心理学报 2026, Vol. 58, No. 2, 308−322
Acta Psychologica Sinica
https://doi.org/10.3724/SP.J.1041.2026.0308
https://cstr.cn/32110.14.2026.0308

“零样本语言学习”：大语言模型能“像人一样”习得语境中的情感吗?

吴诗玉 王亦赟

(上海交通大学外国语学院, 上海 200240)

摘 要 本研究旨在检验大语言模型(LLMs)能否在“零样本”条件下通过阅读附带习得单词所出现的语境情感, 并评估情感效价与语境变异性对词汇学习的影响。研究采用跨模型−人类对比, 4 种 LLMs 与 3 组学习者在统一材料中学习嵌入不同情感(积极/中性/消极)与重复/变化语境的目标词, 并以多项测试衡量情感迁移及词形、词义习得效果。结果显示, LLMs 与人类模式一致, 能将语境情感迁移至目标词, 并在语言生成中保持情感一致; 而且也呈现“积极情感优势” “语境变异优势”, 且在定义生成中出现语境情感与语境变异的交互效应。文章提出“双重机制框架”, 认为 LLMs 在功能层面具备类人的情感语义学习能力, 但其机制基于统计共现与向量优化, 异于人类的具身与社会加工。本研究为情感计算、人机交互伦理与词汇教学提供启示。

关键词 大语言模型, 零样本学习, 情感学习

分类号 B842

1 引言

1.1 “零样本学习”和词汇附带习得

在人类认知系统中, 情感并非附属于理性之后的修饰成分, 而是一种结构性资源(Damasio, 1994)。神经科学与认知心理学研究已明确表明, 情感不仅参与注意调节与记忆巩固(Eysenck & Brysbaert, 2018), 更深度塑造价值判断、道德推理与社会行为选择(Barrett, 2017)。在语言、文化与互动中, 情感并非外在表达的“内容”, 而是驱动交际意图、建构共同视角、协调人际共识的内在机制; 它是一种流动的文化实践：它通过语义联结与语用惯例在社会互动中不断被赋义, 并塑造人类对世界的认知结构(Ahmed, 2004, 2010)。

对情感习得机制的系统研究, 是揭示人类心智独特性的关键路径。个体并非天生拥有完整的情绪判断力, 而是在交互中, 通过对语境、语言与反应的归纳逐步习得情感的识别与表达方式(Wetherell, 2012)。这一过程具有三个关键特征：语境依赖性、共情驱动性与互动可塑性。首先, 情感常隐含于语义−语用结构、非言语线索与语调中, 理解它需依赖复杂的语境建构能力(Louw, 1993; Nevisi et al., 2018)。其次, 情感迁移依赖于“脑对脑耦合” (Hasson et al., 2012)：个体的情绪信号可引发他人的神经共振, 推动跨个体的情绪同步(Ho et al., 2023)。此外, 情感表征并非静态结构, 而是随互动持续建构并更新的认知产物, 从婴儿模仿面部情绪开始, 直至成人在社会情境中对情绪意图的理解和调整(Tamir et al., 2016)。

可见, 情感的习得路径不仅刻画了人类如何“理解世界”, 也定义了我们如何“成为社会化的心智存在”。正是在这一意义上, 情感学习构成了理解人类心智不可替代的理论入口。而任何试图模拟人类智能的系统, 若无法触及这一机制, 将始终与类人认知存在本质距离(Binz & Schulz, 2023)。

当前的大语言模型(LLMs), 如 ChatGPT 和 Gemini, 已展现出卓越的语言生成、推理与问答能力, 让普通用户也能以自然语言与其互动, 从而推

收稿日期: 2025-01-27
吴诗玉, E-mail: shiyuw@sjtu.edu.cn

动了人工智能技术的普及和大众化。但是, 这类系统的强大还不仅体现在能够流畅地生成自然语言、回答复杂问题, 而在于它们所展现的“仿人特性”: LLMs 通过训练能在大规模语料中识别语言模式, 并将其内化于模型参数, 类似于人类通过积累经验, 形成语言知识网络。它们还能在新任务中调用这一内化知识进行推理, 显示出与人类类似的迁移能力。

然而, 决定人工智能能否迈向“类人性”的关键, 或许并非是语言生成的流畅或推理的准确, 而在于其是否具备像人一样习得与传播情感的能力(Ahmed, 2004, 2010; Barrett, 2017)。情感学习不仅关乎情绪信号的识别, 更在于将其迁移至语言使用中, 并在互动中建构情感共识(Hasson et al., 2012; Hatfield et al., 2009)。人类的这一机制建立在具身经验、语境归纳与互动反馈的共同驱动之上(Barsalou, 2008; Ellis & Wulff, 2015), 而当前的 LLMs 仅基于语言输入进行统计学习, 缺乏感官、动机与社会情境的真实嵌入(Bisk et al., 2020)。但值得注意的是, 近年来在快速演化中, LLMs 已开始展现某些令人“惊异”的新兴能力, 尤为引人关注的是所谓的“零样本学习” (Zero-shot Learning): 即在无任何监督数据的条件下, 通过纯粹的强化学习过程, 自主演化出具备推理能力的潜力(Wang et al., 2019)。这种能力在人类语言学习中颇为常见, 例如在毫无预备知识的情境下, 个体可凭借既有语言规则与逻辑推理理解生成全新概念, 产出从未听闻的句子(Chomsky, 1957)。

LLMs 的“零样本学习”为检验其情感学习和传播能力提供了重要机会。本研究将从基于阅读的词汇附带习得入手, 比较 LLMs 和人类学习者对语境情感的学习和传播, 这是因为 LLMs 的“零样本语言学习”跟人语言发展的关键机制——基于阅读的词汇附带习得存在显著相似之处。主要表现在两个层面：首先, 两者均体现出任务驱动下的非专注性学习特征：人类在语篇阅读中通过上下文推理词义(Hulstijn, 2001), 而 LLMs 则依赖输入语境与预训练知识完成生成, 无需显式监督(Brown et al., 2020)。其次, 两者均具备知识整合与迁移能力：人类通过反复接触将新词纳入心理词典, 而 LLMs 可将已习得语言模式泛化至全新语境, 实现类人的跨任务输出。这一机制类同性, 可以使词汇附带习得成为评估 LLMs 情感归纳能力的关键范式。

比较 LLMs 和人类学习者在词汇附带习得中的语境情感学习, 不仅有助于厘清人工智能系统是否具备“类人”的情感学习能力, 也能反向揭示人类词汇学习机制中若干核心变量的心理现实性。基于此, 本研究以统一的实验材料与任务结构为基础, 构建跨人类–模型的对比范式, 探究在“零样本”条件下, 语境中的情感是否能够被归纳、迁移, 并作用于词汇学习过程中的词义建构。下文将首先梳理语境情感学习与传播的相关文献, 介绍主要理论与实证成果, 进而指出现存争议焦点, 最终提出本研究的具体问题。

1.2　语境情感的学习和传播

关于语境情感的学习, 最早兴起于上世纪 80 年代以电脑技术为基础的语料库语言学, 后拓展至认知心理学与应用语言学领域, 先后提出了“传染说”、“迁移说”和“双枣树效应”假说。

“传染说”是语境情感学习领域最早的机制理论, 源于 Sinclair (1987)对某些词(如 *set in*)频繁出现在消极情感语境(如 *rot*、*decay*、*disillusion*)中的观察。Louw (1993)在此基础上提出“语义韵”概念, 指出词汇通过与特定情感词频繁共现而获得稳定的情感色彩。例如, *spark* 常与积极语境(如 *lively*、*justice*)共现, 而 *trigger* 则多与消极语境(如 *sudden*、*bankruptcy*、*crash*)关联。语义韵虽未直接体现于词义, 却对语义加工和情感判断产生深远影响, 这一微妙、近乎隐含的语义也已得到大量语料证据的证实(Nevisi et al., 2018)。然而, “传染说”主要基于对词典中已有且语义稳定词的共现统计, 因此无法解释新词初次接触时情感色彩的“传染”路径, 也难以揭示语境情感如何迁移至学习者的词汇表征中。

Snefjella 等人(2020)由此提出了语境情感学习的“迁移说”。他们通过变量操控和心理过程测量, 一共开展了 5 项实验。让英语母语者阅读多套承载积极、消极和中性情感的小短文, 并学习嵌入其中的目标假词(见 3.2 小节)。结果发现, 尽管所有的目标假词在短文中表达的都是中性语义, 但是如果被试总是在积极情感语境中碰到它们, 就会认为它们在情感上很积极; 相反, 如果总是在消极语境中碰到它们, 就会认为它们很消极。这个结果证实语境情感经阅读接触后成功地从语境迁移到学习者对目标词所形成的语义表征里, 并内化为其词汇知识的一个部分。然而, 尽管“迁移说”在理论与方法上都提供了新视角, 但对“迁移”的发生机制仍有待更精细的验证, 学习者特征、语境变异性及接触频次数等潜在调节因素有待进一步探明。

