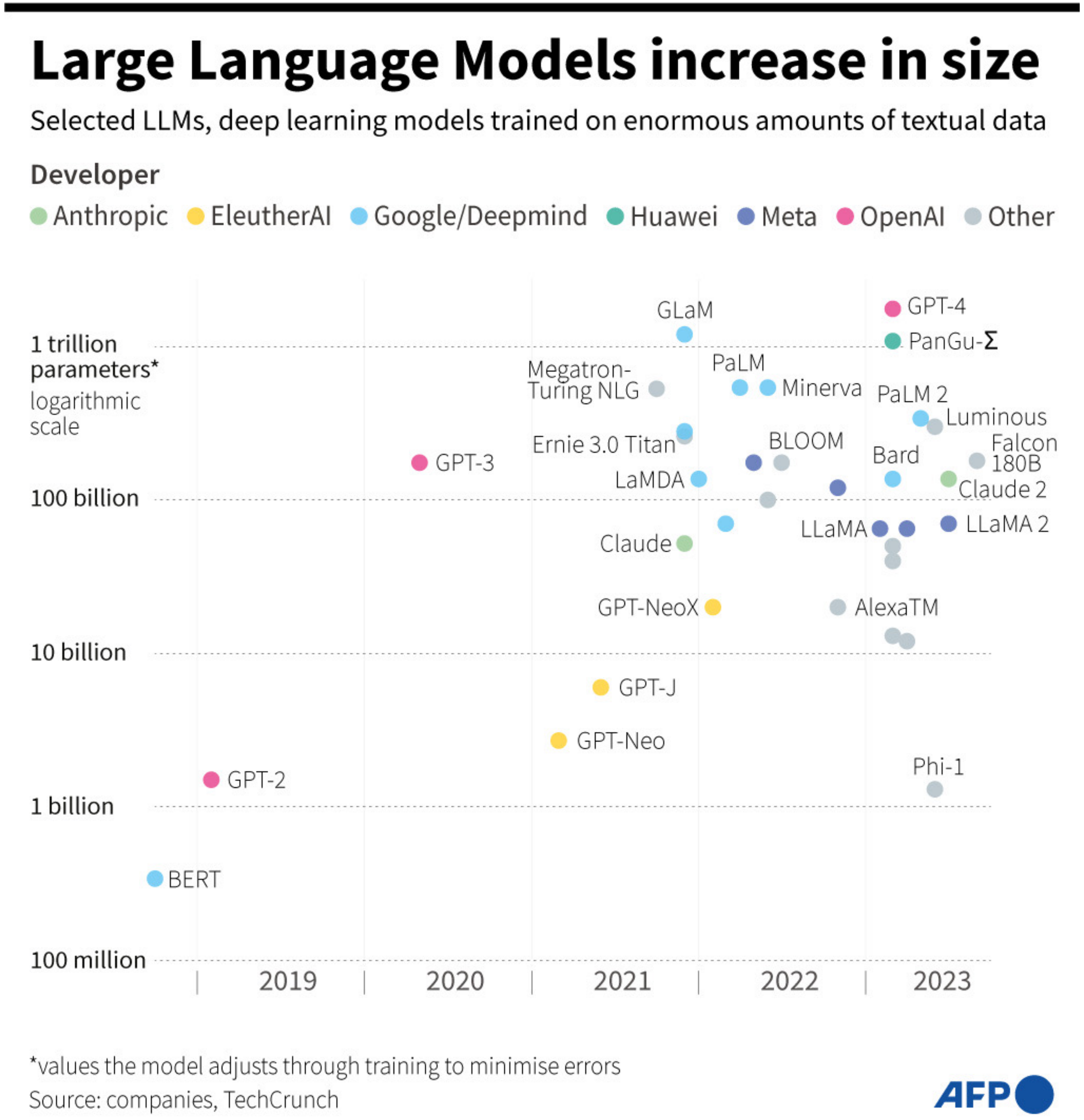
WMT 2023

Machine Translation with Large Language Models

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

Xuan Zhang, Navid Rajabi, Kevin Duh, Philipp Koehn



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

Image from: https://www.barrons.com/news/large-language-models-increase-in-size-2d40e471

MT w/ GPT models

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

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.

