

·第二十七届中国科协年会学术论文·

·计算建模与人工智能·

大语言模型与心理学的历史渊源^{*}

黄林洁琼¹ 张雯^{1,2} 陈珍^{1,2} 李晨曦^{1,2} 李兴珊^{**1,2}(¹ 中国科学院心理研究所, 北京, 100101)(² 中国科学院大学心理学系, 北京, 101408)

摘要 近年来, 大语言模型表现出媲美人类的语言理解和生成能力, 成为人工智能领域的重大突破。在大语言模型的发展历程中, 心理学与其存在深厚的历史渊源。首先, 作为模型架构基础, 人工神经网络早期被心理学家用于模拟人类认知过程; 其次, 计算机科学与心理学在词汇语义表征方面的学科交叉推动了词嵌入技术的发展; 第三, 大语言模型和人类在实时语言加工中存在相似特点; 此外, 跨领域的人员交集与学术传承构成了推动人工智能研究的重要力量。因此, 心理学在语言加工与认知建模方面的耕耘及其与计算机科学的深入合作作为语言智能的突破做出重要贡献。这为科研人员提供重要启示: 跨学科合作与基础研究的深耕是推动创新和实现突破的关键。

关键词 大语言模型 心理学 跨学科合作 语言认知 人工神经网络

语言作为人类认知的核心能力, 其加工机制的复杂性使语言理解与生成成为人工智能领域最具挑战性的任务之一, 被誉为“人工智能皇冠上的明珠”。然而, 令人意外的是, 近年来大语言模型 (large language models, LLMs) 取得了突飞猛进的发展, 展现出接近人类的语言认知能力 (Brown et al., 2020)。这让人不禁深思: 人工智能为何能在如此复杂的语言加工领域取得突破性进展? 本文通过回顾大语言模型的发展历程, 探讨其与心理学的历史渊源, 并在此基础上论述学科交叉在推动技术突破中的重要性, 为未来心理学和人工智能的跨学科交叉研究提供启示。

1 语言加工的复杂性及大语言模型取得的突破

人类语言作为一种抽象的符号系统, 其加工机制具有独特的复杂性。语言理解需要处理多层次的抽象表征。例如, 在词汇层面, 同一个词 (如“bank”) 可能对应完全不同的语义 (“银行”或“河岸”), 这种多义性解析依赖于动态的语境整合 (Kintsch, 1988); 在句法层面, 人类需要处理各种复杂的句式 (如嵌套句、被动句、倒装句、省略句、虚拟语气句、并列复合句、主从复合句等); 在语用层面, 语言行为常包含隐含的意义 (如“原则上可以”实为不可以, 而“原则上不可以”实为可以), 理解这种意图需要调用复杂的社会认知与心理理论

* 本研究得到国家自然科学基金项目 (32371156) 的资助。

** 通讯作者: 李兴珊, E-mail: lixs@psych.ac.cn

DOI:10.16719/j.cnki.1671-6981.20250401

(Ouerchefani et al., 2024)。此外,人类不仅能够理解语言,还能够创造性地使用语言,表达新颖的思想和概念。语言加工的复杂性曾一度使人工智能难以在语言加工领域有所进展(张钊等,2020)。

然而,近年来人工智能在语言领域取得了令人瞩目的突破。大语言模型通过深度学习海量人类语料,能够获得媲美人类的语言理解与生成能力(Vaswani et al., 2017)。其采用 transformer 架构,通过注意力机制捕获文本中的关键信息和内在关系。经过在庞大的文本数据集上进行自监督预训练,大语言模型可以学习人类语言的规律,进行语言生成和理解(Radford et al., 2019);经过监督微调和强化微调,大语言模型还可以学会执行人类的各项任务,按照人类意图和价值观进行文本生成(Gunel et al., 2020; Ouyang et al., 2022)。

