GOSIM AI Paris 2025

NEU, Harvard, Cornell,
Tulane, Futurewei

7B Fully Open Source Moxin-LLM – From Pretraining to GRPO-based Reinforcement Learning Enhancement

May 2025

FUTUREWEI Technologies

Contents

Motivation

- LLMs with superior popularity and capabilities
  - ChatGPT, GPT-4o, OpenAI o1, LLaMA, Mistral ......
- Concerns on the transparency, reproducibility and safety in commercialization
  - Lack necessary components
    - Training code or data
    - Hard for full understanding and reproduction
  - Restrictive licenses
    - May limit further innovations

Model	Params	Tokens	Open dataset?
Open weights, closed datasets			
Llama2	7B	2T	✗
DeepSeek	7B	2T	✗
Mistral-0.3	7B	?	✗
QWEN-2	7B	?	✗
Llama3	8B	15T	✗
Gemma	8B	6T	✗
Phi-3	7B	?	✗

Motivation

- Post training quantization of LLMs
  - The 4-bit quantized models do not perform well
    - In many cases, for a 7B quantized model, it does not understand what the question means
  - The reason may be post-training without enough finetuning on high-quality data
- Fine-tuning quantized LLM requires high quality data
  - The training data are not open
  - Finetuning on large amounts of data is expensive

Moxin 7B

- Follow Model openness Framework (MOF)
  - Rate models based on their completeness and openness
  - Follow principles of open science, open source, open data, open access
- Develop Moxin 7B
  - Release training code, data, and model
  - Make continuous commitments to fully open-source LLMs

MOF Classes

MOF Class	Components Included
Class I – Open Science	• Research Paper • Datasets (any license or unlicensed) • Data Preprocessing Code • Model Parameters (intermediate checkpoints) • Model Metadata (optional) • All Class II Components
Class II – Open Tooling	• Training Code • Inference Code • Evaluation Code • Evaluation Data • Supporting Libraries & Tools (optional) • All Class III Components
Class III – Open Model	• Model Architecture • Model Parameters (final checkpoint) • Technical Report • Evaluation Results • Model Card • Data Card • Sample Model Outputs (optional)

Model Architecture

Adopt the mistral architecture

- More blocks than Mistral-7B
    - 36 blocks v.s. 32 blocks
- Parameters are still around 7B

Parameter	Value
n_layers	36
dim	4096
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8

Pre-train Data

Text data + code data

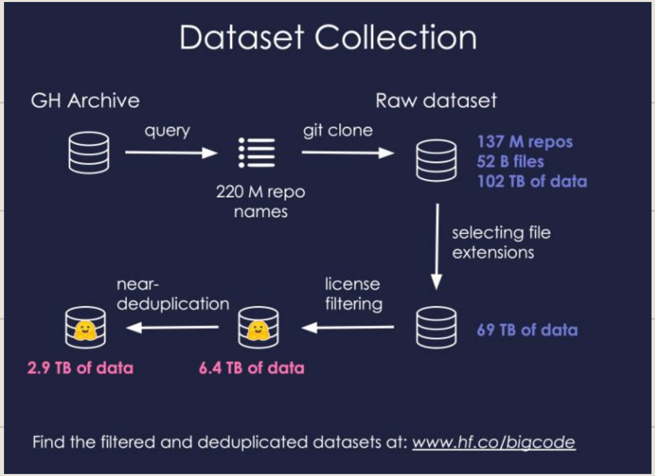
- Text data
  - SlimPajama — a cleaned and extensively deduplicated version of the RedPajama
    - Remove short, low quality documents from RedPajama
    - Prune 49.6% of bytes from RedPajama for deduplication with MinHashLSH
  - DCLM-Baseline
    - Use resiliparse to extract text from CommonCrawl
    - MinHash and near-duplicate Bloom filtering for deduplication
    - Uses fastText OH-2.5 + ELI5 classifier score to filter and keep top 10% of documents

