AUDIT REPORT

Contamination assessment

A review of overlap between training data and evaluation examples.

CLEAN

0.0%

0 / 400 evaluation examples flagged

2,000	400	0
Training examples	Evaluation examples	Matching pairs

Audit configuration

Methods	semantic
Benchmarks	mmlu, gsm8k
Similarity threshold	0.85
Generated (UTC)	2026-11-04 06:59

The rate counts unique evaluation examples, not matching pairs. Similarity identifies potential leakage and should be reviewed in context.

Match index

No matches met the detection criteria.