

生成式人工智能在政务回应中的效能评估 ——基于“图灵测试”与文本分析对政府留言板回复的实证研究*

傅承哲^{①②} 莫庭阳^{**①} 王楚暖^①

①华南师范大学政治与公共管理学院

②华南师范大学粤港澳大湾区跨域治理与公共政策研究中心

摘要: 随着自然语言处理技术和深度学习算法的突破,生成式人工智能已具备对人类语言、行为和思想观念的深刻理解能力,并在政策信息检索、团队冲突解决等场景中展现出超越人类专家的潜力,但政务聊天机器人在实际应用中仍面临技术局限性,亟需系统性评估其能力。研究旨在聚焦大语言模型在公共管理领域的应用效能,构建人工智能赋能政务回应效能评估分析框架,分别通过“图灵测试”与文本质量评估,系统探讨大模型在政民互动中的实践潜力与局限性。首先开展图灵测试,选取人民网领导留言板的咨询、建议、投诉类民众留言,同时调用API要求DeepSeek模拟公务员批量生成回复文本,并通过问卷调查收集用户满意度评价,开展“图灵测试”;其次,基于文本长度、表述逻辑、情感倾向、回应相关性等客观指标,对公务员与大模型回复文本质量进行量化评估比较,揭示不同官员职务层次对回复文本特征的影响。研究发现,通用大模型对民众留言的回复功能已经通过“图灵测试”,大模型生成回复的满意度显著高于公务员回复文本,其在简单政民互动中的回复能力已经初步达到人类水平。相较于公务员回复,通用大模型回复文本具有诸多特征,并在文本长度、语言表述逻辑等方面具有明显优势,但在情感倾向与回应性方面仍有不足,同时在代表不同层级官员回应时存在无差异回复的特征,说明其未能捕捉中国制度下领导职务层级与话语风格的关联,暴露出技术逻辑与行政逻辑的脱嵌问题。最后为发挥生成式人工智能对公共服务供给的提升效能提出若干建议。

关键词: 政民互动;生成式人工智能;大语言模型;DeepSeek;图灵测试;文本分析;领导留言板

DOI: 10.16582/j.cnki.dzzw.2025.08.008

一、引言

随着自然语言处理、深度学习算法等基础技术的飞速发展,基于Transformer架构的大语言模型因其大量预训练基础,获得有效地理解和处理复杂的上下文信息的语言能力。生成式人工智能本质上是一种概率性文本生成系统,通过对海量数据的统计学习形成语义表征与逻辑推理能力^[1],不仅具备强大的文本生成和理解能力,还能在多个领域展现出令人瞩目的应用潜力。依据上述技术基础,计算机科学和社会科学领域学者在探索生成式人工智能的社会科学研究做出了许多尝试,有研究发现,大模型对政治新闻文本的编码能力已经与人类专家相似^[2]。在组织团队冲突情境下,相较于一般民众,大模型在选择解决方案方面与团队管理专家更为相似,说明大模型在复杂社交情境中的应对和决策能力已经超过一般民

众^[3]。在公共管理领域,检索增强生成架构(Retrieval-Augmented Generation, RAG)能够将政策文档知识库与通用大模型联结,提升检索系统的泛化能力和实用性,为用户提供更精准、便捷的政策信息获取服务^[4]。

近年来,大模型在公共管理领域的实践应用得到各方的关注。2023年发布的《生成式人工智能服务管理暂行办法》明确指出:“支持行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等在生成式人工智能技术创新、数据资源建设、转化应用、风险防范等方面开展协作。”《关于加快公共数据资源开发利用的意见》明确提出,要“支持人工智能政务服务大模型开发、训练和应用,提高公共服务和社会治理智能化水平”。2024年,国产开源大模型DeepSeek-V3多项评测成绩超越了Qwen2.5-72B和Llama-3.1-405B等其他开源模型,并

*基金项目:广东省哲学社会科学规划2023年度青年理论学术带头人“揭榜挂帅”项目“在融入国家发展大局中促进港澳青年增强国家认同研究”(项目号:GD23XZZY06);广东省自然科学基金青年提升项目“粤港澳大湾区青年国家认同培育政策的虚拟-现实空间双层赋能机制及其行为优化策略研究”(项目号:2024A1515030062)。

**通讯作者

收稿日期:2025-02-28

修回日期:2025-04-18

在性能上和世界顶尖的闭源模型GPT-4o以及Claude-3.5-Sonnet不分伯仲^[5]。2025年1月20日,使用强化学习训练的DeepSeek-R1系列模型正式发布,其独特之处在于以高度压缩的开发成本,实现671B参数全尺寸模型的私有化部署,同时模型输出包含大量反思和验证的推理过程^[6]。基于这一技术突破,2025年2月,深圳市龙岗区政务服务和数据管理局在政务外网部署上线了DeepSeek-R1全尺寸模型。此后,多个地方政府应用和部署基于DeepSeek等大模型的政务服务系统或智能体^[7],在其政务系统中积极接入DeepSeek以提升工作效率。

以本研究前期探索开发的跨境人才政策咨询服务大模型为例,生成式人工智能的发展普及无疑深刻塑造公共管理实践:公共部门及其实务工作者亟需调整自身的研究实践以面对应用大语言模型的趋势。本文以政务咨询领域中的政务聊天机器人“图灵测试”为具体案例,在整合既有文献的基础上提出评估人工智能赋能政务回应效能的两大方面——主观体验以及客观测度。由此,根据生成式人工智能的返回结果评估其模型性能,用大模型的民众留言回复功能进行“图灵测试”,比较人类回复文本与大模型回复文本的满意度和文本质量差异,尝试探讨大模型应用对公共管理的实务工作带来的机遇与挑战。