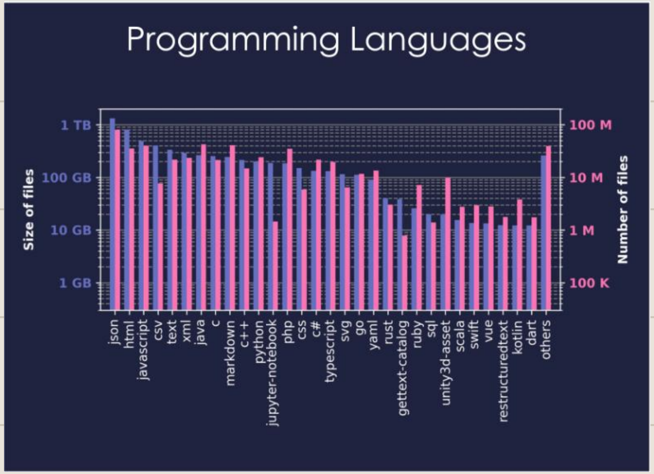
	Tokens	Open source	Curated data sources	Deduplication level
SlimPajama	627B	Yes	Yes	Extensive
RedPajama	1.21T	Yes	Yes	Partial
RefinedWeb-600B	600B	Yes	No	Extensive
RefinedWeb-5T	5T	No	No	Extensive
LLaMA	1.4T	No	Yes	Partial
MPT	1T	No	Yes	Partial
MassiveText	1.4T	No	Yes	Extensive

Pre-train Data

Text data + code data

- Code data
    - The-stack-dedup
        - 137.36M of github repositories were accessible, 51.76B files were downloaded
        - MinHash, Locality Sensitive Hashing, and Jaccard Similarities for deduplication





Pre-train Data

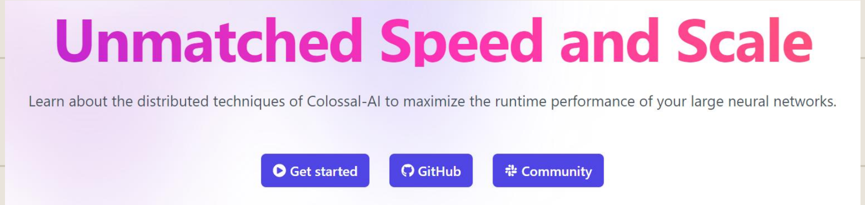
Capability Enhancement data

- Capabilities such as reasoning and knowledge memorizing are essential
- High-quality capability-related data is sparsely distributed in pretraining
- Hard for models to improve these capabilities

- Collect training data of various common evaluation datasets
  - MMLU, ARC, Winogrande, HellaSwag, RACE, OBQA, ...
- Not recommended to train on the training data of evaluation datasets
  - Data contamination
- It is experimental

Pre-train Configuration

- Train on 2T tokens
- Three phases
  - First phase: train with 2K context length
  - Second phase: train with 4K context length
  - Third phase (optional): train with capability enhancement data
- Training framework
  - Colossal-AI
    - A unified deep learning system that provides the fullest set of acceleration technique
  - AdamW optimizer and cosine learning rate decay



Long Context

Techniques for long context

- Sliding window attention (SWA)
  - Do not attend all hidden states
  - Attend to hidden states within a sliding window
  - Handle longer sequences more effectively at a reduced computational cost
- Rolling Buffer Cache
  - Limit the cache size using a rolling buffer cache with a fixed attention span
  - Overwrite past values in the cache
  - Do not increase the cache size

Our model can support 32K context length

- fast inference and low memory cost

Pre-train Evaluation

- Evaluation Tasks
    - AI2 Reasoning
    - HellaSwag
    - MMLU
    - Winogrande
    - PIQA
    - MTbench
- Evaluation Framework
    - Lm-evaluation-harness
        - A unified framework to test LLMs on a large number of different evaluation tasks
        - Support over 60 standard academic benchmarks for LLMs with undreds of subtasks and variants implemented

Task	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM
Astronomy	Solar system, galaxies, asteroids, ...	STEM
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other
College Biology	Cellular structure, molecular biology, ecology, ...	STEM
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM
High School European History	Renaissance, reformation, industrialization, ...	Humanities
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences
High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
High School US History	Civil War, the Great Depression, The Great Society, ...	Humanities
High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
Management	Organizing, communication, organizational structure, ...	Other
Marketing	Segmentation, pricing, market research, ...	Other

