基于大模型的AI芯片评测报告（0.2.1）

一、测试概论

本测试报告的评测逻辑基于上海人工智能实验室《面向国产硬件的大模型评测实施方案（0.1.2）》（以下简称《方案》），涉及的基准值由实验室提供。本报告对送测的计算机视觉领域面向云侧的深度学习训练芯片的基本技术规格，**大模型训练、微调、推理**的功能、性能、稳定性等指标进行测试总结 本次评测 **XXX** 芯片 **XXX**，综合性能为A100芯片的XX%，**测试结果符合预期**。

（此处插入芯片介绍，内容由厂商提供）

在**大模型训练中**，XXX对主流大模型支持度完整。综合性能约为A100的XX%，其中计算性能的得分为XXX，在BERT模型下计算性能较好，能达到基准的XX%。在LLAMA模型上和基准相当。GPT-3模型上则高于基准XX%左右。通信性能的得分为XXX，高于A100基准XX%。模型收敛和训练稳定性较好/一般/较差。

在**大模型微调**中，XXX对主流大模型微调支持度完整，综合性能约为A100的XX%，性能较好。

在**大模型推理**中，XXX文本生成整体性能达到A100的XXX%。XXX芯片的文生图的整体性能达到A100的XXX%。

测试结果汇总表如下：

测试大项	测试小项	评分
大模型训练	功能	
	性能	
	稳定性	
大模型微调	功能	
	性能	
	稳定性	
大模型推理	文本生成	
	文生图	

*得分说明：

1. 未提交数据的测试条目显示为“NA”，记为0分

2. 属于“本身不达标/未提供log/验收未通过”其中一种情况的测试项显示为0分，记为0分
3. 大模型报告不再显示测试项目权重和总分，总分仅在《AI芯片评测报告》中体现

二、测试平台和测试环境

1. 测试平台

上海人工智能实验室牵头的硬件评测基于团体标准评测方法，对送测芯片进行技术规格、软件生态、功能、性能、模型支持等多维度评测，并按季度产出硬件评测报告，评测结论可为各类国产加速卡在不同维度的表现提供参考。目前硬件评测设计了两个层面的实施方案，分别是《AI芯片评测实施方案》和《基于大模型的AI芯片评测实施方案》，以应对市场对硬件在不同场景下的能力要求。（见下图）

基础能力	基本技术规格	芯片基础信息
		算力
		内存
		通讯带宽
		能效比
	软件生态	软件栈
		高性能计算库
		开放性
	算子功能	算子功能
	模型功能	模型功能
	算子性能	CONV
		GEMM
		Transformerblock
		长尾算子
	模型性能	模型性能
	通信性能	AllGather
		AllReduce

芯片基础能力考查

大模型	基础信息	模型介绍
		芯片介绍
		基础信息
	训练	功能指标
		性能指标
	微调	功能指标
		性能指标
	推理	文本生成推理性能指标
		文生图推理性能指标
	其他指标	稳定性指标
		通信指标
		检查点存储|加载性能指标

芯片大模型能力考查

评测结论可作为芯片生产厂商、应用厂商、前场销售及第三方机构对深度学习训练芯片（包含AI芯片模组和AI加速卡等形态）进行设计、采购、评测的参考。一方面评测的结果将为硬件厂商提供宝贵的参考信息，帮助他们了解自身在大模型领域的发展方向和优化需求。另一方面，评测结果也将为用户和开发者提供有价值的参考，帮助他们选择适合自己需求的硬件设备，并优化和改进应用程序的性能和效果。我们期待通过这一评测方案的实施，能够为国产硬件的发展注入新的动力，推动大模型技术的普及和应用，促进语言处理领域的创新和进步。

自2022年以来，实验室已主导进行了5次芯片评测工作，吸引了10余家国产芯片厂商积极参与，累计产出了40余份评测报告。部分企业已经将评测内容、方法和流程引入到自家测试流程中进行指导芯片研发方向。