大语言模型的成功并非一蹴而就,而是得益于一系列关键突破。2017年,谷歌提出了 transformer 架构,这是一种基于自注意力(self-attention)机制的序列转换模型(Vaswani et al., 2017)。相较于此前主流的循环神经网络(recurrent neural network, RNN),它通过自注意力机制直接捕捉文本中的长距离依赖关系,避免了序列逐步计算的局限性,实现了并行化处理和更高效的全局信息交互。通过自注意力机制,transformer 架构能够捕捉全局上下文依赖关系,有效缓解生成过程中的局部短视问题,从而提升文本的连贯性和流畅性,使其更贴近人类的自然语言表达模式。2018年,两大里程碑模型相继问世:谷歌的 BERT 模型采用掩码机制,通过上下文预测被遮蔽的词语(Devlin et al., 2019); OpenAI 的 GPT 则通过前文预测后续词语(Radford et al., 2018)。二者均采用“预训练+微调”的方法,先利用大量人类语料学习通用语言表征,再针对具体任务进行优化,极大地提高了自然语言处理任务的性能,尤其是在逻辑推断、情感分析、长文理解等任务中取得了显著进步。2020年, GPT-3 的发布标志着大语言模型时代的来临,其 1750 亿参数规模展现了惊人的语言理解和生成能力,使其能够在开放对话场景中实现流畅的语言交互(Brown et al., 2020)。2022年, InstructGPT 通过强化微调,使模型输出在表达习惯和价值观上都更符合人类偏

好(Ouyang et al., 2022),为后续 ChatGPT 等对话系统奠定了基础。这一系列突破性进展,共同构建了现代大语言模型的技术体系。

大语言模型的成功固然离不开计算机软硬件的发展,但更为关键的是其背后理论和方法的积累以及跨学科合作的有力支撑。接下来,本文将深入探讨大语言模型与心理学的历史渊源,揭示跨学科合作在大语言模型成功中所发挥的重要作用。

2 大语言模型的架构基础: 神经网络模型在语言认知领域的深耕

大语言模型的发展依赖于深度学习技术的突破,而深度学习是对人工神经网络的拓展与深化。作为一种受生物神经系统启发的计算模型,人工神经网络由相互连接的神经元组成,一般包括输入层、隐藏层和输出层;通过神经元之间的连接形成网络,并通过调整连接的权重从数据中学习复杂模式以做出预测或决策(Dongare et al., 2012)。相比于早期的人工神经网络,现代大语言模型在多个方面实现了飞跃式发展,表现为网络层数逐渐加深、训练数据逐步扩大、参数规模不断扩大、结构逐渐复杂等特点(Smith et al., 2022)。这些进步使大语言模型具备了强大的语言理解和生成能力。因此,人工神经网络是大语言模型的架构基础,大语言模型的突破与人工神经网络的发展密不可分。

人工神经网络的发展是心理学、神经科学、计算科学和数学等多学科长期合作的成果。1943年,神经生理学家 Warren S. McCulloch 和数学家 Walter R. Pitts 首次提出“神经网络”的概念,用形式逻辑模拟神经元连接,为人工神经网络奠定了早期框架(McCulloch & Pitts, 1943)。随后,神经心理学家 Donald Hebb 在 1949 年发表的《行为的组织》中提出著名的 Hebb 学习规则,指出突触连接的强度可随学习或经验而改变,引入了神经网络的学习机制(Hebb, 1949)。1957年,心理学家 Frank Rosenblatt 提出感知机。这是首个物理构建并具备学习能力的人工神经网络,能够通过监督学习完成简单的线性分类任务(Rosenblatt, 1957)。Hebb 和 Rosenblatt 的工作都强调了人工神经网络的学习机制,而学习能力是人工智能技术发展的核心动力,

也是大语言模型成功的重要前提之一。进入 20 世纪 80 年代，人工神经网络的结构日益复杂，逐步用于模拟更复杂的认知过程。1982 年，生物物理学家 John Hopfield 提出具有反馈机制的 Hopfield 神经网络，推动了人工神经网络在联想记忆中的应用。1986 年，心理学家 David Rumelhart 等人提出反向传播（backpropagation）算法，使人工神经网络通过误差反馈调整连接权重，在功能上能够模拟人类神经网络的学习效果（Rumelhart et al., 1986）。同年，心理学家 David E. Rumelhart 和 James L. McClelland 在《Parallel Distributed Processing》中提出并行分布式加工（PDP）模型，系统地运用人工神经网络模拟人类的语言理解等认知过程，推动了心理学与人工神经网络的深度合作（Rumelhart et al., 1986）。进入 21 世纪，基于人工神经网络的模型在计算机领域迎来了新的爆发。2006 年，计算机学家和认知心理学家 Geoffrey Hinton 等人提出深度信念网络，标志着深度学习的兴起（Hinton et al., 2006）；2012 年，Hinton 团队使用深度卷积神经网络在图像识别领域取得突破（Krizhevsky et al., 2012）；近年来，BERT、GPT 等大语言模型蓬勃发展。如今，人工神经网络已广泛应用于自然语言处理、图像识别、医学诊断等各个领域。

