ICS 33.060
CCS M 37

团　体　标　准

T/TAF 312—2025

智能终端大模型计算性能基准评测方法

Benchmark test methods for smart terminal LLMs computing performance

2025-08-11 发布　　2025-08-11 实施

电信终端产业协会　发布

版权声明

邮箱：tafrb@taf.org.cn

电话：010-82052809

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。

本文件由电信终端产业协会（TAF）提出并归口。

本文件起草单位：中国信息通信研究院、高通无线通信技术（中国）有限公司、北京三星通信技术研究有限公司、维沃移动通信有限公司、翱捷科技股份有限公司、中兴通讯股份有限公司、联想（北京）有限公司、紫光展锐（上海）科技有限公司、博鼎实华（北京）技术有限公司、中国移动通信集团终端有限公司、上海移芯通信科技股份有限公司。

本文件主要起草人：刘恩琦、王健宇、周奎翰、刘洋、王彬、高立发、李维成、龙迪、张宏伟、沙通、李丛蓉、马凡、张伟、梁恒康。

智能终端大模型计算性能基准评测方法

1 范围

本文件规定了智能终端大模型计算性能基准测评指标、测试方法。面向终端运算设备（芯片、智能手机、平板电脑、等设备）设计基准测试集，测试终端运算设备的计算性能。

本文件适用于智能手机、可穿戴设备、平板电脑、个人计算机等智能终端产品生成式AI与大模型技术的部署和应用，可不限于本参考框架的指标项或条款项，应符合标准中的定义和规范性描述。

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

YD/T 4515—2023 移动智能终端智能化性能基准测评方法

3 术语和定义

下列术语适用于本文件。

3.1

人工智能 artificial intelligence

表现出于人类智能（如推理和学习）相关的各种功能的功能单元和能力。

3.2

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。

注：大模型的参数量由其功能和模态决定，一般不低于 1 亿。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

生成式人工智能 generative artificial intelligence

基于数据、算法、模型、规则，能够根据使用者提示生成文本、图片、代码、音频、视频等内容的人工智能服务。

3.4

基准测试 benchmark

通过设计科学的测试方法、测试工具和测试系统，实现对一类测试对象的某项性能指标进行定量和可对比的测试。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

CPU：中央处理器（Central Processing Unit）

DSP：数字信号处理器（Digital Signal Processer）
GPU：图形处理（Graphics Processing Unit）
LLM：大语言模型（Large Language Model）
NPU：神经网络处理器（Neural Network Processing Unit）
SDK：软件开发工具包（Software Development Kit）

5 测试架构

5.1 概述

智能终端大模型计算性能基准评测指通过运行终端侧大模型算子、基准大模型来评测智能终端设备相关性能，旨在全面评估智能终端及其计算芯片在执行大模型时的整体和细节计算性能指标，包括但不限于算子级别的计算性能、整体模型、应用的计算性能、推理速度、内存占用、功耗等关键性能指标。具体测试框架见图1。

