大数据·人工智能·区块链研究（二十五）

诠释学视域下GPT语言模型的本质及特征

刘伟伟

（山西大学 哲学社会学学院，山西 太原 030006）

摘　要：GPT语言模型的设计思路具有自然语言理解的诠释学思维特征，但本质上该模型并不具备诠释学语言理解的属人性基础；GPT语言模型将智能视为一种本体论层面以语言作为媒介的整体性系统“涌现”结果，缺乏诠释的“本体论—主体性”地位和“本体论—整体性”结构；GPT语言模型采用的生成式和预训练的设计思路凸显了诠释学理解和解释的“历史性”特征，而数据训练的强化和“思维链”的对话机制使该模型具有“效果历史”和“视域融合”的语言理解特征；GPT语言模型在模拟人类“偏见性”认知方面取得进步，但与人类的“偏见—个性化”和“偏见—创造性”能力相比存在根本差异；GPT语言模型形成了“诠释学循环”的语言对话机制，但其并不具有自身独立的“诠释学循环”实践基础，因此，难以达成诠释学意义上的语言理解共识。

关键词：诠释学；人工智能；理解；解释；自然语言

中图分类号：G252.17；TP18　　**文献标识码：**A　　**文章编号：**1005-9245（2024）03-0066-07

以语言为主线和基础的当代诠释学研究为人工智能的自然语言处理提供了哲学层面的有力支撑，而技术层面的人工智能自然语言处理问题研究为扩展和深化哲学层面的诠释学思想体系提供了丰富的思想养料，“诠释学的洞见获得了人工智能共同体越来越多的关注”[①]。在这一背景下，GPT语言模型（Generative Pre-trained Transformer）的提出、应用及不断发展，不仅在诸多方面契合当代诠释学的思想原则与精神实质，而且为进一步推动“哲学—诠释学”与“技术—人工智能”的跨学科对话和交流提供了可资借鉴的思想启发。

一、GPT语言模型的诠释学语言基础

语言是诠释学的基础，“诠释学经验的普遍性建立在语言基础之上，语言在现代诠释学中发挥了关键的作用”[②]。德国哲学家汉斯-格奥尔格·伽达默尔（H.Gadamer）强调“以语言为主线的诠释学本体论转向”[③]，意在表明诠释学的本体论并非纯粹抽象的形而上学研究，而是建立在语言的理解和解释基础上——诠释学将语言与人的存在相关联，语言由此成为一种世界观。诠释学理解的语言是指人类使用的自然语言而非形式语言，自然语言承载的关

收稿日期：2023-06-10

基金项目：本文系教育部青年基金项目“诠释学视域下数学证明的构造机理研究”（22YJC720008）、国家社科基金一般项目“知识论的逻辑基础研究”（21BZX105）的阶段性成果。

作者简介：刘伟伟，山西大学哲学社会学学院教授、博士生导师。

① S.C.Shapiro.Encyclopedia of Artificial Intelligence，Hoboken：Wiley Publishing Group，1990：363.

② H.Ruthrof. The Roots of Hermeneutics in Kant’s Reflective-Teleological Judgment，Switzerland：Springer，2023：93.

③ [德]汉斯-格奥尔格·伽达默尔：《真理与方法——哲学诠释学的基本特征》，洪汉鼎译，北京：商务印书馆，2010年版，第510、537页。

于人的存在的文化、社会、心理和历史等要素，是诠释学关于理解和解释的核心内容，“诠释学以自然语言、日常语言为领地，洞见到它们是比科学语言更加根本的元语言”①。在诠释学的理解和解释方面，自然语言处理集中体现了人类智能拥有的理解和解释能力的“度”，这使其在一定程度上成为衡量人工智能发展水平与层次的“风向标”。因此，如何以计算的方式将人类使用的自然语言转换为人工智能计算程序能够使用的语言类型，成为人工智能高阶发展的基石，唯有如此才能实现更为理想的人机对话和人机融合。GPT 语言模型的魅力在于其将人工智能系统对自然语言处理的能力向前推进了一大步。以往的人工智能语言模型构造如伽达默尔所言，是一种“关于理解的技艺学”，这种“技艺学”本质上希望建立一种规范的、静态的语言理解和解释系统。针对这一偏见，伽达默尔认为，“我并不想炮制一套规则体系来描述甚或指导精神科学的方法论程序”②。

