Learned Entity Alignment Prompting (LEAP) for Consistent Jargon Translation

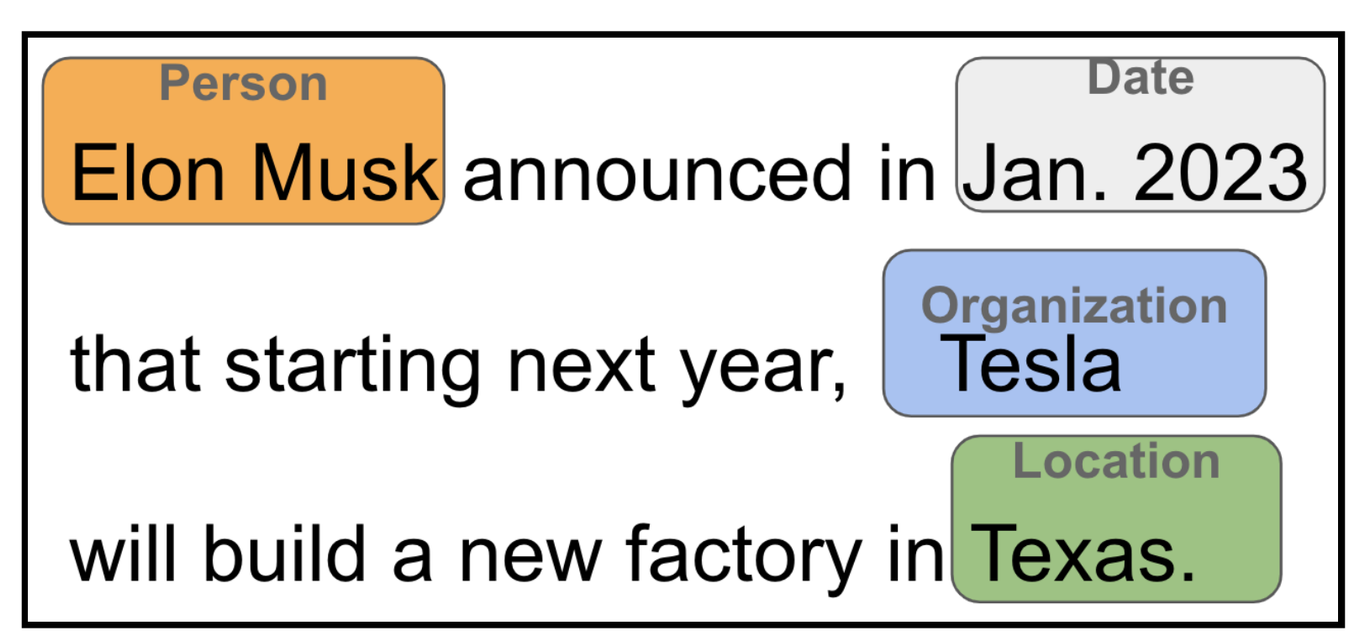
MIT 6.861
Jared Ni, Sibi Raja, Aditi Raju

Research Question

How can LLMs maintain consistent translations for technical jargon?