在此基础上, 吴诗玉和李赞(2024)提出了“双

枣树效应假说”, 他们以中文母语者为被试, 重复了 Snefjella 等人(2020)的实验, 但在实验中还操控了语境变异性这个变量(多篇 vs. 多遍, 见下文), 让每名被试要么阅读承载相同情感负荷的相同材料多遍, 要么阅读多篇承载相同情感负荷的不同材料, 再参加词汇测试。结果证实了语境情感的可迁移性, 但同时发现这种迁移是有条件的：只有重复阅读相同的材料多遍才会发生, 表现出鲁迅《秋夜》中“还有一株也是枣树”的“重复–感染”模式。

最近, 研究者(李赞, 2024; 马拯, 李赞, 2024)在二语语境, 借用 Snefjella 等人(2020)以及吴诗玉和李赞(2024)的实验范式, 检验二语学习者是否也可以有效习得二语目标词所出现的语境的情感。结果发现, 在二语阅读过程中, 语境情感也可以从语境向目标词迁移, 但是这种迁移同时受到语境变异性和学习者二语水平的调节。

概而言之, 现有研究已系统揭示人类学习者可在以阅读为基础的词汇附带习得过程中习得目标词所承载的语境情感, 即语义韵。然而, 情感学习的传播效果, 即情感学习是否进一步促进词汇本身的习得(包括词形与词义), 仍存在明显争议。一方面, 部分研究支持“积极情感优势” (positivity advantage)：与中性或消极语境相比, 积极语境更有利于词汇习得, 表现为更高的识别率、更强的语义整合和更持久的保持效果(Snefjella et al., 2020)。另一方面, 也有研究提出“消极偏倚” (negativity bias)：负面语境因更强的情绪唤醒与认知负荷, 在词汇保持中更具优势(Driver, 2022)。

另一个核心争议聚焦于“语境变异性”的作用(Bolger et al., 2008), 即“多遍”还是“多篇”哪个更有利于词汇学习的问题(吴诗玉, 李赞, 2024)。根据基于实例驱动的理论框架(Instance-based Framework), 词汇在多样语境中的反复出现可生成更多独立语义痕迹, 促进核心意义的抽象。Bolger 等人(2008)和 Lauro 等人(2020)的实验支持该观点, 发现“多篇”语境显著提升词义生成与语义整合表现。然而, 也有研究对变化语境优势提出了质疑。Horst 等人(2011)发现, 重复语境(“多遍”)反而更有助于词义的学习和保持; Joseph 和 Nation (2018)亦指出, 在学习低频词时, 语境多样性并未展现出显著优势; 此外, Balass (2011)在部分语义判断任务中甚至观察到重复语境优于变化语境的结果。

在此背景下, 本研究以“零样本词汇学习”作为实验框架, 比较 LLMs 和人类学习者在词汇附带习得过程中的情感学习, 除可以厘清人工智能系统是否具备“类人的”情感学习能力, 还可反向检验人类词汇习得机制中两个关键变量——情感效应与语境变异性的心理现实性。具体看, 本研究回答以下三个问题：

(1)大语言模型(LLMs)是否能“像人一样”习得语境中的情感?

(2)哪一种语境条件, 变化还是重复, 更能促进 LLMs 的词汇学习?

(3)语境情感是否对人和 LLMs 的词汇学习都造成显著影响?

2 方法

实验分两个阶段：阅读学习和词汇测试。需要说明的是, 在本研究中, “零样本语言学习”所指的并非 LLMs 在完全无输入的状态下完成任务, 而是指模型未接受任何关于测试任务或目标词项的特定监督或微调训练, 仅通过对原始阅读材料的语境处理, 在测试中进行即时输出判断, 符合当前“零样本”任务的广义定义(Brown et al., 2020)。

研究旨在比较 LLMs 与人类学习者的情感学习表现及其影响。人类数据来源于 3 组学习者：(1) Snefjella 等人(2020)实验中的英语母语者; (2)吴诗玉和李赞(2024)实验的汉语母语者; (3)作者近期实验的英语二语学习者(完整数据将在另文中报告)。3 组学习者使用的实验任务与结构保持一致, 所用材料在语言上略有差异：英语母语者与英语二语学习者使用 Snefjella 等人(2020)原始英文材料, 汉语母语者则使用经过严格语义等值翻译的中文版材料, 以确保结构和语境一致性。在词项构建方面, 英文实验使用由 Wuggy 生成的假词, 中文实验则依照相同的构造原则, 以阿拉伯语音译词为基础进行音形改造, 生成词形自然但无语义负载的汉语假词。为便于表述, 本文将上述 3 组实验统称为“基准实验”。

2.1 大语言模型

本研究选取了当前国内外具有代表性的 4 种大语言模型, 分别为文心一言(Ernie Bot 3.5)、ChatGPT (GPT-4)、Gemini (1.5 Pro)与 LLaMA (3.1-8B)。模型选择遵循以下原则(见表 1)：(1)覆盖当前主流技术架构, 包括标准 Transformer 架构(如 LLaMA、ChatGPT)与融合 MoE 机制(Mixture-of-Experts)的混合架构(如 Gemini); (2)体现多样的训练背景与语言适配性, 覆盖中英双语优化模型(如文心一言为

表 1　四种大语言模型简介

模型名称	架构及参数	开源情况	上下文长度(token)	多模态推理能力
百度：文心一言 3.5	Transformer/未公开	闭源	8K	支持文本、图像和视频生成。中等推理能力, 适合中文语境。
OpenAI: ChatGPT 4	Transformer/> 1.75 万亿	闭源	8K	强大的图像生成和数学、逻辑等推理能力, 支持图文结合。
Google DeepMind: Gemini 1.5 Pro	Transformer, MoE/未公开	闭源	2M	强大视觉理解和生成。复杂问题和知识查询推理能力强。
Meta: LLaMA 3.1-8B	Transformer/约 80 亿	开源	128K	文本生成, 缺乏多模态能力。中小规模任务, 基础推理能力。

中文优化模型, ChatGPT 与 Gemini 为英语主导模型); (3)具备不同规模与开源特征, 包括闭源模型与开源模型。这些模型不仅在架构设计、参数规模和上下文窗口长度方面差异显著, 也在对文本情感、语义生成和语境适应能力方面体现出互补性, 具备良好的横向比较价值。

需要说明的是, 这 4 种模型代表了不同的技术路线与语言能力优化方向, 并非在绝对性能上完全可比, 而是在架构、开放性、语言适配、推理能力等关键维度上提供了具有解释力的对照基础。因此, 本研究在统一输入格式、任务类型与输出评估的前提下, 对 4 种模型的表现进行了平行测试, 以探索不同模型在语境情感习得任务中的处理路径和行为模式。模型输出与 3 组人类学习者的表现共同构成 7 组可比数据(4 组 LLMs + 3 组学习者), 为检验语境情感学习与语义韵迁移机制的类人性及其边界提供了坚实基础。

2.2　材料

为方便比较, LLMs 的学习材料与第一、三组基准实验完全一样, 包括：阅读段落、嵌入在段落里的目标词以及词汇知识测试题。详见 Snefjella 等人(2020)以及吴诗玉和李赞(2024)。

2.2.1　目标词

一共考察 9 个目标词的学习, 都是 Snefjella 等人(2020)使用 Wuggy 软件(Keuleers & Brysbaert, 2010)生成的 9 个假词, 分别是：aunith, ceammy, cruce, flyph, mernt, neak, plurk, rotch 和 wurge。这些假词在拼写和读音上与英语真词相似, 符合英语正字法和拼音规则。

每个目标假词再匹配 5 个假词：一个为同音假词(如 crooce 对应 cruce), 用作单词正字法选择任务的干扰项, 另外 4 个非同音假词, 用作另外 4 道词汇知识测试的干扰项(见下文), 一共生成了 45 个干扰项假词。假词都经过了标准化校正, 确保目标词及其填充假词的情感效价不存在显著差异(详见 Snefjella et al., 2020; 吴诗玉, 李赞, 2024)。

2.2.2　阅读材料

阅读材料以三联体(a triplet)为单位设计, 目标词都嵌入到这些三联体里, 每个三联体包含 3 段不同的短文, 如例(1)：

(1) a. Her wonderful teacher performed at the exciting concert.

She was playing the **(NONWORD).**

b. Her new teacher performed at the normal concert.

She was playing the **(NONWORD).**

c. Her annoying teacher performed at the horrible concert.

