ICS 35.240
CCS L 70

中华人民共和国国家标准化指导性技术文件

GB/Z 201—2026

人工智能　大模型选型和应用指南

Artificial intelligence—Large model selection and application guideline

2026-07-30 发布

国家市场监督管理总局
国家标准化管理委员会　发布

资料整理
火龙果科技 荐

目　　次

前言 …… Ⅲ
引言 …… Ⅳ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　缩略语 …… 1
5　选型与应用的原则与流程 …… 2
5.1　原则 …… 2
5.2　流程 …… 2
6　选型准备 …… 3
6.1　策略规划 …… 3
6.2　组织保障 …… 3
6.3　人员保障 …… 4
6.4　现状评估 …… 4
6.5　模型初选 …… 5
7　数据准备 …… 5
7.1　数据采集 …… 5
7.2　数据存储 …… 5
7.3　数据处理 …… 6
7.4　数据集建设 …… 6
8　选型与优化 …… 6
8.1　模型选型 …… 6
8.2　模型训练 …… 7
8.3　模型调优 …… 7
9　部署应用 …… 8
9.1　部署模式选择 …… 8
9.2　模型部署 …… 8
9.3　模型集成 …… 8
9.4　安全运维 …… 8
10　应用评估 …… 9
10.1　模型评估 …… 9
10.2　效益评估 …… 9
11　风险管理 …… 9

资料整理
火龙果科技 荐

11.1 组织能力建设 …… 9
11.2 风险管理目标 …… 9
11.3 风险识别及评估 …… 9
11.4 风险应对 …… 10
附录 A （资料性） 大模型选型参考信息 …… 11
A.1 算力估算 …… 11
A.2 模型分类 …… 11
A.3 模型参数 …… 11
A.4 模型性能 …… 12
A.5 模型压缩 …… 12
参考文献 …… 14

资料整理
火龙果科技 荐

前　言

本文件为规范类指导性技术文件。

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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、南方电网数字电网集团有限公司、阿里云计算有限公司、华为云计算技术有限公司、蚂蚁科技集团股份有限公司、国能信息技术有限公司、华北电力科学研究院有限责任公司、中国海洋石油集团有限公司、中国民航信息网络股份有限公司、北京智源人工智能研究院、深信服科技股份有限公司、烽台科技(北京)有限公司、中煤科工集团上海有限公司、中国铝业集团有限公司、浪潮软件集团有限公司、浪潮通信信息系统有限公司、中国移动通信集团有限公司、中国国际货运航空股份有限公司、北京中科金得助智能科技有限公司、中国南方电网有限责任公司超高压输电公司、普元信息技术股份有限公司、青岛海信信息科技股份有限公司。

本文件主要起草人：董建、张士宗、徐洋、高歌、李晨、李晋伟、徐小天、杨旭、李劲松、袁帅、苗宗利、彭骏涛、郑子木、彭晋、刘颖、叶润国、肖红梅、曹汐、张逸群、汪萌、王源涛、李蕾红、张丹丹、李学武、孟建、赵玉霞、李峰风、蔡洛维、于明亮、孙钦平、王杰、牛拓蒙。

资料整理
火龙果科技 荐

引　言

随着人工智能技术的快速发展，大模型在多个领域展现出巨大潜力，但其复杂性、多样性和潜在风险也对选型与应用提出了更高要求。本文件围绕选型原则与流程、选型准备、数据准备、选型与优化、部署应用、应用评估、风险管理建立实施框架，支持各类使用方根据实际业务需求和技术环境作出合理决策。本文件指导使用方进行人工智能大模型技术应用实施的选型准备、数据准备、选型与优化、部署应用、应用评估、风险管理全流程技术指引，旨在科学地、规范地提升使用方应用人工智能的水平。

资料整理
火龙果科技 荐

人工智能　大模型选型和应用指南

1　范围

本文件确立了人工智能大模型选型原则和应用流程，包括选型原则与流程、选型准备、数据准备、选型与优化、部署应用、应用评估和风险管理，为各类使用方提供了全生命周期的应用指导。

本文件适用于人工智能大模型技术的选型和应用。

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 23011—2022　信息化和工业化融合　数字化转型　价值效益参考模型

