超越工具论：政务 AI 应用的认知误区、适配挑战与数据基础构建

谷剑芳[1]，张阳[2*]，王丹丹[3]，袁云云[4]，金玲[5]，杨奎[1]
1. 河南省政务大数据中心，河南 郑州 450052
2. 河南省数据局，河南 郑州 450052
3. 河南科技大学商学院，河南 郑州 450052
4. 河南省遥感院，河南 郑州 450052
5. 河南省自然资源厅，河南 郑州 450052
DOI:10.61369/ASDS.2026100002

摘　　要： 2025年是一个明显的分水岭。DeepSeek 等国产模型在中文语境下的表现显著提升，私有化部署的门槛又降了下来，各地政务系统引入 AI 的节奏随之加快，智能客服、工单分派、公文辅助一类应用密集上线。问题也随之暴露：不少系统上线即闲置，答非所问，知识库长期不更新。笔者在参与省级政务数据治理与多地调研中感到，这类失败多半不在技术侧，而在决策者对大模型能力边界的判断上。本文据此梳理出三类偏差：最要紧的是把大模型当成查得准、不会错的政策知识库，忘了它的答案其实是按概率拼出来的，不是从库里查出来的；其次是立项时先看指标参数、后看业务痛不痛，又犯了只管建、不管用的老毛病；还有一条常被忽略——按“验收即完”的传统信息化方式管 AI，既没留数据更新的口子，也没留模型迭代的钱。为检验这三类偏差能解释多少结果，本文对比了四个样本：先看福田那批“AI 公务员”；再看漳州的“候喜”；杭州城市大脑是第三个；还有一个不在政务系统里，是沃尔玛。沃尔玛不在政务系统内，放进来是为了隔离一个变量——当数据不缺、预算不缺，工具本身碎片化会造成什么后果。案例分析沿“场景嵌得多深”与“平台统得多广”两条线索展开，由此归纳出需求、场景、数据、平台四个落地条件，并做成一张四维评分表（业务价值40%、数据基础25%、技术可行性20%、风险可控性 15%），供地方在立项前对候选场景打分排序。文章最后讨论支撑条件，认为标准、数据、平台三件事里，从数据供给侧到模型消费侧的那段加工链路最容易被跳过，也最值得先投。相比已有研究，本文想留下的不是又一个技术框架，而是一张能拿去开会用的打分表。

关 键 词： 人工智能；政务服务；大语言模型；认知误区；场景适配；高质量数据集；平台治理

Beyond Instrumentalism: Cognitive Misconceptions, Adaptation Challenges, and Data Infrastructure Construction in Government AI Applications

Gu Jianfang[1], Zhang Yang[2*], Wang Dandan[3], Yuan Yunyun[4], Jin Ling[5], Yang Kui[1]
1. Henan Provincial Government Big Data Center, Zhengzhou, Henan 450052
2. Henan Provincial Administration of Government Services and Data Management, Zhengzhou, Henan 450052
3.School of Business, Henan University of Science and Technology, Zhengzhou, Henan 450052
4. Henan Provincial Institute of Remote Sensing, Zhengzhou, Henan 450052
5. Department of Natural Resources of Henan Province, Zhengzhou, Henan 450052

Abstract: The arrival of DeepSeek and comparable Chinese-built large language models in 2025 changed how fast local governments in China began experimenting with AI. Deployment had become cheap enough, and private deployment secure enough, that smart customer service, work-order triage and drafting assistance moved from pilot to routine inside a single budget cycle. What has received far less attention is how unevenly these systems perform once they are in service. Drawing on the authors' own work in provincial government data governance and on field visits to Shenzhen, Zhangzhou, Hangzhou and Qingdao, this paper argues that the failures are not primarily algorithmic.

作者简介：
谷剑芳，本科，高级工程师，河南省政务大数据中心数据治理部副部长，研究方向为数据治理、数据开发利用、数据要素政策；
王丹丹，硕士研究生，教授，博导，河南科技大学商学院，研究方向为数据要素与数据治理，大数据管理与应用；
袁云云，本科，中级会计师，河南省遥感院财务科会计，研究方向数据要素与数据治理，数据资产管理与应用；
金玲，本科，河南省自然资源厅副处长，研究方向数据要素与数据治理，数据资产管理与应用；
杨奎，硕士，高级工程师，河南省政务大数据中心公共平台部，研究方向为信息系统工程，大数据工程和人工智能应用；
通讯作者：张阳，本科，正高级工程师，河南省数据局数字政府规划建设处处长，研究方向为数字政府规划建设、政务信息化项目建设管理。

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

They are failures of judgment about what a language model can and cannot be asked to do. Three misjudgments recur. The first treats a chatbot as an authoritative policy database, when its answers are sampled continuations of text rather than look-ups against a verified record. The second ranks candidate projects by technical ambition rather than by the size of the administrative problem they would remove, reproducing a bias toward building over using that long predates AI. The third manages AI as a deliverable with an acceptance date, leaving no budget line for the data refresh and retraining a deployed model needs in order not to decay. To test how much these misjudgments actually explain, four cases are compared: Futian District's "AI civil servants", Zhangzhou's "Hou Xi" assistant, Hangzhou's City Brain, and – deliberately, from outside government – Walmart's consolidation of scattered AI agents into a single front end. Walmart is included because it isolates a variable the public-sector cases cannot: what goes wrong when an organisation already has the data and the budget, and still fails because its tools are fragmented. Two questions organise the comparison: how deeply a tool is embedded in the actual work process, and how much of the stack is shared. These are then turned into a four-part scoring sheet (business value 40%, data readiness 25%, technical feasibility 20%, risk controllability 15%) that a local authority can apply before committing funds. A closing section argues that of the three foundations usually cited – standards, data and platforms – the least developed in practice is the chain that turns raw administrative records into material a model can actually learn from, and that this is where marginal investment currently pays off most.

