Additional File 7

Additional Results for Stage 3 Instruction Tuning on Smaller Open Models

Table S7.1. Additional lexical evaluation for Task C timeline summarization.

Metric	Model	Silver→ Silver Zero	Silver→ Silver Few	Silver→ Silver IT	Silver→ Gold Zero	Silver→ Gold Few	Silver→ Gold IT	Gold Zero	Gold Few	Gold IT
ROUGE-1	Gemma4-31B	.826±.008	.823±.008	.886±.006	.805±.011	.794±.013	.865±.013	–	.743±.026	.729±.030
	Qwen3.6-35B-A3B	.749±.010	.791±.009	.879±.007	.717±.014	.754±.016	.867±.011	–	.670±.036	.656±.037
	Qwen3-30B-A3B	.737±.013	.750±.013	.843±.012	.713±.014	.765±.012	.811±.018	–	.136±.040	.296±.041
	Mistral-24B	.736±.011	.761±.011	.878±.007	.679±.021	.724±.017	.864±.011	–	.698±.022	.763±.027
	Meditron-7B	.555±.021	.595±.022	.804±.017	.471±.030	.531±.031	.802±.021	–	.178±.043	.230±.031
ROUGE-2	Gemma4-31B	.711±.010	.693±.011	.804±.009	.675±.013	.649±.017	.776±.015	–	.607±.024	.629±.028
	Qwen3.6-35B-A3B	.621±.011	.671±.011	.797±.009	.575±.016	.622±.018	.782±.012	–	.555±.032	.551±.034
	Qwen3-30B-A3B	.598±.012	.617±.013	.748±.013	.567±.015	.613±.015	.708±.019	–	.112±.033	.229±.034
	Mistral-24B	.600±.012	.640±.013	.795±.009	.540±.020	.597±.019	.778±.013	–	.564±.021	.655±.026
	Meditron-7B	.422±.019	.463±.020	.709±.018	.336±.026	.399±.027	.696±.022	–	.142±.035	.141±.025

We report ROUGE-1 and ROUGE-2 F1 under Silver→Silver, Silver→Gold, and Gold. Results are mean ± bootstrap SD over 2,000 document-level resamples. Higher is better.