Setting (SQuAD), context embed	**Exact Match**	**F1**
Baseline — hard tokens	**0.58**	**0.71**
Hard embedded, avg ×2	0.09	0.21
Soft embedded online, avg ×2	0.05	0.11
Soft embedded regenerate-llm avg ×2	0.07	0.16

Table 1: Baseline against averaging techniques (Prompt–Q–C)

Model	**Compression**	**Exact Match**	**F1**
Mistral-7B (no FT)	×1	49	68
LoRA-FT baseline	×1	59	65
ICAE FT (PwC, authors)	$\times 1.7 \pm 0.7$	41	57
ICAE FT (SQuAD, ours)	$\times 1.7 \pm 0.7$	**69**	**73**

Table 2: ICAE averaging on SQuAD

Encoder	**Decoder**	**Accuracy**	**Average tool call time (s)**
—	Qwen-Full-FT	0.9497	0.5495
—	Qwen-LoRA-FT	0.9153	0.5387
—	Qwen	0.9000	0.5437
del long obs-s	Qwen	0.8911	—
del all obs-s	Qwen	0.8850	—
ICAE (Qwen-LoRA-PT w/ Qwen)	Qwen	0.8855	—
ICAE (Qwen-LoRA-FT)	Qwen	0.9089	**0.4880** (0.13 + 0.35)

Table 3: Qwen and ICAE future variants. FT=Fine-Tuning, PT=PreTraining

Encoder	Decoder	Accuracy	Average tool call time (s)
—	Qwen-Full-FT	0.9497	0.5495
—	Qwen-LoRA-FT	0.9153	0.5387
—	Qwen	0.9000	0.5437
del long obs-s	Qwen	0.8911	—
del all obs-s	Qwen	0.8850	—
ICAE (Qwen-LoRA-PT w/ Q-Full-FT)	Qwen-Full-FT	0.9219	—
ICAE (Qwen-LoRA-PT w/ Qwen)	Qwen	0.8855	—
ICAE (Qwen-LoRA-FT)	Qwen-Full-FT	—	—
ICAE (Qwen-LoRA-FT)	Qwen-LoRA-FT	—	—
ICAE (Qwen-LoRA-FT)	Qwen	0.9089	**0.4880** (0.13 + 0.35)

Table 4: Qwen and ICAE future variants. FT=Fine-Tuning, PT=PreTraining

Encoder	Decoder	Accuracy	Average tool call time (s)
—	Full-FT	0.9484	1.24
—	LoRA-FT	0.9118	1.24
—	Qwen	0.8967	1.23
del long obs-s	Qwen	0.8873	0.44
del all obs-s	Qwen	0.8802	0.39
ICAE (LoRA-PT w/ Full-FT)	Full-FT	0.9219	—
ICAE (LoRA-PT w/ Qwen)	Qwen	0.8808	**1.12 (0.31+0.81)**
ICAE (LoRA-FT)	Full-FT	?	—
ICAE (LoRA-FT)	LoRA-FT	0.9263	—
ICAE (LoRA-FT)	Qwen	0.9020	—

Table 5: No think bug table. Qwen and ICAE future variants. FT=Fine-Tuning, PT=PreTraining

Encoder	Decoder	Acc.	Time (s)	Resolved (/500)
—	Qwen-Full-FT	0.9484	1.24	—
—	Qwen-LoRA-FT	0.9118	1.24	10
—	Qwen	0.8967	1.23	26
ICAE (Qwen-LoRA-FT)	Qwen	0.9020	—	11
ICAE (Qwen-LoRA-FT)	Qwen-LoRA-FT	0.9263	—	3 (overfit?)
ICAE (Qwen-LoRA-FT)	Qwen-Full-FT	?	—	—
del long obs-s	Qwen	0.8873	0.44	1
del all obs-s	Qwen	0.8802	0.39	0
ICAE (Qwen-LoRA-PT w/ Q-Full-FT)	Qwen-Full-FT	0.9219	—	—
ICAE (Qwen-LoRA-PT w/ Qwen)	Qwen	0.8808	**1.12 (0.31+0.81)**	—

Table 6: No think bug table. Qwen and ICAE future variants. FT=Fine-Tuning, PT=PreTraining