Tencent AI
Platform Dept.

Edge-Device LLM Competition @ NeurIPS 2024

Team Tinytron

2024/12/15

Contents

Tencent AI
Platform Dept.

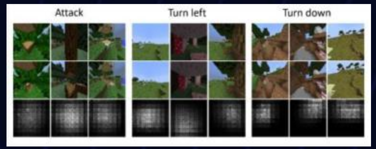
About Us

Tencent AI
Platform Dept.

- Tencent AI Platform Dept. (AIPD): Central Hub for AI Research and Applications in Gaming
    - Founded in 2016, We are a trailblazer of AI + Gaming in China.
    - Research Scope: Decision AI (Massive RL) + Generative AI (Large Models)
    - Industrial Application: 30+ regions, 50M+ DAU, 1B+ daily API calls, 700+ patents

FineArt (2016~)
Adopted in Training Program of China's National Go Team

WeKick (2020)
Winner of Google Research Football Simulation Competition



Juewu-MC (2021)
Champion of NeurIPS 2021 MineRL Competition (Sample Efficient RL)

LuckyJ (2023~)
1st Mahjong AI reaching the 10th dan on Tenhou.net

AI Coaching (2024~)
Voice Coaching for MOBA Game HoK
On-Device Deployment of TTS Model

- Team Tinytron: Close Collaboration between AI Algorithm Researchers and System Engineers

Lvfang Tao

Linhang Cai

Renjie Mao

—— Model Development and Evaluation

Yongguang Lin

Xiaowen Huang

—— System Optimization for Edge Device

Challenge Breakdown

Tencent AI
Platform Dept.

- **Data**: Obtain Best-Possible Training Data from Allowed Sources
  - Cleaning: Heuristic Filtering & Quality Rating
  - Mixing: Efficient Data Composition
  - Generation: Get Data in Target Domain with Pruned Model

- **Training**: Achieve Faster Convergence with Curated Data & Allowed Teacher Model
  - Pruning: Minimizing Loss on Evaluation Tasks
    -> Gradient-Based Pruning with Distillation (Track1)
  - Continued Training / Pretraining: Maximizing Capability Recovery / Improvement on Evaluation Tasks
    -> Continued Pretraining with Distillation (Track1) / Efficient Pretraining (Track2)

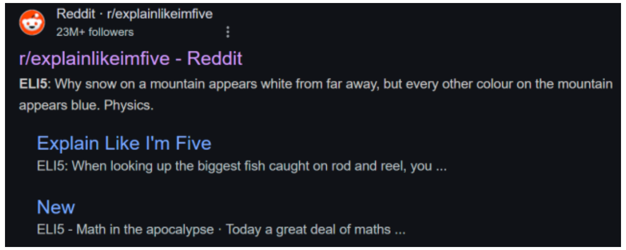
- **On-Device Optimization**
  - Reduce Memory Footprint
  - Reduce Inference Latency

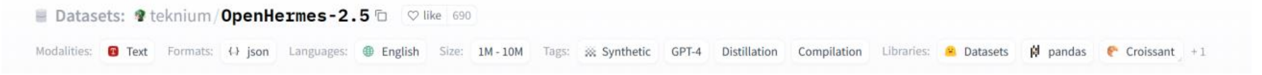
Track1 (Baseline): Data Processing

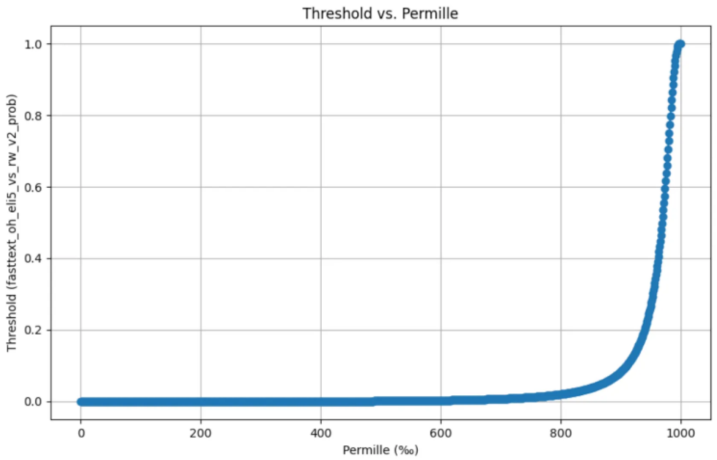
Tencent AI Platform Dept.

