

AI Open Source for Good: Inclusive Access, Equitable Data, and Accessible Compute

Yonghua Lin
VP, BAAI



- ① **AI open-source for Good**
- ② **Our open-source activities for Data**
- ③ **Our open-source activities for Compute**



A sentence from common crawl

“We believe that everyone should have the opportunity to indulge their curiosities,
analyze the world, and pursue brilliant ideas. ”

—— from mission of Common Crawl



What is AI open-source for Good?

Premise: Enable AI to help us make humanity, society, and the environment better.

- **Inclusive AI:** Ensure everyone can enjoy the benefits AI brings.

- **Lower Barriers:** Reduce the obstacles to conducting R&D for “AI for Good,” so that anyone can have the opportunity to innovate and contribute.



What is today's challenge

- **High costs for large-model innovation:** Algorithm innovations for large models can cost millions of US dollars to validate.
- **No single organization can cover everything:** It's impractical for one organization to prepare all necessary components—data, benchmarks, and tooling—for large models.
- **Computational resources remain expensive and inaccessible:** Compute resources are still costly and hard for most R&D teams to obtain.



AI for Good

BAAI, a non-profit AI research institute dedicated to advancing human well-being, champions the ethos of AI for Good through:

- **Reduce the "redundant costs" of innovation:** Open-sourced large model algorithms and models, which can be used for both commercial and research purposes, and ensure reproducibility.
- **Build resources and share for innovation:** Open-sourced high-quality datasets for free use by global researchers, developers, universities, and enterprises.
- **Lower the hardware usage boundary:** Collaborate with organizations to build open-source and unified AI system software, enabling users to freely choose hardware and achieve lower cost.
- **Empower those most in need through AI:** Focus on the accessibility of AI, such as the elderly and children.



- ① AI open-source for Good
- ② Our open-source activities for Data
- ③ Our open-source activities for Compute



If we don't have these open datasets ...

IMAGENET

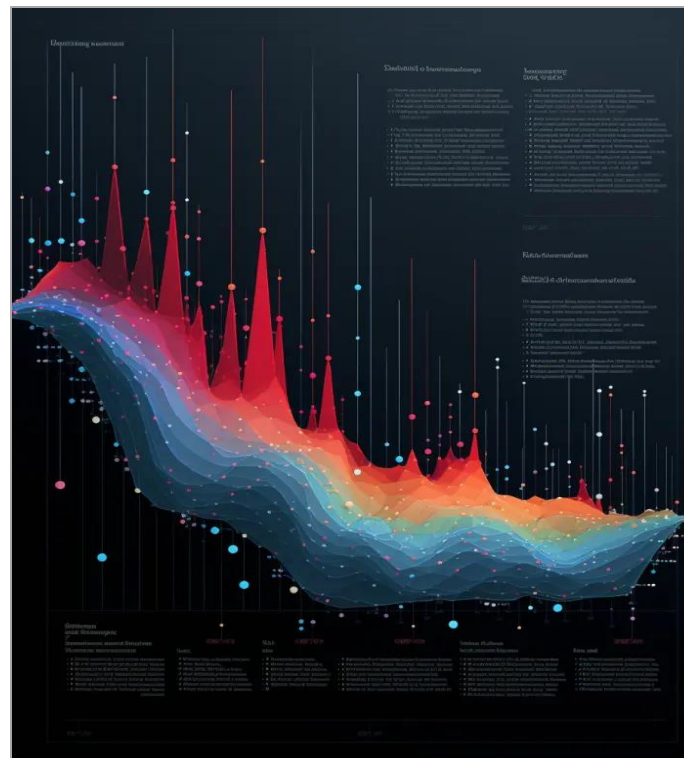
Started from 2007



More than **14 millions** labelled pictures with more than 22 thousands categories.

COMMON CRAWL

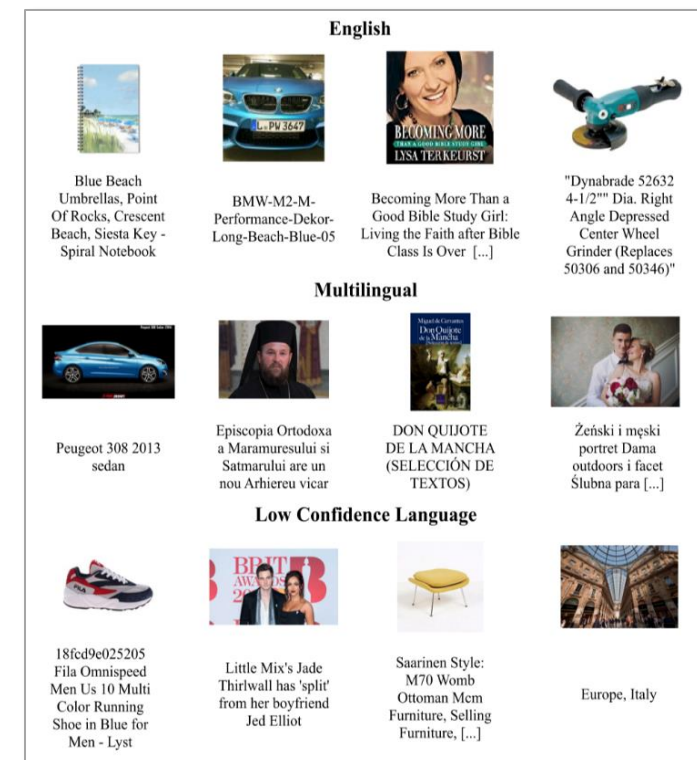
Started from 2007



More than **250 billions** webpages

LAION

Started from 2021



More than **5.8 billions** picture-text pairs

Without these open datasets, advancements in deep learning, large language models and multi-modal models would have slowed down significantly.

BAI 智源研究院 **GOSIM**

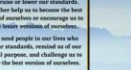
15

“

The people we surround ourselves with have a great influence on the way we think. They either help us to become the best of ourselves or they encourage us to become lesser versions of ourselves.

”

—Matthew Kelly



Infinity-MM

Instruction Tag 1: Logic Reasoning / Structured Image-Text Understanding / Panset context

Q1: <image>+inflow does Matthew Kelly's quote suggest that our choices influence who we become?</image>+Please provide the reasoning process first, and then give the answer.

A1: The quote suggests that individuals have control over their own development by surrounding themselves with people either supportive or unsupportive of high standards. Answer: Our choices about whom we associate with can help us grow into better versions of ourselves (by raising our expectations) or discourage such growth (if those around us lower them).

Instruction Tag 2: Logic Reasoning / Structured Image-Text Understanding / Text understanding

Q2: What qualities should we look for in our relationships with others according to this quote?

A2: We should seek positive influences in our lives, people who raise our standards, remind us about our purpose, and encourage us to improve ourselves.

Instruction Tag 3: Attribute Reasoning / Function Reasoning / Other functional reasoning

Q3: What effect can surrounding oneself with high standards have on personal development?