Pre-train Evaluation

- Base model
    - Moxin-7B-Original: first two phases without training on capability enhancement data
    - Moxin-7B-Enhanced: all three phases with training on capability enhancement data
- Few-shot evaluation
    - AI2 Reasoning Challenge (25-shot)
    - HellaSwag (10-shot)
    - MMLU (5-shot)
    - Winogrande (5-shot)

Pre-train Evaluation

- Few-shot evaluation
  - Moxin-7B-Original outperforms LLaMA 2-7B
  - Moxin-7B-Enhanced achieves competitive accuracy performance
  - Training on capability enhancement data significantly improves the performance

Table 3: Performance comparison for various models in few-shot evaluation.

Model	ARC-C	Hellaswag	MMLU	WinoGrade	Ave
Mistral - 7B	57.59	83.25	62.42	78.77	70.51
LLaMA 3.1 - 8B	54.61	81.95	65.16	77.35	69.77
LLaMA 3 - 8B	55.46	82.09	65.29	77.82	70.17
LLaMA 2 - 7B	49.74	78.94	45.89	74.27	62.21
Qwen 2 - 7B	57.68	80.76	70.42	77.43	71.57
Gemma - 7B	56.48	82.31	63.02	78.3	70.03
Internlm2.5 - 7B	54.78	79.7	68.17	80.9	70.89
Baichuan2 - 7B	47.87	73.89	54.13	70.8	61.67
Yi-1.5-9B	58.36	80.36	69.54	77.53	71.48
Moxin - 7B - Original	53.75	75.46	59.43	70.32	64.74
Moxin - 7B - Enhanced	59.47	83.08	60.97	78.69	70.55

Pre-train Evaluation

- Base model
  - Moxin-7B-Original: first two phases without training on capability enhancement data
  - Moxin-7B-Enhanced: all three phases with training on capability enhancement data
- Zero-shot evaluation
  - AI2 Reasoning Challenge (0-shot)
  - AI2 Reasoning Easy (0-shot)
  - HellaSwag (0-shot)
  - PIQA (0-shot)
  - Winogrande (0-shot)

Pre-train Evaluation

- Zero-shot evaluation
  - Moxin-7B-Enhanced achieves superior accuracy performance
  - Training on capability enhancement data significantly improves the performance under the zero-shot setting

Table 2: Performance comparison for various models in zero-shot evaluation.

Models	HellaSwag	WinoGrade	PIQA	ARC-E	ARC-C	Ave
Mistral - 7B	80.39	73.4	82.15	78.28	52.22	73.29
LLaMA 2 - 7B	75.99	69.06	79.11	74.54	46.42	69.02
LLaMA 2 - 13B	79.37	72.22	80.52	77.4	49.06	71.71
LLaMA 3.1 - 8B	78.92	74.19	81.12	81.06	53.67	73.79
Gemma - 7b	80.45	73.72	80.9	79.97	54.1	73.83
Qwen v2 - 7B	78.9	72.38	79.98	74.71	50.09	71.21
Internlm2.5 - 7b	79.14	77.9	80.52	76.16	51.37	73.02
Baichuan2 - 7B	72.25	67.17	77.26	72.98	42.15	66.36
Yi-1.5-9B	77.86	73.01	80.74	79.04	55.03	73.14
DeepSeek - 7B	76.13	69.77	79.76	71.04	44.8	68.3
Moxin - 7B - Original	72.06	66.31	78.07	71.47	48.15	67.21
Moxin - 7B - Enhanced	80.03	75.17	82.24	81.12	58.64	75.44

Post-train

- **Version 1: Adopt Open-Source Tülu 3 Dataset and Framework**
  - SFT: Tülu 3 SFT Mixture dataset + Open-instruct framework
    - Moxin-7B-SFT
  - DPO: Tülu 3 8B Preference Mixture dataset + Open-instruct framework
    - Moxin-7B-DPO
- **Version 2: Adopt Open-Source Infinity Instruct Dataset**
  - Infinity Instruct: a large-scale, high-quality instruction dataset, millions of instructions
    - Moxin-7B-DPO-II

Post-train Evaluation

- **Zero-Shot Evaluation**
  - Moxin-7B-DPO model can achieve comparable performance with other SOTA instruct models

Table 4: Performance comparison for various models in zero-shot evaluation.