- **Tradeoff: Quality vs Quantity**
  - We adopt the OH2.5-ELI5 fastText classifier open-sourced by DCLM-Baseline [1]
  - We compare pruning outcomes of different quality thresholds

[1] Li, Jeffrey, et al. "DataComp-LM: In Search of the Next Generation of Training Sets for Language Models." NeurIPS 2024 Track Datasets and Benchmarks, www.datacomp.ai/dclm.







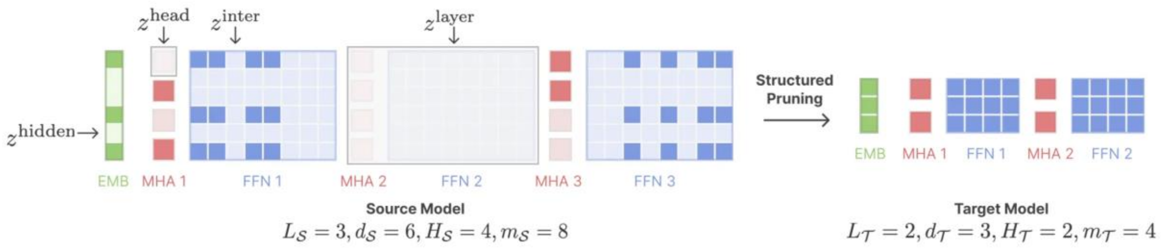
C4 Top Proportion	Commonsense QA	FewCLUE-chid	HumanEval	GSM8K	TruthfulQA	BigBench-Hard	Average Score
5%	29.07	19.33	1.83	17.06	0.2203	21.24	18.43
10%	42.01	16.63	3.66	12.21	0.2375	25.74	20.67
20%	32.92	19.68	3.66	9.40	0.2595	23.68	19.22

- **Data Mix: Step-by-Step Selecting of Optimal Mixing Ratio**
  - C4 (Quality Filtered)
  - Alpaca (Bilingual Augmented)
  - C4 (Code Relevant, Heuristically Filtered with Keywords & Domain Name)

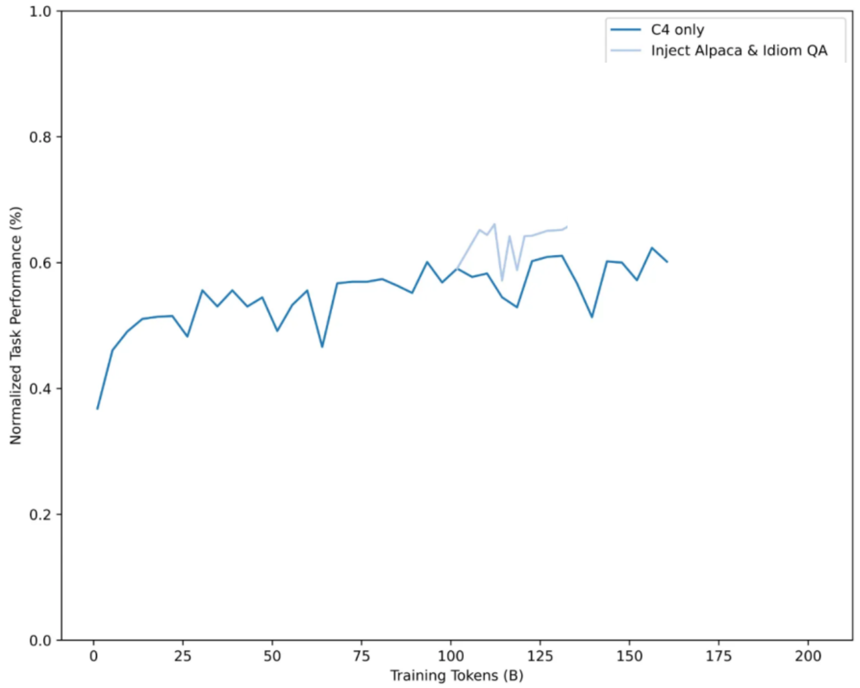
Track1 (Baseline): Pruning with Continued Distillation

Tencent AI
Platform Dept.

