THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
RELAXED SYSTEM LAB

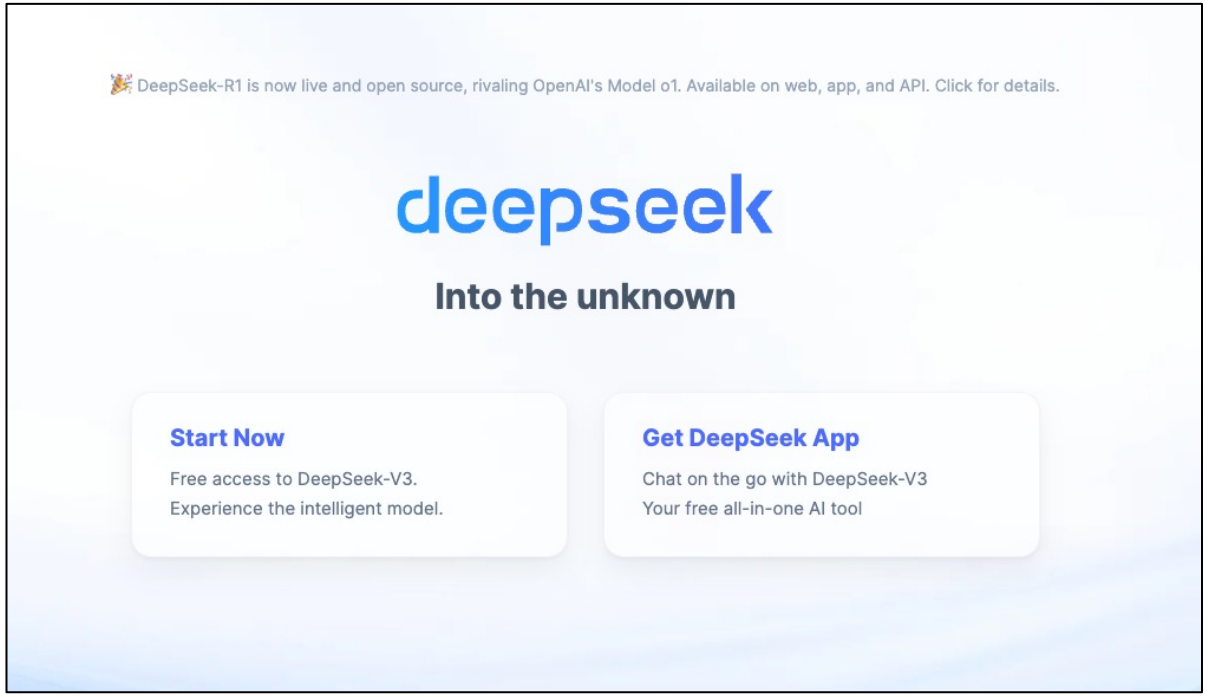
Introduction and Logistic

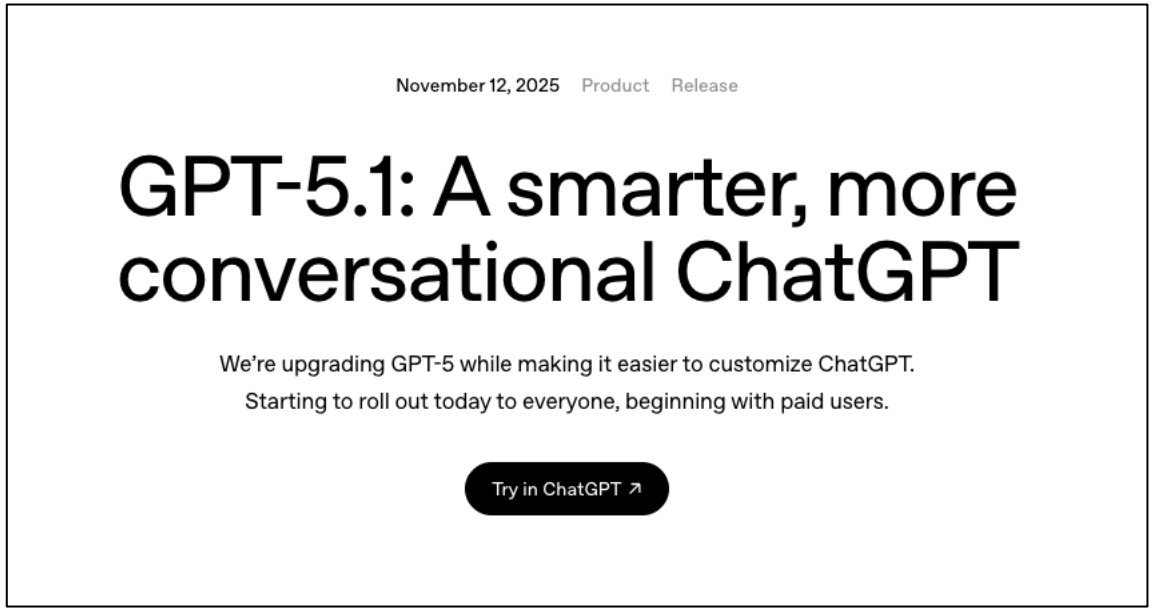
COMP4551

Binhang Yuan

Amazing Progress of AIGC

RELAXED
SYSTEM LAB





$$\min_{x} \mathbb{E}_{\xi} f(\xi, x)$$

RELAXED
SYSTEM LAB

$$\min_{x} \mathbb{E}_{\xi} f(\xi, x)$$

Data

- *(ImageNet) 1.3M Images (est. 160+ GB)*
- *(Llama-3.1) 15 Tillion Tokens (est. 100+ TB)*

Model

- *(GPT-2) 1.3 Billion Parameters (2.6 GB fp16)*
- *(Llama-3.1) 405 Billion Parameters (810GB fp16)*

Compute

- *(GPT-2) est. 2.5 GFLOPS/token*
- *(Llama-3.1) est. 1.2 TFLOPS/token*

The goal of this course:

Unravel the secrets of such foundation models from the both the algorithm and system perspective!

The Path Towards a Foundation Model

RELAXED SYSTEM LAB

Data Preparation → Pre-training → Tuning & Alignment → Evaluation → Deployment

Preparing a massive corpus for pretraining

A multi-stage pipeline:
- Acquisition (e.g. OCR)
- Cleaning;
- Filtering;
- De-duplication;
- Sampling;
- Shuffling.

Training a foundation model by fitting the large corpus by auto-regression

Core infra:
a large-scale parallel training system running over thousands of GPUs

Tune the model so it can follow user instructions effectively and safely.

Various methods:

- SFT
- RLHF
- RLAIF

Evaluate the model from different aspects:

- Comprehension
- Instruction following
- Expression
- Reasoning
- Coding
- Interaction
- Safety

Deploy the model to serve users under different settings:
- Cloud
- Local devices

Require system efforts to optimize the performance, cost, and latency.

Covered by this course!

RELAXED
SYSTEM LAB

Course Overview

Syllabus

RELAXED
SYSTEM LAB

Date	Topic
W1 - 02/03, 02/05	Introduction and Logistics & ML Preliminary
W2 - 02/10, 02/12	Stochastic Gradient Descent & Automatic Differentiation
W3 - 02/17, 02/19	Spring Festival
W4 - 02/24, 02/26	Language Model Architecture & Large Scale Pretrain Overview
W5 - 03/03, 03/05	Nvidia GPU Performance & Collective Communication Library
W6 - 03/10, 03/12	Data-, Pipeline- Parallel Training & Tensor Model-, Optimizer- Parallel Training
W7 - 03/17, 03/19	Sequence-, MoE- parallelism & Mid-Term Review
W8 - 03/24, 03/26	Mid-Term Exam & Generative Inference

Syllabus

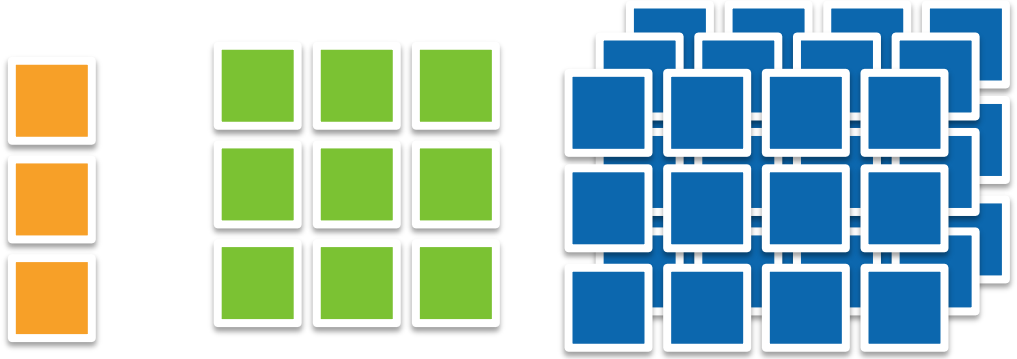
RELAXED
SYSTEM LAB

Date	Topic
W9 - 03/31, 04/02	Inference Alogirhtm Optimizations & Inference System Optimizations
W10 - 04/07, 04/09	Spring Break & Prompt Engineering
W11 - 04/14, 04/16	Inference Scaling & Retrieval Augmented Generation
W12 - 04/21, 04/23	LLM Agent & Parameter Efficient Fine-Tuning
W13 - 04/28, 04/30	RL Alignment & LLM Evaluation
W14 - 05/05, 05/07	Guest Speech (TBD) & Final Review

Machine Learning Preliminary

RELAXED SYSTEM LAB

- Linear Algebra:
  - Vector, matrix, tensor.
- PyTorch Tensors.