Answer: A3: The quote suggests that people who surround themselves with others holding higher standards may be positively influenced by this environment. They might raise their own levels of performance or motivation as they compare themselves against these elevated benchmarks.In Answer: Surrounding oneself with individuals having high standards could lead to increased self-improvement and drive for excellence due to comparison effects. It helps motivate one's growth towards becoming better versions of themselves.

Instruction Tag 4: Fine-grained Perception (cross-image) / Action Recognition / Recognize human actions

Q4: What physical activity is being depicted as part of this quote's illustration?

A4: The silhouette at the bottom left corner shows a person pushing another person shoulder to shoulder. The physical activity being depicted at the bottom right corner is back-to-back teamwork where one person helps another climb upwards using their strength together. This illustrates the concept mentioned in the quote about raising or lowering standards by surrounding ourselves with individuals who can help us achieve higher goals.

Instruction Tag 5: Cause Perception / Image Scene / Identify overall theme

Q5: What message can be inferred from the silhouette figures in front of the mountain?

A5: The message inferred from the silhouette figures in front of the mountain is about striving towards personal growth, overcoming obstacles (or becoming versions) of oneself through effort and determination. This aligns with the quote's theme of surrounding ourselves with positive influences to become better versions of ourselves.

Object / Attribute
Spatial Position
Text Understanding
Knowledge / Reasoning

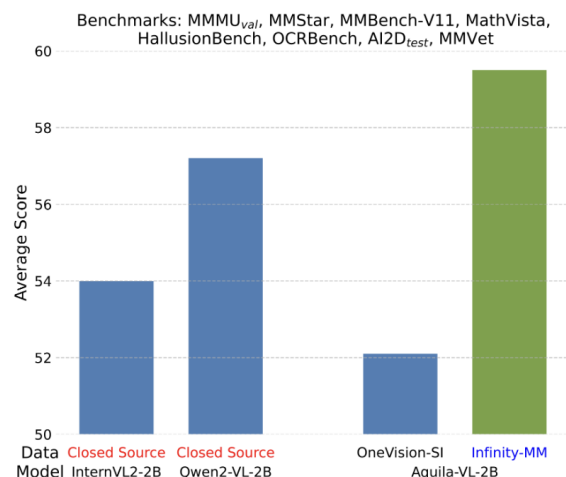


Figure 1. Average score of different VLMs on benchmarks. The Aquila-VL-2B model, trained with Infinity-MM, not only outperforms models trained on other open-source datasets (OneVision-SI) but also surpasses models trained on closed-source datasets.

Model	Params (B)	Average
DeepSeek-VL-1.3B	2.0	39.6
MiniMonkey	2.2	52.7
MiniCPM-V-2	2.8	47.9
PaliGemma-3B-mix-448	2.9	46.5
Phi-3-Vision	4.2	53.6
InternVL2-2B	2.1	53.9
H2OVL-Mississippi-2B	2.1	54.4
XinYuan-VL-2B	2.1	56.1
Qwen2-VL-2B	2.1	57.2
Aquila-VL-2B	2.1	59.5



Chinese Corpora Internet (CCI) 4.0 Released

5TB Chinese Webpage

Leading in scale,
wide-ranging sources

Fuse 3 Qualifier

Precise annotated with
multiple quality scorers

4.2TB CoT data

Large-scale synthesis using
Chains-of-Thoughts

CCI 4.0-M2-Base

2025.05.06

35TB

English & Chinese

Cover 10 billion pages
Multiple quality classifiers

CCI 4.0-M2-CoT

2025.05.06

4200GB

CoT Synthesis

Diverse domain coverage

Huggingface: <https://huggingface.co/datasets/BAAI/CCI4.0-M2-v1>

Datahub: <https://data.baai.ac.cn/details/CCI4.0-M2-v1>

Importance of children's and elderly speech data: large population coverage and high usage demand

Children and the elderly account for over 30% of the population



Age Group Population Distribution in China, 2023 (Source: National Bureau of Statistics)

Significant differences from adult data and numerous challenges

- Children:
- Pronunciation errors
 - Excessive speaking speed
 - Liaison (running words together)
 - Omission of sounds
 - Gender confusion in speech
- Elderly:
- Degeneration of speech organs
 - Speech disorders caused by illness
 - Use of dialects

Current speech and language models show 10%~20% higher recognition error for children's and elderly speech compared to adult speech

模型	在各个数据集上的WER				
	LIBRI_CLEAN	LIBRI_OTHER	CV-en	SO-child	CV-elder
wen-audio	2.32	5.17	24.35	29.20	22.73
wen2-audio	2.25	4.63	9.46	18.83	5.35
LSP	19.94	25.22	36.29	49.20	24.71
speech-gpt	38.29	53.65	135.31	248.10	109.83
almonn	2.26	5.14	55.57	31.82	63.54
vnygpt	15.68	32.81	55.14	72.10	31.18
TU	34.14	37.97	36.30	230.86	103.96
audio-flamingo	-	-	-	-	-
tandagpt	-	-	-	-	-
enseVoice	4.07	8.27	16.46	23.62	11.11
VavLLM	2.21	5.06	15.74	26.89	8.84
QTA	1.34	2.48	5.8	25.75	11.29

WER of children

WER of elderly

Current open-source datasets are scarce and exhibit low coverage

Children's Speech Datasets

Scarcity of data for children under the age of 5, limited availability of dialogue scenarios and Chinese-language content

Elderly Speech Datasets

lack of conversational Chinese speech data

Corpus	Age range	# Speakers	Duration	Style	Year	Trans.	Audio
Ying Corpus	1.5-3.4	1	32	Interaction	2018	Y	Y
CASS CHILD	1-4	25	631	Spontaneous speech	2012	F	N
SLI-CHNC-U1	7-11	927	23.6	Reading	2021	Y	N
SLI-CHNC-U2	4-11	34	27.2	Conversations	2021	Y	N
SpeechKit	7-12	210	72	Reading	2016	Y	Y
Qin	2-3	297	41.5	Conversations	2024	Y	Y

