

# 《面向大模型训练与推理的算力费用核算模型》

## 编制说明

2026 年 09 月

## 一、工作简况

### （一）任务来源

本项目根据中国互联网协会 2026 年团体标准制定计划（标准计划号：240-T/ISC-26），项目名称为“《面向大模型训练与推理的算力费用核算模型》”的任务而进行制订。

### （二）起草单位及主要起草人

本文件起草单位：国家超级计算天津中心、中国科学院计算技术研究所、中国科学院计算机网络信息中心、北京无熵科技文化有限公司、山西云时代技术有限公司先进计算分公司（国家超级计算太原中心）、江苏地金网络科技有限公司、海南智算创新发展研究院、北京清程极智科技有限公司、文昌航天超算智慧管理有限公司、深圳市捷易科技有限公司、华兴中科标准技术（北京）有限公司。

本文件主要起草人：孟祥飞、毛泽鑫、张云泉、姜金荣、江楠、吴青、苏开明、郭卓、徐文洁、穆建芝、翟季冬、唐适之、叶世阳、李利红、李华、任国静、丁月。

### （三）标准制定目的和意义

随着大模型技术快速迭代与规模化落地，人工智能大模型已成为数字经济与智能科技发展的核心引擎，人工智能训练、推理类算力需求持续爆发，大规模训练与推理是大模型研发的关键环节，具有算力需求大、训练周期长、集群规模高、任务连续性要求严苛等特点。行业当前尚无面向大模型场景的专用算力费用核算模型，不同服务商、算力节点、应用主体的计费方法差异较大，引发计价分歧、核算混乱及定价不透明等问题，不利于 AI 算力资源市场化交易、精细化管控与规模化复用，亟需完善标准建设填补标准空白。

制定《面向大模型训练与推理的算力费用核算模型》，旨在针对大模型训练与推理的差异化算力特征，建立科学、统一、可复用的算力费用核算模型与计费规则，明确模型训练、在线推理、增量微调等典型场景的算力消耗统计维度、核算指标、计量方法、成本分摊及计费折算要求，统一行业核算口径与定价依据。

本标准的制定能够填补大模型专项算力费用核算标准空白，破解通用算力计费模式与 AI 模型业务场景不匹配的突出问题，有效规范大模型算力服务交易行为，减少市场计价争议，提升算力计费的科学性、公正性和透明度。同时，可支撑大模型算力资源

精细化管理、成本精准测算、服务定价标准化，降低产学研用各方算力采购与合作对接成本，促进智算资源高效流转与集约利用，健全人工智能算力市场化服务体系，为大模型产业规模化、规范化、高质量发展提供重要的标准支撑。

#### **（四）主要工作过程**

2026 年 08 月本团体标准由中国互联网协会正式立项，立项名称为：《面向大模型训练与推理的算力费用核算模型》。

2026 年 08 月 20 日在中国互联网协会 11 楼 1115 报告厅召开了本团体标准的第一次讨论会。本次会议确定了本标准的框架结构，收集了与会专家及各单位代表对标准草案（初稿）的修改意见或建议，明确了工作组各单位的工作分工，以及会后的具体工作安排。

2026 年 09 月 17 日通过腾讯会议线上召开了本团体标准的第二次讨论会。编制组就本团体标准第一次讨论后的修改情况、意见采纳情况进行了综合阐述。与会专家及代表对该项团体标准的定义内容、技术内容等进行了进一步的深入讨论和修改，编制组认真听取和记录了与会代表提出的问题和调整建议。

## **二、标准编制原则和依据**

### **（一）编制原则**

标准起草小组在编制标准过程中，以国家、行业现有的标准为制订基础，结合我国产业现状，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》的规定及相关要求编制。

### **（二）标准范围**

本文件规定了大模型训练与推理场景下算力费用核算模型的总体架构、核算资源要素、核算单元与数据计量、核算数据采集与处理、算力费用核算模型、训练算力费用核算模型、推理算力费用核算模型、核算规则、核算结果调整及综合定价应用、计费方式与核算结果应用、服务等级协议与费用核算关联、结算与支付以及试验方法等内容。

本文件适用于智算中心、云计算平台、算力调度平台等为大规模深度学习模型训练与推理提供的 GPU、NPU、TPU 等异构算力资源（含裸金属、容器、虚拟化实例）的交易定价与计费，也可作为算力交易双方签订合同的参考。

### （三）标准主要内容与确定依据

#### 1.总体架构

确立费用核算总体要求、八项基本原则、训练与推理核算对象、核算边界及十五个核算维度。

**确定依据：**真实性、一致性、可追溯性等原则与通用财务核算原则一致；“核算与定价适度分离”为起草单位基于多家算力交易平台运营实践提出。核算对象与边界划分由起草单位联合多家智算中心，依据训练与推理任务实际资源流转过程总结形成。

#### 2.核算资源要素、核算单元与数据计量

划分七类核算资源要素，规定计量要求、核算单元选择原则、训练与推理核算单元，以及 SCU 标准算力单位折算条件。

**确定依据：**资源要素分类参照 GB/T 36326 框架；SCU 标准算力单位及性能系数标准由起草单位联合多家智算中心，参照国际算力交易市场 SCU 实践，采用 Transformer、MoE 等典型模型进行跨平台交叉标定试验后形成。

#### 3.核算数据采集与处理

规定数据采集范围、秒级采样精度、统一时间源、清洗校验、聚合归集、异常处理及数据可信性要求。

**确定依据：**数据可信性中防未授权修改、原始记录对应关系、保存期限等要求与 GB/T 35274 保持一致；秒级采样精度高于现有云服务运营标准通常的分钟级要求，系起草单位基于动态批处理与自动扩缩容场景下的连续采集试验，经多家单位重复验证后确定。