GB/T 35273—2020　信息安全技术　个人信息安全规范

GB/T 36073—2025　数据管理能力成熟度评估模型

GB/T 37964—2019　信息安全技术　个人信息去标识化指南

GB/T 45081—2024　人工智能　管理体系

GB/T 45288.1—2025　人工智能　大模型　第1部分：通用要求

GB/T 45288.2—2025　人工智能　大模型　第2部分：评测指标与方法

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 23011—2022、GB/T 36073—2025、GB/T 37964—2019、GB/T 45081—2024、GB/T 45288.1—2025、GB/T 45288.2—2025、GB/T 45288.3—2025 界定的术语和定义适用于本文件。

4　缩略语

下列缩略语适用于本文件。

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

CPU：中央处理器（Central Processing Unit）

DCMM：数据管理能力成熟度评估模型（Data Capability Maturity Model）

FLOPS：每秒浮点运算次数（Floating-point Operations Per Second）

GPTQ：生成式后训练 Transformer 量化（Generative Pre-trained Transformer Quantization）

GPU：图形处理器（Graphics Processing Unit）

IT：信息技术（Information Technology）

LoRA：低秩适应（Low-Rank Adaptation）

MaaS：模型即服务（Model as a Service）

资料整理
火龙果科技 荐

NPU：神经网络处理器（Neural Processing Unit）
QLoRA：量化低秩适应（Quantized Low-Rank Adaptation）
TOPS：每秒万亿次运算（Trillion Operations Per Second）
TPU：张量处理单元（Tensor Processing Unit）
ZeRO：零冗余优化器（Zero Redundancy Optimizer）

5 选型与应用的原则与流程

5.1 原则

5.1.1 合理性原则

人工智能应用过程中宜进行成本效益分析，考虑人工智能项目的投资成本，包括硬件、软件、人力等方面的投入，以及预期的收益回报，如提高生产效率、增加销售额、降低成本等。确保人工智能应用在经济上是可行的，能为使用方带来实际的经济效益。

5.1.2 适用性原则

深入分析业务流程、目标和需求，确保人工智能技术与自身业务紧密结合且适用，使人工智能技术能切实解决业务痛点、提升业务效率或创造新的业务价值，避免盲目跟风应用。

5.1.3 适合性原则

选择人工智能技术时，考虑技术的成熟度、复杂性与使用方的技术能力和资源相匹配，避免过度追求先进技术而导致技术难以落地或实施成本过高。同时，宜关注技术的兼容性，确保其能与使用方现有的信息技术基础设施和系统协同工作。

5.1.4 可解释性原则

对于复杂的机器学习和深度学习模型，宜采用合适的技术和方法，提高人工智能模型的可解释性，有助于发现模型中的潜在问题，如偏差、歧视等，同时也有利于模型的优化和改进。

5.1.5 完善性原则

关注人工智能提供方服务的完整性和持续性。重视模型的开发和部署，确保后续的维护、更新、优化等服务能满足要求。提供完善的用户培训和技术支持，保障用户有效使用人工智能系统，并及时解决使用过程中出现的问题。

5.1.6 安全性原则

保障人工智能系统在各个环节的安全性。数据方面防止数据泄露、篡改；算法方面，避免算法被恶意攻击利用；系统运行方面，确保系统的稳定性和可靠性，防止因人工智能系统故障导致业务中断或产生错误决策等安全问题。

