Baseline vs. Fine-Tuned XLM-R Comparison

Performance on Test Set (67 examples)

Model	Accuracy	F1
Baseline XLM-R (mean, text_only)	68.66%	0.686
Fine-tuned XLM-R (context_text)	71.64%	0.707

Fine-tuning improves test accuracy by ~3 percentage points.

Performance on Gold Set (28 examples)

Model	Accuracy	F1
Baseline XLM-R (mean, text_only)	82.14%	0.816
Fine-tuned XLM-R (context_text)	60.71%	0.594

Fine-tuning drops gold accuracy by ~21 percentage points.

Error Types on Gold (Fine-Tuned)

- False Negatives: 10 (missed sarcasm)
- False Positives: 1 (false sarcasm)
- The model mostly fails by missing sarcastic examples

Key Differences

- Baseline used text_only input; fine-tuned used context_text
- Baseline used frozen weights; fine-tuned updated all weights
- Baseline performed better on gold; fine-tuned performed better on test