Models	HellaSwag	WinoGrade	PIQA	ARC-E	ARC-C	Ave
Mistral 8B Instruct	79.08	73.56	82.26	79.88	56.57	74.27
Llama3.1 8B Instruct	79.21	74.19	80.79	79.71	55.03	73.79
Qwen2.5 7B Instruct	80.5	71.03	80.47	81.31	55.12	73.69
Moxin - 7B - II	79.32	72.93	81.56	80.43	56.91	74.23
Moxin - 7B - SFT	81.44	73.09	81.07	79.8	54.67	74.01
Moxin - 7B - DPO	85.7	73.24	81.56	81.1	58.02	75.92

Post-train Evaluation

- **Few-Shot Evaluation**
  - Moxin-7B-DPO performs competitively

Table 5: Performance comparison for various models in few-shot evaluation.

Model	ARC-C	Hellaswag	MMLU	WinoGrade	Ave
Mistral 8B Instruct	62.63	80.61	64.16	79.08	71.62
Llama3.1 8B Instruct	60.32	80	68.18	77.27	71.44
Qwen2.5 7B Instruct	66.72	81.54	71.3	74.59	73.54
Moxin - 7B - II	61.35	82.1	62.95	77.98	71.095
Moxin - 7B - SFT	60.11	83.43	60.56	77.56	70.42
Moxin - 7B - DPO	64.76	87.19	58.36	76.32	71.66

- **OLMES Evaluation**
  - Adopt the OLMES framework from Tulu3 for evaluation

Table 6: Performance comparison for various models in olmes evaluation.

Models/Datasets	GSM8K	MATH	Humaneval	Humaneval plus	MMLU	PopQA	BBH	TruthfulQA	Ave
Qwen2.5 7B Instruct	83.8	14.8	93.1	89.7	76.6	18.1	21.7	63.1	57.61
Gemma2 9B Instruct	79.7	29.8	71.7	67	74.6	28.3	2.5	61.4	51.88
Moxin - 7B - II	71.04	21	78.21	72.35	63.27	27.98	44.33	56.22	54.42
Moxin - 7B - DPO	81.19	36.42	82.86	77.18	60.85	23.85	57.44	55.27	59.38

CoT Enhancement

- **SFT on Reasoning data**
  - SFT on Openthoughts: Feed questions to DeepSeek R1 and collect the reasoning response
  - SFT on OpenR1-Math-220k: Feed questions to DeepSeek R1 and collect the reasoning response
- **RL with GRPO**
  - Version 1: Adopt the DeepScaleR framework
    - Moxin-7B-RL-DeepScaleR
  - Version 2: Adopt the Areal framework
    - Moxin-7B-RL-AReal

CoT Evaluation

- **Math Evaluation**
  - Moxin-7B-RL-DeepScaleR achieves outstanding performance
    - RL is effective for small LLMs such as 7B models
  - Moxin-7B-RL-DeepScaleR performs better than Moxin-7B-RL-AReal

Table 7: Performance comparison for various models on reasoning evaluation.

Models/Datasets	MATH 500	AMC	Minerva Math	OlympiadBench	Ave
Qwen2.5-Math-7B-Base	52.4	52.5	12.9	16.4	33.55
Qwen2.5-Math-7B-Base + 8K MATH SFT	54.6	22.5	32.7	19.6	32.35
Llama-3.1-70B-Instruct	64.6	30.1	35.3	31.9	40.48
Moxin-7B-RL-AReal	68.6	50	16.9	31.7	41.8
Moxin-7B-RL-DeepScaleR	68	57.5	16.9	30.4	43.2

Model Release

- Develop multiple models
- Release multiple models

Table 8: Our developed models and their names in our releases.