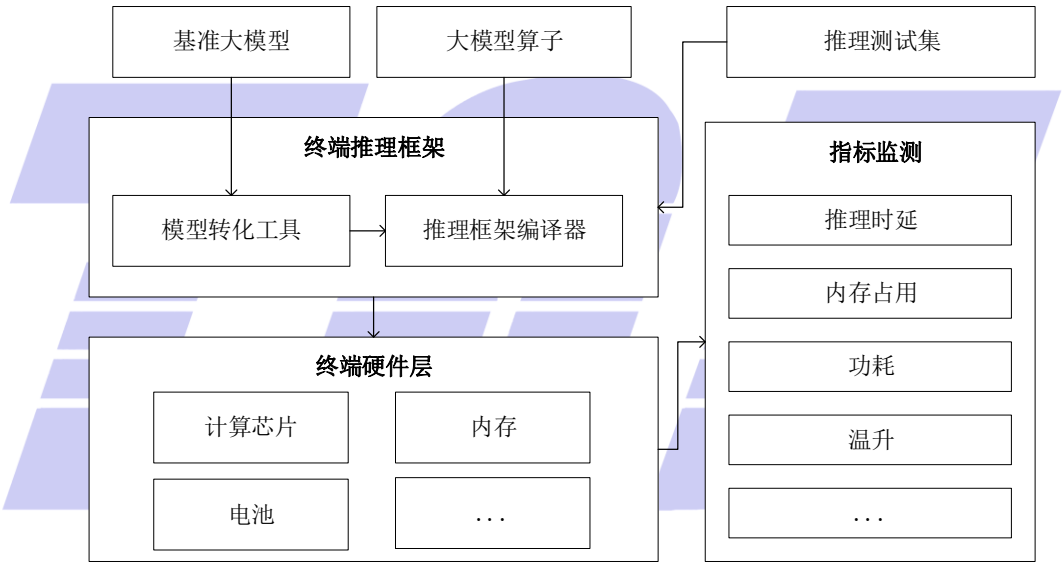

图1 大模型计算终端基准测试架构

5.2 基准大模型

基准大模型是指在基准测试过程中用于评估终端设备计算性能的大模型。应选用领域中具有代表性和可靠性的模型，覆盖不同的应用场景和计算负载。

5.3 大模型算子

大模型算子是构成大模型的基本计算单元，算子的性能直接影响整个模型的性能。测试过程中宜选用大模型的基础算子，包括卷积算子、矩阵乘积算子、注意力算子等。

5.4 推理数据集

作为推理计算的测试数据集，可为图片、文本等格式的数据或文件。

5.5 终端推理框架

终端推理框架是指针对终端设备进行模型部署和优化的一种工具，通常由模型转换工具和推理框架编译器等组成。在测试过程中应指明使用的终端推理框架。

模型转化工具：将大模型进行量化压缩和优化。大模型在模型转化的过程中可以通过模型量化进一步减小模型大小，即将值的精度从全浮点降低到半精度浮点（Float16)、8位整数（Int8）或更低，进一步减低模型所需的内存空间。