如上所述，在人工神经网络的发展历程中，心理学家不仅提出了重要的理论，也推动了人工神经网络在认知建模中的应用，致力于让其更好地模拟人类认知过程。而作为心理学的一个重要分支，心理语言学将人工神经网络应用于人类的语言加工研究，为其模拟人类的语言加工过程提供了重要的理论支持和架构启发。基于人工神经网络的语言加工认知模型在 20 世纪 80 年代兴起，在 1995 年左右达到鼎盛。其中，PDP 模型激发了心理语言学领域的广泛研究，包括词汇识别、词汇命名和言语感知等方面（Coltheart et al., 2001; McClelland & Elman, 1986; McClelland & Rumelhart, 1981; Seidenberg & McClelland, 1989）。作为一种连接主义模型，PDP 强调信息并非存储在某个单元中，而是分布在整个网络的权重和连接模式中；认知过程则通过多个单元并行加工以及它们之间的激活和抑制实现。例如，McClelland 和 Rumelhart（1981）提出的交互激活

模型，模拟了不同情景下的词汇识别过程。该模型认为视觉词汇识别主要涉及视觉特征水平、字母水平和词汇水平。视觉输入首先激活与其一致的字母探测器，这些字母探测器进一步激活与之匹配的词探测器。反过来，被激活的词探测器会反馈到字母水平，增强相关字母探测器的激活，使这些字母更容易被感知。该模型能够模拟词优效应：与单个字母或无意义字符串中的字母相比，真词内的字母因为接收到词汇水平的反馈而有更多激活，因此更容易被识别。

此外，心理学家还利用人工神经网络探讨语言加工中的核心争论。其中一个备受关注的议题是：语音和拼写形式一致的规则词与不一致的不规则词是否依赖相同的加工机制。对此，心理语言学提出不同理论并通过人工神经网络模型进行验证与模拟。例如，双通路模型（dual-route cascaded model; Coltheart et al., 2001）强调二者依赖不同通路生成发音，规则词通过字母-音素映射规则进行拼读，而不规则词则需要通过心理词典检索发音。与此不同，三角模型（triangle model; Seidenberg & McClelland, 1989）则提出一种统一的加工机制，认为所有词都通过调整神经网络的连接权重实现发音，而非依赖拼读规则或词典检索。

再者，基于人工神经网络的语言加工认知模型还可用于模拟人类词汇学习的动态过程。以三角模型为例，研究者通过操控训练的词汇量（如使用 1000 个或 10000 个词）对模型进行训练，发现模型的表现会随着词汇量的增长而提升，这种学习轨迹与儿童语言发展过程相似。因此，研究者可以据此将模型在不同训练阶段的表现与儿童的词汇学习阶段进行类比，从而更深入地探讨词汇学习的机制与发展过程（Seidenberg & McClelland, 1989）。可见，为了让人工神经网络的结构更贴近人类的认知过程，心理语言学家开展了很多针对性的工作。在这一过程中，人们对于人工神经网络如何进行语言理解和生成有了更深的理解。

总体来说，作为大语言模型的架构基础，人工神经网络的发展与心理学家的重要贡献紧密相关。心理学家不仅在人工神经网络的发展中提出了重要理论，还长期利用人工神经网络模拟人类的认知加

工过程。为了使人工神经网络更好地模拟人类的语言理解和生成,心理语言学领域不断发展和完善人工神经网络的架构,从而为人工智能在语言处理领域取得重大突破提供了重要的方法基础。

3 大语言模型的语义表征:学科交叉推动发展的词嵌入技术

词嵌入(word embeddings)技术通过训练大量的文本数据,将词汇的语义及其语义关系转化为向量形式,使复杂抽象的语言能够变成计算机可以理解和处理的形式。因此,这一技术的发展是人工智能能够在语言领域取得突破的一个关键前提。词嵌入技术的发展依赖于词汇语义表征方法的不断演进,而该演进过程得益于心理学、语言学与计算机科学等多个学科的交叉融合。