Corpus	Language	Age range	# Speakers	Duration	Year
Preschooler Corpus (Dietrich et al., 2006)	English	1.5	5	303	2006
Lyons Corpus (Dietrich and Freedman, 2008)	English	1.5	4	183	2008
TRAILS (Kamranfar et al., 2019)	English	1.5 - 2.4	296	40	2019
CV Children's Read and Perceptual Speech Corpus (Hager et al., 2019)	English	1.5 - 2.0	663	-	2019
CHIL-Child Speech Corpus (Shafait et al., 2007)	English	1.5-4.0	1,100	-	2007
CV Story Corpus (Hager et al., 2019)	English	1.5-2.0	108	40	2019
MyST Corpus (Friedman et al., 2024)	English	1.5-2.0	1,371	303	2024
PE-SSTAR Children's Speech Corpus (Scherer et al., 2005)	English	4-6.6	159	14.5	2005
The CMU Kid Corpus (Scherer et al., 1997)	English	6-11	76	-	1997
TEENOTES (Lewand and Doolington, 1993)	English	6-15	101	-	1993
CEU children's speech corpus (Lee et al., 1999)	English	5-16	436	-	1999
Speeches202 (Zhang et al., 2011)	English	5-16	129	9	2011
New-Yorker children's speech corpus (Wachs and Berman, 2022)	English	7-12	28	5.5	2022
British English Corpus (Dietrich, 1992)	English	2-4	89	90	1992
CHILDE (Garnica and Mennen, Stanford, 2006)	English	2-6	89	-	2006
BISC Child (Pinto-Ferreira et al., 2021)	English	5-11	174	-	2021
DAEMON-CEU Corpus (Cachia et al., 2008)	French	7-16	-	-	2008
SANNOES (Korpi et al., 2024)	Finnish	6-12	47	-	2024
CPSC (Petersen and Gurevitz, 2012)	Finnish	6-11	52	-	2012
Swedish NICE Corpus (Boll et al., 2015)	Swedish	8-15	5,960	-	2015

Source: FlagEval语音语言模型评测榜单 2024



SeniorTalk: The world's first Chinese conversational speech dataset focused on the elderly population



All samples collected for the dataset have been processed for privacy protection and anonymization, with each speaker providing a signed authorization form

 <https://github.com/flageval-baai/ChildMandarin>  <https://huggingface.co/datasets/BAAI/ChildMandarin>

Figure 1. Gender and Age Distribution of the Dataset

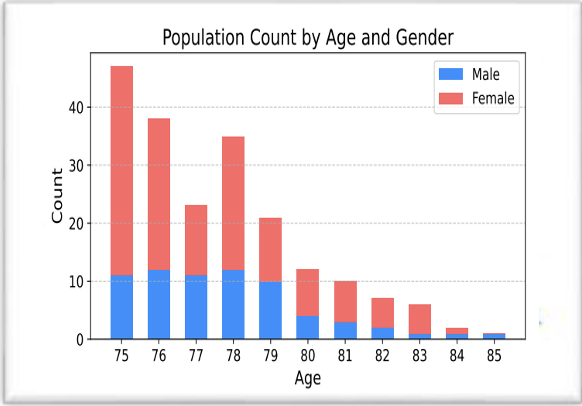


Figure 2. Population Distribution Across Provinces in the Dataset

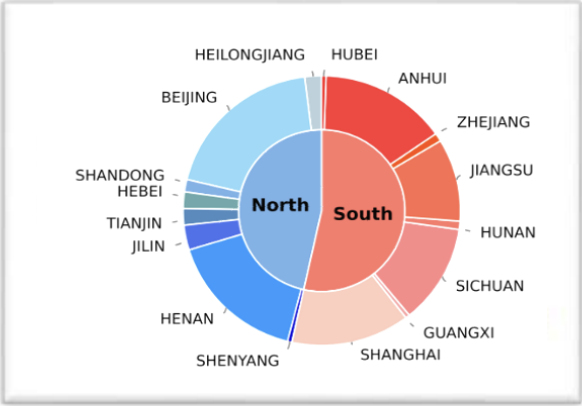


Figure 3. Duration Distribution of Dataset Samples

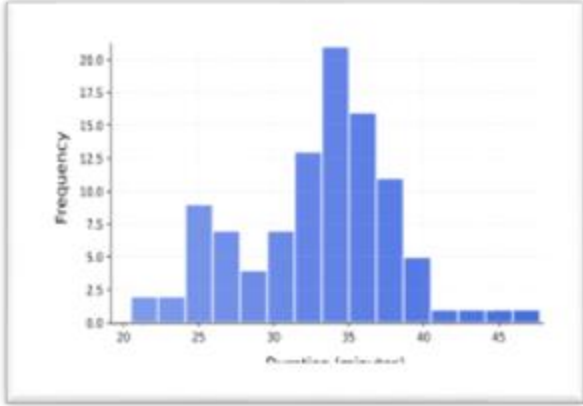
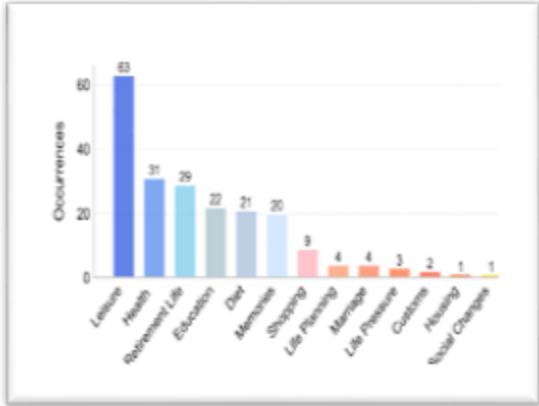


Figure 4. Topic Distribution of Dataset Samples



- Multi-level annotation supports the development of various urgently needed elderly-friendly speech analysis technologies, including speech recognition, speaker verification, and speech event detection

Annotation Level	Annotation Dimension	Associated Tasks	Representative Instances
Speaker Metadata	Demographic Age	Elderly Speech Analysis	75
	Geographic Origin (Province)		Jiangsu, Henan
	ID Card Gender		Female/Male
Session	Temporal Segmentation	Speaker Diarization	[48.475 - 73.582] spk_001
	Overlapping Speech	Speech Separation	trans1(trans2)[+]
Utterance	Raw Transcription	Speech Recognition	[Mandarin utterance]
	Accent Intensity (0-3)	Ordinal Classification	Neutral (0) / Strong (3)
Token	Special Markers	Paralinguistic Analysis	[MUSIC], [NOISE], [LAUGHTER]

喜欢重复一些词语，方言 红烧 读成 红绍

Ground truth: 对 有 一 些 菜 反 正 我 喜 欢 吃 红 烧 菜 炒 菜 再 你 像 红 烧 红 烧 肉
Zero-shot: 一 有 一 些 菜 反 正 我 喜 欢 吃 红 绍 菜 超 菜 一 你 想 红 绍 热 WER: 37.50 % N=24 C=15 S=5 D=4 I=0
Fine-tuning: 对 有 一 些 菜 反 正 我 喜 欢 吃 红 烧 菜 炒 菜 这 你 像 红 烧 红 烧 肉 WER: 4.17 % N=24 C=23 S=1 D=0 I=0

语言组织困难，结巴

Ground truth: 嗯 多 去 外 边 游 游 游 游 乐 一 下
Zero-shot: 都 去 外 边 有 了 一 下 WER: 61.54 % N=13 C=5 S=3 D=5 I=0
Fine-tuning: 嗯 多 去 外 边 游 游 游 游 了 一 下 WER: 7.69 % N=13 C=12 S=1 D=0 I=0

方言口语 散步 觉了（感觉是）

Ground truth: 我 要 是 散 步 觉 了
Zero-shot: 我 要 是 三 不 就 来 WER: 62.50 % N=8 C=3 S=4 D=1 I=0
Fine-tuning: 我 要 是 散 步 觉 得 WER: 12.50 % N=8 C=7 S=1 D=0 I=0