Keywords : **artificial intelligence; government services; large language models; cognitive misconceptions; contextual adaptation; High-Quality datasets; platform governance**

引言

（一）研究背景与现实问题

数字技术的快速演进正深刻重塑政府治理模式与公共服务形态。以大语言模型为代表的新一代人工智能技术，凭借其在自然语言理解、语义推理与文本生成方面的突出能力，为政务服务智能化升级、行政流程优化再造和辅助决策支持提供了前所未有的技术驱动力。2025年以来，以 DeepSeek 为代表的国产大语言模型凭借在中文语境下的显著性能提升、灵活可控的私有化部署方式及对数据安全的高度重视，迅速成为政务领域智能化探索的重要工具。深圳、杭州、福州、青岛等数字化治理基础较好的地区，已在智能客服、政策咨询、公文辅助、热线工单处理等场景中积累了一定实践经验。

然而，技术热潮之下，政务 AI 的深度落地并非一帆风顺。从实践观察来看，当前政务 AI 建设中存在三类较为普遍的结构性问题：一是认知层面，政务决策与执行主体对大语言模型的技术本质、能力边界和适用范围存在系统性误判，导致应用路径选择偏差；二是实施层面，技术导向优先于业务需求导向的路径依赖，使“重建设、轻应用”的历史顽疾在 AI 时代以新形式延续；三是机制层面，沿用传统软件项目的静态工程思维，忽视了政务 AI 系统持续运营、数据迭代与人机协同的内在复杂性。上述问题相互叠加，导致大量政务 AI 项目在建设规模扩张的同时，实际效能释放严重不足，甚至出现“上线即搁置”“形象工程”等现象，造成公共资源的显著浪费与公众信任的潜在损耗。

这一现实困境表明，推动政务 AI 从浅层试点走向深度落地，既需要对大语言模型技术本质形成准确认知，也需要在场景选择、数据治理与平台建设等基础能力上实现系统性突破。如何在正确认知 AI 能力边界的前提下，构建符合政务场景特性的应用框架与评估工具，成为当前政务信息化研究亟待回答的核心问题。

（二）相关研究综述

人工智能用在公共服务和政府治理上，这几年讨论得不少。但具体到选型、落地和运维，已有研究给出的答案还不够用。下面梳理的三点不足，就是本文要试着补上的地方。

关于 AI 进政府会遇到什么阻力，Wirtz 等人早在2019年就用技术接受模型加制度理论梳理过一遍，归结到三样东西：单位自己能不能接住、制度上给不给空间、公众买不买账[1]。这个判断今天看依然站得住，但它停在“要不要用”的层面，没往下回答“用在哪”。Sun 等人的一项跨国比较发现，丹麦和中国的项目卡住的地方几乎一模一样——数据底子薄，部门之间不通气[2]。不过要注意，他们研

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

究的是上一波以规则引擎和机器学习为主的 AI，大模型带来的不确定性并不在他们的问题域内，这也正是本文想接着往下做的地方。

国内的研究更贴近工程实践：孟庆国、鞠京芮关心的是平台型政府的技术框架与实施路径[3]，施敏、杨海军盯的是隐私保护这一侧，具体到大模型的数据脱敏与采集环节[4]。两篇合起来看，一个讲怎么搭，一个讲怎么防，但都默认数据已经就绪——这个前提在真实项目里往往不成立。

真正对本文有直接支撑的是关于大模型能力边界的讨论。Bender 那篇被广泛引用的“随机鹦鹉”[5]点破了要害：模型输出的是按概率拼接的语言形式，流畅不等于理解。Huang 等人的综述给出的数字也不乐观——主流模型回答事实性问题时，出错的比例仍然可观，更麻烦的是这些幻觉常以高置信度的语气出现，使用者很难当场识别[6]。Ji 等人划出的高风险区更麻烦：打官司、看病、解释政策，恰恰都是知识密集的事[7]。把这三篇的结论叠起来看，会得到一个对政务场景很不友好的推论：幻觉最高发的任务类型，恰好与政务 AI 最想做的事高度重合。这一点在已有文献里没有被认真对待过，多数讨论停在“这个技术有什么特点”，没有往下写“那么在政务里该怎么管”。

在政务 AI 场景适配与数据治理研究方面，现有研究已认识到数据质量对 AI 效能的决定性作用。Janssen[8] 等指出，政府数据的碎片化、异构性与敏感性构成了公共部门 AI 应用的独特挑战，呼吁建立政务数据治理的专门框架。国内学者从数字政府建设实践出发，对政务数据共享、数据标准化和数据安全合规进行了较为深入的探讨，但对如何将高质量数据集建设与具体 AI 应用场景相结合，尚缺乏系统性的操作框架。此外，现有研究对政务 AI 平台建设模式的讨论多集中于技术架构层面，对“统分结合”模式在实践中如何兼顾共性集约与个性适配的机制设计，尚存在明显研究空白。