词汇语义表征方法的发展大致经历了五个关键阶段。(1)“分布即意义”思想的提出。20世纪中期,语言学家 Zellig Harris 和 John R. Firth 提出词汇语义的分布式表征思想,认为词义应通过上下文构建(Firth, 1957; Harris, 1954),为后来的分布式语义模型(distributional semantic models, DSMs)提供了理论基础。心理学家 Osgood (1952)提出的语义差异法通过量表评分和因子分析,为词义的多维量化表征提供了方法灵感。(2)符号表征方法的发展。典型代表是 WordNet (Miller, 1995),它将词义表示为同义词集,同义词集之间通过语义关系形成语义网络,具有较强的可解释性,但在处理抽象概念等方面存在局限。(3)静态向量表征的建立。20世纪末,认知心理学家 Kevin Lund 和 Curt Burges 开发了 HAL 模型(Lund & Burgess, 1996),通过词-词共现关系构建高维向量表征词义;Thomas K. Landauer 和 Susan T. Dumais 提出 LSA 模型(Landauer & Dumais, 1997),采用词-上下文共现矩阵并结合奇异值分解方法,实现了高维语义空间的有效压缩。两者是第一代分布式语义模型的代表,生成的高维稀疏向量通常需要降维处理。在深度学习的推动下,第二代分布式语义模型如 word2vec (Mikolov et al., 2013),通过人工神经网络直接生成词汇的低维密集向量,词汇和隐藏层之间的权重反映了词义信息。但其产生的词向量是静态的,难以区分不同语境中

的词义。(4)动态语义表征的发展。以 BERT (Devlin et al., 2019) 和 GPT (Radford et al., 2019) 为代表的大语言模型依赖于多层编码器网络,可生成动态的、上下文依赖的词向量,能够捕捉不同语境下的词义。

(5)跨模态表征的兴起。近来,语言模型开始与其他人工智能技术(如计算机视觉)相结合,将不同模态的数据映射到同一向量空间中,生成词汇的“跨模态嵌入”(Zhang et al., 2024)。

如上所述,多学科交叉融合共同推动了词汇语义表征方法的持续创新。一方面,计算机科学贡献了计算框架和优化技术,大规模数据处理与人工神经网络等技术的进步提供了核心驱动力;另一方面,心理学和语言学提供了理论灵感和方法启发,尤其是心理学家对词汇语义量化方法的探索为分布式语义模型奠定了基础。词汇语义表征方法的发展促成了词嵌入技术的进步,使得海量语料文本的快速高效处理成为可能,推动了人工智能在语言领域的重大突破。

4 大语言模型和人类在实时语言加工中的共同特点

大语言模型的算法与人类语言认知机制也展现出诸多的共同特点,主要体现在渐进性、预测性和动态注意力分配三个方面。人类和大语言模型都表现出渐进式加工的特征。人类在语言交流中能够快速、高效地理解复杂的语言结构,而无需等待全部信息接收完毕,这是因为大脑会随着语言信息的逐步输入实时进行语义整合和句法分析(Frazier & Rayner, 1982)。在语义整合方面,当输入词与前文语境不一致时,如句子“He spread the warm bread with socks”中的“socks”会诱发较大的 N400 效应(Kutas & Hillyard, 1980)和更长的注视时间(Rayner et al., 2012),这表明大脑会根据上下文语义对新输入词进行实时整合。在句法分析方面,人类认知系统表现出典型的渐进式处理特征。以花园路径句“While Mary bathed the baby played in the crib”为例,读者会首先将“the baby”解析为“bathed”的直接宾语,直到遇到歧义词“played”时才意识到需要重新分析句子结构。这种句法重构过程表现为个体在歧义词上产生更长的阅读时间和更多的回视行为

(Joseph & Liversedge, 2013), 并在解歧词上诱发 P600 成分 (Osterhout & Holcomb, 1992), 反映了实时的句法再分析机制。目前主流的生成式大语言模型为了建立词汇对前文的因果依赖, 也采用了类似自左向右的串行加工方式, 逐词 (token-by-token) 进行即时加工。当输入一个 token 时, 模型会基于该 token 本身的含义以及前文的语义背景, 对其进行即时理解和解释。如果输入与前文矛盾, 模型会降低该序列的概率分数, 从而调整其对后续输入的处理策略 (Vaswani et al., 2017), 这种串行加工模式与人类语言加工的时序特性形成了功能对应。值得注意的是, 并不是所有的模型均采用渐进式加工的方式。例如, BERT 模型就是利用上下文共同预测被掩码的词。虽然这种方法可以对书面文本进行非实时、离线的处理, 但在需要实时加工的自然口语处理中, GPT 所采用的渐进式加工策略更合理。这也许是人类和目前被广泛使用的大语言模型 (例如 GPT) 均采用渐进式加工的原因。

