

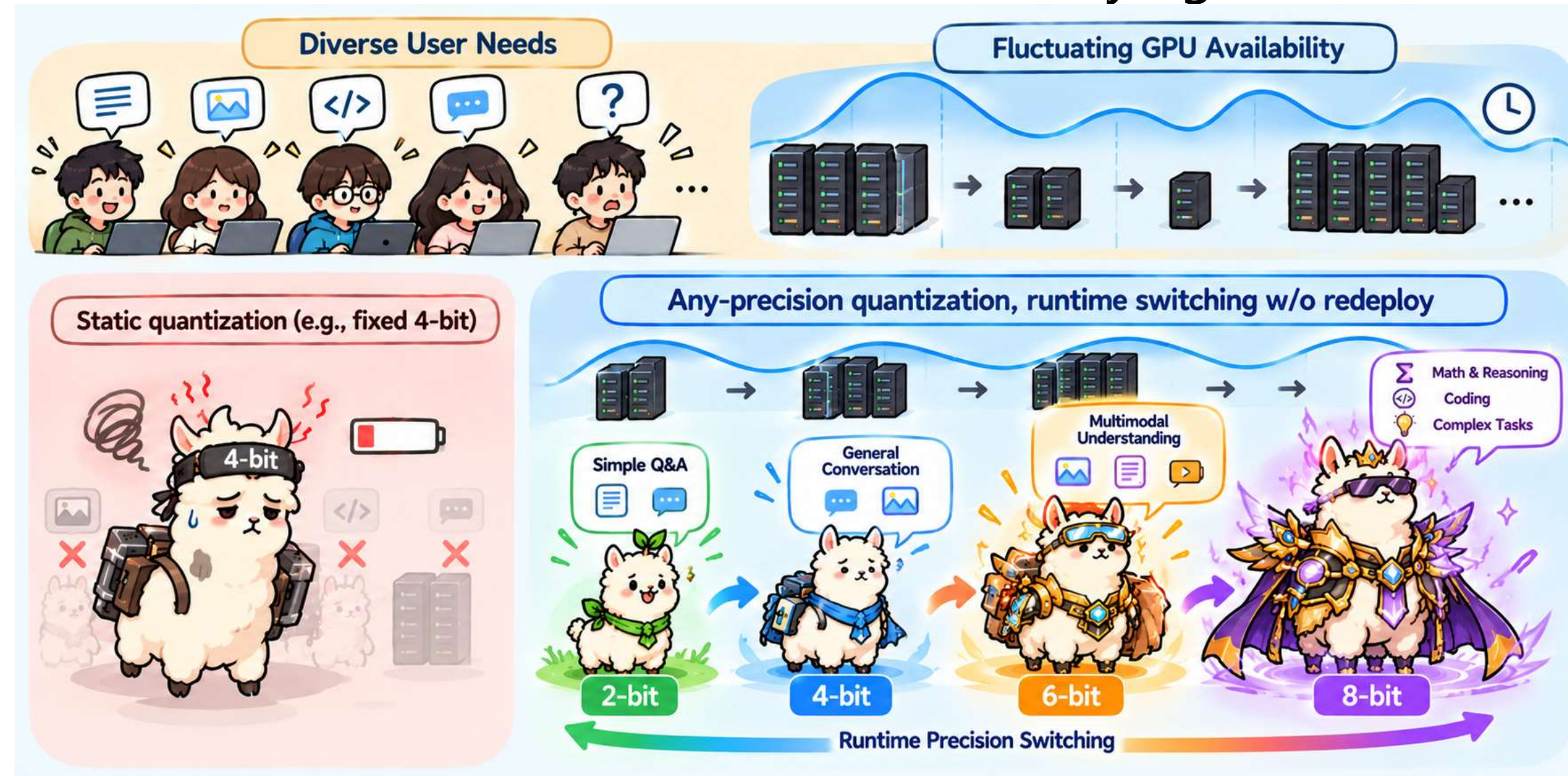
MoBiQuant: Mixture-of-Bits Quantization for Token-Adaptive Any-Precision LLM

Dongwei Wang*, Jinhee Kim*, Seokho Han*, Denis Gudovskiy, Yohei Nakata, Tomoyuki Okuno, KhayTze Peong, Kang Eun Jeon[†], Jong Hwan Ko[†], Yiran Chen[†], Huanrui Yang[†]

