Comparison of F1 Test Scores per User and Baseline

Legend:
- ● User Mean
- ✖ Baseline (majority voting)

Datasets (x-axis): ttcw, slm, hanna, confederacy, pronvsprompt

Y-axis: F1 Test Score (0.00 to 1.00)