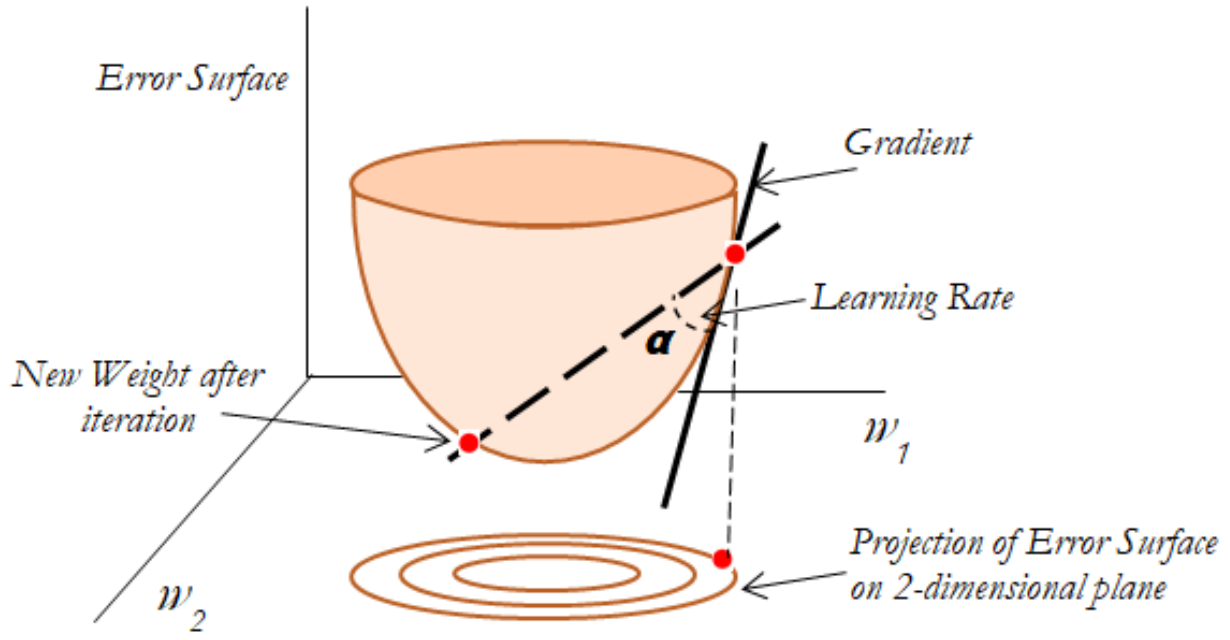


PyTorch

Stochastic Gradient Descent

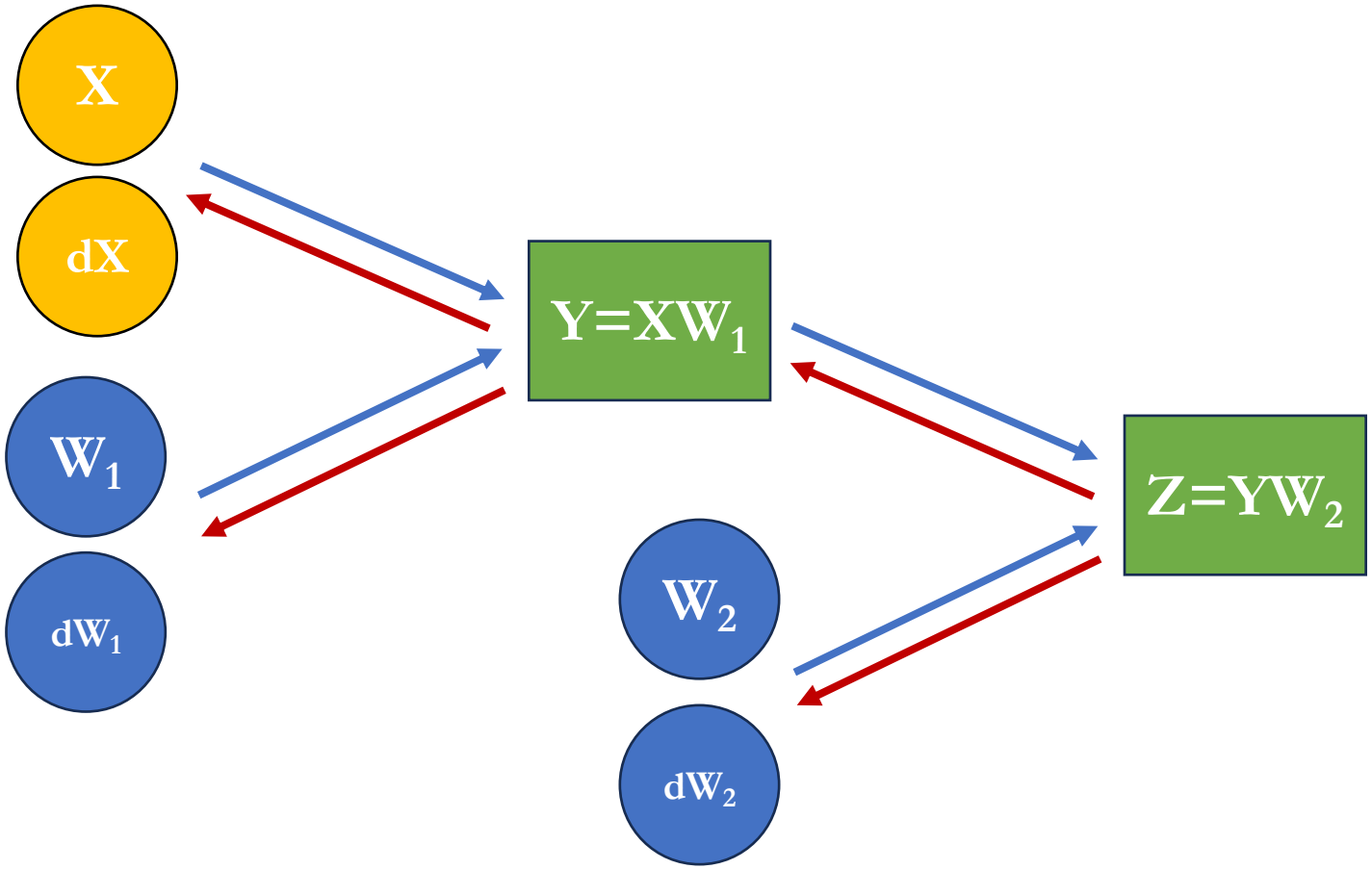
RELAXED SYSTEM LAB

- Then, suppose we have:
  - $f: \mathbb{R}^d \rightarrow \mathbb{R}$;
- Definition of a derivative/gradient :
  - $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \in \mathbb{R}^d$
- Where:
  - $\frac{\partial f}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(x_1, x_2, \ldots, x_i + \epsilon, x_{i+1}, \ldots, x_d) - f(x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_d)}{\epsilon}$
    $= \lim_{\epsilon \to 0} \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$



Auto-Differentiation & PyTorch Autograd

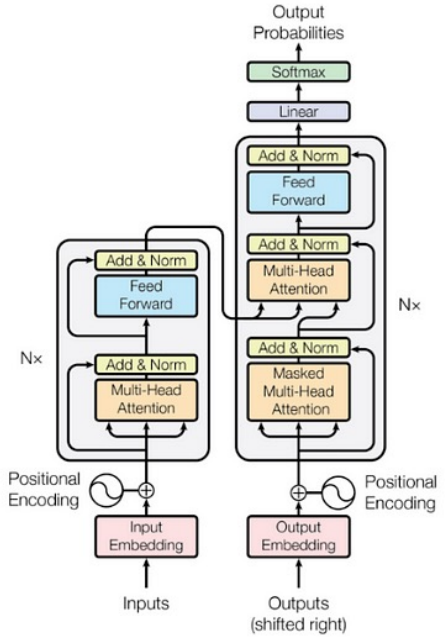
RELAXED SYSTEM LAB



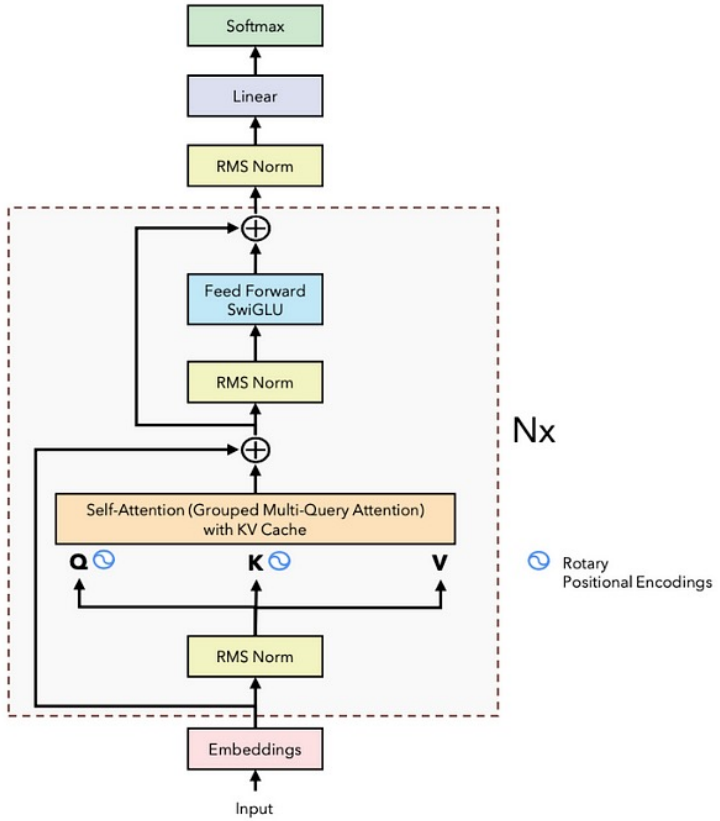
Transformer Architecture

RELAXED SYSTEM LAB

Transformer vs LLaMA



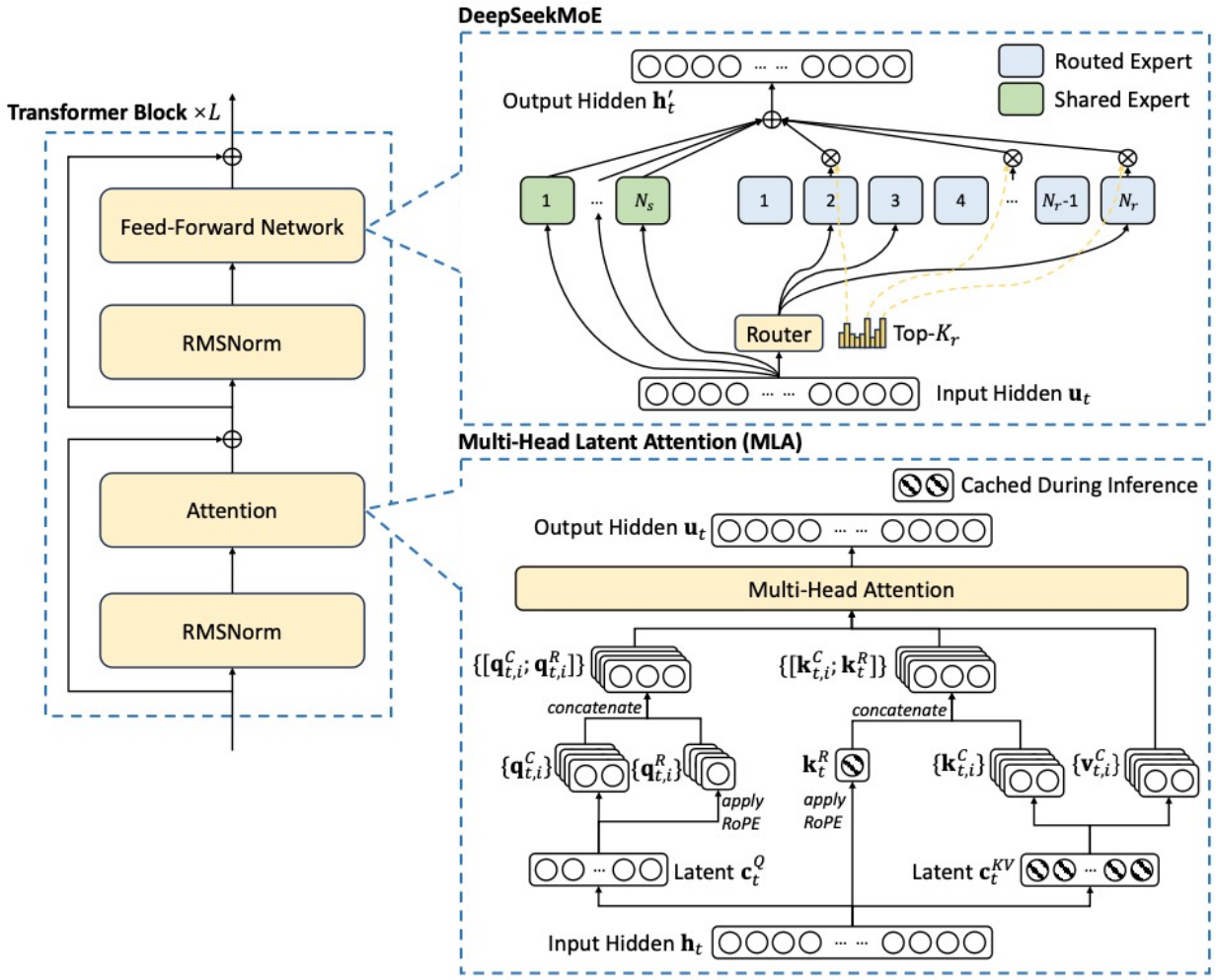
Transformer
("Attention is all you need")



