ICS 35.240.99
CCS L67

团　　体　　标　　准

T/ISC XXX—XXXX

政务领域人工智能大模型统一服务平台成熟度模型

Maturity Model for Unified AI Large Model Service Platforms in the Government Sector

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

2026-04-16

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 协 会　　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：

本文件主要起草人：

引　　言

当前，全球数字化转型持续深化，人工智能已成为驱动数字经济发展、推进治理体系和治理能力现代化的核心技术支撑。大模型在语言理解、内容生成、智能交互等方面的技术能力不断成熟，为政务服务创新、政务效能提升提供了重要支撑。在我国数字政府建设进程中，政务大模型平台建设稳步推进，构建统一规范、安全高效、集约共享的政务领域人工智能大模型统一服务平台，已成为政务信息化高质量发展的重要方向。

当前政务大模型应用正从单点试点向平台化、集约化服务演进，行业发展仍面临突出问题：一是平台建设缺乏统一标准规范，算力、数据等资源分散独立，集约化利用水平不高；二是平台服务能力水平不一，模型调用、智能体应用开发等核心能力缺乏科学统一的成熟度评价依据；三是政务场景对安全合规、内容可控、审计溯源有严格要求，通用大模型相关规范无法全面适配政务领域专属监管与安全管控需求。

为破解上述问题，规范政务领域人工智能大模型统一服务平台建设，推动政务大模型应用高质量、规范化发展，制定本文件。本标准结合政务业务实际需求，构建包含算力调度、模型训练、模型推理、智能体应用开发、模型安全、模型运营六大核心能力的成熟度模型体系，明确平台成熟度分级规则与能力判定依据，为各级政务部门、平台建设与服务单位提供规范统一、科学可量化的建设指引与评估依据。

政务领域人工智能大模型统一服务平台成熟度模型

1 范围

本文件规定了政务领域人工智能大模型统一服务平台成熟度模型，包括六方面：一是算力调度服务能力，二是模型训练服务能力，三是模型推理服务能力，四是智能体应用开发服务能力，五是模型安全服务能力，六是 API 及运营服务能力。

本标准适用于：

a）政务领域人工智能大模型统一服务平台建设单位与服务提供商评估自身平台建设与服务能力；

b）政务管理部门对政务领域人工智能大模型统一服务平台提出建设、监管与验收要求；

c）第三方机构开展政务领域人工智能大模型统一服务平台成熟度评估与能力定级。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 45288.1—2025 人工智能 大模型 第1部分：通用要求

GB/T 45288.3—2025 人工智能 大模型 第3部分：服务能力成熟度评估

3 术语和定义

GB/T 41867—2022、GB/T 45288.3—2025、GB/T 45288.3—2025界定的以及下列术语和定义适用于本文件。为了便于使用，以下重复列出了GB/T 45288.1—2025中的某些术语和定义。

3.1

模型优化 model optimization

提升模型执行速度，泛化能力，或改善利益相关方所关心的其他特性的方法。

[来源：GB/T 41867—2022，3.2.19]

3.2

大模型 large-scale model

大规模深度学习模型 large-scale deep learning model

基于大量数据训练得到,具有复杂计算架构,能处理复杂任务,且具备一定泛化性的深度学习模型。

[来源：GB/T 45288.1—2025，3.1]

注：大模型的参数有其功能和模态决定，一般不低于 1 亿规模。大模型训练使用的数据总量受参数量的影响，达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

3.3

提示词 prompt

提示语

使用大模型进行微调或下游任务处理时,插入到输入样本中的指令或信息对象。

[来源：GB/T 45288.1—2025，3.5]

3.4

大模型服务 large-scale model service

开发、应用大模型及大模型系统的服务，以及以此为手段提供支持需方业务活动的业务。

[来源：GB/T 45288.3—2025，3.2]

