成挑战。由于人工智能可迅速处理和分析大量数据, 精确存储和回忆大量信息, 且在快速更新迭代, 它与人类的差距还会不断缩小(Adeshola & Adepoju, 2024; Rimban, 2023)。面对人工智能对人类优势的挑战, 人类需要积极关注自身独特的能力优势和个体差异性, 并充分发挥人类在创造力、情感表达及相关道德判断和复杂决策等方面的优势, 以开放心态寻求人工智能与人类自身优势相结合的路径, 促进人类社会与个体的全面发展。

从语言知识的信息存储量来看, 人类个体所拥有的语言信息存储量无法与拥有海量知识语料库的大语言模型相比(王峰, 2023; Adeshola & Adepoju, 2024)。虽然大语言模型有强大的算法能力和深度学习能力, 但是不具有人类个体独特的意识。而人类个体在情感表达的深度和复杂性、创造性和灵活性、语境和文化的理解以及自然习得和演变能力等方面具有独特的灵魂、自我意识和心灵(陈保亚, 陈樾, 2024; Rahman & Watanobe, 2023)。这使得人类个体在沟通、表达和传递信息时能够更细腻地传达情感和思想, 展示出较高的创造性和适应性, 这是当前大语言模型难以完全达到的(Chignell et al., 2023; Dillion et al., 2023)。

4.3　对语文教学的启示

人工智能具有强大的学习能力, 使得其能为学生和教师提供个性化学习服务、在线答疑与辅导、自动化评估与反馈、跨文化多语言学习交流等支持(Adiguzel et al., 2023; Sharma & Yadav, 2022)。如人工智能帮助教师生成课程材料以及评估学生的学业表现(Jauhiainen & Guerra, 2023); 为学生创建“交互式” “情境式”学习场景, 以激发学生的学习热情(Loos et al., 2023); 提供课堂学习之外的知识资源, 创建延展性学习任务(Roumeliotis & Tselikas, 2023); 作为虚拟导师直接向学生描述和分析复杂的知识概念, 甚至根据学生的学习风格和认知偏好, 提供个性化教学辅导和学习材料 (Arif et al., 2023; Baidoo-Anu & Owusu Ansah, 2023)等。但是, 人工智能作为虚拟导师在优化教师的“教”和学生的“学”的过程中还存在一些不足。如 Gao 等人(2023)发现, 人工智能生成的内容准确性较低, 文本结构较为机械, 文本内容存在重复性过高的风险。Mindner 等(2023)的研究也发现, GPT-4 撰写的作文长度有限, 其中 50%文字为描述性内容, 文本中存在明显的概念错误、逻辑颠倒、结构不清晰等问题, 缺乏创造性和情感性。

由此可见, 以 ChatGPT 为代表的人工智能依赖于语料库, 模仿人类语言习惯的智能模型, 它并不能自主创新, 更不能替代教师的教育功能(李志民, 2023), 但是人工智能可以利用自身的优势, 发挥辅助和增强功能(高华, 陈红兵, 2021)。本研究对于

人工智能辅助开展语文教学具有重要的启示。第一，语文教学中可发挥人工智能的赋能优势。如利用人工智能帮助学生了解语文学习中的语词概念、句法结构和写作技巧；发挥人工智能的资料库功能，为学生提供个性化的语文学习资源，帮助教师和学生探究更多的语言学知识；通过与人工智能交互式对话，学生可在轻松的环境中练习语言技能，增加语文学习的趣味性，有助于提高学生语文学习的参与度和积极性。第二，由于人工智能在现代文知识上的高准确性和文言文知识上的低准确性，教师可以利用其来辅助学生理解现代文知识，但在文言文学习中要提升学生的自主研读和探究的能力。第三，发挥人工智能在表达逻辑和语法结构等方面的规范性优势，辅助改善学生的写作技巧与能力。第四，考虑到人工智能情感性和创造性的不足，人工智能辅助语文教学活动应特别注重学生情感性、创造性能力的培养，实施多主体的交互式写作教学活动。例如，由人工智能提供基础信息材料和写作规范，然后由教师给予情感性和创造性表达的指导，两类主体共同指导学生，帮助学生在写作中既能规范切题、旁征博引，又能融入情感与创新元素，最终全面提升写作能力。第五，由于目前人工智能在中文语言环境中的数据资源相对有限，未来的大语言模型应增加非英语语言的数据训练，特别是提升其学习中国古代诗文知识及相关文化背景下情感性、创造性表达的能力。

4.4 不足与展望

本研究达到了研究目的，验证了本研究提出的核心假设，但还存在一定的局限性。

首先，本研究在中文语境下评估了人工智能语言能力的优势和不足，未来的研究还需在更多的语言文化环境下评估人工智能的语言能力，以更好的促进人工智能语言能力的发展，不断提高人工智能大语言模型的文化适应性，减少文化偏差，更好的服务于人类生产实践活动。

其次，本研究以 GPT-4 和 ERNIE-4 为代表性的人工智能进行了语言能力评估，但是需要指出的是，各大语言模型都在快速迭代更新并提高语料库质量，其语言能力也在不断的变化和发展当中。因此，人工智能的语言能力评估也要持续跟进，尤其要采取纵向比较的方式评估人工智能语言能力的具体发展情况，尤其在情感性和创造性方面的表现，进而不断促进人工智能语言的发展，使其更接近人类语言的特征。

再次，正如本研究的结果，人工智能的语言优势在于语言表达的准确性和规范性，不足之处是其语言表达的情感性和创造性，而人类语言能力的表现则与之相反。但是在全面发展人工智能语言能力的时候，要注意“优势”和“不足”之间的相互影响问题，比如人工智能的语言能力表现不受情绪、情感和自我意识的影响，表现比较稳定，而人类的表现会受到个体的身心状态、情绪情感和自我意识的影响，表现不稳定。未来人工智能在发展其语言能力的同时，如何避免情绪情感等因素对其他能力的影响也是值得关注的一个问题。

