

团 体 标 准

JH/T/DZJN XX-YYYY

预训练语言模型训练阶段能耗测评方法

Energy consumption evaluation methods for the training stage of
pre-trained language models

(征求意见稿)

2025-XXXX-XX 发布

2025-XX- XX 实施

中国电子节能技术协会 发布



版权保护文件

版权所有归属于该标准的发布机构，除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

目 次

前言	3
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 测量条件	2
5 测评方法	3
附录 A（资料性） 预训练语言模型训练阶段能耗测评示例	5
参考文献	7

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这些专利的责任。

本文件由中国电子节能技术协会科技创新与安全工作委员会提出。

本文件由中国电子节能技术协会归口。

本文件起草单位：

本文件主要起草人：

预训练语言模型训练阶段能耗测评方法

1 范围

本文件规定了预训练语言模型训练阶段能耗测评的测量条件、测量要求和测评方法。

本文件适用于通用预训练语言模型产品的选型及模型的优化,指导第三方测评机构开展预训练语言模型训练阶段能耗测评工作。

本文件不涉及模型性能及效果测评。

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 2589-2020 综合能耗计算通则

GB/T 29239-2024 移动通信设备节能参数和测试方法 基站

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 42018-2022 信息技术 人工智能 平台计算资源规范

GB/T 45288.1-2025 人工智能 大模型 第1部分:通用要求

GB/T 45652-2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

ISO/IEC 2382:2015 information technology —— Vocabulary

3 术语和定义

下列术语和定义适用于本文件。

3.1

能耗 energy consumption

智能终端或应用在特定时间段内消耗的总电能,单位为毫焦(mJ)。

[来源: GB/T 2589-2020, 3.5]

3.2

模型训练 model training

利用训练数据,基于机器学习算法,确定或改进机器学习模型参数的过程。

[来源: GB/T 41867-2022, 3.2.18]

3.3

迭代 iteration

针对一批样本,重复地执行系列步骤直至完成训练的过程。

[来源: GB/T 41867-2022, 3.2.5]

3.4

人工智能加速卡 artificial intelligence accelerate card

专为人工智能计算设计、符合人工智能服务器硬件接口的扩展加速设备。

[来源：GB/T 42018-2022, 3.6]

3.5

功耗 power consumption

智能终端或应用在规定时间内消耗的总电能，单位为毫瓦（mW）。

[来源：GB/T 29239-2024, 3.1.1]

3.6

大模型 large-scale 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]

3.7

预训练数据 pre-training data

用于生成式人工智能预训练的数据。

[来源：GB/T 45652-2025, 3.6]

3.8

批（量）训练 batch training

一种训练方法，仅在一个（训）期后，才调整连接权重。

[来源：ISO/IEC 2382:2015, 2120690]

4 测量条件

4.1 物理环境

测量时物理环境应符合以下要求：

——环境温度：15～35摄氏度（℃）；

——相对湿度：25%～75%；

——气压：86～106千帕（kPa）。

4.2 测量设备

运行被测模型的测量平台应选择独立的服务器（不采用分布式算力）。服务器应符合 GB/T 45288.1-2025 的要求，同时应满足被测模型在训练阶段所需的配置。服务器硬件的温度应处于工作温度规格内，且空载温度应与模型训练时温度相近。

应采用高精度功率分析仪采集电源端的功率数据。功率分析仪应符合如下要求：

——电源要求：单项交流220伏特（V），电源消耗功率为60瓦特（W）；

——输入信号范围：直流信号（DC）～1兆赫兹（MHz）交流信号；

——采样率：≥200兆样本每秒（MS/s）；

——测量频率范围：直流（DC）～100兆赫兹（MHz）；

- 功率测量误差: $\leq \pm 0.15\%$;
- 数据更新周期: ≤ 100 毫秒 (ms) ;
- 测量电流范围: 0 ~ 20安培 (A) 。

针对测量过程功率采样, 在所规定最小时间段内以不大于 1 秒为均匀时间间隔读取功率。

4.3 软件及数据要求

1. 软件应符合模型训练的要求, 包括操作系统、驱动程序、深度学习框架以及相关依赖库等。
2. 应采用适合模型训练的统一、规范的数据集。

4.4 其他要求

1. 测评方应具有相关资质, 并出具测评报告。
2. 测评前由送测方向测评方提出书面申请, 并按要求提交被测模型及相关材料。
3. 被测模型训练脚本原则上由测评方主导、送测方协助完成编制。

5 测评方法

5.1 概述

测评采用绝对能耗测量和相对能耗测评两种方法。绝对能耗测量方法直接测量被测模型的能耗, 相对能耗测评方法采用标准能耗比(被测模型能耗与基准值的比值)评估被测模型的能耗水平。预训练语言模型训练阶段能耗测评方法示例见附录 A。

5.2 绝对能耗测量方法

绝对能耗测量步骤如下:

- (1) 记录被测模型基本信息, 包括模型名称、结构以及参数量等。
- (2) 安装、部署模型训练所需的软硬件以及环境条件, 并详细记录, 如: CPU、人工智能加速卡以及驱动软件、内存和硬盘的型号及数量、操作系统以及模型训练阶段的依赖库等。
注: 不采用虚拟机或容器形式进行软硬件部署。
- (3) 按照模型要求准备经过预处理的统一、规范的训练数据集。
- (4) 编写被测模型的训练脚本。训练脚本应在指定位置保存每次训练过程的详细日志, 包含当前时间和迭代次数等。训练脚本应确保模型权重处于随机初始化, 不应加载预训练权重。
- (5) 安装、校准功率分析仪。
- (6) 启动硬件设备直至状态稳定。
- (7) 执行第4步中编写的训练脚本至模型收敛。
- (8) 待模型训练完成后, 从功率分析仪计算训练期间的总能耗。
- (9) 基于训练时长, 使用功率分析仪测量服务器空载时的能耗。
- (10) 将训练期间的总能耗减去服务器的空载能耗, 得到被测模型的单次训练能耗。

重复步骤1到步骤10, 对被测模型进行多次测量(至少3次)。并计算平均能耗作为绝对能耗最终结果。

5.3 相对能耗测评方法

采用标准能耗比表示相对能耗，即被测模型绝对能耗平均值与基准值的比值，计算方法见公式 (1)。

$$S = \frac{E_{obj}}{E_{std}} \quad (1)$$

式中：

S ——被测模型在训练过程中的相对能耗；

E_{obj} ——被测模型在训练过程中的绝对能耗平均值；

E_{std} ——基准值。

评测方根据具体应用场景选择基准值，并根据S值确定被测模型训练阶段的能耗级别。

示例：

S值	功耗级别
>2	高
≥ 1 且 ≤ 2	中
<1	低

为了进行不同模型能耗的比较，测评方应该使用满足各个模型训练要求的统一平台进行能耗测评，在此基础上确定各个模型的能耗级别。

附录 A

(资料性)

预训练语言模型训练阶段能耗测评示例

A.1 测量基本信息

编号：NH-P-NLP-T-20251231-00001。

送测产品：Qwen2.7-7B-base。

送测单位：xxx 公司。

送测人员：张三。

测评开始日期：2026 年 1 月 1 日。

测评完成日期：2026 年 1 月 5 日。

测评机构：xxx。

测评人员：李四。

测评地点：xx 大厦 101 室。

A.2 测试目的

1.测量模型 Qwen2.5-7B-base 训练过程中的能耗。

2.确定该模型的能耗级别。

A.3 测评依据

JH/T/DZJN XX-YYYY 预训练语言模型训练阶段能耗测评方法

A.4 测量环境

A.4.1 硬件条件

CPU 型号及数量：Intel Xeon Gold 6348 28 核 2.6G Hz；1 颗。

人工智能加速卡型号及数量：Nvidia A100 NVLink 版 80G；4 张。

内存型号及数量：DDR4 3200MHz LRDIMM 16G；12 条。

硬盘型号及数量：SATA SSD 2.5 英寸 3.84T；1 块。

电源型号及数量：1600W 80 Plus 铂金牌 PSU（支持冗余模式）；4 个。

主板型号及数量：Intel C621A series chipset（LBG-R）；1 块。

A.4.2 软件条件

操作系统：Ubuntu 22.04。

数据存储方式：以 csv、txt 文件格式存储在 /data 目录下。

人工智能加速卡驱动：CUDA 11.8。

深度学习框架：Pytorch 1.11。

被测模型结构：包含 RoPE、SwiGLU、RMSNorm 的 Transformer 架构，注意力头数 28（Q 为 28、KV 为 4），最大上下文长度 131072 个 token，精度为 BF16。

被测模型参数量：7B。
训练数据：CC-Stories-R 31GB。
训练批数：12。

A.4.3 环境条件

温度：24℃。
湿度：50%。
气压：101KPa。

A.5 测量工具

功率分析仪：Power Monitor。

A.6 测评步骤

Qwen2.5-7B-base 模型按 JH/T/DZJN XX-XXXX 5.2 和 5.3 的方法进行测评。为减小误差，模型在同一设备上进行 3 次训练，计算平均能耗、平均运行时长以及相对能耗。测试过程中无设备故障和训练终止等异常情况。

A.7 绝对能耗测量结果

表 A.1 绝对能耗测量结果

测量轮数	训练时长（小时）	训练能耗（KW·h）
1	30.1	2.95×10^5
2	30	3×10^5
3	29.9	3.05×10^5

平均时长为 30 小时，平均绝对能耗为 3×10^5 KW·h。

A.8 相对能耗测评结果

基准值： 2.5×10^5 KW·h。
S 值：1.2。
能耗级别：中。

A.9 测评有效期

本次测评结果有效期为 2 年，自测评完成之日起。

参考文献

- [1] GB/T 2589-2020 综合能耗计算通则
 - [2] GB/T 41867-2022 信息技术 人工智能 术语
 - [3] GB/T 42018-2022 信息技术 人工智能 平台计算资源规范
 - [4] GB/T 29239-2024 移动通信设备节能参数和测试方法 基站
 - [5] GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求
 - [6] GB/T 45652-2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
 - [7] ISO/IEC 2382:2015 information technology —— Vocabulary
-