Machine Translation with Large Language Models

Prompting, Few-shot Learning, and Fine-tuning with QLoRA

Xuan Zhang, Navid Rajabi, Kevin Duh, Philipp Koehn
Large Language Models increase in size
Selected LLMs, deep learning models trained on enormous amounts of textual data

LLMs have overtaken much of NLP. How about Machine Translation?

*values the model adjusts through training to minimise errors
Source: companies, TechCrunch

MT w/ GPT models

<table>
<thead>
<tr>
<th>System</th>
<th>COMET-22</th>
<th>COMETkiwi</th>
<th>ChrF</th>
<th>BLEU</th>
<th>COMET-22</th>
<th>COMETkiwi</th>
<th>ChrF</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT-Best</td>
<td>85.0</td>
<td>81.4</td>
<td>58.5</td>
<td>33.4</td>
<td>87.2</td>
<td>83.6</td>
<td>64.6</td>
<td>38.4</td>
</tr>
<tr>
<td>text-davinci-002</td>
<td>73.2</td>
<td>73.1</td>
<td>46.1</td>
<td>23.3</td>
<td>82.0</td>
<td>79.0</td>
<td>56.0</td>
<td>28.6</td>
</tr>
<tr>
<td>text-davinci-003</td>
<td>84.8*</td>
<td>81.2*</td>
<td>56.8</td>
<td>30.9</td>
<td>85.6*</td>
<td>82.8*</td>
<td>60.2*</td>
<td>31.8*</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>84.8*</td>
<td>81.1</td>
<td>58.3*</td>
<td>33.4*</td>
<td>84.2</td>
<td>81.0</td>
<td>59.6</td>
<td>30.9</td>
</tr>
<tr>
<td>WMT-Best</td>
<td>81.0</td>
<td>77.7</td>
<td>61.1</td>
<td>33.5</td>
<td>86.7</td>
<td>82.0</td>
<td>41.1</td>
<td>44.8</td>
</tr>
<tr>
<td>text-davinci-002</td>
<td>74.1</td>
<td>73.1</td>
<td>49.6</td>
<td>20.6</td>
<td>84.0</td>
<td>79.0</td>
<td>32.1</td>
<td>36.4</td>
</tr>
<tr>
<td>text-davinci-003</td>
<td>81.6*</td>
<td>78.9*</td>
<td>56.0*</td>
<td>25.0</td>
<td>85.8*</td>
<td>81.3*</td>
<td>34.6</td>
<td>38.3</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>81.2</td>
<td>78.3</td>
<td>56.0</td>
<td>25.9*</td>
<td>84.4</td>
<td>78.7</td>
<td>36.0*</td>
<td>40.3*</td>
</tr>
<tr>
<td>WMT-Best</td>
<td>86.0</td>
<td>81.7</td>
<td>68.9</td>
<td>45.1</td>
<td>89.5</td>
<td>84.4</td>
<td>58.3</td>
<td>32.4</td>
</tr>
<tr>
<td>text-davinci-002</td>
<td>77.5</td>
<td>76</td>
<td>58.7</td>
<td>34.9</td>
<td>85.4</td>
<td>80.9</td>
<td>51.6</td>
<td>25.1</td>
</tr>
<tr>
<td>text-davinci-003</td>
<td>84.8*</td>
<td>81.1*</td>
<td>64.6</td>
<td>38.5</td>
<td>86.7*</td>
<td>82.2*</td>
<td>54.0*</td>
<td>27.5*</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>84.8*</td>
<td>81.0</td>
<td>66.5*</td>
<td>41.0*</td>
<td>77.6</td>
<td>70.4</td>
<td>41.1</td>
<td>19.0</td>
</tr>
<tr>
<td>WMT-Best</td>
<td>89.5</td>
<td>80.7</td>
<td>81.2</td>
<td>64.8</td>
<td>85.7</td>
<td>79.5</td>
<td>74.6</td>
<td>58.4</td>
</tr>
<tr>
<td>text-davinci-002</td>
<td>66.6</td>
<td>67.9</td>
<td>45.8</td>
<td>25.9</td>
<td>64.2</td>
<td>67.6</td>
<td>44.6</td>
<td>24.5</td>
</tr>
<tr>
<td>text-davinci-003</td>
<td>84.6</td>
<td>77.9</td>
<td>65.7</td>
<td>42.5*</td>
<td>78.5</td>
<td>76.1</td>
<td>58.9</td>
<td>35.6</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>84.7*</td>
<td>78.5*</td>
<td>65.2</td>
<td>42.0</td>
<td>81.6*</td>
<td>79.8*</td>
<td>60.7*</td>
<td>37.3*</td>
</tr>
</tbody>
</table>