人类与大语言模型在语言加工过程中均对下一个词进行预测。例如, 在视觉情景范式中, 当被试分别听到 “the boy will eat” 和 “the boy will move” 时, “eat” 会让被试提前注视到屏幕上呈现的 “cake” 图片, 而 “move” 则不会产生这种效果, 这表明人类能够利用前文预测即将出现的词语 (Altmann & Kamide, 1999)。大语言模型在预测机制上的突破性进展始于 2018 年。谷歌的 BERT 采用掩码机制, 通过上下文预测被遮蔽的词语, 而 OpenAI 的 GPT 则通过前文预测后续词语。尽管二者的实现方法不同, 但都实现了与人类的预测类似的功能。研究发现, 在给定足够上下文时, GPT-2 模型与人类预测下一个词的准确率表现相当 (Goldstein et al., 2022), 这表明大语言模型在预测功能上已经达到了与人类相似的水平。

在信息选择方面, 人类和大语言模型均采用了动态的注意力分配机制。人类在处理语言信息时, 会根据语境和任务需求, 对关键信息进行策略性的注意力分配。这种动态的注意力分配机制在语言加工的不同层级都得到了实证研究的验证。在句子加工层面, Birch 和 Rayner (1997) 的研究发现, 当词汇出现在强调句中 (如 “It was the mayor who refused to answer a reporter's question”), 读者会对焦点词

“mayor” 的注视时间更长, 这表明更多的注意力被分配到了该词上 (Birch & Rayner, 1997)。在语篇加工层面, Kintsch 和 van Dijk (1978) 发现读者对表达篇章全局结构的命题 (如主题句) 的回忆率比局部命题 (如细节信息) 高大约 2~3 倍, 且表现出更慢的遗忘速度。Hyönä 等人 (2002) 使用眼动追踪技术也发现熟练读者会对篇章中的主题句产生更长的阅读时间和更多的回视行为, 表明读者会对主题句分配更多的注意力, 进行更深入的加工和理解。大语言模型通过自注意力机制实现了类似的动态权重分配, 在生成过程中, 模型会为更重要的词汇赋予更高的注意力分数 (Vaswani et al., 2017)。具体而言, 在预测下一个词时, 模型会动态调整所有词的注意力分数。通过深度学习不断更新参数, 模型会为那些对预测下一个词更重要的词赋予更大的注意力分数, 从而在预测时产生更大的影响。这种动态调整语言加工时注意力分配方式的算法, 虽然与人类的神经机制存在差异, 但在信息选择上与人类注意力功能类似。

需要强调的是, 语言加工的渐进性特征与预测性的结合, 构成了大语言模型实现无监督学习的核心基础。在预训练过程中, 这种特性使得模型能够通过上文预测后续词或通过上下文预测遮蔽词, 继而将实际语料中的真实词汇与模型预测结果进行比对。基于这种预测误差, 人工智能系统动态调整深度神经网络的权重参数, 从而完成知识表征的自主构建。这种基于语言生成机制的学习范式, 使人工智能系统能够实现自监督学习。与手工标注海量文本相比, 自监督学习使得人工智能系统利用海量文本进行训练更具可行性。可以说, 正是语言本身所具备的这种渐进可预测特性, 为人工智能在自然语言处理领域取得突破性进展提供了根本性的实现路径。

总体而言, 大语言模型与人类在语言处理上展现出渐进性、预测性和动态注意力分配三大功能相似性, 揭示了跨形态智能系统对语言普适性原则的遵循。人类语言加工过程中的实时加工、主动预测和焦点选择机制, 与大语言模型的自回归生成、掩码预测和自注意力机制虽然在具体机制方面上存在差异, 却展现出了高度的功能趋同。二者存在共同的特点说明, 对语言本质规律的把握可能是实现通

