Stage 3 Performance by Language (Multilingual Model)

Legend:
- English (4,130 pairs)
- Chinese (3,434 pairs)

Annotation: Chinese outperforms English in the multilingual model

Metric	English	Chinese
Accuracy	24.6%	29.6%
Macro F1	13.9%	16.6%
Weighted F1	15.2%	21.7%

Y-axis: Score (%)