LLMs with zero-shot evaluation lag behind dedicated MT models.

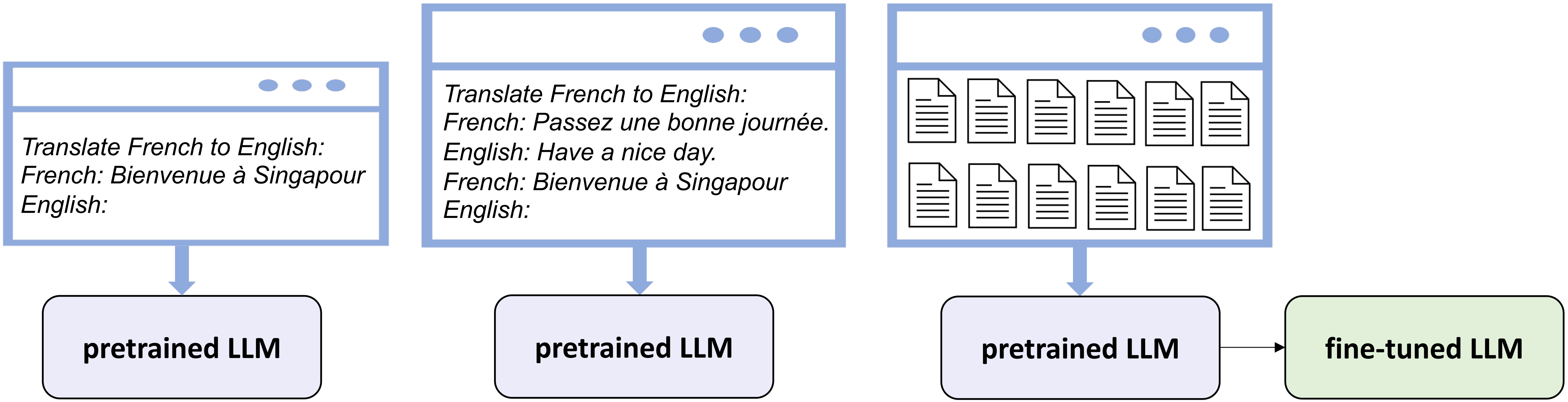
How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, Hendy et al. arXiv 2023

Previous work on MT w/ LLMs

Fine-tuning LLMs for MT tasks is underexplored.