二、文献回顾

(一) 基于大模型模拟能力的技术赋能政民互动过程

大语言模型(LLMs)凭借强大的推理和规划能力,为模拟人类行为提供了新的机遇。通过角色扮演,大模型驱动的智能体能模拟特定情境下的个体反应,成为研究人类行为的有效工具。与此同时,多智能体的交互被广泛用于解决问题和模拟复杂动态:大模型可被定制为具备特定知识和技能智能体,通过协作实现目标、完成任务;它

们的简单互动也能涌现复杂集体行为,模拟现实中的复杂社会动态。这些模拟为对社会现象的理解与预测提供了新工具,也为政策制定和社会管理提供了支持。Mou等将大语言模型对人类与社会的模拟归纳为三种类型——个体模型、场景模拟和社会模拟(参见图1)^[8]。

个体模型是指通过模拟单一个体的认知、情感与行为模式,构建具有特定人格、知识储备与决策逻辑的虚拟主体的模拟过程。其核心在于还原人类心理特征。场景模拟是指聚焦特定情境下的多主体互动,通过环境变量与动态规则设定,再现真实社会,关注交互过程的因果性与情境适应性。社会模拟是指基于群体行为动力学与复杂系统理论,建模宏观社会现象。通过整合个体行为规则与网络结构特征,揭示社会趋势涌现机制及政策干预效应。三者分别对应微观个体、中观场景与宏观系统的分析层级,共同支撑对社会智能的多尺度解构与推演。

在传统的政民互动过程中,一方面,公共服务与行为输出受公共服务动机与公共决策心理共同决定;另一方面,满意度、政府信任和权威感知合法性等公民认知与情绪体验也对产生公共服务供给产生反馈作用^[9]。而数字技术在治理公共服务供给中的深刻嵌入则重塑政民互动过程,技术应用同时带来行政效率提升与心理负担增加等积极与消极影响,例如接诉即办改革实现数字技术赋能,有助于实现政民良性互动,促进公共价值共创^[10]。另一方面,数字技术与组织结构的“结构性”失衡、数字化公共接触中的“互动性”障碍以及街头官僚与公民的数字化应对能力不足等问题,增加了政民互动中的繁重体验^[11];信息技术在公共部门中的广泛应用,增加了技术数据排斥宽容与同理心、街头官僚与公民心理距离扩大等风险^[12,13]。此外,公共部门数字技术应用可能在政治合法性基础薄弱的特殊群体中引发政府

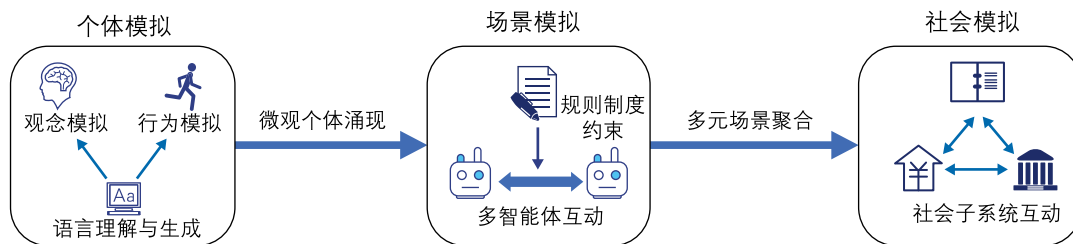


图1 大语言模型的社会模拟能力

权力扩张的数字利维坦想象，引发公共价值冲突等隐患^[14]。基于政民互动等相关领域的研究基础，本研究以大语言模型的模拟能力为技术核心，将技术要素嵌入政民互动过程，构建基于大模型模拟能力的技术赋能政民互动模型（参见图2），从大模型的实践手段与认识工具双重属性出发讨论技术应用下政民互动实践模式和学术研究方法的更新迭代。

在治理实践层面，大模型能够以其自然语言理解和生成功能嵌入等公共服务供给场景，以政务问答机器人等具体形式优化公共服务供给的质量与效率，提升公共服务满意度。生成式人工智能等数字技术可为个体提供定制化的海量知识和信息供给，有利于突破传统媒介形成的信息壁垒，更为全面地了解公共政策问题，进而鼓励和提升其参与行为与能力。而在学术研究层面，研究者或实务工作者首先能够基于多Agent互动框架搭建人工社会，在特定场景中进行政策干预模拟，预测政策实施的效果；同时能够高效接收与处理民众反馈的海量多模态数据，实时监测、深入解析政策干预后的行为和态度反应，为政策制定、优化调整提供科学依据，提升公共政策效能。

（二）政务大模型的技术发展历程、应用情况与现实问题

政务聊天机器人的技术构想并非新鲜事物，早在20世纪90年代，即有学者畅想以自然语言与计算机进行交

互，并断言随着个人计算机的广泛使用以及客户对制造商提供自然语言界面的愿望，公众对于对话机器人（conversational agents）的需求变得越来越迫切^[15]。然而，受限于当时的计算能力和自然语言处理技术，这一构想长期停留在理论阶段。进入21世纪，随着机器学习和深度学习技术的突破，尤其是基于规则的简单问答系统和关键词匹配技术的应用，政务聊天机器人开始在部分公共服务领域试水，主要提供如办证咨询、政策解读等基础服务，然而这些早期系统因回复机械化、语义理解能力有限而难以满足复杂多样的民众需求。

随着生成式人工智能的发展普及，早年计算科学家构想的智能化政务聊天机器人逐渐成为现实：以DeepSeek等为代表的大语言模型凭借强大的上下文理解和文本生成能力，能够模拟人类语言风格，提供更加自然、流畅的交互体验。由此构建的政务大模型通过自然语言处理技术实现的语义理解与知识推理能力，为政务聊天机器人构建了核心技术框架，使其通过对话的形式回应民众诉求，实现政务服务。政务聊天机器人逐渐从简单的问答工具演变为能够处理复杂政务咨询、政策解读甚至情绪安抚的智能助手，以其信息整合优势和自然语言处理算法赋予的语言理解能力，不仅能够提供政策信息咨询等政务服务中降低服务获取难度和成本，还以社交平台私聊等人机互动形式帮助民众跨越政策信息获取和理解的信息鸿沟，提升其政务服务满意度和政策认知程度。然而，受限于算法技术的有限性，聊天机器人仍可能出现服务失败问题^[16]。技术特性^[17]、人机互动方式^[18]以及互动的环境、组织、技术等诸多外部因素^[19]都将影响用户对政务聊天机器人的信任。