5.2 流程

大模型选型和应用的基本流程见图1。

资料整理 火龙果科技 荐

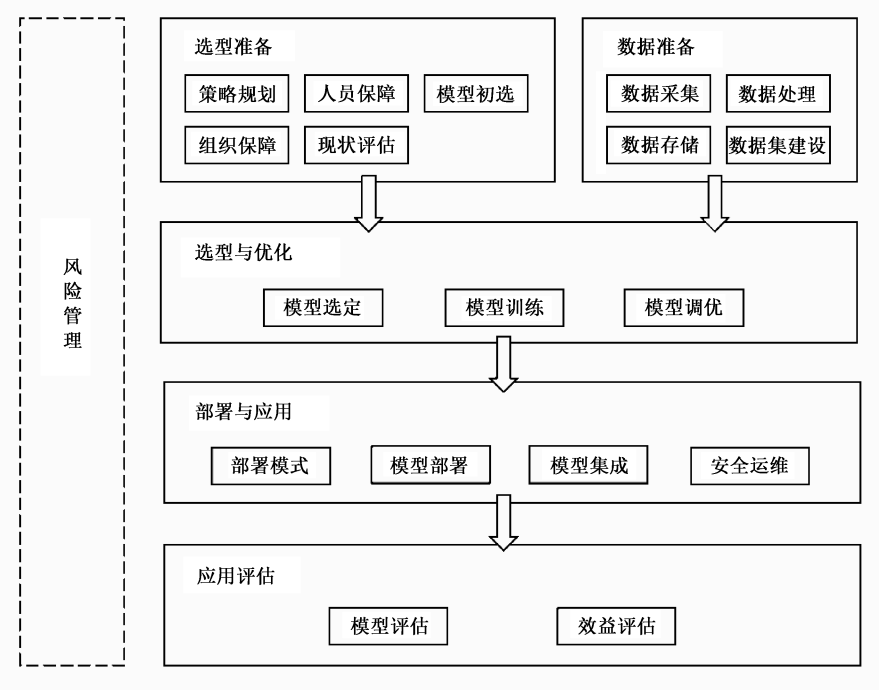

图 1　大模型选型和应用基本流程

大模型选型和应用包含如下流程。

a）　选型准备：使用方需加强策略规划、组织保障、人员保障、现状评估及模型初选环节，通过市场调研、同行交流、案例分析、试用体验等方式评估当前引入人工智能的基础条件，综合考虑成本投入与预期效益，确定引入人工智能的目标场景与中长期规划。

b）　数据准备：加强使用方自主数据采集、存储及处理能力，必要时可采购外部的行业专属数据集，或采用数据反转、模拟合成等技术手段，进行数据集建设，形成使用方专属数据集。

c）　选型与优化：根据具体应用场景，从性能、成本等多维度筛选合适的模型，并通过提示词工程、检索增强生成、微调或综合调优等技术针对性优化来提升模型在特定场景下的效果。

d）　部署应用：根据使用方技术能力、系统部署规模、业务安全需求、综合运行成本等确定部署方式，并为业务安全稳定运行提供充足的人、财、物保障。

e）　应用评估：定期组织开展应用的模型效果与效益综合评估，结合行业发展趋势，调整人工智能应用总体规划。

注：图 1 中各环节流程根据实际情况进行裁剪。

6　选型准备

6.1　策略规划

使用方宜将人工智能战略纳入整体业务规划中，明确人工智能应用的目标、重点领域和实施路径，定期评估人工智能技术对使用方现有业务流程、组织架构和商业模式的影响，制定相应的转型策略，以确保人工智能应用与使用方长期发展目标相一致。

6.2　组织保障

使用方宜建立适应人工智能应用的组织架构，设立专门人工智能项目团队，包括项目经理、数据科学家、算法工程师、业务分析师等角色，并明确各角色的职责和权限。宜建立由企业主要负责人或主管

资料整理
火龙果科技 荐

技术与业务的高级管理者牵头的联合工作机制，统筹技术、市场、业务、运营、采购、财务、人事等多部门间协作，畅通各部门成员之间沟通渠道，共同推动人工智能项目的顺利实施。宜符合 GB/T 45081—2024 的相关要求，确保人工智能技术的负责、可持续发展，以及在开发、部署和应用过程中的安全性和合规性。

6.3 人员保障

人工智能项目团队人员能力，包括如下内容。

a) 专业技术能力：具备相应的技术能力，如掌握机器学习、深度学习算法，熟悉数据处理和模型开发工具等；不同岗位的人员根据其职责要求具备不同层次的技术能力，以满足人工智能项目的开发和运维需求。

b) 业务理解能力：具备较强的业务理解能力，能深入理解使用方的业务流程、需求和目标，将人工智能技术与业务需求相结合，确保人工智能应用能为使用方带来实际价值。

c) 安全保障能力：具备一定的信息安全知识，了解信息系统安全防护基本知识与常见工具，规划人工智能应用安全建设方案，制定安全管理制度，开展安全运维，提高人工智能融合应用的安全性。

d) 伦理道德意识：遵循伦理道德规范，了解人工智能可能带来的伦理问题，如算法偏见、隐私侵犯等，并积极采取措施避免相关问题的发生，保障人工智能应用的合法性和合理性。