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

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

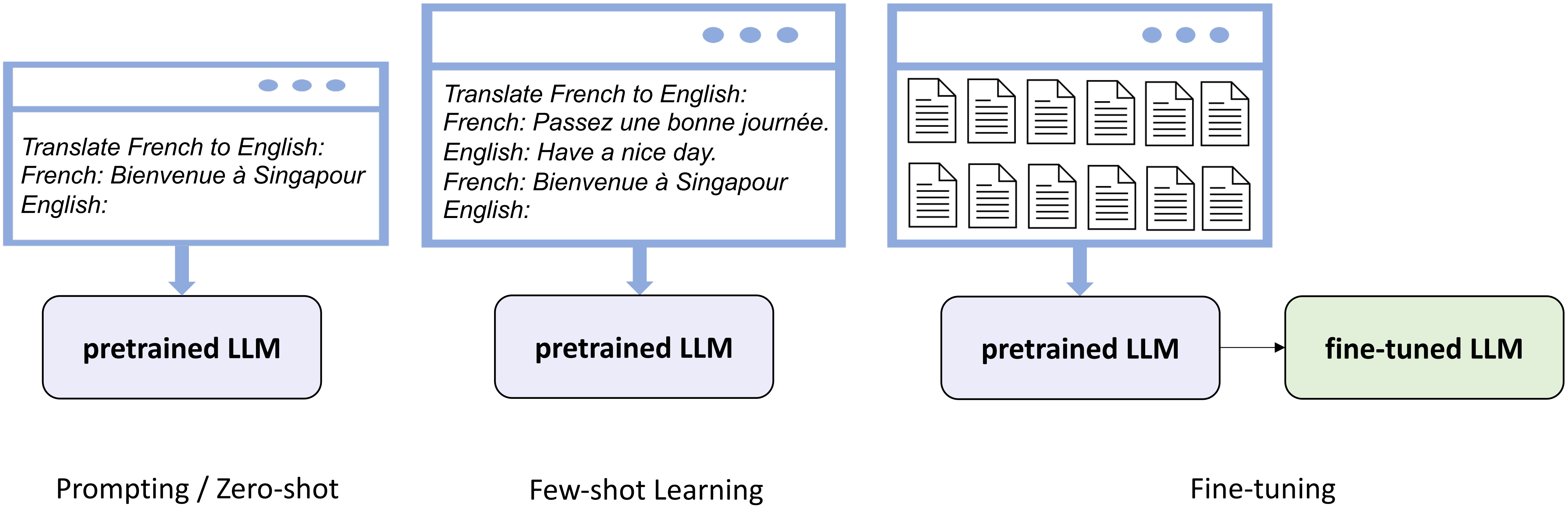


Prompting / Zero-shot

Few-shot Learning

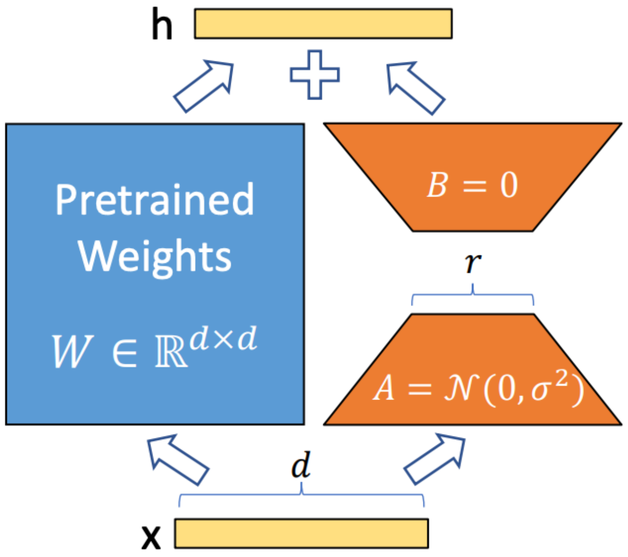
Fine-tuning

MT w/ LLMs:
prompting, few-shot learning, 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.