在其他各类具体应用方面，政务大模型同样实现了多领域渗透与差异化部署，并深刻塑造着公共治理过程。以行政经验积累为例，政务大模型的突破性发展通过对多模态政务数据进行特征提取与知识凝练，可以形成可迁移的政府治理知识图谱，提高政府治理效能，实现更高质量服务水平，赋能数字政府^[20]和“服务型”政府建设^[21]。以环境影响评价领域为例，杭州、厦门等地方政府采用“DeepSeek-R1+专业增强知识库”的技术架构，成功将环评报告生成周期从传统作业模式的数月

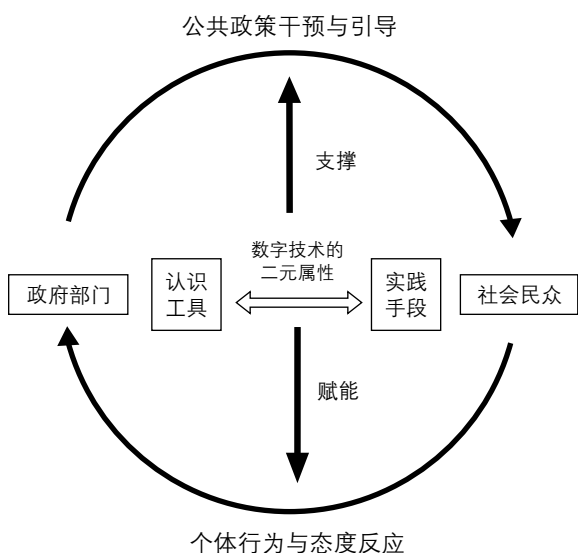


图2 基于LLM模拟能力的技术赋能政民互动模型

量级压缩至30分钟级,并通过空间数据集成实现生态环境准入条件的智能化比对^[7]。此类典型案例揭示了大模型技术在标准化流程、高数据密度场景中的比较优势。然而,其应用边界亦逐步显现:涉及多元价值冲突调解、跨组织协调或社会关系重构的高复杂性治理场景仍需保持人类决策者的主导地位。^[22]例如,基于DeepSeek架构的政务搜索引擎虽显著提升了信息检索效率,但其算法透明度缺陷与生成内容可信度风险可能引致公众认知偏差,需通过分级分类的数据治理框架与人类监督机制实现风险缓释^[23]。说明在新兴技术落地的同时,制度性适配问题随之日益凸显。地方政府在规模化部署过程中面临责任归属模糊、采购评估机制失效及治理架构滞后等系统性挑战,折射出技术工具应用与现有治理制度之间的深层张力。

(三) 政务大模型的服务失败问题与服务质量评估

囿于自然语言理解和处理能力,政务聊天机器人在实践中呈现多种多样的服务失败问题。部分研究尝试通过回复文本的拟人化从侧面缓解其技术问题对用户带来的负面情绪反映^[24]。生成式人工智能(Generative AI)的接入为政务聊天机器人带来了显著的技术升级,使其代表政府机构实时为公众提供各类与政务服务、政策法规和办事程序相关的问答服务^[25]。张朝阳等将预训练语言模型与知识图谱进行有效融合,提升政务信息问答服务的精准度和个性化水平^[26]。在政务咨询场景中,DeepSeek模型在中文场景的卓越理解推理能力有助于根据不同群体的表达习惯给出更有针对性的回应^[27]。

目前,学界缺乏针对政务聊天机器人的政务咨询回复能力开展的系统评估。在已有研究中,用户体验和满意度是衡量客服聊天机器人回复质量的关键因素,满意度的形成机制涉及技术特性、可用性设计及社会认知等多个维度。首先,技术特性(如智能性、拟人性和个性化)通过增强用户对系统的信任间接提升满意度。例如,政务智能聊天机器人的智能性和拟人性不仅直接影响用户的初始信任,还通过提升感知有用性和易用性产生中介效应;而便利性则完全依赖这两者的中介作用。这表明优化机器人的技术特性能够强化用户对服务价值的认可^[17]。其次,可用性设计缺陷可能显著降低客户满意度。Wei和Landay通过

评估语音智能设备发现,系统状态不透明、反馈冗长或缺乏智能确认(如未明确告知操作是否成功)会导致用户困惑和挫败感^[28]。此外,社会认知理论指出,用户对聊天机器人的评价遵循“温暖-能力”双维度模型^[29]。温暖(如友好、同理心的语言)决定用户是否感知到服务的善意,而能力(如解决问题的准确性和速度)影响用户对效率的认可。综上,客服聊天机器人的回复设计需整合技术优化、可用性改进与社会认知策略,通过智能且人性化的交互提升客户满意度。

与此同时,另一文献关注人机互动过程中聊天机器人的回复语言特征对用户满意度的影响:例如使用口语化的表达方式、关注用户的语言特征和情感是评估聊天机器人功能的重要指标^[28]。回复内容的情感与理性平衡显著影响用户感知:高学德和康静在政府回应研究中发现,理性导向和混合导向的回应模式(如结合规则解释与共情话语)比单纯情感导向更能提升公民满意度^[30]。另外,互动过程中的用户自主性同样深刻影响服务效能,在线零售中的人机交互若赋予用户控制感(如自定义搜索选项)和个性化响应(如基于历史行为的推荐),可通过提升自主性间接增强幸福感^[31]。

