# A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models

**Haoran Xu**, Young Jin Kim, Amr Sharaf, Hany Hassan Awadalla

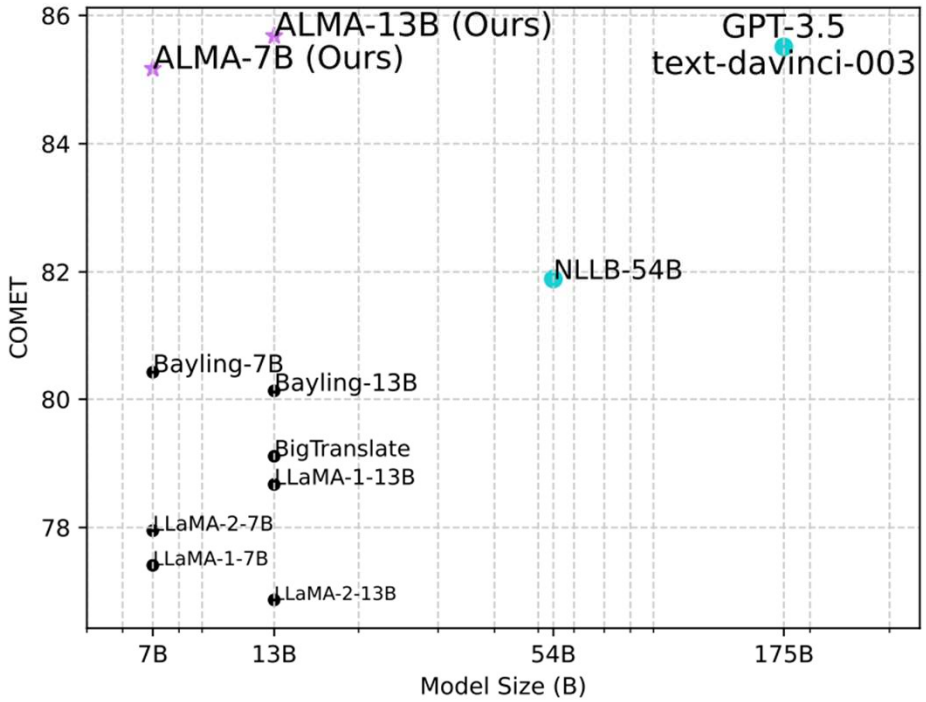
April 2024



# ALMA Overview

What is ALMA *(**A**dvanced **L**anguage **M**odel-Based Transl**A**tors)* ?

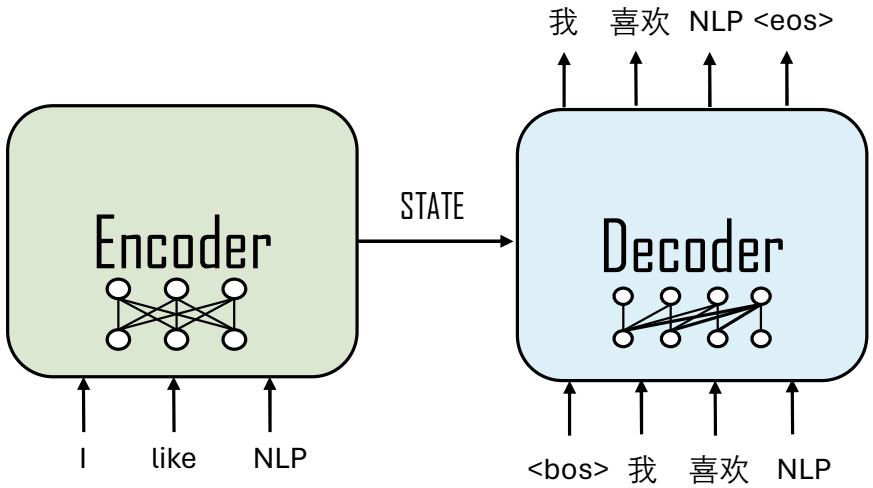
The first open-source LLM-based translation models which have the comparable performance with GPT-3.5.