Table 2: Zero-Shot evaluation results with three GPT models on 8 language pairs from WMT22 Testset. The best scores across different systems are marked bold. * denotes the best results among GPT systems.
## Previous work on MT w/ LLMs

<table>
<thead>
<tr>
<th>LLMs</th>
<th>Methods</th>
<th>Datasets</th>
<th>Language pairs</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang et al., Jan 2023</td>
<td>GLM-130B</td>
<td>K-shots, FLORES, WMT21, Multi-domain</td>
<td>en, de, zh</td>
<td>Performance depends on the number and quality of prompt examples.</td>
</tr>
<tr>
<td>Hendy et al., Feb 2023</td>
<td>ChatGPT</td>
<td>K-shots, WMT21, WMT22</td>
<td>18 language pairs</td>
<td>LLMs are worse than dedicated MT models.</td>
</tr>
<tr>
<td>Bawden and Yvon, May 2023</td>
<td>BLOOM</td>
<td>K-shots, WMT14, FLORES-101, DiaBLA</td>
<td>-</td>
<td>Few-shot results are close to SOTA.</td>
</tr>
<tr>
<td>Moslem et al., May 2023</td>
<td>GPT-3.5, BLOOM, BLOOMZ</td>
<td>K-shots, TICO-19</td>
<td>en, ar, es, fr, rw, zh</td>
<td>Few-shot results are better than dedicated MT models.</td>
</tr>
<tr>
<td>Sia and Duh, May 2023</td>
<td>GPTNeo-2.7B, BLOOM-3B, XGLM-2.9B</td>
<td>K-shots, WMT19, Biomedical, MTNT, FLORES</td>
<td>en, fr, de, pt</td>
<td>Better performance is achieved with prompts from the same domain.</td>
</tr>
<tr>
<td>Wang et al., Oct 2023</td>
<td>GPT-3.5, GPT-4</td>
<td>K-shots, mZPRT, WMT22, IWSLT</td>
<td>en, de, zh, ru</td>
<td>Promising and better results are obtained for document-level translation.</td>
</tr>
<tr>
<td>Zhu et al., Oct 2023</td>
<td>ChatGPT, GPT-4, OPT-175B, LLaMA2-7B-chat, Falcon-7B, XGLM-7.5B, BLOOMZ-7.1B</td>
<td>K-shots, FLORES101</td>
<td>102 languages, 606 translation directions</td>
<td>GPT-4 beats NLLB in 40.91% of translation directions. GPT-4 lags behind commercial MT systems.</td>
</tr>
</tbody>
</table>
MT w/ LLMs: prompting, few-shot learning, fine-tuning

Translate French to English: French: Bienvenue à Singapour
English: Have a nice day.

Prompting / Zero-shot

Few-shot Learning

Fine-tuning
MT w/ LLMs: prompting, few-shot learning, fine-tuning

Translate French to English:
French: Bienvenue à Singapour
English: 

Prompting / Zero-shot

Translate French to English:
French: Passez une bonne journée.
English: Have a nice day.
French: Bienvenue à Singapour
English: 

Few-shot Learning

Translate French to English:
French: Bienvenue à Singapour
English: 

Fine-tuning

Expensive to fine-tune the entire model.
Fine-tuning w/ QLoRA (Quantization + Low-Rank Adaptation)

LoRA reparameterization. Only A and B are trained.

Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

Hu et al., LoRA: Low-Rank Adaption of Large Language Models, ICLR 2021
Tim et al., QLoRA: Efficient Finetuning of Quantized LLMs, arXiv 2023
Datasets

**Language pair:** French - English

**Fine-tuning dataset:** WMT14 Europarl + News Commentary

**Dev:** newstest2013

**Test:** newstest2014

<table>
<thead>
<tr>
<th></th>
<th>#sents</th>
<th>#docs</th>
<th>avg.sents/doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>2,366,117</td>
<td>21,430</td>
<td>144</td>
</tr>
<tr>
<td>dev</td>
<td>3000</td>
<td>126</td>
<td>24</td>
</tr>
<tr>
<td>test</td>
<td>3003</td>
<td>169</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics.
Baseline and LLMs