注：大模型系统是大模型与大模型平台的集成，是与大模型服务相关的活动、过程等的集合。

4 符号和缩略语

下列符号和缩略语适用于本文件。

API 应用程序编程接口（Application Programming Interface）

CPT 继续预训练（Continual Pre-Training）

SFT 有监督-模型微调（Supervised Fine-Tuning）

DPO 有监督-直接偏好优化（Direct Preference Optimization）

5 政务领域人工智能大模型统一服务平台成熟度模型框架

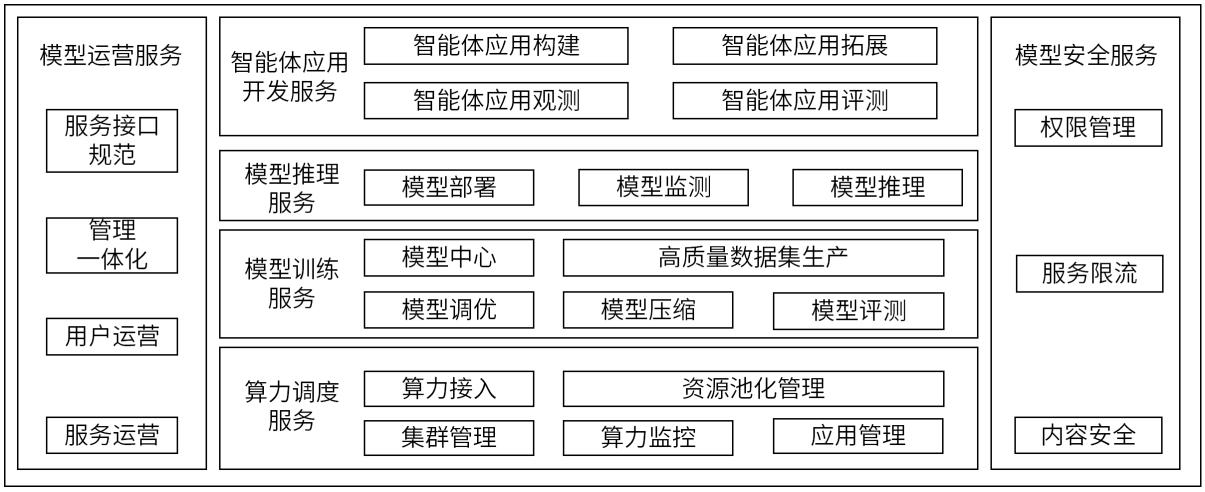

图 1 政务领域人工智能大模型统一服务平台成熟度模型框架图

6 政务领域人工智能大模型统一服务平台成熟度模型分级定义

表 1 政务领域人工智能大模型统一服务平台成熟度模型分级定义表

级别	英文	中文	定义
1 级	Exploration level	探索级	探索级政务大模型统一服务平台初步尝试构建大模型服务平台基础能力，在有限资源下实现单一模型的简单调用与问答功能。

			平台尚无完整的训练-部署-评测闭环，数据管理依赖人工处理，知识以静态规则为主，智能体应用尚未形成，安全与运营能力薄弱。整体处于技术验证阶段，仅支撑小范围内部测试，未面向实际政务业务开放服务。
2级	Basic Level	基础级	基础级政务大模型统一服务平台具备标准化的模型接入与基础服务能力，支持合规大模型的统一纳管和API调用。可开展监督微调等基本模型优化，部署流程初步自动化，并提供简单的Token用量监控。智能体应用可通过低代码方式搭建基础对话流程，覆盖政策咨询、常见问题解答等单一场景。平台建立初步权限控制与内容过滤机制，但复杂任务仍需人工介入，跨系统协同与动态知识更新能力有限。
3级	Innovation Level	创新级	创新级政务大模型统一服务平台构建训推一体、弹性调度的平台架构，支持模型压缩、多轮对话优化与自动/人工结合的评测体系。智能体应用可集成知识库、政务工具插件与MCP服务，覆盖坐席辅助、公文生成、办事引导等多个政务场景，并支持多终端接入。平台实现知识动态管理与检索增强，具备内容安全干预、服务限流与调用链路观测能力，初步形成“开发—部署—反馈—优化”的服务闭环。
4级	Excellence Level	优秀级	优秀级政务大模型统一服务平台能力深度融入政务业务流程，算力资源可根据业务峰谷智能调整，模型训练与推理关键环节实现高效协同。知识库与事项办理、审批流等核心业务深度融合，支持高精度、上下文感知的智能服务。服务体系打通跨部门协作链条，可支撑政策解读、民生诉求分办、应急响应等复杂场景，并具备智能预判与决策辅助能力。运营层面建立多维度评估模型，服务效率、准确性与安全性显著提升，具备区域级规模化推广价值。
5级	Leading Level	引领级	引领级政务大模型统一服务平台实现平台能力的全面智能化与生态化，构建覆盖多模态、多场景、多层级的政务大模型服务中枢。平台不仅支持主动服务、个性化推荐、智能引导等创新模式，更形成“感知—响应—治理”一体化的智能治理体系，全面赋能跨域复杂政务应用。在安全合规、技术先进性、服务体验和生态影响力等方面达到行业领先水平，成为全国政务智能化基础设施的标杆，具备广泛示范效应和持续引领能力。

7 算力调度服务

7.1 算力接入

a) 支持算力资源信息的自动采集与注册，包括CPU、AI加速卡、IP地址、服务器机器名称等；
b) 支持机器配置信息查询；
c) 支持对算力集群的创建功能，包括集群主节点及集群工作节点的注册，所创建集群能够汇总各节点所提供的算力资源，如GPU卡、CPU、内存数；
d) 支持根据任务需求自定义配置算力资源，包括但不限于AI加速卡类型、内存、显存等；

e） 支持多维度的算力调度策略配置，包括性能优先、能效优先、成本优先等调度模式；

7.2 资源池化管理

a） 支持物理GPU切分，按算力与显存两个维度，实现灵活的GPU资源切分，以满足不同AI任务的算力需求。
b） 支持资源池的动态扩容/缩容，新增GPU节点可自动加入资源池并按策略参与全局调度；
c） 支持建立硬件兼容性抽象层，屏蔽不同GPU厂商（如NVIDIA、华为昇腾）的驱动与架构差异，实现池内异构资源的统一纳管；
d） 支持动态异构资源调度能力，根据训练任务阶段自定义分配CPU、GPU、NPU资源，支持查看训练集群资源利用率，包括CPU/GPU使用率、内存使用率、存储使用率等；
e） 支持监控池化GPU的健康状态（如温度、ECC错误率），自动隔离故障设备并触发替换策略。

7.3 集群管理

a） 支持快速筛选查询目标集群列表，支持对集群操作状态、运行状态、集群资源快速查询；
b） 支持对算力集群的节点扩容和缩容操作，支持对已有负载节点的扩缩容校验；
c） 支持集群及算力资源多维度查询功能，支持查询集群数、节点数、服务数及AI加速卡类型，AI加速卡、CPU、内存总量、分配量及使用量；
d） 支持自动部署集群应用服务网关，支持查看集群应用服务网关访问路径。