LLaMA

Transformer with MoE

RELAXED SYSTEM LAB

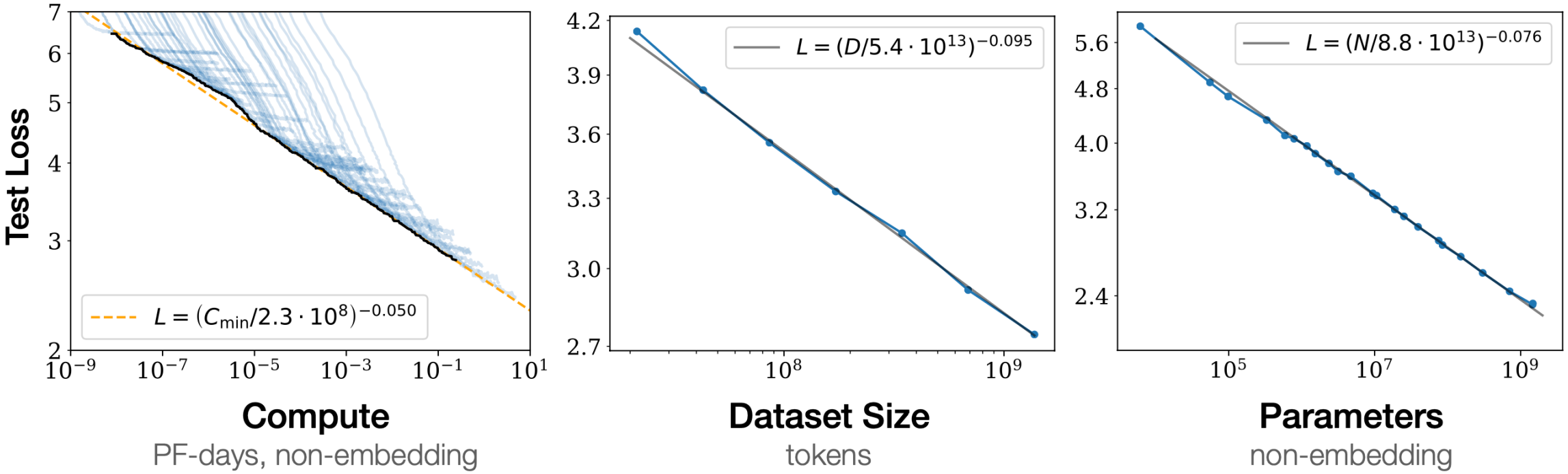


Deepseek V3

Large Scale Pretrain Overview

RELAXED SYSTEM LAB

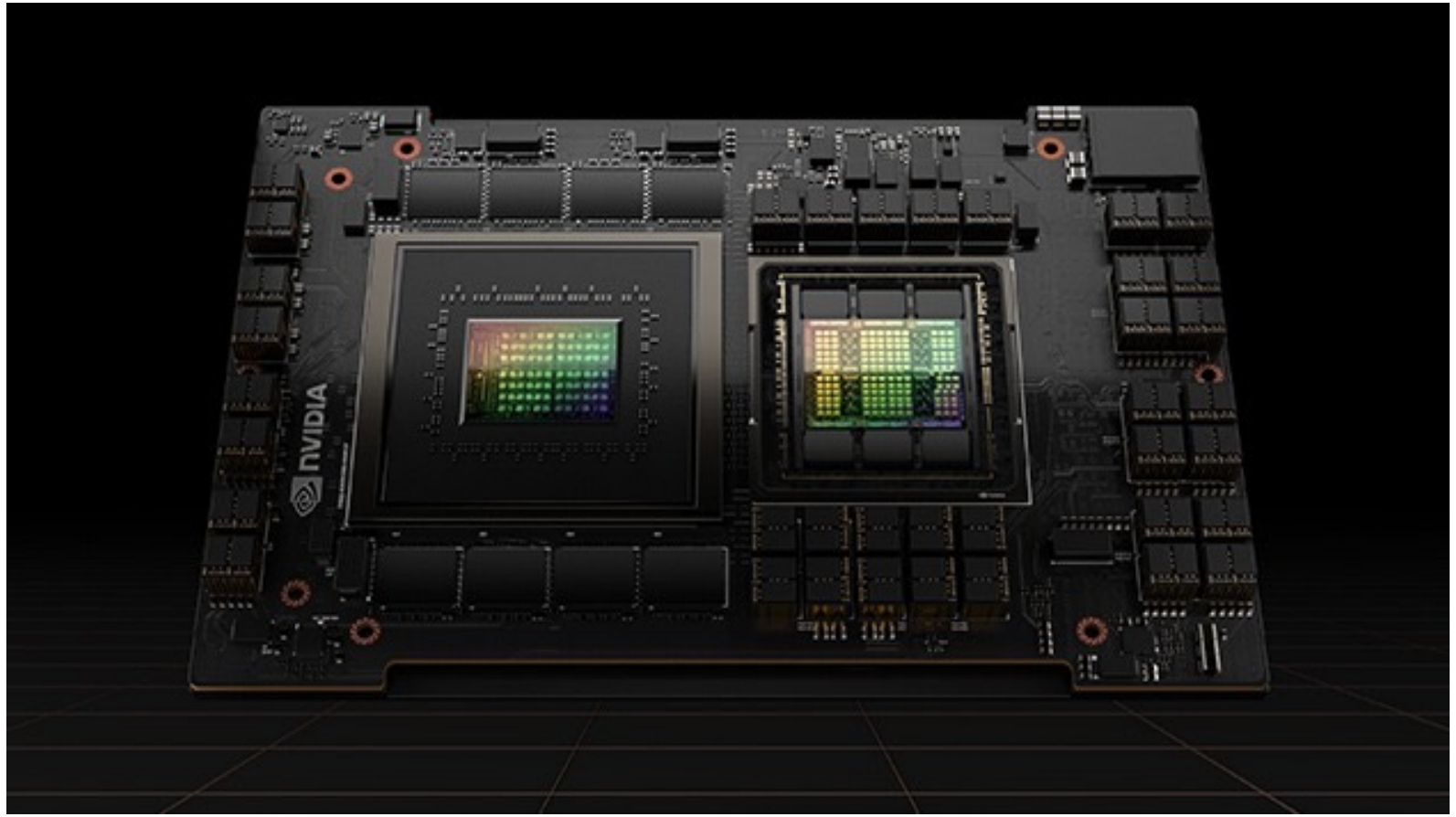
Scaling Laws for Neural Language Models

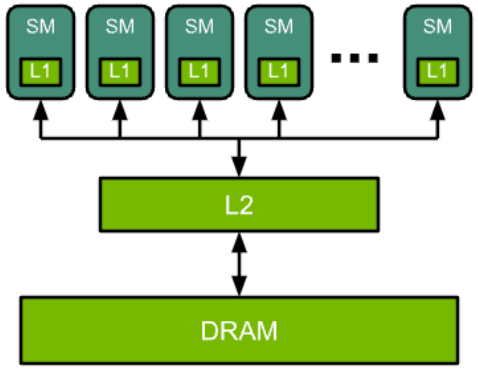


https://arxiv.org/pdf/2001.08361.pdf

Nvidia GPU Performance

RELAXED SYSTEM LAB



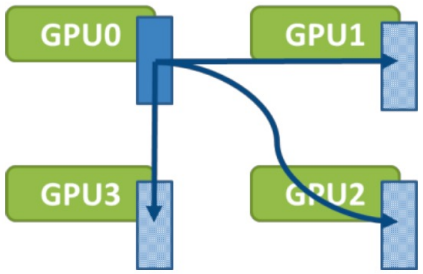


https://www.nvidia.com/en-us/data-center/h100/

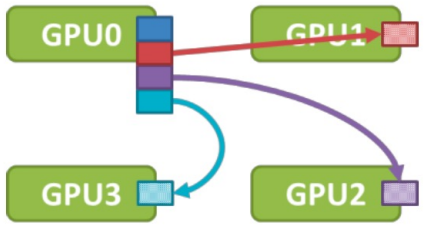
Nvidia Collective Communication Library

RELAXED SYSTEM LAB

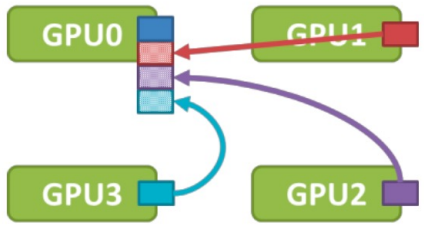
Broadcast



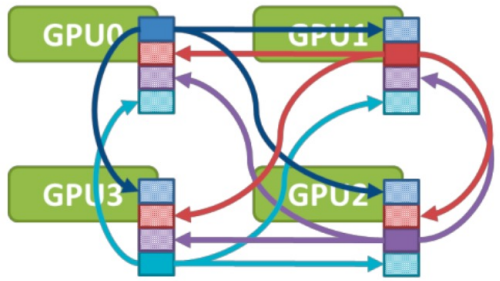