需要指出的是，在将哲学诠释学的“理解”与“解释”迁移到人工智能语言模型设计思路的过程中，有必要保持一种审慎的态度。诠释学本身是从人的存在出发展开的理解和解释，这种理解和解释的主体是人，人是兼具生物与社会双重属性的特殊存在。据此，GPT 语言模型在根本上与人的语言禀赋和能力的内在机制差距甚大，其缺乏以观察、体验和交往为媒介的人类自然语言运行的基础，这一点使 GPT 语言模型的理解和解释具有一种隐喻和类比意义。当人类惊诧 GPT 语言模型在与人类进行文本对话过程中展现的突出能力时，诠释学思想带给人类的启发在于，GPT 语言模型的自然语言“理解”并不意味着该模型本身具备了类人的“语言—思维”活动机制，而是指 GPT 语言模型在处理自然语言以及在与人对话过程中呈现出行为和现象层面类似人类的自然语言使用特征，但并不意味着该模型具有诠释学语言的内在理解属性。如前文所述，基于诠释学的考量，GPT 语言模型在对自然语言进行理解和解释的过程中，必然会模拟和借鉴人类对于自然语言的学习、使用习惯以及人类关于自然语言的思维机制。这种模拟和借鉴在以二进制为基础的人工智能计算系统中获得了“革命性”的影响力，这种“革命性”主要是就其完全异于人类理解和解释基础的计算系统本身而言的，“GPT 语言模型并不具有一种真正的自然语言理解”③。尽管如此，我们仍要肯定 GPT 语言模型对人工智能发展具有的重大意义。

一方面，GPT 语言模型的自然语言理解能力的形成和塑造并非遵循自上而下的设计思路，而是如同人类自然语言理解经历了漫长、复杂的迭代发展过程。例如，GPT 语言模型基于海量数据供给、利用特定算法对语言数据进行特征标记，进而发现和把握语言数据之间的内在规律，在特征标记过程中隐含的算法层次越多，对自然语言数据的识别能力就越强；GPT 语言模型在大规模自然语言真实文本的系统训练方面取得了长足进展，使其在语用层面能更好地实现不同语境中概念的内涵把握、语词歧义甄别、隐语成分补充等目标；GPT 语言模型建立了较为理想的 LLM（Large Language Model，大型语言模型）交互接口，这一接口能够在一定程度上满足人类的自然语言指令要求，无需人类适应其指令表达习惯，上述进展使 GPT 语言模型的诠释学理解能力获得空前提升。

另一方面，GPT 语言模型的自然语言文本“解释—创造”能力实现大幅提升，离不开作为支撑的大规模预训练数据在量和质方面的突破。事实上，在诠释学领域，自然语言的解释作为语言的“外在表达”根植于本体论意义上人的存在——诠释学意义上的理解与解释相互结合，理解的展开具有解释的参与，而解释是理解的外在展现。对于 GPT 语言模型而言，诠释学意义上的理解表现为其对文本语言的加工和处理，诠释学意义上的解释通过其以自然语言为形式的结果生成和输出得以表征。例如，GPT 语言模型在“生成文本”语法的规范性、语义的明晰性等方面进步明显，表明其“对话—解释”的自然语言语用模型取得了很大进步；GPT 语言模型的“即时学习能力”（In-context Learning）使其在自然语言文本解释的语境性、时效性和准确性等方面的优势得以凸显；GPT 语言模型可以实现对话文本解释的连续性和持续性，并且

① 牛文君：《诠释学与社会科学的逻辑 —— 重思哈贝马斯和伽达默尔之争》，《社会科学》，2022 年第 6 期。

② [德]汉斯-格奥尔格 · 伽达默尔：《真理与方法 —— 哲学诠释学的基本特征》，洪汉鼎译，北京：商务印书馆，2010 年版，第 2 页。

③ Min Zhang，Juntao Li. A Commentary of GPT-3 in MIT Technology Review 2021，Fundamental Research，2021（6）：832.