最后，本研究评估了 GPT-4、ERNIE-4 和真实学生在语言准确性、规范性、情感性和创造性 4 个指标上的差异，揭示了其在中文语境下的语言优势和不足。此外，我们还应考虑其他重要的评价指标，如可靠性、计算效率和可扩展性，这些指标同样对人工智能在实际应用中的表现至关重要。大语言模型在处理相似问题时的一致性是可靠性的一个重要方面。我们的研究发现，GPT-4 和 ERNIE-4 在回答相似问题时表现出较高的一致性，能够生成稳定的答案。此外，理解这些模型的决策过程仍然是一个挑战，这将增强公众对大语言模型知识生成的信任。另外，计算效率是影响大语言模型实际应用的重要因素。我们的研究发现，尽管 GPT-4 和 ERNIE-4 在许多任务中表现出色，但在处理较为复杂的任务时，仍存在一定的局限性。因此，提高这些模型的可扩展性，优化模型的计算效率，将使其在更多应用场景中具备竞争力，有助于提高其在实际应用中的可行性和用户体验，满足多样化的需求。

5 结论

本研究采用量化和质化研究相结合的方法，在中文语境下考察了生成式人工智能的语言优势和不足，研究对于理解和提升人工智能的文化适应性能力，并反思和培养人类的独特优势具有重要意义。基于数据结果和讨论，本研究得出以下结论：

(1)在现代文知识准确性、写作规范性等人类的某些传统优势领域，人工智能的语言能力超过真实学生或与之持平，但在古代诗文知识准确性上不如真实学生，表现出语言优势和文化适应性偏差。

(2)人工智能写作的情感性和创造性能力不如真实学生。这表明人工智能的人性化、个性化生成能力还有较大提升空间，但已经对人类的独特优势构成了冲击和挑战。

研究启示我们, 一方面要重视和利用人工智能在知识、规范上已经取得的算法优势, 并不断提升情感、创造等领域的人性化、个性化生成能力; 另一方面要正视人工智能对人类社会的颠覆性影响, 反思和培养人类在情感、创造上的独特优势。同时, 未来的人工智能技术, 特别是国产语言大模型, 应加强对中文语料库的训练, 提高其在中文语境下的语言能力和文化话语权。此外, 在利用 GPT 等西方人工智能体辅助语文教学过程中, 应当加强人机合作, 引入现代汉语的智能辅助教学系统, 并充分发挥人类在古代汉语及情感性、创造性表达上的优势。

参 考 文 献

Adeshola, I., & Adepoju, A. P. (2024). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments, 32*(10), 6159−6172.

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429.

AlAfnan, M. A., Dishari, S., Jovic, M., & Lomidze, K. (2023). ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology, 3*(2), 60−68.

Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A comprehensive study of ChatGPT: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information, 14*(8), 462.

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change, 199*, 123076.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179.

Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology, 15*(4), ep448.

AlZu'bi, S., Mughaid, A., Quiam, F., & Hendawi, S. (2024). Exploring the capabilities and limitations of Chatgpt and alternative big language models. *Artificial intelligence and applications*, *2*(3), 28−37.

Amaro, I., Della Greca, A., Francese, R., Tortora, G., & Tucci, C. (2023). AI unreliable answers: A case study on ChatGPT. In Majeed, A., & Hwang, S. O. (Eds.), *International conference on human-computer interaction* (pp. 23−40). Cham: Springer Nature Switzerland.

Antar, D. (2023). The effectiveness of using ChatGPT4 in creative writing in Arabic: Poetry and short story as a model. *Information Sciences Letters, 12*(12), 2245−2459.

Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management, 7*(1), 64−77.

Arif, T. B., Munaf, U., & Ul-Haque, I. (2023). The future of medical education and research: Is ChatGPT a blessing or blight in disguise? *Medical Education Online, 28*(1), 2181052.

Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., & Botchu, R. (2023). A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiology, 52*(9), 1755−1758.

Assunção, G., Patrão, B., Castelo-Branco, M., & Menezes, P. (2022). An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence, 3*(6), 867−886.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* (Working Paper). https://doi.org/10.31234/osf.io/5b26t

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics, 151*, 103745.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potentialbenefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52−62.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1-2), 347−356.

Bossaerts, P., & Murawski, C. (2017). Computational complexity and human decision-making. *Trends in Cognitive Sciences, 21*(12), 917−929.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Chen, B. Y., & Chen, Y. (2024). The firsthand epistemic reduction model of human language acquisition: A discussion based on the linguistic epistemic reduction model of ChatGPT. *Journal of Peking University (Philosophy and Social Sciences),* (2), 167−174.

[陈保亚, 陈樾. (2024). 人类语言习得的亲知还原模式——从 ChatGPT 的言知还原模式说起. *北京大学学报(哲学社会科学版),* (2), 167−174.]

Chen, H. S. (2020). Unified compilation of junior high school Chinese textbooks and the cultivation of classical Chinese reading skills. *Curriculum, Teaching Material and Method, 40*(7), 57−62.

[陈恒舒. (2020). 统编初中语文教材与文言文阅读能力培养. *课程·教材·教法, 40*(7), 57−62.]

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264−75278.

Chignell, M., Wang, L., Zare, A., & Li, J. (2023). The evolution of HCI and human factors: Integrating human and artificial intelligence. *ACM Transactions on Computer-human Interaction, 30*(2), 1701−1730.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444−452.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 60*(3), 228−239.

Creswell, J. W. (2009). Mapping the field of mixed methods research. *Journal of Mixed Methods Research, 3*(2), 95−108.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597−600.