Project
Page
here



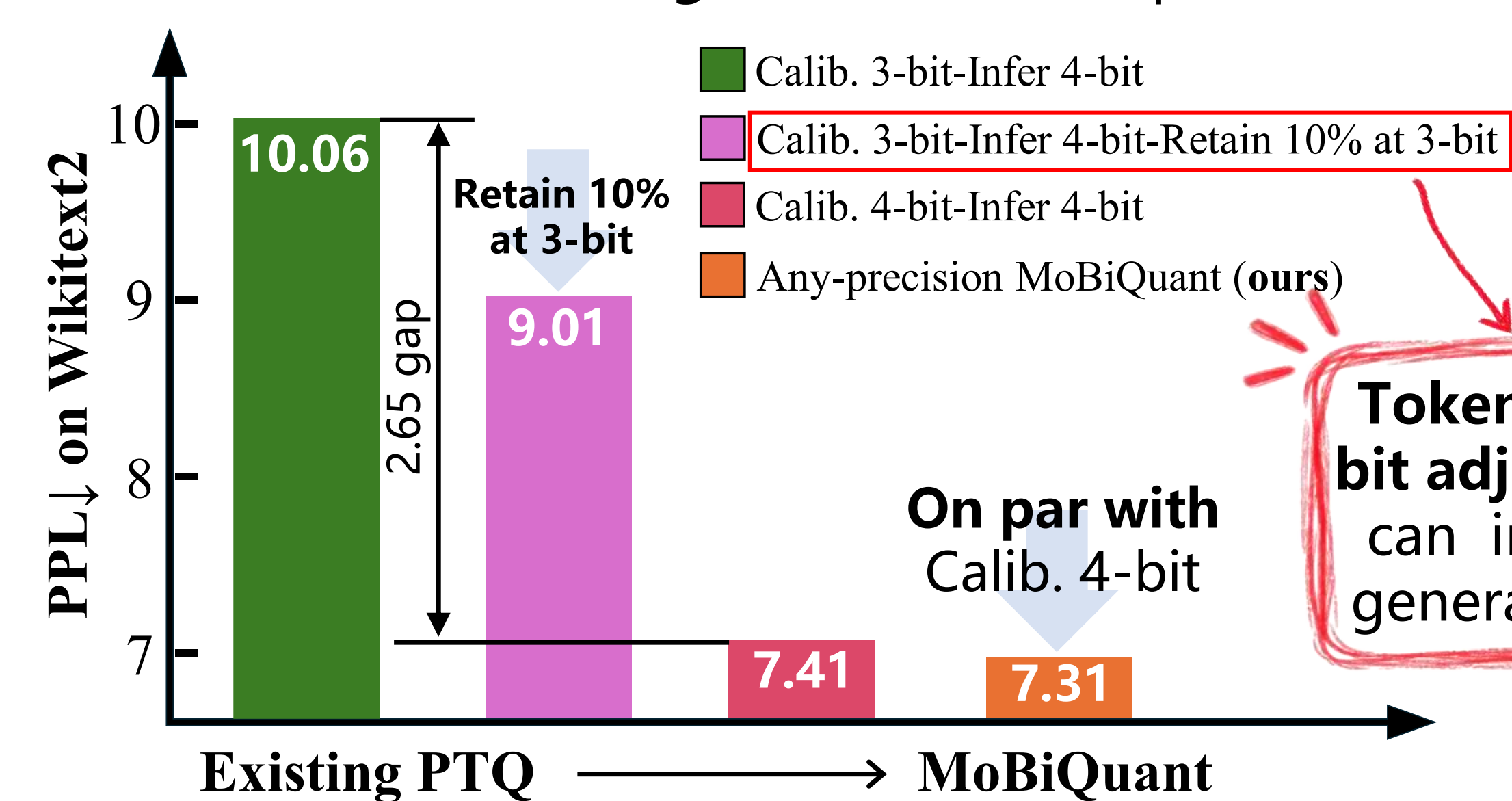
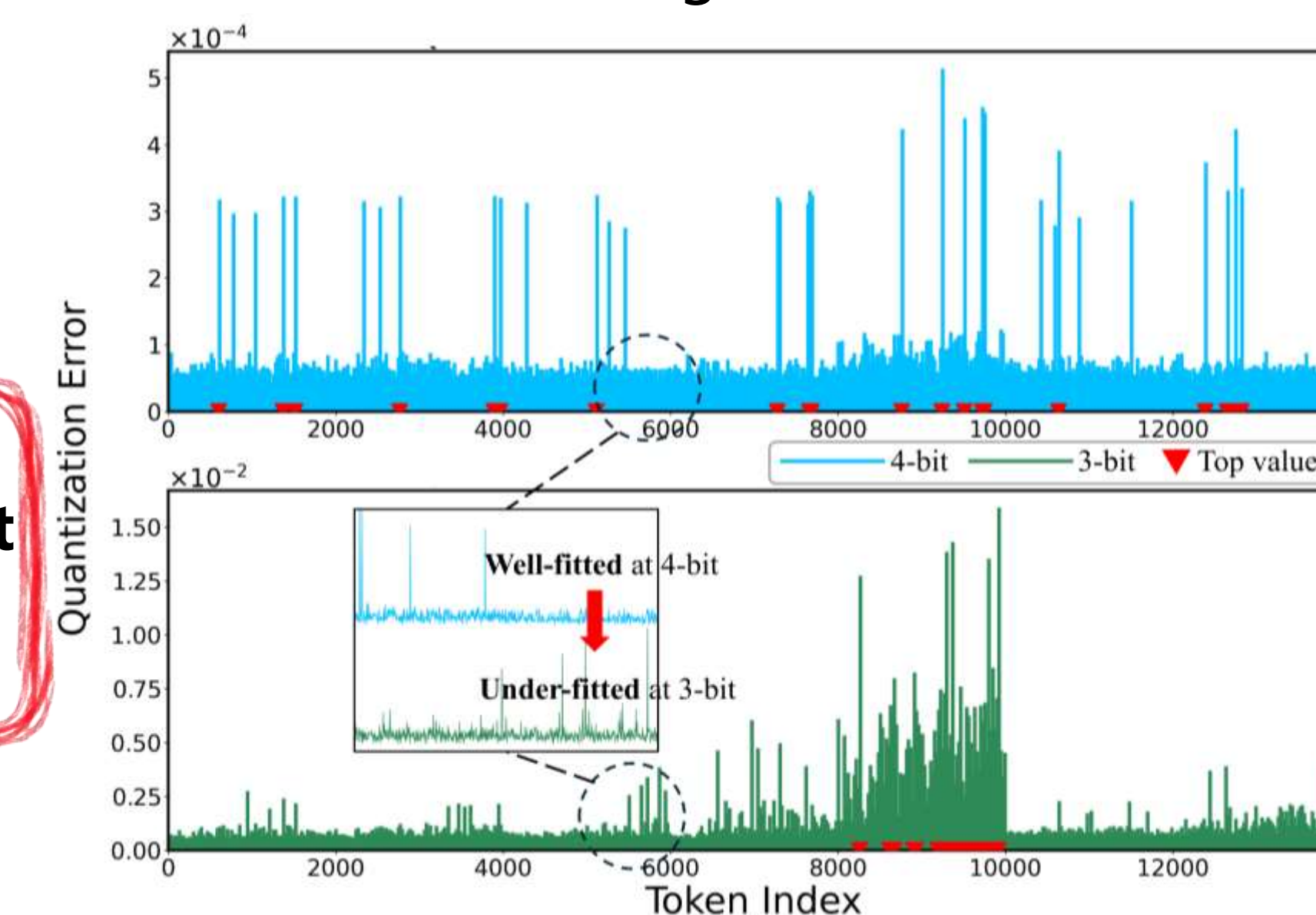
□ An ideal Quant(LLM) should be able to **adjust** its precision at **runtime** based on available GPUs and varying user demands.