6.4 现状评估

6.4.1 基础能力

使用方宜系统性定位现状、识别短板、规划路径，从而确保智能化升级的科学性和有效性。使用方能从以下方面开展人工智能模型基础能力评估与诊断。

a) 综合参考 DCMM 评估、数字化转型成熟度模型与评估、人工智能使用方智能化成熟度模型、人工智能管理能力成熟度模型等标准，评估自身数字化、智能化水平，识别转型升级瓶颈。结合经济性分析与风险评估，科学确定人工智能应用需求。

b) 参考行业典型案例，统筹衔接企业数字化转型工作，合理设置应用目标，确定人工智能应用核心场景和技术导入优先级，确保人工智能应用精准支撑主营业务发展。

c) 构建统一技术底座和场景化应用套件相结合的硬件支撑体系，通过算力、存储、网络优化升级，建设与人工智能应用需求相适应的基础设施，并保留适度可扩展性。

d) 加快核心软件智能化改造升级，优化基础软件内核，植入智能调度算法，提升资源分配效率。部署集成数字孪生、大模型等数智技术类应用软件。

6.4.2 算力资源

使用方宜从以下方面开展算力资源估算，算力需求估算与规划建议参考附录 A 的 A.1。

a) 盘点、统计、整合现有 IT 基础设施，包括但不限于服务器、存储设备、网络设备等。按设备类型（CPU、GPU、TPU）、部署位置、使用部门等维度分类统计，推荐参考业界标准如 FLOPS 或 TOPS 作为衡量标准。区分人工智能专用计算设备和通用计算设备，统计并分析现有设备的能耗情况，为后续算力规划提供依据和参考。

b) 结合企业发展实际，制定阶段化、梯度上升的算力部署规模，优先选择能实现瞬时响应、可扩缩容的算力服务。

c) 采用云计算服务快速构建智能化基础服务能力。具备良好数字化基础的企业推荐设自有算力基础设施。

资料整理 火龙果科技 荐

d) 集成基于业务特征的云边端协同模式智能调度系统，整合多元异构芯片资源，满足不同场景需求。

6.4.3 成本性能估算

企业宜从以下方面开展成本估算：

a) 开展模型预期效果评估，包括内容有用性、合规性、准确性等；

b) 开展模型预期成本评估，对比各种模型、部署模式的建设成本与运行成本，包括训练成本，推理成本，部署、运维及升级成本等；

c) 开展模型预期性能评估，包括训练速度、推理速度、响应速度、生成速度等。

6.5 模型初选

宜选择成熟供应商的产品开展任务效果验证，通过基准测试了解模型上限表现，并根据测试结果开展调优及提示词优化等工作，直至达到预期的效果满意度；效果验证满意后，考虑选择降低规格参数的大模型及配套支持进行多轮降配验证，根据 GB/T 45288.3—2025 中 6.1.1 和 6.1.2 规定的关于软硬件平台及配套工具链能力，综合考虑性价比、生态社区等因素后完成模型初选。

必要时宜先拟定多技术路线，模型性能、模型分类及适用场景建议参考 A.2 和 A.4。

7 数据准备

7.1 数据采集

使用方可开展数据采集，数据类别可分为内部数据，如财务运营、系统日志等；外部数据，如行业统计、宏观经济等。使用方从以下方面开展数据采集。

a) 宜确定数据来源，包括内部业务数据、外部公开数据等。根据 GB/T 23011—2022 的要求，宜确保数据采集过程符合法律法规和使用方隐私政策，采集的数据具有代表性和多样性，为人工智能模型提供充足的数据基础。

b) 宜搭建使用方基础数据库，形成覆盖全业务场景的数据资源池。构建各类知识库，有效支撑使用方人工智能数据集需求。

c) 宜建立数据审批管控流程，保障数据符合国家相关法律法规、部门规章以及相关国家标准的要求。

d) 宜明确数据采集的多样性、代表性要求，建立数据偏差检测机制，如统计检验、公平性指标，并提出缓解策略，如重采样、对抗训练等。

e) 宜对采集到的数据进行标识，设立数据资产管理目录，对采集的数据进行分类分级管理，并建立数据质量保障体系。在数据采集、传输、存储中宜根据项目要求遵循隐私保护的要求，根据 GB/T 37964 对数据进行去标识化、加密等处理。

f) 设计并使用高质量数据集全生命周期的工具链体系，涵盖数据采集、清洗、增强、标注、合成、存储、传输、分析与应用等环节，根据 GB/T 36073—2025 的要求为使用方人工智能应用持续提供高质量、高效、安全的数据集支持。重点使用深度推理数据集建设工具、智能标注工具、专家协同标注工具、融合机理与仿真的数据合成工具、数据集质量评估和提升工具、数据集安全监测工具、数据集安全应用工具等。

