ML Flow Default Readability Metrics:

ari_grade_level():

- Estimates the grade level needed to understand the text.
- Target: Lower values (e.g. <= 8) are preferred for general readability, but target mostly dependent on readability of input text.

flesch_kincaid_grade_level():

- Estimates the grade level needed to understand the text.
- Target: Lower values (e.g. <= 8) are preferred for general readability, but target mostly dependent on readability of input text.

ML Flow Default Accuracy Metrics

exact_match()

- Measures how often the corrected output exactly matches the gold standard.
- Target: Values close to 1.0 (0.9+).

rouge1()

- Measures unigram (word-level) overlap between prediction and target.
- Target: Values close to 1.0 (0.9+).

rougeL()

- Measures the longest common subsequence between prediction and target.
- Values close to 1.0 (0.9+).

Custom Accuracy Metrics

semantic_similarity_metric

- Measures meaning similarity using TF-IDF and cosine similarity.
- Target: Values close to 1.0 (0.9+). Indicates perfect semantic alignment with gold standard text.

Custom Grammar and Correction Metrics

grammar_error_count_metric

- Average number of grammar errors in model output.
- Target: 0 – no grammar mistakes.

grammar_error_rate_metric

- Grammar errors per word – normalizes error count by output length.
- Target: 0 – no errors per word.

grammar_improvement_metric

- Measures reduction in errors from input to output (positive = improved).
- Target: Higher positive values – model is fixing more than it breaks

grammar_score_metric

- Heuristic grammar score out of 100; penalizes errors.
- Target: 100 – no errors detected.

Custom Style and ClarityMetrics

readability_improvement_metric

- Measures improvement in sentence simplicity and structure.
- Target: positive value – model makes text more readable.

Custom GPT-Based Judgment

llm_judge_metric

- LLM assigns a score from 0-10 based on correction quality (fixes, style, meaning).
- Target: 10 – perfect correction with no issues.