综上,系统评估政务聊天机器人的政务咨询功能需要结合上述两种分析路径,以回复文本为评估内容,以各类客观指标评估文本特征,对政府回应话语开展类型学划分,同时在分析评估文本质量时兼顾理性与情感、用户自主性及技术可用性等诸多因素,以用户主观感知为政务咨询质量的根本评估依据。因此,为系统评估生成式人工智能应用对政务回应效能的赋能作用,本文整合两种分析路径,以公务员回应为参照,相应提出生成式人工智能赋能政务回应效能评估的两大阶段(参见图3):第一阶段,围绕“人工智能能否赋能”的问题,关注大模型赋能政务回应的主观体验,即政务回应中大模型是否有替代人工回复的可能性,通过问卷形式测量民众对政务回应服务的主体主观感知以及满意度;其次围绕“如何评估人工智能评估效果”的问题,关注大模型赋能的政务回应客观测度,在证明大模型具备主观体验层面上的可替换性后,测度大模型赋能的多个维度,即大模型能够在哪些方面赋能民众理解政务回复文本,



图3 人工智能赋能政务回应效能评估分析框架

以各类具体指标作为评价回复文本的客观测度标准，从而评估回应文本的质量和特点。在操作流程方面，首先以人民网领导留言板的民众留言为素材，要求DeepSeek等生成式人工智能代表地方政府和领导回复民众留言，进而通过表述逻辑、情感倾向等客观文本指标系统比较大模型回复与真实公务员回复文本，捕捉二者的语言特征及其间差异，初步揭示以生成式人工智能作为政务聊天机器人回应民众政务需求的优势与不足。其次，通过发放问卷以“图灵测试”方式评估大语言模型对人类行为与语言的模拟程度，并询问民众对大模型回复与真实公务员回复的满意度，最终揭示影响政务咨询服务供给效能的关键因素。

三、人工智能赋能的主观体验：政务大模型回复民众留言的“图灵测试”

（一）政务大模型回复民众留言的“图灵测试”研究设计

关于政务大模型的问答功能，已有学者尝试自行搭建智能问答模型评估性能。例如，华斌等基于政策文本内涵建立知识模型，并从文本语义相似度计算与知识元计量两个层面对模型问答结果进行质量检验，发现对于事实类问题具有较好的应用效果，可以为事实问答的研究提供新的思路^[32]。而目前政务大模型作为公共服务供给领域的垂直大模型，需要在通用大模型的基础上，通过知识增强、模型微调或者增量预训练来构建。基于通用大模型的技术开发路径为使用通用大模型生成回复文本，以此为评估大模型在政务沟通中的回复功能提供了契机。因此，本研究参考已有研究的“图灵测试”^[33]，评估大语言模型对人类行为与语言的模拟程度，以政务问答为具体实践场景，比较人类回复文本和大模型回复

文本的回复满意度，以探索性研究方式为政务大模型问答功能的相关研究提供启示，并回应图2中大模型作为实践手段自动生成治理工具的积极作用。

“图灵测试”的核心要义在于通过人机对话方式，系统评估人工智能系统是否具备生成与人类无显著差异的交互响应能力。根据Turing的经典论述^[33]，该测试通过结构化对话场景设计，判断某一智能机器是否能够使人类评估者难以有效辨识其身份属性，即通过模拟人类对话的方式，检验机器是否能够以与人类无法区分的方式进行回答，从而使人类评估者无法区分回答者是机器还是人类，其标准化实施流程包含以下关键环节：

第一，实验环境构建。建立严格隔离的双盲测试场域，将评估者与被评估对象置于物理隔离状态，通过标准化文本交互界面进行对话。

第二，对话任务设计。人类评估者向被评估对象（智能机器）提出一系列问题，这些问题旨在引发深入的对话，其提问内容可以涵盖语义理解、逻辑推理、情境应变及社会文化认知等维度。

第三，评估指标测试。当系统满足以下阈值标准时，即判定其通过“图灵测试”：如果人类评估者中能准确识别出被评估对象的机器身份的比例低于70%；或超过30%的人类评估者将机器回答误认为是人类回答，则证明该机器通过了“图灵测试”，初步具有人类智能。

大模型回复文本生成后，面向一般民众发放问卷，进行图灵测试：研究采用被试内实验设计，所有受访者均阅读咨询、建议、投诉3条民众留言内容，每条留言后以随机顺序先后展示由公务员和大模型回复的2条文本。具体而言，首先呈现同一类型的民众留言内容，其次在不告知文本生成方式前提下先后展示公务员回复文

本和大模型回复文本(参见表1),并询问其对各段文本的满意程度。最后告知受访者其中一段文本由大模型生成,并要求受访者猜测其中何者由大模型生成。

具体留言示例如下:建议加快建设桥底空间规划:我区辖区内很多高速公路桥下空间处于丢荒状态(如A镇某高速公路桥下空间),望领导们能参考周边城市经验,合理利用桥下空间,打造公共设施场所。

大模型回复文本生成后,面向一般民众发放问卷,进行“图灵测试”。由于被试内设计所需的被试量少,且能够排除被试差异带来的误差以提高实验的敏感性^[34],因此研究采用被试内设计。

(二) 政务大模型回复民众留言的“图灵测试”结果

在实验样本方面,研究使用G*Power v.3.1计算样本量,在保证中等效应量(Cohen's $f=0.25$)、显著水平为0.05及检验力达到0.95的情况下,仅需209条样本。由于研究采用被试内实验设计,共获得80名被试,每个被试均需阅读3条民众留言内容和6条回复文本,并做出3次判断选择,对6条文本进行满意度评价,因此后续分析满意度的回归模型中样本量为480,超过所要求209条。样本的描述性统计结果如表2所示,在满分为5,最低分为1的情况下,样本的满意度均值达到4.038,说明受访者对所有回复文本整体上具有较高的满意度。

在收集问卷之后,计算所有受访者将公务员回复文本判断为大模型回复的数量及其总体比例。结果如表3所示,在所有类型回复文本中,受访者的判断错误概

表2 “图灵测试”样本的描述性统计结果

变量	样本量	均值	标准差
性别	80	1.350	0.477
年龄段	80	2.663	0.822
户口类型	80	1.738	0.440
婚姻状况	80	1.350	0.477
就业部门	80	4.130	1.253
满意度	80	4.038	0.859