ChildMandarin: Chinese Conversational Speech Dataset for Young Children



<https://github.com/flageval-baai/ChildMandarin>



<https://huggingface.co/datasets/BAAI/ChildMandarin>

Fig 1. Gender and Age Distribution of the Dataset

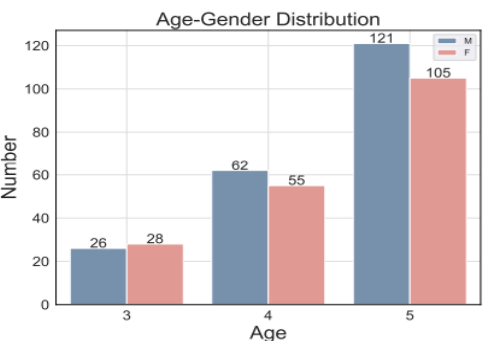


Figure 1: Distribution of speakers by age and gender in our dataset

Fig 2. Topic Distribution of Dataset Samples

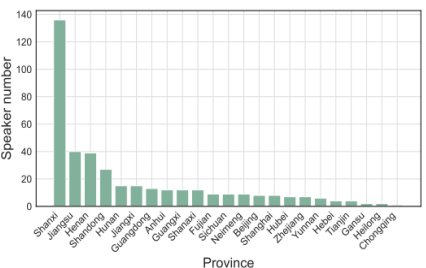
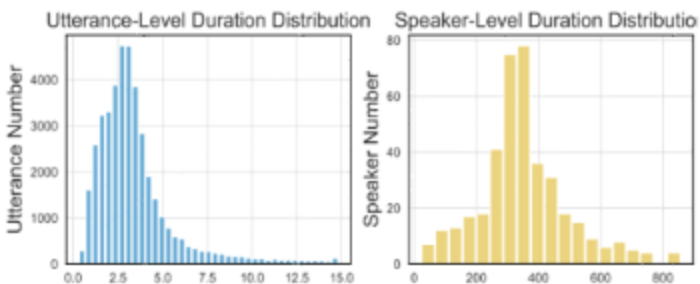


Figure 3: Geographic distribution of speakers in our dataset



Parent-Guided Dialogues:

- Natural pronunciation with rich child-specific speech phenomena

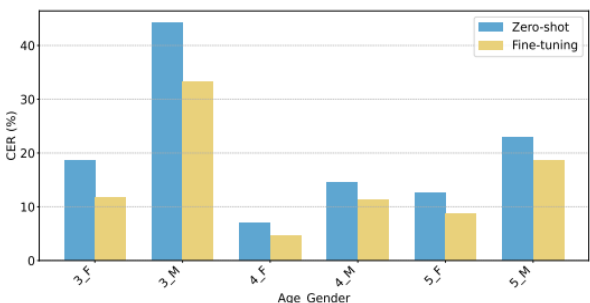
Compliance:

- Data is anonymized and supervised by parents
- Authorization forms are signed by guardians

Effectiveness and Applicability of the Dataset for Speech Recognition and Speaker Verification Tasks

Speech Recognition Performance Analysis (by Age and Gender)

Fig 3. Fine-tuning with the Dataset Significantly Reduces Error Rates Across Multiple Age Groups



Significant Domain Adaptation Effects (Before and After Adaptation)

Utterance: 184_5_F_L_NANJING_IPHONE_001_012	
Ground truth: 兔子会穿上很大的毛衣	
Zero-shot: 兔子会穿上很大的毛衣	CER: 9.09 % N=11 C=10 S=1 D=1 I=0
Fine-tuning: 兔子会穿上很大的毛衣	CER: 0.00 % N=11 C=11 S=0 D=0 I=0
Utterance: 080_5_F_L_CHENGDU_Android_005	
Ground truth: 我跟同学一起玩橡皮泥	
Zero-shot: 我跟同学一起玩橡皮泥	CER: 10.00 % N=10 C=9 S=1 D=0 I=0
Fine-tuning: 我跟同学一起玩橡皮泥	CER: 0.00 % N=10 C=10 S=0 D=0 I=0
Utterance: 320_5_F_L_SHANXI_iPhone12_001_013	
Ground truth: 因为怪兽都是男生没有女生	
Zero-shot: 一位怪兽都是男生没有女神	CER: 25.00 % N=12 C=9 S=3 D=0 I=0
Fine-tuning: 因为怪兽都是男生没有女生	CER: 0.00 % N=12 C=12 S=0 D=0 I=0
Utterance: 403_5_M_L_ZHENGZHOU_Android_005	
Ground truth: 我要唱好多	
Zero-shot: 我要差好多	CER: 20.00 % N=5 C=4 S=1 D=0 I=0
Fine-tuning: 我要唱好多	CER: 0.00 % N=5 C=5 S=0 D=0 I=0
Utterance: 235_3_M_M_CHIFENG_opporeno3pro_001_108	
Ground truth: 周末日当午汗滴禾下土谁知盘中餐粒粒皆辛苦	
Zero-shot: 周末日到五大低和下土直他哪他秘密接心	CER: 75.00 % N=20 C=5 S=13 D=2 I=0
Fine-tuning: 周末日当午汗滴禾下土谁知盘中餐粒粒皆辛苦	CER: 10.00 % N=20 C=18 S=0 D=2 I=0

- ① AI open-source for Good
- ② Our open-source activities for Data
- ③ Our open-source activities for Compute