综合来看，现有研究在以下三个方面存在不足：第一，缺乏针对大语言模型技术特性与政务场景特殊需求之间适配关系的系统分析；第二，缺乏面向政务 AI 场景甄选的可操作评估工具；第三，在政务 AI 可持续发展的基础能力建设路径上，理论框架与实操指导之间仍有较大距离。本研究旨在针对上述三项不足作出针对性回应。

（三）研究问题与研究方法

1. 研究问题。文献读完，有三个问题绕不开：

问题一：地方在引入大模型时反复踩的坑，究竟是“不懂”，还是“懂”得过于自信？这两种情况的表现和后果不一样，得分开说。

问题二：同样是大模型，为什么福田的“AI 公务员”能铺到 240 个场景，而不少地方的系统上线后无人使用？差别主要出在场景选择，还是出在平台支撑？

问题三：标准、数据、平台这三件事，如果财力只允许先做一件，该做哪一件？三者之间是否存在严格的先后次序，还是可以并行？

2. 研究方法。本文用文献梳理与案例对比两种办法。

案例方面，按“理论复制”的思路选了五个样本，其中四个来自政务系统，沃尔玛作为企业侧对照。在多案例对比分析方面，以“理论复制”逻辑为指导，选取深圳福田“AI 公务员”、漳州“候喜”、杭州城市大脑、广州政务 AI 平台及沃尔玛 AI 整合转型共五个案例，构成覆盖政务场景纵深嵌入、跨域平台整合、基础设施集约建设及跨行业对照的多维样本。案例选取遵循以下标准：一是具有较为明确的应用目标与建设路径；二是有相对可查的实施过程信息；三是在场景类型、建设规模或组织形态上存在显著差异，能够支撑比较分析。需要指出的是，受制于公开信息的可及性，案例分析所依据的数据主要来自政府官方发布、权威媒体报道及行业研究报告，部分案例的效果数据尚缺乏独立第三方评估支撑，相关结论应在此限制下加以理解。

在分析框架上，本研究以“认知层—路径层—机制层”三层结构为主轴，围绕认知误区识别、场景适配要素归纳与基础能力建设路径三个维度展开论证，最终形成“需求导向—场景适配—数据底座—平台治理”的政务 AI 落地分析框架。

（四）研究贡献与文章结构

1. 研究贡献。本文想做的事有三件。

第一件是把“认知误区”从口号变成可辨认的具体表现——我们给出了每类误区的判别特征和对应的失败案例，决策者可以拿去对照自查，这是已有研究没做的。

第二件是那张四维评分表，它不新在概念上，新在能直接打分：权重、判分口径、一票否决条件都写死了，可以直接用于立项评审。

第三件是路径建议，我们没有按“标准—数据—平台”平铺，而是给出了优先次序和理由，也说明了“统”的代价。需要说明的是，这套东西是在有限的试用范围内调出来的，换到数据基础差异大的地区，权重需要重新校准，不能照搬。

2. 文章结构。全文共分 5 个部分。第 2 节梳理中国政务 AI 整体应用现状，勾勒技术应用的宏观背景；第 3 节系统分析阻碍政务 AI 深度落地的三类核心认知误区，通过多案例比较归纳政务 AI 成功落地的核心准则，并提出场景评估矩阵；第 4 节探讨政务 AI 可持续发展的基础能力体系构建路径；第 5 节总结研究结论，说明研究局限并提出未来研究展望。

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

一、政务 AI 整体应用现状

政务 AI 的推进节奏在2025年出现了一次明显变化。此前各地的尝试多是单点试验——一个智能问答框、一套人脸识别门禁、一个能查进度的语音热线，彼此不通气。2025年之后，DeepSeek 一类模型把本地部署的门槛压到县区级也能承受，加上私有化部署方案成型，数据安全这一关才算真正过得去。

从落地顺序看，最先跑通的是面向公众的高频咨询类场景。原因不难理解：问题相对标准、答案可以预先准备、答错了后果可控。沪、浙、粤三地已经把智能问答嵌进一网通办的办事环节，用户在填表环节就能拿到指引，这类应用目前运行得比较稳定。城市治理类场景铺得更广，但深浅不一。杭州城市大脑是少数真正打通了数据源的例子——它把交通、城管、应急的数据拉到一处，再按实时流量调信号灯，这套做法后来被不少城市照搬。更多的项目则只做到了可视化大屏，算法照旧跑在各部门自己的系统里，数据并没有真正合流。行政办公侧的普及速度超出预期。写材料、整纪要、翻文件这几件事，因为不直接对外产生效力，推进阻力最小，也是基层感受最明显的。法院走得更快：判决书可以先由模型出草稿，关键要素由它去抽取，这些在多地已经日常化。不过这里有个前提：法院的案件数据本身结构化程度高，换到数据基础差的部门，同样的工具往往推不动。