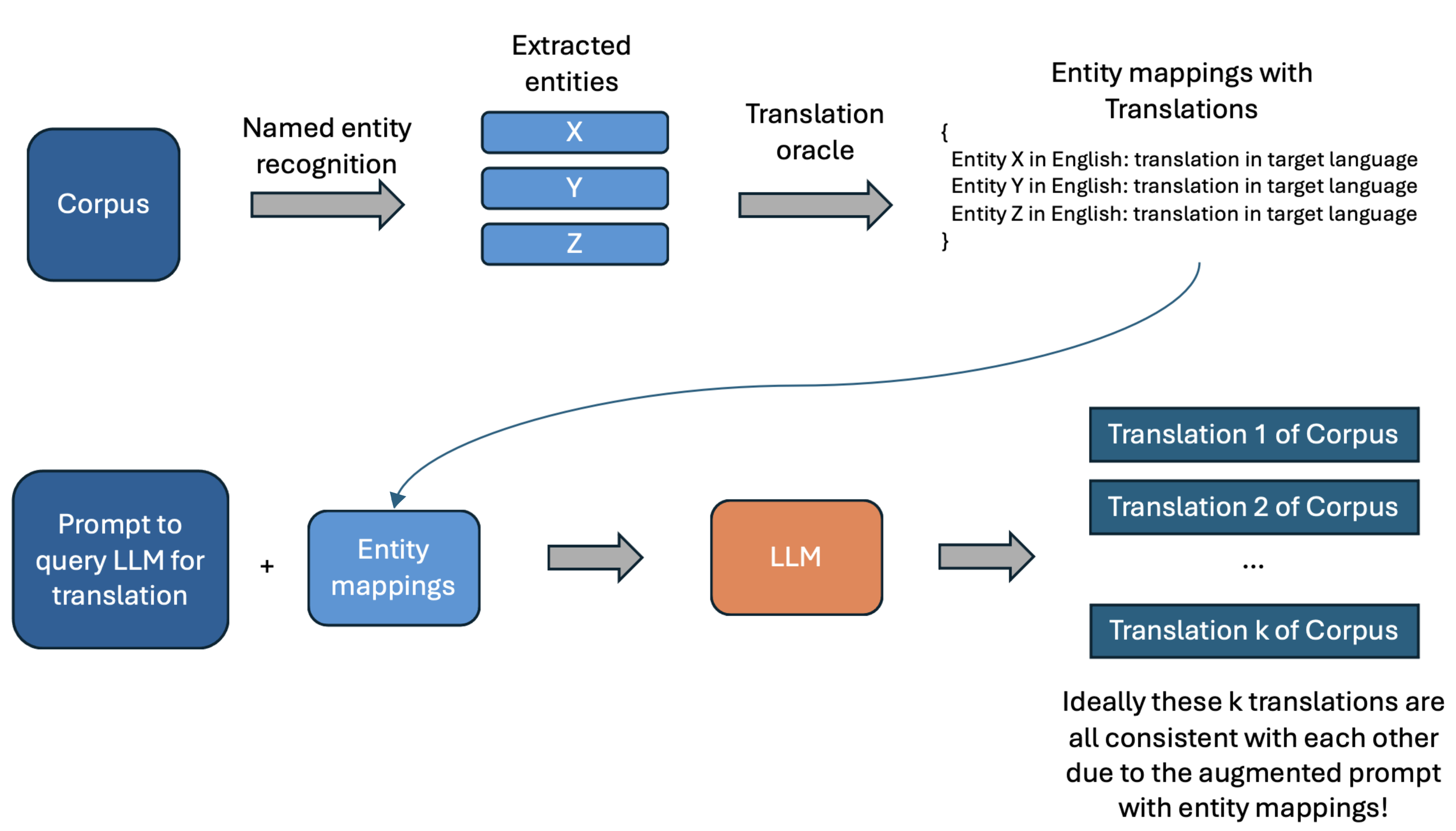
Motivation

Maintaining translation consistency—particularly for proper nouns and domain-specific terminology (jargon)—is a significant challenge in machine translation. Inconsistent translations, especially in fields that require high language accuracy (such as law and medicine), can lead to confusion, lack of clarity, and incorrect interpretations. Recent work has shown to achieve more consistent translations at the cost of large memory and retrieval overhead. We aim to take a different approach driven by **Named Entity Recognition (NER)**.