推理框架编译器：主要通过推理框架编译器执行推断计算，支持转换工具优化过的大模型，并提供计算所需的硬件资源的调度和使用：CPU、GPU、DSP、NPU等。

5.6 终端硬件层

测试对象，终端大模型处理涉及的硬件，包括CPU、GPU、NPU、AI硬件加速单元，内存,电池等。

5.7 指标监测

指标监测包括监测大模型推断计算性能的通用硬件性能指标和推理效果指标。

通用性能指标详见第6章内容。

6 测试环境

测试环境应符合以下要求。

——环境温度：（23±5）℃或满足被测设备产品要求说明书的要求。

——相对湿度：（20%～60%）%RH或满足被测设备产品要求说明书的要求。

——大气压强：（86～106）kPa或满足被测设备产品要求说明书的要求。

——供电电源：（220±11）V，（50±1）Hz或满足被测设备产品要求说明书的要求。

——其它：无影响仪器正常工作的电磁干扰及机械振动。

7 通用测试指标

7.1 推理时延

推理时延指推理任务从执行到终止的运行时间,即从内存发送样本数据到模型输出推理结果的时间间隔，单位为 s。

计算方法：记录开始推理的时间戳和结束推理的时间戳，两者之差即为推理时延。

7.2 内存占用

内存占用为测试过程所需的内存值，包括内存峰值占用和内存平均占用，单位为 GB。

计算方法：监控推理过程中的内存峰值或平均占用量。

7.3 功耗

功耗测试为测试过程中的平均电流，单位为 mA。

计算方法：监控推理过程中总电量消耗，平均电流可以通过使用功耗测试仪测量设备在一段时间内的电流消耗并取平均值得到。

——计算平均电流值：

$$I_{avg} = \frac{\sum_{i=1}^{n} I_i}{n}$$

式中：

I_i——每次测量的电流值；

n——总测量次数。

——计算总电量消耗：

$$E = V_{avg} \times I_{avg} \times t$$

式中：

V_{avg}——平均电压；

I_{avg} ——平均电流；

t ——测试时间。

7.4 温升

推理过程中硬件组件的温度的升高量，单位为℃。

计算方法：使用温度传感器监测关键组件（如 CPU、GPU）的温度，温升量等于测试结束的温度减去测试前的温度。

8 测试方法步骤

8.1 测试准备

在进行基准测试之前，需要确保所有的测试设备、软件和工具都已经准备就绪，包括但不限于：

——对于被测设备，应屏蔽测试无关的其他应用、后台功能、调整屏幕亮度、记录初始电量等，使得每次测试前终端的运行状态保持一致；

——确保被测设备电源充足，对于移动设备，宜连接电源进行测试；

——测试应进行被测设备的硬件检查，确保测试设备的硬件资源（如CPU、GPU、内存、存储等）处于正常工作状态；

——测试前应清除被测设备缓存，避免测试过程中缓存数据对测试结果的干扰；

——测试前应对设备进行内存与存储清理，确保设备在测试时不会因为内存不足或存储空间不足而影响性能；

——对于被测设备，应关闭网络连接，如Wi-Fi、蓝牙、移动数据等，避免外部网络流量对测试造成干扰；

——配置测试环境，确保环境条件符合第5章测试环境的要求。

8.2 测试布置

测试系统的连接示意如图2所示，测试电脑通过待测设备的接口与其相连，功耗测试仪与待测设备连接，通过基准测试工具对待测设备计算推理时间及性能评价指标进行测试，功耗软件测试工具对待测设备进行示波器反馈电流电压结果进行功耗分析。

基准测试工具应至少具备下述功能：向待测设备下发大模型算子计算任务，向待测设备下发大模型性能计算任务，从待测设备读取第6章中所述的推理时延、内存占用、温升等计算通用测试指标所需的数据。

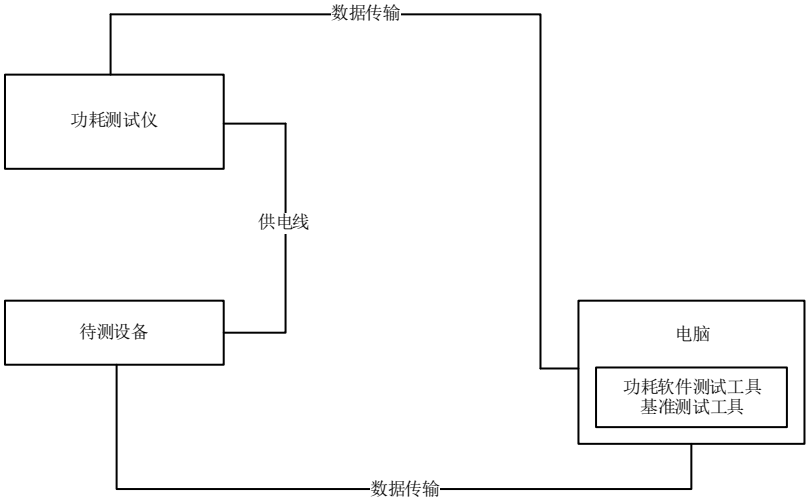

图2　测试布置示意图

8.3　测试执行

执行测试时，应按照以下步骤进行：

a)　按照7.1节要求进行测试前的准备；

b)　按照7.2节要求进行测试布置；

c)　执行数据预处理，具体方法如下：

- 对于文本数据，应进行分词和标准化。采用与模型一致的tokenizer进行文本分词（如BPE、WordPiece等）；统一大小写、去除无效符号和停用词（如去掉HTML标签、特殊字符、无关的标点符号等）；并确保数据集的输入格式与模型的输入要求一致（如特定的token长度、padding处理等）；
- 对于图像数据，应进行图像大小调整和归一化处理。采用一致的图像尺寸、色域、分辨率等；
- 对于音频数据，应进行音频数据的标准化处理。采用一致的通道数、采样率、采样深度和音频时长等；
- 执行上述数据预处理后，应将数据特性进行记录。

d)　运行基准测试工具，记录所有相关的性能指标。

9　大模型算子基准评测方法

9.1　被测设备操作系统

根据被测设备支持的操作系统，可以为 Windows、macOS、Linux、Android、iOS 或者 HarmonyOS。

9.2　算子调用 SDK

被测方应提供算子调用接口函数及其详细说明。

9.3　参考实现

本节给出大模型算子的基准测试参考实现，包括测试中包含的算子类型、计算芯片类型、输入数据

精度和推理计算方式，具体参考实现见图 3。

<table>
<tr><th>算子类型</th><th>计算芯片类型</th><th>数据精度</th><th>推理计算方式</th></tr>
<tr><td>矩阵乘法</td><td rowspan="2">CPU</td><td>FP32</td><td rowspan="3">单核运行性能</td></tr>
<tr><td rowspan="2">卷积</td><td rowspan="3">FP16</td></tr>
<tr><td rowspan="3">GPU</td></tr>
<tr><td>池化</td><td rowspan="3">INT8</td><td rowspan="4">并行计算性能</td></tr>
<tr><td rowspan="2">归一化</td><td rowspan="3">NPU</td></tr>
<tr><td>INT4</td></tr>
<tr><td>激活函数</td></tr>
</table>