2. 测试芯片

测试芯片为XXXXX，基准值选取英伟达A100，芯片简要介绍如下，具体技术参数详见：基本技术规格。

说明：

（此处插入芯片介绍，内容由厂商提供）

本报告基准芯片使用英伟达A100主要针对 AI、数据分析和 HPC 应用场景，在不同规模下实现出色的加速。英伟达A100是目前人工智能模型领域使用最多的芯片。

3. 测试环境

3.1 硬件信息

本报告的测试基于以下环境：

序号	Item	Info/Version
2	CPU型号	
3	节点内连接	
4	节点间连接	
5	操作系统	
6	GCC	
7	Python	
8	Pytorch	

3.2 软件栈版本

测试使用的代码库及其版本信息:

序号	Name	Version/Branch
1	InternLM	
2	Bert-Pytorch	
3	Alpaca-Lora	
4	LLaMa	
5	Stable-Diffusion	

4. 测试方法

上海人工智能实验室与各芯片厂商联合开展评测。在制定好实施方案并确定所测模型后，厂商需要严格按照实施方案对自家芯片进行测试，并将验证数据与日志整理并发送到实验室制定的数据上传地址。在整个模型适配过程中，厂商只可进行有限的代码修改，以确保模型能够在不同芯片上公平公正的评测，也防止厂商通过不合理的手段提高测试数据。在产生测试结果并提交数据后，实验室会对厂商所提交的数据进行上机抽查验证与日志验证，确保数据真实可靠。验证结束后，实验室会与厂商同步测试结果与芯片表现。同时，所有测试代码均已开源，测试过程、数据均可复现。

三、测试内容和测试数据

1. 基本技术规格

基础技术规格内列举的项目详见基础模型评测报告，不再重复计分。

表4 基础技术参数

项目	基准值	测试值
FP16算力 (TFLOPS)	312	
FP32算力 (TFLOPS)（非Tensor Core）	19.5	
FP32算力 (TFLOPS)（Tensor Core）	156	
INT8算力 (TOPS)	624	
INT16算力 (TOPS)	312	
BF16算力 (TFLOPS)	312	
TF32算力 (TFLOPS)	156	
容量 (GB)	80	
带宽 (GB/s)	2039	
主机-设备带宽 (h2d)(GB/s)	22.3	
设备-主机带宽 (d2h)(GB/s)	15.0	
节点内卡间带宽 (GB/s)	600	
单张显卡内数据转移带宽 (d2d) (GB/s)	1587.6	
最高浮点算力能效比 (TFLOPS/W)	0.557	
最高整型算力能效比 (TOPS/W)	1.116	

2. 大模型训练

2.1 功能指标

测试内容：

测试训练芯片以及其软件栈是否支持大模型的训练以及是否满足对应技术要求。

测试方法：

参照《方案》附录A待测大模型列表，给定模型测试数据集、超参配置、要求训练轮数以及测试精度要求，当训练轮数达到训练要求的轮数时，测试模型训练目标（Loss、Loss与标准值方差、PPL）满足要求。

最后模型支持率，公式如下：

$$模型支持率 = \sum \alpha_i \cdot [模型_i是否支持]$$

其中，α_i 为第i个模型的权重系数，每个模型权重系数如列表所示，根据模型发布时间、技术热度、参数规模等决定。

测试数据：

模型名称	权重	Loss		Loss与标准值方差		PPL		得
		基准值	测试值	基准值	测试值	基准值	测试值	
	15%							
	40%							
	30%							
	15%							
总分								