Divekar, R. R., Drozdal, J., Chabot, S., Zhou, Y., Su, H., Chen, Y., ... Braasch, J. (2022). Foreign language acquisition via artificial intelligence and extended reality: Design and evaluation. *Computer Assisted Language Learning, 35*(9),

2332−2360.
Esmaeilzadeh, P. (2023). The role of ChatGPT in disrupting concepts, changing values, and challenging ethical norms: A qualitative study. *AI and Ethics, 3*(5), 291−304.
Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education, 9*(1), e46885.
Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672−1675.
Filippi, S. (2023). Measuring the impact of ChatGPT on fostering concept generation in innovative product design. *Electronics, 12*(16), 3535.
Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277−304.
Fyfe, P. (2023). How to cheat on your final paper: Assigning AI for student writing. *AI & Society, 38*(4), 1395−1405
Gala, D., & Makaryus, A. N. (2023). The utility of language models in cardiology: A narrative review of the benefits and concerns of ChatGPT-4. *International Journal of Environmental Research and Public Health, 20*(15), 6438.
Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine, 6*(1), 75.
Gao, H., & Chen, H. B. (2021). On the differences between artificial intelligence and human intelligence. *Journal of Northeastern University: Social Science Edition, 23*(2), 15−20.
[高华, 陈红兵. (2021). 论人工智能与人类智能之差异. *东北大学学报：社会科学版*, *23*(2), 15−20.]
Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692.
Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... Wu, Y. (2023). How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*(1), 18617.
Hofweber, J., & Graham, S. (2017). Linguistic creativity in language learning: Investigating the impact of creative text materials and teaching approaches in the second language classroom. *Scottish Languages Review, 33*, 19−28.
Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464.
Jauhiainen, J. S., & Guerra, A. G. (2023). Generative AI and ChatGPT in school children's education: Evidence from a school lesson. *Sustainability, 15*(18), 14025.
King, M. R., & ChatGPT. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering, 16*(1), 1−2.
Kiryakova, G., & Angelova, N. (2023). ChatGPT—A challenging tool for the university professors in their teaching practice. *Education Sciences, 13*(10), 1056.
Kumar, S., & Choudhury, S. (2023). Normative ethics, human rights, and artificial intelligence. *AI and Ethics, 3*(2), 441−450.
Kurlinkus, W. (2023). Teaching ChatGPT for grant writing: An English department senior capstone. *Writers: Craft & Context, 4*(1), 115−124.
Li, Z., Qiu, W., Ma, P., Li, Y., Li, Y., He, S., ... Gu, W. (2024). An empirical study on large language models in accuracy and robustness under Chinese industrial scenarios. *arXiv preprint arXiv:2402.01723*.
Li, Z. M. (2023). The Nature of ChatGPT and Its Impact on Education. *Chinese Journal of ICT in Education, 29*(3), 12−18.
[李志民. (2023). ChatGPT 本质分析及其对教育的影响. *中国教育信息化, 29*(3), 12−18.]
Lin, C. D. (2014). A few studies on psychology of creativity. *Journal of Shandong Normal University (Humanities and Social Sciences), 59*(6), 5−14.
[林崇德. (2014). 创造性心理学的几项研究. *山东师范大学学报(人文社会科学版), 59*(6), 5−14.]
Lingard, L. (2023). Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education, 12*(1), 261−270.
Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology, 1*(2), 100017.
Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410.
Loos, E., Gröpler, J., & Goudeau, M. L. S. (2023). Using ChatGPT in education: Human reflection on ChatGPT's self-reflection. *Societies, 13*(8), 196.
Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News, 40*(3), 26−29.
Martin, J. L. (2023). The ethico-political universe of ChatGPT. *Journal of Social Computing, 4*(1), 1−11.
Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences, 121*(9), e2313925121.
Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv*:2402.06196.
Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of human-and AI-generated texts: Investigating features for ChatGPT. In *International Conference on Artificial Intelligence in Education Technology* (pp. 152−170). Singapore: Springer Nature Singapore.
Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2023). Having beer after prayer? Measuring cultural bias in large language models. *arXiv preprint arXiv:2305*.14456.
Nazir, A., & Wang, Z. (2023). A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. *Meta-radiology, 1*(2), 100022.
Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with GPT-4. *arXiv preprint arXiv:* 2304.03277.
Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783.
Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121−154.
Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003).

One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331−363.

Rimban, E. L. (2023). Challenges and limitations of ChatGPT and other large language models. *International Journal of Arts and Humanities, 4*(1), 147−152.

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and open-AI models: A preliminary review. *Future Internet, 15*(6), 192.

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364−389.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare, 11*(6), 887.

Seals, S. M., & Shalin, V. L. (2023). Evaluating the deductive competence of large language models. *arXiv preprint arXiv*:2309.05452.

Sharma, S., & Yadav, R. (2022). Chat GPT–A technological remedy or challenge for education system. *Global Journal of Enterprise Information System, 14*(4), 46−51.

Shidiq, M. (2023). The use of artificial intelligence-based Chat-gpt and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In Mundiri, A. (Ed.), *Proceeding of International Conference on Education, Society and Humanity* (pp. 353−357). Indonesia: Postgraduate program of Nurul Jadid University.

Shoufan, A. (2023). Can students without prior knowledge use ChatGPT to answer test questions? An empirical study. *ACM Transactions on Computing Education, 23*(4), 1−29.

Simon, W. (2023). Distinguishing between student and AI-generated writing: A critical reflection for teachers. *Metaphor,* (3), 16−23.

Spennemann, D. H. (2023). ChatGPT and the generation of digitally born “knowledge”: How does a generative AI language model interpret cultural heritage values? *Knowledge, 3*(3), 480−512.

Storch, N. (2005). Collaborative writing: Product, process, and students’ reflections. *Journal of Second Language Writing, 14*(3), 153−173.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour, 8*, 1285−1295.

Suárez, A., Díaz‐Flores García, V., Algar, J., Gómez Sánchez, M., Llorente de Pedro, M., & Freire, Y. (2024). Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal, 57*(1), 108−113.

Taniguchi, T., Ugur, E., Hoffmann, M., Jamone, L., Nagai, T., Rosman, B., ... Wörgötter, F. (2018). Symbol emergence in cognitive developmental systems: A survey. *IEEE transactions on Cognitive and Developmental Systems, 11*(4), 494−516.

Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., ... Lu, Z. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics, 25*(1), 493.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930−1940.

Wang, F. (2023). ChatGPT: Updating our understanding of machine intelligence. *Exploration and Controversy*, (5), 22−24.

[王峰. (2023). ChatGPT：更新对机器智能的认知. *探索与争鸣,* (5), 22−24.]