技术路线上，目前主流做法是“底下用通用模型，中间挂本部门的知识和规则，上面再按场景做微调”，人机分工也基本固定下来：模型出初稿，人来核。这个分工在可预见的时期内不会变。需要指出的是，覆盖面扩大不等于效能提升。跨部门数据拿不到，是最常被提到也最难解决的一条；其次是县乡一级没人维护知识库，省里训的模型到了基层容易失准；政务数据敏感度高，模型安全合规的要求比一般信息系统严，很多通用做法不能直接套用。这几条在后面两节会分别展开。

总体看，这几年政务 AI 的推进顺序是清楚的：先在外围的咨询、辅助类场景跑通，再往核心业务走。这个顺序不宜颠倒，也解释了为什么同样是上大模型，有的地方很快见到效果，有的则长期停在演示阶段。阻碍政务 AI 深度落地的核心症结在认知。DeepSeek 等大语言模型（LLMs）的突破性进展，点燃了2025年以来全社会各领域的 AI 应用热情，政务领域亦不例外。深圳龙岗政务服务大厅率先推出基于 DeepSeek 的“数字公务员”示范[9]，引发了各地政府的学习与效仿热潮，以期快速提升政务服务智能化水平。然而，在技术热潮涌动之际，一个深层次的挑战已然显现：政务主体对 AI 技术（尤其是 LLMs）的基本原理、核心能力边界、适配路径及长期运维逻辑普遍存在关键性认知偏差。这种系统性认知不足，源于政务决策与执行层面对 AI 相关通识知识的普遍欠缺，不仅导致其在应用路径选择上易陷于盲目性决策，更成为掣肘 AI 技术真正深度融入政务场景、实现可持续价值的核心瓶颈。

（一）能力认知偏差：过度泛化，误将 LLMs 视为“全域精准知识库”

GPT、DeepSeek、通义千问等生成式大语言模型（LLMs）的涌现，极大提升了社会对 AI 能力的预期。在政务领域，LLMs 的对话交互能力使其在办事查询、政策咨询等场景展现出显著便捷性，却也滋生了一种危险的认知偏差，即将 LLMs 等同于“全域精准、静态的知识库”。这种偏差直接导致对其可靠性的不合理预期。

现实风险已清晰显现：一是信息可信度危机。LMYS.org 2025年8月 Vicuna 评测显示，主流 LLMs 如 Gemini-2.0-Flash、OpenAI o3-mini-high、DeepSeek-V2.5均存在幻觉现象[10]。用户常遭遇 LLMs 以高度自信口吻输出虚构事实或错误信息。二是专业领域失准。美国律师因提交 ChatGPT 编造案例险酿司法事故[11]；DeepSeek-R1关于国产动画的回答充斥事实谬误[12]；更严峻的是，某医院 AI 处方系统忽略患者转氨酶异常，推荐可能加重肝损的降脂药瑞百安[13]，暴露其缺乏医学判断与风险规避能力。技术根源在于 LLMs 的运作本质。LLMs 的核心机制是基于海量语料进行分词、向量化，并利用注意力机制建模词汇关联与位置关系。其生成的“逻辑合理”文本，本质是统计概率驱动下的上下文拟合，而非基于事实核查或深度理解的理性推理。尽管模型持续优化，其“推理”过程仍具有自强化特性，极易在缺乏真实依据时生成“貌似合理”却偏离事实的“幻觉”内容，与人类基于认知与情感的判断存在根本差异。Emily M. Bender、Timnit Gebru 等研究者指出，LLMs 根据概率信息将训练数据中的语言形式片段拼接成序列，但并不具备人类的理解能力，“就像个随机鹦鹉”[14]。

LLMs 不等于知识库，其输出是概率产物，非精准数据库检索结果，不具备可靠的事实核查与专业判断能力。在现有技术框架下，统计模型驱动的自递归推理必然伴随一定程度的幻觉风险。在涉及法律、医疗、政策解读等高严谨性、高后果责任的政务场景中，过度依赖 LLMs 作为最终知识源或决策依据存在重大安全隐患。构建高质量政务语料库虽能降低错误概率，却无法解决 LLMs 基于概率生成“合理错误”的本质缺陷，需结合其他机制（如事实核查、规则约束、人类监督）构建安全边界。

（二）实施路径偏差：技术本位，忽视政务场景适配性的单向思维

“重建设、轻应用”的顽疾，长期困扰着政府信息系统建设，亦是催生数字形式主义的温床。尤其在综合类大型平台（如数据中台、监管平台）建设后，常陷入用户注册率高、实际活跃度低、功能长期闲置的窘境。这种脱离现实业务需求、盲目追求技术先进性的“为建而建”思路，导致技术与应用严重脱节，造成巨大的资源浪费与项目烂尾风险。

AI 项目的特殊性加剧了风险。相较于传统信息化项目，AI 项目（尤其是大模型应用）具有技术复杂度高、算法结果不确定性大、算力资源消耗巨大三大特征，其建设成本、实施难度与失败风险显著提升。比如，某地住房城乡建设厅投资1200万元建设的6个政务服务信息系统，片面追求硬件“高端大气”，却因缺乏实际业务支撑和用户需求牵引，部分系统建成即停用或长期闲置，沦为典型的“面子工程”与资源浪费[15]。政务服务的核心价值在于便民利民与行政效能提升。然而，脱离具体业务场景的 AI 大模

