REPORT: TEST_N2_M2_MULTICLASS

Model: Qwen2.5-VL-M2-Classification

Date: 2026-02-17 | Dataset: M2 (Multiclass 1-4)

1. PERFORMANCE SUMMARY (Macro Avg)

Metric	Baseline	Fine-Tuned (Avg 4)
Accuracy	0.5118	0.6943 ± 0.0113
F1 Macro	0.3552	0.3389 ± 0.0158
Prec Macro	0.3620	0.3724 ± 0.0285
Rec Macro	0.4551	0.3296 ± 0.0123
Inv. Rate	12.32%	0.00%
Infer Time (s)	0.405	0.705
Train Time (m)	N/A	150.3 ± 0.3

2. PER-CLASS BREAKDOWN (Baseline vs Fine-Tuned)

Metric	Class	Baseline	Fine-Tuned (Avg ± Std)
Precision	1 (Flaming)	0.1739	0.2217 ± 0.1035
	2 (Denig.)	0.7921	0.7440 ± 0.0049
	3 (Sexual)	0.4318	0.5237 ± 0.0120
	4 (Racism)	0.0500	0.0000 ± 0.0000
Recall	1 (Flaming)	0.6667	0.1250 ± 0.0481
	2 (Denig.)	0.5405	0.8834 ± 0.0186
	3 (Sexual)	0.4130	0.3098 ± 0.0274
	4 (Racism)	0.2000	0.0000 ± 0.0000
F1-Score	1 (Flaming)	0.2759	0.1595 ± 0.0659
	2 (Denig.)	0.6426	0.8077 ± 0.0085
	3 (Sexual)	0.4222	0.3885 ± 0.0189
	4 (Racism)	0.0800	0.0000 ± 0.0000

3. CONFUSION MATRICES (Avg)