7.4 算力监控

a） 支持选择不同集群，监控不同集群的算力资源使用情况；
b） 支持节点和AI加速卡的选择和信息查询，如AI算力使用率、CPU、显存占用情况，支持查看实时信息及历史趋势；

7.5 应用管理

a） 支持应用配置时，选择算力集群、工作负载配置和服务校验；
b） 支持查询工作负载列表和实例列表，包括AI加速卡类型、AI加速卡数、CPU、内存资源占用等；
c） 支持将适配不同AI硬件的AI服务由系统调度部署到对应类型的AI算力节点；
d） 支持查看已创建的服务列表，支持查看服务所关联的pod数量、服务的集群内部地址以及集群外部访问地址；

8 模型训练服务

8.1 模型中心

a） 应支持政务大模型合规模型准入，支持通过网信办算法备案的模型准入管理，涵盖语言大模型、多模态大模型等类型，确保模型服务符合监管要求；
b） 应支持政务大模型总体概览，支持在控制台首页展示模型总量、模型列表、类型分布等，提供全局资源态势视图；
c） 应支持政务大模型元数据查看，支持查看模型详情，包含模型名称、训练数据规模、适用场景、版本迭代记录等关键信息；
d） 应支持政务大模型调用方式查看，支持展示模型的 API 调用协议及认证配置说明，降低开发者接入门槛；

e) 应支持政务大模型体验，支持在模型中心内嵌交互式测试窗，允许用户输入样例数据实时验证模型效果，无需编码即可完成功能评估；
f) 应支持专用小模型统一纳管，提供 OCR、ASR 等专用小模型的集中化管理能力。

8.2 高质量数据集生产

8.2.1 数据集接入

a) 应支持接入多种来源的政务数据，如块存储、文件存储和对象存储等；
b) 应支持接入结构化和半结构化数据，如 csv、tsv、txt、parquet 等数据类型；
c) 宜支持政务公文模板数据导入，支持提供符合政府公文格式（如通知、报告、函件等）的数据集上传模板，便于进行政务文本生成模型的训练与微调。

8.2.2 数据预处理

a) 应支持结构化数据的清洗,如数据拆分、异常值检测、缺失值填充等；
b) 应支持非结构化数据的清洗,根据特定规则剔除不符合要求的非结构化数据,如内容去重等；
c) 应支持自定义数据预处理功能,如用户自定义预处理算法等。

8.2.3 数据标注

a) 应支持多种数据类型的标注工具或模板,如文本类、表格类、图片类、音视频类等;
b) 应支持对标注标签、标注属性等标注信息的管理,如编辑、删除和查询等;
c) 应支持可视化标注,标注信息在原始数据直观呈现;
d) 应支持团队标注的管理,如任务管理、人员管理等:
e) 应支持对标注的评估,如准确性、有效性等;
f) 应支持对标注数据、标签等标注结果导出;
g) 宜支持智能标注,如调用算法或外部服务自动标注数据、通过训练算法自动标注等。

8.2.4 数据集管理

a) 应支持政务数据类型管理，明确政务领域数据集是用于 SFT、DPO、CPT 训练的训练集或是评测集，并针对不同政务数据集具备对应数据集上传模板要求，支持提供如 excel、alpaca、sharegpt 等模板文件；
b) 应支持对政务训练数据和评测数据的分级管理，支持数据上传、删除等操作，支持本地上传导入方式，支持多种文件类型的数据集导入，并明确文件导入的最大数量。

8.2.5 数据集质检

a) 应支持政务领域数据质检任务的配置能力，支持对已上传训练数据集的质检数量选择、质检人员分配及可用率阈值设置，满足质检任务的前置参数化配置要求；
b) 应支持政务领域数据集质检标注与任务管理能力，支持对质检任务中每条数据的“通过”“不通过”等打标操作，并支持质检任务完成后的提交或删除操作，满足质检流程闭环管理的能力要求。

8.2.6 任务管理

a) 应支持政务领域数据处理任务的全生命周期管理能力，支持数据处理任务的新建、查看任务详情、停止、运行、重新运行、删除等操作；

b) 应支持对政务领域数据集处理任务的配置，包括输入任务名称、选取数据工具、选择数据集、进行工具参数、运行时参数配置；
c) 应支持政务数据任务列表的可视化管理能力，查看已分配数据流任务的处理情况和进度；
d) 应支持政务数据任务执行过程中的异常情况进行监控与排除，提供日志查询浏览功能。

8.3 模型调优

8.3.1 创建模型调优任务

a) 应支持创建政务大模型调优任务，支持输入调优任务名称、选择调优数据集、选取基模型、选择训练方式、选择调优方法、设置调优工具参数、运行参数等。
b) 应支持 API 或命令行方式进行政务大模型调优；
c) 宜支持控制台低代码方式进行政务大模型调优。

8.3.2 调优方式

a) 应支持多种政务大模型调优训练方式，支持选择不同的基模型及个性化调优方式，包括但不限于监督微调、强化学习等；
b) 应支持全参微调和部分参数微调的微调方法；

8.3.3 训练数据

a) 应支持政务大模型训练数据选择，支持根据数据管理库中添加数据集，并可指定具体的训练数据集。

8.3.4 超参配置

a) 应支持政务大模型训练超参配置，包括但不限于配置循环次数、学习率、批次大小、序列长度、L2 正则化等；
b) 应支持政务大模型训练工具参数配置，包括但不限于数据集路径、模型名称、训练类型等。