Scatter



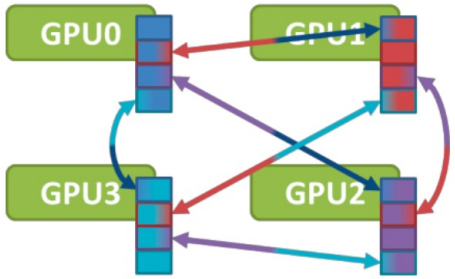
Gather



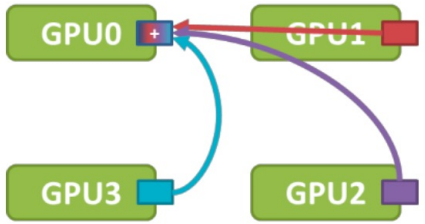
All-Gather



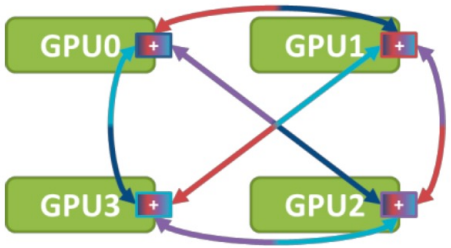
All-to-All



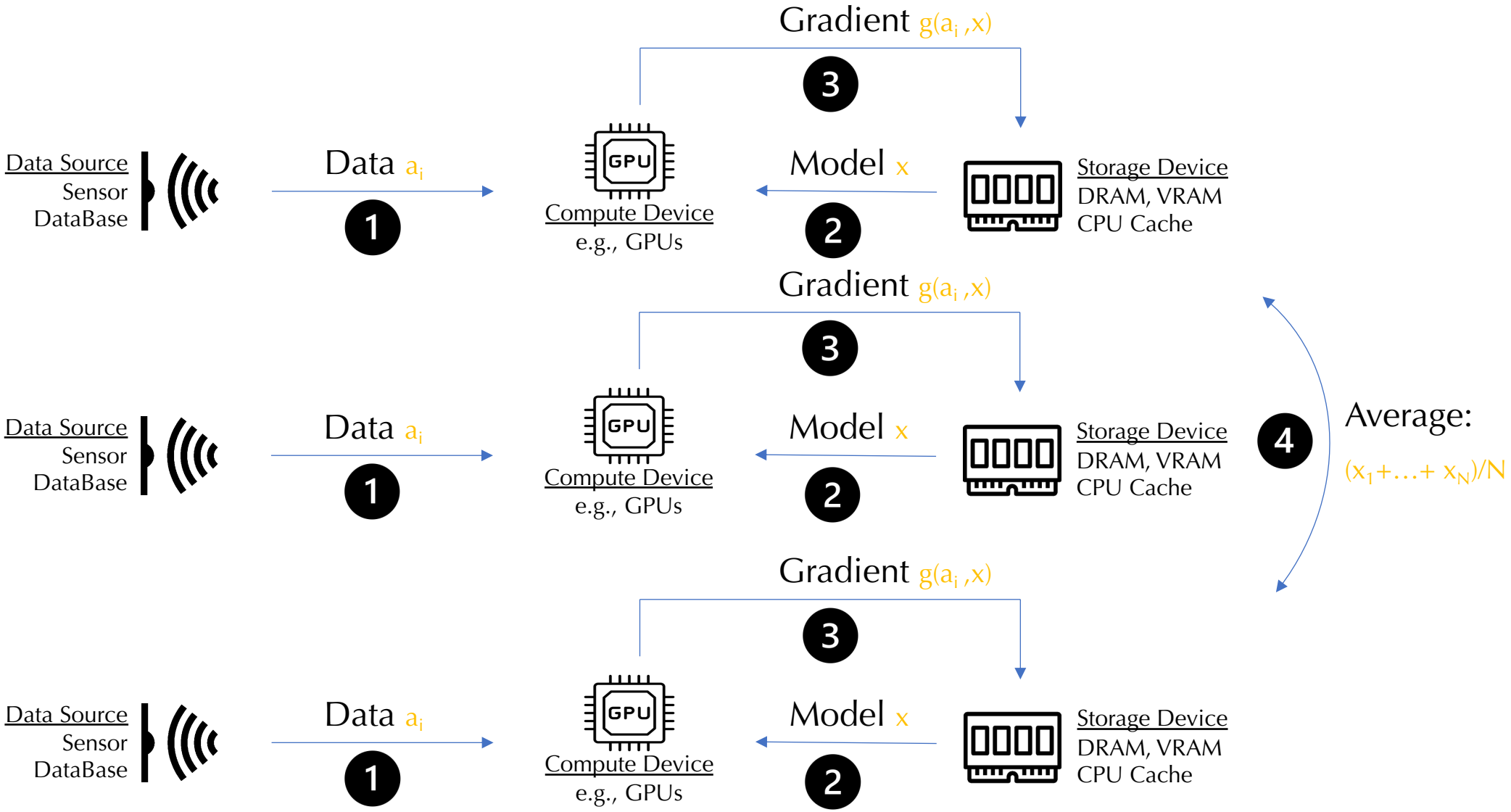
Reduce



All-Reduce

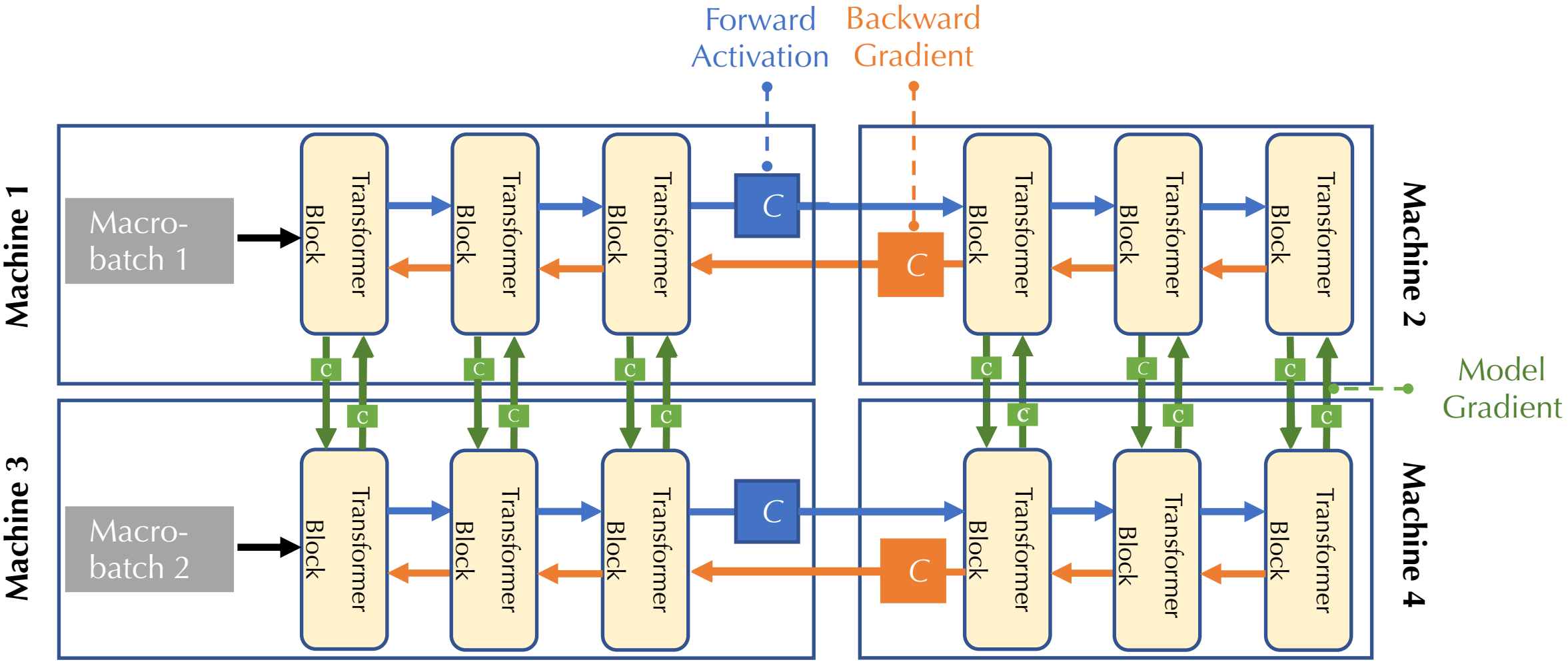


Data Parallelism



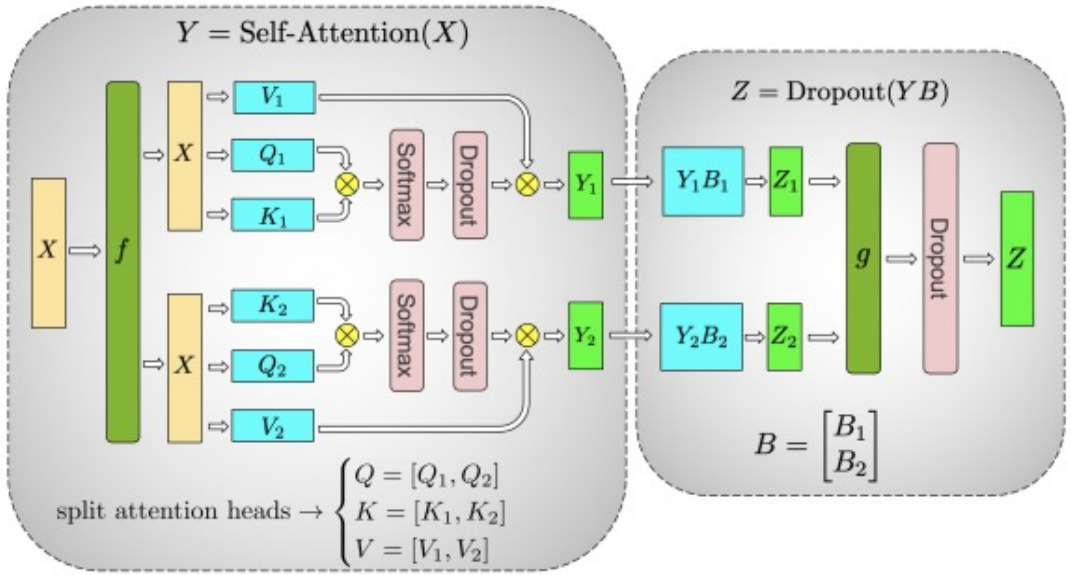
Pipeline Parallelism

RELAXED SYSTEM LAB

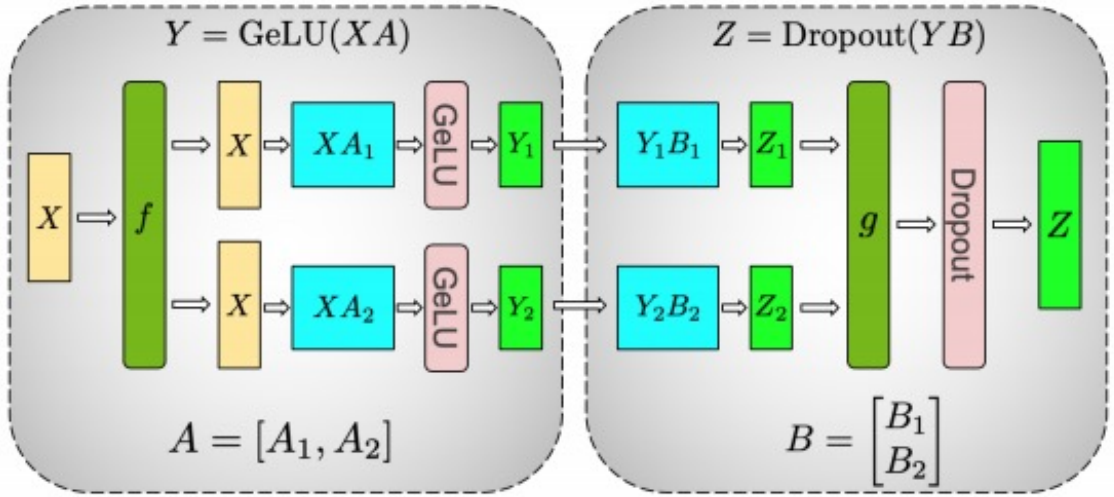


Tensor Model Parallelism

RELAXED SYSTEM LAB



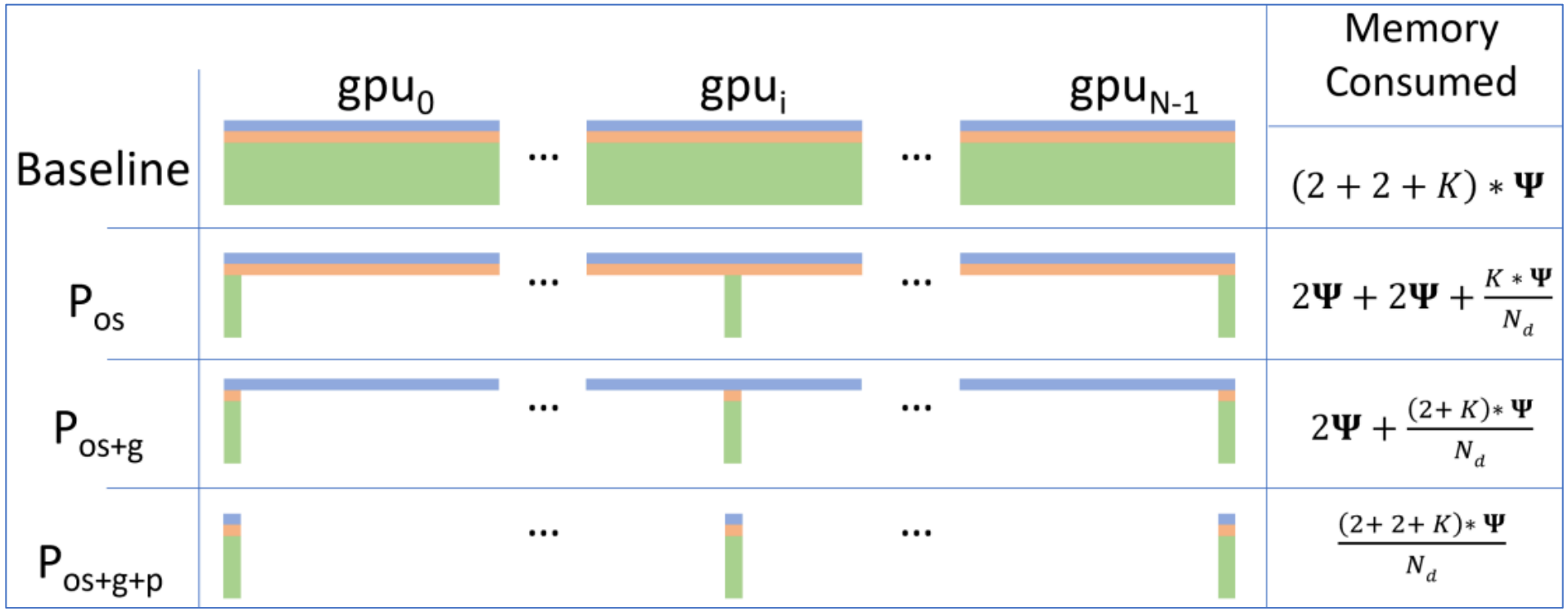
(b) Self-Attention



(a) MLP

Optimizer Parallelism

RELAXED SYSTEM LAB

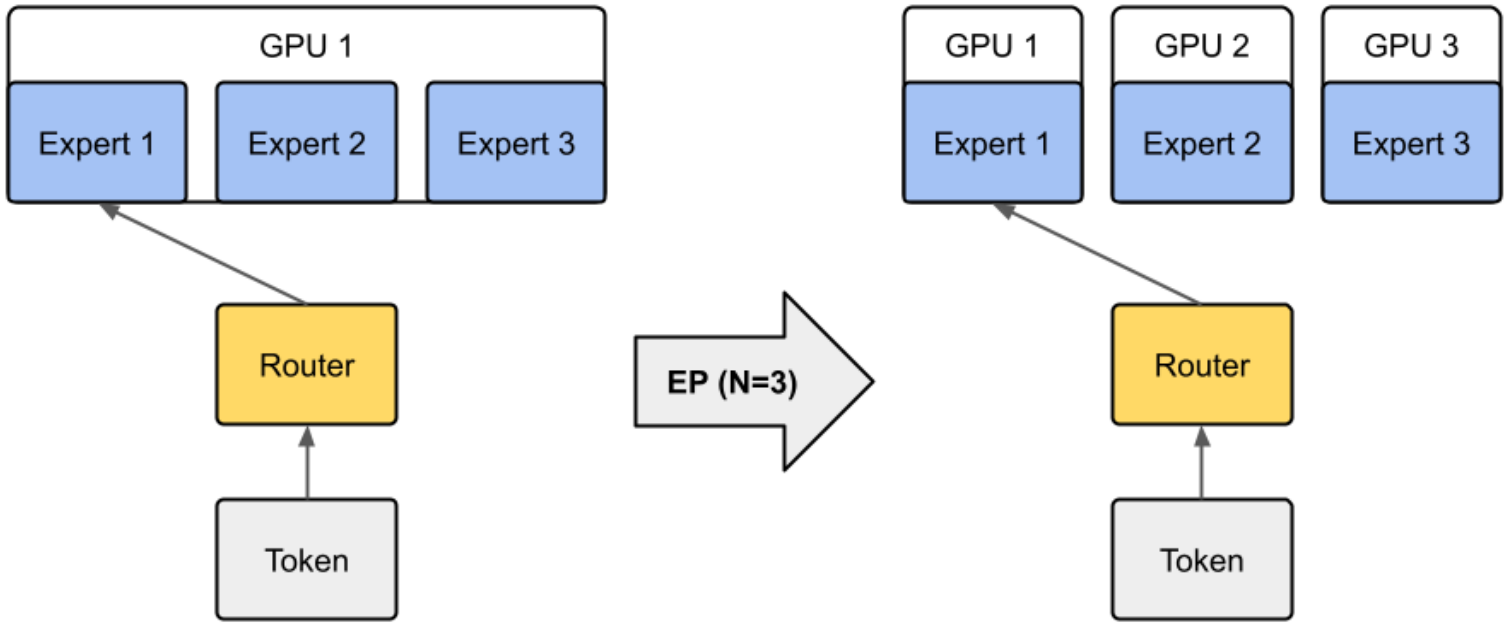


Zero Redundancy Optimizer (ZeRO)