表3 各类型回复文本判断错误率

留言类别	错误率
建议	37.5%
咨询	53.75%
投诉	46.25%
总体	45.83%

率均高于30%,平均而言,所有留言判断的错误率达到45.83%,即接近一半的受访者将公务员回复文本误认为大模型回复文本,说明在简单政务回应这一应用场景中,即使是未经微调的大模型也能通过“图灵测试”,也可以认为其已经具有一定程度的人类智能。

随后以回复类别对回复满意度进行回归分析。结果如表4显示:相较于公务员回复文本,大模型生成回应甚至取得了更高的满意度,其系数达到0.01的显著性水平。说明相较于公务员回复文本,大模型生成的回复文本得到更高的满意度。总而言之,在大模型模拟的政务留言中,民众对大模型模拟留言的满意度已经与真实的政务留言水平相当。

表1 回复文本比较(以建议类型留言为例)

回复主体	回复文本
公务员	尊敬的网友: 您的留言收悉。现回复如下: 我区已利用桥下空间共10处。下一步,我区相关部门将在满足桥下空间安全管理规定的原则下,充分借鉴国内外桥下空间利用优秀案例,结合实际情况,持续分期策划推动空间利用项目,盘活桥下空间资源,持续改善人居环境。 感谢您的留言,特此回复。
大模型	尊敬的市民: 感谢您的宝贵建议。针对您提出的桥底空间规划问题,区政府高度重视。我们将组织相关部门进行调研,借鉴周边城市的成功经验,制定合理的规划方案。同时,加强与高速公路管理部门的沟通协调,确保项目顺利推进。我们的目标是将桥下空间打造成集休闲、娱乐、健身于一体的公共设施场所,为居民提供更好的生活环境。再次感谢您的关注和支持!

表4 大模型生成回复与人类回复的满意度差异

变量	满意度
回复类别 (公务员回复=0, 大模型回复=1)	0.233*** (0.072)
留言种类固定效应	已控制
个体固定效应	已控制
Con	4.352*** (0.326)
F值	2.14
R	0.306
N	480

注: ***表示回归系数达到0.01的显著性水平, 下同

值得一提的是, 本文仅仅简单设定对话背景信息, 生成文本过程未采用提示词技巧, 也并未以对话文本对模型进行微调, 仍然取得较高的用户满意度。类比于有学者提出的极端反推模式^[35], 即通过极端个案反推总体状态的方法, 可以据此推断, 集成各类政策文件与政府内部信息、以人类生成的优质回复文本作为有标注的特定任务数据进行微调后的垂直领域政务大模型将具有更为明显的优势。

四、人工智能赋能客观测度：政务大模型回复民众留言的文本质量评估

(一) 留言回复评价指标与大模型回复文本生成过程

为了系统评估政民互动中政府回应质量, 并测试生成式人工智能的政务回应能力, 需要围绕文本本身寻找评估指标, 系统比较公务员与大模型回复文本。因此, 在人民网领导留言板中选取近3年时间的留言, 并随机得到地方政府回复的民众留言6000条, 同时保存地方政

府对各留言的回复文本; 随后要求大模型扮演公务员回应的民众留言, 由于DeepSeek服务器繁忙等问题, 部分留言未能得到大模型回复, 因此共得有效样本5846条, 其中投诉留言3007条, 咨询留言1438条, 建言留言1401条。最后以客观指标分析比较公务员回应与大模型回应的文本特征。参考已有相关研究, 选取表述逻辑等4个指标进行具体评估(各指标及其测量如表5所示)。

与前述文本比较部分相同, 具体到留言的生成过程中, 通过调用API方式要求大模型作为人民网领导留言板的政务咨询助手, 代表被留言的政府部门及其领导回复民众留言, 以此生成回复文本。具体步骤如下: 首先, 在人民网领导留言板中选取咨询、建议、投诉类型的民众留言, 同时保存地方政府对各留言的回复文本。其次, 根据民众留言对象、种类和具体文本, 设计与大模型对话的系统信息(System)和提示词(Prompt), 要求大模型扮演公务员回应民众留言。系统信息(System)具体为: 你是人民网领导留言板中负责回答民众留言、投诉和建议的政务大模型, 请代表被留言的政府部门或领导回答以下问题。提示词(Prompt)为: 这是一篇给{row['地方领导']}的留言, 是关于{row['留言主题']}的{row['留言类型']}, 具体留言内容如下:{row['留言内容']}, 请你代表该领导及其部门回答这一留言。其中, 留言主题代表民众留言反映问题所属的政策领域, 包括三农、交通、企业、医疗、城建、就业、政务、教育、文娱、旅游、治安、环保、金融等13类。而留言类型指代留言本身的类别, 分为咨询、投诉和建议三类。模型设定参数如表6所示。

表5 政府回复文本评估指标

指标名称	测量方式
文本长度	以Python语句计算字符长度
表述逻辑	调用百度AI开放平台基于DNN (deep neural network) 训练的语义表达模型接口进行文本逻辑性分析, 返回的结果为对留言文本的评分。该指标取值越小表示某一文本表述逻辑越好, 越符合人类语言逻辑, 语句越通顺
文本倾向	调用百度AI开放平台训练的情感分析接口计算, 返回的结果包括正向、负向情感程度和情感极性分类结果。正面(负面)情感程度取值为[0, 1], 取值越大表示该文本正面(负面)情感越强烈
回应相关性	回复文本与民众留言的相关程度。使用Jaccard系数度量, 即计算A文本和B文本交集元素个数占A和B并集元素个数的比例 ^[36, 37] 。该指标取值为[0, 1], 取值越大表示A、B文本相似程度越大