7.2 数据存储

使用方宜从以下方面开展数据存储。

a) 选择合适的数据存储方式，如数据库、数据仓库等。建立数据存储管理机制，保障数据的安全

资料整理 火龙果科技 荐

存储、高效访问和备份恢复，防止数据丢失或损坏。

b) 按照 GB/T 35273—2020 中第 5 章和第 6 章的要求对数据进行采集、传输和存储，并采用有效的技术措施对采集数据中的个人信息进行保护。

7.3 数据处理

使用方从以下方面开展数据处理。

a) 数据清洗：宜对采集数据中的异常、冗余、缺失等数据质量问题进行数据清洗转换等操作，用于提升训练数据集的质量；清洗后的数据通过数据质量评估工具进行验证，确保其完整性、一致性、准确性等指标符合模型训练的需求。

b) 数据增强：人工智能模型开发方宜根据模型训练需求，基于有限的数据生成更多等价数据，丰富训练数据的分布，提高模型泛化能力。

c) 数据标注：人工智能模型开发方按照模型训练需求，对训练目标或对象进行标注，并对数据标注结果进行质检，高质量标注数据集有助于提升模型训练效果。

d) 特征工程：人工智能模型开发方根据模型训练需求，对经过预处理的数据集的特征进行选择、提取的过程。特征工程以提取高质量特征为目标，从而保障模型性能水平。避免可能存在的模型训练计算量大、训练时间过长、样本不均衡、过拟合或欠拟合的问题。

e) 数据融合：宜采用跨模态表示学习方法（如联合非负矩阵分解、变分自编码器或图神经网络），解决不同模态数据在维度、密度、标注难度上的异构性问题。需确保模态间预测一致性，减少融合过程中的信息损失，通过轻量级融合模型提升非完美对齐状态下的鲁棒性。

f) 数据质量评估：人工智能模型开发方宜对采集数据进行数据探索性分析以及数据质量评估，推荐采用可视化工具初步分析数据集的数量、属性、特征、分布以及确认采集的数据质量是否符合构建阶段的采集需求。

7.4 数据集建设

使用方宜从以下方面开展数据集建设。

a) 数据集划分：将数据划分为训练集、验证集和测试集。训练集用于模型训练，验证集用于模型调优，测试集用于最终评估模型的泛化能力。合理划分数据集能保证评估结果的客观性和可靠性。

b) 数据合成：综合利用扩散模型等数据合成技术，充实多模态场景数据；建立多维度数据验证机制，保障生成合成数据与真实数据的特征一致性、逻辑合理性。优先生成高危作业、数据稀缺、隐私敏感场景的数据。

c) 数据集采购：对于特定场景有定制化需求的情况，能通过购买、合作或交换等方式从正规数据供应商、公共数据库或第三方平台获取数据资源、配套工具链及专业服务，平衡数据的成本、质量与合规性。

8 选型与优化

8.1 模型选型

使用方宜从以下方面开展模型选型。

a) 利用初选选定的人工智能模型，确定模型的结构、参数等。在模型构建过程中，遵循最佳实践和行业标准，确保模型的合理性和有效性，同时考虑模型的可扩展性以适应未来业务变化。

b) 模型参数规模、算力需求的确定方法见 A.3。

资料整理 火龙果科技 荐

8.2 模型训练

模型在训练过程中，会分别涉及决策参数、权重占比、逻辑合理性等问题，在训练建模过程中，需要重点关注以下几点。

a) 避免偶发性参数的选择。在将客户的一些偶发性操作行为的识别，作为参数触发条件，推进到模型决策的下一个阶段，进而影响用户的选择等。宜对不同决策参数设置的合理性进行评估，以避免对用户造成不适当的引导或反感，对用户的安宁权造成影响。避免模型根据用户某个偶发性的操作行为（如误点击）作为决策树中重要的参考因子进行频繁推荐。

b) 权重合理。将特定参数或操作行为作为模型建模训练基础时，将某一参数设置过高或过低的权重或占比，以得出最终决策依据。相应的，在发现此类情况下，推荐对该权重或占比的合理适当性进行复核，以保证模型的科学合理性。

c) 逻辑合理。在传统评分逻辑下，较多依赖的是因果关系推导，通常情况下不会引发逻辑合理性的问题。但伴随着大数据技术、统计学的应用等，宜特别注意避免陷入到单纯统计学意义上的决策逻辑中。例如将根本无法进行关联的行为操作作为决策的基础。