# Introduction

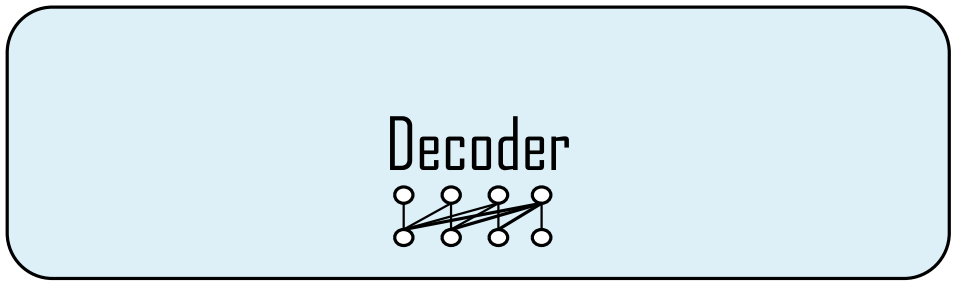
What does the mainstream machine translation model look like?

Encoder-decoder, millions of parallel data to train the model.



# Introduction

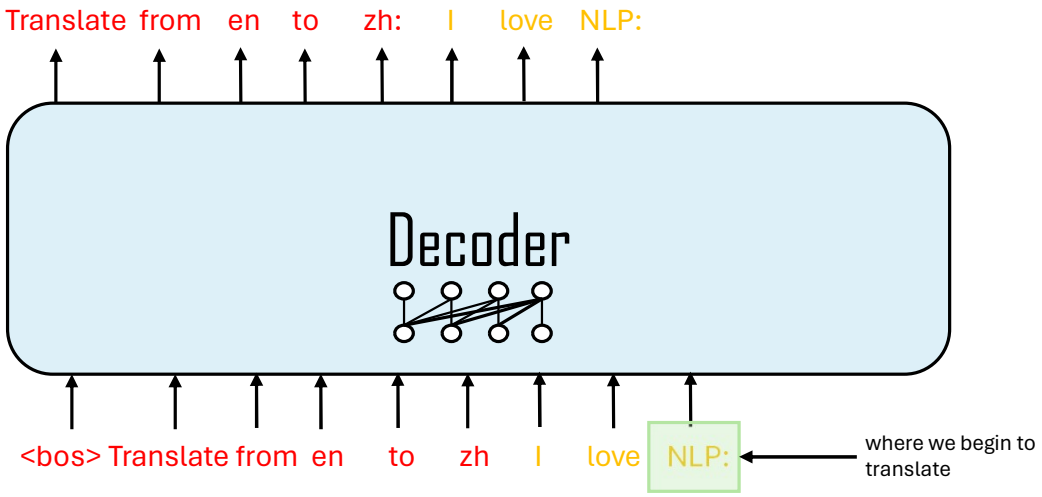
How to use decoder-only models (LLMs) for translation?




# Introduction

How to use decoder-only models (LLMs) for translation?

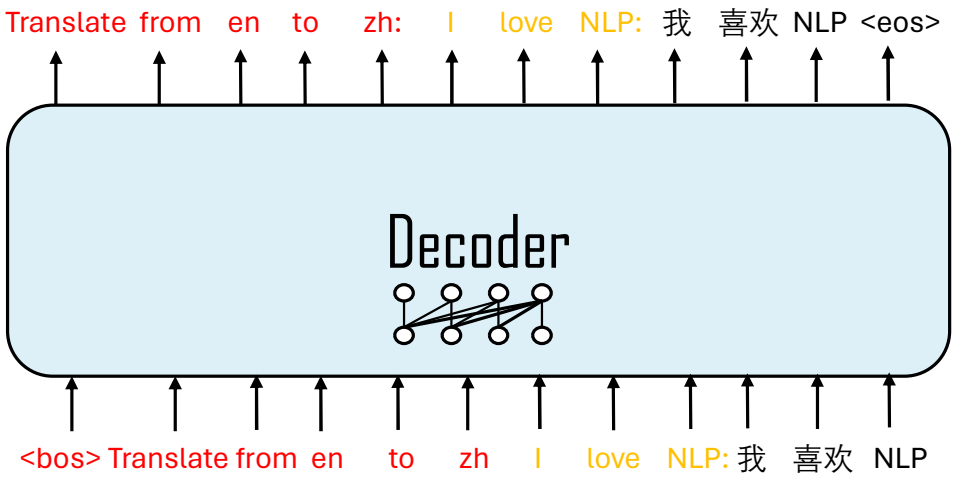
We need a **prompt** to instruct the model to translate:



# Introduction

## How to use decoder-only models (LLMs) for translation?

We need a **prompt** to instruct the model to translate:

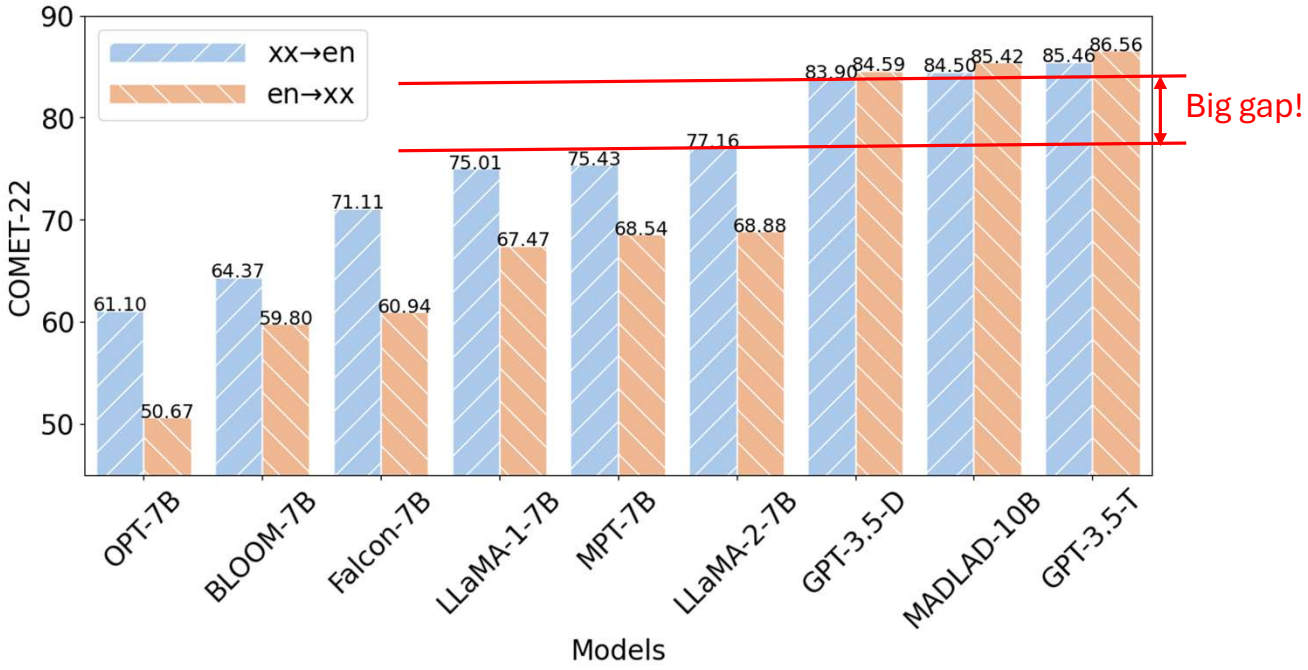


# Introduction

LLMs Performance in machine translation?

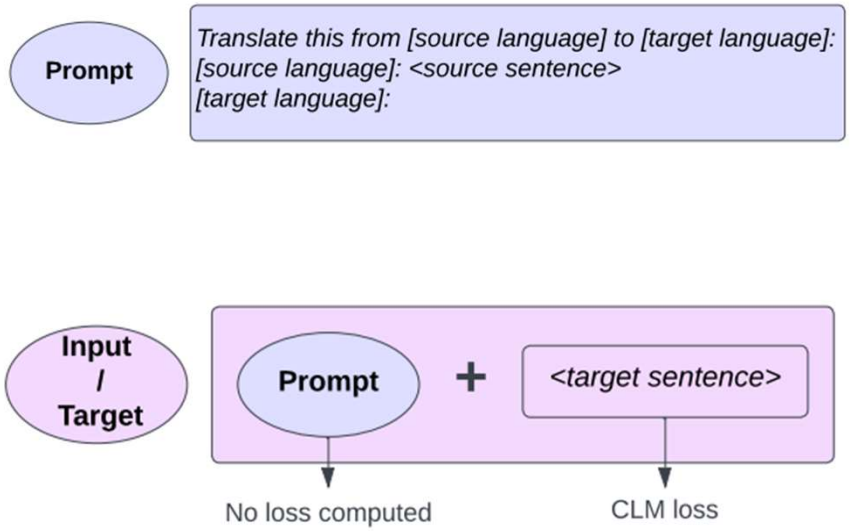
Averaged **zero-shot** (just grab and test) performance on WMT'22:

Avg. performance on de,cs,is,zh,ru



# Challenges of MT in LLMs

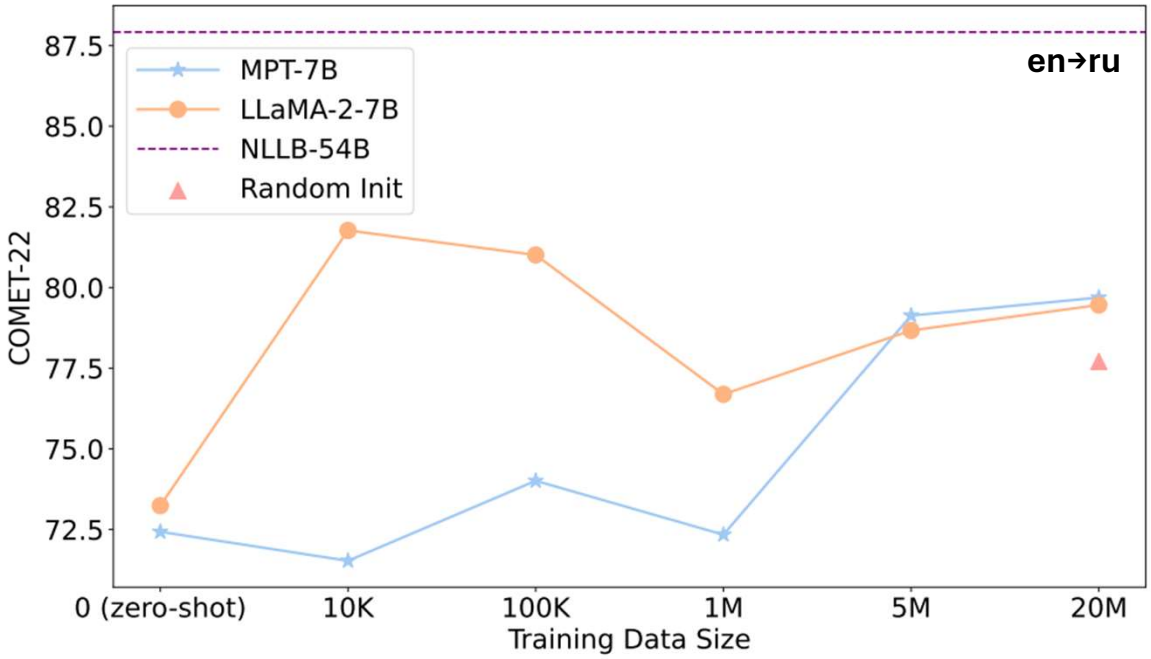
Previously: Fine-tune the model on millions of parallel data
Inertial thought: do the same!



# Challenges of MT in LLMs

How much parallel data do we need? Is more always better?

Anti-intuitive: 10K looks like enough. 20M may lead to catastrophic forgetting.



# A New Training Recipe

**Motivation 1:**

LLMs are trained on English-centric data and lack knowledge of other languages. It should learn general multilingual linguistic knowledge.