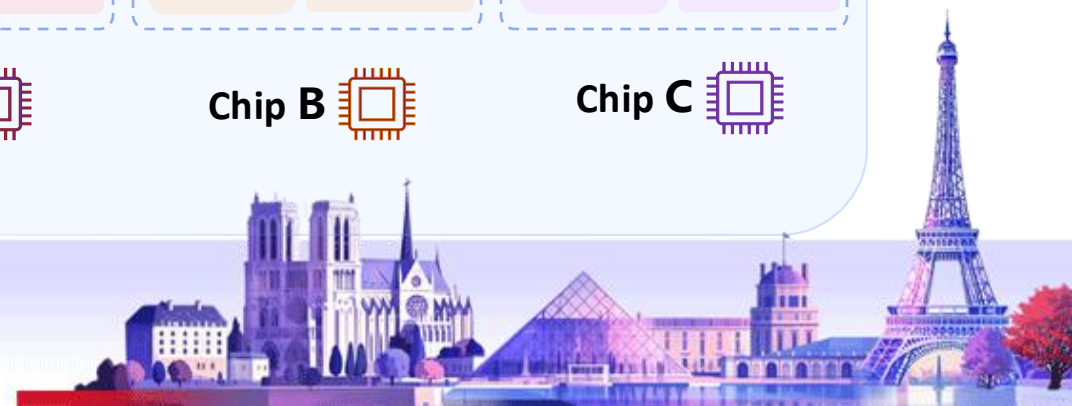
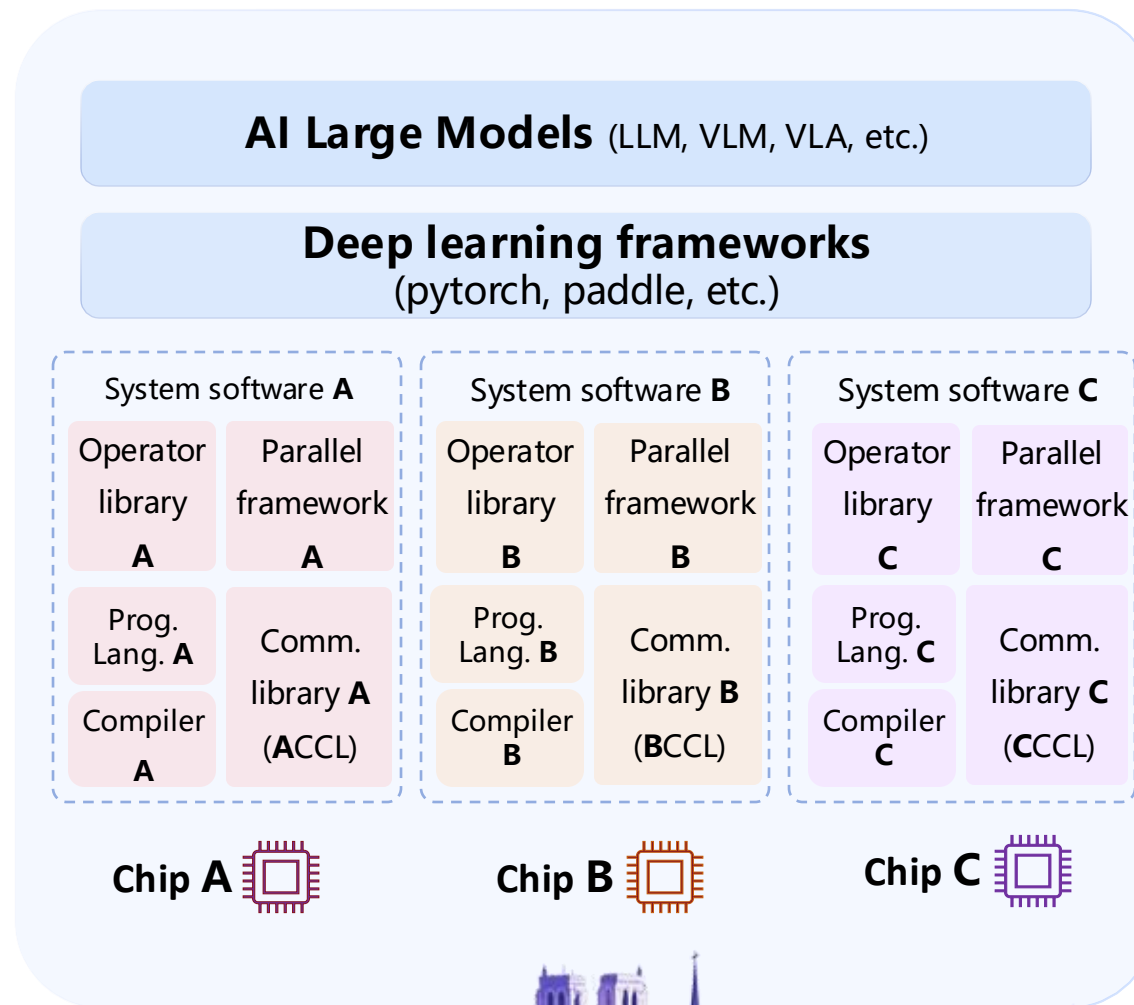
Fragmented software ecosystems hinder the large-scale industrial deployment of AI chips across different architectures

Proprietary and fragmented software ecosystems from different vendors have become the biggest barrier to users adopting alternative AI chips.

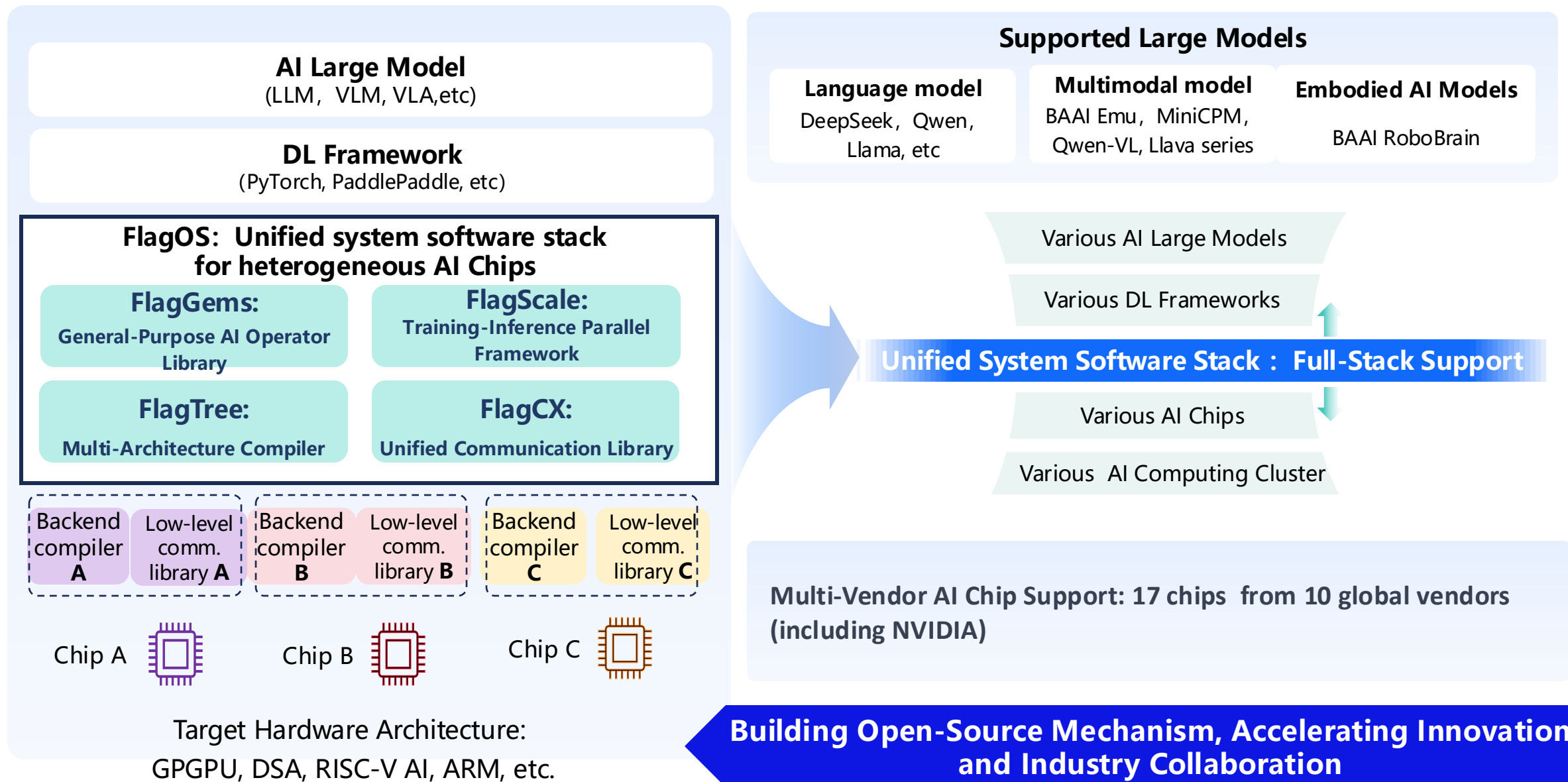
- **Limited by each chip company's capabilities and investments:**

Lack of comprehensive library support, and difficulty keeping pace with the rapid evolution of AI models and optimization technologies.