8.3.5 训练管理

a) 应支持任务列表查看政务大模型训练详情，可查看任务类型、调优任务、基础模型、训练状态等信息；
b) 应支持可视化政务大模型训练任务详情，包括不限于可视化损失率、模型训练过程中实时产生的日志、资源消耗情况等。

8.4 模型压缩

8.4.1 模型量化

a) 应支持对训练完成的政务大模型进行模型量化，支持训练任务选择、模型选择、量化精度选择、量化方法选择、参数配置；
b) 应支持包括 Int4、Int8 等量化方式，并支持多种量化方法。

8.4.2 模型剪枝

a) 应支持模型剪枝能力，通过结构化或非结构化剪枝技术移除冗余参数，实现模型体积压缩与推理效率提升。

8.4.3 模型蒸馏

a) 应支持知识蒸馏能力，通过将大型教师模型的知识迁移至小型学生模型，实现模型规模减小的同时保持较高性能表现。

8.4.4 压缩管理

a) 应支持混合压缩策略，支持将剪枝、量化、知识蒸馏等多种压缩技术按需组合（如先剪枝后量化、蒸馏后量化等），并提供协同优化的流程或工具链；
b) 应支持提供详细的压缩过程日志和结果报告，包括但不限于压缩配置、资源消耗、压缩率、精度变化等关键指标；
c) 应支持在压缩过程中或压缩后对政务大模型精度进行查看，如准确率、F1 值、BLEU 等，进行自动化评估，并提供与原始模型的对比结果；
d) 应支持提供政务大模型压缩前后的资源消耗，如模型大小、内存占用、FLOPs、推理延迟的量化对比分析能力；
e) 应支持提供清晰的 API 接口和/或命令行工具，使得压缩流程易于集成到现有的政务大模型训练和部署流水线中。

8.5 模型评测

8.5.1 打分评测

a) 应支持政务大模型评测维度模板创建，支持自定义设置评测参数模板，包括但不限于模板名称、测评维度方向名称、打分量级、维度说明、量级描述、量级对应分值、分值注释、增加维度等；
b) 应支持测评维度模板管理，支持对已成功创建的政务大模型评测维度模板进行查看、删除、修改等操作。

8.5.2 人工测评

a) 支持人工测评任务管理，支持展示任务名称、任务状态、评估对象、创建人、创建时间，可对任务进行编辑、删除等操作；
b) 支持创建人工测评任务，支持输入任务名称、评测类型、填写描述、选择评测数据，选择评估方法、评估指标和评估量级；
c) 支持查看评测状态，支持对执行中评测进行删除操作；
d) 支持保存评测结果并生成人工测评报告；
e) 支持查看人工评测任务的评测类型、评测数量、人工标注量、未标注量以及评测进度信息等；
f) 支持预置评估指标场景，支持根据预设的评测规则进行评测，评测规则包括但不限于通用场景、文本生成场景、分类任务场景等，对参评模型基于评测数据生成的输出进行自动评分。

8.5.3 自动测评

全过程无需人工参与，通过内置的深度学习指标（包括BLEU、ROUGE和F1）和AI评测器，自动对模型的输出效果进行评分。

a) 支持自动化评测任务管理，支持查看任务详情、查看评测结果以及删除等操作；
b) 支持自动评测任务的创建，支持输入任务名称、选择训练任务、评测模型、评测数据、评测规则、选择集群、卡类型、输入副本数量、每副本卡数；
c) 支持根据不同的任务，选择评测模型进行智能评测，基于 Prompt 评分模板对参评模型的输出进行评分；

d) 支持生成自动评测数据报告，支持单个模型评测结果或多个模型对比表现结果生成，自动评测报告应包含数据集 ID、评测问题、参考答案、推理答案、指标类型、打分指标和结果等。

9 模型推理服务

9.1 模型部署

9.1.1 模型选择

a) 应支持预制模型库中政务大模型的一键部署能力，具备标准化模型的快速上线与集成功能；
b) 应支持自定义政务大模型的全流程部署能力，涵盖从模型训练、调优到版本迭代的端到端服务化部署功能；
c) 应支持政务大模型部署时自定义选择算力集群及算力资源池，支持多资源池同时部署同一模型服务；
d) 应支持政务大模型部署时可选择多副本，具备多运行实例部署同一模型服务，保障模型服务的高可用；
e) 应支持政务大模型模型服务、模型运行实例的查看，实现模型运行状态、模型运行实例的展现，确保实时查看可用的模型服务及模型服务并发。

9.1.2 部署调用

a) 应支持政务大模型部署后的多协议适配调用能力，要实现与业务系统、开发框架及应用工具的兼容对接，满足自然语言处理、代码生成、知识检索等复杂场景的模型调用需求；
b) 应支持政务大模型模型服务的申请，并针对政务大模型模型服务申请进行模型服务实例 apikey 的下发与授权；
c) 应支持政务大模型模型服务实例查看，支持政务大模型模型服务实例按日展示总调用次数、成功调用次数、输入 tokens 数、输出 tokens 数。

9.1.3 部署服务扩缩容

a) 应支持部署服务的弹性扩缩容能力，通过自助调节实例数量动态调整计算资源，确保高并发场景下的性能稳定与资源成本平衡。

9.1.4 部署服务下线

a) 应支持部署服务的自动化下线能力，要实现通过控制台或接口触发服务终止、资源回收及数据清理操作，确保服务下线后无残留资源占用和数据泄露风险。

9.2 模型监测

a) 应支持调用记录，支持查看政务大模型在过去一段时间内的使用情况，包括但不限于支持查看调用次数和调用量的趋势和波动，支持查看失败次数和失败率，及时发现异常；
b) 应支持性能指标，支持查看政务大模型的多种常见性能指标，包括但不限于首 Token 延时、调用时长、RPM（每分钟调用次数）、TPM（每分钟消耗 Token 数）及失败率等。