- **Baseline:**
  trained-from-scratch 12-layer transformer with 4B parameters

- **LLMs:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Release Time</th>
<th>Data</th>
<th>Size (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-Neo (Black et al., 2021)</td>
<td>Mar, 2021</td>
<td>English-centric</td>
<td>1.3; 2.7</td>
</tr>
<tr>
<td>OPT (Zhang et al., 2022)</td>
<td>June, 2022</td>
<td>English-centric</td>
<td>1.3; 2.7; 6.7</td>
</tr>
<tr>
<td>LLaMA2 (Touvron et al., 2023)</td>
<td>July, 2023</td>
<td>English-centric</td>
<td>7; 13</td>
</tr>
<tr>
<td>XGLM (Lin et al., 2021)</td>
<td>Nov, 2022</td>
<td>Multilingual</td>
<td>1.7; 2.9; 4.5; 7.5</td>
</tr>
<tr>
<td>BLOOMZ (Muennighoff et al., 2022)</td>
<td>Nov, 2022</td>
<td>Multilingual</td>
<td>1.7; 3; 7.1</td>
</tr>
</tbody>
</table>

Table 2: Overview of evaluated LLMs.
Prompted Fine-tuning

• sentence-level prompt:
  
  French: [fr sent] English: [en sent] <eos>

• document-level prompt¹:
  
  French: <BEG> [fr sent1] <SEP> [fr sent2] <SEP> <BRK>
  English: <BEG> [en sent1] <SEP> [en sent2] <SEP> <BRK>

  French: <CNT> [fr sent1] <SEP> [fr sent2] <SEP> <END>
  English: <CNT> [en sent1] <SEP> [en sent2] <SEP> <END>

¹ Junczys-Dowmunt, Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation, WMT 2019
Prompting vs. Fine-tuning LLMs

![Graph showing BLEU and COMET scores for different models](image)

- **Baseline (B)**
- **B BLEU**
- **P BLEU**
- **FT BLEU**

- **Baseline (B)**
- **B COMET**
- **P COMET**
- **FT COMET**
Prompting vs. Fine-tuning LLMs

- High COMET: LLMs produce semantically coherent translations
Prompting vs. Fine-tuning LLMs

• Fine-tuning boosts LLM performance on average by 8 BLEU points. BLOOMZ-7.1B: 20.13 BLEU leap
Prompting vs. Fine-tuning LLMs

- Baseline surpasses most prompted LLMs, except for LLaMA2
Prompting vs. Fine-tuning LLMs

- 8 out of 15 fine-tuned LLMs exceed Baseline, best: fine-tuned BLOOMZ-7.1B
Prompting vs. Fine-tuning LLMs

• No clear advantage is discerned comparing English-centric and multilingual LLMs
Prompting vs. Fine-tuning LLMs

- Bigger models do not necessarily outperform smaller ones: fine-tuned BLOOMZ-1.7B outperforms OPT-13B
Document-level Translation
Most LLMs struggle at document translation with prompting, except for LLaMA2.
Document-level Translation

• Fine-tuning enhances the BLEU scores of prompted counterparts by an average of **16.33** BLEU. (sentence-level: **8** BLEU improvement)
Document-level Translation

- Fine-tuning enhances semantic coherency
Fine-tuning w/ vs. w/o QLoRA

- QLoRA marks a 21-fold acceleration with 1370 times fewer trainable parameters.

<table>
<thead>
<tr>
<th></th>
<th>params(%)</th>
<th>#GPUs</th>
<th>time(hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No QLoRA</td>
<td>27.40</td>
<td>4</td>
<td>52</td>
</tr>
<tr>
<td>QLoRA</td>
<td>0.02</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3: Fine-tuning xglm-2.9b with and without QLoRA to achieve the BLEU score of 30.05. Only the self-attention layers are tuned. The rank $r$ for QLoRA approximation is set to 2.
Qualitative Study

French

L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés.

English reference

The UN has reported even higher numbers with 979 dead and 1,902 injured.

BLOOMZ-7.1B P

L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés. copy without translating

BLOOMZ-7.1B FT

The UN gives a higher figure with 979 dead and 1 902 wounded.<eos>.<eos>.<eos>. duplicating

LLaMA2-13B P

979 deaths and 1,902 injuries, according to the UN's latest tally.

LLaMA2-13B FT

The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The

duplicating

It is necessary to post-process generations from fine-tuned LLMs.
Conclusions

• The proficiency of LLMs in machine translation varies. **LLaMA2** consistently outperforms its counterparts. Other LLMs, when relying solely on k-shot learning, often lag behind the trained-from-scratch baseline model.

• Fine-tuning invariably enhances performance, especially for document-level translation. It can transform a seemingly inadequate model into a top-tier translation model.

• QLoRA is a superior alternative to original fine-tuning methods. Fine-tuning LLMs with QLoRA can be a promising and new paradigm for MT practice.