用智能的重要路径,因此心理学对语言加工机制的探索或许会为将来人工智能的发展提供关键的理论支持和实践指导。

5 计算机科学与心理学领域的人员交集与学术传承

大语言模型的发展得益于心理学家和计算机科学家等多领域研究者的跨学科合作与学术传承。正如前文所述,无论是人工神经网络的进步,还是词嵌入技术的发展,除了人工智能领域研究学者在技术上的推动,心理学和认知科学领域的学者也在模型架构与设计方面提供了重要的理论支持和方法启发。其中,不少研究学者同时具备心理学与认知科学的学术背景。例如 Geoffrey Hinton, David E. Rumelhart 和 James L. McClelland 三位研究学者,他们基于自身研究为大语言模型的发展做出了重要贡献。此外,三位研究学者还培养了众多科学家继续探究人工智能和大语言模型的发展。这些学者中包括 Ilya Sutskever (前 OPEN AI 首席科学家) 和 Ruslan Salakhutdinov (前苹果 AI 总监) 等为人工智能的发展起到决定性作用的工业界学者。基于 Neurotree (Neurotree.org) 网站和三人的官方网站,我们追踪了三人直接指导的本科生、硕士生、博士生和博士后的发展现状。经过统计和整理,Geoffrey Hinton 指导的 63 位学生,James L. McClelland 的 26 位学生和 David E. Rumelhart 的 11 位学生被纳入分析。结果表明,有 64 人继续在高等教育机构和研究所开展研究(其中 10 人同时具有高科技公司的任职),32 人在高科技公司任职。其中相当一部分研究学者基于心理学和认知科学的实验,开展与 AI 和人机交互相关的研究。这些学者在语言认知领域受到良好的训练,很多人继续从事和语言加工相关的研究,是人工智能研究的生力军。由此可见,一代代科学家在语言认知领域的持续努力,也是人工智能能够在语言认知领域取得突破进展的原因之一。

6 总结与启示

大语言模型近年来取得飞跃式发展,表现出与人类媲美的语言理解和生成能力,成为人工智能领

域的一项重大突破。本文通过回顾大语言模型的发展历程,指出大语言模型和心理学深厚的历史渊源是人工智能能够在语言处理领域取得突破的关键因素之一。首先,纵观人工神经网络的发展历史,心理学家不仅提出了重要的理论机制,也积极推动了人工神经网络在认知建模中的应用,为大语言模型的突破提供了坚实的理论与方法基础。其次,词嵌入技术作为大语言模型能够理解和处理语言的一个重要前提,其发展体现了多学科协作的力量。除了计算机科学在技术进步中的核心作用,语言学和心理学也提供了思想启发,并且心理学家通过大规模自然语言语料构建高维语义向量的方法,推动了词汇语义表征技术的持续进步。第三,大语言模型与人类在语言处理上具有共同的特点,表现在渐进性加工、预测后续内容和注意力动态分配等方面。人类和大语言模型具备这些共同特点的原因目前还不清楚,有可能是学者在开发大语言模型的过程中有意或无意地借鉴了人类语言加工的机制,也可能是为了适应人类语言的特点,人类和大语言模型独立发展出了类似的加工机制。无论如何,人类和大语言模型具备很多共同特点这个事实表明,心理学在语言加工机制方面的研究成果有望为人工智能的未来发展提供理论支持和实践指导。第四,计算机科学和心理学等领域的人员交集与学术传承构成了推动人工智能研究的重要力量。以 2024 年诺贝尔物理学奖得主 Geoffrey Hinton 为例,他虽以人工智能研究闻名,但其早期接受了心理学领域的教育,也在心理学领域做了大量的学术研究。Hinton 从心理学转向人工智能,一方面反映了人工智能的发展仅依赖心理学理论并不足够,还需要技术的进步;另一方面,他拥有的心理学视角可能启发了其在人工智能领域的工作;这也从侧面印证了跨学科合作对于人工智能发展的深远意义。

值得一提的是,大语言模型的突破也与语言在人类文明中的核心地位紧密相关。人类文明和知识的积累与传承在很大程度上是以语言的形式实现的。例如,人类几千年积累的文明多是以书籍的形式记录和传承,人类的沟通交流也主要是以语言和文字的形式进行。因此,要想真正具备与人类媲美的智能,人工智能系统就需要吸收人类积累的文明成果。而