- **Solving Pruning Mask with Distilled Gradient**
  - Like Sheared LLaMA[1], we perform gradient-based optimization to solve the structured pruning problem.
  - Weights and masking variables are jointly optimized.
  - Key difference is we compute Kullback–Leibler divergence against teacher logits for more accurate and noise-tolerant gradient, guiding quick recovery of the pruned model.
- **Longer-Term Continued Distillation**
  - The pruning experiment is inefficient as two copies of teacher model weights persist along the training. (1 copy is frozen, the other is active)
  - When pruning mask becomes stable, we perform structured pruning, then use internal training framework (based on Megatron-Core) for compute-efficient continued pretraining on the pruned model (via distillation).



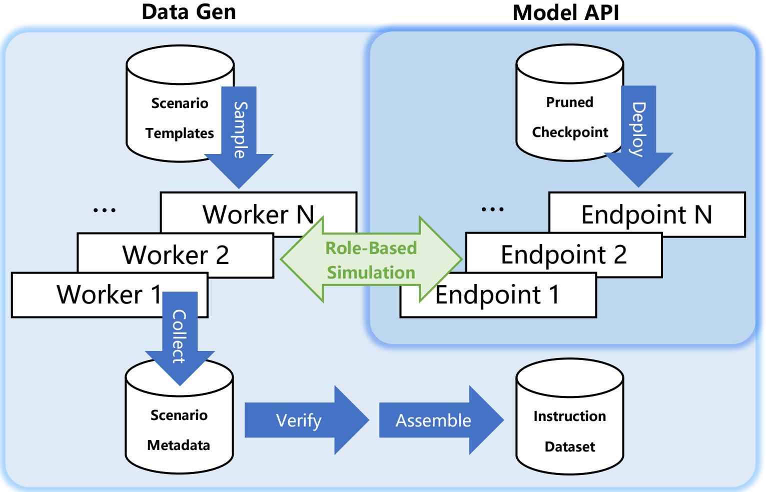
[1] Xia, Mengzhou, et al. "Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning." The Twelfth International Conference on Learning Representations.



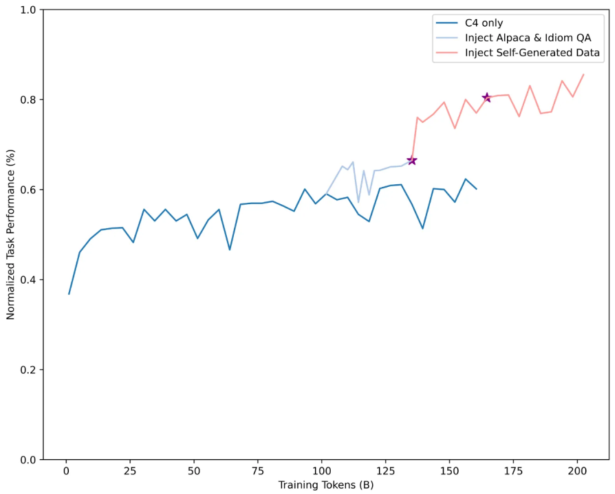
Baseline: Continued Distillation on Qwen 3.9B