2.2 性能指标

（1）计算性能

测试内容：

大模型在不同配置（单节点、多节点）下训练的效率和速度。

定义训练计算性能的核心指标TGS(tokens/gpu/second)，表示每秒单块GPU能够处理token的数量。

其中TGS计算公式如下：

$$TGS = \frac{tokens \cdot GlobalBatchSize}{每轮迭代时间 \cdot GPU卡数}$$

测试方法：

测试的模型来自《方案》附录A。流程如下：

1. 准备模型训练所需的参数、数据集，训练过程不能对设定参数进行修改。

2. 启动模型训练，执行M（M<50）次迭代（iter）或步数（step）训练作为暖身轮。

3. 至少执行一个完整的训练轮（epoch）或超过规定迭代（iter）或步数（step），根据《方案》6.3.6中定义计算模型的训练性能。

模型训练TGS性能单项得分，其计算公式如下：

$$TGS性能单项得分 = \frac{TGS}{TGS_{baseline}}$$

其中，该模型TGS性能得分算术平均的公式为：

$$模型TGS性能 = \frac{1}{N} \cdot \sum_{i=1}^{N} 模型_i的TGS性能单项得分$$

最后，大模型性能总分，其计算公式如下：

$$TGS性能总分 = \sum_{i=1}^{N} 权重W_i \cdot 模型_i的TGS性能$$

测试数据：

模型名称	卡数	权重	TGS(带宽XXXGb/s)		性能得分
			基准值	测试值	
总分					

（2）通信性能

测试内容：

测试5种不同的通信操作（all_reduce、all_gather、all_to_all、broadcast、pt2pt）在8卡和32卡下通信性能。对于大模型来说all reduce是影响训练性能的核心因素，因此权重占比最高。基准数据采用DeepSpeed Communication Benchmarking Suite在A100下获取。

测试方法：

每一个测试配置（节点配置和通信负载）下的计算公式如下：

$$通信单项配置得分 = \frac{基准耗时}{测试耗时} \cdot 0.34 + \frac{测试吞吐}{基准吞吐} \cdot 0.33 + \frac{测试带宽}{基准带宽} \cdot 0.33$$

通信性能的总计算公式如下：

$$\text{通信性能评分} = \alpha_i \cdot \text{配置}_i\text{的通信性能}$$

其中，α_i 表示第i项配置的权重系数。

测试数据：

操作	卡数	权重	Size （Bytes）	Description（数据量*字节数）	Duration（ms）		Throughput (Gbps)	
					基准值	测试值	基准值	测试值
all_reduce	8	25.00%	15.86 GB	4259924984x4	126.905		2148.338	
all_reduce	32	25.00%	15.86 GB	4259924984x4	373.565		729.815	
all_gather	8	6.30%	512.0 MB	134217728x4	16.915		2031.292	
all_gather	32	6.30%	128.0 MB	33554432x4	47.117		729.249	
all_to_all	8	6.30%	4 GB	1073741824x4	17.54		1958.977	
all_to_all	32	6.30%	4 GB	1073741824x4	314.13		109.38	
broadcast	8	6.30%	15.86 GB	4259924984x4	72.62		1877.133	
broadcast	32	6.30%	15.86 GB	4259924984x4	196.439		693.939	
pt2pt	8	6.30%	15.86 GB	4259924984x4	66.02		2064.778	
pt2pt	32	6.30%	15.86 GB	4259924984x4	762.939		178.673	
总分								

（3）检查点存储/加载性能

测试内容：

测试大模型训练过程中的Checkpoint保存和加载的性能。

测试方法：

通过日志输出的Checkpoint保存总耗时、Checkpoint加载总耗时分别与基准值比较计算该项评分：

$$\text{检查点存储/加载性能得分} = \frac{\text{测试}Ckpt\text{保存总耗时}}{\text{基准耗时}} \cdot 0.7 + \frac{\text{测试}Ckpt\text{加载总耗时}}{\text{基准耗时}} \cdot 0.3$$

测试数据：

模型名称	卡数	save-model time (ms)		save-optimizer time (ms)		all save time (s)		all load time (s)		sav
		基准值	测试值	基准值	测试值	基准值	测试值	基准值	测试值	
	8									
	32									
	32									
	128									
总分										

2.3 稳定性指标

测试内容：

测试大模型训练的稳定性，包括前100个step的Loss波动，用于反映模型训练在精度上的稳定性，以及3天内是否出现训练过程终止、显存溢出、Loss不收敛等与预期不符导致训练无法继续的崩溃情况。

测试方法：

1. 前100个step或iter的Loss与基准值的方差。

$$\text{稳定性方差} = \frac{1}{100}\sum_{i=1}^{100}(L_i - L_{mean})^2$$

其中 L_i 为第i个step的Loss与基准值的差值， L_{mean} 为这100个差值的平均值。

$$\text{稳定性方差得分} = 1 - min(1.0, \text{稳定性方差})$$

2. 记录72小时的训练崩溃次数C，以及每次崩溃后的平均恢复时间R。

$$\text{崩溃率得分} = \frac{1}{(C+1)^2}$$

3. 最长无故障工作时间（h）。

$$\text{最长无故障得分} = \frac{h}{72}$$

$$\text{稳定性得分} = \text{崩溃率得分} \cdot 0.6 + \text{最长无故障得分} \cdot 0.2 + \text{稳定性方差得分} \cdot 0.2$$