具备能够加工人类语言能力的大语言模型，在吸收以语言形式传递的知识方面具有天生的优势，这也是人工智能在语言领域取得突破的一个重要原因。

人工智能在语言处理领域的突破性进展，充分反映了基础研究对于技术突破的重要性。大语言模型的成功并非一蹴而就，而是多个基础学科长期积累的结果。以词汇语义表征技术为例，虽然词嵌入技术的出现是一次技术飞跃，但这背后是从 20 世纪中叶“分布即意义”理论提出以来长达几十年的研究积淀。在 20 世纪末，认知心理学家发展了第一代分布式语义模型，从早期对词义的多维量化表征方法，到基于词与词、词与上下文共现关系的高维向量表征，再到语义空间的压缩优化，这一系列基础性研究成果为词嵌入技术的出现做了重要铺垫。由此可见，人工智能技术的突破不仅依赖技术本身的进步，还离不开心理学、认知科学、语言学等基础研究领域的理论支持。值得注意的是，尽管当前大语言模型已相当强大，但仍存在“一本正经地胡说八道”等问题，表现为常识和逻辑上的错误，与人们梦寐以求的通用人工智能仍有较大差距。为解决这些问题，我们应继续重视基础研究，并充分应用其成果。

大语言模型的成功也为未来心理学和人工智能的跨学科交叉研究提供重要启示。一方面，心理学能够启发人工智能的发展。例如，研究者可以借助心理学的研究范式，评估人工智能模型是否具备类似人类的认知能力（Aher et al., 2023）。此外，心理学在认知机制方面的研究成果也可能为人工智能算法提供灵感。Marr（1982）提出信息系统研究的三个层次，即计算理论、计算表征与算法、物理实现。虽然人类和人工智能在计算表征与算法以及具体的物理实现方面存在差异，但二者具有相同的计算理论，即均致力于理解、产生和学习语言，都涉及知识的参与和加工系统的制约。因此，心理学的研究成果有望启发人工智能算法的发展。另一方面，人工智能的发展也可以促进心理学的研究。如前文提到的，心理学研究为了澄清规则词和不规则词的加工机制，借助人工神经网络对不同理论进行模拟和检验。未来研究也可以利用人工智能模型帮助厘清更多心理学的核心争论，如“先天与后天”等经典

心理学议题。此外，人工智能的广泛应用催生众多新的心理学研究问题。例如，人工智能对人类认知的正向和负向影响以及由此引发的伦理与社会问题。

综上，人工智能在语言处理领域取得重大突破并非偶然，而是计算机科学、心理学等多个学科协同努力以及基础学科长期积累的结果。这一发展历程为科研人员提供重要启发：跨学科深度合作以及基础学科的深耕至关重要。未来研究可以更加注重不同学科之间的深度协作，从而为理论创新和技术进步以及实现关键突破提供重要支持。

参考文献

- 张钹, 朱军, 苏航. (2020). 迈向第三代人工智能. *中国科学: 信息科学*, 50(9), 1281–1302.
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). *Using large language models to simulate multiple humans and replicate human subject studies*. Paper presented at the Proceedings of the 40th International conference on machine learning.
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Birch, S., & Rayner, K. (1997). Linguistic focus affects eye movements during reading. *Memory and Cognition*, 25, 653–660.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1), 204–256.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies.
- Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology*, 2(1), 189–194.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Oxford.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2), 178–210.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380.
- Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2020). *Supervised contrastive learning for pre-trained language model fine-tuning*. ArXiv.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.

- Hebb, D. O. (1949). *The organization of behavior*. Wiley & Sons.
- Hyönä, J., Lorch Jr, R. F., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94(1), 44–55.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554.
- Joseph, H. S. S. L., & Liversedge, S. P. (2013). Children's and adults' on-line processing of syntactically ambiguous sentences during reading. *PLoS ONE*, 8(1), e54141.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review*, 95(2), 163–182.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394.
- 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.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2), 203–208.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, 88(5), 375–407.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5, 115–133.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. ArXiv.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.
- Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49(3), 197–237.
- Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806.
- Ouerchefani, R., Ouerchefani, N., Rejeb, M. R. B., & Le Gall, D. (2024). Pragmatic language comprehension: Role of theory of mind, executive functions, and the prefrontal cortex. *Neuropsychologia*, 194, 108756.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading*. Psychology Press.
- Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. MIT Press.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4), 523–568.
- Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Z., Liu, & Catanzaro, B. (2022). *Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model*. ArXiv.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., L., Kaiser, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Zhang, C., Peng, B., Sun, X., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Fei, C., Yin, C. H., Yan, L. K., & Wang, T. (2024). *From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models*. ArXiv.