拥有“思维链”（Chain-of-thought）的复杂推理能力。上述特征充分表明，GPT 语言模型对自然语言的解释与以往的人工智能语言模型相比具有“分水岭”意义。值得注意的是，这种解释能力是对于模仿层面而言的，GPT 语言模型解释能力背后依托的仍是机器系统的计算符号表征，对语言的熟练使用并不代表 GPT 语言模型具有人类的语言思考能力。

二、GPT 语言模型的诠释学本体论立场

从当代诠释学视域看，理解的本体论与语言的本体论相互交融，理解并非仅是一种“能力”，还与人的存在紧密结合，这使本体论的诠释学超越了客观主义的诠释学方法论立场。与之类似，经历了 20 世纪 70−80 年代人工智能研究的“寒冬”后，科学界认识到如不从本体论出发考察自然语言的形成、构造及其使用机制，仅以方法论路线（以符号主义为代表）为圭臬，难以真正解决人工智能的自然语言理解和解释问题。因此，GPT 语言模型将自然语言处理视为一个复杂的巨系统任务，并将其与云计算、机器学习、大数据以及知识谱系构造等结合，为学术界从诠释学视角出发考察人工智能的自然语言本体论提供了前提。

第一，GPT 语言模型的诠释学“本体论—智能性”涌现。一方面，在本体论诠释学领域，自然语言的意义并非认识论层面的发现或构造，而是一种本体论层面理解的意义现象学显现，“本体论诠释学的主要特征之一在于自然语言的意义是非决定性的”[①]。与之类似，GPT 语言模型是在本体论“构造”的层面涌现（Emergence）其“智能”。具体而言，GPT 语言模型采用“大数据 + 大算力 + 强算法 = 大模型”的设计思路，其本质是在语言的本体论层面尽最大可能创造出语境“可能性”的空间，进而为基于自然语言的智能“涌现”提供条件和基础。另一方面，诠释学由方法论上升到本体论的动因之一在于打破了二元论的世界观，重建整体性的世界观。本体这一概念意味着诠释学注重整体意义的揭示，而方法论的单向度演绎将导致理解对象脱离理解的世界整体基础，即远离语言理解与解释的开放性、历史性和相对性，难以实现诠释活动的总体目标。近年来，人工智能的语言模型经历了从概率预测模型到“Transformer”预训练语言模型再到“大语言模型”的转变，转变原因在于当语言模型较小时，其智能水平和计量参数间存在程度上的正向相关性，但大模型的提出可以突破统计规律性，产生超越常规的指令任务完成能力，即涌现更高层次的智能。换言之，GPT 语言模型的智能是一种信息处理不同单元、结构间整体协同配合的演化结果——语言数据的累积为智能的涌现提供前提，局部结构和局部功能的改进与突破为智能的涌现提供可能性，“言语认知和自然语言处理，能够从产出类人‘输出’的大语言模型中涌现出来”[②]。这一点符合诠释学关于语言理解“本体论—整体性”的立场，这一整体性立场为 GPT 语言模型自然语言处理的智能性“涌现”奠定了基础。

第二，GPT 语言模型的诠释学“本体论—主体性”反思。本体论诠释学认为，理解与人的存在密不可分，使作为文本语言理解的主体成为理解活动的核心。GPT 语言模型作为一种人机自然语言对话系统，涉及人与机器的主客体关系问题。以往机器语言的对话系统之所以在“智能”层面不尽人意，根本原因在于机器系统并非如同人类的自然意义上的主体存在，这一点使机器系统的“理解”无法与其自身的存在相关联。因此，在这种条件下的机器系统对于自然语言的处理是有局限的、单维的，缺乏语言主体面向交往语境要素的开放性、包容性。GPT 语言模型力求模拟和实现语言诠释主体的运行机制和行为效果，主要表现在 GPT 语言模型能够不断学习和模拟人类诠释主体的抽象思维能力。例如，GPT 语言模型可以进行任务泛化，即模型可以根据指令生成新的答案。尽管如此，GPT 语言模型在诠释学的“本体论—主体性”层面仍缺乏理解的主体特异性，其本身并不具有任何情绪或意向的主体性，其呈现的中立性和客观性只是人为施加的算法规则参数调整，具体而言：一是 GPT 语言模型缺乏诠释的“本体论—主体性”作用机制，没有人类语言诠释的因果超越性和跨时空

① A.K.Gangadean. Between Worlds—The Emergence of Global Reason，Pieterlen：Peter Lang International Academic Publishing Group，1998：94.

② Mathias Risse. Political Theory of the Digital Age：Where Artificial Intelligence Might Take Us，Cambridge：Cambridge University Press，2023：x.