d) 道德风险管控。在建模阶段，宜设立复核小组，原则上人数不少于 3 人，并建立相应的复核流程管控机制。以尽可能规避道德风险，以及个人喜恶偏好等原因造成侵害公众权益，乃至违反社会普适的价值观的可能。

e) 提升模型鲁棒性。对于公共服务、交通驾驶、金融服务、卫生健康、教育等重要领域中的安全攸关场景，在训练建模阶段中需使用对抗训练、数据增强等方法提升模型鲁棒性。

f) 提升模型的可解释性。以提升人工智能模型可解释为目标，在保证系统性能能够满足任务需求的前提下，优先考虑使用可解释性较强的模型。

g) 隐私保护。需要考虑的要素如下：

1) 在模型搭建和模型训练过程中需确保隐私数据的保密性，防止数据隐私的直接泄露；

2) 注意关键数据特征的保护，防止间接的隐私泄露，如梯度变化、输出向量等均存在暴露用户信息的风险；

3) 防止模型逆向攻击，通过模型输出的置信度逆向反推输入样本或数据的部分或全部特征；

4) 确保模型搭建和模型训练过程在可信环境下进行，例如通过硬件隔离或加密手段对涉及隐私数据的运算和操作进行保护。

8.3 模型调优

结合场景特征和业务目标，设定模型选型所用的可量化指标，为模型选型和调优提供依据。基于业务场景需求，开展模型比测选型。合理选取包括模型微调、提示词工程、上下文工程、检索增强生成等人工智能技术。

a) 提示词工程：宜使用提示词工程来引导模型行为给出回复，用于提高回复内容的逻辑性、稳定性。构建涵盖常规问题、边缘案例的提示词库，建立多维度评价指标。实施数据源权威性评价与内容监测机制。

b) 上下文工程：应对可能影响大模型或智能体行为的上下文信息进行系统识别，并按照信息来源和性质进行分类。宜避免引入与当前任务无关或冗余的上下文信息，以降低模型理解与推理复杂度。包括但不限于：

1) 任务指令类信息；

2) 业务规则与约束类信息；

3) 用户和角色属性类信息；

4) 环境状态与历史交互信息。

资料整理 火龙果科技 荐

c) 检索增强生成：宜利用检索式增强结合外部知识数据，实现更可控地回答垂直、封闭领域问题，提高大模型的准确性。

d) 模型微调：宜使用参数高效微调，如 LoRA、QLoRA 等轻量化微调来优化模型在特定任务上的表现，实现适配特定任务，提高大模型的准确度。

e) 综合调优：实施差异化模型优化策略，综合采用参数微调、架构搜索、大小模型协同等手段。

9 部署应用

9.1 部署模式选择

使用方宜根据业务应用规模、信息化管理水平、基础设施建设情况等选择适合的部署模式，模型性能计算方法见 A.4：

a) 中小微规模使用方优先考虑直接使用 MaaS 模型服务，直接调用模型服务 API，快速构建智能化基础服务能力，降低前期技术投入成本；

b) 具备良好数字化基础且对数据安全要求较高的使用方，推荐依托自有算力基础设施及 IT 资源，私有化部署人工智能应用，实现资源集约化利用；

c) 鼓励龙头企业、大型企业等大规模使用方结合实际需求牵头建设智算中心，集成 GPU、TPU、NPU 等专用芯片，优化矩阵运算与并行计算能力，支持万卡级集群调度。

9.2 模型部署

不同的模型在不同的领域所适用的业务场景、客群范围等均有差异，一旦部署出现偏差可能会引发潜在风险。使用方对模型部署的过程进行记录和管理，并在部署完成后安排部署人员以外的质检团队进行验收。为了管控模型应用的风险，在应用策略上，宜遵循稳妥审慎的原则，逐步开放适用的范围，而不宜直接应用到所有客户群体，以便于进一步根据客户使用情况对模型进行调优及改善。模型部署阶段考虑以下因素。

a) 模型部署在可信环境，例如 SGX、TrustZone 等。

b) 保障算法库、深度学习推理框架的安全性和可靠性。

c) 当计算、存储资源有限时（如部署在智能终端），根据 GB/T 45288.3—2025 中的 6.2.2.6 考虑使用模型剪枝、知识蒸馏等技术压缩模型，并尽量减小对模型精度的影响，压缩方案见 A.5。

d) 业务规模大的使用方，考虑基于模型服务平台来构建部署。对模型有灵活控制权，并推荐根据特定的业务需求进行深度优化。

e) 对数据安全和隐私有特殊要求的需求场景，宜选择本地机房部署，或者选择提供高级安全和合规特性的云服务。

f) 模型发布管理：模型发布前宜履行分级审批程序，明确发布范围与回滚预案；宜采用灰度发布策略，按客群、区域或业务模块逐步扩大应用范围，持续监测关键指标异常并建立自动熔断与版本回退机制，确保风险可控。