Named Entity Recognition example

Methodology



LEAP workflow

Corpus (Datasets):

- Casehold: law dataset of ~53,000 legal case contexts
- PMC Patients: medical dataset of ~167,000 patient summaries extracted from case reports in PubMed Central (PMC)
- We test 500 examples from each dataset on every model and language
- We truncate each piece of text to the first 50 words

Named Entity Recognition:

- SpaCy library

Translation Oracle:

- Google Translate API

Models (LLMs):

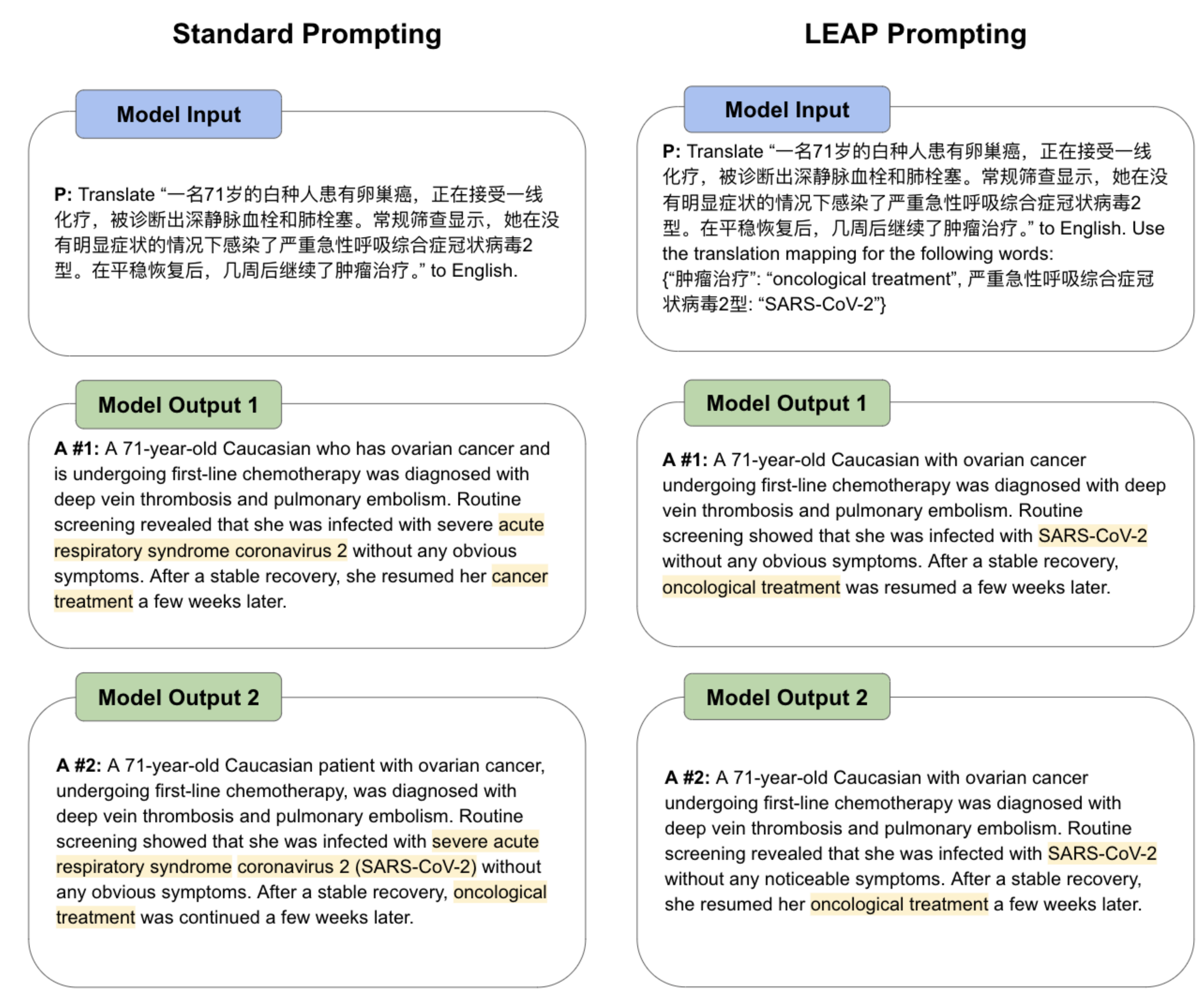
- GPT 4o-mini
- Claude 3.5 Haiku
- Mistral 70 Billion

Languages (En → Lang):

- French
- Simplified Chinese

Hyperparameter:

- Prompt each model to translate a document **k = 3** times



Regular vs LEAP example on ChatGPT's output

Evaluation Metrics

Jargon Translation Consistency (JTC): Measures how consistently domain-specific terminology is translated across multiple translations.