She was playing the **(NONWORD).**

三联体中的短文都包含两句话, 第一句通过形容词和名词来操控语境的情感, 分为积极、中性和消极三个水平, 同一个三联体的第二句话则完全相同, 包含要学习的目标词。第一和第二句话所构成的语境能确保学习者推断出每个目标词的词义, 即外延义(denotation)。所有的目标词表达的外延意义都是中性情感, 且是日常生活中常见的物品(大语义范畴), 如一种乐器、植物和厨房用具, 比如例(1)表示的是一种乐器。一共有 9 个外延义。

每个目标词分布(嵌入)在 5 个不同的三联体中, 因此, 一共设计了 45 组三联体(9 × 5), 共 135 篇短文(45 × 3)。每名学习者在实验中一共需要阅读 45 篇短文, 每个目标词在整个实验中出现了 5 次, 这可以确保学习者通过阅读习得该目标词(如 Blythe et al., 2012; Godfroid et al., 2018)。阅读材料均按 Snefjella 等人(2020)事先设计的模板生成, 以对一些潜在的干扰变量进行控制, 比如语境长度和目标词出现的频率。

语境情感(积极、中性和消极)为被试内变量, 采用拉丁方块, 根据 3 个语境情感水平和目标词的 9 种外延义对上述 135 篇阅读材料(45 组三联体)进行交叉平衡, 一共生成了 9 套材料。在每套材料中, 每个目标词均出现在 5 个不同的篇章, 并且始终出现在同一种语境情感且表达同一种外延义。通过 9

套材料, 确保了每个目标词出现在了每一种语境情感, 表达了每一种外延义, 实现了全面且均衡的覆盖。另外, 为检验语境变异假说, 实验还操控了语境变异性(多篇 vs. 多遍)这个变量。但是, 把它设计为被试间变量, 即对同一个目标词要么让实验对象阅读相同的语篇 5 遍, 要么阅读 5 篇不同的语篇, 因此, 最终一共生成了 18 套实验材料(9 × 2)。该设计旨在避免因任务量过大而导致的注意力分散和记忆负荷上升, 同时保持与前期基准研究一致, 以增强可比性; 在统计分析中引入混合效应模型, 以控制由个体差异引发的变异。

2.2.3 词汇测试题

一共设计了 5 种词汇知识测试任务, 以从不同维度来全面评估 LLMs 对目标词的学习效果, 分别为: 情感效价评分任务、句子产出任务、正字法选择任务、定义匹配任务和定义生成任务。除句子产出任务以外, 其他 4 项任务跟基准实验完全一样。

情感效价评分任务, 用于测试 LLMs 对语境情感的学习效果。一共 18 个测试词, 包括 9 个目标词(如 *flyph*)和 9 个填充词(如 *snicle*)。在测试时, 按伪随机顺序依次呈现, LLMs 首先需要判断该词是否曾在前面的阅读中见过, 如果为“否”, 直接跳到下一词, 如果为“是”, 则对该词进行情感评分(1 表示“悲伤”, 9 表示“愉快”)。

句子产出任务也用来测试 LLMs 对语境情感的学习效果。在此任务中, 一共有 18 个词按伪随机顺序依次呈现, LLMs 首先需要判断该词是否曾在前面的阅读中见过, 如果为“否”, 直接跳到下一词, 如果为“是”, 则使用该词生成一个不超过 30 个词的句子。其中, 9 个为目标词, 另外 9 个为填充词。这一测试任务是相较于基准实验的增设任务, 之所以增加这项测试任务, 是因为语言生成是 LLMs 的核心功能之一。通过句子产出, 可以在更贴近真实语言使用场景的情况下, 评估 LLMs 对语境情感的学习效果, 更好地反映模型在实际应用中的表现。

正字法选择任务, 用于测量 LLMs 对目标词的词形知识的掌握。在此测试中, LLMs 一共需判断 18 个词是否曾在阅读中出现过, 9 个为目标词(如 *flyph*), 另外 9 个为与填充假词(如 *fliph*)。通过评估 LLMs 区分目标词与填充词(同音词)的能力, 来测量其对目标词词形的记忆水平。

定义匹配任务, 用于评估 LLMs 对目标词的词义识别(meaning recognition)。在此任务中, LLMs 需在给定的定义中选择与单词匹配的选项, 一共有 18 个测试词, 9 个为目标词, 9 个为填充假词。定义生成任务, 用于测试 LLMs 对目标词的词义回忆(meaning recall)。在此任务中, 一共有 18 个测试词, 9 个为目标词, 9 个为填充假词, LLMs 首先需要判定是否在前面的阅读中碰到过该词, 如果答案为否则直接跳到下一页, 如果答案为是则生成其定义, 包括中文翻译、英文释义或同义词。

尽管定义匹配和定义生成任务都测试了 LLMs 对目标词的词义学习效果, 但是, 定义生成需要从记忆中提取词义, 相比定义匹配中的语义识别(从给定的选项中选择与单词匹配的定义), 它的难度更高, 考察了更高级的词汇语义知识(Laufer & Aviad-Levitzky, 2017; Stewart et al., 2024)。

2.3 程序

本研究以第三项基准实验(英语二语学习者)的数据为蓝本, 按被试一一对应地进行 LLMs 的人工模拟与数据采集。该基准实验共纳入 306 名中国英语学习者, 每名被试阅读 45 篇短文并完成 4 种词汇测试。据此, 我们为每一名人类被试构建一条对应的 LLM 测试流程: 每个独立测试会话(session)被视为一个唯一观测单元, 功能上等同于人类实验中的“一名被试”。因此, 每种大语言模型(共 4 种)均独立完成 306 次会话, 总计生成 1224 名“模型被试”。每一次会话中, 阅读材料的呈现顺序严格复现其对应的人类被试顺序, 既确保了与人类数据的可比性, 也保留了材料呈现的随机性。

为有效保障会话间的独立性, 本研究采用严格的会话隔离规范: 每个会话均在全新对话窗口中启动, 模型在其中依次完成 45 篇短文的阅读及随后 5 项测试(包括句子产出任务); 一旦该会话结束, 立即彻底清空对话上下文与历史缓存, 并重新开启新会话, 确保不会携带任何先前会话的信息。此程序有效防止跨会话“信息污染”, 满足混合效应模型(见下文)对观测独立性的核心假设, 并保证不同 LLMs 的结果具备可比性与可解释性。所有测试均在零样本条件下完成, 提示语(见附件)仅用于控制任务流程与回答格式, 不提供任何示例、线索或训练性输入, 以避免信息泄露并确保结果可复现与可核查。

在任务流程上, 当 LLMs 在每一个测试会话中完成 45 篇短文的阅读(代表 1 名被试)后, 立即接受 5 项词汇测试, 顺序为: 正字法选择任务、情感效价评分任务、定义生成任务、定义匹配任务和句子产出任务。在整个过程中, 模型均未接收任何示例或反馈, 确保所有回答严格基于零样本条件生成。

3　结果

3.1　语境情感的习得

从单词情感效价评分任务和句子产出任务来报告 LLMs 对语境情感的习得效果。由于没有英语母语者的原始数据, 因此, 只概要总结 Snefjella 等人(2020)所报告的结果。而对中文母语者和二语学习者, 为方便比较, 把他们的数据跟 LLMs 的数据合并分析。

3.1.1　情感效价评分

Snefjella 等人发现(2020), 英语母语者对出现在积极语境情感中的目标词的情感效价评分总是最高, 而对在消极语境情感中的目标词的评分则总是最低。统计推断的结果进一步证实英语母语者在 3 种语境情感下(消极 vs. 中性 vs. 积极)所作评分存在显著差别, 证实了语境的情感成功地迁移到目标词。

图 1 呈现了两组学习者(中文母语者、二语学习者)以及各种不同的大语言模型在目标词学习时的语境情感学习的结果。

从图 1 可以看到, 两组学习者之间情感效价评分的表现在总体模式上非常接近, 更重要的是, 各个大语言模型, 不管是文心一言, 还是 ChatGPT、Gemini 或 LLaMA 的情感效价评分在总体模式上也都与这两组学习者非常相似：积极语境的情感效价评分总是最高, 而消极语境下的评分则总是最低。

使用 R 语言的 ordinal 包的 *clmm* ()函数, 对两组学习者以及各个大语言模型的情感效价评分(有序数据)拟合累积链接混合效应模型(Christensen, 2023)。模型的固定效应因素有两个：语境情感(消极、中性、积极)和学习者类型(中文母语者、二语学习者以及各个大语言模型)。随机效应因素有 3 个：被试(或模型模拟的被试)、词汇测试项和目标词对应的外延意义。模型拟合的结果显示, 不管哪种学习者类型, 即不管是两组学习者(母语和二语), 还是各个不同的大语言模型, 对出现在 3 种不同语境情感下的目标词所作情感效价评分都存在显著差异($zs < -9.00$, $ps < 0.0001$), 表现为：积极语境 > 中性语境 > 消极语境。正如前面所介绍的, 由于这些词本身为假词, 不表示任何意义, 在嵌入到各个语篇之后所表达的也都是中性的、具体的、大语义范畴(如一艘船、一件厨房用具), 因此, 在阅读这些语篇之后, 实验对象所作情感效价评分存在差异的来源只可能来自于语境, 即他们从单词出现的语境中学会了单词的情感。