Motivation of MoBiQuant

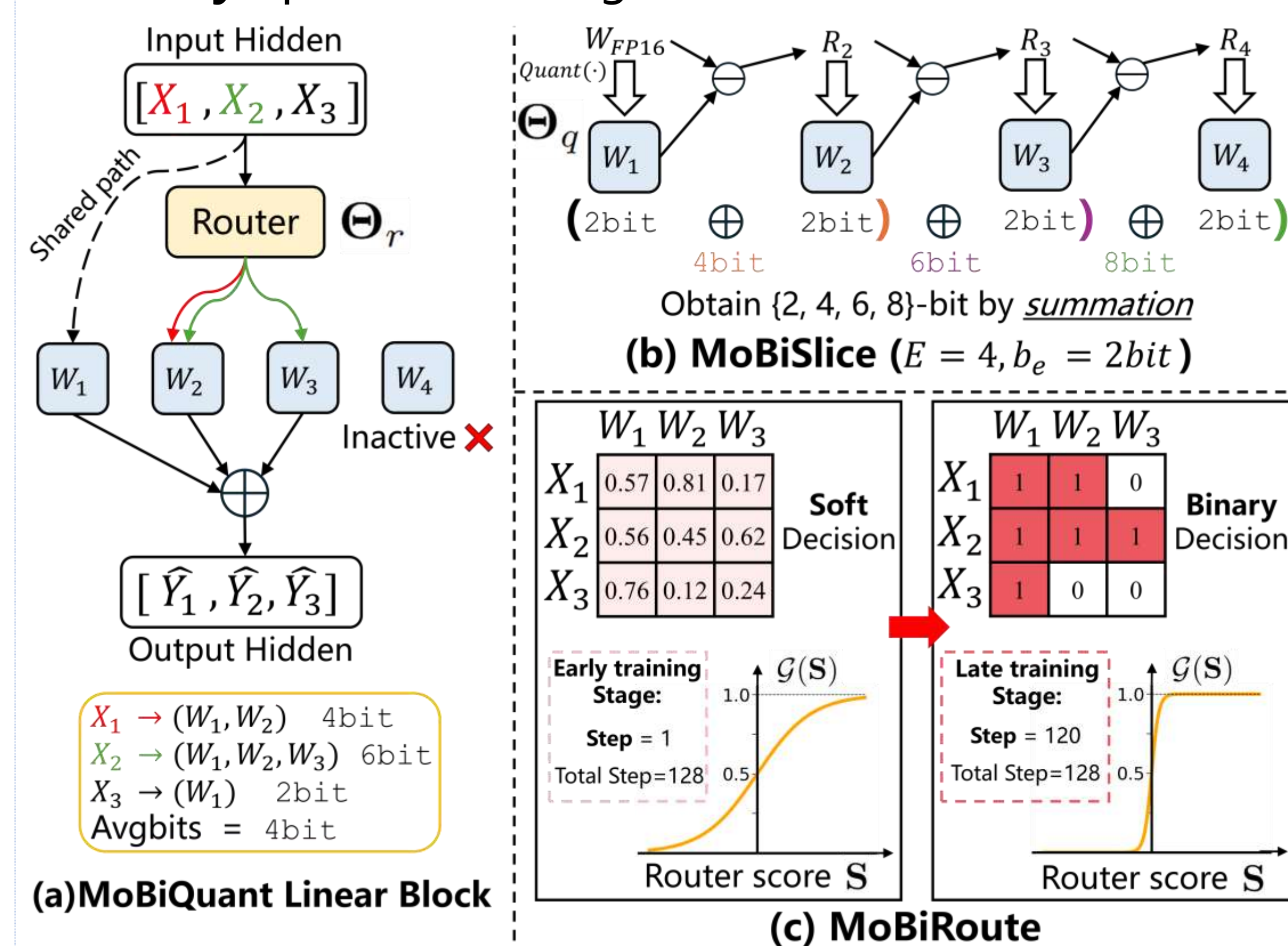
□ However, existing PTQ methods optimize for fixed bit-widths and **fail to generalize** across precisions.

□ We identify a bit-dependent **outlier migration** phenomenon that hinders generalization.



Overall Framework

- MoBiSlice: Multi-bit quant/dequant with **shared** scaling.
- MoBiRoute: Route tokens to different bit-slices.
- **Jointly** optimized during PTQ calibration.



Tailored Kernel Design

- Bit-major pack for **on-demand** fetching.
- Shared scaling enables **fast dequant**.
- Kernel fusion **reduces** routing overhead.