**Motivation 2:**

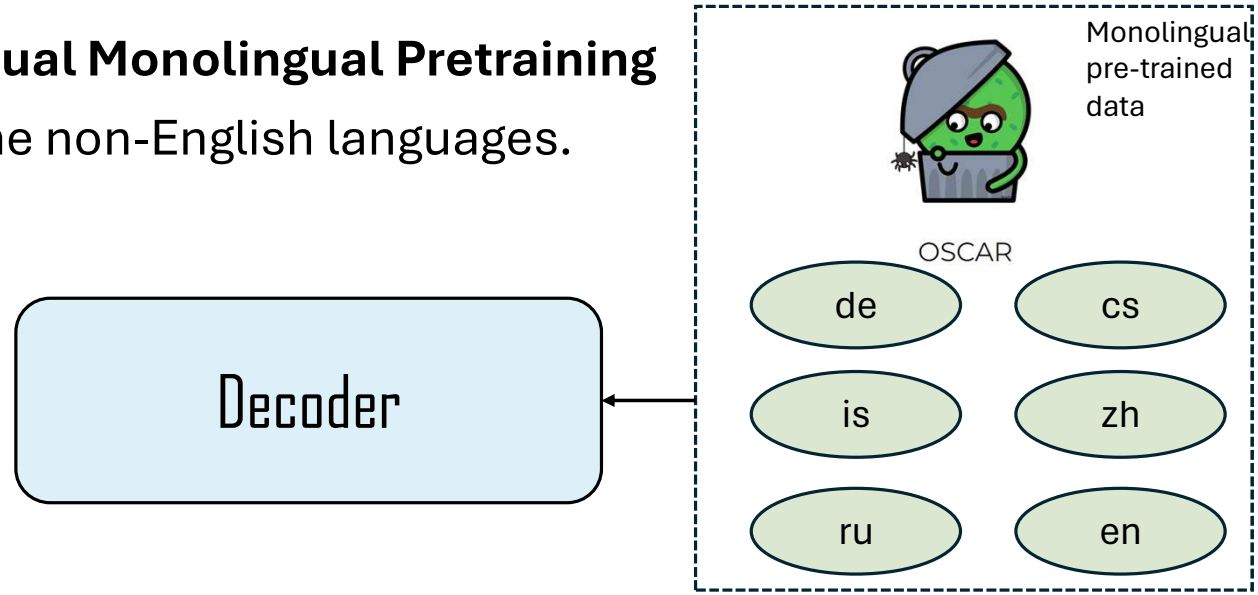
Around 10K parallel data is sufficient for well-pretrained LLM, so why not use small but very high-quality data for training?

# A New Training Recipe

**Step 0: Using** de,cs,is,zh,ru for all steps and all experiments.

**Step 1: Continual Monolingual Pretraining**

Fine-tune on the non-English languages.

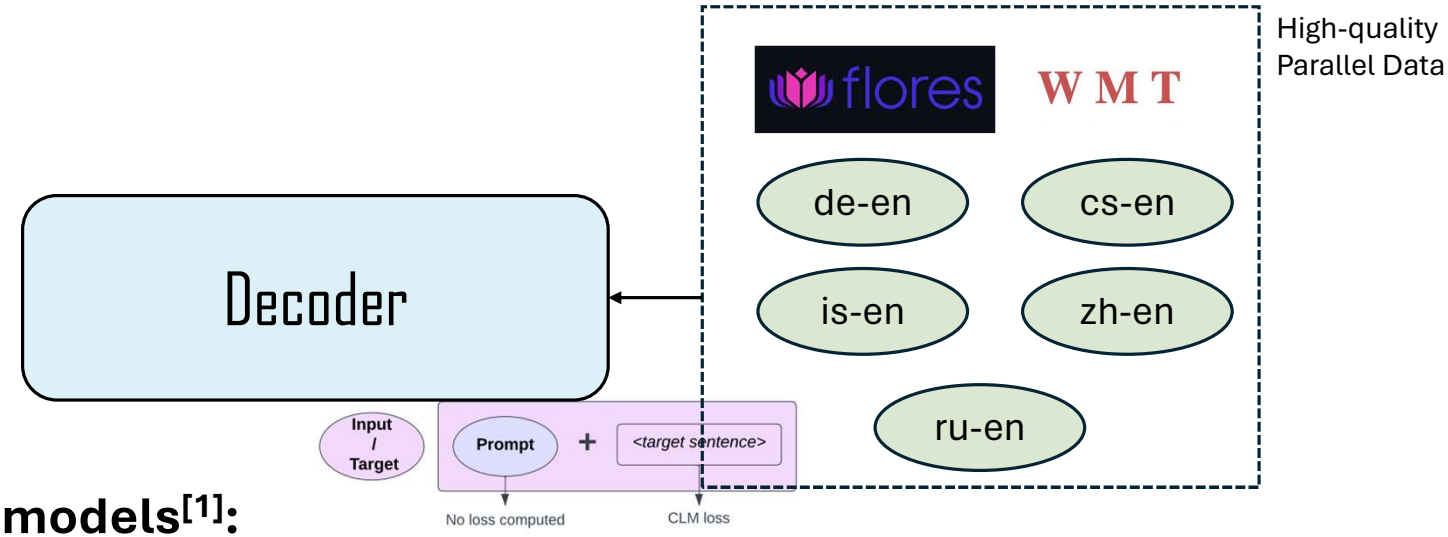



[1] A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models (ICLR 2024)

