# METHODOLOGY

## SO2 - Assess Translation Performance Across Leading LLMs and NMT Models

### 1. Dataset Creation

A stratified sampling strategy was employed on the corpus previously created via SO1:

| domain | direction | sentence_type_fine | available_rows | sampled_rows |
|---|---|---|---:|---:|
| bible | en->fj | statement | 1532 | 20 |
| bible | fj->en | statement | 1532 | 20 |
| conversational | en->fj | greeting | 5 | 5 |
| conversational | en->fj | question | 16 | 16 |
| conversational | en->fj | statement | 112 | 20 |
| conversational | fj->en | question | 17 | 17 |
| conversational | fj->en | statement | 115 | 20 |
| definition | en->fj | example | 93 | 20 |
| definition | en->fj | headword | 770 | 20 |
| definition | fj->en | example | 46 | 20 |
| definition | fj->en | headword | 817 | 20 |
| dictionary | fj->en | statement | 8959 | 20 |
| idiom | fj->en | idiom_literal | 126 | 20 |
| legal | en->fj | definition | 5 | 5 |
| legal | en->fj | legal_clause | 560 | 20 |
| legal | en->fj | obligation | 136 | 20 |
| legal | en->fj | right | 31 | 20 |
| legal | fj->en | definition | 6 | 6 |
| legal | fj->en | legal_clause | 685 | 20 |
| legal | fj->en | right | 41 | 20 |
| medical | en->fj | information | 76 | 20 |
| medical | en->fj | instruction | 9 | 9 |
| medical | en->fj | prevention | 2 | 2 |
| medical | en->fj | symptom | 37 | 20 |
| medical | en->fj | treatment | 4 | 4 |
| medical | en->fj | warning | 3 | 3 |
| medical | fj->en | information | 127 | 20 |
| medical | fj->en | instruction | 6 | 6 |
| medical | fj->en | prevention | 2 | 2 |
| medical | fj->en | symptom | 40 | 20 |
| medical | fj->en | warning | 2 | 2 |

These translations were verified by multiple human annotators

### 2. Translation Execution

Each selected sentence was translated using various machine translation systems.

### 2.1 Large Language Models (LLMs)

LLMs (prominent models: GPT-5.2 and Gemini 1.5 Pro) were prompted using the format:

*"Translate the following sentence to [target language]: [sentence]"*

### 2.2 Neural Machine Translation Systems (NMTs)

Translation was performed using APIs and open-source toolkits such as Google & Microsoft Translate. Preprocessing and tokenization steps were standardized across systems where applicable.

## 3. Automatic Evaluation

Machine-generated outputs were compared to human reference translations using standard automatic evaluation metrics:

- BLEU
- CHRF++
- TER

These metrics provided quantitative assessments of translation accuracy and fluency.

## 4. Human Evaluation

Where feasible, bilingual speakers assessed the translations. Each output was rated based on:

- **Fluency** (grammatical correctness and naturalness) on a 1–5 scale
- **Adequacy** (faithfulness to the source meaning) on a 1–5 scale
- **Cohesion/Discourse** (for long texts), evaluated through qualitative feedback or an extended scale

Multiple evaluators were used to ensure consistency and reduce subjective bias.

## 5. Performance Quantification

For each model or system, the following performance indicators were calculated:

- Average BLEU, CHRF++, and/or COMET scores
- Mean human evaluation scores

## 6. Reporting

Results were compiled into tables and visualizations to highlight:

- Comparative performance across systems
- Specific strengths and weaknesses
- Observations in domain-specific or low-resource contexts
- Recommendations or future research