- **Lack of compatibility between AI-chip vendors:**
Migrating to a different AI chip requires high R&D costs.



FlagOS: Open-sourced & unified system software stack for various AI chips



FlagOS: Provide better performance on various HW

FlagGems: Triton-based largest high-performance operator library

<https://github.com/FlagOpen/FlagGems>

- ✓ 170+ optimized operators for AI large models — the largest in the world
- ✓ Better performance than CUDA in average.
- ✓ Supported 17 chips from 10 vendors.

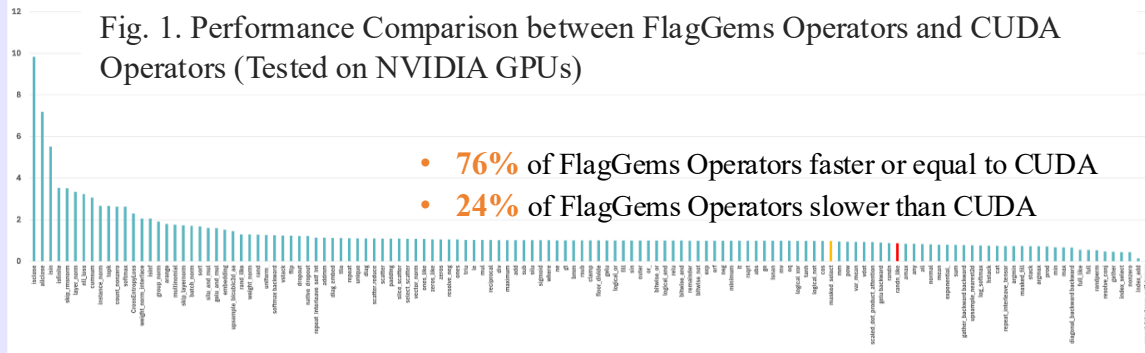
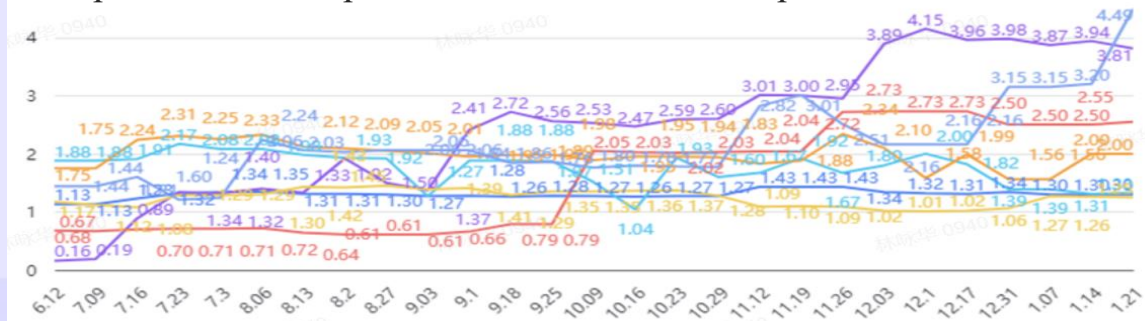


Fig. 2. The FlagGems operator library achieves an average 2x speedup compared to native operators on 6 different AI chips.



添加FlagGems的替换使用例子

In most cases, FlagGems provides better operator performance on Nvidia and other AI chips with same code set.



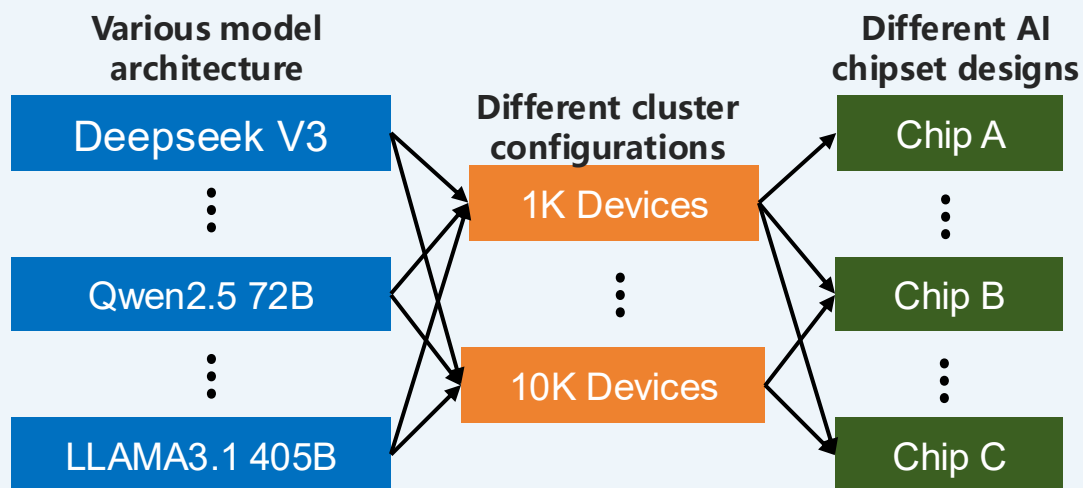
FlagOS: **Automatically** provide better performance on various HW

The parallel framework **FlagScale** provides automatic parallel optimization. It **automatically generates optimization strategies** for any given model, computation cluster, and AI chipset — allowing users to achieve best model training and inference efficiency very easily.