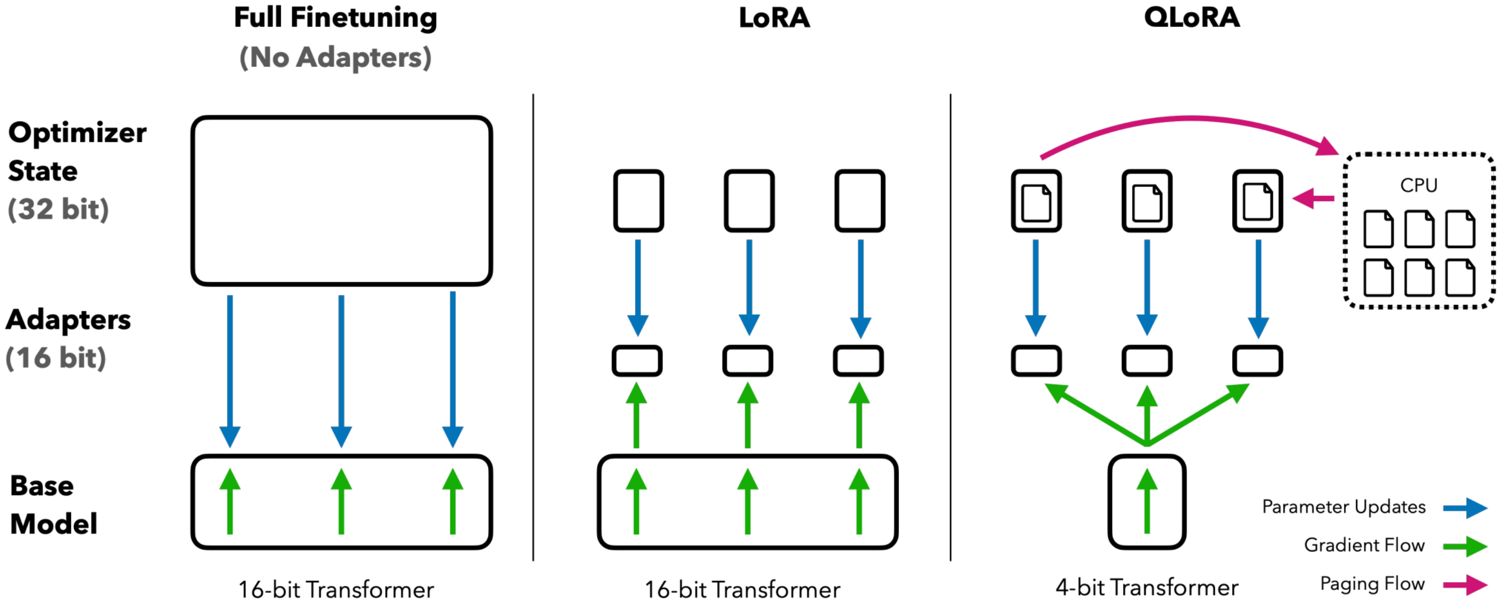


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

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

	#sents	#docs	avg.sents/doc
train	2,366,117	21,430	144
dev	3000	126	24
test	3003	169	18

Table 1: Dataset statistics.

Baseline and LLMs

- **Baseline:**

  trained-from-scratch 12-layer transformer with 4B parameters

- **LLMs:**

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

Table 2: Overview of evaluated LLMs.

Prompted Fine-tuning

- sentence-level prompt:

French: [fr sent] English: [en sent] <eos>

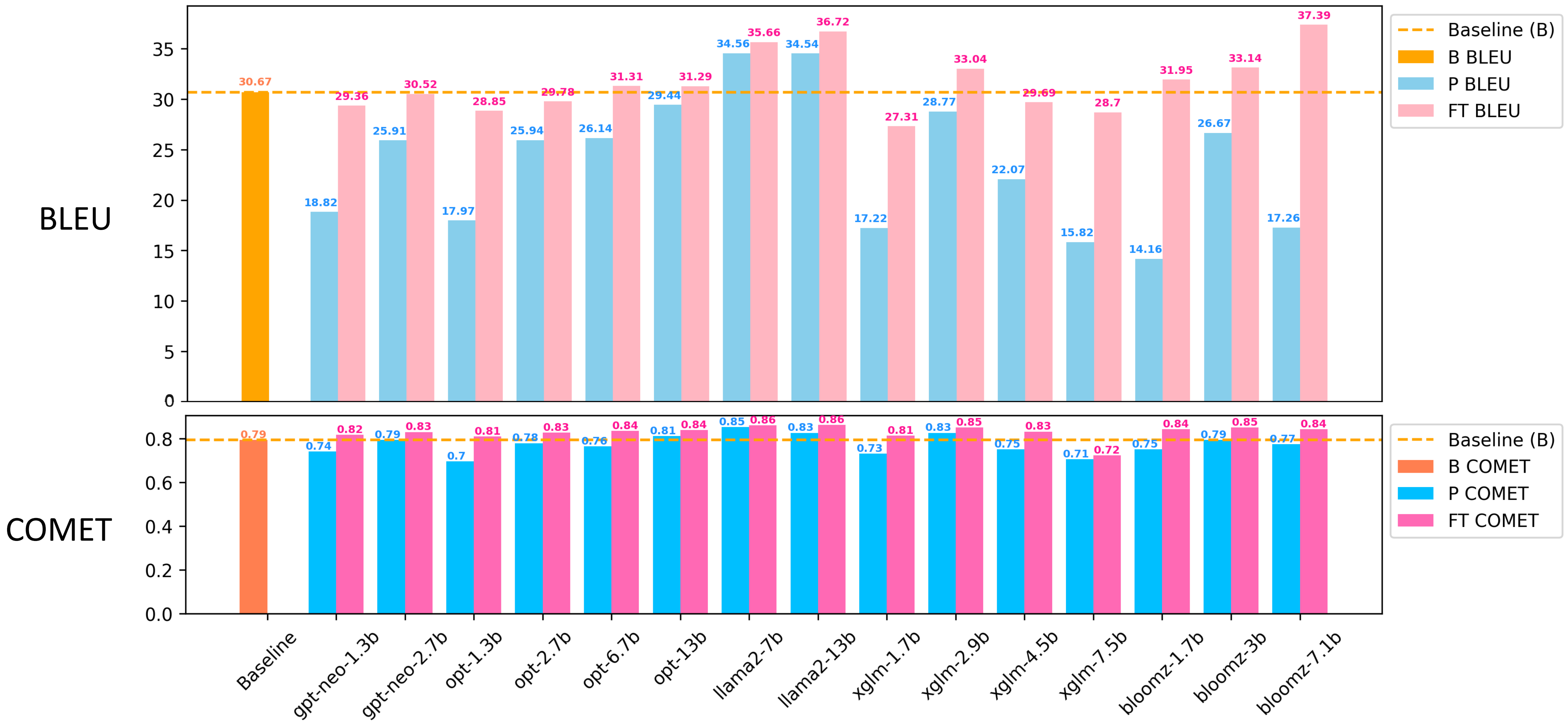
- document-level prompt[1]:

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>

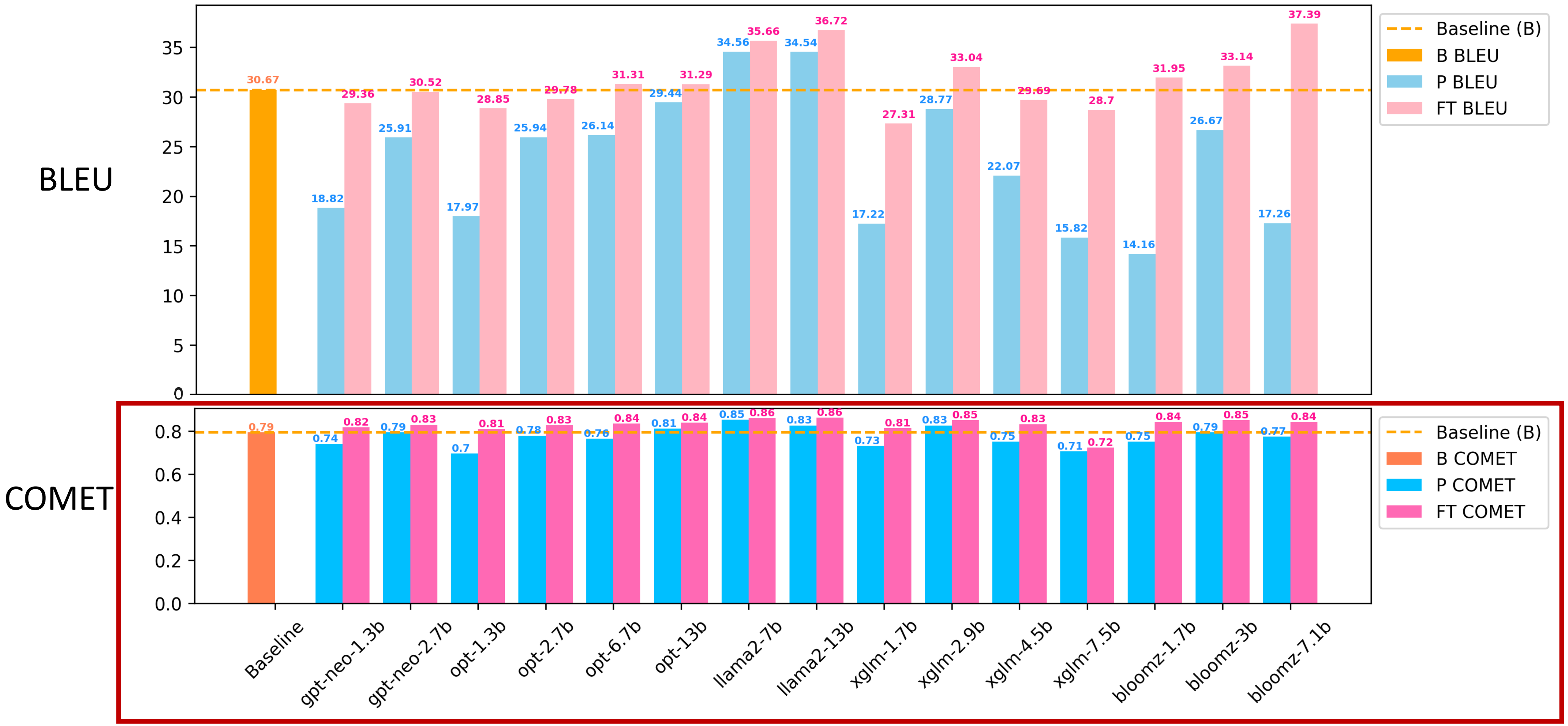
[1] Junczys-Dowmunt, Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation, WMT 2019