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

型建设，极易陷入追求技术复杂性的迷思。复杂、难用的 AI 交互设计，反而可能增加公众操作负担，背离“简化流程”初衷。“智能外壳”下若缺乏对流程优化、数据贯通等治理本质的支撑，非但无法提升效率，甚至可能因系统冗余、响应迟缓而降低服务效能。华而不实的 AI 项目不仅浪费公共财政，更会严重侵蚀公众对政府数字化能力的信任与满意度，损害政府公信力。

AI 立项必须以解决具体政务痛点、满足真实用户需求、提升核心业务效能为根本出发点，而非技术本身的先进性。坚决摒弃“为 AI 而 AI”的思维。地方政府需立足本地实际（业务成熟度、数据基础、财力人力），对 AI 项目的实用性、可行性、投入产出比进行严格论证。避免贪大求全、盲目跟风上马高风险项目。优先选择业务边界清晰、价值可衡量、技术风险可控的细分场景进行试点验证（如特定事项智能问答、表单自动填充），通过敏捷迭代、持续优化积累经验，再考虑规模化推广。避免一次性投入巨资建设“大而全”却无用的平台。AI 赋能政务服务的关键在于技术与治理的深度融合。必须同步推进业务流程再造、数据治理标准化、人员能力提升等治理工作，确保 AI 技术真正嵌入并服务于政务效能提升的目标，而非成为新的负担或“花瓶”。

（三）运维机制偏差：静态工程观，规避政务系统复杂性所需的动态闭环

当前受媒体渲染影响，将 AI 平台简单视作“即插即用”型工具，误以为部署即可输出稳定服务。这种认知严重低估了政务 AI 系统的复杂性与生命特性，本质上是将动态生命体错认为静态工程品。

政务 AI 平台（尤其基于大模型）是开放式系统，其核心能力（模型、高质量数据集、训练机制）需要持续运营而非一次性建设。AI 的精准性、适用性（如政策解读、问题解答）高度依赖持续注入的多维度、高质量政务数据流。这要求建立常态化的数据归集、清洗、标注、校核与训练机制。AI 功能价值的实现必须深度嵌入并适应动态变化的政务业务流程与规则，而非独立存在。比如，河北省石家庄市依托 DeepSeek 大模型对政务服务智能客服进行全面升级，通过百万级数据量政务知识库训练和本地化场景适配，提供 7×24 小时在线咨询解答 [16]。

在 AI 的应用实践中，真正有挑战的不是建立机器学习模型，而是在模型复杂而庞大的情况下，如何构建一个可以持续集成的机器学习系统，以及在生产环境中持续运行该系统 [17]。沿用传统软件项目“验收即结束”的思维，会导致后续数据更新停滞、模型迭代乏力、知识库陈旧，AI 性能迅速退化甚至失效；也会导致对持续性数据工程（DataOps）、模型运维（MLOps）所需资源（预算、编制、算力、专业团队）的长期规划和投入保障的缺失。AI 持续优化依赖技术方与业务方的高频深度协同（问题反馈收集、业务规则同步、场景适配调整）。静态思维下，协同机制往往流于形式或难以持续。非专业人士对 AI“易用性”和“智能水平”的认知混乱，叠加技术复杂性带来的“黑箱”效应，易导致决策者产生不切实际的预期或低估长期投入的必要性。这使得政策制定和项目规划脱离实际，最终在运维阶段暴露问题，造成项目“形同虚设”或“半途而废”。

所以与其说“AI 即服务”，不如把话说白：这类项目的钱，大头在验收之后。立项时就要把三件事写进方案，缺一件不建议开工。

一是数据的持续供给。模型上线只是起点，政策法规在改、办事指南在变，语料停更三个月，回答的准确率就会明显下滑。建议把数据归集、清洗、标注的责任落到具体处室，并明确更新频率——写进合同，而不是写在倡议里。石家庄的做法可以参考：依托 DeepSeek 升级智能客服时，同步建了百万级数据量的政务知识库，并配了本地化适配的维护流程，覆盖企业开办、执业登记、食品药品等200余项高频事项 [16]；真正难的其实不是建这个库，是让它一直新。

二是运维的责任主体。技术方撤场后谁来接，这个问题多数项目答不上来。比较可行的做法是设跨职能的“数字专员”，编制放在业务处室而非信息中心，负责收集问题、同步规则变更、协调模型再训练。同时给业务人员做一轮 AI 素养培训，目的不是让他们懂算法，是让他们知道什么样的错误必须反馈。

三是考核口径。建议把“模型更新频率、问题解决率、用户满意度”三项列为后评估的硬指标，与运维经费续拨挂钩。只考核“建成了几个”，结果一定是建完就停。

二、政务 AI 成功落地的核心准则与方法论

做信息化是为了办成事，不是为了有个系统。AI 也一样，先说清楚要解决谁的什么问题，再谈用什么模型。

（一）AI 应用要坚持需求导向

找场景不是列清单。调研中见过不少地方把“AI 应用场景”做成一张几十项的表，逐项比谁列得多，最后真正跑起来的不过三四项。问题往往出在一开始就问错了：不该问“AI 还能干什么”，而该问“哪些事现在办得最费劲”。