3.1.2　句子产出的情感分析

各个语言模型产出的句子都达到了很高的准确率, 除文心一言为 89.10%以外, 其他都超过 90%。使用 Python 的自然语言处理工具 TextBlob 对大语言模型产出正确的句子进行情感分析。TextBlob 会为每个句子生成情感得分, 范围从−1 至 1。负值表示负面情感, 正值则表示正面情感, 得分越大, 情感的积极程度越强。图 2 呈现了分析的结果。

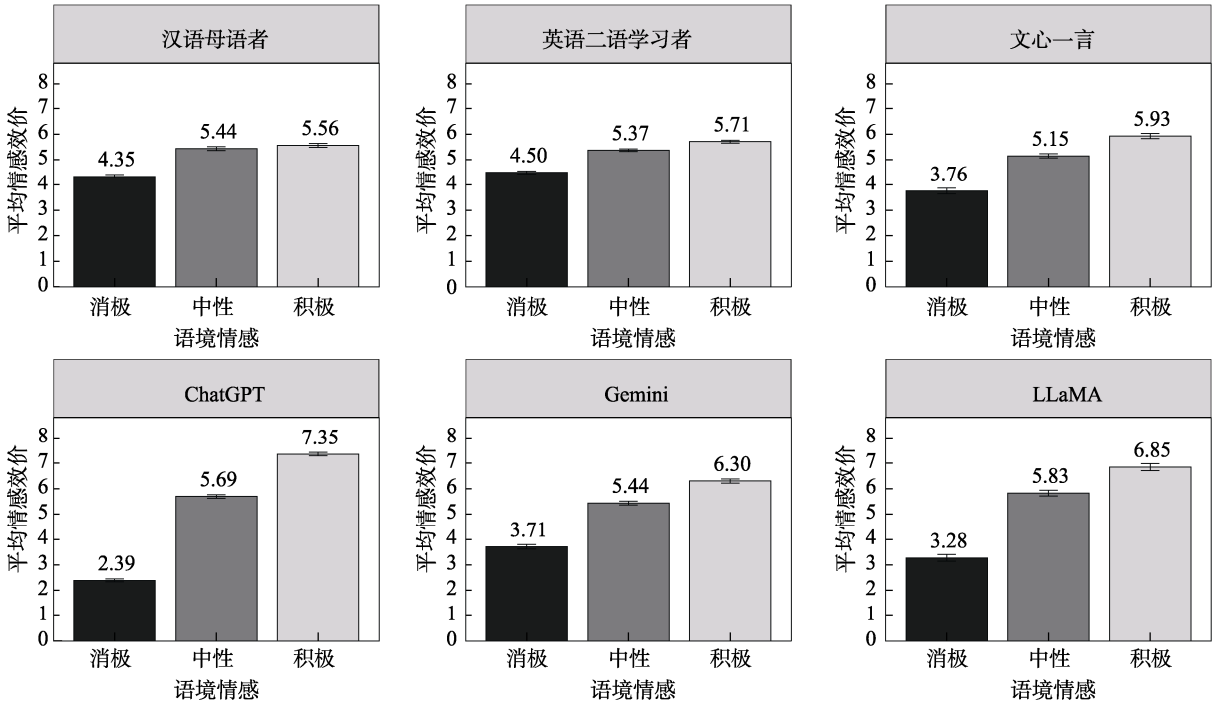

图 1　语言学习者和 LLMs 的情感学习结果

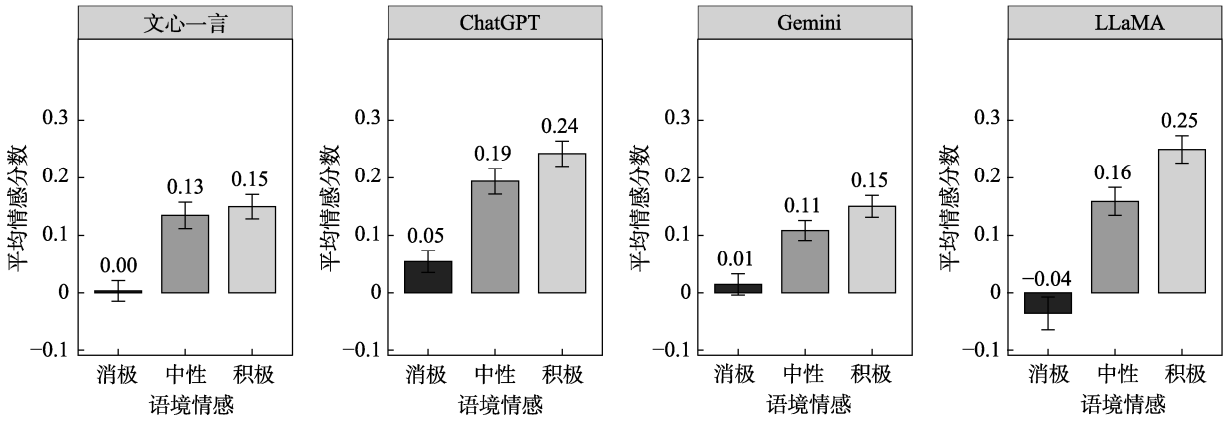

图 2 LLMs 所产出的句子的情感分析的结果

总体模式上图 2 所展示的结果跟图 1 非常接近, LLMs 使用在积极语境下的目标词造句时, 所造的句子的情感值总是最高, 而使用在消极语境下的目标词所造的句子的情感值则总是最低。更具体看, LLaMA 的情感学习看起来效果最为明显, 其次为 ChatGPT, 再次是 Gemini, 最后是文心一言(Ernie Bot)。

使用 R 语言的 lme4 包的 *lmer* ()函数, 对各个大语言模型所产出句子的情感值拟合混合效应模型(Baayen et al., 2008)。模型的固定效应因素有两个：语境情感(消极、中性、积极)和 LLMs 类型。随机效应因素有 3 个：模型模拟的被试、词汇测试项和目标词对应的外延意义。模型拟合的结果显示, 除了文心一言以外, 其它 3 种大语言模型在 3 种语境情感之下所产出的句子的情感值都存在显著差异($zs < -2.14$, $ps < 0.05$), 表现为：积极语境 > 中性语境 > 消极语境。就文心一言来说, 尽管积极和中性之间不存在显著区别($b = -0.004$, $SE = 0.01$, $z = -0.29$, $p = 0.95$), 但是中性和消极之间以及积极和消极之间差别显著($zs < -7.10$, $ps < 0.0001$)。

这些结果说明, 尽管各个大语言模型之间存在个体差异, 但是, 从语言产出的角度看也都成功地习得了单词出现的语境情感, 即语义韵。

3.2 词形和词义的习得

由于 Snefjella 等人(2020)的实验没有考察语境变异性的影响, 因此, 此处只比较第二、三项基准实验的两组学习者(中文母语者、二语学习者)和 LLMs 的表现。词形和词义的习得通过 3 种测试任务来测量, 其中正字法选择任务测试词形知识, 而单词定义匹配任务和单词定义生成任务测试词义知识。所有测试任务的答案都按“正确”和“错误”编码, 比如语言学习者或 LLMs 选择了正确的词形或者为目标词正确地匹配了定义或者提供了正确的单词定义, 答案编码为“正确”, 反之编码为“错误”。

运用 lme4 包中的 *glmer* ()函数, 使用逻辑回归的混合效应模型来拟合学习者以及 LLMs 在这 3 个测试任务的成绩。模型有两个自变量：语境(重复 vs. 变化)和语境的情感(消极 vs. 中性 vs. 积极)。但语境情感并不是当作分类变量进入模型, 而是把语言学习者或 LLMs 对每个目标词所作的情感效价评分进入模型, 这样可以更加直观地评估语境情感(学习效果)的影响(参见 Snefjella et al., 2020)。在把这个变量进入模型前先作标准化处理, 以方便模型的拟合及结果的解释。模型的随机效应因素有 3 个：被试(学习者, 或模型模拟的学习者)、词汇测试项和目标词对应的外延意义。

3.2.1 正字法选择任务

汉语母语学习者和二语学习者的平均准确率分别为 89.17%和 82.65%。但是大语言模型除 LLaMA 低于 1 为 98.8%以外, 其他的准确率都达到 1, 展现了 LLMs 更强的词形学习能力。

3.2.2 单词定义匹配任务

图 3 呈现了语言学习者和各个大语言模型在不同语境条件下(重复 vs. 变化)完成单词定义匹配任务时的平均准确率。

从图 3 看, LLMs 的学习效果总体上好于语言学习者, 混合模型的结果也给予了证实($b = 1.88$, $SE = 0.07$, $z = 24.45$, $p < 0.001$)。但是, LLMs 又与学习者表现出一种非常一致的模式：“多篇” (变化)语境下的学习效果都要好于“多遍” (重复)语境, 也获得了混合效应模型结果的证实($b = 1.13$, $SE = 0.21$, $z = 5.720$, $p < 0.001$)。

