Translating with LLMs

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

November 2025

1 Objectives and context

In this practical, you will evaluate the translation capabilities of local LLMs accessed via `Ollama`. The focus is on *contextual* machine translation : translating a sentence given its surrounding discourse. You will :

- Construct a small contextual test set from a TED talk translation dataset.
- Translate it with *two* different LLMs served locally through `Ollama`.
- Evaluate translation quality using the COMET metric.
- Compare three prompting strategies :
  1. A direct translation prompt.
  2. A chain-of-thought style prompt (difficulty assessment, translation, post-editing).
  3. An *n*-best generation prompt that produces 5 candidate translations, followed by automatic or LLM-based selection of the best candidate.
- Compare the impact of (i) the model, (ii) the prompting strategy, and (iii) the selection method.

2 Work to do

Building a Test Set The first step consists in constructing a contextual test set using the TED Talk data available on Hugging Face 🤗 for the language pair of your choice. Because the aim of this practical is to evaluate translation performance *in context*, it is essential to exploit the available metadata (typically timestamps) to reconstruct the discourse structure rather than treating the dataset as a collection of isolated sentences. You are free to decide how many examples you can reasonably extract and how much surrounding context to include for each segment. In your report, you should discuss the various trade-offs involved in these choices, such as dataset size, context window length, computational constraints, and the impact of these decisions on the quality and representativeness of your evaluation.

Prompting strategy We will experiment with several prompting strategies. The first is a simple, direct prompt that asks the model to produce a translation of either a single sentence or a sentence accompanied by its context. The second is a CoT style prompt in which the translation process is explicitly decomposed into three steps : (i) identifying potential translation difficulties, (ii) producing a context-aware (or context-free) translation, and (iii) revising the output to correct any errors. The third strategy is a sampling-based approach : starting from the direct prompt, the model is instructed to generate five distinct candidate translations using different sampling probabilities. We then apply two alternative methods to select the best translation from these candidates : either using an external confidence-based metric such as COMET, or asking an LLM directly to choose the most appropriate translation.

There now exist dedicated tools for implementing prompt "chains", which allow you to structure multi-step prompting workflows without manually stitching together sequences of string concatenations. Such frameworks make it easier to design, organise, and reuse complex prompts—particularly those involving multi-stage reasoning or iterative refinement, as required in this lab. Becoming familiar with these tools is now considered an essential part of working effectively with modern LLM pipelines, and you are encouraged to draw on them where appropriate in your implementation.

Models I strongly recommend using small models with `Ollama`, as a careful implementation of the various inference algorithms allows relatively large models to run on personal computers. In particular, you may consider :

— `mistral-small`

— `llama3.2`

— `eurollm` (available in various sizes and with different levels of quantisation)

Make sure to keep in mind the points discussed in class (instruction fine-tuning, controlled output, etc.).

Expected Work In your report, you should provide a precise description of how you constructed your test dataset, detailing the selection process, the use of context, and any preprocessing decisions. You must also present the full set of prompts you experimented with, explaining their design and intended effects. Furthermore, your report should describe how you implemented the translation pipeline used to process the test set, including any tooling or automation frameworks employed. You should conclude by comparing and analysing the different prompting strategies, highlighting their relative strengths, weaknesses, and the patterns you observed in the resulting translations.