Developed Models	Names in Releases
Moxin-7B-Enhanced	Moxin-7B-Base
Moxin-7B-SFT	
Moxin-7B-DPO	Moxin-7B-Instruct
Moxin-7B-DPO-II	
Moxin-7B-RL-DeepScaleR	Moxin-7B-Reasoning
Moxin-7B-RL-AReal	

Model Release

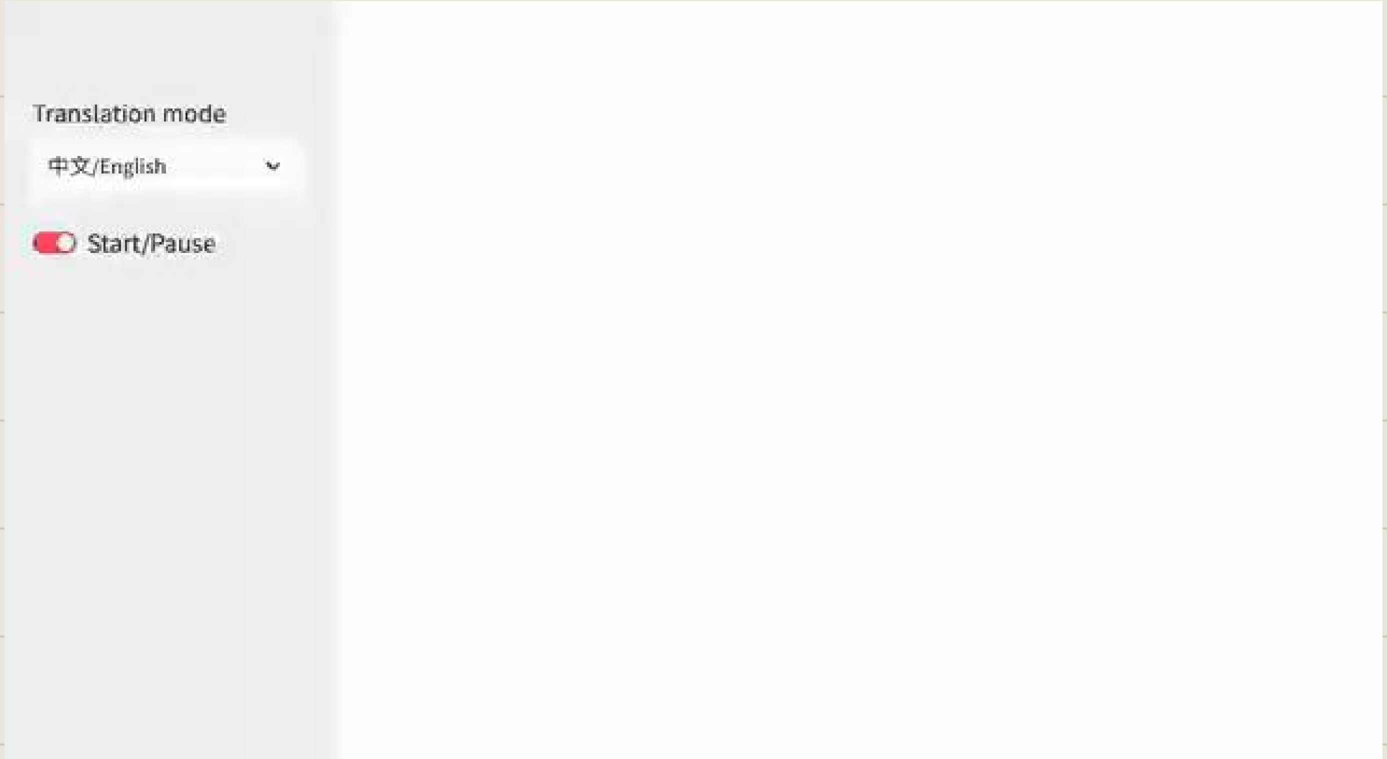
7B Fully Open Source Moxin-LLM – From Pretraining to GRPO-based Reinforcement Learning Enhancement

Pu Zhao[1], **Xuan Shen**[1], **Zhenglun Kong**[2], **Yixin Shen**[3], **Sung-En Chang**[1], **Timothy Rupprecht**[1], **Lei Lu**[1], **Enfu Nan**[1], **Changdi Yang**[1], **Yumei He**[4], **Weiyan Shi**[1], **Xingchen Xu**[5], **Yu Huang**[6], **Wei Jiang**[7], **Wei Wang**[7], **Yue Chen**[7], **Yong He**[7], **Yanzhi Wang**[1,8]