考虑到上述变量影响的一致性, 为方便比较, 首先把两组学习者的数据以及各个大语言模型的数据分别合并, 再构建混合效应模型, 进行统计分

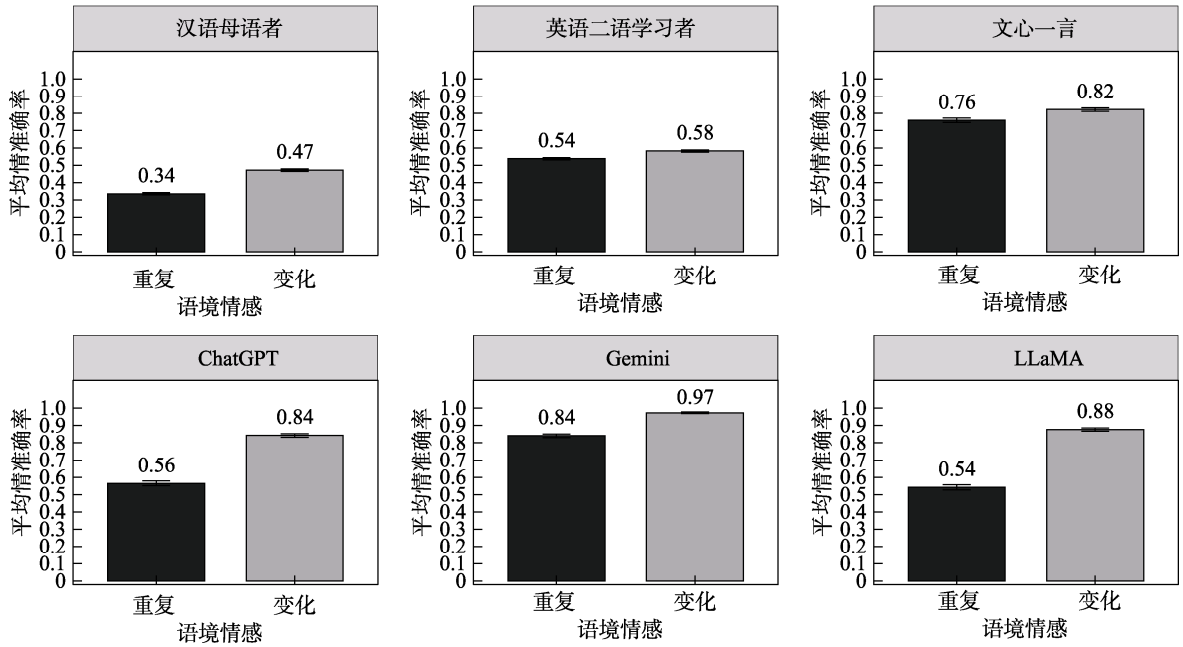

图 3　语言学习者和 LLMs 在重复和变化语境下完成单词定义匹配任务的平均准确率

析[1]。结果发现, 语境变异性和语境情感效价对两组实验对象产生了非常相似的影响：(1)积极情感优势, 即：随着语境情感效价的提高, 两组学习者和各个大语言模型的词义识别能力以近乎接近的幅度(0.21 vs. 0.23)显著增强(zs > 7.64; *p*s < 0.001)。(2)变化语境优势, 即：不管是对两组学习者还是各个大语言模型, 变化语境都比重复语境产生了显著更好的学习效果(学习者：$b = 0.71$, $SE = 0.28$, $z = 2.56$, $p = 0.01$; 大语言模型：$b = 2.02$, $SE = 0.25$, $z = 7.94$, $p < 0.001$)。(3)不管是对两组学习者还是各个大语言模型, 语境情感效价和语境的变异性都不存在交互效应(*p*s > 0.15), 即语境情感效价的影响不取决于语境变异性的不同水平。

需要指出的是, 尽管总体趋势上 LLMs 与两组学习者高度相似, 但是两者也表现出了一些具体的差别。比如, LLMs 总体的学习效果要好于两组学习者(准确率显著更高, 图 3)。另外, 尽管都表现出变化语境优势, 但是, LLMs 的优势要比学习者更为显著(效应量：2.02 vs. 0.71)。就各个模型的具体表现看, Gemini 的词义识别显著好于其他各个模型(*p*s < 0.001), 而其他各个模型之间没有显著区别(*p*s > 0.1)。

3.2.3　单词定义生成任务

图 4 呈现了两组学习者和各个大语言模型在两种不同的语境条件下(重复 vs. 变化)完成单词定义生成任务时的平均准确率。

跟单词定义匹配任务类似, LLMs 的学习效果总体上也要好于两组学习者(准确率更高) ($b = 2.48$, $SE = 0.08$, $z = 31.35$, $p < 0.001$), 但是, LLMs 仍然与学习者表现出一种非常一致的模式：“多篇” (变化)语境下的学习效果都要好于“多遍” (重复)语境($b = 1.29$, $SE = 0.21$, $z = 6.11$, $p < 0.0001$)。

但是, 不管是对学习者, 还是对 LLMs, 语境的情感效价与语境变异性都存在显著的交互效应(学习者：$\chi^2(1) = 12.51$, $p < 0.001$; LLMs：$\chi^2(1) = 19.15$, $p < 0.001$), 说明语境情感效价的影响还要取决于语境变异性的不同水平(重复 vs. 变化)。图 5 呈现了学习者和 LLMs 在单词定义生成任务中这两个变量的交互效应。

从图 5 可以看到, 语境情感效价和语境的变异性对学习者和 LLMs 在完成单词定义生成任务(高级词义的学习)的影响在总体趋势上非常相似：(1)清晰的变化语境优势, 即：不管是学习者还是 LLMs, 在变化语境下的高级词义学习效果总体上要好于在重复语境下(学习者：$b = 0.85$, $SE = 0.26$, $z = 3.30$, $p < 0.001$; LLMs: $b = 2.89$, $SE = 0.27$, $z = 10.80$, $p < 0.001$)。(2)变化语境下显著更强的积极情感优势, 即：在变化语境下, 语境情感效价越大, 所带来的高级词义的习得优势显著强于在重复语境下的优势(学习者：$b = 0.26$, $SE = 0.04$, $z = 6.96$, $p < 0.001$; LLMs: $b = 0.56$, $SE = 0.08$, $z = 7.43$, $p < 0.001$)。