9.3 模型推理

a) 应支持推理引擎，实现政务大模型的高性能推理，如：vLLM、TensorRT-LLM、MindIE、DeepSpeed（DeepSpeed-MII 、DeepSpeed-Inference）、TGI（Text Generation Inference）等；

b) 应支持推理效果参数调节，以控制输出结果确定性、多样性，如：Top-k、Top-p、Temperature 等；
c) 应支持流水线并行、张量并行等多种并行技术；
d) 应支持混合专家（Mixed-of-experts, MoE）并行；
e) 应支持 kvcache 中间件进行推理加速；
f) 应支持批处理策略优化，实现推理吞吐提升，如：连续批处理（Continuous Batching）、 动态批处理（Dynamic Batching）等；
g) 应支持投机采样方式部署，支持 token 级别采样，从而实现推理加速；
h) 应支持对模型推理服务的状态查询和展示，如内存、显存、I/O 等指标；
i) 应支持基于平台算力承载能力与资源负载阈值，配置化管控用户输入规模，支持设置输入字符长度上限、文本分片阈值与单轮输入数据量限制。

10 智能体应用开发服务

10.1 智能体应用构建

10.1.1 模型配置

a) 应支持参数差异化配置能力，支持开放政务大模型参数自定义面板，如温度系数、最长回复长度、上下文轮数等；
b) 应支持多模型灵活选型能力，支持提供多种政务大模型选择，支持在创建流程中选择不同政务大模型，并允许用户根据场景需求自主选定；
c) 应支持对特定政务大模型启用增强能力，如开启思考模式提升复杂任务表现，同时自动隐藏当前模型未支持的参数项。

10.1.2 低代码开发

a) 应支持多节点低代码配置能力，支持通过可视化界面以低代码方式配置多样化节点类型，包括但不限于开始/结束节点、知识库调用节点、大模型服务节点、意图分类节点、条件判断节点、函数计算节点、循环控制节点、自定义组件节点、MCP 节点等；
b) 应支持政务业务流程节点配置，在低代码画布中预置或支持自定义“表单填报”、“多级审批”、“公文套红”、“电子签章调用”、“数据报表生成”等政务通用业务流程节点，方便快速构建办事流程类应用。
c) 应支持即时交互验证能力，支持在政务场景应用创建完成后提供嵌入式测试窗口，允许用户实时输入问题并查看模型响应，以快速验证智能体基础功能；
d) 应支持模板参数继承与改写能力，当通过模板创建政务场景应用时，支持自动继承原模板的模型配置与参数，同时允许用户在应用管理界面重新调整模型选择及参数值；
e) 应支持工作流 DSL（或 YML 文件）的导入与导出能力，在工作流配置页面支持“导入 DSL（或 YML 文件）”和“导出 DSL（或 YML 文件）”功能，保障配置的可迁移性与可复用性。

10.1.3 智能体应用创建

智能体应用是一种对提示词依赖较高，基于上下文对话，自主决策并调用工具来完成复杂任务的对话式 AI 应用。

a) 应支持智能体应用创建，支持基于上下文对话，集成 RAG、工具插件、应用组件、MCP 及长期记忆等功能；

b) 应支持模板化智能体管理，具备应用模板仓库，提供模板搜索、预览及一键复制功能，复制时需触发配置弹窗要求用户填写新应用名称和部署空间。

c) 应支持典型政务场景智能体模板，具备面向政务场景的应用模板仓库，提供如“政策咨询与解读机器人”、“民生诉求智能分办助手”、“公文内容校对与润色工具”、“会议纪要生成助手”、“营商环境智能问答”、“应急指挥信息摘要生成”等模板，用户可一键复制并基于自身业务数据进行配置。

10.1.4 工作流应用创建

工作流应用将复杂的任务拆分成一系列有序执行的步骤，以降低系统复杂度。

a) 应支持任务型工作流创建能力，通过依赖单一系统输入变量（query），适用于文本搜索、数据处理、翻译、问答等一次性任务。

b) 应支持对话性工作流应用的创建能力，支持多轮对话上下文管理能力，支持适配智能客服、虚拟助手等场景，通过历史记录实现情感化响应与深度问题推理。

c) 应支持创建智能体群组，支持创建一个包含多个智能体的群组，支持配置多个子智能体的执行顺序与信息传递方式；

d) 应支持政务领域场景智能体样板间服务，支持智能体的快速创建，并对对话开场白、下一步问题建议、文字转语音、引用和归属、内容审查等内容进行配置；

e) 应支持全链路测试验证能力，支持在画布中直接启动测试功能，实时验证节点逻辑、数据流转换及智能体协同效果；

f) 宜支持政务协同工作流，支持创建跨部门、多角色的复杂审批与处理流程，如“企业补贴申领审批流”、“市民投诉建议处理流”，并确保流程状态可追踪、责任可追溯。

g) 应支持政务热线场景的对话型工作流，支持在对话中集成业务系统查询接口（如社保、公积金、税务），实现“问答即办理”的智能化服务。

h) 宜支持政务服务“边聊边办”对话型工作流，支持在对话中进行情形引导、填写表单、提交材料、签字确认等流程，支持调用相关业务接口（如办事要点查询、办事条件预检、表单数据查询等），为办件人线上办事提供与线下窗口办事类似的交互办事体验。

i) 宜支持辅助受理审批工作流，为工作人员提供业务受理审批知识解答、材料智能预审、文书自动生成等服务，支持申报材料内容识别、审批依据自动匹配、关联数据智能核验，提高受理审批效率。