The Historical Origins of Large Language Models and Psychology

Huang Linjieqiong¹, Zhang Wen^{1,2}, Chen Zhen^{1,2}, Li Chenxi^{1,2}, Li Xingshan^{1,2}

(¹ Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101)

(² Department of Psychology, University of Chinese Academy of Sciences, Beijing, 101408)

Abstract In recent years, large language models (LLMs) have made significant advancements. Through deep learning, LLMs have learned from vast amounts of human language data and demonstrated human-like language understanding and generation abilities. Through techniques such as supervised fine-tuning and reinforcement learning, LLMs can handle a variety of human tasks and generate text according to human intentions, marking a major breakthrough in the field of artificial intelligence (AI). This paper reviews the development of LLMs, demonstrates their historical roots in psychology, and highlights the critical role of interdisciplinary collaboration, offering insights for future research at the intersection of AI and psychology.

First, psychologists have played a foundational role in the development of artificial neural networks—the backbone of LLMs. Early neuropsychologists such as Donald Hebb and psychologist Frank Rosenblatt focused on learning mechanisms within neural systems, thereby laying the groundwork for machine learning. Long before the deep learning era, psychologists extensively used artificial neural networks to model human cognition. Researchers such as James L. McClelland and David E. Rumelhart continuously refined network architectures to simulate language processing, fostering deep integration between psychology and artificial neural networks. These contributions provided essential theoretical and methodological foundations for the development of LLMs.

Second, the technique of word embeddings is central for enabling LLMs to understand language, and its development has benefited from interdisciplinary collaboration among psychology, linguistics, and computer science. Word embedding technique enables abstract language to be transformed into a form that computers can understand and process. Early psychological and linguistic research introduced the concept of distributed representations of lexical semantics and developed initial quantitative methods. Psychologists later used large-scale corpora to construct high-dimensional semantic vectors, advancing semantic representation techniques. Computer scientists, building on this foundation, implemented these ideas via neural network-based embedding techniques capable of capturing contextual meaning. The evolution of lexical semantic representation methods has facilitated the development of word embedding techniques, enabling the rapid and efficient processing of massive text corpora and contributing to major breakthroughs in language-related AI.

Third, the algorithms of LLMs and cognitive mechanisms of human language processing share several key characteristics, mainly in terms of incremental processing, predictive ability, and dynamic attention allocation. Although the real-time processing, active prediction, and selective attention mechanisms shaped by human biological evolution differ in specifics from the autoregressive generation, masked prediction, and self-attention mechanisms used by LLMs, they exhibit a high degree of functional convergence. This convergence highlights the crucial role of language itself in the development of AI. The deep analogy between the two suggests that understanding the fundamental principles of language may be a vital pathway to achieving general intelligence. Therefore, psychological research into language processing mechanisms could provide essential theoretical foundations and practical guidance for the future development of AI.

Finally, the development of LLMs has greatly benefited from interdisciplinary collaboration and academic legacy among researchers in fields such as psychology and computer science. Whether in the advancement of artificial neural networks or the evolution of word embedding techniques, progress has not only relied on technical innovations from AI researchers, but also on crucial theoretical support and methodological insights from scholars in psychology and cognitive science. Moreover, these researchers have trained a new generation of scientists who continue to explore the development of AI and LLMs. The intersection and academic inheritance between disciplines like computer science and psychology have become a driving force behind AI research. Thus, the continuous efforts by generations of scientists in the field of language cognition are one of the key reasons AI has achieved breakthroughs in this domain.

In conclusion, the major breakthroughs of AI within the field of language processing are not accidental, but rather the result of long-term accumulation and collaborative efforts across multiple disciplines including computer science and psychology. This developmental trajectory offers an important insight for researchers: deep interdisciplinary collaboration and long-term accumulation in fundamental disciplines are crucial. Future research should emphasize on deep cross-disciplinary cooperation and fundamental studies, providing essential support for theoretical innovation, technological advancement, and critical breakthroughs.

Key words large language models, psychology, interdisciplinary cooperation, language cognition, artificial neural networks