<https://github.com/flagopen/flagscale>

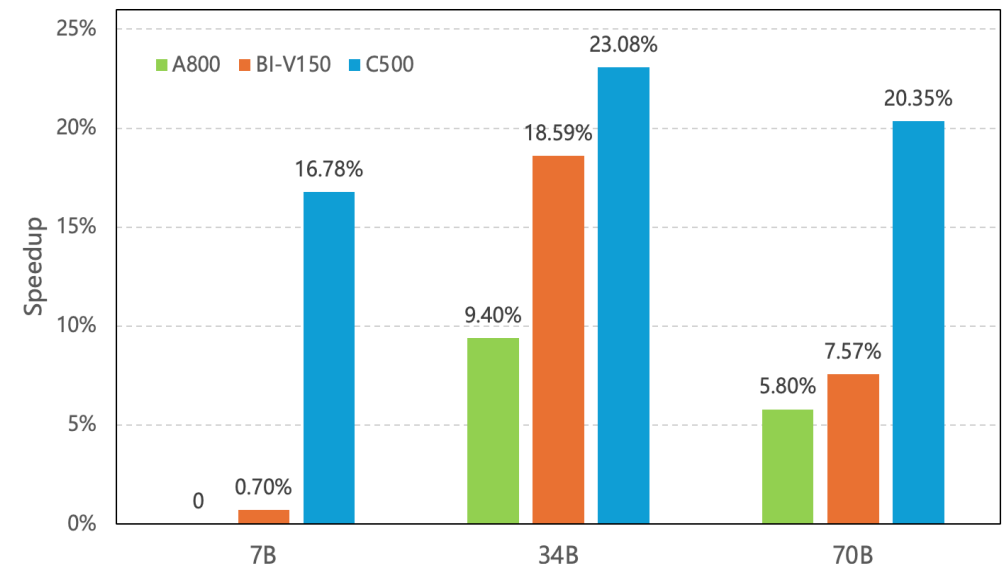
- To address the **multi-dimensional optimization challenges** posed by varying model architectures, cluster configurations, and AI chip designs, FlagScale employs automated optimization to replace reliance on human experts.

N × M × K multi-dimensional challenge:



- Compared to expert (manual) optimizations, FlagScale achieves up to a **23.08%** speedup for training, with an average acceleration of **11.3%**.

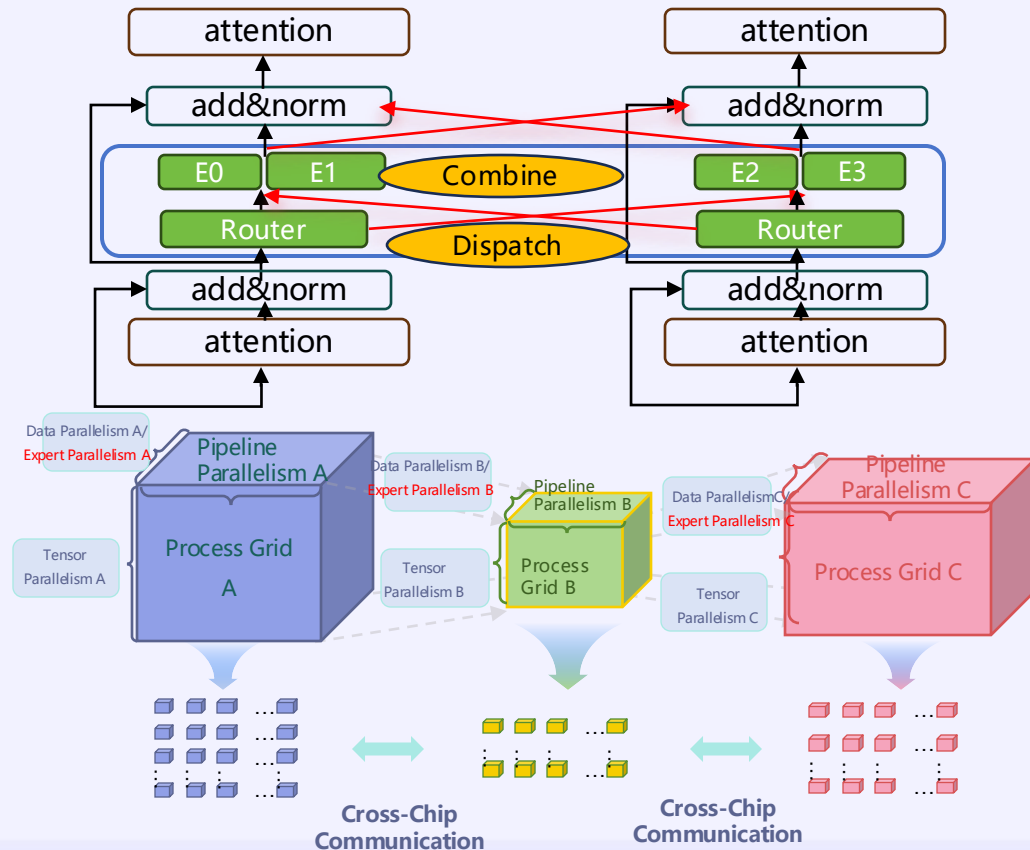
Fig. 1. Comparison of automatic parallel-optimization results versus human-expert optimization for training tasks of various models on three different AI clusters.



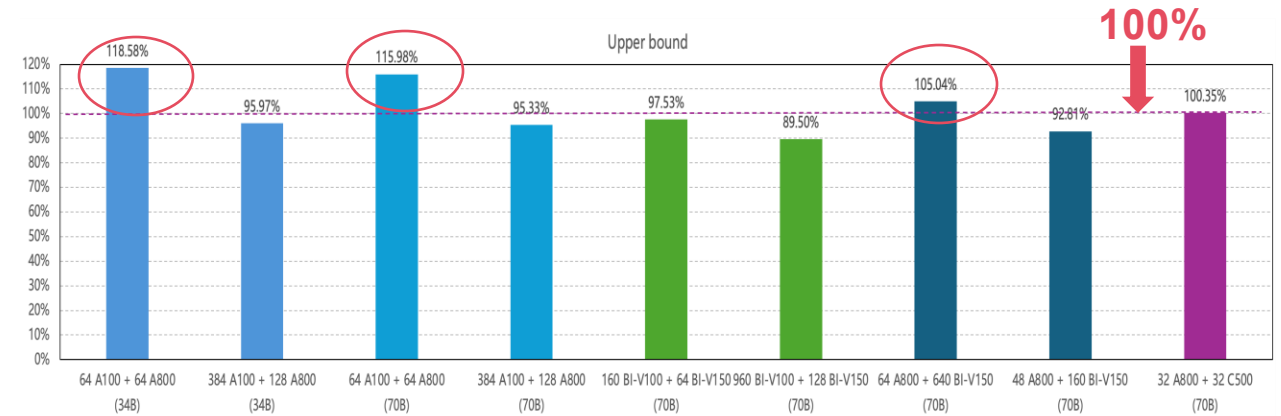
FlagOS: Easily satisfy various computation scenarios — lowering the barriers of computation resource restriction

Don't have enough chips? — FlagScale enables single-model training on cluster composed of mixed chipsets

Supports heterogeneous mixed training for DeepSeek-V3



• **System efficiency of single-model training on mixed HW:** FlagScale achieves high system efficiency in various mixed configuration. By employing combined optimization strategies, FlagScale even can attain over 100% efficiency ("1+1>2") — for example, enabling larger micro-batches in a larger cluster and avoiding unnecessary recomputations.



• **Consistent training performance:** Across varying model sizes and heterogeneous configurations, heterogeneous mixed-training delivers model performance equivalent to that of homogeneous training.

