ICS 13.020.10
CCS Z 04

团　　体　　标　　准

JH/T/DZJN XX-YYYY

预训练语言模型推理阶段能耗测评方法

Energy consumption evaluation methods for the inference 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

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：

预训练语言模型推理阶段能耗测评方法

1 范围

本文件规定了预训练语言模型推理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推理阶段能耗测评工作。

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

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

推理 inference

从给定的前提进行论证并得出结论。

注 1：在人工智能领域中，一个前提是一个事实、一个规则、一个模型、一个特征或原始数据。

注 2：术语“推理”既指过程也指结果。

[来源：GB/T 41867-2022, 3.2.30]

3.3

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]

3.4

功耗 power consumption

智能终端或应用在单位时间内消耗的总电能，单位为毫瓦（mW）。

[来源：GB/T 29239-2024, 3.1.1]

3.5

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]

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推理阶段所需的配置。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推理时温度相近。

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：≤ ± 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. 软件应符合模型推理的要求，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. 应采用适合模型推理的统一、规范的数据集。
3. 除被测模型外，不应使用其他的推理引擎。

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. 被测模型推理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量两种方法。绝对能耗测量方法直接测量被测模型的能耗，相对能耗测量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。预训练语言模型推理阶段能耗测评方法示例见附录 A。

5.2 绝对能耗测量方法

绝对能耗测评步骤如下：

(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、部署模型推理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、操作系统以及模型推理阶段的依赖库等。
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、规范的推理数据集。
(4) 编写被测模型的推理脚本。推理脚本应在制定位置保存每次推理过程的详细日志，包含当前时间、批量数和迭代次数等。
(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步中编写的推理脚本。
(8) 待模型推理完成后，从功率分析仪计算推理期间的总能耗。
(9) 基于推理时长，使用功率分析仪测量服务器空载时的能耗。
(10) 将推理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推理能耗。

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)。

$$S = \frac{E_{obj}}{E_{std}} \tag{1}$$

式中：

S ——被测模型在推理过程中的相对能耗；

E_{obj}——被测模型在推理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合选择基准值，并根据S值确定被测模型推理阶段的能耗级别。

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推理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）

预训练语言模型推理阶段能耗测评示例

A.1 测量基本信息

编号：NH-P-NLP-I-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 1 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 推理过程中的能耗。

2.确定该模型的能耗级别。

A.3 测评依据

JH/T/DZJN XX-YYYY 预训练语言模型推理阶段能耗测评方法。

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；1 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：Csv、txt 文件直接存储在 /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 FP32。

被测模型参数量：7B。

推理测试数据：MMLU。

推理批数：1。

推理引擎：未使用。

A.4.3 环境条件

温度：24℃。

湿度：50%。

气压：101KPa。

A.5 测量工具

Qwen2.5-7B-base 模型测评用到的工具为功率分析仪。

A.6 测评步骤

Qwen2.5-7B-base 模型按 JH/T/DZJN XX-YYYY 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 次推理，计算平均运行能耗、平均运行时长以及相对能耗。测试过程中无设备故障、推理终止等异常情况。

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	推理时长（小时）	推理能耗（KW・h）
1	0.09	285
2	0.1	288
3	0.11	291

平均时长为 0.1 小时，平均绝对能耗为 288 KW・h。

A.8 相对能耗测评结果

基准值：200 KW・h。

S 值：1.44。

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 41867-2022 信息技术 人工智能 术语
[2] GB/T 42018-2022 信息技术 人工智能 平台计算资源规范
[3] GB/T 45288-2025 人工智能 大模型 第 1 部分：通用要求
[4] GB/T 2589-2020 综合能耗计算通则
[5] GB/T 29239-2024 移动通信设备节能参数和测试方法 基站