Expert Parallelism

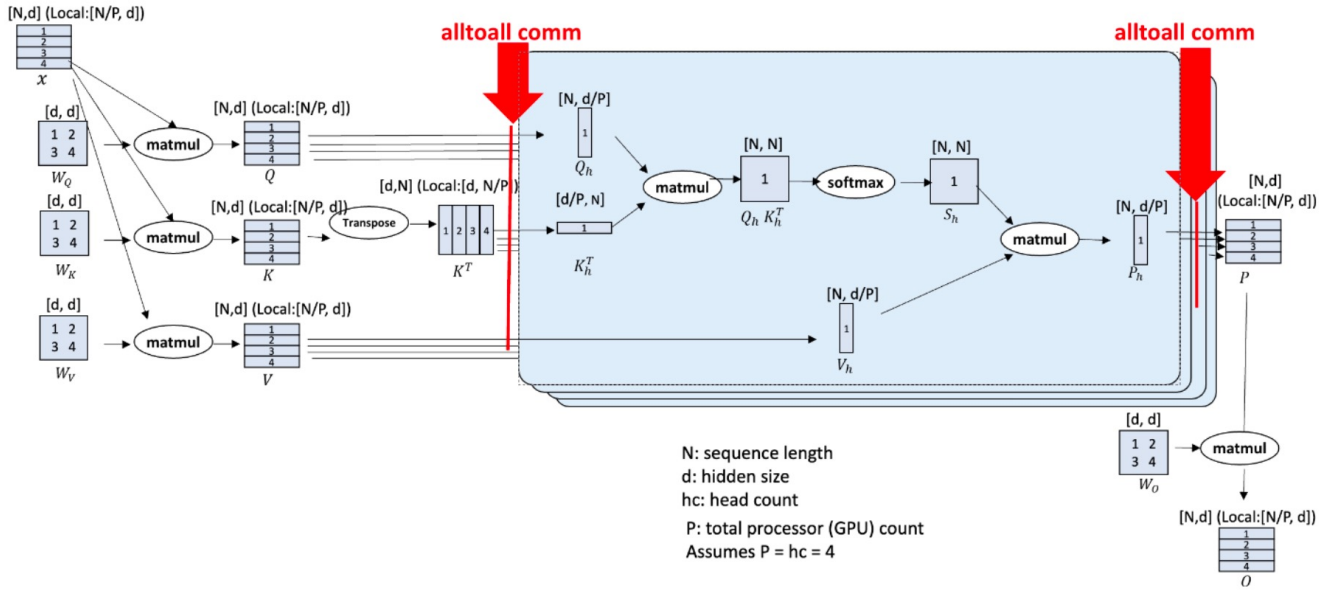
RELAXED SYSTEM LAB

Expert Parallelism applied on Mixture-of-Experts.



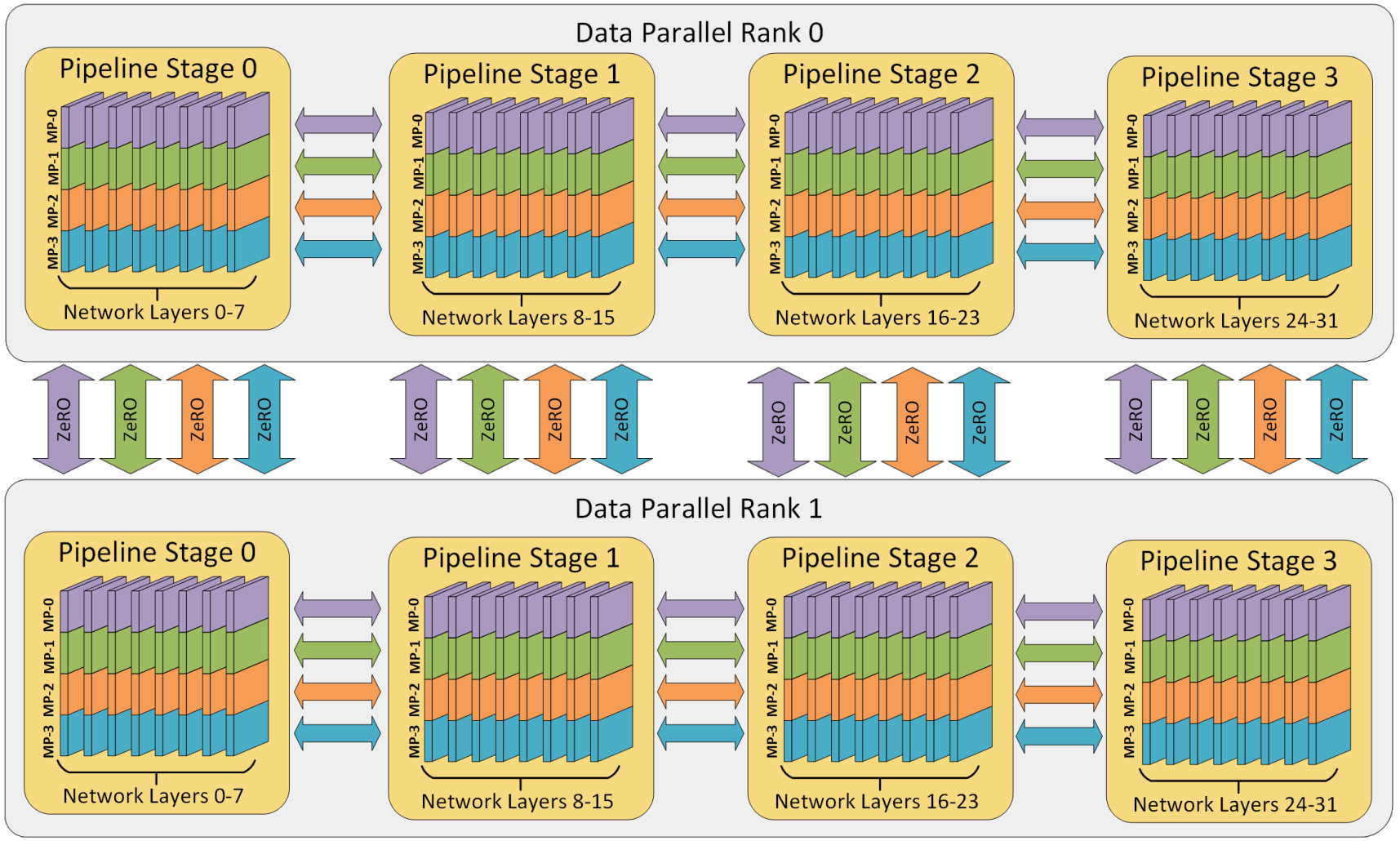
Long Sequence Parallelism

RELAXED SYSTEM LAB



Data-, Pipeline-, Tensor Model-, Optimizer- Parallelisms

RELAXED SYSTEM LAB



https://www.deepspeed.ai/getting-started/

Generative Inference & Hugging Face

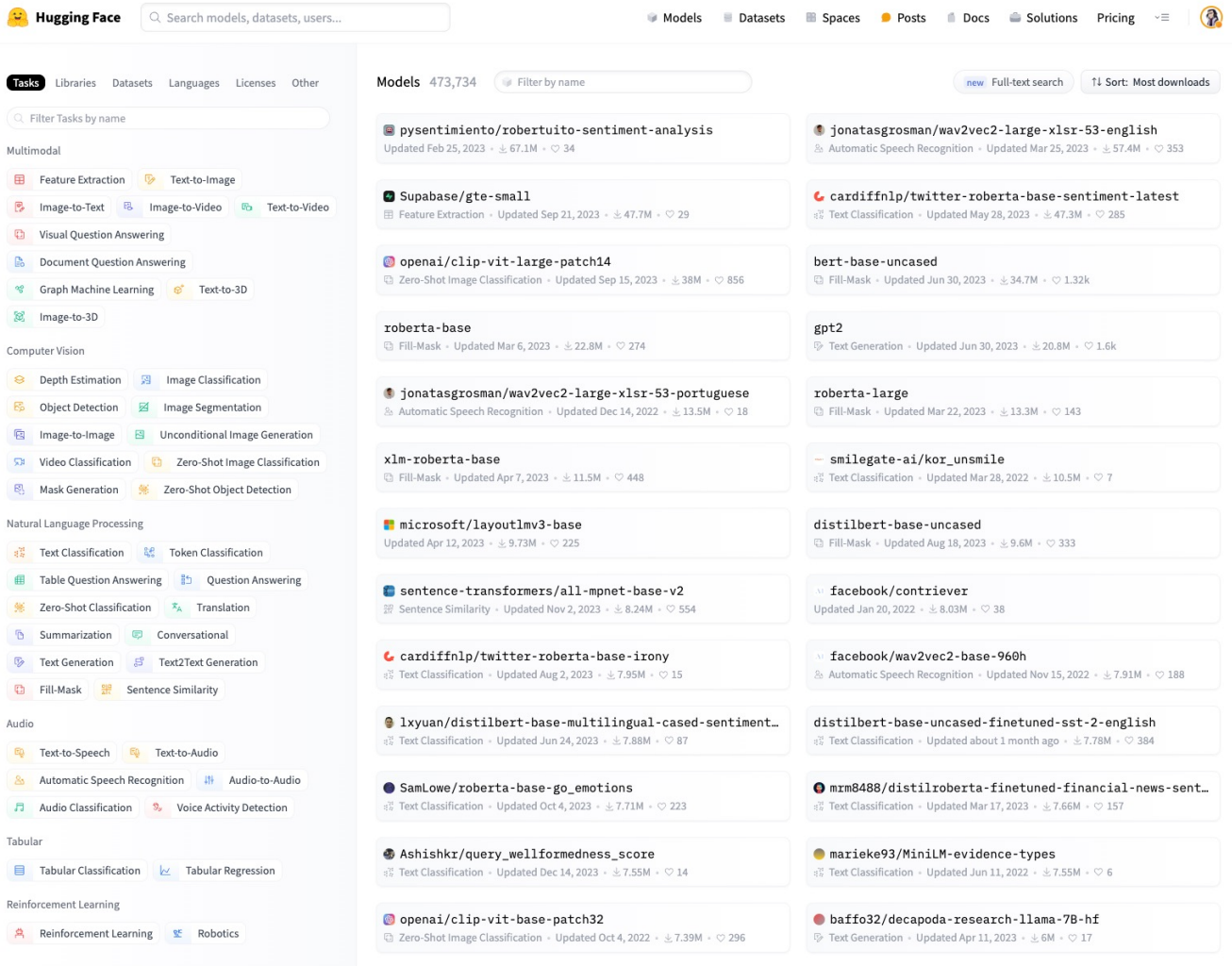
RELAXED SYSTEM LAB

```python
from transformers import AutoTokenizer
import transformers
import torch

model = "meta-llama/Llama-2-7b-chat-hf"

tokenizer = AutoTokenizer.from_pretrained(model)
pipeline = transformers.pipeline(
    "text-generation",
    model=model,
    torch_dtype=torch.float16,
    device_map="auto",
)

sequences = pipeline(
    'I liked "Breaking Bad" and "Band of Brothers". Do you have any recommendations of other shows I
    do_sample=True,
    top_k=10,
    num_return_sequences=1,
    eos_token_id=tokenizer.eos_token_id,
    max_length=200,
)
for seq in sequences:
    print(f"Result: {seq['generated_text']}")
```

