ICS 33.050
CCS M30

团　体　标　准

T/TAF 327—2026

面向智能手机的端侧大语言模型技术要求

Technical requirements for on-device large language model for smartphones

2026-02-09 发布　　2026-02-09 实施

电信终端产业协会　发布

版权声明

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：中国移动通信集团终端有限公司、中国信息通信研究院、维沃移动通信有限公司、中兴通讯股份有限公司、安谋科技（中国）有限公司、翱捷科技股份有限公司、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、紫光展锐（上海）科技有限公司、博鼎实华（北京）技术有限公司、联想（北京）有限公司、小米通讯技术有限公司、南德认证检测（中国）有限公司深圳分公司、上海移芯通信科技股份有限公司。

本文件主要起草人：王绍颖、董千洲、黄云霞、王健宇、范洪源、傅蓉蓉、高立发、曹宇琼、李根、李丛蓉、彭程、潘正、周世乐、董雾、刘妍、王彬、龙迪、曾勇波，聂大伟、吴术霞、耿琦、王骏超、梁恒康。

引　言

在大语言模型席卷全行业的大背景下，移动终端凭借其个人专属化、多模态感知以及强大的通信计算能力，成为未来移动大语言模型的理想载体。随着端侧大语言模型的持续进化，智能终端将不再局限于实现文生文、文生图、文生视频等基础应用，而是逐步发展为一个全方位的“移动智能体”。它不仅能够完成通信任务，更将成为人工智能的最佳载体，为用户开启一个崭新的 AI 交互领域。各大终端品牌厂商纷纷加大在大语言模型端侧部署方面的创新力度，领先的芯片厂商也在不断提升移动平台的 AI 能力。

大语言模型技术不断深化与落地，轻量化、多模态等技术的持续升级，有效推动了端侧智能的发展。通过将计算任务从云端迁移到终端，端侧智能具备了更高的独立性、低时延和高可靠性，逐渐成为实现万物智能的关键途径，能够更好地满足大语言模型在隐私保护、实时响应、网络负载和灵活部署等方面的需求。构建面向智能手机的端侧大语言模型技术，不仅能够提升用户体验，减少对云端计算的依赖，促进技术创新和发展，还能满足行业对数据安全和隐私保护的严格要求。随着技术的不断成熟与应用的持续拓展，端侧大语言模型将在未来发挥更加重要的作用。

为了规范和评估终端侧大语言模型应用场景和任务，提高消费者用户的使用体验，亟待相关标准制定工作，规范并促进大语言模型在智能手机上的应用，促进国内相关行业的发展。

面向智能手机的端侧大语言模型技术要求

1 范围

本文件规定了面向智能手机的端侧大语言模型技术要求。

本文件适用于指导模型开发商、第三方测评机构对端侧大语言模型进行模型能力进行测试评估等工作。

2 规范性引用文件

本文件无规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

端侧大语言模型 on-device large language model

部署在端侧的大语言模型，一种规模庞大、基于大量数据训练得到参数众多的模型，主要用于处理文本相关任务，如文本生成、分类、翻译等，通过对大量文本数据进行学习来提升在相应文本任务中的性能且具备一定泛化性的深度学习模型。

4 缩略语

下列缩略语适用于本文件：

GB：吉字节（Gigabyte）

INT：整数（Integer）

NPU：神经网络处理器（Neural Network Processing Unit）

5 概述

5.1 前置条件

5.1.1 端侧大语言模型部署

本文件不规定端侧大语言模型部署方式和推理框架，智能手机内置大语言模型本文件不考虑。

5.1.2 端侧大语言模型体积

大语言模型体积指大语言模型部署在端侧时的空间占用。

计算方法：记录模型文件在端侧所占用的存储空间大小，单位GB。

表1所示为模型参数量在1B及以下、1B-3B、3B及以上的模型INT4与INT8量化下，模型文件体积不宜超过的大小。

表1　大语言模型体积

量化方式	模型规模		
	1B及以下	1B-3B	3B及以上
INT4量化	0.5GB	1.5G	≤RAM*80%大小
INT8量化	1GB	3GB	≤RAM*80%大小

5.2　技术要求指标

技术要求指标请见表2。

表2　技术要求指标

一级指标	二级指标	三级指标
模型性能	加载时延	加载时延
	推理速度	首词响应时延
		出词速度
	内存占用	内存峰值、平均内存占用
	增量功耗	增量功耗

6　模型性能

6.1　加载时延

模型加载时延指把大语言模型文件从存储设备加载到内存中所需的时间。

计算方法：记录初始化开始的时间戳t_0和初始化结束时的时间戳t_1，两者之差即为模型加载时间t，单位s，计算公式：

$$t = t_1 - t_0 \quad \cdots\cdots (1)$$

下表3展示了参数量在 1B 及以下、1B－3B、以及3B及以上的模型，在INT4和INT8量化条件下，端侧加载延时的建议上限。所有数据均为模型在端侧加载20次后的平均耗时，指标要求请见表3。