把两侧的能力摆在一起看会更清楚。大模型这边，真正成熟的本事集中在三处：读得懂用自然语言写的诉求，能从长文本里抽要点，能按模板生成格式固定的文稿。它明显不擅长的，是需要唯一正确答案、且答错了要担责的事。政务这边的特点恰好相反——规则硬、留痕严、一个事项常常牵扯好几个部门，多数环节不允许“大概齐”。两者一叠，能马上见效的范围其实很窄：咨询量大而答案相对固定的（政策问答、办事指引）；材料多而格式统一的（表单填报、证照识别）；以及需要在大量文本里找线索的（工单分派、投诉归类）。至于需要援引具体法条、直接影响相对人权利义务的环节，现阶段更适合让它写初稿、人来定稿，而不是直接对外输出。这条边界划清楚，后面讨论场景才有意义。

先说 AI 到底能干什么。把它和政务业务放在一起比对会发现，真正成熟的能力集中在三处：读得懂用自然语言写的诉求；能从长材料里抽出要点；能按固定模板生成文稿。把这三样放到具体岗位上看，最先受益的就是窗口咨询、热线应答、材料预审这类高频、重复、答案相对固定的环节——青岛12345把 AI 接进工单流转之后，话务诉求量同比下降了19.91%[18]，靠的就是这一

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

类替换。反过来，需要引用具体条款、直接影响相对人权利义务的事，现阶段更适合让它写初稿、人来定稿。

第二种用法是在跨部门的链条上找机会。企业开办、不动产过户这类“一件事”，群众要跑好几个部门、同一份材料反复交，是被抱怨最多的地方。模型能做的是把各部门的材料清单和政策条文通读一遍，把重复项和互相打架的条款标出来，由人决定砍哪一条。这一步在市县两级的效果差别很大：数据通的地方，几天就能出结果；数据不通的地方，模型只是把问题重新描述了一遍，事情照样卡在协调会上。

第三种用法，是让 AI 反过来治理数据本身。清洗、格式转换、跨库比对这类活儿，规则写清楚之后机器比人快，也不容易漏。我们试过的做法是：先让模型把明显矛盾的记录挑出来，附上出处，交给人去判。人的时间于是能花在判断上，而不是逐条翻台账。这一步的前提得说清楚——训练用的语料和更新机制必须先到位，否则机器校出来的东西同样不可信。

（二）AI 应用要适应场景需要

光讲道理不够，下面拿几个真实项目做对比，看跑起来的和趴窝的项目之间，究竟差在哪几个变量上。

案例一：福田那70个“AI 公务员”是怎么铺到240个场景的

先看深圳福田。福田在2025年年初上线了首批70个“AI 公务员”[19]。它们不是机器人，是挂在通用大模型上的一批小助手：每个助手只管一类事，比如公文初稿、民生诉求分拨，各自带着本部门的业务数据和业务规则。全区一共铺了240个场景，从公文处理到招商引资都有。数量看着惊人，做法其实很朴素——先把通用模型接进来，再灌业务数据，最后按具体事项切成一个个边界清楚的小体。

案例二：沃尔玛——把散落的 AI 工具收成一个入口

沃尔玛的教训在另一个方向。它很早就在用 AI：商品搜索一套、排班一套、供应商管理又一套。工具各自好用，放在一起却很别扭——入口不同、上下文不连通，员工要在几个界面之间来回跳，数据也各存各的。2025年7月，沃尔玛动手把这些工具重做成面向四类用户的“超级代理”，把搜索、推荐、客服收进同一个入口[20]。这次整合并没有换更强的模型，改的是架构。这一点对本文很重要：沃尔玛既不缺数据也不缺预算，它出问题纯粹是因为工具散着装——在政务案例里，这个变量和“数据不通”混在一起，很难单独剥出来。

两个案例摆在一起，能读出一条共同的规律：决定成败的不是模型好坏，是应用之间有没有被组织起来。福田的做法可以拆成三步——先接一个通用大模型，再灌入各部门自己的业务数据和规则，最后按具体事项做成独立的智能体。70个“AI 公务员”覆盖240个场景，看着是规模问题，实质是组织问题：每个智能体的边界划得很清楚，只做本部门那一类事，所以规则能写死、责任能落到人。反过来，如果一上来就做一个“什么都能答”的通用助手，规则写不进去，出错也找不到人负责。沃尔玛走的是相反的路：它先有了很多 AI 工具，问题恰恰出在“多”上——产品搜索、员工调度、供应商管理各有一套，接口不同、上下文不通，员工要在几个系统之间来回切。2025年7月的整合并不是换更好的模型，而是把入口合并成面向四类用户的“超级代理”[20]。值得注意的是，沃尔玛既不缺数据也不缺预算，它的失败完全来自架构碎片化——这个变量在政务案例里很难单独剥离出来，所以本文特意把它放进来做对照。由此得到的可迁移判据有两条：一是场景边界是否清楚到能写进规则，二是多个应用之间是否共用上下文。至于“技术先进度”，在这两个案例里都不是决定性因素，甚至没有出现在失败原因里。需要提醒的是，福田模式的隐含前提是部门数据已经打通，对数据基础薄弱的地区，照搬很难达到同样效果。

（三）政务 AI 场景甄选的价值引领与多维评估矩阵构建

为帮助决策者在立项前筛掉不成熟的想法，我们做了一张评分表（下称“评估矩阵”）。它的设计依据来自两份政策文件：《数字中国建设整体布局规划》强调应用实效，《生成式人工智能服务管理暂行办法》则划定了安全合规底线；前者对应业务价值，后者对应风险可控，中间的数据基础与技术可行性，是把已有研究的共识操作化之后补进去的。