[1] 把两组学习者或把各个大语言模型分开所获得的结果总体上一致。

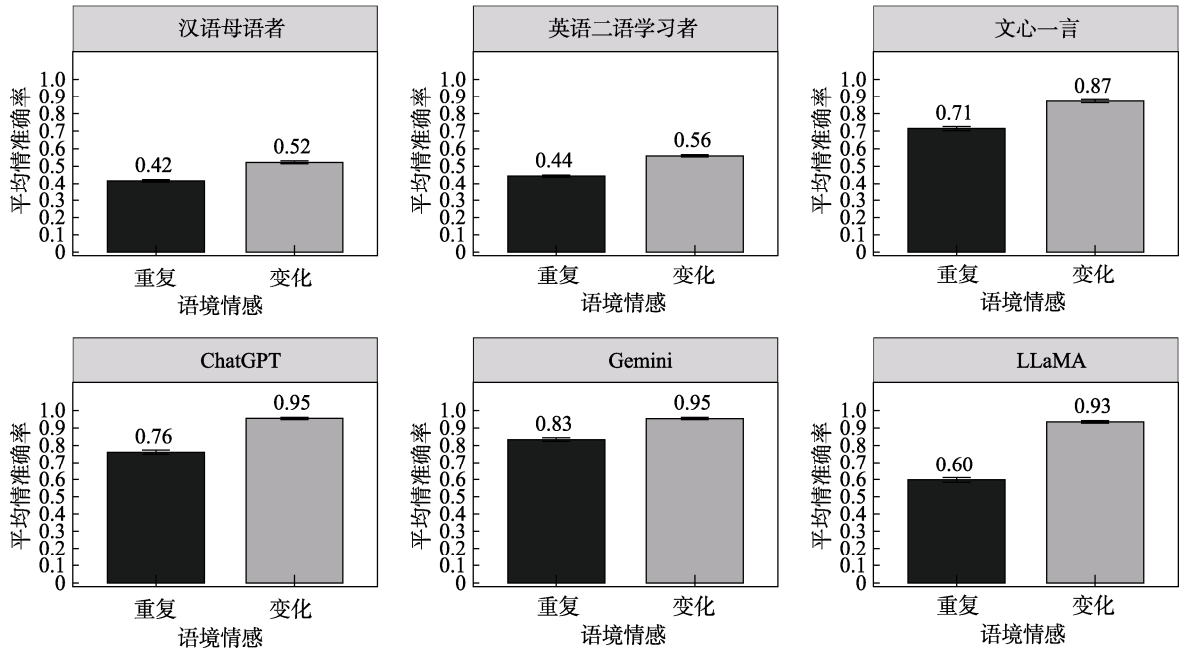

图 4 语言学习者和 LLMs 在重复和变化语境下完成单词定义生成任务的平均准确率

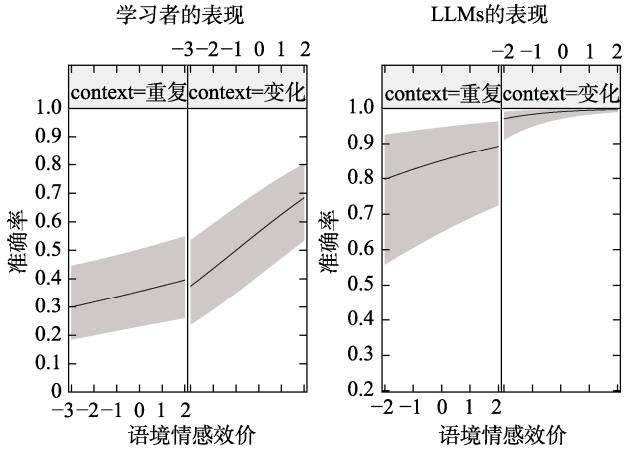

图 5 语境情感效价和语境变异性在学习者和 LLMs 单词定义生成的交互效应

由(2)进一步带来了(3), 即积极情感优势的放大效应：在变化语境下, 随着情感效价的提高, 学习者和 LLMs 的表现会越来越好, 进一步拉大跟重复语境之间的差距。

同样需要指出的是, 尽管在总体趋势上 LLMs 与两组学习者表现出了高度的相似性, 但是两者也表现出了一些具体的差别。比如, 在高级词义的习得上, LLMs 总体的学习效果要好于两组学习者(见图 5)。另外, 尽管都表现出变化语境优势, 但是 LLMs 的优势要比学习者在这两种语境之间的差异更为显著。就各个模型的具体表现看, 在变化语境下, 各个模型的表现都要优于文心一言(ps < 0.02), 而在重复语境下表现最好的是 Gemini, 显著好于 LLaMA 和文心一言(Ernie Bot) (ps < 0.01), 其次是 ChatGPT, 显著好于 LLaMA (p = 0.01), 其它模型之间没有区别。

4 讨论

本研究以一项基于阅读的词汇学习任务(LLMs 的零样本学习), 对比人类学习者与 LLMs 在语言学习(词汇附带习得)过程中的情感学习表现, 回答 3 个问题, 即 LLMs 是否可以像人一样习得单词出现的语境情感, 以及语境变异性和语境情感对 LLMs 和人的词汇附带习得的影响。下面分别从语境的情感及词形词义的习得来讨论结果。

4.1 语境情感的习得

从结果上看, 各个大语言模型以及所有语言学习者(母语及二语), 都通过阅读接触, 附带习得了目标词所出现的语境情感：当在积极语境下碰到过目标词时, 在测试中他们(LLMs 和学习者)就会认为该目标词的情感效价积极, 反之则认为其消极。对这一行为层面的趋同性, 社会心理学家称作为“情感传染” (emotional contagion), 即个体之间情感状态的共享(Hatfield et al., 2009; Ho et al., 2023), 而心理语言学家则称此为“迁移”, 即情感从语境迁移到了目标词本身(见 Snefjella et al., 2020)。问题是：不管是传染还是迁移, 它们是如何发生的?

为此, 我们提出一个“双重机制框架” (Dual-Mechanism Framework), 认为人类与 LLMs 通过本

质上迥异的路径, 最终呈现出在功能上等效的行为模式。首先, 对于人类学习者而言, 情感学习是一个具身的(embodied)、情境化的(situated)过程, 其内在机制可由具身认知理论(Embodied Cognition)得到有力阐释。该理论提出, 语言意义的建构根植于个体与环境的感官运动互动及其身体经验(Barsalou, 2008), 并高度依赖于语言输入所触发的具身模拟(embodied simulation)。例如, 当学习者阅读“Her wonderful teacher performed at the exciting concert”时, 他们并非被动接收文字符号, 而是主动激活与积极情感相关的感知−动作系统及情境记忆, 在脑中想象性地重现如兴奋的情绪体验、温暖的社会互动等多模态表征。目标词(如 *aunith*)正是通过与这类丰富的具身经验反复共现, 才被逐步“传染”上情感色彩, 从而实现语境情感的迁移。这一过程深深植根于人类的生物性与社会性互动中, 表明情感学习并非独立的认知模块, 而是与动机调控、注意力分配和记忆深度编码等过程紧密关联的动态系统(见 Eysenck & Brysbaert, 2018; Pessoa, 2008)。

然而, 对于大语言模型而言, 情感学习并非依赖具身认知系统, 而是一个计算的、基于统计分布与向量空间优化的过程。不同于人类通过感官与身体经验来建构语义, LLMs 的“学习”源于对大规模语料中词汇共现模式的抽取与再现(Bisk et al., 2020; Landauer & Dumais, 1997)。其核心操作可被形式化地理解为：在处理包含情感线索的大量语境时, 模型通过自注意力机制不断调整目标词在高维语义空间中的向量位置。当一个新词(如 *aunith*)反复出现在积极语境中, 它的向量会逐渐靠近 *exciting*、*wonderful* 等积极词汇的语义簇, 从而在数学上获得积极的情感取值。由此, 情感效价评分可解释为目标词向量与情感锚点向量的相似度计算, 其分值高低反映的是语义空间的邻近性而非主观体验; 而在句子生成中, 被定位于“积极区域”的词向量则更倾向于在自回归采样中触发积极词汇的联结, 从而系统性地再现情感一致性。这一机制说明, LLMs 的情感迁移实质上是基于统计共现的模式补全, 并不等同于人类的具身体验。然而, 本研究发现, 它们所展现出的“积极情感优势”等行为效应与人类高度一致。这种现象凸显了一种“异构同功”机制：即通过完全不同的技术路径, 纯粹依赖分布式语义建模与梯度优化的计算过程, LLMs 同样能够表现出类人的情感迁移功能。这一“双重机制框架”, 不仅为理解人工智能的类人特性提供了新的理论支点, 也通过计算建模的路径, 反衬出人类情感学习中具身体验的不可或缺性; 同时也提醒我们, 行为相似性不能简单等同于机制同一性(Binz & Schulz, 2023; Hagendorff et al., 2023)。

值得注意的是, 从基于使用的语言习得理论(Usage-Based Learning Theory)视角看(Ellis & Wulff, 2015; Tomasello, 2000), 人类与 LLMs 在情感语境学习中均表现出频率驱动与分布概括的共同特点：两者都通过重复接触语境−词汇共现模式, 逐步抽象并形成稳定的情感−语义联结。然而, 这一表面相似性的背后隐藏着的却是机制差异。对人类学习者而言, 频率效应是在具身模拟与社会互动中实现的, 频率强化了情感体验与词汇形式之间的深度编码; 而对 LLMs 而言, 频率直接体现为词汇向量在高维空间中的统计分布优化, 是一个基于分布式表征与自注意力机制的计算过程。

我们进一步发现, LLMs 在使用所“学”目标词进行句子产出时, 不仅能保持语法与语义的正确性, 其产出的情感走向也与学习阶段所处语境的情感取向高度一致。这一结果尤为关键：它显示 LLMs 并非只在表征层面建立了“情感−词汇”联结, 而是在语言产出阶段依然能够重现这种对应关系, 从而展现出与人类高度趋同的“情感一致性”与“积极情感优势”。在理论上, 这一现象具有“出人意料”的意义:现有的具身认知与情感迁移理论均以感官体验、身体参与和情绪调控为前提, 并未假设缺乏具身通道的纯计算系统能够表现出类比人类的情感迁移特征(Barsalou, 2008; Snefjella et al., 2020)。然而, 本研究的证据表明, LLMs 即便在零样本条件下, 仍可依托其分布式表征与自注意力机制, 在高维语义空间中对语境−词汇的统计共现关系进行条件化建模, 并在自回归生成中通过概率性模式补全并将目标词引导至匹配的情感区域。换言之, LLMs 的“类人”情感产出并非源于具身体验, 而是计算过程的产物。我们认为, LLMs 这种表现不仅将“类人”情感学习的边界从输入与表征扩展至输出与生成, 也为情感−语言交互的理论建构提供了新的可检验路径; 同时也进一步验证了我们提出的“双重机制框架”：人类的情感迁移依赖于由具身经验驱动的感知−情绪系统, 而 LLMs 的相似表现则依赖于对语言统计结构的结构化建模与泛化再现。