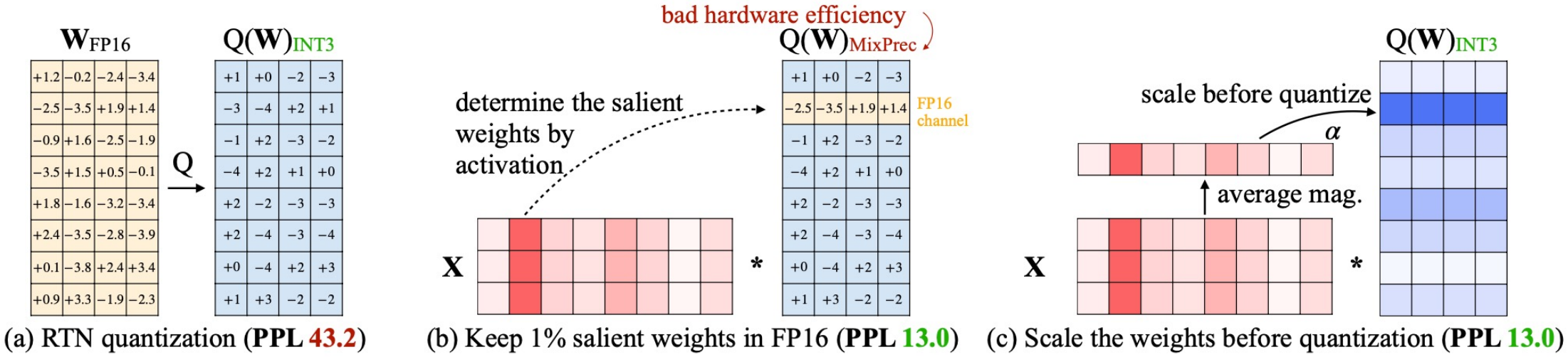


https://huggingface.co/models

Generative Inference Algorithm Optimization

RELAXED SYSTEM LAB

Quantized LLM Inference

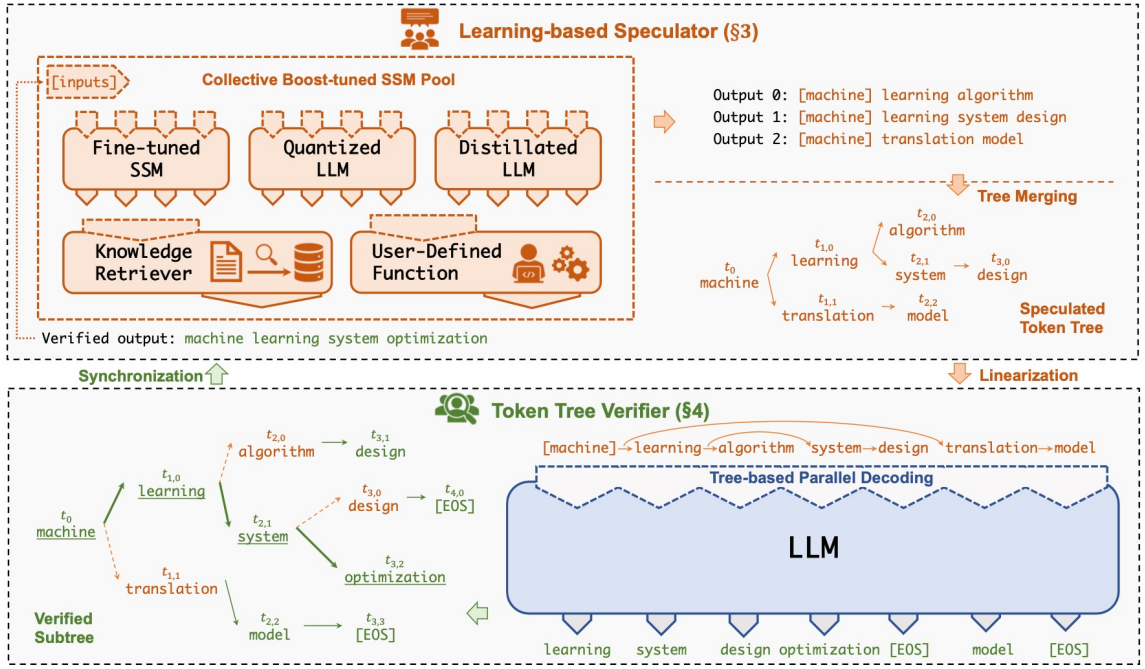


(a) RTN quantization (**PPL 43.2**) (b) Keep 1% salient weights in FP16 (**PPL 13.0**) (c) Scale the weights before quantization (**PPL 13.0**)