10.2 智能体应用拓展

10.2.1 Prompt 提示词配置

10.2.1.1 Prompt 模板创建与管理

a) 应支持预置 Prompt 模板调用能力，支持提供预置 Prompt 模板库，支持查看模板详情（内容、变量、ID）并通过界面操作直接复制或调用系统模板；

b) 应支持自定义 Prompt 模板动态变量定义能力， 支持通过 ${变量名} 语法声明动态变量，并自动识别提取模板中的变量形成结构化列表；

c) 应支持 Prompt 模板唯一标识自动生成能力，在保存新建的自定义 Prompt 模板时，支持自动分配全局唯一的模板 ID，用于后续识别、检索与调用；

d) 应支持自定义 Prompt 创建能力，支持自定义输入 Prompt，并支持大模型辅助优化成提示词框架样式。

10.2.1.2 Prompt 样例库管理

a) 应支持样例数据批量导入能力，支持提供标准化 Excel 模板下载及文件上传接口，支持批量导入符合规范；
b) 应支持样例库动态维护能力，支持提供界面化操作支持对样例库及库内单条样例进行增删改查，确保样例数据的实时性和准确性。

10.2.1.3 Prompt 优化

a) 应支持政务大模型辅助的 Prompt 优化能力，支持大模型进行多轮优化 Prompt；
b) 应支持基于样例数据的 Prompt 反馈优化，支持提供的输入输出样例数据或总结提示词添加至 Prompt 中，并通过多轮自动化评估、反思和优化，生成符合期望的 Prompt 内容；
c) 应支持优化结果的模板化与应用绑定，支持优化后的 Prompt 可保存为 Prompt 模板或直接用于创建智能体应用，确保优化效果可复用并适配实际业务场景。

10.2.2 知识库创建

a) 应支持多格式数据导入能力，支持提供本地上传接口，支持多种格式导入，如.docx、MD、HTML、PDF、WPS 等格式的私有知识文件至智能体应用数据池；
b) 应支持非结构化数据（如文本、HTML、PDF 等）处理能力，支持内置智能文本切分算法，自动解析上传文件的原始内容并分割为可检索的知识片段；
c) 应支持政务知识库元数据定义能力，支持允许用户创建知识库时自定义名称与描述，明确标识知识库领域范围；
d) 应支持政务文件标签配置，支持控制台调试知识库设置标签或请求参数中指定标签；
e) 宜支持政务知识库的统建统管，支持将各部门的政策文件、法律法规、办事指南、历史问答等知识进行统一汇聚、标准化处理与授权共享，形成平台级的政务知识底座；
f) 宜支持数据库接入能力，提供筛选指定数据表、字段，同时支持全量同步与增量同步；
g) 宜支持自定义文本切分规则，包括按标题、按自然段、按章节条等多种文本切分方案；
h) 宜支持多模态文档解析能力，支持对视频、音频、图片、复杂表格等数据的解析能力。

10.2.3 知识检索增强（RAG）集成

a) 应支持构建状态监控能力，支持提供知识库构建进度监控机制，确保用户感知知识库就绪状态；
b) 应支持一键式知识绑定能力，支持智能体管理界面选择对应知识库，支持自动注入检索指令至 Prompt 并支持从列表勾选关联知识库；
c) 应支持政务知识检索增强效果优化能力，包括不限于针对政务大模型理解有误、政务知识未检索到、检索不相关、检索不完整等结果具备优化策略。

10.2.4 模块化应用组件

a) 应支持政务智能体或工作流应用的模块化封装，支持在应用管理界面将智能体或工作流应用发布为独立组件，供其他应用调用；
b) 应支持组件的接入与绑定，支持在政务智能体应用编辑界面添加官方或自定义组件，并完成绑定；
c) 应支持参数传参方式的差异化配置，实现业务透传模式下由智能体使用者或工作流上游节点提供参数值，模型识别模式下在政务智能体场景中由大模型基于上下文自动推断填充参数，以及在工作流场景中强制通过上游节点显式传递参数值，确保输入的确定性与可控性。

10.2.5 工具与服务拓展

a) 应支持统一的工具/服务管理界面，支持官方工具库、MCP 服务库、自定义插件的统一管理；
b) 应支持多协议工具接入，支持遵循 MCP 协议的标准服务，也支持传统 API/SDK 方式的插件接入；
c) 应支持工具/服务的授权与调用统计，支持基于 APIkey 的访问控制及调用次数分析；
d) 应支持政务专用工具/服务，如政务政策查询、表单预填、材料检查等场景化能力；
e) 应支持工具在政务智能体应用或工作流中的添加、测试与发布。