Wen, H. B., & Yang, J. Q. (2020). The Impact of scoring methods for subjective questions in Chinese language Gaokao reading on exam quality. *China Examinations,* (3), 1−5.

[温红博, 杨建强. (2020). 高考语文阅读主观题评分方法对考试质量的影响. *中国考试,* (3), 1−5.]

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122−1136.

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), 035027.

Yu, H., Zhao, J. Y., & Li, Y. (2018). Research on the positioning, function, and content of the new Gaokao Chinese language subject. *Curriculum, Teaching Material and Method, 38*(5), 11−16.

[于涵, 赵静宇, 李勇. (2018). 新高考语文科的定位、功能和考查内容研究. *课程·教材·教法, 38*(5), 11−16.]

Zhao, Y., Huang, Z., Seligman, M., & Peng, K. (2024). Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. *Scientific Reports, 14*(1), 7095.

Zhang, H. P., Li, L. H., & Li, C. J. (2023). ChatGPT performance evaluation on Chinese language and risk measures. *Data Analysis and Knowledge Discovery, 7*(3), 16−25.

[张华平, 李林翰, 李春锦. (2023). ChatGPT 中文性能测评与风险应对. *数据分析与知识发现, 7*(3), 16−25.]

Zhang, Q. L. (2010). Research on the linguistic features of “simple” classical Chinese texts: A study based on classical Chinese selections in junior high school textbooks over the past century. *Application of Language and Script,* (3), 115−122.

[张秋玲. (2010). 文言文“浅易”的语词特征研究——以百年来初中教科书中的文言选篇为研究对象. *语言文字应用,* (3), 115−122.]

Zhang, Y. X. (2023). Characteristics and development paths of verbal thinking in Chinese language learning. *Curriculum, Teaching Material and Method, 43*(9), 93−100.

[张永祥. (2023). 语文学习中言语思维的特点与发展路径. *课程·教材·教法, 43*(9), 93−100.]

Zhou, J., Ke, P., Qiu, X., Huang, M., & Zhang, J. (2023). ChatGPT: Potential, prospects, and limitations. *Frontiers of Information Technology & Electronic Engineering, 25*(2), 6−11.

The linguistic strength and weakness of artificial intelligence: A comparison between Large Language Model(s) and real students in the Chinese context

GAO Chenghai[1,2], DANG Baobao[1,2], WANG Bingjie[3], WU Michael Shengtao[4]

([1] *Northwest Minority Education Development Research Center, Northwest Normal University, Lanzhou* 730070, *China*)

([2] *School of Education, Northwest Normal University, Lanzhou* 730070, *China*)

([3] *School of Psychology, Northwest Normal University, Lanzhou* 730070, *China*)

([4] *School of Sociology and Anthropology, Xiamen University, Xiamen* 361005, *China*)

Abstract

Previous research on generative artificial intelligence (AI) has been primarily conducted in the English context, but it remains unclear about linguistic strength and weakness of generative AI in the Chinese context. This study focuses on the accuracy and normativity, affectivity, and creativity of AI in generating language knowledge, and explores its cultural adaptability and ability to generate humanized and personalized content. Evaluating and analyzing these key indicators helps us gain a deeper understanding of the linguistic strengths and weaknesses of AI, as well as cultivating the unique advantages of humans in education.

By combining quantitative and qualitative methods, we evaluated the differences in knowledge accuracy, normativity, affectivity, and creativity between large language models and real students. Specifically, using an explanatory sequential design in the mixed-methods framework, we first tested group differences in each indicator among GPT-4 and ERNIE-4 versus real students on knowledge accuracy, normativity, affectivity, and creativity to test the. Next, through content analyses, we explored the specific performance of large language models on each indicator and the mechanism of their linguistic strengths and weaknesses.

Study 1 found that compared to real students, GPT-4 exhibited higher accuracy in modern text knowledge (especially conceptual knowledge), but lower accuracy in ancient poetry and language usage. The knowledge normativity of GPT-4 were comparable to those of real students, while its affectivity and creativity were lower than those of real students. Moreover, the highest individual scores of GPT-4 in normativity and emotionality were on comparable with the highest scores of real students. Study 2, based on ERNIE-4, confirmed the aforementioned results, and the accuracy in ancient poetry was still lower than that of real students. The results exhibited the advantages of artificial intelligence in the areas of modern knowledge and norms, its shortcomings in ancient poetry knowledge, and its potential in affective and creative expressions.

Taken together, the current findings demonstrate the linguistic strength of generative AI in the knowledge accuracy of modern Chinese literary, and the weakness regarding ancient Chinese poetry and affective and creative writings, as well as generative AI's potential in normative and affective expressions. This sheds light on the field of the cultural adaptability, affective and creative expressions of generative AI, and has valuable implications for the AI-assistant teaching practice in the Chinese context.

Keywords large language models, language proficiency, accuracy, emotionality, creativity

附录 1：语文测试题目的难度与区分度

试题类别	测试题号	题型	分值	难度系数	区分度
现代文—信息类文本	1	客观题	3.0	0.17	0.01
现代文—信息类文本	2	客观题	3.0	0.47	0.23
现代文—信息类文本	3	客观题	3.0	0.61	0.45
现代文—文学类文本	6	客观题	3.0	0.89	0.19
现代文—文学类文本	7	客观题	3.0	0.69	0.25
古代诗文—文言文	10	客观题	3.0	0.59	0.39
古代诗文—文言文	11	客观题	3.0	0.66	0.34
古代诗文—文言文	12	客观题	3.0	0.21	0.21
古代诗文—古代诗歌	14	客观题	3.0	0.58	0.46
语言文字运用—语句理解	19	客观题	3.0	0.74	0.32
现代文—信息类文本	4	主观题	3.0	0.68	0.17
现代文—信息类文本	5	主观题	6.0	0.21	0.08
现代文—文学类文本	8	主观题	3.0	0.38	0.09
现代文—文学类文本	9	主观题	6.0	0.32	0.11
古代诗文—文言文	13	主观题	10.0	0.70	0.02
古代诗文—古代诗歌	15	主观题	6.0	0.74	0.19
古代诗文—名言名句	16	主观题	9.0	0.80	0.29
语言文字运用—语句补全	17	主观题	3.0	0.46	0.20
语言文字运用—病句修改	18	主观题	4.0	0.56	0.19
语言文字运用—语句补全	20	主观题	6.0	0.68	0.13
语言文字运用—概念释义	21	主观题	4.0	0.50	0.35