Prompting vs. Fine-tuning LLMs



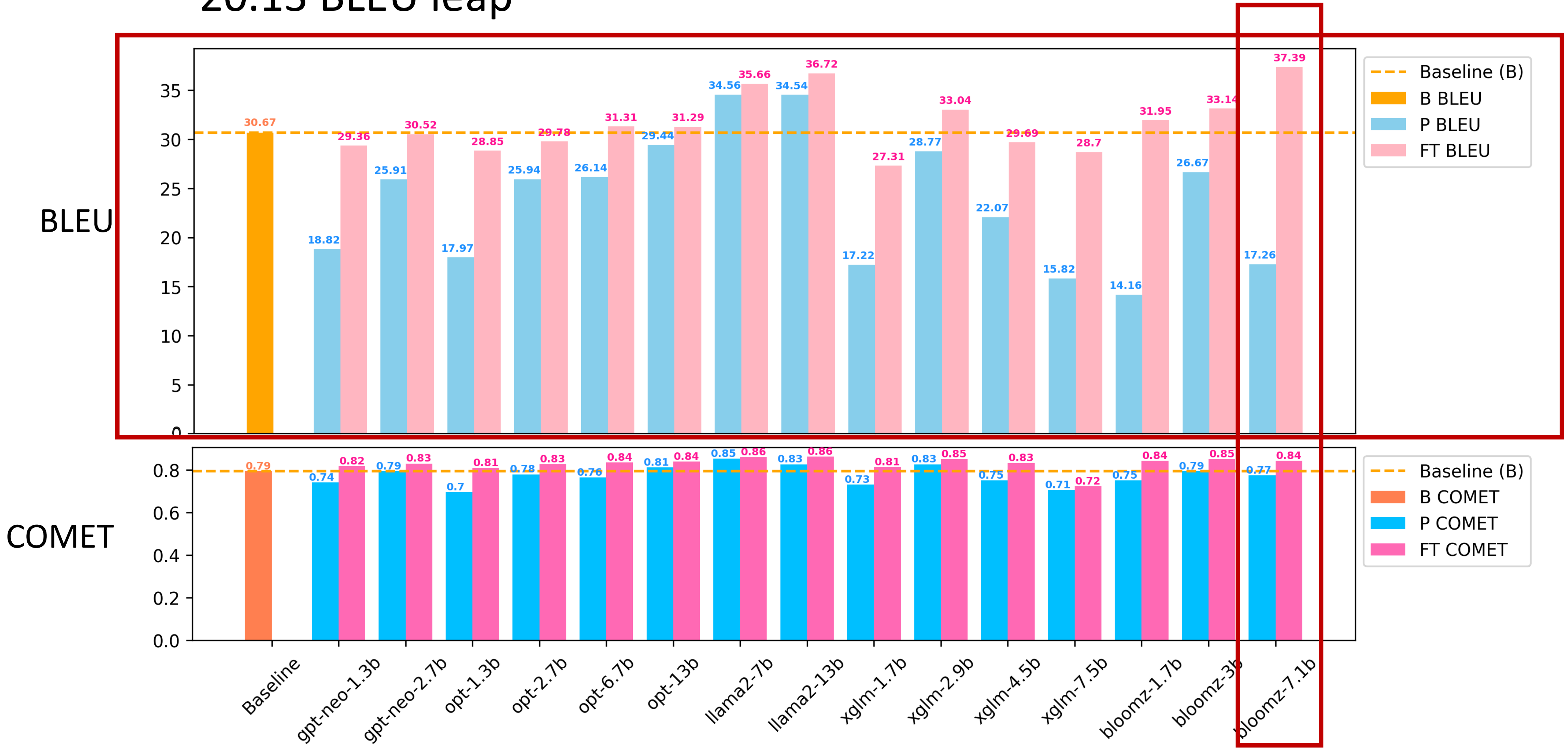
Prompting vs. Fine-tuning LLMs

- High COMET: LLMs produce semantically coherent translations