Track1 (Final): Data Enrichment with Pruned Model

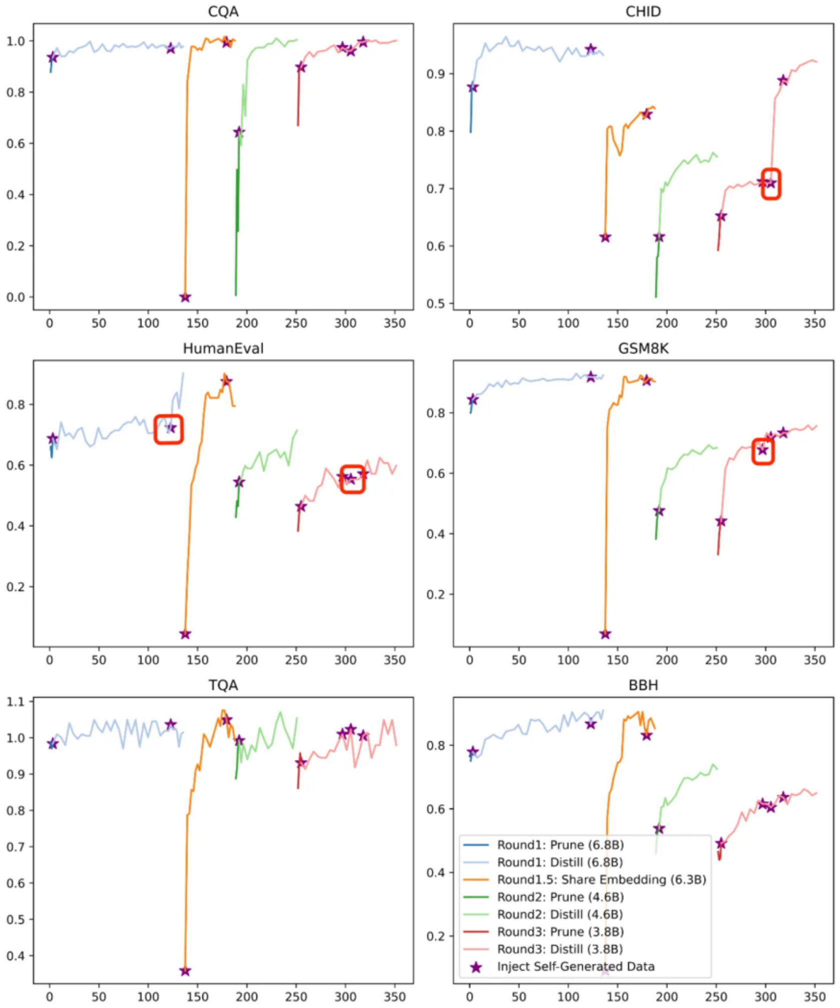
Tencent AI
Platform Dept.

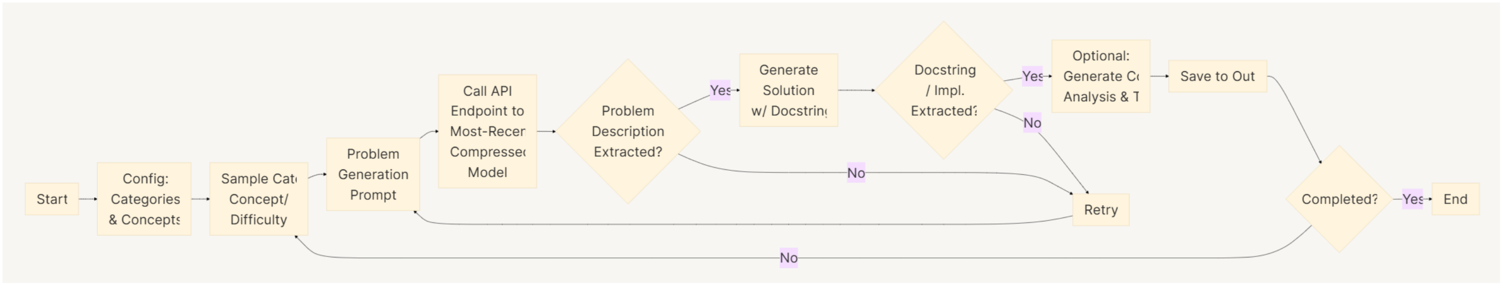


Batch Generation of Instruction Dataset
(with Pruned Student Model)



Continued Distillation on Qwen 3.9B
(with Student-Generated Data)