10.3 智能体应用观测

a) 应支持全链路调用追踪能力，支持实时追踪政务应用内各流程工具的执行顺序与时间戳，通过可视化链路图展示节点调用路径及耗时分布；
b) 应支持为政务智能体分配所需模型 API 服务运行情况的管理监控、统计分析等运营服务，支持对模型服务所需算力及模型进行动态监管和模型服务扩容；
c) 应支持智能体分配政务大模型推理可观测能力，支持监控大模型服务的响应延迟、运行状态、单次调用用时及调用时间点，并在观测面板动态更新健康指标；
d) 应支持智能体分配政务大模型思考过程查看能力，思维链可视化能力，包括决策依据、知识检索路径及临时结论生成步骤；
e) 应支持政务应用级资源聚合能力，支持在观测管理界面统计应用总量、总调用次数、LLM 累计 Token 消耗量及平均调用时长，形成全局负载视图；
f) 应支持细粒度观测管理能力，支持按应用维度筛选观测数据，提供应用列表的调用详情下钻分析入口，关联展示各实例的实时性能指标。

10.4 智能体应用评测

a) 应支持自动评测，支持利用政务大模型基于指定应用知识库来自动创建评测集，评估智能体的回答并产出评测报告与调优建议。
b) 应支持人工评测智能体应用，支持通过针对特定业务场景来人工构建评测集，并对应用的回答进行人工分析与评分，产出评测报告。
c) 应支持多应用横向评测对比，支持查看各应用总体得分、问题汇总等，支持查看各应用的性能差异对比。

11 模型安全服务

11.1 权限管理

11.1.1 用户管理

a) 应支持主账号创建，主账号拥有访问和管理的所有权限；
b) 应支持子账号创建，支持通过主账号或拥有特定系统策略的子用户在控制台中创建、管理并进行授权，实现不同子用户拥有不同资源访问权限的目的。

11.1.2 角色管理

a) 应支持用户权限配置，支持选择管理角色和访客角色选择，使其在自定义业务空间内获得角色对应的相关功能的使用权限；
b) 应支持权限自定义策略配置，支持可视化或脚本编辑权限操作，并支持选择新增授权用户执行。

11.2 服务限流

a) 应支持语言类模型限流保护功能，支持统计政务大模型每分钟调用次数（QPM)或每分钟消耗Token数(TPM)进行统计，并按照所有调用该模型的总和计算限流，如果超出调用限制，用户的调用请求将会因为限流而失败，用户需等到不满足限流条件时才能再次调用；
b) 宜支持多模态政务大模型限流保护功能，支持根据作业提交接口每秒的请求次数或同时处理中任务数量进行统计，并按照所有调用该模型的总和计算限流，如果超出调用限制，用户的调用请求将会因为限流而失败，用户需等到不满足限流条件时才能再次调用。

11.3 内容安全

a) 应支持输入输出内容安全管控能力，满足通过政务大模型输入输出管控工具对内容进行实时过滤与风险拦截，确保内容合规性与安全性；
b) 应支持红线知识库标准化过滤能力，满足基于预设敏感规则库对输入输出内容进行筛查，保障内容符合监管要求与政治中立性；
c) 应支持对不同内容的安全问题进行分类，如国家安全、公共安全、社会伦理道德等类别，并对每种类别的回复作出限定；
d) 应支持输入内容白名单动态管理能力，满足用户对敏感词库、干预词库的自主修改、更新及白名单输入控制，实现策略的灵活适配；
e) 应支持涉政问题标准化回复能力，满足在对话涉及领导人、重大政策等敏感话题时，提供客观、准确、政治中立的标准回复内容；
f) 应支持突发安全事件干预能力，满足通过语义干预、文本干预、关键词干预等手段快速阻断风险输出，降低安全事件影响范围；
g) 应支持输出内容安全检测与处置能力，满足对政务大模型输出内容进行审核，并提供兜底回复、不上屏等处置方案，防止违规内容外泄；
h) 应支持安全策略动态更新能力，满足策略模板管理、安全词表维护及白名单输入控制，确保安全规则持续优化与业务需求匹配；
i) 应支持输入输出日志管理能力，满足对政务大模型系统输入输出日志的记录、分析、存储与管理，实现操作过程的可追溯性与可审计性；
j) 应支持内容生成溯源能力，满足对政务大模型生成内容的来源追踪与生成路径记录，确保内容生成过程透明化与责任可归因。

12 模型运营服务

a) 应具备统一的服务接口规范，支持服务的动态编排组合，同时提供可复用的服务封装能力；
b) 应具备管理一体化能力，能够实现对推理服务的管理和调度，包括服务的创建、分配、监控和优化等；
c) 应具备用户支持与服务能力，提供用户支持和服务，包括技术咨询、培训服务、使用指南、在线测试、故障处理等，帮助用户更好地使用推理服务；
d) 应具备服务运营数据分析能力，能够对推理服务的运营数据进行收集、分析和挖掘，为优化服务提供数据支持；
e) 应支持政务大模型的调用量统计分析，不限于用量排行、调用成功率、使用率等指标；
f) 应支持政务大模型服务的性能统计分析，包括资源占用、服务稳定性统计等指标；
g) 应支持政务大模型以及政务AI原生应用的线上使用情况统计与分析能力。

13 政务领域人工智能大模型统一服务平台成熟度模型分级判定依据

13.1 算力监控服务分级判定依据

表 2 算力监控分级划分依据表

级别 要求		探索级	基础级	创新级	优秀级	引领级
算力接入	a	√	√	√	√	√
	b	√	√	√	√	√
	c		√	√	√	√
	d			√	√	√
	e				√	√
	f					√
资源池化管理	a		√	√	√	√
	b			√	√	√
	c				√	√
	d					√
	e	√	√	√	√	√
集群管理	a	√	√	√	√	√
	b			√	√	√
	c				√	√
	d					√
算力监控	a	√	√	√	√	√
	b			√	√	√
应用管理	a		√	√	√	√
	b	√	√	√	√	√
	c			√	√	√
	d				√	√