性，其表面看似创造性的文本输出结果仍依赖计算的概率统计和分析。例如，GPT 语言模型基于大规模反馈式训练和语词频率分析挖掘语义关联的基本思路具有行为主义的明显特征，无法复现人类主体思维的抽象运行机制，从而使人工矫正和人工反馈成为 GPT 语言模型不可摆脱的桎梏。二是 GPT 语言模型采用经验论的数据整合模式，难以实时把握自然语言使用主体的心理、文化、思维类型等语境信息，缺乏诠释学理解的“本体论—整体性”结构特征，而人类能够从自身存在出发，依据理解的整体语境迅速展开对特定认知对象的判断。三是 GPT 语言模型缺乏完整的诠释学本体论基础架构，而理性与非理性、规则与非规则是基于人的存在的诠释学本体论基本内涵。尽管 GPT 语言模型在现有技术条件下已接近人类理性思维的极限，但其以理性扩张弥补非理性缺陷的技术化“图腾”与人类基于现实交互、感性与理性相统一的智能表征机制仍相去甚远。

三、GPT 语言模型的诠释学历史原则

诠释学认为，文本语言的理解和解释具有历史性，读者对文本的理解是重铸文本历史的过程，这使文本语言理解的孤立性、静止性和片面性在诠释学领域得到一种历史的、动态的、发展的思维模式拓展。GPT 语言模型的结构“塑造”在一定程度上彰显了历史性维度对于对话语言文本生成的重要价值所在，而如何超越以往记忆空白的、没有思维连贯性的、缺乏对话逻辑组合能力的人工智能模型，是 GPT 语言模型研发者技术创新的初衷所在。

第一，GPT 语言模型的“效果历史”作用。“效果历史”是伽达默尔诠释学思想中的一个重要概念。伽达默尔认为，文本不是脱离历史的文本，而是建立在历史理解链条基础之上的文本——理解者对文本语言的理解是一种参与文本意义历史创造的过程，是文本与理解者主客交融的历史性“时间间距”发挥作用的过程，“效果历史意识具有对传统的开放性”[①]。具体而言，任何理解在诠释学中都是建立在前理解基础上的活动，文本理解的目的不是去除这种前理解，相反，前理解构成人的理解活动的积极力量。对于 GPT 语言模型而言，模拟并塑造文本语言理解的“效果历史”作用主要体现在三方面。一是 GPT 概念中蕴含的生成式和预训练概念内涵具有语言理解的历史性、动态性和即时性特征。GPT 语言模型对于作为自然语言历史性理解规范的数据参数量具有依赖性。据测算，GPT-4 的数据参数量在 1750 亿 −2800 亿间，这使其具备了在不同场景、不同任务目标下的“时—空”语境适应性——GPT 语言模型的语境长度（Context Length）可以达到上万的标记（Token），从而使 GPT 语言模型在诠释学理解的意义上能够获得更强大的“历史性—前理解”语言能力训练保障。二是 GPT 语言模型对语言历史性“记忆—前理解”的把握能力得到加强，能够以更加贴近使用者目的的方式展开语言对话反馈，提高了使用者交流体验的满意度。特别是在 GPT 语言模型联网上线后，海量增加的用户对话数据将作为一种历史性语料的积累为其进一步完善发挥作用。三是 GPT 语言模型的人机自然语言对话博弈产生了语言理解历史性“视域融合”的现实结果。“视域融合”意味着人的理解的开放性和可扩展性，前理解与理解相互交织，理解的视域并不与历史矛盾，而是连接过去、现在和未来，“诠释学以一种过程性的思维来看待视域融合”[②]。这一点在 GPT 语言模型的对话“思维链”机制中得到较充分的体现——GPT 语言模型能在一定程度上基于过去的对话内容形成新的对话回应，极大地避免了机器语言输出的非连贯性，这种融合对话双方视域的“思维链”机制将取得更佳的对话交流效果，从而使 GPT 语言模型展现出超越以往类人机器的智能水平。

第二，GPT 语言模型的历史性理解“偏见”。文本语言的理解“偏见”是诠释学的一个基础概念，也是诠释学历史性的一种具体表现——诠释者对文本的理解是处于一定历史时空范围内与文本之间互动的结果，这种互动并非抽象、无限制的，而是建立在诠释者特定的思维结构前提下，这种前置的思维结构即诠释学所谓的偏见，“由于效果历史的作用，每一个理解者都有他自己的偏见”[③]。基于

① [德]汉斯-格奥尔格·伽达默尔：《真理与方法——哲学诠释学的基本特征》，洪汉鼎译，北京：商务印书馆，2010年版，第510页。

② Paul S. Chung. The Hermeneutical Self and an Ethical Difference，London：James Clarke & Company Limited，2012：65.

