

Metric

Cosine distance from reference embedding

Semantic similarity score

Entailment score

LLM-as-judge meaning preservation

What it captures

Overall meaning divergence

Nuanced meaning preservation

Logical consistency with original intent

Full meaning assessment

Alert threshold

> 0.15 from baseline

< 0.80 on 0-1 scale

< 0.85 on entailment probability

< 4.0 on 1-5 scale

Tooling

Sentence transformers, vector stores

Sentence-BERT, cross-encoders

NLI models (e.g., DeBERTa-MNLI)

Any frontier LLM with a defined scoring rubric