# Translation Evaluation Methodology

## Overview

This methodology evaluates machine translations from LLM (GPT-4) and Google Translate against human reference translations using two complementary metrics.

## Evaluation Metrics

### 1. METEOR Score

- Evaluates exact matches, stems, synonyms, and paraphrases
- Range: 0-1 (higher is better)
- Better correlates with human judgments
- Recognizes valid paraphrases and handles word order variations

### 2. BLEU Score

- Industry standard focusing on n-gram precision
- Compares phrase overlaps between candidate and reference translations
- Range: 0-1 (higher is better)
- Penalizes incorrect length translations

## Process

1. **Data Preparation**: UTF-8-sig CSV with 'english' source and 'label' reference columns
2. **Translation Generation**: Create LLM and Google translations while preserving placeholders
3. **Metric Calculation**: Compute METEOR and BLEU scores against references
4. **Statistical Analysis**: Calculate means and standard deviations

## Interpretation

- METEOR better evaluates meaning preservation
- BLEU better measures phrase accuracy and fluency
- Using both provides comprehensive quality assessment

## Limitations

- Metrics approximate human judgment
- Multiple valid translations may exist
- Automated metrics can't fully capture nuanced language
- Results affected by placeholders, domain terminology, and reference quality