Bangla - Hindi MT improvement using LLM

G M Arafat Rahman, Abbas Awarkeh, Mohammad AL TAKACH

Motivation:

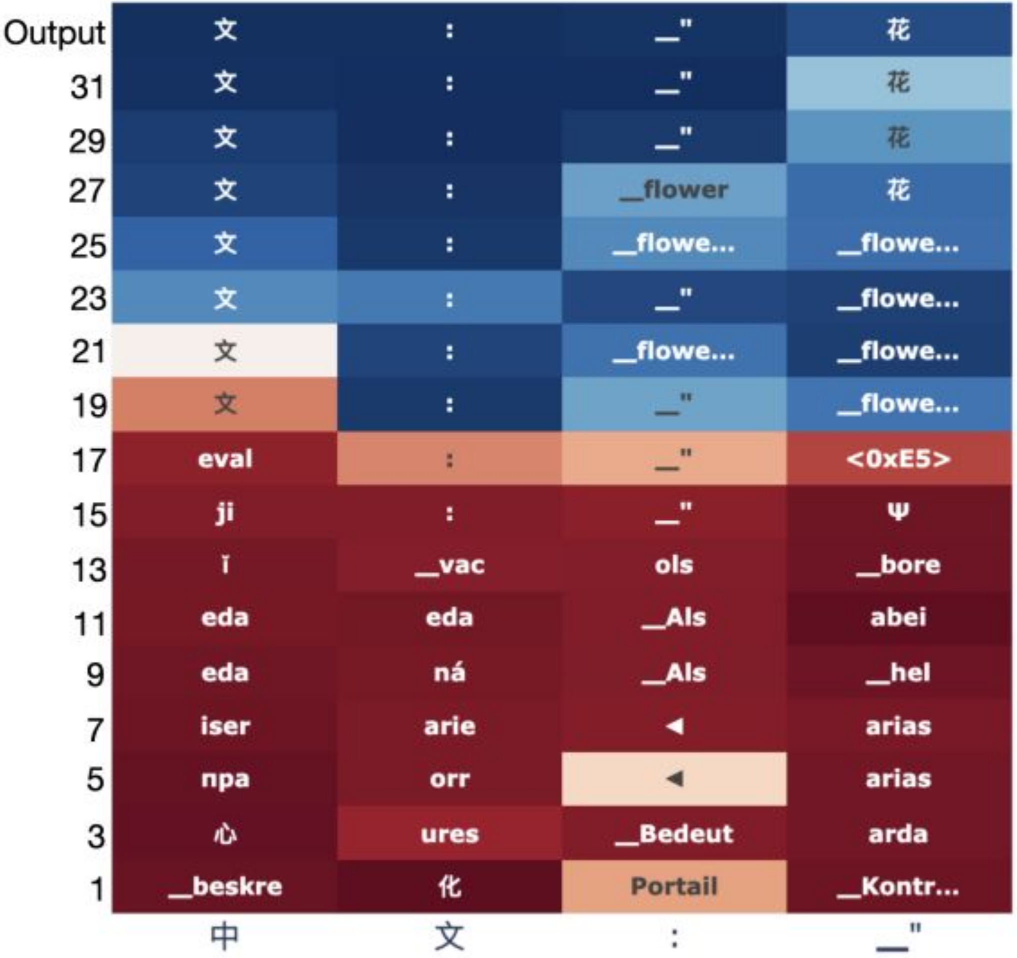
- Most modern large language models (LLMs) are trained on massive corpora of mostly **English text**
  - Intuitively, one way to achieve strong performance on non-English data in a data-efficient manner is to use **English as a pivot language**, by first translating input to English, processing it in English, and then translating the answer back to the input language.

Motivation Cont.

- **Chris Wendler et al. showed that the anatomy of Llama-2’s forward pass**, suggesting that, in middle layers, **the transformer operates in an abstract “concept space”**
- In this interpretation, the latent embeddings’ proximity to English tokens observed through the logit lens follows from **an English bias in concept space**, **rather than from the model first translating to English and then “restarting” its forward pass from there**

Logit Lens:

- In a transformer, **each input token’s embedding vector is gradually transformed layer by layer without changing its shape**. After the final layer, an **“unembedding” operation turn**s the vector into a next token distribution.
- applying the “unembedding” operation prematurely in intermediate, non-final layers—**a technique called logit lens**



	中	文	:	_"
Output	文	:	_"	花
31	文	:	_"	花
29	文	:	_"	花
27	文	:	_flower	花
25	文	:	_flowe...	_flowe...
23	文	:	_"	_flowe...
21	文	:	_flowe...	_flowe...
19	文	:	_"	_flowe...
17	eval	:	_"	<0xE5>
15	ji	:	_"	Ψ
13	ĭ	_vac	ols	_bore
11	eda	eda	_Als	abei
9	eda	ná	_Als	_hel
7	iser	arie	◀	arias
5	npa	orr	◀	arias
3	心	ures	_Bedeut	arda
1	_beskre	化	Portail	_Kontr...

Figure: Illustration of Logit Lens

Source: https://arxiv.org/abs/2402.10588

What we have now:

- Training data: WMT 21
- Test data: WMT 21
- **Mistral-7b-bnb-4bit quantized fine tuned model**

Final goal:

1. Establish the fact that for mistral’s forward pass, in middle layers, the transformer operates in an abstract “concept space” and, **the latent embeddings’ proximity to English tokens shows an English bias in concep**t space when translating from Bangla to Hindi and vice versa.
2. We have 4 million **Bengali-Hindi** parallel corpora, how much data we will use for fine-tuning and for testing will be an exploring point.

References:

GPT-4 Technical Report - [2303.08774] GPT-4 Technical Report (arxiv.org)

Do Llamas Work in English? On the Latent Language of Multilingual Transformers - https://arxiv.org/abs/2402.10588