[1]Northeastern University, [2]Harvard University,
[3]Cornell University, [4]Tulane University, [5]University of Washington,
[6]Roboraction.ai, [7]Futurewei Technologies, [8]AIBAO LLC

Homepage with all codes: *https://github.com/moxin-org/Moxin-LLM*
Base model: *https://huggingface.co/moxin-org/moxin-llm-7b*
Instruct model: *https://huggingface.co/moxin-org/moxin-Instruct-7b*
Reasoning model: *https://huggingface.co/moxin-org/moxin-Reasoning-7b*

Real-Time Translation and Agent



Large-Scale MoE Model Deployment

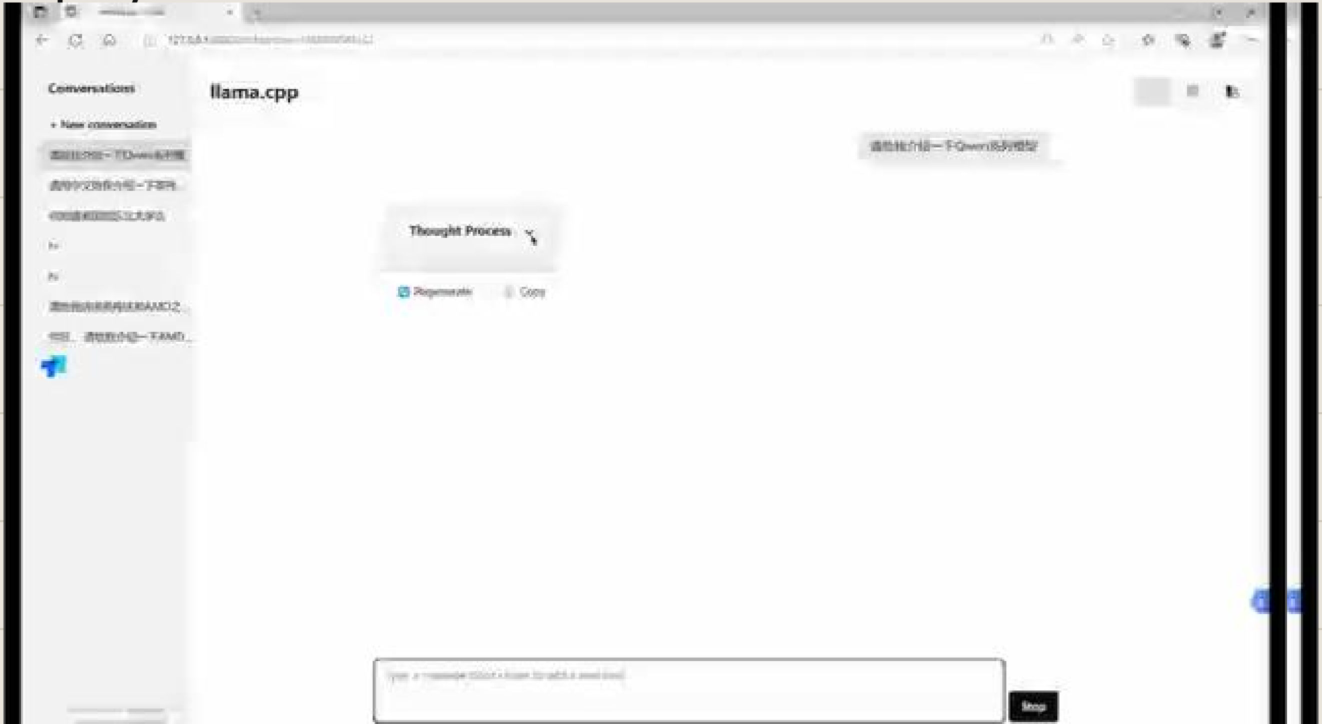
- Results on the recent Qwen 3

Benchmark (Metric)	Qwen3 **Ours** IQ2_S	Qwen3 Original Q8_0
Architecture	MoE	MoE
# Activated/Total Params	22B/235B	22B/235B
# Model Size (GiB)	72.17	232.77
Winogrande	75.9	77.8
MMLU(EM)	84.9	86.8
Hellaswag	83.4	84.7