表6 模型参数

参数名称	参数含义	取值
model	选用模型	DeepSeek-R1
temperature	生成文本的创造性	0.1
top_k	预测词汇的随机性	4
max_tokens	输出的最大文本量	1024
stream	流式输出	False

(二) 大模型与公务员留言回复文本质量评估与比较结果

由于利用大模型对每条得到公务员回复的民众留言均生成相应的大模型回复文本，因此公务员回复与大模型回复样本量相同，一一匹配，采用T检验比较二者在文本长度等文本指标的差异，以此评估、比较二者的回复质量。表7结果显示：大模型回复的文本长度和语言逻辑性显著高于公务员回复。而相较于大模型回复，公务员回复与民众留言的相似程度更大。在情感倾向方面，公务员回复的正面情感更为明显，而大模型回复的负面情感更为明显。上述结果说明：相较于公务员回复，大模型回复文本更为详细，且更符合人类语言逻辑，但其回复对民众留言内容的回应性不及公务员，且表露出更为消极负面的情感倾向。

在简单比较公务员与大模型回复文本的整体差异之外，回复文本在留言类型（投诉、建言和咨询）、留言类别（问题的领域）、被留言官员层级等诸多方面的异质性同样不应忽视。根据符号性回应理论^[38]，官僚机构通过差异化的语言符号建构组织合法性。高层官员回复多采用“顶层设计”“统筹推进”等宏观叙事框架，基层回复则聚焦“限时办结”“专人负责”等执行性话语，这种分层话语体系在组织社会学中被视为维持科层

表7 模型参数

变量	公务员回复 (N=5846)	大模型回复 (N=5846)	T检验结果
文本长度	282.848	666.648	-383.800***
表述逻辑	8715.869	1797.959	6917.91***
留言相关性	0.092	0.086	0.005***
正面情感程度	0.499	0.303	0.195***
负面情感程度	0.501	0.697	-0.195***

权威的技术手段^[39]。具体到政民互动领域，部门层级、参与部门数量、利益相关方、满足网民诉求的所需代价等方面显著影响政府回应^[40]，同时也有实证研究发现，不同层级的政府部门在回应公众环保诉求时采取不同的话语模式^[41]。因此可以推断，在代表不同职务层次的政府领导回应民众留言时，公务员可能采取不同的语言表达策略和话语模式。然而，大模型回复是否保留这一回复特征仍未得到实证证明。因此，下文根据留言对象身份（官员职级）进行分样本分析，比较公务员与大模型在代表不同职务层级官员时的回复文本差异。

首先，以公务员回复文本为样本，通过回归分析揭示不同官员职务层次的回复文本差异。表8结果显示，相较于基层官员，在代表高层官员回复民众留言时，公务员回复倾向于采用更短的文本回复，表露更为积极的情感态度，但回应性与表述逻辑均较差。这一发现与已有研究结论相类似^[40,41]，说明公务员回应的话语模式受其代表的官员层级影响。这种层级差异可能源于政府科层体系中的权责配置与资源分配逻辑：高层政府官员需要维系政府的整体权威形象，其话语往往承担政治表态功能，同时又面临更复杂的议程设置和决策压力，其回应往往以情感安抚替代实质性解决方案，因此出现情感积极而回应性较弱的现象。而基层官员因直面公众诉求且受考核制度约束，需在有限行政资源下构建更详尽的政务回复，更为直接有效地回应民众政务需求，因此呈现具有较强回应性的长文本回复策略。

其次，以大模型回复文本为样本进行回归，表9结果显示，在代表不同职务层次官员回复民众留言时，大

表8 不同官员职务层次的公务员回复文本差异

变量	长度	正面情感倾向	回应性	表述逻辑
被留言官员职务层次	-10.187*** (1.725)	0.010*** (0.002)	-0.001*** (0.0002)	817.300*** (256.57)
Con	328.275*** (9.866)	0.462*** (0.011)	0.093*** (0.002)	4503.981*** (1467.370)
F值	34.87	24.44	6.15	10.15
R	0.009	0.004	0.001	0.002
N	5846	5846	5846	5846

表9 不同官员职务层次的大模型回复文本差异

变量	长度	正面情感倾向	回应性	表述逻辑
被留言 官员职务 层次	0.534 (1.035)	0.002 (0.002)	-0.0002 (0.0001)	42.430 (96.226)
Con	663.459*** (5.918)	0.293*** (0.010)	0.087*** (0.001)	1579.610*** (550.322)
F值	0.270	1.29	1.43	0.19
R	0.0001	0.0002	0.0002	0.0001
N	5846	5846	5846	5846

模型回复在所有文本指标上均未出现显著差异,而是倾向于采用通用化表述代表所有层级官员回复民众留言。这一现象表明,尽管提示词中已经输入领导层级相关信息,但由于政府回应领域的训练数据不足等问题,生成式人工智能缺乏对科层制下层级分类和职务层级权责的深刻理解,对政府官员的角色感知能力不足,未能捕捉中国制度下的官员职务层级与语言风格的关系。因此,由于技术逻辑与行政逻辑的脱嵌,在政民互动实践中公务员对民众留言回复呈现“差异化回应”微观语言特征的同时,而大模型则采用“无差异回复”模式。

五、结论与建议

在各地政府政务系统广泛接入DeepSeek的背景下,充分挖掘生成式人工智能赋能公共服务供给的技术优势有其现实意义,本文通过评估生成式人工智能赋能政务回应的实践案例,以人民网领导留言板回复为具体应用场景,通过对大模型返回的文本结果进行分析,以此评估其应用效能,研究发现:通用大模型对民众留言的回复功能已经通过“图灵测试”,模型生成回复的满意度显著高于公务员回复文本。相较于公务员回复,通用大模型回复文本具有诸多特征,并在文本长度、语言表述逻辑等方面具有明显优势,但在情感倾向与回应性方面仍有不足,同时由于缺乏对科层制层级分类和职务层级权责的理解,大模型在代表不同层级官员回应时存在“无差异回复”的特征。

