Table 1: **Model Performance on Argument Detection and Classification Tasks** - 10,000 examples in the training set, 1000 validation examples, validation scores reported. outputs were constrained to "Topicality", "Disadvantages", "Advantages" or "Counterplans". We use the same fine-tuning settings for LoRA, ReFT and Orthogonalization as in the original experiments. Results averaged over 3 runs, with error reported in parenthesis

Model	Technique	Accuracy (%)	F1 Score	Precision	Recall
Mistral-7B	Base	72.8(5)	0.70(4)	0.68(3)	0.69(3)
Mistral-7B	LoRA	78.2(4)	0.76(3)	0.75(4)	0.76(4)
Mistral-7B	ReFT	77.9(6)	0.75(2)	0.74(3)	0.75(3)
Mistral-7B	Orthogonal	73.1(6)	0.71(3)	0.70(3)	0.71(3)
LLaMA3-8B	Base	74.5(6)	0.72(3)	0.70(2)	0.71(3)
LLaMA3-8B	LoRA	81.2(3)	0.79(2)	0.78(3)	0.79(2)
LLaMA3-8B	ReFT	80.8(5)	0.78(3)	0.77(2)	0.78(3)
LLaMA3-8B	Orthogonal	75.2(7)	0.73(4)	0.71(3)	0.72(4)

Table 2: **Performance of Models on Stance Detection Tasks** - 10,000 examples in the training set, 1000 validation examples, validation scores reported. Outputs were constrained to either "aff" or "neg"

Model	Technique	Accuracy (%)	Precision	Recall	F1-Score
Mistral-7B	Base	77.3(6)	75.8(7)	76.2(5)	76.0(4)
Mistral-7B	LoRA	82.4(5)	81.2(6)	81.7(4)	81.4(5)
Mistral-7B	ReFT	81.9(4)	80.7(5)	81.1(6)	80.9(5)
Mistral-7B	Orthogonal	78.8(5)	77.2(6)	77.5(4)	77.3(4)
LLaMA3-8B	Base	79.5(7)	77.9(8)	78.2(7)	78.0(6)
LLaMA3-8B	LoRA	85.3(8)	83.8(7)	84.1(5)	83.9(6)
LLaMA3-8B	ReFT	84.1(6)	82.5(7)	83.0(6)	82.7(5)
LLaMA3-8B	Orthogonal	80.5(7)	79.2(6)	79.5(6)	79.3(5)
LLaMA3-8B	LoRA (1M Ex)	87.1(9)	85.4(8)	85.8(7)	85.6(6)

Table 3: **Performance of Pretrained and Non-Pretrained Models on Downstream Tasks** - The entire dataset is used for both pretraining and fine-tuning. ArgAnalysis35K is an argument quality classification task

Model	Dataset	PT	R-1	R-2	R-L	PPL	LJ-Q	LJ-S	Steps
Mistral-7B	DebateSum	No	26.3(4)	7.5(3)	23.1(6)	130.5(42)	7.3(3)	7.0(3)	60k
Mistral-7B	DebateSum	Yes	30.4(14)	9.4(9)	26.5(11)	19.8(17)	7.8(3)	7.6(3)	40k
LLaMA3-8B	DebateSum	No	24.2(5)	6.9(3)	21.9(8)	95.7(45)	7.0(3)	6.7(3)	60k
LLaMA3-8B	DebateSum	Yes	32.2(15)	9.9(10)	27.4(12)	21.8(25)	8.0(2)	7.8(2)	35k
Mistral-7B	ArgAnalysis35K	No	44.8(3)	21.2(5)	40.5(8)	25.2(11)	7.2(3)	7.0(3)	50k
Mistral-7B	ArgAnalysis35K	Yes	48.2(15)	24.9(10)	43.4(12)	21.8(5)	7.8(2)	7.6(2)	35k
LLaMA3-8B	ArgAnalysis35K	No	42.4(6)	19.6(2)	38.8(13)	27.3(13)	7.0(3)	6.8(3)	50k
LLaMA3-8B	ArgAnalysis35K	Yes	48.9(17)	25.1(12)	44.2(14)	22.7(7)	8.0(3)	7.8(3)	30k
Mistral-7B	Multi-LexSum	No	40.1(5)	20.3(4)	37.7(9)	30.1(23)	7.1(2)	7.0(2)	65k
Mistral-7B	Multi-LexSum	Yes	45.2(12)	22.5(7)	40.8(10)	22.0(11)	7.7(3)	7.5(3)	45k
LLaMA3-8B	Multi-LexSum	No	38.2(8)	19.5(3)	36.0(10)	32.5(25)	7.0(2)	6.8(2)	65k
LLaMA3-8B	Multi-LexSum	Yes	46.8(14)	23.0(8)	41.7(13)	23.9(14)	7.8(3)	7.6(3)	40k

Table 4: **Inter-Annotator Agreement: Human vs. GPT-4 Ratings on Summaries** - 50 expert debaters rated 6 summaries each

ID	Format	Human-AQ	GPT4-AQ	Human-SQ	GPT4-SQ
S1	Policy	8.2	8.0	8.0	7.9
S2	Lincoln-Douglas	7.6	7.8	7.4	7.6
S3	Public Forum	7.2	7.4	7.1	7.3
S4	Policy	8.5	8.3	8.3	8.2
S5	Lincoln-Douglas	7.8	7.7	7.5	7.6
S6	Public Forum	7.4	7.2	7.2	7.1

Table 5: Statistical Results

Metric	Pearson
Argument Quality	0.78
Support Quality	0.74