9.3 模型集成

宜开发具体的模型应用接口和用户界面，提供可视化展示、数据分析结果和决策支持。

9.4 安全运维

使用方宜从以下方面开展安全运维。

a) 确保模型在各种输入条件下，模型推理的基本功能宜通过主、客观测试，模型版本管理功能正常运行。

资料整理 火龙果科技 荐

b) 对模型推理时涉及的传感器、仪表进行实时监控；根据 GB/T 45288.3—2025 中的 6.3.1，模型推理的服务效率、稳定性等性能指标，宜保证基础设施的计算资源的使用率、服务可用性、容灾恢复时间等关键性指标在安全范围内稳定运行。

c) 对模型部署集成后的用户体验反馈、人机协助便捷度、模型异常处置的日志进行记录、分析。

10 应用评估

10.1 模型评估

使用方宜从以下方面开展模型评估：

a) 根据 GB/T 45288.1—2025 和 GB/T 45288.2—2025 中大模型的通用能力的评测指标开展大模型技术能力与应用效果评估；

b) 根据 GB/T 45288.3—2025 大模型服务能力成熟度，结合行业大模型的业务指标及真实用户反馈，开展行业大模型应用建设度及成熟度的成效评估。

10.2 效益评估

使用方宜从以下方面开展效益评估：

a) 宜对生产力价值进行评估，衡量人工智能所带来的生产力提升，例如平均通话处理时间、文档处理时间和使用工具所节省的时间；

b) 宜对成本节约进行评估，测量传统成本与人工智能解决方案之间的成本对比；

c) 宜对创新与增长指标进行评估，通过衡量文档处理能力、知识扩展性以及工作沟通质量的改善，评估人工智能在推动新产品、服务或商业模式方面的促进创新作用；

d) 宜对客户体验进行评估，通过衡量流失减少、收入增长、访问量、网站停留时间等指标，评估人工智能对客户满意度和忠诚度的影响；

e) 组织能力建设：评估人工智能应用对员工技能结构、协作模式、创新文化的促进作用，如人工智能素养培训覆盖率、跨职能团队组建数量等。

11 风险管理

11.1 组织能力建设

组织能力包括：

a) 宜具备风险管理的组织能力，宜设立多层管理机制；

b) 宜具备明确的各角色责任划分；

c) 宜具备高效的沟通与协调机制。

11.2 风险管理目标

宜结合业务场景及政策法规等内外部环境，设定明确的风险管理目标。

11.3 风险识别及评估

风险识别及评估包括如下内容。

a) 构建内置模型安全防御体系，建立模型协同验证机制。定期开展模型功能、性能、安全测试评估。

b) 建立覆盖系统开发部署全周期的可信环境，实施软硬件安全审查与风险评估。

资料整理
火龙果科技 荐

c) 建立覆盖数据全生命周期的防护体系，通过加密技术、访问控制与动态审计保障数据安全。
d) 加强权限管理，依据用户角色和业务需求，精准分配人工智能应用的操作权限。
e) 具备识别风险影响对象、风险来源等的能力。
f) 具备评估潜在事件和影响、后果程度等的能力。

11.4 风险应对

明确风险事件处置流程和岗位职责，制定泄露、出现重大偏差等情况下的风险应急预案，具体包括：

a) 明确应急响应的触发条件（如模型性能显著下降、数据泄露、对抗攻击等），制定详细的响应步骤（如模型下线、数据隔离、漏洞修复等），并明确各岗位的职责和协作流程；
b) 定期组织模型相关人员进行风险应急预案相关培训和应急演练，至少每年一次；
c) 记录、并存档演练记录；
d) 提前建立模型下线或更替机制，保证在发生攻击或泄露等情况下，能采取有效措施进行应对；
e) 发生风险事件后立即采取有效措施防止风险扩大，并视风险程度向相关监管部门报告；
f) 记录事件内容，分析和鉴定事件产生的原因，评估事件可能造成的影响，总结经验教训，制定防止再次发生的补救措施；
g) 根据相关法律法规变化情况，以及事件处置情况，及时评估更新应急预案；
h) 使用方审查验证、测试和确认结果，以根据风险处理计划了解导致风险的事件和条件，审查和改进风险管理的流程和控制计划。

