

Aggregate Quantitative Comparison Between Baseline and AIMS
Across Lexical, Semantic, and Factual Evaluation Metrics