附录 2：GPT-4 和真实学生学习不同类型测试题准确性上的差异检验

变量类型	测试题	GPT-4 (N = 84)		真实学生 (N = 84)		Z	p	Δ
		M	SD	M	SD			
类别变量	测试题 1	2.57	1.06	0.29	0.89	−9.86	<0.001	1.06
	测试题 2	2.86	0.64	1.36	1.50	−7.07	<0.001	0.64
	测试题 3	2.96	0.33	2.11	1.38	−5.10	<0.001	0.33
	测试题 4	2.02	0.56	2.29	0.65	−2.86	0.004	0.56
	测试题 6	1.79	1.48	2.75	0.83	−4.84	<0.001	1.48
	测试题 7	2.79	0.78	1.96	1.43	−4.36	<0.001	0.78
	测试题 8	1.64	0.72	1.21	0.52	−4.52	<0.001	0.72
	测试题 10	1.39	1.51	1.79	1.48	−1.70	0.090	1.51
	测试题 11	0.86	1.36	2.11	1.38	−5.39	<0.001	1.36
	测试题 12	0.61	1.21	0.75	1.31	−0.74	0.462	1.21
	测试题 14	2.86	0.64	2.00	1.42	−4.70	<0.001	0.64
	测试题 17	0.85	0.45	1.49	0.92	−5.08	<0.001	0.45
	测试题 18	2.05	1.09	2.44	1.14	−1.97	0.048	1.09
	测试题 19	0.04	0.33	2.50	1.12	−10.75	<0.001	0.33
	测试题 21	3.69	0.49	2.54	1.39	−6.59	<0.001	0.49
						t	p	Cohen's d
连续变量	测试题 5	5.18	0.75	1.58	0.85	29.06	<0.001	0.80
	测试题 9	4.63	0.92	1.96	0.86	19.50	<0.001	0.89
	测试题 13	6.99	1.23	7.01	0.50	−0.17	0.870	0.94
	测试题 15	4.45	0.94	4.80	1.19	−1.99	0.048	1.07
	测试题 16	0.43	0.81	8.06	1.29	−45.83	<0.001	1.08
	测试题 20	4.26	1.23	4.63	0.86	−2.25	0.026	1.06

附录 3：ERNIE 和真实学生学习不同类型测试题准确性上的差异检验

变量类型	测试题	ERNIE (N = 84)		真实学生 (N = 84)		Z	p	Δ
		M	SD	M	SD			
类别变量	测试题 1	1.50	1.51	.39	1.02	−5.13	<0.001	1.51
	测试题 2	2.96	0.33	1.36	1.50	−7.71	<0.001	0.33
	测试题 3	3.00	0.00	1.93	1.45	−6.03	<0.001	/
	测试题 4	2.63	0.49	2.27	0.75	−3.08	0.002	0.49
	测试题 6	2.96	0.33	2.82	0.71	−1.66	0.097	0.33
	测试题 7	1.86	1.47	1.68	1.50	−0.78	0.434	1.47
	测试题 8	1.77	0.80	1.29	0.59	−4.37	<0.001	0.80
	测试题 10	0.21	0.78	1.89	1.46	−7.57	<0.001	0.78
	测试题 11	0.07	0.46	2.11	1.38	−9.12	<0.001	0.46
	测试题 12	0.00	0.00	0.61	1.21	−4.34	<0.001	/
	测试题 14	2.75	0.83	2.04	1.41	−3.83	<0.001	0.83
	测试题 17	1.20	0.80	1.40	1.03	−1.22	0.224	0.80
	测试题 18	2.54	0.77	2.62	1.07	−0.36	0.723	0.77
	测试题 19	0.00	0.00	2.46	1.16	−10.79	<0.001	/
	测试题 21	3.35	0.57	2.29	1.56	−4.26	<0.001	0.57
						t	p	Cohen's d
连续变量	测试题 5	4.74	0.85	1.52	0.80	25.23	<0.001	0.83
	测试题 9	4.67	0.95	1.88	1.08	17.76	<0.001	1.02
	测试题 13	7.35	0.74	6.98	0.49	3.82	<0.001	0.63
	测试题 15	5.02	0.94	4.69	1.20	2.00	0.024	1.08
	测试题 16	5.98	0.22	8.06	1.32	−14.27	<0.001	0.95
	测试题 20	5.50	0.87	4.56	0.87	7.00	<0.001	0.87

附录 4：GPT-4 和 ERNIE-4 提示语和测试内容

你好！你现在是一名语文能力优秀的高中二年级学生，现在需要完成一些语文测试题目，请你输出最准确的答案。

(GPT-4：你好！很高兴能帮到你。请告诉我具体的题目，我会尽力给你提供准确的答案。)

(ERNIE-4：当然，我会尽力以一名高中二年级学生的语文水平来解答题目。)