- E = set of entities (jargon terms)
- T = set of translations
- c_e = count of entity e in source text
- $c_{t,e}$ = count of translated entity in translation t
- K = number of times we prompt

We contribute JTC as a novel consistency metric in addition to LEAP

$$JTC = 1 - \frac{\sum_{e \in E}(K \cdot c_e - \sum_{t \in T} c_{t,e})}{K \cdot \sum_{e \in E} c_e}$$

Jaccard Similarity: Measures the similarity between different translations by comparing their word/character sets

- Preprocessed differently for Chinese (jieba segmentation) vs other languages (word splitting)
- Ranges from 0 (completely different) to 1 (identical)
- Considers unique terms only

$$Jaccard(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

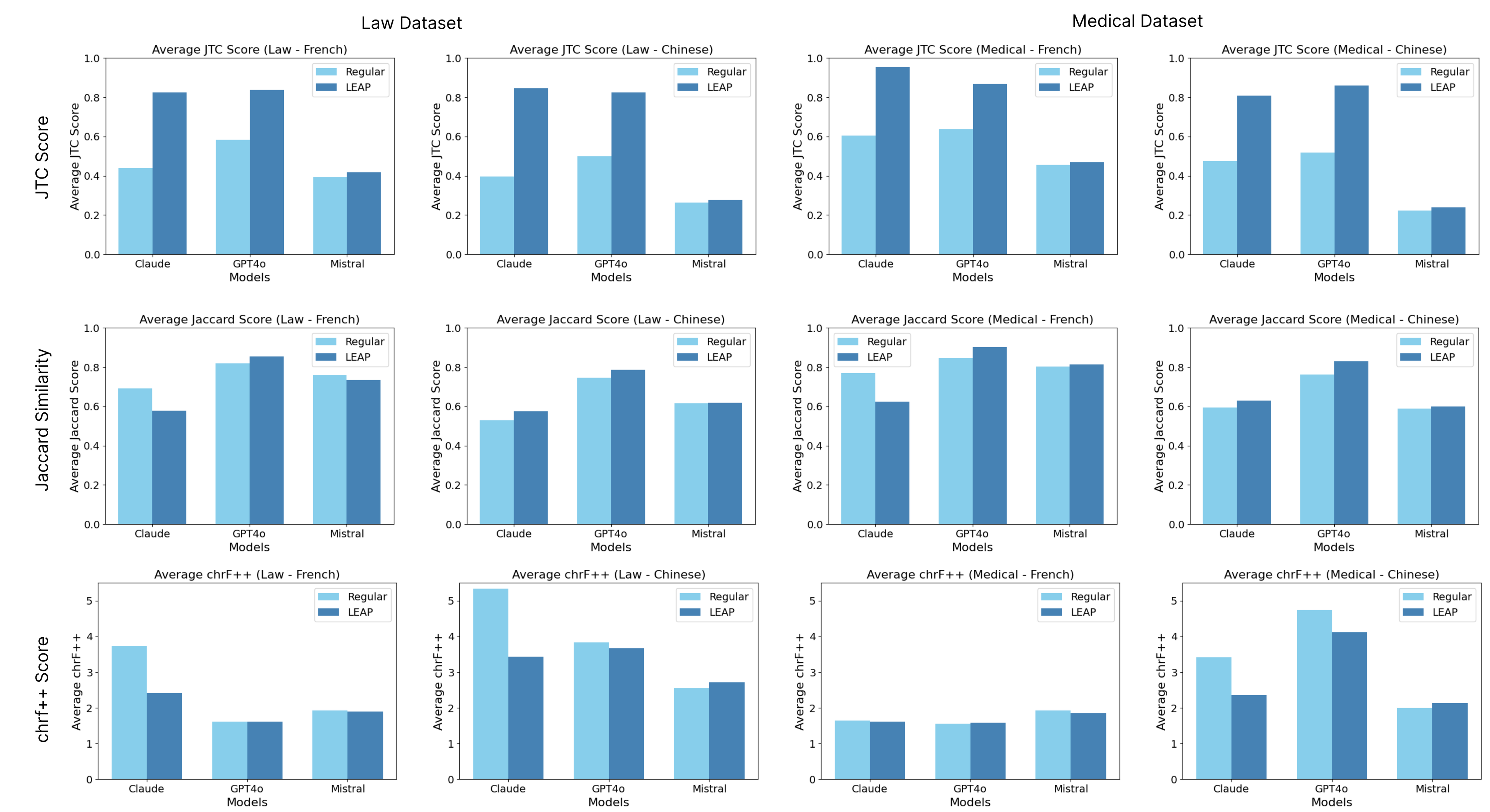
chrf++: A character-level metric that compares translations against a reference translation