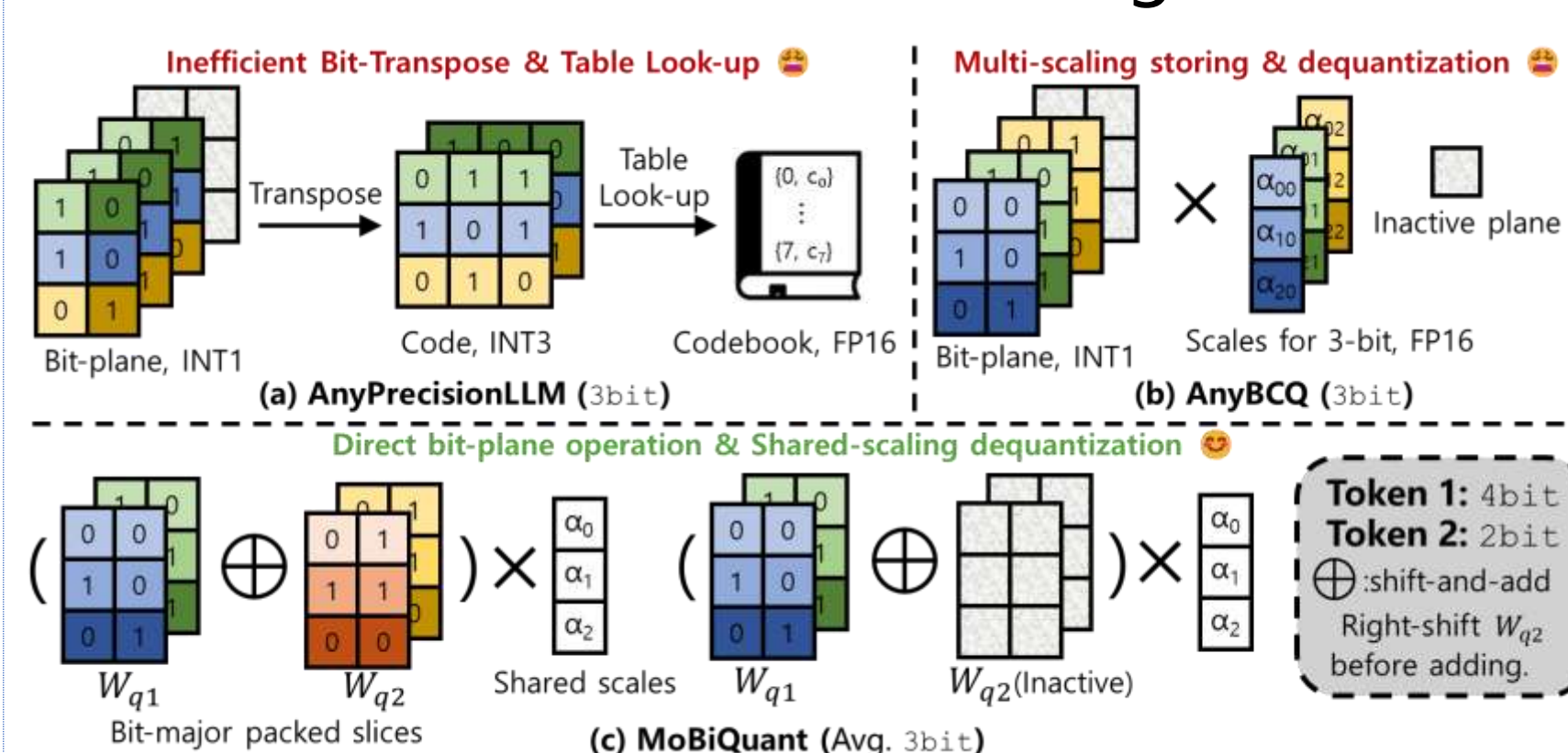
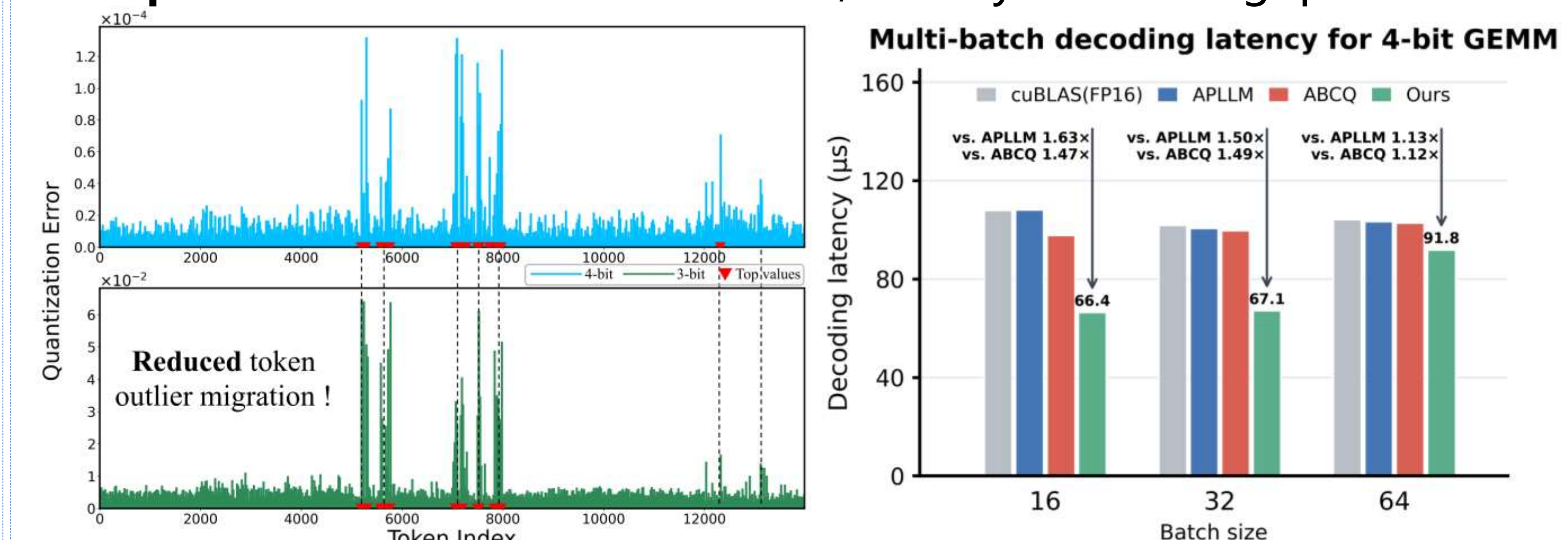