测试数据：

模型名称	卡数	前100个step的Loss波动		72小时内崩溃次数		最长无故障工作时间		前100个step的Loss波动得分	72小时内崩溃次数得分
		基准值	测试值	基准值	测试值	基准值	测试值		
InternLM-7B	8	0		0		72			
InternLM-65B	32	0		0		72			
总分									

3. 大模型微调

3.1 功能指标

测试内容：

测试训练芯片以及其软件栈是否支持大模型的微调以及是否满足对应技术要求。

测试方法：

参照《方案》附录B待测大模型列表，给定模型测试数据集、超参配置、要求训练轮数以及测试精度要求，当训练轮数达到训练要求的轮数时，测试模型训练目标（Loss、Loss与标准值方差、PPL）满足要求。

最后模型支持率，公式如下：

$$\text{模型支持率} = \sum \alpha_i \cdot [\text{模型}_i\text{是否支持}]$$

其中，α_i 为第i个模型的权重系数，每个模型权重系数如列表所示，根据模型发布时间、技术热度、参数规模等决定。

测试数据：

模型名称	Loss		Loss与标准值方差		PPL		得分
	基准值	测试值	基准值	测试值	基准值	测试值	
Alpaca-Lora	0.950		0	-	-	-	
InternLM-7B/LLaMa1-7B	1.350		0	-	-	-	
总分							

3.2 性能指标

（1）计算性能

测试内容：

大模型在不同配置（单节点、多节点）下微调的效率和速度。

测试方法：

测试的微调模型来自《方案》附录B。流程及性能计算公式参考大模型训练。

测试数据：

模型名称	权重	卡数	TGS		TGS（关闭int8量化）		TGS得分
			XXXGb/s		XXXGb/s		
			基准值	测试值	基准值	测试值	
Alpaca-Lora	0.25	8					
Alpaca-Lora	0.25	32					
InternLM-7B	0.25	8			-		
InternLM-7B	0.25	32					
总分							

（2）通信性能

大模型微调通信底层原理与预训练一致，直接使用大模型训练通信性能评分。

（3）检查点存储/加载性能

大模型微调检查点存储与加载原理与预训练一致，直接使用大模型训练检查点存储/加载性能评分。

3.3 稳定性指标

大模型微调稳定性评分直接使用大模型训练稳定性评分。

4. 大模型推理

4.1 文本生成推理性能

测试内容：

测试FP16/INT8 LLaMa v1 7B&65B 或 FP16/INT8 LLaMa v2 7B&70B大模型在不同并发、不同输入和输出长度下的性能指标：

（1）模型初始化加载时间：通过配置不同参数量大模型的并发数和batchsize，来测试不同组合下的模型初始化加载时间，响应时间和吞吐量，以评估服务的承载能力和性能。

（2）响应时间：指用户从发送请求的时刻到用户收到响应结果的时刻所经过的时间，它直接影响用户体验。对于大语言模型来说，因响应结果是流式输出，目长度不确定，所以我们把这个指标转化为：首字延迟（first token latency ）和 推理耗时，分别度量第一个token生成时所消耗的时间，以及平均每个token所需要的时间。

（3）吞吐量（token throughput）：指每秒生成的token数量，它与响应时间、并发数密切相关，用来衡量服务的承载能力。

测试方法：

首字延迟是影响用户体验的最主要因素，而吞吐又直接与推理成本相关。我们筛选了线上推理服务的数十万条日志，512的长度能够覆盖线上用户93%的需求，通过不同的长度组合模拟真实用户输入输出，取该组首字延迟和吞吐平均值反映不同情况下推理性能。然后根据不同需求场景下对首字延迟的要求，分为500毫秒以下、500毫秒至1秒、1秒至3秒、3秒以上一共四档，取满足该档要求的最大吞吐值，然后分档与基准值进行对比计算得分。