- prec = character n-gram precision
- rec = character n-gram recall
- β = 2 (weighs recall twice as much as precision)
- n-gram size = 6 (default)

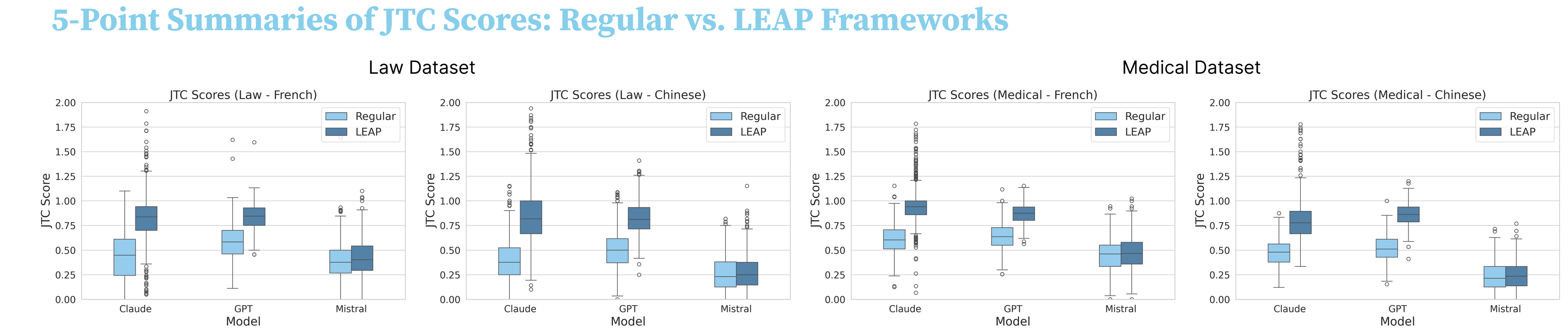
$$chrF++ = (1 + \beta^2) \cdot \frac{prec \cdot rec}{(\beta^2 \cdot prec) + rec}$$