#### 4.算力费用核算模型

建立综合费用计算框架，分解为直接费用、间接费用分摊、其他费用及 SLA 调整金额，规定间接费用分摊权重。

**确定依据：**费用分解框架参照通用成本核算原则；间接费用分摊权重（计算资源占用时间、标准算力使用量、Token 工作量、组合权重等）为起草单位基于多家智算中心运营数据，分析各指标与资源消耗相关性后形成。

#### 5.训练算力费用核算模型

规定按资源实际占用量和按有效算力两种核算方式，涵盖并行训练、Checkpoint、资源利用率、失败重试、断点续训等规则。

**确定依据：**按资源占用量核算为行业通用做法；有效算力核算及避免重复调整机制高于现有标准要求，由起草单位联合多家单位，通过对比不同硬件平台相同训练任务的实际吞吐率与费用核算结果，经交叉验证后形成。失败重试责任划分基于多家单位故障处理数据与费用争议案例总结。

## 6.推理算力费用核算模型

规定基于资源占用、请求与 Token 工作量、请求数量及混合核算方式，涵盖共享资源分摊、长上下文、多轮交互、流式推理、动态批处理、空闲资源等规则。

**确定依据：**混合核算模式（资源容量费用+业务工作量费用）无国标出处，由起草单位通过构造不同并发与 Token 比例的推理测试负载，标定容量成本与工作量成本比例后形成。Token 区分输入输出、长上下文 KV Cache 计量、多轮逐轮计量等均为多家单位自主试验所得，验证了仅按请求次数或总量核算的失真问题。

## 7.核算规则

系统规定按实际使用量、预留量、运行时长、任务工作量、Token、请求量、实例、综合资源及混合核算等十余种规则，以及最低资源量、超额使用、利用率调整、共享分摊、失败重试、异常中断、SLA 未达标等特殊处理。

**确定依据：**各项规则为起草单位基于多家智算中心与云服务商运营实践总结，部分参考通用会计原则。其中资源利用率调整、责任归属划分等经多家单位联合讨论与模拟验证后形成。

## 8.核算结果调整及综合定价应用

明确基础核算与交易性因素分离，规定算力性能、时间、服务质量、合约周期、安全合规等调整因素及综合交易价格形成框架。

**确定依据：**算力性能调整中性能系数标定参照国际 SCU 实践并自主试验；核算与定价分离原则为起草单位运营经验总结；综合交易价格公式为自主构建。

## 9.计费方式与核算结果应用

规定按量、包周期、任务、Token、请求、有效算力、抢占式、混合及辅助计费方式。

**确定依据：**计费方式基于行业通用云计费模式，针对大模型场景细化，由起草单位结合多家平台实际计费规则总结形成。

## 10.服务等级协议与费用核算关联

建立“基础核算→SLA 评价→核算调整→最终结果”逻辑，给出训练与推理 SLA 指标、测量方法、补偿方式及服务质量系数。

**确定依据：**SLA 框架与 GB/T 31167、GB/T 36326 保持一致；TTFT、Token 生成速率等具体指标为自主试验确定；服务质量系数公式为起草单位提出，经多家单位联合验证。属高于国标的细化要求。

## 11.结算与支付

规定账单生成、内容、结算周期、支付方式、欠费处理及争议处理。

**确定依据：**账单内容与财务结算通用要求一致，支付方式基于行业实践；欠费处理中数据与 Checkpoint 保护要求为起草单位结合数据管理要求提出。

## 12.试验方法

规定计量准确性、核算结果验证、训练与推理核算、异构算力折算、SLA 及数据异常等试验的覆盖场景与验证内容。

**确定依据：**试验方法框架参照 GB/T 45087 和 GB/T 47514；具体试验场景与验证内容由起草单位联合多家单位设计，覆盖训练与推理典型负载，用于验证核算模型的一致性、可重复性与可追溯性。

## 13.附录 A、B（资料性）

给出主流 GPU/NPU 性能系数参考表及训练计算精度折算系数表。

**确定依据：**由起草单位联合多家智算中心，采用统一软件环境、统一模型（Llama、Qwen、DeepSeek 等）和统一测试方法进行跨平台标定试验，以多家单位测试结果加权平均后形成参考区间。

## 三、国内外情况简要说明

经查，目前国内外尚无专门针对大模型训练与推理的算力费用核算标准，现有云计算相关国标侧重云服务通用运营、安全能力等要求，缺少大模型异构算力核算、训练推理差异化计费等专项规定。本文件编制引用 GB/T 35274、GB/T 36074、GB/T 36326、GB/T 37739 等现行标准，保证标准框架合规。本文件结合国内智算、算力交易实际业务，划分训练、推理场景，明确核算资源要素、核算单元、费用归集分摊规则，建立异构算力折算、SLA 核算联动、试验验证、账单结算体系，回应算力计量口径不一、共享资源分摊缺乏依据等现实问题。综上，本标准达到国内先进水平。

#### 四、与有关的现行法律、法规和强制性国家标准的关系

本标准的制定过程、技术要求的选定、试验方法的确定、检验项目设置等符合现行法律法规和强制性国家标准的规定。

#### 五、重大分歧意见的处理经过和依据

本文件在制定过程中未出现重大分歧意见。

#### 六、废止现行有关标准的建议

本标准不涉及对现行标准的废止。

#### 七、其他应予说明的事项

无。

团体标准编制组

2026 年 09 月