

Comparison Report: Stock vs MaxWarpC (No OutlierGuard)

This report compares the baseline GPT-2 model (Stock) against the MaxWarpC configuration (tap -9, warp+detect+denoise enabled, no outlier suppression). The MaxWarpC run used significantly stronger warp parameters to test whether aggressive basin reshaping improves performance.

Metric	Stock Baseline	MaxWarpC noOutlier	Δ (Change)
Top-1 Accuracy	0.324	0.355	+0.031
Top-2 Accuracy	0.610	0.621	+0.011
Macro F1	0.323	0.355	+0.032
ECE (Calibration)	0.112	0.080	-0.032 (better)
Brier Score	0.750	0.743	-0.007
Overconfidence >0.70	0.7%	1.2%	slight ↑

Interpretation:

- MaxWarpC achieved a +3.2 point lift in Macro F1 (0.323 → 0.355) and +3.1 point in Top-1 accuracy.
- Calibration improved significantly (ECE 0.112 → 0.080).
- Gains were broad across classes, with better stability and recall.
- Slight increase in moderate overconfidence, but still very low overall.

Conclusion: MaxWarpC (no outlier suppression) outperformed Stock across accuracy, F1, and calibration. This represents a validated +3 F1 point success, demonstrating that aggressive warp can yield measurable downstream improvements.