先看业务价值，权重给到40%。理由很直接：一个场景如果做与不做的差别不大，后面三项再漂亮也没有立项必要。打分时建议只看三件事——这件事现在要花多少人天、出错率有多高、投诉集中在哪。三者有一项突出就往下走，三项都平平基本可以直接否掉。需要提醒的是，不少地方把“提升治理能力”“赋能科学决策”写进立项理由，这类表述无法证伪，建议在打分表里明确不计分。

数据基础给25%，排序在技术之前，是因为我们在实践中反复遇到同一件事：模型选得没错，场景也对，最后卡在数据上——要么拿不到，要么拿到的是三年前的。判断标准建议简化为两条：一是这批数据谁有权给、走什么流程、多久能到手（很多好想法死在“走流程”上）；二是数据有多旧、多久更新一次。至于完整性、准确性，可以用抽检代替全量评估：按批次随机抽样，错误率超过约定阈值的退回重标。

技术可行性给20%，是四项里权重最低的，也是刻意的。大模型应用的工程门槛这两年下降很快，真正因为“技术做不到”而失败的项目，比因为“数据不给”或“没人维护”失败的要少。这里只问三件事：现有算力够不够跑、团队有没有人能接手改、出问题能不能回滚。

风险可控性给15%，但它是“一票否决项”，不参与加权。涉及个人信息、可能影响相对人权利义务、或直接生成对外文书的，无论前三项分数多高，都得先把评估、备案这套合规流程走完，并且留一道人工审核。反过来，内部辅助、不对外、错了也无后果的场景，可以走简化流程。

表：AI 评估矩阵表

评估维度	指标解释和权重说明	参考权重
业务价值	看三件事：现有人天投入、当前出错率、投诉集中度。三者皆平则否掉。“提升治理能力”一类无法证伪的表述不计分。	40%
数据基础	看两条：数据能否拿到（权责与流程）、数据有多新（更新频率）。完整性、准确性用抽检代替全量评估，按批次随机抽样，超阈值退回重标。	25%

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

技术可行性	只问：算力够不够、有没有人接手改、出问题能否回滚。权重低于数据基础，因为实践中卡在技术上的项目少于卡在数据上的。	20%
风险可控性	不参与加权，一票否决。涉及个人信息、影响相对人权利义务、直接生成对外文书的，须过安全评估与算法备案，并保留人工审核环节。	15%

需要说明的是，这组权重是我们依据实际试用情况调出来的，不同省份数据基础差异较大，建议正式采用前先在本地试评两三个已知结果的项目，校准后再推开。评分表解决的只是“先做哪个”，做起来之后能不能持续，取决于后面一章讨论的那些支撑条件。

三、政务 AI 可持续发展的基石与体系化构建路径

前面回答了“选什么场景”，这一节回答“靠什么撑住”。数据、标准、平台是三件绕不开的事，但它们的紧迫程度并不相同，本节按这个顺序展开，并说明我们认为应该先做哪一件。

（一）标准规范先行，引领政务 AI 健康有序发展

标准先行这句话说了多年，真正卡住的不是没有标准，是标准落地时没人认。在省级数据共享平台的实际对接中，常见的情况是：接口规范发了，但各部门历史系统的改造排不上队，新标准只能在新建模块上生效，老数据依旧按老口径出。所以与其列一张涵盖架构、组件、协议、语料、流程、指标的全清单，不如先抓两件能验收的事。

一是模型与平台之间的调用接口，先统一到能在省内互认的程度；二是数据标注的质量要求，明确到“什么算标错、抽检比例多少、谁签字”这一层。其余的可以边做边补，一次把标准做全的项目我们没见过成功的。评价环节建议硬一点：明确每半年对在用模型评一次，评不过的就下线，而不是“优化或淘汰”这种两头都算的话。

（二）夯实数据根基，保障高质量政务数据供给

高质量数据集这个提法出自2024年底一份关于数据产业的指导意见，但落到政务口，难的不在“集”，在“炼”。政务数据的原始状态很差：省直部门的数据大多躺在各自业务库里，字段口径各定各的；市县一级还有大量扫描件、PDF 红头文件、Excel 台账。这些东西直接喂给模型，效果通常比不喂更糟——模型会把表格表头当正文读，把扫描件的水印和页码当关键词。因此，笼统地讲“全生命周期管理”没有意义，真正该单独立项的是供给侧到消费侧之间那段加工链路，可以拆成四步：先做面向训练的预处理工程，用自动化处理加人工抽检来消化扫描件和非结构化文本；再做标准化的接入通道，让共享平台能按模型需要的格式自动供数；第三是效果反馈，模型答错了要能回溯到是哪批语料的问题；第四是给标注成本单独算账——标注这件事靠业务骨干，不靠外包，得有长效机制留人。