表3　加载时延要求

量化方式	模型规模		
	1B及以下	1B-3B	3B及以上
INT8量化	4s	6s	10s
INT4量化	2s	4s	8s

6.2　推理速度-首词响应

首词响应时延是指用户感受到大语言模型推理服务的响应时间，即从用户发送样本数据到大语言模型生成并返回第一个字符所需的响应时间。

计算方法：记录将文本输入到大语言模型的时间戳t_0和大语言模型返回第一个字符的时间戳t_1，两者之差即为首词响应时延t，单位s。

$$t = t_1 - t_0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2)$$

本文件的测试样本token长度分别为128、1024 和4096。针对参数量在1B及以下、1B－3B、以及3B及以上的模型，规定在INT4和INT8量化条件下，经过20次测试所得的平均首词响应时延不宜高于表4所列上限。

表4　首词响应要求

模型规模	量化方式	被测设备（内置NPU）		
		128 tokens	1024 tokens	4096 tokens
	INT8	1.5s	2.0s	3.0s
	INT4	1.2s	1.8s	2.5s
1B－3B	INT8	2.5s	3.0s	4.0s
	INT4	1.5s	2.5s	3.2s
3B及以上	INT8	5.0s	5s	5s
	INT4	5.0s	5s	5s

6.3　出词速度

出词速度表示大语言模型单位时间内生成的字符数是多少。

计算方法：记录大语言模型生成第一个字符时的时间戳和生成最后一个字符时的时间戳，两者之差记为t；记录生成的字符数为N，出词速度S，计算公式如下，单位为tokens/s。

$$S = \frac{N}{t} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3)$$

表5展示了参数量在1B及以下、1B－3B以及3B及以上的模型，在原始精度、INT4和INT8量化条件下，端侧推理20次后获得的平均出词速度下限要求。模型出词速度不宜低于下表5所列数值。

表5　出词速度要求

模型规模	量化方式	被测设备（内置NPU）
1B及以下	INT8	16 tokens/s
	INT4	20 tokens/s
1B－3B	INT8	5 tokens/s
	INT4	12 tokens/s
3B及以上	INT8	4 tokens/s
	INT4	6 tokens/s

6.4　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。

计算方法：监控推理过程中的内存占用。假设调用大语言模型前的内存占用为d_0，调用大语言模型并直到大语言模型推理结束，监控到内存占用达到的最大值为d_1及平均值d_2。则内存计算方式如下：

a） 内存峰值：

$$d_{max} = d_1 - d_0 \quad \cdots\cdots (4)$$

式中：

d_{max}——内存峰值；

d_1 ——内存占用达到的最大值；

d_0 ——调用大语言模型前的内存占用。

b） 内存平均占用：

$$d_{avg} = d_2 - d_0 \quad \cdots\cdots (5)$$

式中：

d_{avg}——内存平均占用；

d_2 ——平均值；

d_0 ——调用大语言模型前的内存占用。

本文件规定测试端侧大语言模型平均20次的内存峰值宜不超过智能手机内存的90%，内存平均占用宜不超过智能手机内存的80%。

6.5 增量功耗

增量功耗测试为测试端侧大语言模型运行过程中的平均电流。

计算方法如下：

a） 监控被测设备待机中总电量消耗为N_1，平均电流I可以通过使用功耗测试仪测量设备在一段时间内（t）的耗电量N_1计算得到，单位为mAh。

b） 监控推理过程中总电量消耗为N_2，平均电流I可以通过使用功耗测试仪测量设备在一段时间内（t）的耗电量N_2计算得到，单位为mAh。（以4V电压下作为标准）。

$$I = \frac{N_2 - N_1}{t} \quad \cdots\cdots (6)$$

表6示为模型参数量在1B及以下、1B-3B、3B及以上的模型在原精度、INT4与INT8量化下，增量功耗不宜过高，均为在端侧推理20次的平均增量功耗，测试时间t不小于1h。

表6 增量功耗要求

量化方式	模型规模		
	1B及以下	1B-3B	3B及以上
INT4量化	600mAh	900mAh	1100mAh
INT8量化	700mAh	1000mAh	1200mAh

电信终端产业协会团体标准

面向智能手机的端侧大模型技术要求

T/TAF 327—2026

*

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