Benchmark (Metric)	Qwen3 **Ours** IQ2_S	Qwen3 Unsloth Q2_K
Architecture	MoE	MoE
# Activated/Total Params	22B/235B	22B/235B
# Model Size (GiB)	72.17	79.80
Winogrande	75.9	75.2
MMLU(EM)	84.9	84.4
Hellaswag	83.4	83.1

Large-Scale MoE Model Deployment

- Deployment of Qwen3 235B on AMD AI Max+ 395 notebook



Large-Scale MoE Model Deployment

- Results on DeepSeek V3 and R1

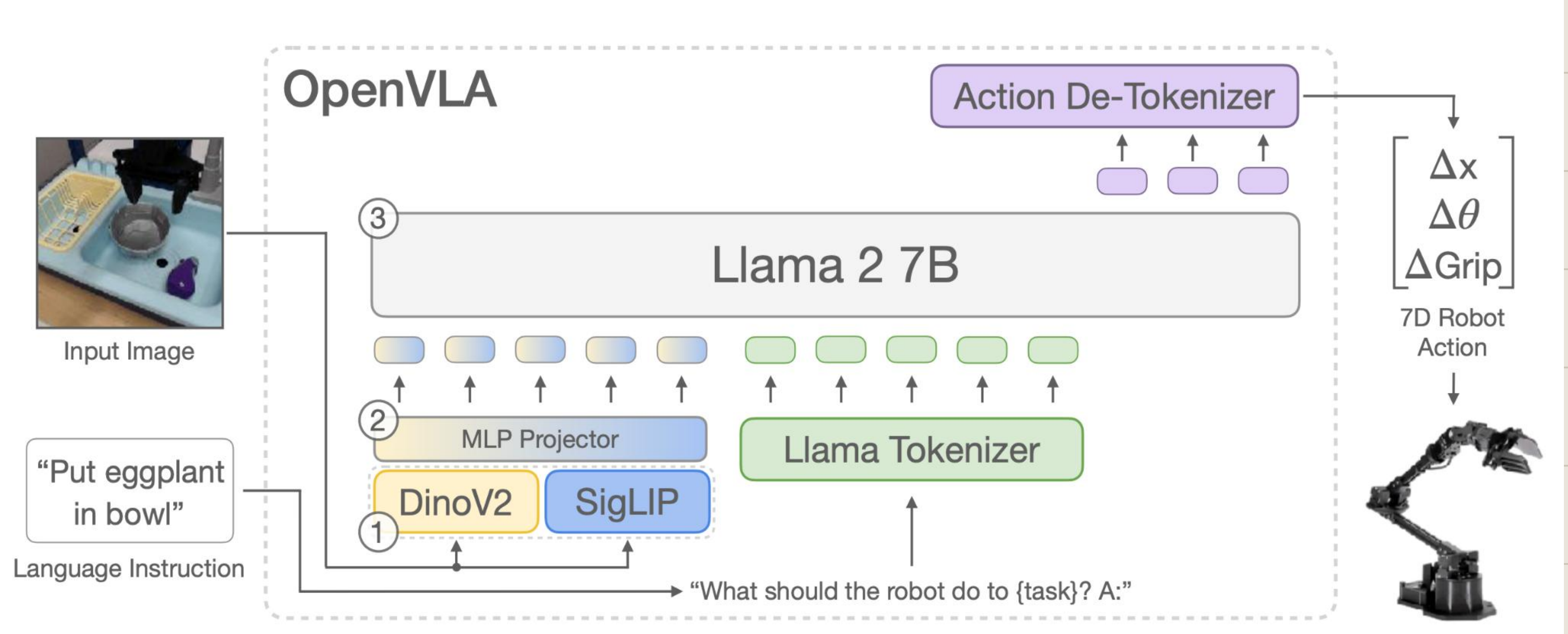
	DeepSeek R1-distill LLaMA 70B	Ours 85GB	Full-Size DeepSeek V3-0324
MMLU	77.86	79.5	84.9
Winogrande	76.6	77.0	76.9
Speed	4.5 tokens/s	8.34 tokens/s	--

