

Scope

How to assess model capabilities.

Metrics

Perplexity, factuality, toxicity.

Benchmarks

MMLU, HellaSwag, TruthfulQA.

Limitations

Overfitting, brittleness, hallucinations.