图3　参考实现示意图

9.4　推理指标

大模型算子的推理指标应包括第 6 章通用测试指标内容。

10　大模型计算性能基准评测方法

10.1　被测设备操作系统

根据被测设备支持的操作系统，可以为 Windows、macOS、Linux、Android、iOS 或者 HarmonyOS。

10.2　推断框架 SDK

被测方应提供提供推断框架及其详细说明，提供接口函数包括初始化 Init()，预处理 PreProcess()，加载模型 LoadModel()，运行 Run()和后处理 PostProcess()。

10.3　前置信息披露

包含模型原始准确率，及转换后模型精度（浮点/定点），模型精度包括但不限于 FP32、FP16、FP8、INT8、INT4 及混合精度等。

10.4　模型文件

提供原始训练模型及其相关信息包括模型类别，输入输出节点名，前处理时均值及其归一化参数，张量信息(输入及输出)，通道信息(RGB/BGR)，数据格式(NHWC)等。

10.5　参考输出

由于评估的应用场景及网络存在差异，评估的指标也各不相同。基准测试会根据被测对象提供相应参考输出来测量被测设备或芯片的大模型计算与支持能力。

10.6 测试方法

10.6.1 参考测试用例 1：大语言模型文本问答测试

10.6.1.1 测试描述

此测试旨在评估大语言模型处理文本问答任务的能力。模型需要理解问题的上下文并生成准确的答案。

10.6.1.2 推理集要求

推理数据集应包含多样化的问题集，涵盖广泛的主题和复杂性。数据集应从公开可用的资源中获取，例如 MMLU 或其他类似的问答数据集，数据集应进行标注，标注格式应至少包含以下数据字段：

——问题：清晰、具体的问题描述；
——选项：包含多个选项的列表；
——答案：正确答案。

10.6.1.3 模型要求

评测模型可参考选取下列大语言模型：

——Llama2 7B；
——Llama3 8B；
——ChatGLM 7B；
——Phi-3 Mini；
——Qwen2 1.5B；
——Gemma 2B。

10.6.1.4 测试步骤

测试步骤：

a) 从推理集中依次选取问题；
b) 记录模型开始推理前的时间戳；
c) 将问题输入模型，获取模型生成的答案；
d) 记录模型输出答案的时间戳；
e) 重复步骤 a)-d)，直到完成所有选定测试数据的推理；
f) 计算并记录所有测试数据的性能指标。

10.6.1.5 测试指标

参考测试指标可参考如下指标：

——通用测试指标：指标名称与计算方法见第 6 章；
——首词响应时间：用户完成执行操作到接收到响应所需的时间，单位为 s；
——出词速率：从第一个词出现至结束，单位时间平均的出词量，单位为 tokens/s；
——正确率：模型生成的答案与标准答案的匹配度。

10.6.2 参考测试用例 2：大模型文生图测试

10.6.2.1 测试描述

此测试旨在评估大模型生成图像的能力。模型需要理解输入关键词并生成相符的图片。

10.6.2.2 推理集要求

数据集格式应满足下列要求：
——输入提示词应小于 10 个；
——数据格式中应包含描述文本；
——可选的公开数据集：mg18。

10.6.2.3 模型要求

评测模型可以选择下列模型：
a) Stable Diffusion；
b) DALL E2。

10.6.2.4 测试步骤

测试步骤：
a) 从推理数据集中依次选取提示词；
b) 记录模型开始推理前的时间戳；
c) 将提示词输入模型，获取模型生成的图片；
d) 记录模型输出图片的时间戳；
e) 重复步骤 1)-4)，直到完成所有选定测试数据的推理；
f) 计算并记录所有测试数据的性能指标。

10.6.2.5 测试指标

参考测试指标可参考如下指标：
——通用测试指标：指标名称与计算方法见第 6 章；
——吞吐量：模型单位时间内输出图片的数量，单位为图片数/s。

电信终端产业协会团体标准

智能终端大模型计算性能基准评测方法

T/TAF 312—2025

*

电信终端产业协会发布
地址：北京市西城区新街口外大街 28 号
电话：010-82052809
电子版发行网址：www.taf.org.cn