4.2　词形及语义的习得

上面的结果显示, LLMs 和人类学习者的词汇习得总体上展现出非常一致的模式, 都显著地受到

语境情感和语境变异性的影响。首先, 不管是语义识别还是语义生成, LLMs 和学习者的表现都体现出显著的积极情感优势：如果 LLMs 和学习者认为一个目标词在情感上越积极, 他们的词义习得效果就越好。由于这种积极性来自于语境情感, 因此, 这实际说明积极的语境情感促进了词义的习得。积极情感对人的学习的促进作用, 心理学领域提出过很多解释。比如, 有学者提出积极情感能够激发学习者的动机并提升注意力, 进而促进学习者与所学材料的持续互动(MacIntyre & Vincze, 2017)。还有学者提出, 积极的语境可以作为一种助记线索, 促进学习者从记忆中提取词汇的意义, 如果目标词与语境的积极情感产生关联, 有助于形成更加生动且易于检索的记忆痕迹(Lana & Kuperman, 2024; Snefjella et al., 2020)。而 LLMs 表现出的积极情感优势, 则从“零样本词汇学习”的角度展现了大语言模型的仿人特性, 另一方面也支持了情感会对人类学习造成影响。就本研究来说, 积极情感优势的背后则是情感与语境的紧密关联, 这种关联能有效促进词汇的深度加工与精确习得。

其次, 不管是语义识别还是语义生成, LLMs 和学习者的表现都体现出显著的变化语境优势：在变化语境下, 词义学习效果总是好于重复语境。这个结果与语境变异假说相一致, 从人工智能的角度支持了多样化语境的优势。Jones 等人(2012)提出的语义独特性模型(Semantic Distinctiveness Model, SDM)从词义表征的角度提供了 LLMs 这种表现的一个强有力的解释。在 SDM 中, 单词是单词——上下文矩阵中文档分布的一部分。该矩阵利用如频率、文档数量或语义独特性等索引将单词表示为向量。随着单词在不同上下文出现次数的增加, 其向量表示也会扩展。在阅读中, 读者会不断将向量中存储的信息与单词出现的新鲜语境进行比较。如果新鲜语境与向量表示中已经存储的信息一致, 则新鲜语境只会以较弱的方式进行编码。相反, 如果新鲜语境包含了向量表示中未存储的新的独立信息, 则新鲜语境就会以更强的方式进行编码。学习者每次在文本中碰到一个单词时, 都会启动一个新的认知历程, 将新单词与心理词典中已存储的上下文知识进行比较。如果与其存储的上下文知识不同, 则读者对该单词的心理表征会获得更新; 相反, 如果与存储在记忆向量中的上下文知识一致, 则读者对该单词心理表征的更新就较少。基于此, 读者能获取的该单词的上下文信息越丰富, 其心理词典的更新就越丰富(Jones et al., 2017)。这有助于读者及 LLMs 获得该单词的去语境化的知识, 促进核心意义知识的学习。

另外, 在语义生成任务的表现, LLMs 与人类学习者也都展现了显著不同于语义识别的结果, 主要表现在语境变异和语境情感的显著交互效应所造成的变化语境优势的放大：变化语境下, 随着情感效价的提高, 学习者和 LLMs 的表现都越来越好, 拉大跟重复语境之间的差距。对人类学习者而言, 这个结果说明变化语境为同一目标词提供了更丰富的情境线索, 使得与动机、注意、深度编码相关的具身模拟更频繁地被触发, 而积极语境进一步提升了心境与提取线索的匹配度, 因而在回忆式产出中形成更稳固、更可检索的高层语义表征(Schmidt, 1990; Tulving & Thomson, 1973)。对 LLMs 而言, 同一目标词在多样语境中的反复出现, 会让其上下文依赖的向量表示在高维语义空间内获得更分散的情感/语义信号：自注意力在不同上下文对齐不同的情感线索(如与 *exciting*, *wonderful* 等词簇的同现), 从而推近目标词向量与积极词汇簇之间的几何邻近关系, 提升情感维度的可线性分离性与条件似然(参见 Andrews et al., 2009; Landauer & Dumais, 1997)。另外, 积极语境在通用语料中具有更高的词汇密度与共现连通性(正向词汇在网络中更“稠密”), 这使得目标词在训练/推理时更容易“落入”积极区域的“吸引盆地”, 在自回归生成中优先采样到与该区域一致的描述性用语, 因而在定义生成这类高要求的回忆式输出上体现出更强的“情感一致性”与更高的正确率(参见 Clark et al., 2019; Radford et al., 2019)。换言之, 人类的放大效应源于具身模拟驱动的深度编码与提取匹配, LLMs 的放大效应则来源于分布式表征在变化语境下对情感共现信号的条件化聚合与向量几何上的高密度邻域效应。

不过需要注意的是, 尽管在总体模式上, LLMs 与人类学习者在词汇习得(词形及词义)上表现出高度相似性, 但是在具体测试指标上, 包括词形测试、单词定义匹配任务和定义生成任务, LLMs 所表出的准确率都要显著优于人类学习者。我们认为, 这种优势并非表明 LLMs 比人类“更聪明”, 而是揭示了二者在学习机制上的根本性差异。人类的学习是一个资源受限的过程, 受到工作记忆容量、注意力波动、动机水平以及已有知识结构的强烈影响(Eysenck & Brysbaert, 2018)。在有限的实验接触中, 人类学习者可能无法完全整合所有信息。相反,

LLMs 作为以预测和模式匹配为核心目标的计算架构, 在其上下文窗口内近乎完美地执行着统计优化任务。它们不受疲劳、分心或动机问题的困扰, 能够毫不动摇地利用输入文本中的每一个统计线索(Bisk et al., 2020)。具体而言, 在当前任务框架下, LLMs 的行为仿佛一个在无认知约束条件下的理想化统计学习者。它们所展现出的更强、更纯粹的“语境变异优势”和“积极情感优势”, 可以被视为对分布语义学假说(即词汇意义源于其语境分布)的有力验证。LLMs 放大了这些效应, 恰恰证明了这些学习原则本身是高效且强大的。因此, LLMs 的表现从一个独特的角度说明, 人类词汇习得中观察到的类似效应, 其底层很可能同样遵循着高效的统计学习原则, 只是受限于人类生物与认知条件。这一“机器超越人类”的发现强化了我们前面提出的“双重机制框架”, 并凸显了利用 LLMs 作为计算模型来分离和检验人类学习模型中各个变量的潜力。

需要指出的是, 尽管本研究纳入了当前最具代表性的 4 种大语言模型展开对比分析, 研究结果仍具有一定的可推广性限制。首先, 4 个模型在架构与规模上差异较大, 涵盖了标准 Transformer 架构(如 LLaMA 与 ChatGPT)以及融合 MoE 机制的混合架构(如 Gemini), 其在推理路径和参数激活机制上的差异可能影响模型的语境情感加工方式。其次, 语言优化方向也存在不同, 文心一言为中文优化模型, 而 ChatGPT 与 Gemini 主要基于英文语料训练, 这种训练背景可能导致模型对语境中的情感信号反应存在语言特异性。因此, 尽管我们在输入结构与测试任务设计上实现了统一, 但不同模型间的语言适配能力与语义表征机制差异, 仍需在今后研究中进一步验证其对语境情感习得能力的一致性与稳健性。此外, 当前研究仅选择了 4 种具有代表性的大语言模型, 未来研究可进一步扩大模型覆盖范围, 纳入更多架构类型和语料背景的模型, 以系统检验不同模型在语境情感习得任务中的稳定性与差异性。

5　结论

本研究以“零样本语言学习”为实验框架, 比较了 4 种代表性大语言模型与 3 组人类学习者在基于阅读的词汇附带习得中的情感学习表现, 以及语境情感与语境变异性对学习效果的影响。结果表明, LLMs 不仅能够像人类一样从语境中“吸收”并迁移情感, 在词形与词义学习中也展现出与人类高度一致的“积极情感优势”和“语境变异优势”, 甚至在多个指标上表现出超越人类学习者的准确性。这一发现表明, LLMs 已具备类人层面的语境敏感性与情感一致性, 凸显了分布式语义建模在复杂学习任务中的效力。

在理论层面, 我们提出并验证了一个“双重机制框架”, 以阐释行为相似性背后的异构认知基础：人类的情感学习依赖于具身经验驱动的感知−情绪系统, 其过程深深嵌入在感官运动体验、情境记忆与社会互动中; 而 LLMs 的情感迁移则源于高维语义空间中对语言共现模式的统计建模与再现, 是一种基于自注意力与向量空间优化的计算过程。两者在功能上呈现出相似的结果, 却体现了生物认知与机器计算之间的根本差异。该框架不仅为理解“类人”智能的情感表现提供了理论工具, 也为探索语言−情感交互的跨机制实现路径提供了系统性的比较基础。

更重要的是, 本研究在宏观层面揭示了情感学习在人工智能发展中的双刃剑意义。一方面, LLMs 的类人情感学习和迁移能力为教育、医疗和社会治理等领域的人机互动带来前所未有的契机, 它们能够在无情感标注的条件下捕捉并再现语境−词汇的情感联结, 从而提升生成内容的自然度与可接受性。另一方面, 这一能力也提示我们必须警惕其伦理风险：模型在继承和放大语料中的文化偏见与情感倾向时, 可能在不知不觉中影响其情感输出的价值导向(Caliskan et al., 2017)。换言之, LLMs 的“情感一致性”不仅是认知科学与分布语义理论的实验验证, 同时也是人工智能伦理评估的前沿议题。

综上所述, 本研究的发现不仅推动了我们对人类与机器在情感学习机制上的深度理解, 也为未来制定人工智能系统的情感规范与风险防控策略提供了坚实的实证依据与理论支点。

参　考　文　献

Ahmed, S. (2004). *The cultural politics of emotion*. New York: Rouledge.