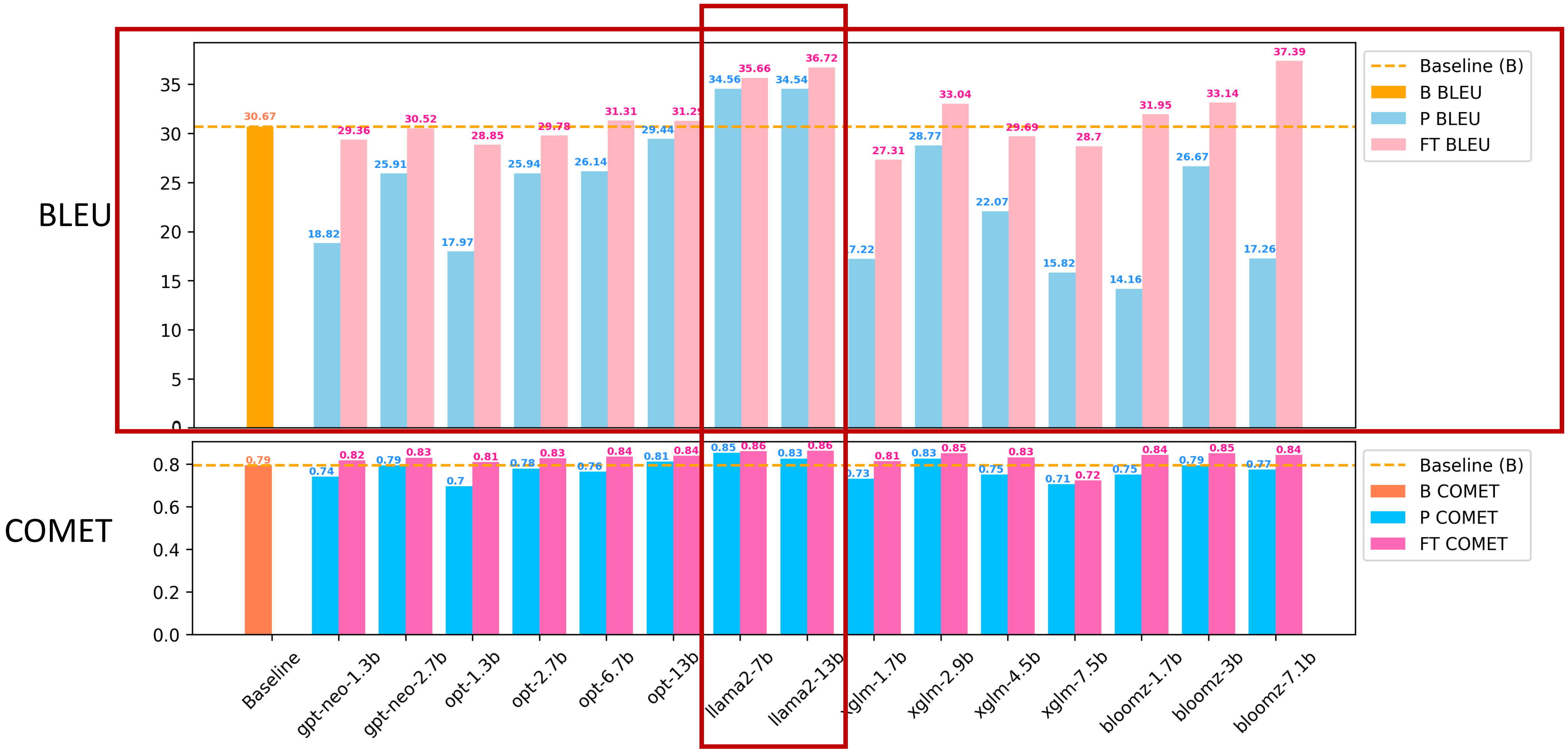
Prompting vs. Fine-tuning LLMs

- 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