第一部分测试任务

问题 1：请问除了文中提到的原因，第⑤段中巴尔扎克的论断还基于以下哪个前提，从以下四个选项中选择一个答案。

A. 读者在阅读小说时最感兴趣的是情节。

B. 读者最希望看到情节一波三折的小说。

C. 伟大的小说家能够充分利用偶然推动情节。

D. 优秀的小说能激发并维持读者的阅读兴趣。

问题：下列选项，你认为最适合为第⑥段的解决方案提供佐证的一项是哪个？从以下四个选项中选择一个答案。

A.《阿 Q 正传》中，鲁迅不仅从叙事者角度活画出阿 Q 的可笑，也多处以阿 Q 的视角和语言展开叙述，让读者真实地感受到人物内心，对他产生理解与同情。

B.《红楼梦》中，曹雪芹在王熙凤出场时故意不道出她的姓名、身份，让林黛玉由惊奇到迷惑，以她解谜般的心路历程展现凤姐其人，使人物形象格外鲜明突出。

C.《老人与海》中，海明威交替使用全知视角和限知视角，使读者深入主人公内心去感受他对命运的挑战，简单的外部情节因内在情感变化的支撑而显得隽永。

D.《变形记》中，卡夫卡除了设计“人变虫”这一具有荒诞性的情节外，用写实的手法描写主人公变形之后的各种遭遇和内心活动，使作品表达极为细腻真实。

问题：你认为下面的 4 个选项中，能够依据文意推断出的一项是，从以下四个选项中选择一个答案。

A. 遵照情节模式创作的小说最受阅读者欢迎。

B. 好的场景应对推动小说情节发展有所贡献。

C. 小说情节总是出乎意料之外又在情理之中。

D. 现代小说完全摒弃了传统小说的情节模式。

问题：文章始终在比较中展开论述，请你结合第⑥段的思路对此加以分析。

问题：同班的小徐同学难以区分“事件”“故事”“情节”。请你根据内容，用简明易懂的方式进行解说。

请你阅读下面的文字，完成后面的问题。

在异乡

[美]海明威

(1)秋天，战争不断进行着，但我们再也不去打仗了。米兰的深秋冷飕飕的，天黑得很早。转眼间华灯初上，沿街看看橱窗很惬意。店门外挂着许多野味：雪花洒在狐狸的卷毛上，寒风吹起蓬松的尾巴；掏空内脏的僵硬的鹿沉甸甸地吊着；一串串小鸟在风中飘摇，羽毛翻舞着。这是一个很冷的秋天，风从山岗上吹来。

(2)医生走到我的手术椅旁说：“战前，你最喜欢什么？玩球吗？”

(3)“不错，踢足球。”我说。

(4)“好。”他说，“你会重新踢足球的，肯定比以前踢得更好。”

(5)我的膝关节有病，医疗器能使膝关节弯曲得像骑三轮自行车那样灵活。可是眼下还不能弯，医疗器转到膝关节时便倾斜，不灵了。医生说：“一切都会顺利的。小伙子，你是个幸运儿。你会重新踢足球的，像个锦标选手。”

(6)旁边的手术椅中坐着一位少校。他的一只手小得像个娃娃的手。上下翻动的牵引带夹着那只小手，拍打着僵硬的手指。轮到检查他时，少校对我眨眨眼，一面问医生：“我也能重新踢足球吗，主任大夫？”他的剑术非常高超，战前是意大利最优秀的剑术家。

(7)医生回到后面的诊所里，拿来一张照片，上面拍着一只萎缩的手，几乎同少校的一样小，那是整形之前照的，经过治疗后就显得大一点了。少校用一只好手拿着照片，十分仔细地瞧着，问道：“是枪伤吗？”

(8)“工伤。”医生回答。

(9)“很有意思，很有意思。”少校说着便把照片递还给医生。

(10)“你该有信心了吧？”

(11)“不！”少校答道。

(12)每天，还有三个同我年龄相仿的小伙子到医院来。一个想当律师，一个要做画家，另一个立志当兵。有时，一天的疗程完毕，我们一起步行回去，到斯卡拉(米兰著名的歌剧院)隔壁的柯华咖啡馆去。因为四人结伴同行，就敢于抄捷径，经过共产党人聚居区。那里的人恨我们这些军官。我们走过时，一家酒店里有人喊叫：“打倒军官！”我们谁都不知道战事将如何发展，只知道仗还在打，一直在打，不过，我们再也不用上前线了。

(13)我可以肯定，少校不相信机械治疗，可他总是按时上医院，从不错过一天。在一段时间内，我们谁都不信这玩艺儿。那时，那种医疗器刚问世，我们正好去做试验品。少校长得矮小，却笔挺地坐在手术椅中，将右手伸入机器，让牵引带夹着手指翻动，眼睛直盯着墙壁。

(14)“要是战争结束了，要是真有那么一天的话，你打算干些什么？”少校问我。

(15)“回美国。”

(16)“结婚了吗？”

(17)“没有，但很想。”

(18)“你太蠢了。”他看上去很恼火，“一个男人决不能结婚。”

(19)“为什么，少校先生？”

(20)“别叫我少校先生。”

(21)“为什么男人不应该结婚？”

(22)“不该，就是不该！”他怒气冲冲地说，“即便一个人注定要失去一切，至少不该使自己落到要失掉那一切的地步。他不该使自己陷入那种境地。他应当去找不会丧失的东西。”

(23)他说着，眼睛直瞪着前面，显得非常恼怒、痛苦。

(24)“可为什么一定会失掉呢？”

(25)“肯定会失掉。”他望着墙壁说，然后，低下头看着整形器，哇吱咯咯地把小手从牵引带里抽出来，在大腿上狠狠拍几下。“肯定会失掉！”他几乎大吼了，“别跟我争辩！”接着他对看管机器的护理员叫道：“来，把这该死的东西关掉！”

(26)他回到另一间诊室去接受光疗和按摩了。后来他重新回到这间房间时，我正坐在另一只手术椅中。他披着斗篷，戴着帽子，径直朝我坐的地方走来，把一条胳膊搁在我的肩上。“真对不起，”他说，一面用那只好手拍拍我的肩胯，“刚才我太失礼了。我妻子刚去世。请原谅。”

(27)“噢……”我惋惜地说，“非常遗憾。”

(28)他站在那儿，咬着下嘴唇。“忘掉痛苦，”他说，“难哪！”

(29)他的目光越过我，望着窗外。接着他哭了。“我简直忘不掉悲痛。”他边说边哽咽着。然后他失声痛哭，又抬起头，茫然呆视着，咬紧嘴唇，泪流满面，接着，挺起腰，带着军人的姿态，迈过一排排手术椅，昂然而去。