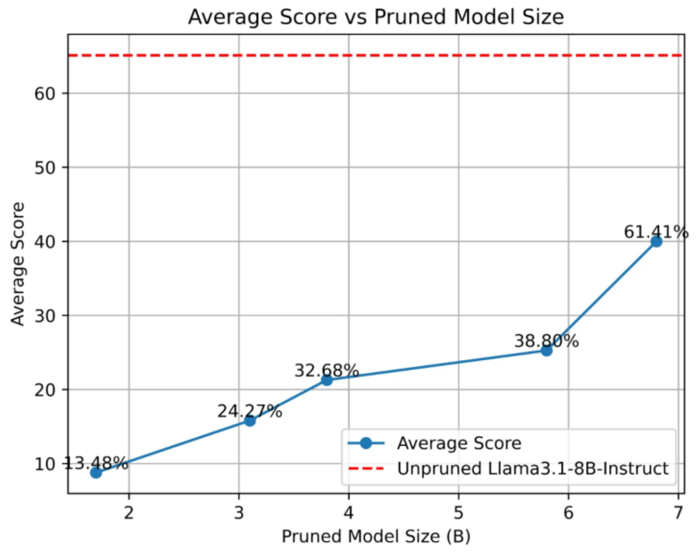


Construction of the CodeExercises Dataset

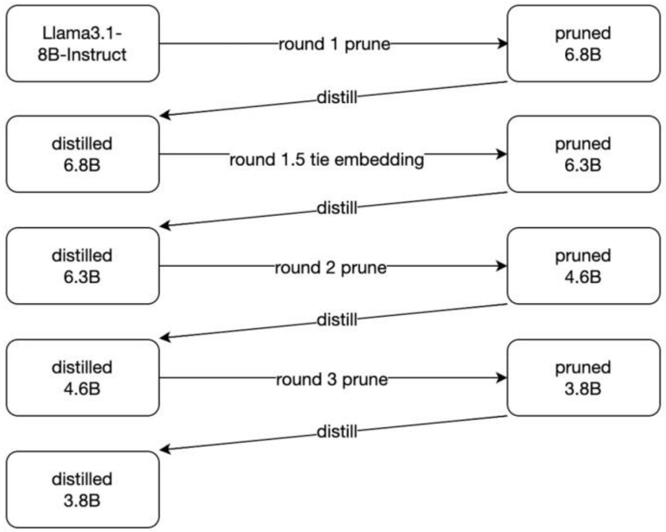
Iterative Pruning with Distillation on Llama 3.8B
(with Student-Generated Data)

Track1 (Final): Progressive Pruning & Continued Distillation

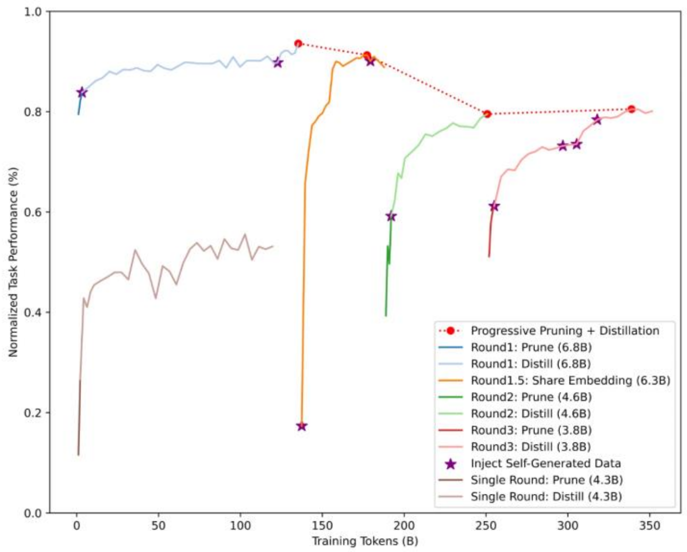
Tencent AI
Platform Dept.



Simple Comparison
(at fixed LR=1e-4)



Progressive Pruning & Continued Distillation
For Llama3.1-8B-Insturct



Normalized Average Performance
(6 Public Tasks)

Track2: Data Processing

Tencent AI
Platform Dept.