Results

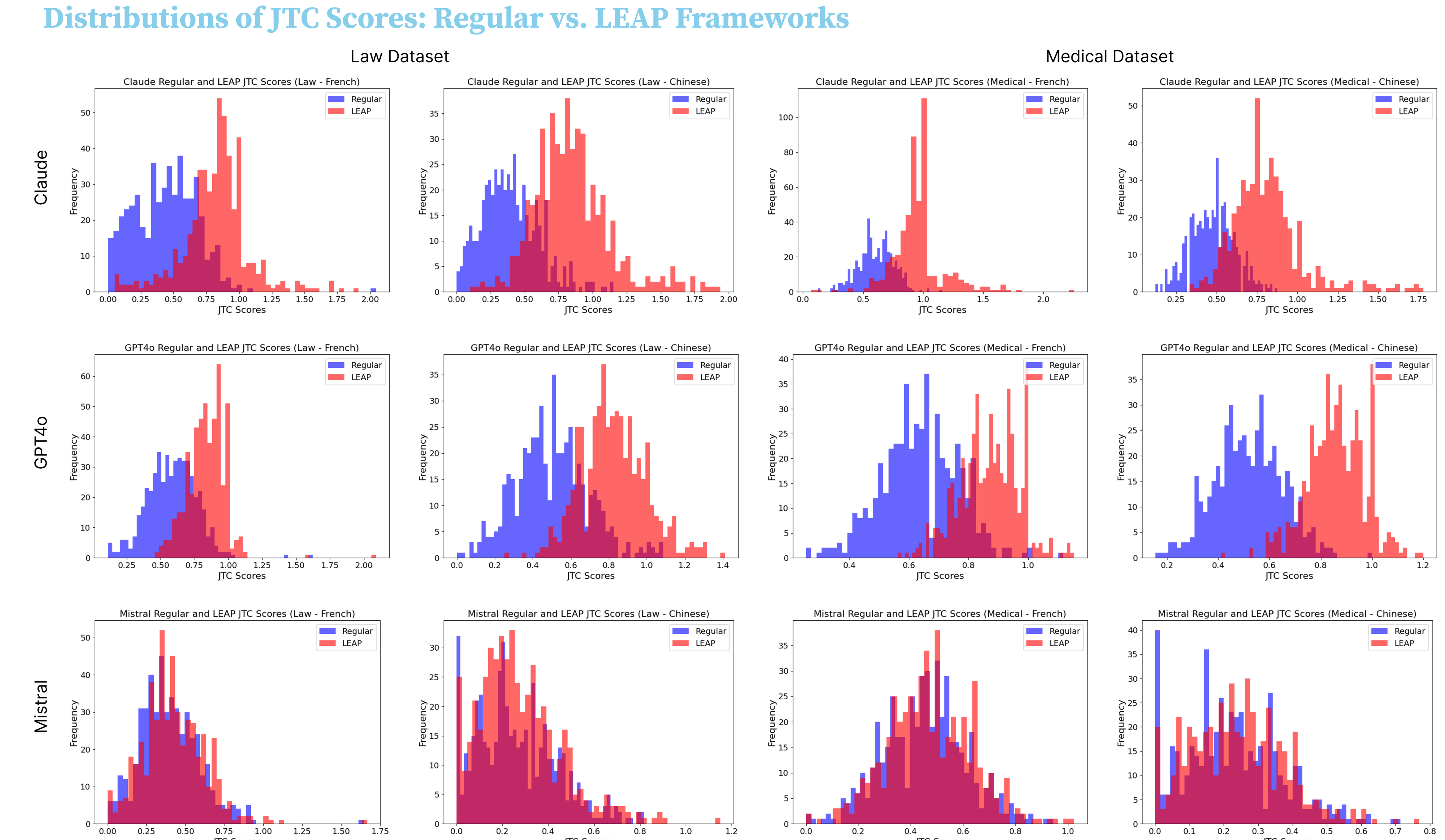
JTC, Jaccard, and chrf++ Scores: Regular vs. LEAP Frameworks



5-Point Summaries of JTC Scores: Regular vs. LEAP Frameworks



Distributions of JTC Scores: Regular vs. LEAP Frameworks



We conclude that the LEAP framework has better jargon translation consistency. LEAP's JTC scores are consistently better than the regular framework. There remains room for improvement for Jaccard Similarity and chrf++.