在大模型应用提升公共部门运行效率和政务咨询等公共服务供给速度的同时,同样不应忽视的是,随着生

成式人工智能在政务咨询、公文写作、报告生成、公共决策支持等方面生成能力的日益优化,公共部门的职业替代风险也随之产生。为应对势不可挡的职业替代挑战,尝试基于上述研究发现对公共管理实践与学术研究的技术转型提出可行路径:

一是需要构建完善公共组织工作分类体系。随着政务大模型快速发展并逐渐替代公共组织内部程序化、高重复性行政任务,公共组织需重新定位公务员的核心能力需求,根据现行公共部门行政工作内容按照不同特征和属性进行分类,确保分类的准确性和实时性;在政务大模型日常应用中实时收集记录运行日志,实时分析技术迭代对岗位能力需求的影响;综合把握最新技术演进动向、岗位需求、人才供给状况,根据上述分类体系为公务人员的选拔、培训提供科学依据。

二是以调整训练数据、价值观对齐、模型微调等手段确保政务大模型的公共价值导向。研究发现当前大模型在回应民众诉求时,其回复未能捕捉科层制下官员职务层级与语言风格的关联。因此,在训练数据构建阶段,需通过系统标注数据将官员职务序列、部门权责清单、行政程序规范等制度性知识转化为机器可识别的结构化数据,破解模型对条块关系等本土行政实践的理解盲区。此外,作为公共服务的供给工具,政务大模型必须始终以服务公众利益、促进社会福祉、维护公平正义为根本目标。公共价值导向要求模型不仅要追求技术效率和性能优化,还要确保其输出与社会公认的价值观和伦理标准一致,实现技术理性与价值理性的耦合。

三是亟需基于通用大模型开发公共管理中各类具体领域的垂直大模型。首先,需要根据政务咨询、公文写作、政策制定等公共行政与公共服务供给的具体需求进行深入分析。其次,广泛收集、预处理政策文本和学术论文等相关资料数据,为模型训练奠定坚实基础。再次,在精确选择基础模型后,根据各工作领域的特性和要求,对模型架构进行定制化微调,通过科学训练方法和策略,对定制后的模型进行系统的训练与调优,确保其在专业领域内的性能和准确性达到最优。最后,通过接口封装实现与其他大模型的无缝集成,同时建立持续的监控与维护机制,以及根据实际应用中的反馈和业

务变化,对模型进行不断的迭代与优化,持续提升其功能。垂类大模型能够结合特定群体的政策认知特征和心理决策机制,在政务咨询服务供给中以行为工具提升受众的政策认知,降低政策执行成本,提升政策效应。具体到政务咨询大模型方面,以本研究前期探索开发的跨境人才政策咨询服务大模型为例,未来的政务回应效能提升和应用开发需要通过意图识别实时解析用户提问的隐含需求,并结合上下文语境生成个性化回应,并在多轮对话的动态交互中,系统捕捉并分析公众在政务咨询中表露行为倾向与认知偏好,总结特定政策领域的政务回应特征,探索优化政务回应模式与语言表达框架,为提升政务回应的行为效应提供参考。

尽管本研究为生成式人工智能在政务领域的应用提供了实证支持,但在实验设计、评估指标等方面仍有待未来研究进一步完善提升。例如,仅仅采用正负情感倾向、语言逻辑性等少数指标,仍有待探讨大模型文本的回复话语模式等相关议题。此外,主要通过图灵测试和文本质量评估来衡量大模型的效能,但未充分考虑用户在实际政务回应互动中的长期使用体验及其动态反馈。未来研究中,可以设计更长时间的实验,纳入更多理论模型和关键变量,基于政务大模型提供政务回应服务的真实操作场景,系统评估民众在制度、技术、个人多重因素共同形塑的人机互动过程中产生的对技术应用的认知与评价及其对相关政策的认知、态度与行为倾向^[14],从而更为全面地揭示政务大模型应用的效能提升及其作用机制。

参考文献:

- [1]张熙,杨小汕,徐常胜. ChatGPT及生成式人工智能现状及未来发展方向[J]. 中国科学基金, 2023, 37(05): 743-750.
- [2]Heseltine M, von Hohenberg B C. Large language models as a substitute for human experts in annotating political text[J]. Research & Politics, 2024, 11(01): 1.
- [3]Mittelstädt J M, Maier J, Goerke P, et al. Large language models can outperform humans in social situational judgments[J]. Scientific Reports, 2024, 14: <https://doi.org/10.1038/s41598-024-79048-0>.
- [4]沈思,冯暑阳,吴娜,等. 融合大语言模型的政策文本检索增强生成研究[J/OL]. 数据分析与知识发现, 1-18[2025-04-22]. <http://kns.cnki.net/kcms/detail/10.1478.g2.20241118.1718.002.html>.
- [5]DeepSeek-V3: Technical report[EB/OL]. (2024-12-01)[2025-04-11]. <https://arxiv.org/abs/2412.19437>.
- [6]DeepSeek LLM: Scaling open-source language models with longtermism [EB/OL]. (2024-01-01)[2025-04-11]. <https://arxiv.org/abs/2401.02954>.
- [7]董昌其,李大宇,米加宁. 大模型嵌入政务服务: 能力边界、协同治理与发展路径——基于地方政府大规模部署 DeepSeek的观察[J]. 电子政务, 2025(08): 13-21.
- [8]Mou X, Ding X, He Q, et al. From individual to society: A survey on social simulation driven by large language model-based agents[EB/OL]. (2024-01-01)[2025-04-11]. <https://arxiv.org/abs/2412.03563>.
- [9]张书维,李纾. 行为公共管理学探新: 内容、方法与趋势[J]. 公共行政评论, 2018, 11(01): 7-36, 219.
- [10]曹海军,王丽娟. 服务逻辑主导下的数字赋能、政民互动与价值共创——以北京市“接诉即办”为例[J]. 理论探讨, 2023(06): 76-83.
- [11]黄小勇,刘倪. 数字行政负担的生成机理及消减策略——基于“组织-技术-用户”的分析框架[J]. 电子政务, 2025(04): 74-84.
- [12]Kalu N. Leadership and discretionary decision-making in a technocratic administration: Confronting a new praxis[J]. Administrative Theory & Praxis, 2001, 23(03): 311-336.
- [13]张书维,张梓丙,王戈. 行为公共管理视角下的信息技术与街头官僚问责制: 理论述评与模型构建[J]. 甘肃行政学院学报, 2024 (01): 16-27, 124-125.
- [14]莫庭阳,傅承哲,张健新. 数实相融视角下粤港澳大湾区跨境发展政策的融合机制及其优化策略[J]. 数字治理评论, 2024(00): 146-175, 204-205.
- [15]Wilks Y. Machine Conversations[M]. Dordrecht: Kluwer, 1999: 7-10.
- [16]Honig S, Oron-Gilad T. Understanding and resolving failures in human-robot interaction: Literature review and model development[J]. Frontiers in Psychology, 2018(09): 861.
- [17]冉龙亚,陈涛,孙宁华. 技术特性对政务智能聊天机器人初始信任的影响[J]. 情报杂志, 2022(07): 138-144.