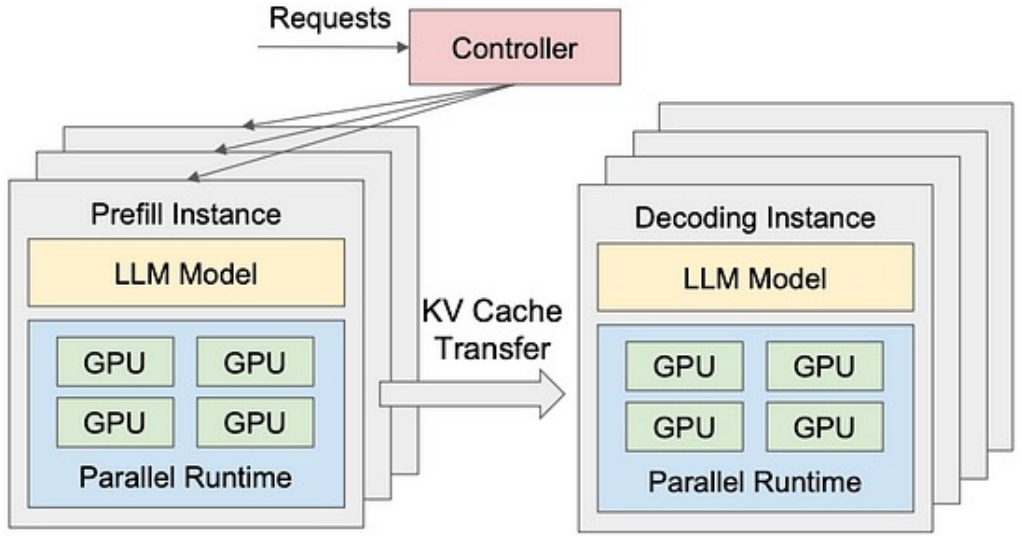
https://arxiv.org/pdf/2306.00978.pdf

Generative Inference System Optimization

RELAXED SYSTEM LAB



Speculative Decoding

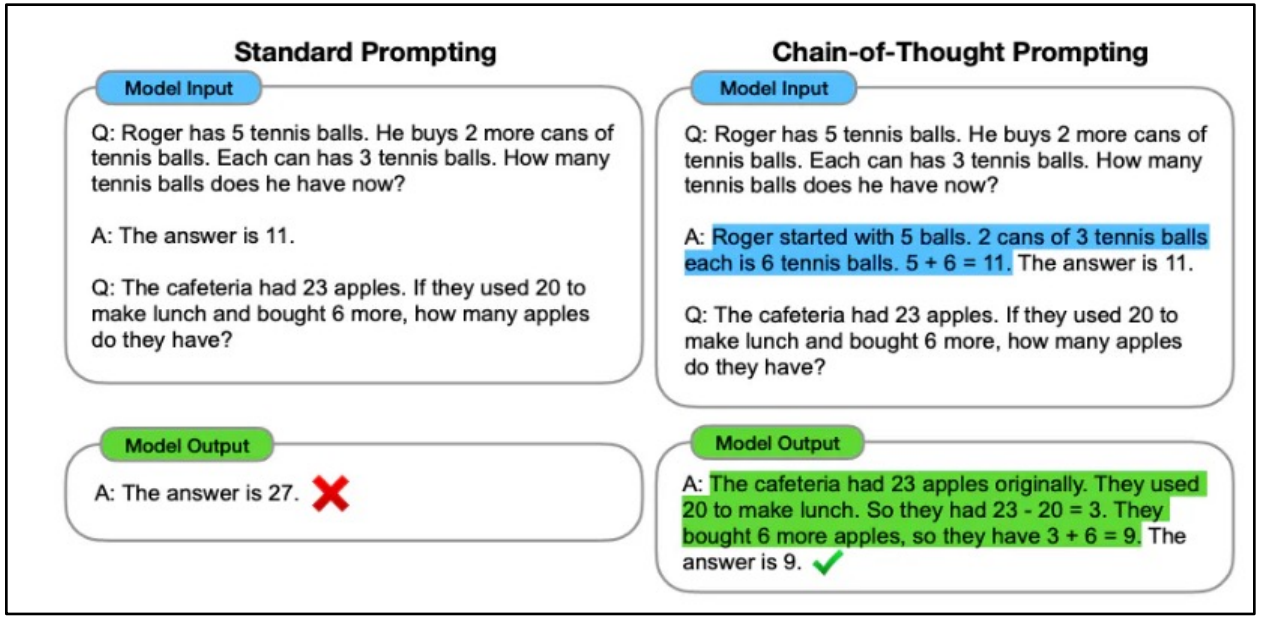


Prefill-decode Disaggregated Inference

Prompt Engineering Overview & Practices

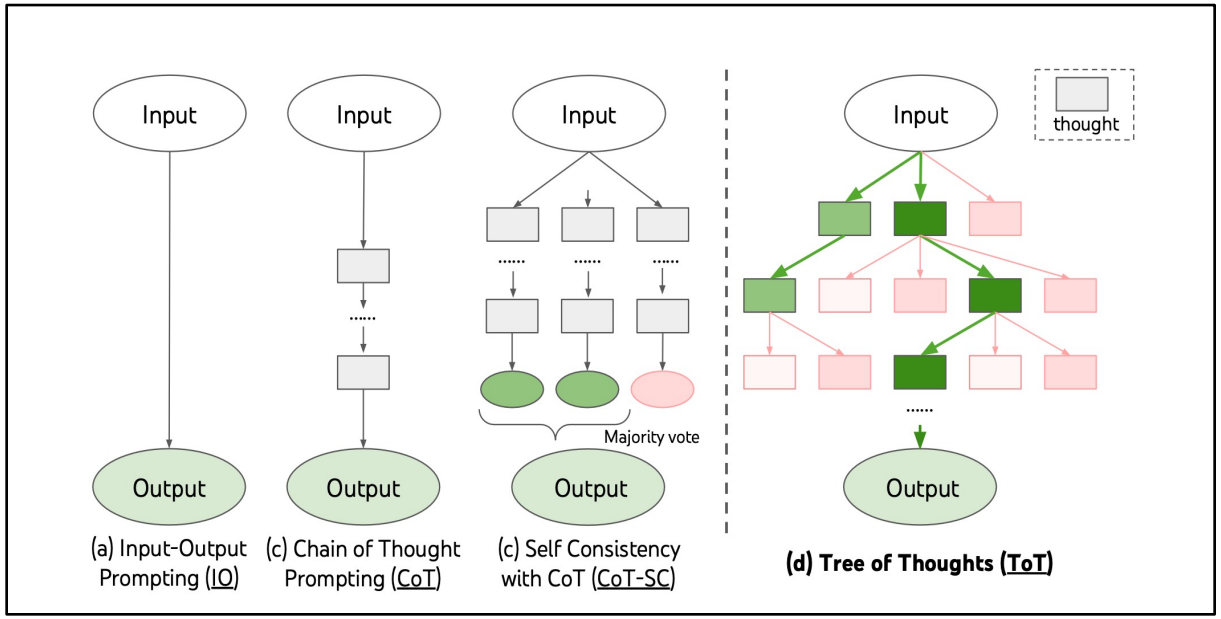
RELAXED SYSTEM LAB

Chain of Thoughts



https://arxiv.org/pdf/2201.11903.pdf

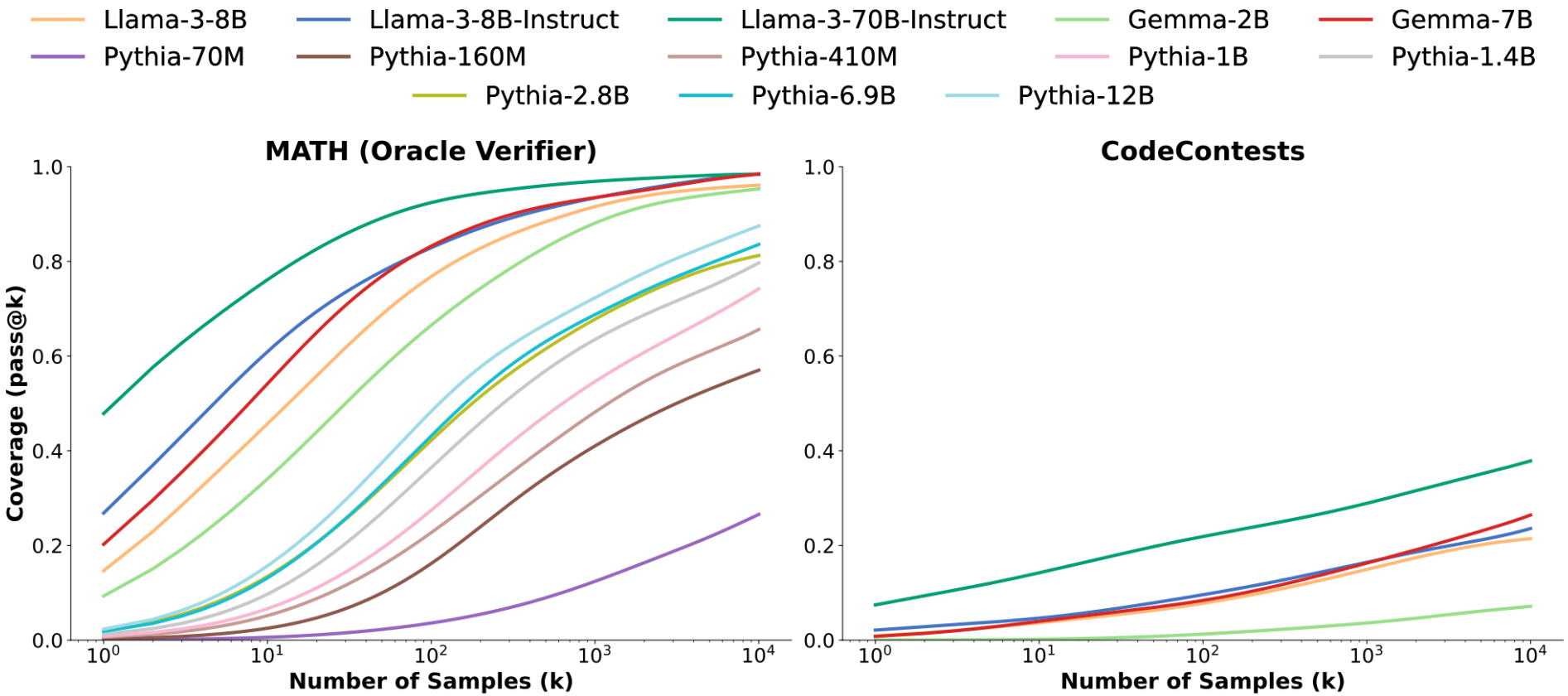
Tree of Thoughts



https://arxiv.org/pdf/2305.10601.pdf

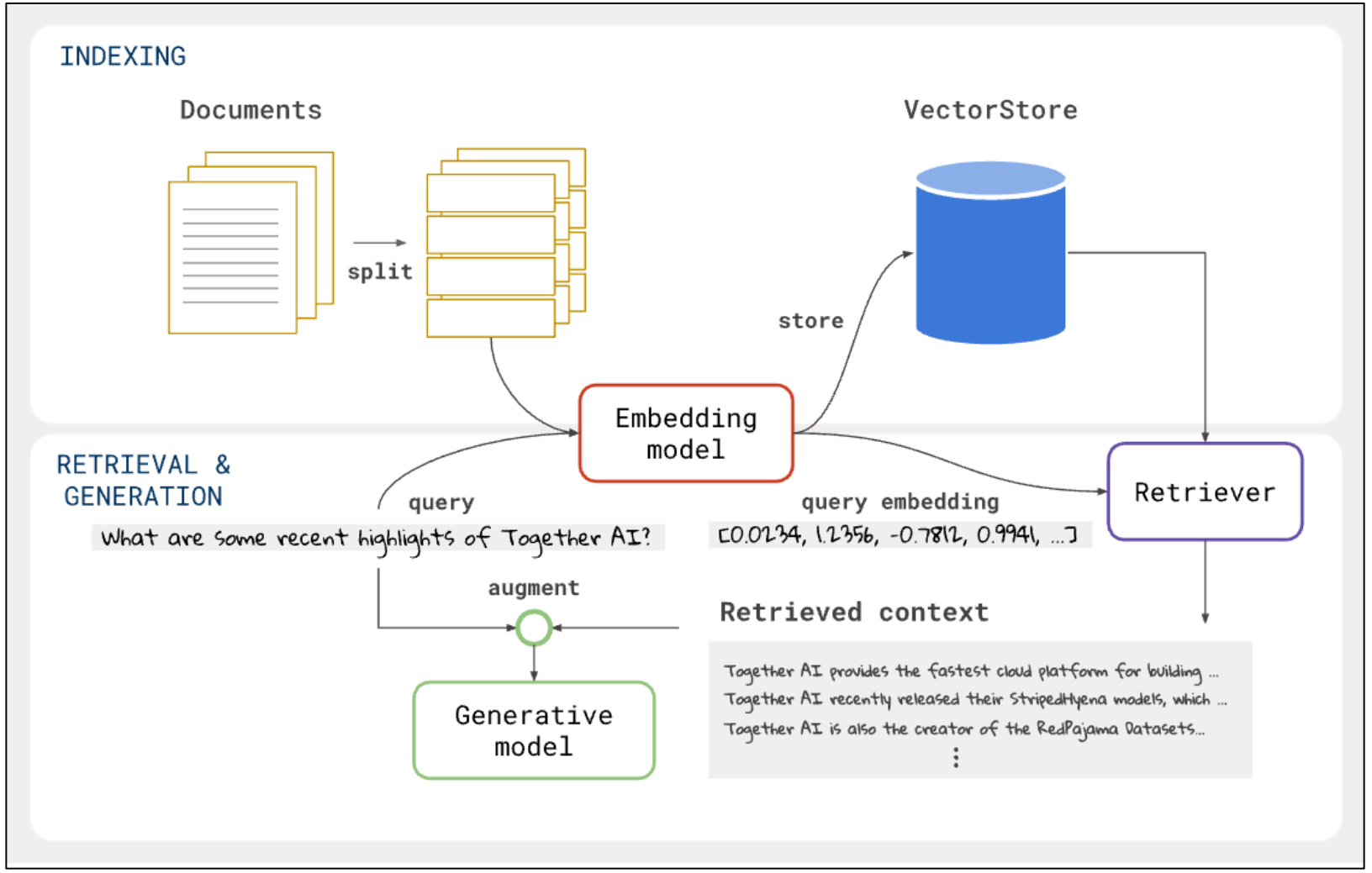
Inference Time Scaling

RELAXED
SYSTEM LAB



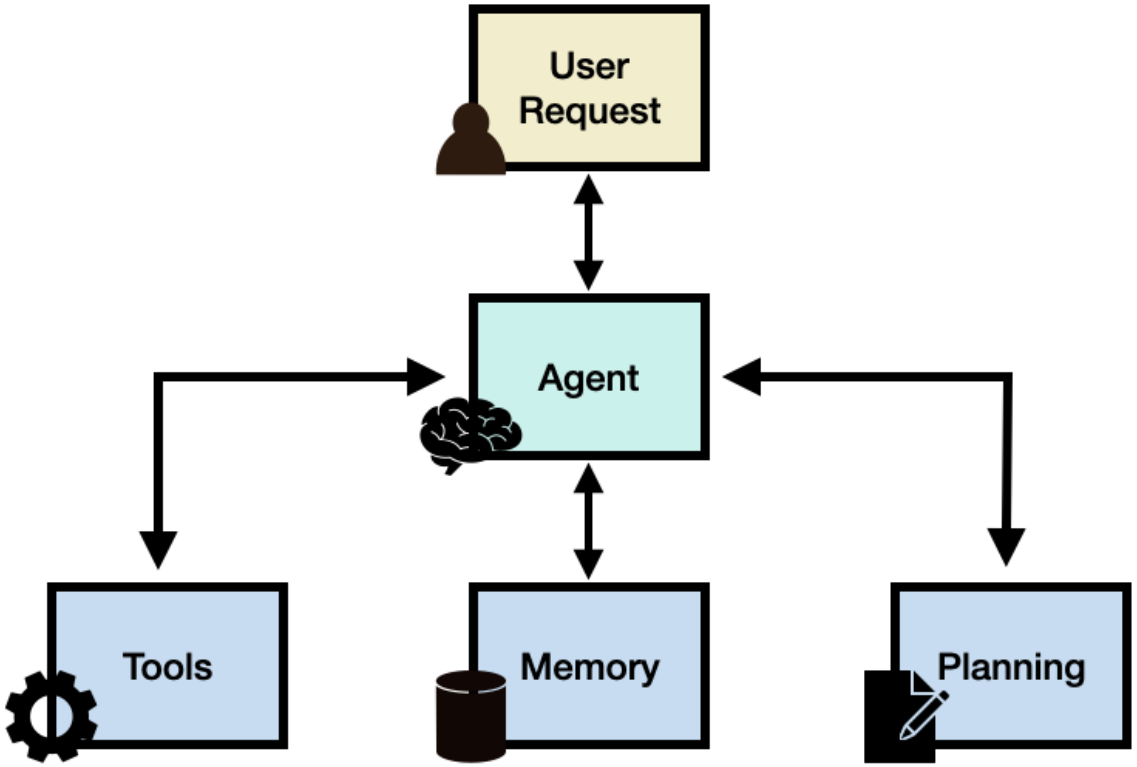
Retrieval Augmented Generation

RELAXED SYSTEM LAB



LLM Agent

RELAXED
SYSTEM LAB

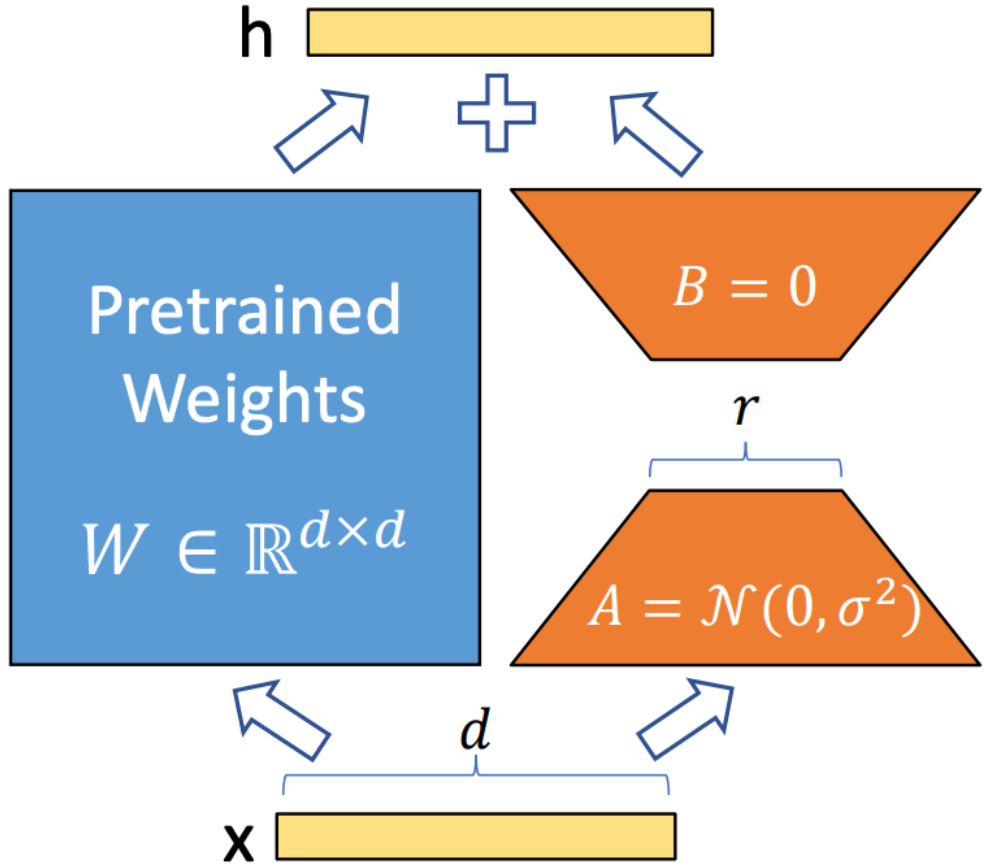


Parameter Efficient Fine-tuning (LoRA)

