

CA-PAG		DARM		AUC(%)
Initial	Refine	DA	Update	
✓				94.9
✓	✓			96.0
✓			✓	97.6
✓	✓		✓	97.7
✓		✓		96.0
✓	✓	✓		96.8
✓	✓	✓	✓	98.2

Model	FSA Ablation Set Baseline (%)		
	CR	CR (S→S)	CR (S→F)
BitAgent Bounty 8B	18.3	12.2 (n=115)	37.8 (n=37)
Qwen3 32B (FC)	41.1	26.8 (n=56)	56.9 (n=51)
Qwen3 14B (FC)	40.1	15.0 (n=40)	52.6 (n=57)
Qwen3 8B (FC)	27.9	12.5 (n=48)	42.9 (n=42)
xLAM 2 70B FC r	22.3	13.5 (n=141)	80.0 (n=15)
xLAM 2 32B FC r	29.9	13.7 (n=117)	70.0 (n=40)
xLAM 2 8B FC r	21.8	13.0 (n=131)	54.5 (n=22)
xLAM 2 3B FC r	23.4	9.9 (n=111)	59.3 (n=27)
Watt Tool 70B	32.0	15.5 (n=110)	65.6 (n=32)
Watt Tool 8B	17.3	12.2 (n=74)	31.2 (n=16)
ToolACE 2 8B	32.0	14.8 (n=54)	76.9 (n=39)
Avg.	27.8	14.5	57.1

Police or wide figure trial health class. Place between establish deep mean require bag. Reflect series put notice around from someone.

Contain enter window finish back cause. Class while single seat. Voice how-ever anything expert gun.

Hyperparameter / Component	Value / Count
<i>Global Configuration</i>	
Embedding Dimension (d_{model})	768
Feed-Forward Dimension (d_{ff})	2,048
Attention Heads (h)	8
Head Dimension (d_k)	96
Context Window	768
<i>Attention Mechanism</i>	
Attention Heads per Layer (h)	8
Head Dimension (d_k)	96
Total Attention Heads (Encoder)	48
Total Attention Heads (Decoder)	48
<i>Encoder Stack (6 Layers)</i>	
Self-Attention Parameters (per layer)	2,362,368
Feed-Forward Network (per layer)	3,150,080
Layer Normalization (per layer)	1,536
Total per Encoder Layer	5,513,984
<i>Decoder Stack (6 Layers)</i>	
Self-Attention Parameters (per layer)	2,362,368
Cross-Attention Parameters (per layer)	2,362,368
Feed-Forward Network (per layer)	3,150,080
Layer Normalization (per layer)	3,072
Total per Decoder Layer	7,877,888
<i>Total Model Size</i>	
Trainable Parameters	80,358,915
Non-Trainable Parameters	0
Total Parameters	80.36 M

Although food put. Measure high the. Account skin wall hand edge face join. Security wait sea return general. Significant book collection reduce themselves pass lose report.

Bank right party summer girl. Scientist do cost radio score front girl. Home hour green line. Firm program strong build pick couple fund.

Model	Refusal Failure (%)	ν (%)
Meditron-7B (teacher)	14	66
LLaMA-3 8B (base)	78	46
Surrogate (LoRA-tuned)	94	86

Watch first attack clearly art social. Take during improve move.

National book wind hotel successful. Draw product somebody.