Table 1: DRAM read traffic across bit-widths.

Bit	Planes	DRAM Read	Δ /Plane
2	1	12.73 MB	—
4	2	24.05 MB	+11.32 MB
6	3	35.40 MB	+11.35 MB
8	4	46.68 MB	+11.28 MB

Highlight Experimental Results

- **Reduced outlier migration** across bit-widths.
- **Improved Cross-Bit Generalization**, latency and throughput.



Model	Bits	WikiText2, PPL ↓				MMLU, Acc. ↑				Throughput, tokens/sec. ↑		
		FP16	AP	ABC	MoBiQ	FP16	AP	ABC	MoBiQ	AP	ABC	MoBiQ
LLaMA2-7B	2	5.47	2e3	15.38	10.91	45.30	25.11	29.87	33.20	279	312	404
	3	5.47	6.13	6.25	6.07	45.30	41.85	41.56	43.35	241	268	311
	4	5.47	6.05	5.91	5.82	45.30	41.74	43.84	43.28	211	225	256
LLaMA3-8B	2	6.12	1e3	19.01	58.12	65.02	24.66	35.32	40.13	228	245	352
	3	6.12	8.60	8.08	7.97	65.02	55.53	58.28	59.71	194	212	257
	4	6.12	6.70	6.84	7.31	65.02	64.04	63.15	63.29	171	186	209
Gemma2-9B	2	6.84	19.62	12.72	12.21	70.74	33.09	43.36	44.28	163	185	266
	3	6.84	7.94	7.95	7.73	70.74	65.30	65.38	66.41	144	164	199
	4	6.84	7.06	7.23	7.07	70.74	69.23	68.81	69.08	125	146	164
Qwen3-32B	2	7.61	12.58	12.07	11.50	81.85	71.23	70.70	72.68	43	52	75
	3	7.61	8.93	8.42	8.14	81.85	78.95	79.72	80.43	40	46	56
	4	7.61	7.87	7.81	7.75	81.85	80.82	80.70	80.97	35	42	45