RELAXED SYSTEM LAB

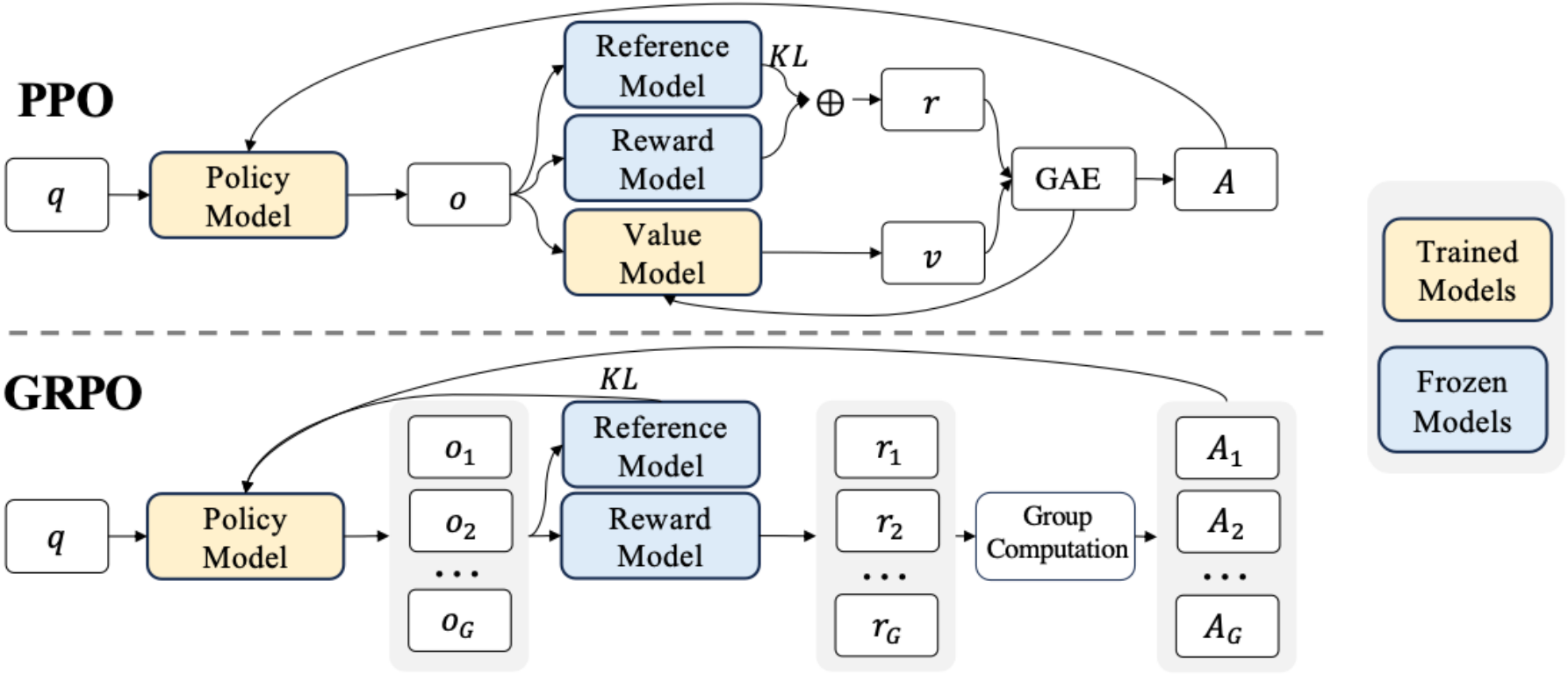
Low-Rank Adaptation (LoRA)

https://arxiv.org/pdf/2106.09685.pdf



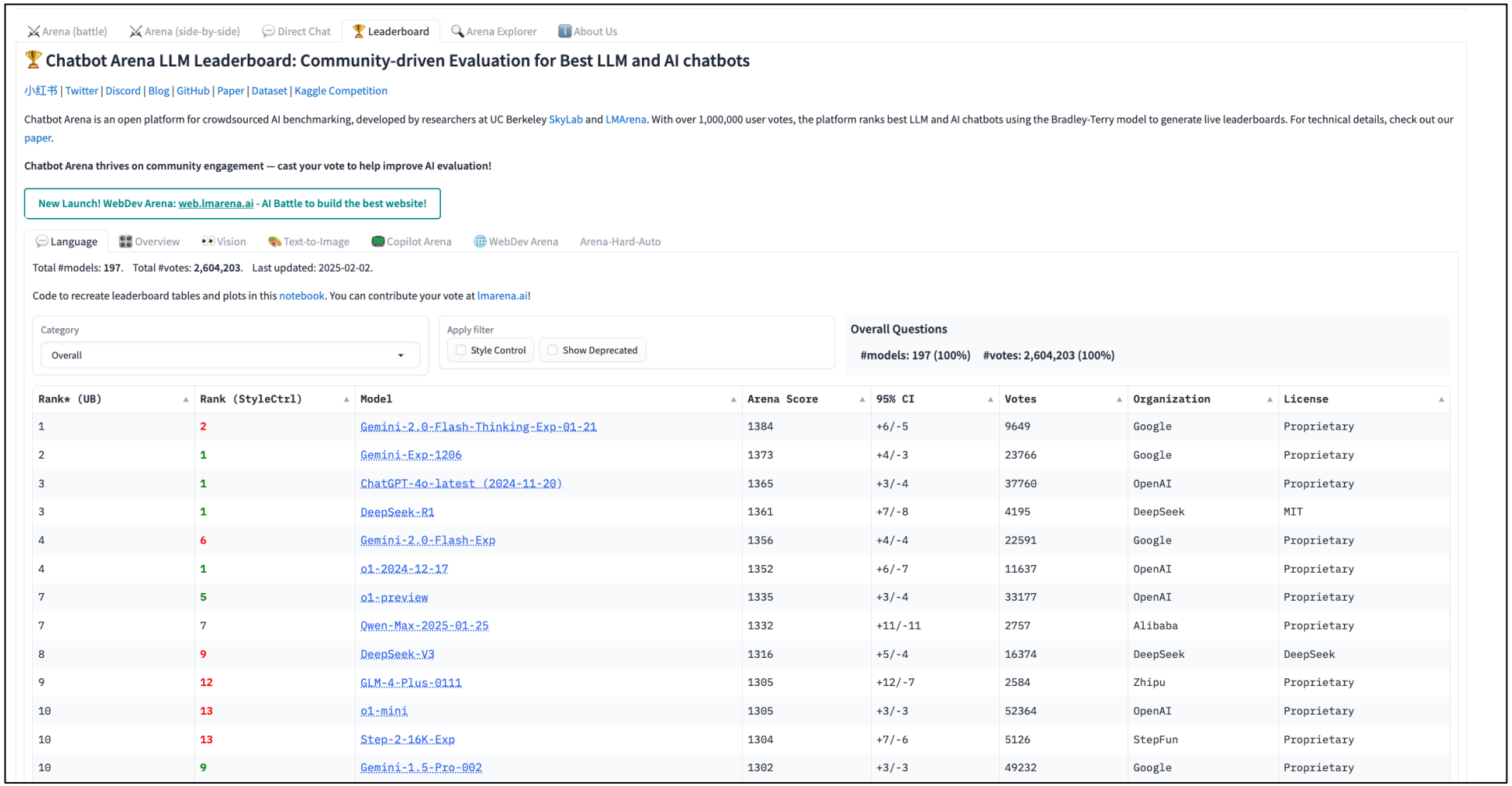
RL Alignment

RELAXED SYSTEM LAB



LLM Evaluation

RELAXED SYSTEM LAB



⚔ Arena (battle) | ⚔ Arena (side-by-side) | 💬 Direct Chat | 🏆 Leaderboard | 🔍 Arena Explorer | ℹ️ About Us

🏆 Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

小红书 | Twitter | Discord | Blog | GitHub | Paper | Dataset | Kaggle Competition

Chatbot Arena is an open platform for crowdsourced AI benchmarking, developed by researchers at UC Berkeley SkyLab and LMArena. With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. For technical details, check out our paper.

Chatbot Arena thrives on community engagement — cast your vote to help improve AI evaluation!

New Launch! WebDev Arena: web.lmarena.ai - AI Battle to build the best website!

💬 Language | 🎛️ Overview | 👀 Vision | 🎨 Text-to-Image | 🟩 Copilot Arena | 🌐 WebDev Arena | Arena-Hard-Auto

Total #models: **197**. Total #votes: **2,604,203**. Last updated: 2025-02-02.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

Overall Questions

#models: 197 (100%) #votes: 2,604,203 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	2	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-5	9649	Google	Proprietary
2	1	Gemini-Exp-1206	1373	+4/-3	23766	Google	Proprietary
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+3/-4	37760	OpenAI	Proprietary
3	1	DeepSeek-R1	1361	+7/-8	4195	DeepSeek	MIT
4	6	Gemini-2.0-Flash-Exp	1356	+4/-4	22591	Google	Proprietary
4	1	o1-2024-12-17	1352	+6/-7	11637	OpenAI	Proprietary
7	5	o1-preview	1335	+3/-4	33177	OpenAI	Proprietary
7	7	Qwen-Max-2025-01-25	1332	+11/-11	2757	Alibaba	Proprietary
8	9	DeepSeek-V3	1316	+5/-4	16374	DeepSeek	DeepSeek
9	12	GLM-4-Plus-0111	1305	+12/-7	2584	Zhipu	Proprietary
10	13	o1-mini	1305	+3/-3	52364	OpenAI	Proprietary
10	13	Step-2-16K-Exp	1304	+7/-6	5126	StepFun	Proprietary
10	9	Gemini-1.5-Pro-002	1302	+3/-3	49232	Google	Proprietary

Guest Speech

RELAXED
SYSTEM LAB

Logistics

Grading Policy

RELAXED
SYSTEM LAB

- 4 Homework:
  - 4 × 5% = 20%
- Mid-term Exam (30%):
  - In-class exam (80 Minutes): **2025/03/24**
- Final Exam (50%):
  - 2 hour
  - Time: **TBD**

Temporary Homework Schedule

RELAXED SYSTEM LAB

- **Homework-1**
  - Release: 2025/02/13
  - DDL: **2025/02/22 23:59**
- **Homework-2**
  - Release: 2025/03/04
  - DDL: **2025/03/12 23:59**
- **Homework-3**
  - Release: 2025/03/25
  - DDL: **2025/04/10 23:59**
- **Homework-4**
  - Release: 2025/04/22
  - DDL: **2025/05/06 23:59**

Homework

RELAXED
SYSTEM LAB

- Delay penalty (20% each day after the DDL)
- The homework must be finished by yourself:
  - Discussion is welcomed;
  - Copy source code or any other content from other people is strictly fobidden;
  - Using generative AI is cool.

Exams

RELAXED
SYSTEM LAB

- Close-book;
- **<u>One A4 page of cheating sheet allowed</u>**;
- The mid-term exam covers the topics during the first half of the semester;
- The final exam covers all the topics.

No Attendance Requirement for the Lecture!

The Laboratory will be hosted with Announcements!

GitHub

https://github.com/Relaxed-System-Lab/HKUST-COMP4551-2026spring