③ B.H.McLean. Biblical Interpretation and Philosophical Hermeneutics，Cambridge：Cambridge University Press，2012：184.

对人类自然语言使用偏见能力和机制的模仿，GPT 语言模型从两方面进行了改进。一是 GPT 语言模型将机器自主学习与监督学习相结合，在模拟人类“偏见性”的理性认知方面取得了进步。例如，针对空问题或假问题，原有的人工智能系统由于未内置意义判断的相关参数，有可能出现合逻辑但反事实的问题输出结果，GPT 语言模型由于采用人为特征标记策略，其对话生成的纠错性和优化性能力得到较大幅度提升。二是 GPT 语言模型在模拟人类社会伦理规则方面的“偏见性”认知能力得到提高。例如，GPT 语言模型通过数据训练增加了人为的价值观介入与干预，在人类反馈强化（Reinforcement Learning from Human Feedback）等技术手段的支撑下，最终的文本输出结果可以在“真 / 假”“有害 / 无害”以及“有用 / 无用”等考量指标方面获得增益。

如前文所述，“偏见”在诠释学中并不意味着人类思维的先天缺陷，理解者与解释者由于“偏见”具有思想的特殊性与创造性，使文本的理解和解释成为充分彰显人的主体性与能动性的活动。尽管如此，GPT 语言模型在诠释学意义上并不具有人类语言理解和解释的“偏见”独特性，其主要原因有三点。第一，虽然 GPT 语言模型的创造者希望这一模型能够以差异化、动态生成的方式灵活调整和应对每一项参与界面对话用户的语言指令任务，但其基于普遍中立知识的历史记忆库数据训练模式仍使其倾向于提供一种能够为人类通常表达习惯和知识陈述方式普遍接受的观点。因此，人们很难在 GPT 语言模型的对话输出中观察到符合情境要素特征、具有独特语言风格与态度的应答结果。第二，GPT 语言模型在认知层面并不具有与人类同步的“偏见”基础，尽管 GPT 语言模型最新的联网功能使其在语言数据的全域性方面获得极大扩张，但人类作为创造历史记忆的生命体，会展现每个人独特倾向的个性，这是人类个体能够在历史传承过程中具有创造性的根本动能。从本质看，GPT 语言模型并不具备这种“偏见—个性化”潜力，导致其“偏见”认知较容易带来真正的偏见，并且容易使“人—机”在达成语言交往的共识理解时出现障碍。第三，GPT 语言模型的“偏见”认知机制在强调认知特异性的同时，仍在可靠性和真实性方面存在极大不确定性，其使用的隐性神经网络在可信度和可解释性方面存疑，并且其能力涌现缺乏事实验证。目前，GPT 语言模型仍是一种语言“黑箱”运行机制，同时，其在智能通用性定位上更加偏向英美文化背景，与英美文化背景下的对话贴合度更高，在多民族语言材料兼容性方面较为薄弱。

四、GPT 语言模型的“诠释学循环”

“诠释学循环”是诠释学理解与解释过程中的普遍现象——从诠释学看，语言文本的部分与整体、一般与特殊之间的关系并不是绝对的，而是相互依存，可以在彼此间进行相互转换，“从个别理解整体并从整体理解个别这一诠释学原则……转变为理解的艺术”[①]。从人工智能建构的诠释学意义出发进行考察，GPT 语言模型的贡献在于其脱离了孤立且缺乏连续性的语言模型底层建构逻辑，将语言的输入与输出视为循环往复的博弈过程，“对解释意义的阐释涉及到了不可避免的解释学循环”[②]。人工智能着重模仿人类智能的一个关键点在于智能主体能够与“他者”之间建立起有效、持续且往复的语言交往关系——语言交往关系的动态循环性是人类智能生成、演化和表征的重要标志，也是 GPT 语言模型不断改进和创新的重要目标。