Ahmed, S. (2010). *The promise of happiness*. London: Duke University Press.

Andrews, B., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463−498.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390−412.

Balass, M. (2011). *Learning words in context: An ERP investigation of word experience effects on familiarity and meaning acquisition* [Unpublished doctorial dissertation].

University of Pittsburgh.
Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1−23.
Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617−645.
Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *Computer Science*. https://doi.org/10.48550/arXiv.2306.03917
Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... Turian, J. (2020). Experience grounds language. *Computer Science*. https://doi.org/10.48550/arXiv.2004.10151
Blythe, H. I., Liang, F., Zang, C., Wang, J., Yan, G., Bai, X., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words: An eye movement study. *Journal of Memory and Language*, *67*(2), 241−254.
Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*(2), 122−159.
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877−1901.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183−186.
Chomsky, N. (1957). *Syntactic structures*. Berlin: De Gruyter Mouton.
Christensen, R. H. B. (2023). *ordinal: Regression models for ordinal data*. R package version 2023.12-4.1. https://CRAN.R-project.org/package=ordinal
Clark, E., Celikyilmaz, A., & Smith, N. A. (2019, July). *Sentence mover's similarity: Automatic evaluation for multi-sentence texts*. Paper presented at the meeting of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.
Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Grosset/Putnam.
Driver, M. (2022). Emotion-laden texts and words: The influence of emotion on vocabulary learning for heritage and foreign language learners. *Studies in Second Language Acquisition*, *44*(4), 1071−1094.
Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten, & J. Williams (Eds.), *Second language acquisition research series: Theories in second language acquisition* (pp. 75−94). Routledge.
Eysenck, M. W., & Brysbaert, M. (2018). *Fundamentals of cognition* (3rd ed.). Routledge.
Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., ... Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, *21*(3), 563−584.
Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., ... Schulz, E. (2023). Machine psychology. *Computer Science*. https://doi.org/10.48550/arXiv.2303.13988
Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114−121.
Hatfield, E., Rapson, R. L., & Le, Y. L. (2009). Emotional contagion and empathy. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 19–30). Boston, MA: MIT Press.
Ho, M. H., Kemp, B. T., Eisenbarth, H., & Rijnders, R. J. P. (2023). Designing a neuroclinical assessment of empathy deficits in psychopathy based on the Zipper Model of Empathy. *Neuroscience and Biobehavioral Reviews*, *151*, 105244.
Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*, 17.
Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258−286). Cambridge University Press.
Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2), 115−124.
Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 67, pp. 239−283). United States: Elsevier Science & Technology.
Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, *166*, 190−211.
Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627−633.
Lana, N., & Kuperman, V. (2024). Learning concrete and abstract novel words in emotional contexts: Evidence from incidental vocabulary learning. *Language Learning and Development*, *20*(2), 158−173.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211−240.
Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729−741.
Lauro, J., Schwartz, A. I., & Francis, W. S. (2020). Bilingual novel word learning in sentence contexts: Effects of semantic and language variation. *Journal of Memory and Language*, *113*, 104123.
Li, Z. (2024). *Semantic prosody acquisition and its influence on the learning of L2 novel word forms and meanings* [Unpublished doctorial dissertation]. Shanghai Jiao Tong University.
[李赞. (2024). *语义韵习得及其对二语新颖词词形和词义学习的影响* (博士学位论文). 上海交通大学.]
Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157−176). Amsterdam, The Netherlands: John Benjamins.
Ma, Z., & Li, Z. (2024). Acquiring semantic prosody in L2 novel word learning: The effect of context variability and gender. *Modern Foreign Languages*, *47*(6), 790−801.
[马拯, 李赞. (2024). 二语新颖词语义韵的习得：语境变异性及性别的影响. *现代外语*, *47*(6), 790−801.]
MacIntyre, P. D., & Vincze, L. (2017). Positive and negative emotions underlie motivation for L2 learning. *Studies in Second Language Learning and Teaching*, *7*(1), 61−88.
Nevisi, R. B., Hosseinpur, R. M., & Darvish, F. Z. (2018). The impact of L1/L2-based explicit output task instruction on Iranian EFL learners' semantic prosody learning. *Journal of Language Horizons*, *2*(2), 51−74.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148−158.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129−158.
Sinclair, J. (1987). *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London, England: Collins ELT.
Snefjella, B., Lana, N., & Kuperman, V. (2020). How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language*, *115*, 104171.
Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, *41*(1), 89−108.
Tamir, M., Schwartz, S. H., Cieciuch, J., Riediger, M., Torres, C., Scollon, C., ... Vishkin, A. (2016). Desired emotions across cultures: A value-based account. *Journal of Personality and Social Psychology*, *111*(1), 67−82.
Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1−2), 61−82.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352−373.
Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, *10*(2), 1−37.
Wetherell, M. (2012). *Affect and emotion: A new social science understanding*. Los Angeles: SAGE.
Wu, S. Y., & Li, Z. (2024). How semantic prosody is acquired in novel word learning: Evidence from the “Double-Jujube Tree” effect. *Acta Psychologica Sinica*, *56*(5), 531−541.
[吴诗玉, 李赞. (2024). 新颖词语义韵的发生机制：“双枣树”效应的证据. *心理学报*, *56*(5), 531−541.]

Zero-shot language learning: Can large language models (LLMs) acquire contextual emotion in a human-like manner?

WU Shiyu, WANG Yiyun

(*School of Foreign Languages, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract

Emotion is a structural resource in human cognition that guides attention, memory, and social coordination. During incidental vocabulary acquisition (IVA), readers often internalize the affective tone of surrounding discourse and transfer it to novel words (“contextual emotion transfer/semantic prosody”). Recent LLMs appear to display analogous behavior despite lacking embodiment, raising the question of whether they can acquire contextual emotion in a human-like manner and whether the same contextual factors shape both human and model learning. Building on usage-based and distributional accounts, we expected two robust regularities to hold across agents: a positivity advantage (higher contextual valence predicts better learning) and a variability advantage (varied contexts outperform repeated ones). We further hypothesized that, in more demanding recall (definition generation), contextual valence would interact with variability, such that positive emotion would amplify the benefits of varied contexts.

We conducted zero-shot, parallel evaluations with four representative LLMs (Ernie Bot 3.5, ChatGPT/GPT-4, Gemini 1.5 Pro, LLaMA 3.1-8B) and three human cohorts matched to prior IVA paradigms (English L1, Chinese L1, English L2; 306 participants). Each agent learned nine pseudowords embedded in 45 two-sentence texts spanning positive, neutral, and negative contexts; context variability was manipulated between repeated versus varied exposures. After reading, LLMs completed (a) valence rating and sentence production (emotion transfer) and (b) orthographic choice, definition matching, and definition generation (form/meaning). LLMs were evaluated in strictly isolated zero-shot sessions with no task-specific supervision or fine-tuning. Ordinal mixed-effects models (CLMM) analyzed ratings; linear/logistic mixed-effects models analyzed production and accuracy, with random effects for participant/LLM session, item, and denotation class.

Contextual emotion transferred reliably to targets: across humans and LLMs, ratings followed positive > neutral > negative, and generated sentences aligned in polarity with the learning context. For vocabulary learning, both groups exhibited a positivity advantage—higher contextual valence significantly predicted better meaning performance—and a variability advantage—varied contexts significantly outperformed repeated

contexts in definition matching and definition generation. In recall, valence interacted with variability: positive emotion amplified gains under varied exposure for both humans and LLMs, yielding the largest improvements in definition generation. LLMs frequently matched or exceeded human accuracy in form recognition and often reached higher overall accuracy on meaning tasks while preserving the same qualitative patterns. These effects held in mixed-effects analyses controlling for participant/session, item, and denotation, and were observed without providing LLMs with examples, feedback, or fine-tuning.

The study showed that LLMs did acquire contextual emotion and reproduced core human regularities (positivity and variability advantages; valence-by-variability interaction in recall). We interpret the convergence via a Dual-Mechanism perspective: human emotion learning is embodied and socially situated, whereas LLM "emotion" arises from distributional co-occurrence and vector-space optimization; distinct mechanisms can yield functionally similar behavior. The findings advance theories of emotion–language interaction and support context variability as a general driver of vocabulary learning. Practically, emotion-sensitive LLM behavior can enhance educational and communicative applications, while necessitating safeguards against unintended amplification of corpus-borne affective biases.

Keywords large language models, zero-shot learning, emotion learning