Model	Diff Average	CEVAL (5shot)	MMLU (5shot)	PIQA (0shot)	Hellaswag (10shot)	ARC-C (25shot)	BOOLQ (0shot)
Aquila-34B (A800+A100, Hetero-DP)	-0.74%	-0.25%	-1.25%	0.84%	-0.31%	-3.76%	0.32%
Aquila-34B (A800+A100, Hetero-PP)	-0.66%	0.10%	-0.11%	-0.77%	0.13%	-2.92%	-0.40%
Aquila-70B (A800+A100, Hetero-PP)	0.44%	-0.17%	0.67%	0.22%	-0.06%	1.90%	0.07%
Aquila-70B (BI-V100+BI-V150, Hetero-PP)	-0.62%	-1.81%	0.28%	-0.44%	0.19%	-1.26%	-0.65%
Aquila-70B (BI-V150+NVIDIA, Hetero-PP)	-0.04%	-3.02%	-0.81%	0.14%	0.37%	4.21%	-1.16%
Aquila-70B (C500+NVIDIA, Hetero-PP)	-0.13%	-2.91%	0.22%	0.88%	0.05%	0.60%	0.39%

FlagOS: satisfy various computation scenarios — lowering the barriers of computation resource restriction

New hardware innovation without SW support?

Rapidly enable groundbreaking innovations across diverse computing paradigms

RISC-V ISA + Reconfigurable Chip Architecture

Built on FlagOS, the HW with new instructions and novel architecture can support 50+ AI model operators in very short time. It enabled the new servers benchmark-ready in record time.



New NPU: New IP Design

Built on FlagOS, it enables rapid performance validation of target applications using large model operators during hardware design, achieving hardware-software co-innovation.

From cloud to edge migration without R&D cost

FlagOS-Edge: Edge-Side Execution of Embodied Foundation Models (ILUVATAR + BAAI)

Embodied Foundation Models require On-Robot execution with powerful AI Chips. This is a new frontier. Powered by FlagOS, BAAI's newly released RoboBrian achieves rapid deployment on ILUVATAR's edge GPU modules within days.



Based on FlagOS, user can deploy target version of large models on new hardware in 5 steps <https://huggingface.co/FlagRelease/>

FlagRelease: Released open-sourced models, codes, and docker images for different HW. Supported Qwen, DeepSeek, miniCPM, Qwen-VL, etc.

Users don't need migration. We prepared everything for you !!

To deploy a new model on target chipset HW easily
— 5 steps:

1. Get docker image and model:

```
flagscale pull --image <image> --ckpt <ckpt_path>
```

2. Run docker image:

```
docker run -it -v <ckpt_path>:<ckpt_path> <image>  
/bin/bash
```

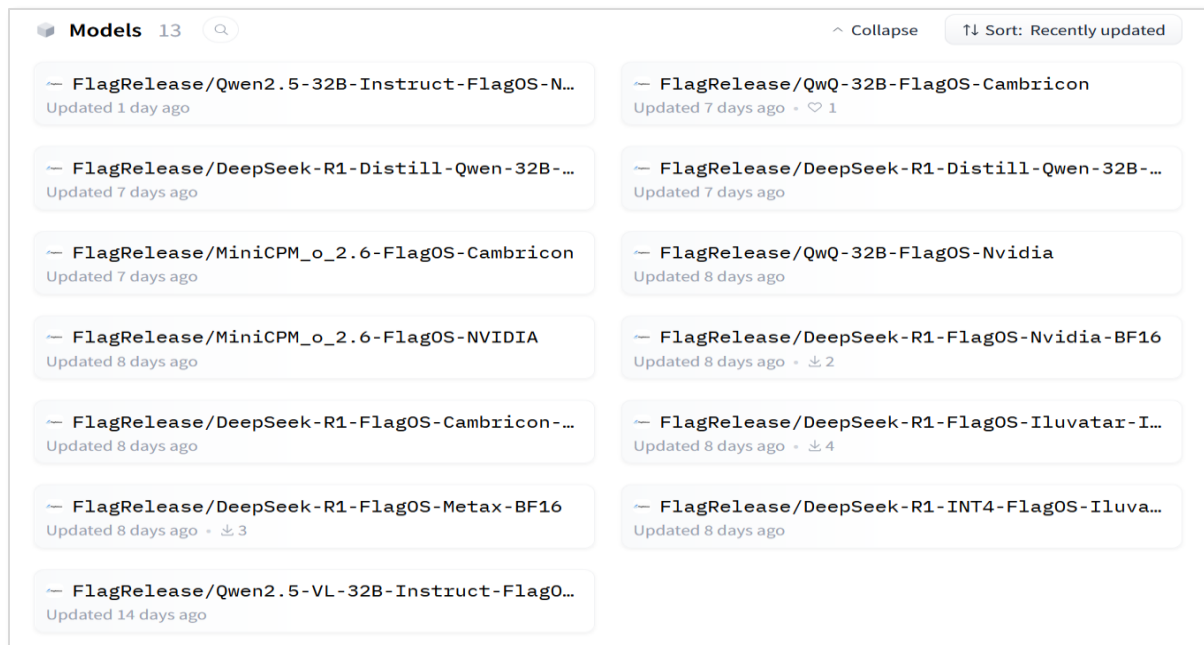
3. Install FlagGems:

```
pip install <FlagGems>/. --no-deps
```

4. Install FlagScale with target chipset patch:

```
python <FlagScale>/tools/patch/unpatch.py --device-type  
<device> --commit-id <commit>; pip install <FlagScale>/.  
5. Model deployment:
```

```
flagscale serve deepseek_r1
```



All the migrated released models have provided readme with benchmark reports to ensure **performance consistence** with original model running on CUDA.

Metrics	DeepSeek-R1-H100-CUDA	DeepSeek-R1-FlagOS-Metax-BF16
GSM8K (EM)	95.75	95.38
MMLU (Acc.)	85.34	85.38
CEVAL	89.00	89.23
AIME 2024 (Pass@1)	76.67	76.67
GPQA-Diamond (Pass@1)	70.20	71.72
MATH-500 (Pass@1)	93.20	93.80



Look for open-source collaboration

Welcome to chat with us at BAAI booth



- FlagScale

- **Large model developer:** Use FlagScale to conduct data experiments and model architecture innovations, or perform large-scale pre-training (even across different chips)
- **Large model service provider:** Automatic deployment of large models for different chips by FlagScale and develop AI applications on top of them
- **Large model developer or user:** Optimize the performance of the training or inference implementation of FlagScale
- **Chip vendors:** Enable training and inference capabilities for new chips through the adaptation of FlagScale

- FlagCX

- **System software innovator:** Optimize existing communication operators or add new ones to meet emerging distributed training and inference requirements
- **Chip vendors:** Adapt to new chips and enable their interconnection with other chips
- **System software innovator:** Explore new communication programming paradigms to achieve cross-chip universal adaptive communication technology
- **System software innovator:** Jointly develop communication library standards to achieve better interoperability



THANK YOU