第一，GPT 语言模型的“诠释学循环”构造基础。“诠释学循环”是 GPT 语言模型获得人们对其语言理解能力认可的重要依据——但它与人类的理解相比仍相去甚远。GPT 语言模型之所以能够产生高质量的自然语言对话输出，主要源于其自我监督的概率统计和预测功能，这使其能不断学习语词之间以及句法结构之间的内在关联——基于自然语言的表达方式和习惯学习，以“诠释学循环”的方式应用到自然语言输入的分析和自然语言输出的构造过程中，这种循环性表现在该模型的语言对话机制方面并非一次性、单向性的，而是具有连续性和整体性的。一是 GPT 语言模型在实现“诠释学循环”的过程中，预训练数据的选择具有至关重要的作用，高质量的预训练数据能引导高质量类人语言表征能力的产生，低质量的预训练数据则会产出背离人类自然语言表达习惯和表达语法的文本，由此语言数据的预训练成为 GPT 语言模型建立“诠

① [德]汉斯-格奥尔格· 伽达默尔：《诠释学II：真理与方法》，洪汉鼎译，北京：商务印书馆，2007年版，第70页。

② 郭贵春：《后现代科学实在论》，北京：知识出版社，1995年版，第20页。

释学循环”的重要起点。二是GPT语言模型未将其“自身”视为该模型与用户之间交流的孤立要素，而是在一种循环的动态整体机制中判断和定位语言文本的输入与输出，这使该语言模型获得了一种类人语言理解和解释的“诠释学循环”定位。三是GPT语言模型采用的自然语言机器学习范式超越了严格规则的编程模型，并在一定程度上形成基于案例模仿的“诠释学循环”意义上的泛化理解能力。尽管这种能力很多时候缺乏可解释性，但这与人类超规则的复杂意识活动类似，进而使GPT语言模型相对非确定性的自然语言理解能力得到较大提升。

第二，GPT语言模型的“诠释学循环”对话机制。诠释作为一种对话的艺术，不仅需要与文本作者展开对话，而且需要与他者展开对话。因此，诠释是一种“你—我—他”相互交织的有机对话系统。对GPT语言模型而言，其语言对话机制呈现显著的“诠释学循环”特征。一是GPT语言模型的对话系统在提问与应答的逻辑进程中可以不断拓展和延伸，在此条件下的问答环节与过程形成了诠释学循环意义上的对话场景。例如，GPT语言模型的人类反馈强化学习机制可以根据人类喜好进行对话的概率优化，在不断循环的对话中训练学习模型的适应性；指令微调（Instruction Fine-tuning）被用来模拟自然语言的对话丰富性和变化性，能在语言模型微调的基础上降低机器对话的错误率。二是GPT语言模型的创造性能力来自诠释学“整体性—循环性”的进一步强化，这使其在人机互动的质量方面得到显著提升——在此过程中，使用者的真实操作与语言模型的调整和完善间建立了紧密关联，从而使过去机器语言系统的弱诠释学意义迎来向强诠释学意义转变的曙光。例如，GPT语言模型通过海量无标注数据增强学习模型的常识训练，继而展开有标注数据的“Fine-tune”训练，这种半监督学习使机器系统基于“诠释学循环”的语境理解能力得到极大提高。

尽管如此，GPT语言模型的解释能力在对人类自然语言使用进行模仿的过程中，其表象的类人语言对话逼真性和逻辑性无法掩盖由于本体论意义上的存在缺位带来的语言解释失真性和非完整性的显著缺陷——GPT语言模型的“解释—创造”能力并非无懈可击，例如，GPT语言模型的“自注意”机制使“文本—解释”的自洽性得到提高，但其仍缺乏人类的情感思维和因果推理能力——当GPT语言模型在处理部分复杂语言指令及多模态语言任务时，经常出现反常识的情况——就诠释学意义而言，理解和解释是构成“诠释学循环”的必要环节，二者缺一不可，且这种循环性是动态的、发展的。由于GPT语言模型在自然语言解释能力方面存在先天不足，所以其在支撑人工智能的“通用性”目标实现方面任重道远。