资料整理
火龙果科技 荐

附 录 A
（资料性）
大模型选型参考信息

A.1 算力估算

大模型算力需求的核心评估指标包括：模型参数量（N）、训练数据量（D）和总浮点运算量（FLOPs）等。模型参数量（N）能通过 Transformer 架构的层数（L）、隐藏层维度（h）等计算得出。如，GPT-3（1 750 亿参数）的 $L=96$ 层，$h=12\ 288$，$N\approx12Lh$。训练数据量需满足 $D\geqslant15N$（如万亿参数模型需 $D\geqslant1.5\times10^{13}$ tokens），总所需量化位数$\approx6\times N\times D$，指导硬件的估算，见公式（A.1）。

$$M_{VM}=\frac{b_{PS}\times q_{QBW}}{8}\times1.2 \qquad \cdots\cdots(A.1)$$

式中：

M_{VM} ——显存，单位为千兆字节（GB）；

b_{PS} ——参数规模；

q_{QBW} ——量化位数。

A.2 模型分类

大模型应用场景分类见表 A.1。

表 A.1 大模型应用场景分类

技术	业务场景
通用对话大模型	智能客服、聊天助手
多模态大模型	图文生成、图像理解
推荐大模型	电商、内容平台、音乐流媒体
代码大模型	代码生成、代码解释
医疗大模型	病历分析、医学问答
法律大模型	合同审查、法条检索
金融大模型	风控、舆情分析

A.3 模型参数

大模型微调 LoRA 微调不同参数规模对应训练配置见表 A.2。

表 A.2 大模型参数规模对应训练配置

模型参数量	最低显存需求	最低算力需求
1.5B	3G	3.5×10^{13} FLOPs
7B	13G	1.7×10^{14} FLOPs

资料整理 火龙果科技 荐

表 A.2　大模型参数规模对应训练配置（续）

模型参数量	最低显存需求	最低算力需求
8B	15G	2×10^{14} FLOPs
14B	26G	3.5×10^{14} FLOPs
32B	48G	8×10^{14} FLOPs
70B	130G	1.7×10^{15} FLOPs
671B	1 250G	1.7×10^{16} FLOPs
……	……	……

A.4　模型性能

A.4.1　基础并发量

基础并发量的计算见公式(A.2)。

$$B_{bc}=\frac{H}{I} \qquad \cdots\cdots(A.2)$$

式中：

B_{bc}——基础并发量；

H ——硬件每秒推理次数；

I ——单次推理平均耗时，单位为秒(s)。

若已知系统吞吐量和延迟要求，则基础并发量计算见公式(A.3)。

$$B_{bc}=\mathrm{TPS}\times(\mathrm{TTFT}+\mathrm{TPOT}\times\mathrm{NGT}) \qquad \cdots\cdots(A.3)$$

式中：

TPS(Transactions Per Second)　——每秒事务处理量；

TTFT(Time to First Token)　——首 Token 生成时间(建议<1.4 s)，单位为秒(s)；

TPOT(Time Per Output Token)　——每 Token 生成耗时，单位为秒(s)；

NGT(Number of Generated Toksns)——生成 Token 数。

A.4.2　基于服务器压力的平均和峰值并发数

基于服务器压力的平均和峰值并发数的计算见公式(A.4)。

$$C=\frac{n\times L}{T} \qquad \cdots\cdots(A.4)$$

式中：

C ——平均并发数；

n ——活跃用户数；

L ——会话平均时长；

T ——考察时间段。

A.5　模型压缩

使用方宜采用量化技术减少显存消耗，见表 A.3。

资料整理
火龙果科技 荐

表 A.3　优化技术降低硬件门槛

技术方案	显存压缩率	性能损失	适用场景
8-bit 量化	50%	＜2%	企业级推理(如客服系统)
4-bit GPTQ	75%	＜5%	边缘设备(如手机端大模型)
ZeRO Offload	60%～80%	延迟增加 20%	低成本微调(单卡训练 70B)

资料整理
火龙果科技 荐

参 考 文 献

[1] GB/T 41867—2022 信息技术 人工智能 术语
[2] GB/T 42129—2022 数据管理能力成熟度评估方法
[3] GB/T 43439—2023 信息技术服务 数字化转型 成熟度模型与评估

资料整理
火龙果科技 荐