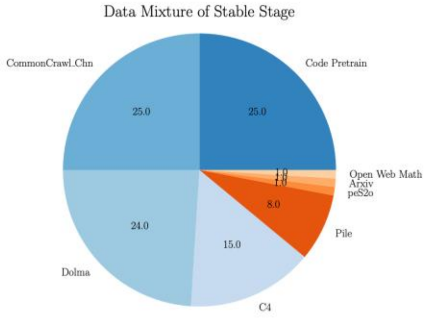
- **Baseline**
  - Heuristic Filtering & Quality Rating (same with Track1)
  - Split into chunks (C4 English)
  - MAP-NEO Chinese pipeline (C4-zh)

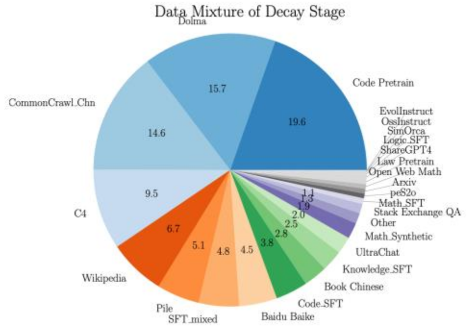
- **Improvement: for math and multi-round ability**
  - Further filtering C4-zh based on sentence structure
  - Construct idiom cloze multiple-choice QA pairs by replacing idioms in Chinese text with "__"
  - Rule-based generation (Simple math, multi-round QA pairs, list/dict manipulation)

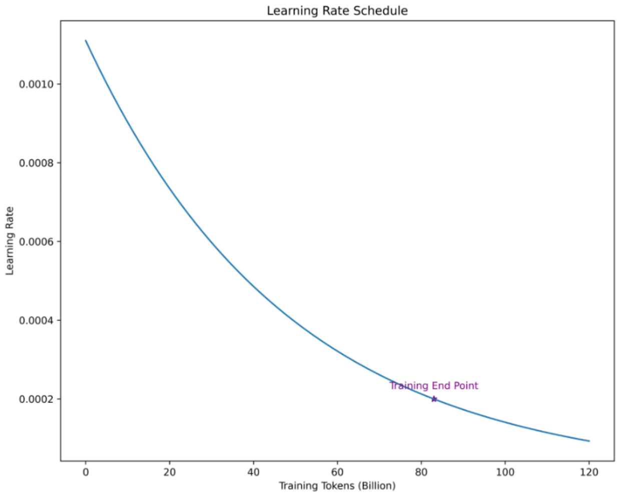
Track2: Training

Tencent AI
Platform Dept.

- **Baseline: Curriculum Learning with Learning Rate Annealing (Referencing MiniCPM)**
  - Stable stage
    - Relatively coarse-grained data
    - High & Stable learning rate
    - Doubling batch size during training
  - Decay stage
    - Add high quality data
    - Alter proportion of chunks with different qualities
    - Exponential (half-life) decay