13.2 模型训练服务分级判定依据

表 3 模型训练服务分级判定依据

级别 要求		探索级	基础级	创新级	优秀级	引领级
模型中心	a		√	√	√	√
	b	√	√	√	√	√
	c	√	√	√	√	√
	d		√	√	√	√
	e			√	√	√
	f				√	√

高质量数据集生产 - 数据集接入	a	√	√	√	√	√
	b	√	√	√	√	√
	c			√	√	√
高质量数据集生产 - 数据预处理	a	√	√	√	√	√
	b	√	√	√	√	√
	c			√	√	√
高质量数据集生产 - 数据标注	a		√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d				√	√
	e					√
	f			√	√	√
	g					√
高质量数据集生产 - 数据集管理	a		√	√	√	√
	b		√	√	√	√
高质量数据集生产 - 数据集质检	a			√	√	√
	b			√	√	√
高质量数据集生产 - 任务管理	a	√	√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d				√	√
模型调优 - 创建模型调优任务	a		√	√	√	√
	b		√	√	√	√
	c			√	√	√
模型调优 - 调优方式	a		√	√	√	√
	b				√	√
模型调优 - 训练数据	a		√	√	√	√
模型调优 - 超参配置	a			√	√	√
	b			√	√	√
模型调优 - 训练管	a	√	√	√	√	√

理						
	b			√	√	√
模型压缩 - 模型量化	a			√	√	√
	b			√	√	√
模型压缩 - 模型剪枝	a			√	√	√
模型压缩 - 模型蒸馏	a				√	√
模型压缩 - 压缩管理	a				√	√
	b					√
	c				√	√
	d					√
	e					√
模型评测 - 打分评测	a			√	√	√
	b			√	√	√
模型评测 - 人工测评	a		√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d			√	√	√
	e				√	√
	f					√
模型评测 - 自动测评	a			√	√	√
	b			√	√	√
	c				√	√
	d					√

13.3 模型推理服务分级判定依据

表 4 模型推理服务分级判定依据表

级别 / 要求		探索级	基础级	创新级	优秀级	引领级
模型部署 - 模型选择	a	√	√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d				√	√
	e	√	√	√	√	√
模型部署 - 部署调	a		√	√	√	√
	b		√	√	√	√

用	c			√	√	√
模型部署 - 部署服务扩缩容	a			√	√	√
模型部署 - 部署服务下线	a		√	√	√	√
模型监测	a	√	√	√	√	√
	b			√	√	√
模型推理	a	√	√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d				√	√
	e			√	√	√
	f				√	√
	g					√
	h	√	√	√	√	√
	i				√	√

13.4 智能体应用开发服务分级判定依据

表 5 智能体应用开发服务分级判定依据表

级别 / 要求		探索级	基础级	创新级	优秀级	引领级
智能体应用构建 - 模型配置	a		√	√	√	√
	b		√	√	√	√
	c			√	√	√
智能体应用构建 - 低代码开发	a		√	√	√	√
	b			√	√	√
	c		√	√	√	√
	d			√	√	√
	e				√	√
智能体应用构建 - 智能体应用创建	a		√	√	√	√
	b			√	√	√
	c			√	√	√
智能体应用构建 - 工作流应用创建	a			√	√	√
	b			√	√	√
	c				√	√
	d			√	√	√
	e			√	√	√
	f				√	√

	g					√
	h					√
	i					√
智能体应用拓展能力-Prompt模板创建与管理	a		√	√	√	√
	b			√	√	√
	c				√	√
	d					√
智能体应用拓展能力-Prompt样例库管理	a			√	√	√
	b				√	√
智能体应用拓展能力-Prompt优化	a				√	√
	b				√	√
	c					√
智能体应用拓展能力-知识库创建	a		√	√	√	√
	b		√	√	√	√
	c		√	√	√	√
	d			√	√	√
	e				√	√
	f				√	√
	g				√	√
	h					√
智能体应用拓展能力-RAG集成	a			√	√	√
	b			√	√	√
	c			√	√	√
智能体应用拓展能力-模块化应用组件	a			√	√	√
	b			√	√	√
	c				√	√
智能体应用拓	a			√	√	√
	b			√	√	√

展能力-工具与服务拓展	c			√	√	√
	d				√	√
	e			√	√	√
智能体应用观测	a			√	√	√
	b			√	√	√
	c				√	√
	d				√	√
	e				√	√
	f					√
智能体应用评测	a			√	√	√
	b			√	√	√
	c				√	√

13.5 模型安全服务分级判定依据

表 6 模型安全服务分级判定依据表

级别 / 要求		探索级	基础级	创新级	优秀级	引领级
权限管理-用户管理	a	√	√	√	√	√
	b		√	√	√	√
权限管理-角色管理	a		√	√	√	√
	b			√	√	√
服务限流	a		√	√	√	√
	b			√	√	√
内容安全	a		√	√	√	√
	b		√	√	√	√
	c			√	√	√
	d			√	√	√
	e			√	√	√
	f				√	√
	g			√	√	√
	h				√	√
	i				√	√
	j					√

13.6 模型运营服务分级判定依据

表 7 模型运营服务分级判定依据表

级别 要求		探索级	基础级	创新级	优秀级	引领级
模型运营服务	a		√	√	√	√
	b		√	√	√	√
	c		√	√	√	√
	d			√	√	√
	e			√	√	√
	f				√	√
	g				√	√