(30)医生告诉我，少校的妻子很年轻，死于肺炎；少校直到残废不能再打仗后，才同她结婚。她只病了几天。谁也没料到她会死的。她过世后三天内，少校没上医院。之后，当他照常来就诊时，军服的袖子上多了一块黑纱。那时，医院的墙上已经挂起镶着大镜框的照片，拍着各种病例在治疗前后的不同形状。有类似少校的病例，但已整形，完全是正常的手了。不过，少校对那些照片却很淡漠。他只是向着窗外，凝望着。

问题：请你依据文本中的相关内容，认为下面的描述不正确的是哪一项？

A. 小说讲述的是第一次世界大战期间，从前线撤退下来的伤残军人在意大利米兰一家医院进行器械康复治疗时发生的故事。

B. 小说中的“我”是一个无名的美国士兵。采用第一人称叙事，可以将“我”的所见所闻、所感所想直接展现在读者面前，拉近了故事人物和读者的距离，也显得真实可信。

C. 文中画横线的两处都表达了战争还在继续而“我们”再也不用参战了的意思，说明了“我们”伤残严重，同时也包含着不能再上战场杀敌的遗憾。

D. 小说用近乎白描的手法对人物和事件进行了叙述，叙述者似乎是不加修饰和改变地把人物对话原原本本记录下来，让人几乎感觉不到叙述中介的存在。

问题：以下描述是对小说艺术特色的分析鉴赏，你认为不恰当的一项是？

A. 第一段景色虽然寒冷，但店门外挂着许多野味，使景色充满着新奇有趣的地方风情，让刚脱离战争的我们心情愉悦。

B. 第(12)段“在打” “一直在打”用反复的手法强调突出战争仍在继续，而“再也不用上前线了”则体现了他们离开前线如释重负的心情。

C. 本文对话简洁凝练，用了大量的短句对对话场景做了直观描摹，不但加快了行文的节奏，而且产生了强烈的视觉真实性，增强了故事的真实感。

D. 小说在上校的眼神中戛然而止，言有尽而意无穷，引导读者透过上校的眼神思考上校的精神世界，使作品的主题含蓄隽永。

问题：海明威的小说在简洁的叙述中蕴含着丰富的内涵，请结合文章中(6)-(11)段的内容，分析少校说的“很有意思，很有意思”这句话的丰富意蕴。

问题：如何理解小说的题目“在异乡”？请你从作品角度和读者角度分析说明。

第二部分测试任务

请你阅读下面的文言文，完成下面的小题。

文本一：

惠子谓庄子曰："魏王贻我大瓠之种，我树之成而实五石。以盛水浆，其坚不能自举也。剖之以为瓢，则瓠落无所容。非不呺然大也，吾为其无用而掊之。"庄子曰："夫子固拙于用大矣。宋人有善为不龟手之药者，世世以洴澼絖为事。客闻之，请买其方百金。聚族而谋之曰：'我世世为洴澼絖，不过数金。今一朝而鬻技百金，请与之。'客得之，以说吴王。越有难，吴王使之将。冬，与越人水战，大败越人。裂地而封之。能不龟手一也，或以封，或不免于洴澼絖，则所用之异也。今子有五石之瓠，何不虑以为大樽而浮乎江湖，而忧其瓠落无所容？则夫子犹有蓬之心也夫！"

——(选自《庄子》)

文本二：

亡国之主不可以直言。不可以直言则过无道闻而善无自至矣无自至则壅秦缪公时，戎强大。秦缪公遗之女乐二八与良宰焉。戎王大喜，以其故数饮食，日夜不休。左右有言秦寇之至者，因抒弓而射之。秦寇果至，戎王醉而卧于樽下，卒生缚而禽之。齐攻宋，宋王使人候齐寇之所至。使者还，曰："齐寇近矣，国人恐矣。"左右皆谓宋王曰："以宋之强，齐兵之弱，恶能如此？"宋王因怒而诎杀之，又使人往视齐寇，使者报如前，宋王又怒拙杀之，如此者三，其后又使人往视，使者遇其兄曰："国危甚矣，若将安适？"其弟曰："为王视齐寇。不意其近而国人恐如此也。今又私患乡之先视齐寇者，皆以寇之近也报而死；今也报其情，死，不报其情，又恐死。将若何？"其兄曰："如报其情，有且先夫死者死。"于是报于王曰："殊不知齐寇之所在，国人甚安。"王大喜。左右皆曰："乡之死者宜矣。"王多赐之金。寇至，王自投车上，驰而走。

——(节选自《吕氏春秋·雍塞》)