使用 FP16/INT8 LLaMa-V1/2 进行推理。

测试数据：

编号	模型	并发数	Batchsize	Input tokens	Output tokens	卡数		吞吐			首字延迟（ms）	
						基准值	测试值	基准值	测试值	实际值	基准值	测试值
1		1	1	256	128	1						
2		1	1		512	1						
3		1	1		1024	1						
4		1	1	512	128	1						
5		1	1		512	1						
6		1	1		1024	1						
7		1	1	1024	1024	1						
8		8	8	256	128	1						
9		8	8		512	1						
10		8	8		1024	1						
11		8	8	512	128	1						
12		8	8		512	1						
13		8	8		1024	1						
14		8	8	1024	1024	1						
15		16	16	256	128	1						
16		16	16		512	1						
17		16	16		1024	1						
18		16	16	512	128	1						
19		16	16		512	1						
20		16	16		1024	1						
21		16	16	1024	1024	1						
22		32	32	256	128	1						
23		32	32		512	1						
24		32	32		1024	1						
25		32	32	512	128	1						
26		32	32		512	1						

27		32	32		1024	1					
28		32	32	1024	1024	1					
29		64	64	256	128	1					
30		64	64		512	1					
31		64	64		1024	1					
32		64	64	512	128	1					
33		64	64		512	1					
34		64	64		1024	1					
35		64	64	1024	1024	1					
36		1	1	256	128	8					
37		1	1		512	8					
38		1	1		1024	8					
39		1	1	512	128	8					
40		1	1		512	8					
41		1	1		1024	8					
42		1	1	1024	1024	8					
43		8	8	256	128	8					
44		8	8		512	8					
45		8	8		1024	8					
46		8	8	512	128	8					
47		8	8		512	8					
48		8	8		1024	8					
49		8	8	1024	1024	8					
50		16	16	256	128	8					
51		16	16		512	8					
52		16	16		1024	8					
53		16	16	512	128	8					
54		16	16		512	8					
55		16	16		1024	8					
56		16	16	1024	1024	8					
57		32	32	256	128	8					
58		32	32		512	8					
59		32	32		1024	8					
60		32	32	512	128	8					
61		32	32		512	8					
62		32	32		1024	8					
63		32	32	1024	1024	8					
64		64	64		128	8					

64		64	64		128	8					
65		64	64	256	512	8					
66		64	64		1024	8					
67		64	64		128	8					
68		64	64	512	512	8					
69		64	64		1024	8					
70		64	64	1024	1024	8					
71		80	80		128	8					
72		80	80	256	512	8					
73		80	80		1024	8					
74		80	80		128	8					
75		80	80	512	512	8					
76		80	80		1024	8					
77		80	80	1024	1024	8					
78		96	96		128	8					
79		96	96	256	512	8					
80		96	96		1024	8					
81		96	96		128	8					
82		96	96	512	512	8					
83		96	96		1024	8					
84		96	96	1024	1024	8					

Input tokens 256							最大吞吐得分	模型加载时间得分
编号	模型	首字延迟范围（ms）	最大吞吐（90%）		模型加载时间（s）（10%）			
			基准值	测试值	基准值	测试值		
0		(0, 500]						
1		(500, 1000]						
2		(1000, 3000]						
3		(3000, +∞)						
4		(0, 500]						
5		(500, 1000]						
6		(1000, 3000]						
7		(3000, +∞)						
总分								

Input tokens 512							最大吞吐得分	模型加载时间得分
编号	模型	首字延迟范围（ms）	最大吞吐（90%）		模型加载时间（s）（10%）			
			基准值	测试值	基准值	测试值		
0		(0, 500]						
1		(500, 1000]						
2		(1000, 3000]						
3		(3000, +∞)						
4		(0, 500]						
5		(500, 1000]						
6		(1000, 3000]						
7		(3000, +∞)						
总分								

4.2 文生图推理性能

测试内容：

图像扩散模型在不同生成提供分辨率图像的耗时。

测试方法：

根据图片生成的时间计算单位时间内的吞吐量作为衡量文生图模型的推理性能。

使用 Stable Diffusion 1.5/2.1，图编译/原生模型加载时间。

测试数据：

Batch size	分辨率	内存占用（GB）		推理耗时（s）		吞吐量（fps）		模型加载时间		吞吐量得分	模型加载时间得
		基准值	测试值	基准值	测试值	基准值	测试值	基准值	测试值		
1	512*512										
	640*640										
	768*768										
2	512*512										
	640*640										
	768*768										
总分											