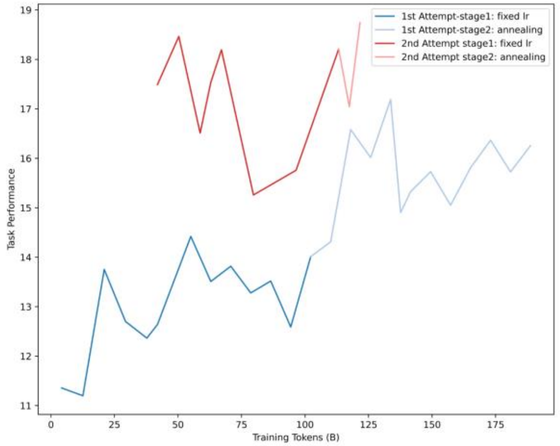


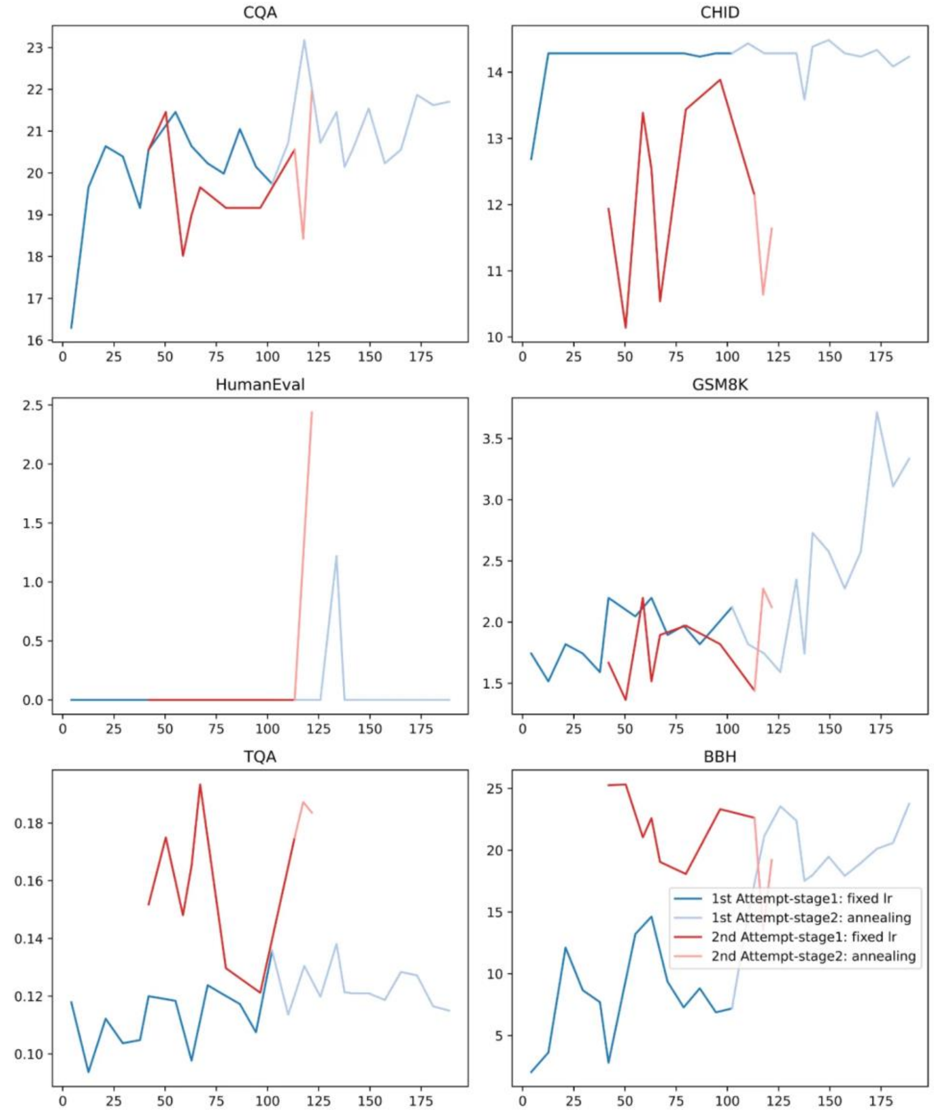
- **Improvement: Mitigating Overfitting**
  - Attention & hidden layers dropout
  - Limit the maximum repetitions for each dataset

Track2: Training recipe

Tencent AI
Platform Dept.

- **1st Attempt: Baseline**
  - Architecture: QKV bias on / QK LayerNorm
  - Basic Data Filtering / Mixing
  - Curriculum Learning with Learning Rate Annealing as MiniCPM
- **2nd Attempt: Final**
  - Enhanced Data Filtering / Mixing / Generation
  - Mitigating Overfitting: Adopt Dropout in attention & hidden layer





System Optimization for Edge-Device LLM Inference

Tencent AI
Platform Dept.

- **Memory optimization**
  - Modifying **mlc-chat-config.json** for reduction of memory occupancy of intermediate tensors
    - Set **max_batch_size** to 1 since the client only needs one user interaction
    - Set **prefill_chunk_size** to 16 for balancing prefilling and memory occupancy
    - Set **context_window_size** to 512 or 768 as needed
  - Force the function getting global memory to return 16GB