第三，GPT语言模型“诠释学循环”的实践误区。GPT语言模型建立的自洽的语言对话机制并不能在实践层面最终达成人类自然语言理解和解释的“诠释学循环”目标。事实上，语言作为一种能力是人类与他人和世界在实践交往基础上形成的诸多能力之一，这意味着语言是人类存在的表现（Performance），而非存在的本质——判断人存在的本质离不开人类的实践活动。从诠释学循环的整体论立场看，理解、解释和实践是完整的有机循环，这表明诠释的基础和媒介虽然是语言，但诠释的目标和动力源于实践，实践是诠释学语言循环的根本遵循，“解释的实在性在于它的实践性”[①]。由此观之，GPT语言模型的形成和演化并不具有自身独立的“诠释学循环”实践基础。例如，GPT语言模型虽希望在模型优化的基础上借助“奖励”程序贯彻“有益性”“诚实性”和“无害性”等诠释学意义层面的实践要求，但上述实践要求缺乏相应的实践检验作为回应。根源在于GPT语言模型并不具备人类独立的语言实践活动能力。因此，难以完成“诠释学循环”的完整“拼图”——由于语言实践活动的缺失，导致GPT语言模型与用户间并不存在真正意义上的语言交流理解共识，而界面语言互动也不是一种类似人类主体基于实践的“诠释学语言循环”关系。由此可见，GPT语言模型强调的只是一种经验主义的说明，缺乏真正的诠释学理解，因为诠释学理解不仅是对文本语义的一种简单说明，还包括对文本意义的某种创造——“创造性”的语言理解具有“自我”属性，应建立在交流双方自我理解的基础上并且形成循环的语言对话机制。

五、结　语

基于语言理解和解释的诠释学研究，为人们

① 郭贵春：《后现代科学实在论》，北京：知识出版社，1995年版，第20页。

更加全面地看待 GPT 语言模型的本质和特征提供了哲学层面的视角。在人工智能飞跃式发展的时代背景下，人们清晰地认识到语言问题并不仅仅是语言学问题，也不仅仅是哲学层面的抽象概念问题，而是基于人脑进化系统的复杂认知问题，且这种认知源于特定的文化与历史语境——上述基本要素共同构成关于语言理解和解释的诠释学整体语境，“诠释学能够为理解自然语言或者表征社会世界知识的人工智能系统设计指明边界、方向甚至于标准”①。由此而言，尽管目前人类并不认为 GPT 语言模型是人工智能未来发展的终极方案，但这种技术性的实践探索至少为多维度、深层次地把握人类语言的本质和意义提供了更加丰富的养料。诠释学在哲学层面的抽象思辨，连接关于人类语言理解的“元理论”基础，对于这种语言基础的考察能使人们在人工智能语言模型的技术设计与规划过程中更好地沿着类人语言认知的路径向前发展，避免陷入纯粹形式主义和逻辑主义的非人类语言模型构建误区，打通诠释学语言理解、解释与人工智能语言模型构建间的隔绝屏障，营造两者良性互动、协同共进的全新格局，从而为人工智能的未来发展开辟新的进路。

The Essence and Characteristics of GPT Language Model from Hermeneutical Perspective

LIU Wei-wei

（School of Philosophy and Sociology，Shanxi University，Taiyuan Shanxi 030006）

Abstract: The design idea of the GPT language model embodies the hermeneutic thinking characteristics of natural language understanding，but in essence，the model does not have the humanistic basis of hermeneutic language “understanding”. The GPT language model regards “intelligence” as the result of the emergence of a holistic system using language as a medium at the ontological level，but the model still lacks “ontology-subjectivity” status and “ontology-integrity” structure in the hermeneutic sense. The “generative” and “pre-training” design ideas adopted in the GPT language model highlight the “historical” characteristics of hermeneutic understanding and interpretation，while the strengthening of data training and the dialogue mechanism of “thought chain” make the model possess the “understanding” characteristics of “effective history” and “fusion of horizons”. The GPT language model has made some progress in simulating human “biased” cognition，but there are still fundamental differences compared to human “biased personalization” and “biased creativity” abilities. The GPT language model forms a linguistic “dialogue” mechanism of “hermeneutic cycle”，but it does not have its own independent practical basis of “hermeneutic cycle”，making it difficult to reach a “consensus” of language understanding in the hermeneutic sense.

Key words: Hermeneutics；Artificial Intelligence；Understanding；Dialogue；Natural Language

[责任编辑：曹晶晶]
[责任校对：李　蕾]

① S.C.Shapiro. Encyclopedia of Artificial Intelligence，Hoboken：Wiley Publishing Group，1990：374.