# A New Training Recipe

**Step 2: High-Quality Parallel Data Fine-tuning**

We use WMT test + FLORES for training (50K data)



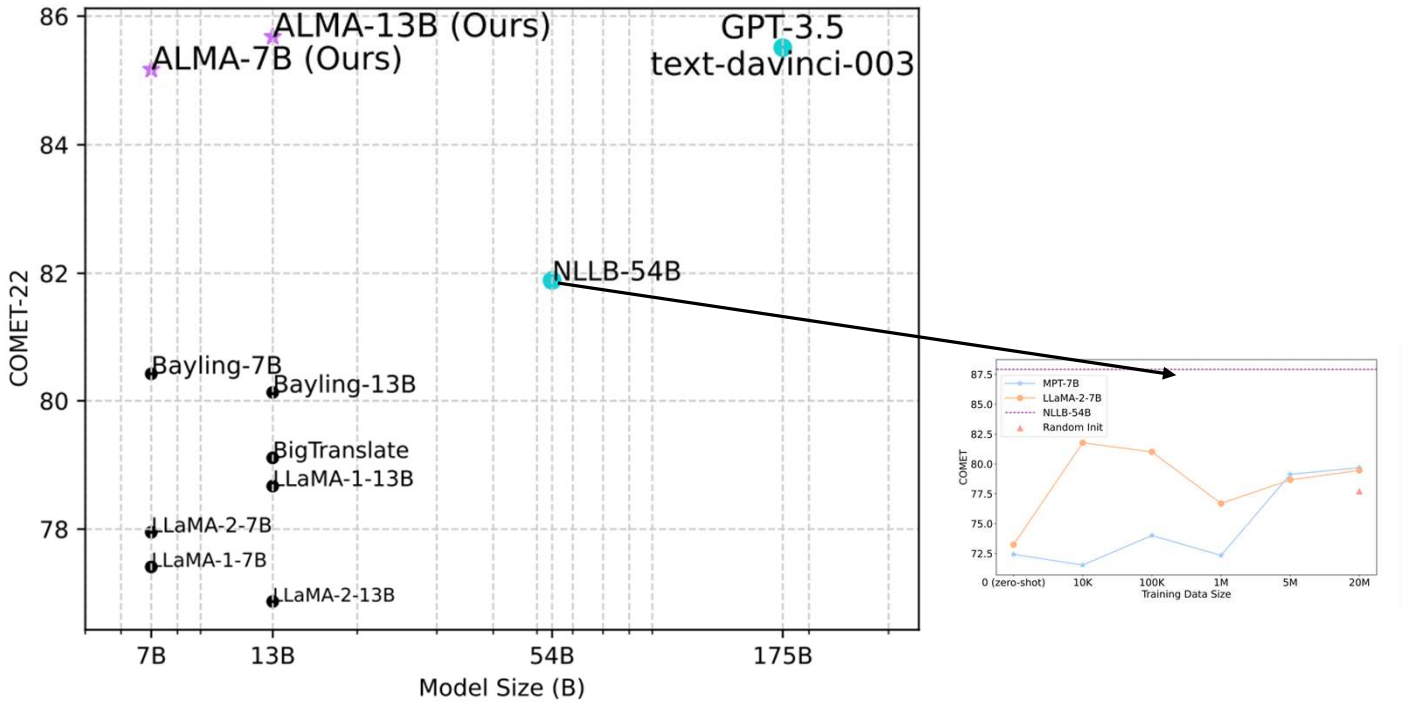

**ALMA models[1]:**

We fine-tuned LLaMA-2 with the new training recipe.

[1] A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models (ICLR 2024)

# A New Training Recipe

What does the performance look like now?



JOHNS HOPKINS UNIVERSITY

Microsoft

# Questions?