Collaboration on the efficient deployment

HUAWEI AMD intel 海飞科 HEXAFLAKE

Embodied AI

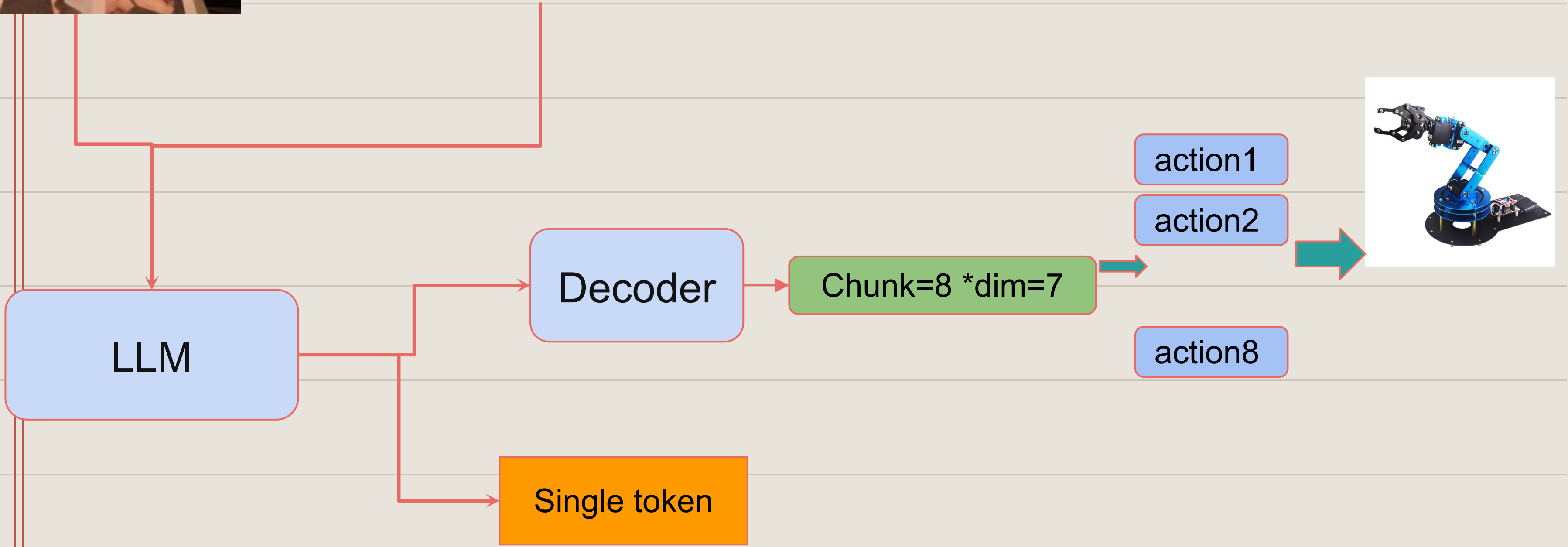
- Previous work OpenVLA



Embodied AI

Our method

In: What action should the robot take to move coke can to taylor swift



Embodied AI

Methodology

-
-

Embodied AI

Advantages

Embodied AI

- Comparison results on accuracy (success rate)

Method	libero_spatial_no_noops(SR)	libero_object_no_noops(SR)	libero_goal_no_noops(SR)	libero_10_no_noops(SR)	Average
openvla	84.7%	88.4%	79.2%	53.7%	76.50%
openvla-oft	96.20%	98.30%	**96.20%**	90.70%	95.35%
DiT Policy (fine-tuned)	84.20%	96.30%	85.40%	63.80%	82.40%
Ours chunk 8	**98.00%**	**99.50%**	96.00%	**94%**	**96.88%**
Ours chunk 16	96.00%	98.5%	94.00%	91.10%	94.90%

Bold is the highest success rate

Embodied AI

Comparison results on speed

Orin Board

Method	Throughput(Hz)	Latency (Sec) ↓
Openvla fp16	1.19	0.84
Openvla int4	**2.88**	0.347
Ours chunk 8,bf 16	**23**	0.347

Chunk 8 means predict 8 actions at one model forward

Chunk 16 means predict 16 actions at one model forward

Dim means one action has 7 dims, for a single-arm robot, its actions consist of seven dimensions for movement a={x, y, z, roll, pitch, yaw

THANK YOU!