- [18]Ciccio R D, Silva S C E, Alparone F R. Millennials' attitude toward chatbots: An experimental study in a social relationship perspective[J]. *International Journal of Retail & Distribution Management*, 2020, 48(11): 1213–1233.
- [19]姚清晨, 黄璜. 聊天机器人在公共部门管理中的应用研究综述 [J]. *图书情报知识*, 2022(05): 144–156.
- [20]潘琳, 李兴腾, 李珮. 政务大模型赋能数字政府治理: 驱动逻辑与优化路径——基于“结构-过程-功能”的分析框架 [J]. *重庆工商大学学报: 社会科学版*, 2024, 42(02): 1–14.
- [21]汪波, 牛朝文. 从ChatGPT到GovGPT: 生成式人工智能驱动的政务服务生态系统构建[J]. *电子政务*, 2023(09): 25–38.
- [22]Giest S N, Klievink B. More than a digital system: How AI is changing the role of bureaucrats in different organizational contexts[J]. *Public Management Review*, 2022, 26(02): 379–398.
- [23]宋华琳. 政务搜索引擎的法治建构——基于DeepSeek等人工智能大模型的应用 [J]. *求索*, 2025 (02): 131–141.
- [24]张雁冰, 吕巍, 张佳宇. AI营销研究的挑战和展望[J]. *管理科学*, 2019(05): 75–86.
- [25]Soares M A C, Parreiras F S. A literature review on question answering techniques, paradigms and systems[J]. *Journal of King Saud University – Computer and Information Sciences*, 2020, 32(06): 635–646.
- [26]张朝阳, 沈建辉, 叶伟荣. 融合预训练语言模型的知识图谱在政务问答系统中的应用研究[J]. *数字通信世界*, 2024(09): 188–190.
- [27]魏钰明, 贾开, 曾润喜, 等. DeepSeek突破效应下的人工智能创新发展与治理变革[J]. *电子政务*, 2025(03): 2–39.
- [28]Wei Z, Landay J A. Evaluating speech-based smart devices using new usability heuristics[J]. *IEEE Pervasive Computing*, 2018, 17(02): 84–96.
- [29]Fiske S T, Cuddy A J C, Glick P. Universal dimensions of social cognition: Warmth and competence[J]. *Trends in Cognitive Sciences*, 2007, 11(02): 77–83.
- [30]高学德, 廉静. 数字治理背景下政府回应话语模式对公民满意度的影响——基于情感和理性视角[J]. *电子政务*, 2024(12): 65–78.
- [31]沈鹏熠, 万德敏, 许基南. 在线零售情境下人机交互感知如何影响消费者幸福感——基于自主性的视角[J]. *南开管理评论*, 2021, 24(06): 26–40.
- [32]华斌, 康月, 范林昊. 政策文本的知识建模与关联问答研究[J]. *数据分析与知识发现*, 2022(11): 79–92.
- [33]Turing A M. Computing machinery and intelligence[J]. *Mind*, 1950, 59(236): 433–460.
- [34]舒华, 张亚旭. 心理学研究方法: 实验设计和数据分析[M]. 北京: 人民教育出版社, 2008: 151.
- [35]King G, Keohane R O, Verba S. *Designing Social Inquiry: Scientific Inference in Qualitative Research*[M]. Princeton: Princeton University Press, 1994.
- [36]Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901, 37(142): 547–579.
- [37]Salton G, McGill M J. *Introduction to modern information retrieval*[M]. New York: McGraw–Hill, 1983.
- [38]Edelman M. The symbolic uses of politics[J]. *Western Political Quarterly*, 1964, 19(19): 194.
- [39]周雪光. 中国国家治理的制度逻辑——一个组织学研究[M]. 北京: 生活·读书·新知三联书店, 2017: 63.
- [40]张会平, 邓凯, 郭宁, 等. 主体特征和信息内容对网民诉求政府回应度的影响研究[J]. *现代情报*, 2017, 37(11): 17–21, 27.
- [41]常多粉, 孟天广. 动之以情还是晓之以理?——环境治理中网络问政的政府回应话语模式[J]. *社会发展研究*, 2021, 8(03): 44–62, 243.

作者简介:

傅承哲, 管理学博士, 华南师范大学政治与公共管理学院副教授, 华南师范大学粤港澳大湾区跨区域治理与公共政策研究中心副主任, 全国港澳研究会会员, 广东省习近平新时代中国特色社会主义思想研究中心华南师范大学基地研究员, 研究方向为粤港澳大湾区研究、行为公共管理、政策评估。

莫庭阳, 华南师范大学政治与公共管理学院硕士研究生, 研究方向为行为公共管理、政策评估、文化心理。

王楚暖, 华南师范大学政治与公共管理学院硕士研究生。