- **Inference speed optimization:**
  - Transpose weight layout for flexible computing
  - Prefill Matmul operation
    - Template: dl.gpu.Matmul()
  - Modify Decode Matmul operation
    - Template: dl.gpu.GEMV()

```
uint64_t TotalDetectGlobalMemory(DLDevice device) {
  // Get single-card GPU size.
  TVMRetValue rv;
  DeviceAPI::Get(device)->GetAttr(device, DeviceAttrKind::kTotalGlobalMemory, &rv);
  int64_t gpu_size_bytes = rv;
  // Since the memory size returned by the OpenCL runtime is smaller than the actual available
  // memory space, we set a best available space so that MLC LLM can run 7B or 8B models on Android
  // with OpenCL.
  if (device.device_type == kDLOpenCL) {
+   int64_t min_size_bytes = 16LL * 1024 * 1024 * 1024;  //  Minimum size is 16 GB
    gpu_size_bytes = std::max(gpu_size_bytes, min_size_bytes);
  }
  return gpu_size_bytes;
```

```
class NoQuantize:  # pylint: disable=too-many-instance-attributes
+   """Configuration for no quantization but transpose"""

    name: str
    kind: str
    model_dtype: str  # "float16", "float32"

    def __post_init__(self):
        assert self.kind == "no-quant"
+       self._func_cache = {}
+
+
+   def quantize_model(
+           self,
+           model: nn.Module,
+           quant_map: QuantizeMapping,
+           name_prefix: str,
+       ) -> nn.Module:
+       # return model
+
+       class _Mutator(nn.Mutator):
+           def __init__(self, config: NoQuantize, quant_map: QuantizeMapping) -> None:
+               super().__init__()
+               self.config = config
+               self.quant_map = quant_map
```

Results

Tencent AI
Platform Dept.

Category	Model	Release Date	Total Params	Training Tokens (B)	EdgeLLM-Public (6 datasets)	EdgeLLM-Eval (8 datasets)	Extended (12 datasets)
Tinytron Submission (November)	Track1: Llama3.1-8B-Instruct-Tinytron	2024/11/20	3826965504	324	57.01	55.41	58.61
	(Progressive pruning + distill+ self-generate data)		**Normalized (to uncompressed)**		82.97%	82.67%	84.28%
	Track1: Qwen2-7B-Instruct-Tinytron		3861627392	205	55.96	55.05	57.08
	(Soft pruning + distill + self-generate data)		**Normalized (to uncompressed)**		96.80%	88.42%	86.86%
	Track1: Phi-2-Tinytron-preview		2906924544	78	46.48	45.87	49.24
	(Distill across different vocabs, student use Qwen Vocab)		**Normalized (to uncompressed)**		104.31%	105.03%	108.84%
	Track1: Average score of three compressed models		3531839147	/	**53.15**	**52.11**	**54.98**
	(Track1 average, aligned with official ranking criteria)		**Normalized (to uncompressed)**		93.20%	90.38%	91.38%
	Track2: Cauchy-3B-preview		3110275328	221	**15.88**	**18.79**	**18.66**
Tinytron Baselines (October)	Llama3.1-8B-Instruct-Tinytron-preview	2024/10/20	4311419904	95	40.91	42.72	48.82
	(Distill baseline: no progressive pruning, no self-gen data)		**Normalized (to uncompressed)**		59.54%	63.72%	70.19%
	Qwen2-7B-Instruct-Tinytron-preview	2024/10/11	3861627392	137	44.28	46.15	51.25
	(Distill baseline: no self-generated data)		**Normalized (to uncompressed)**		76.60%	74.13%	78.00%
	phi-2	/	2779683840	/	44.56	43.67	45.24
	(Original uncompressed model)		**Normalized (to uncompressed)**		100.00%	100.00%	100.00%
	Track1: Average score of three compressed models	/	3650910379	/	**43.25**	**44.18**	**48.44**
	(Track1 average, aligned with official ranking criteria)		**Normalized (to uncompressed)**		75.84%	76.63%	80.50%
SOTA Distillation Methods	**Sparse-Llama-3.1-8B-ultrachat_200k-2of4 (2:4 Semi-structured sparse, layer-wise distillation)**	2024/11/25	8030261248	13	53.34	53.44	57.95
	Llama-3.2-1B-Instruct (Pruning 3.1 8B base as initialization + post training)	2024/9/25	1235814400	9000	38.99	40.07	43.35
	Llama-3.2-3B-Instruct (Pruning 3.1 8B base as initialization + post training)	2024/9/25	3212749824	9000	56.49	56.09	59.96