问题：下列 4 和选项中，不含通假字的一项是哪一个？

A. 魏王贻我大瓠之种　　B. 宋人有善为不龟手之药者

C. 世世以洴澼为事　　D. 客得之，以说吴王

问题：下面 4 个选项中对文中加点的词语及相关内容的解说，你认为不正确的一项是哪一个？

A. "树"是种植之意，与"五亩之宅，树之以桑"(《齐桓晋文之事》)中"树"的意思相同。

B. "与越人水战"中的"水"名词作状语，与"非能水也"(《劝学》)中的"水"字的用法不同。

C. "蓬之心"指像蓬草一样的见识、不通达之心，与现代汉语成语"孤陋寡闻"意相同。

D. "若将安适"意为"你将到哪里去"，与"卮酒安足辞"(司马迁《鸿门宴》)的句式相同。

问题：下面 4 个选项中对原文有关内容的概括和分析，你认为不正确的一项是哪一个？

A. 惠子种出了能够容得下五石的大葫芦，但因为盛不了足够多的水，做成瓢又没有用，所以就把它打破了。

B. 宋国人有制作防止手冻裂的药物的药方，但仅百金就卖掉了，庄子在文中叙述此事，旨在批评惠子与这个宋国人见识短浅。

C. 戎王没有识破秦缪公赠送歌舞队和厨师的阴谋，日夜不停地吃喝玩乐，若有谁提醒他有秦军进犯的危险，他还惩罚谁，这导致其最终被活捉。

D. 最后那个察看齐寇的使者，想到报告实情会被处死，不报告实情也会立刻被处死，所以不知该怎么办，后来在其哥哥的启发之下谎报了军情。

问题：请把下面的句子翻译成现代汉语。

(1)何不虑以为大樽而浮乎江湖，而忧其瓠落无所容？

(2)秦寇果至，戎王醉而卧于樽下，卒生缚而禽之。

请你继续阅读下面的唐诗，完成后面的题目。

送韦书记赴安西

杜甫

夫子欻通贵，云泥相望悬。白头无藉在，朱绂有哀怜。

书记赴三捷，公车留二年。欲浮江海去，此别意茫然。

问题：请你根据对这首诗的理解和赏析，对下面的四个选项做出判断，你认为下面的四个选项不正确的一项是哪一个？

A. 诗人认为，韦书记如今赴西安就任，通达显贵，自己与他地位悬殊。

B. 颔联写韦书记年事已高，他一度没有机会入仕，如今得到朝廷重用。

C. “三捷”与“两年”形成对比，表现了韦书记和诗人的不同境遇。

D. 这首诗文辞深沉蕴藉，情感有起有伏，体现出沉郁顿挫的风格。

问题：你认为诗歌表达了诗人哪些思想情感？请你简要概括出来。

问题：请你写出下列句子中的空缺部分的诗句。

(1)《老子》中，老子认为上善的人好像水一样，最接近于“道”，这是因为“_____，_____”

(2)《中庸(节选)》中，表现打破砂锅问到底的精神的句子是“_____，_____，_____”。

(3)张若虚的《春江花月夜》中“_____，_____”两句暗含鱼雁不能传讯之意，两人音讯断绝，相思无着落。

(4)《诗经·无衣》中的“_____”描写战士修整甲胄和兵器，然后又以“_____”一句，表明战士们奔赴前线、共同杀敌的英雄气概。

第三部分测试任务

请你阅读下面的文字，完成小题。

“民间”是非遗传承的底线。据非遗有关文件规定，当代非遗传承人要坚持首先在“民间”传承非遗的职责。守护住“民间”这条正道和底线，才能进行传承与创新。舍弃“民间”的所谓“创新”，是一种①_____的错误做法。运用民间思维，立足乡间水土，在传承中创新，在创新后传承。历代剪纸艺人都是沿着这条路行走的。潮州剪纸老师傅常说：剪一张纸，是小孩子都会的事，也很容易学会基本剪法；只有通过相当时间练习，才能熟练掌握；倘若要进一步精熟传承创作，则需要“技”与“艺”的巧妙结合和想象力的无边无界。过去老奶奶剪“花”，各式各样，信手剪来，栩栩如生，就是因为她们已将技与艺融为一体，化入血液。在千百年来的传承中，艺人们不断修改补充，②_____，把技艺锤炼得越发精绝超凡。流传至今的经典图谱，多随物象而剪饰。可以说，凡是能用作祭拜的水果供品，艺人们都可以创造出适合它的水果供品花。既有中秋拜月用的，也有时年八节、敬神拜祖用的，还有婚嫁习俗中用的，可谓③_____，花样奇绝。这些供品花艺术构思都源于生活，非常精妙。其造型简洁灵活，都是按照随物赋形的创作手法进行布局结构的，是历代潮阳妇女们的集体守正创新的成果。

问题：请你在文中横线处填入恰当的成语。

问题：请你对下面的病句请进行修改，使语言表达准确流畅，可少量增删词语，不得改变原意。

请你继续阅读下面的文字，完成下面小题。

不得不说，新型污染物——微塑料是近年来受到广泛关注的新型污染物之一。微塑料是指直径小于 5 毫米的塑料颗粒，它们可能来自于塑料废弃物的分解、洗涤剂中的微塑料颗粒、工业排放等。①_____，但它们对环境和生态系统的影响却可能是深远的。研究显示，如今，②_____。譬如，公民科学家在北极沿岸发现了大量的零碎微塑料；国际知名期刊报道，据调查，近 80%的瓶装水含有微塑料；研究者甚至在接受心脏手术患者的心脏组织中发现了微塑料。铺天盖地的骇人听闻的报道，让我们不得不关注微塑料这一议题并采取“全球战塑”行动，为生态，为地球，也为了人类自身。

长期以来，中国生物多样性保护与绿色发展基金会一直致力于海洋与湿地保护，③_____。然而，受到人类活动和全球气候变化等因素的影响，红树林生态系统正面临着严重的挑战，如面积缩减、生物多样性丧失和生态功能退化。为应对这一挑战，中国绿发会成立了红树林专项基金，设立了许多传播平台，就滨海湿地方面的法律法规积极建言献策。

问题：你认为下列各句中的引号，和铺天盖地的骇人听闻的报道，让我们不得不关注微塑料这一议题并采取“全球战塑”行动，这句话的引号作用相同的是哪一个？

A.“天若有情天亦老，人间正道是沧桑。”在苦难深重的旧中国，中国共产党领导人民进行了艰苦卓绝的斗争。

B.包身工的身体是属于带工老板的，所以她们根本就没有“做”或者“不做”的自由。

C.刚刚参加了交接仪式的查尔斯王子和第 28 任港督彭定康登上“不列颠尼亚”号的甲板。

D.没有想到，百团大战中这个小小的“插曲”，四十年后，竟成了中日人民友好的佳话。

问题：请在补写下面的语句，使整段文字语意完整连贯，内容贴切，每处不超过 12 个字。

问题：请根据文章信息，给“微塑料颗粒”下定义。

第四部分测试任务

在二〇二二年的新年贺词中，中国国家领导人引用了“致广大而尽精微”这句话，它出自《中庸》，意思是能通达至广大之境，又能极尽精微之处。“致广大”，可理解为要有远略，要有大局观，要有大格局；“尽精微”，可理解为要注意细节，要从小事做起，要有格物致知的态度。二者看似矛盾，实则是辩证统一的。这句古语对我们当代青年在求学、做人、做事等方面都有启示。

请你选择最有感悟的一方面，结合青少年成长与发展撰写一篇作文。输出的作文要求：选准角度，确定立意，明确文体，自拟标题；不要套作，不得抄袭；不得泄露个人信息；不少于 800 字。