（三）构建 AI 平台，强化共性支撑与集约发展

平台这块，争议一直不在“要不要统”，而在“统到哪一层”。杭州城市大脑和广州的政务大模型平台给出了两种答案。杭州从城市治理切入，拿 DeepSeek-R1 当底座，把十几个行业的模型放到同一套算力上来调度[21]；广州则先把算力、存储、网络、安全做成统一底座，再在上面长出工单分派、公文助手、政务服务“慧办小精灵”等十几类智能体[22]。两者的共同点在于底座统、应用分：算力、工具链、数据通道统一供应，行业模型各自训练。差别在起步点——杭州从场景倒推，广州从基础设施正推。对多数省份而言，后一条路更可复制：先把算力和数据通道统起来，让各部门不再各自买卡，再谈模型。需要一并说明的是，“统”也有代价——统一底座意味着单个部门的需求要排队，紧急需求响应变慢，这需要在立项时就把优先级规则写清楚，否则很容易在具体项目上反复扯皮。

四、结论与展望

（一）研究结论

本文从人工智能基本概念切入，首先明确 AI 在政务服务领域应用的正确认知，进而结合实际应用场景，总结了 AI 的创新案例，分析其面临的技术挑战。通过剖析典型案例，强调了技术适配场景、场景反哺技术以及充分激活数据价值的关键作用。研究认为，推动政务服务从数字化向智能化、人性化和精准化迈进，需要系统性谋划，包括挖掘核心应用场景、加工高质量数据集以筑牢数据基础、增强算力供给能力、采用统分结合的方式建设 AI 平台、完善运营服务体系等。

（二）研究展望

本研究虽构建了政务 AI 应用的系统性分析框架并提出实践建议，但选取的代表性案例多集中于特定地域与行业，且案例以成功实践为主，研究结论基于当前大语言模型与 AI 技术特征，部分结论可能受限于现阶段技术水平，对未来技术突破带来的新挑战与机遇预见不足。虽强调高质量数据集的重要性并设计 AI 评估矩阵，但对数据采集、清洗、标注等具体操作流程及工具的技术分析较为概括，未能充分展开实操层面的细节探讨。未来可重点从政务大模型的伦理规制体系构建、联邦学习、差分隐私等技术在政务数据处理中的适配性、主动感知技术的演进、以及人机对抗方向等方面进一步深化研究。

参考文献

[1]Wirtz ,Weyerer ,Geyer .Artificial Intelligence and the Public Sector—Applications and Challenges[J].International Journal of Public Administration,2019,42(7):596-615.

[2]Sun Q T ,Medaglia R .Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare[J].Government Information Quarterly,2019,36(2):368-383.

[3] 孟庆国、鞠京芮 . 人工智能支撑的平台型政府：技术框架与实践路径 [J]. 电子政务 ,2021(9):37-46.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

[4] 施敏，杨海军 . 大语言模型数据隐私保护的难点与探索 [J]. 大数据 ,2024(5):174-182.

[5] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: can language models be too big?[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021: 610-623. DOI: 10.1145/3442188.3445922.

[6] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[PP/OL]. arXiv(2023-11-09)[2026-09-09]. https://arxiv.org/abs/2311.05232. DOI: 10.48550/arXiv.2311.05232.

[7] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38. DOI: 10.1145/3571730.

[8]Janssen M ,Brous P ,Estevez E , et al.Data governance: Organizing data for trustworthy Artificial Intelligence[J].Government Information Quarterly,2020,37(3).

[9] 赵娟 , 孟天广 . 数字政府的纵向治理逻辑 : 分层体系与协同治理 [J]. 学海 ,2021(2):90-99.

[10] 刘泽垣 , 王鹏江 , 宋晓斌 , 张欣 , 江奔奔 . 大语言模型的幻觉问题研究综述 [J]. 软件学报 ,2025,36(3):1152-1185.

[11] 王秀平 . 人工智能司法应用法律风险及法治应对 [J]. 政法论丛 ,2024(2):151-160.

[12] 韩旭 , 张正彦 , 刘知远 . 知识指导的预训练语言模型 [J]. 中兴通讯技 ,2022,28(2):10-15.

[13] 李雪晴 , 王石 , 王朱君 , 朱俊武 . 自然语言生成研究综述 [J]. 计算机应用 ,2021(5):1227-1235.

[14]Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big?[C]Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,2021:610-623.

[15] 欧阳康 . 基层数字形式主义：表现、成因与对策 [J]. 国家治理 ,2024(13):10-14.

[16] 曾维和 .DeepSeek 嵌入政务智能生态的风险及治理 [J]. 理论探讨 ,2025(5):37-47.

[17] 林锋 . 持续训练在机器学习平台中的应用研究 [D]. 上海：上海交通大学电子信息与电气工程学院 ,2021.

[18] 王静 , 张红 . 人工智能在数字政府建设中的应用及其法治化路径 —— 基于 DeepSeek 在数字行政中应用场景的分析 [J]. 中外法学 ,2025,37(3):625-643.

[19] 霍旻含 .AI“公务员”，表现很亮眼 [N]. 人民日报海外版 ,2025-02-24(8).

[20] 毛伟杰 . 动态能力视角下跨国零售企业供应链数字化转型路径研究 ——以沃尔玛为例 [J]. 广西民族大学硕士电子期刊 ,2025-06-16—2025-07-15,2025(7).

[21] 赵路 , 屠友军 . 杭州全面启动城市大脑 3.0 建设 [N]. 浙江日报 ,2025-04-01(1).

[22] 王燕 , 方晴 . 广州：全国首个城市“三维数字底座”将这样支撑数字政府改革 [N]. 广州日报 ,2022-08-25(1).

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License