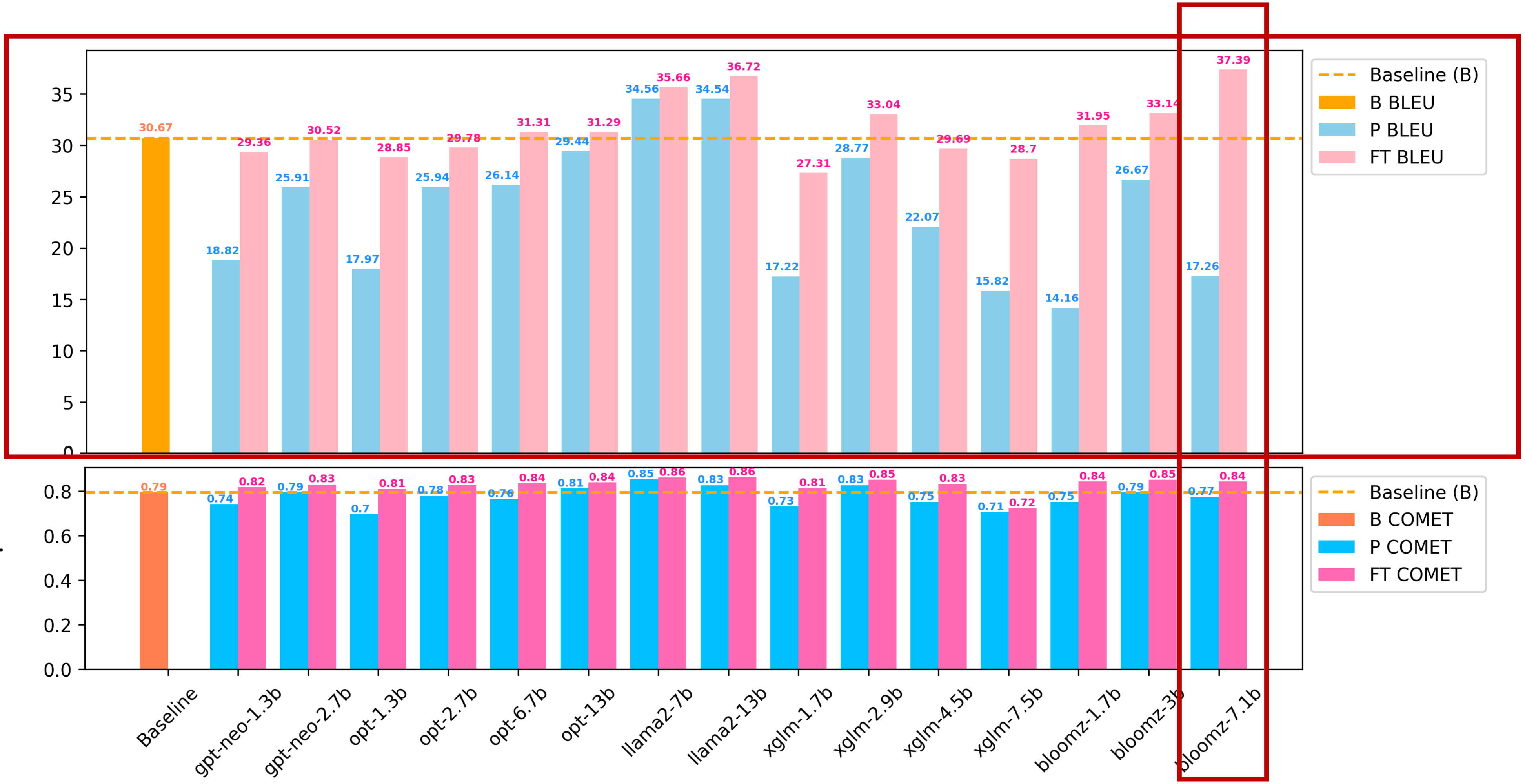


Prompting vs. Fine-tuning LLMs

- 8 out of 15 fine-tuned LLMs exceed Baseline, best: fine-tuned BLOOMZ-7.1B

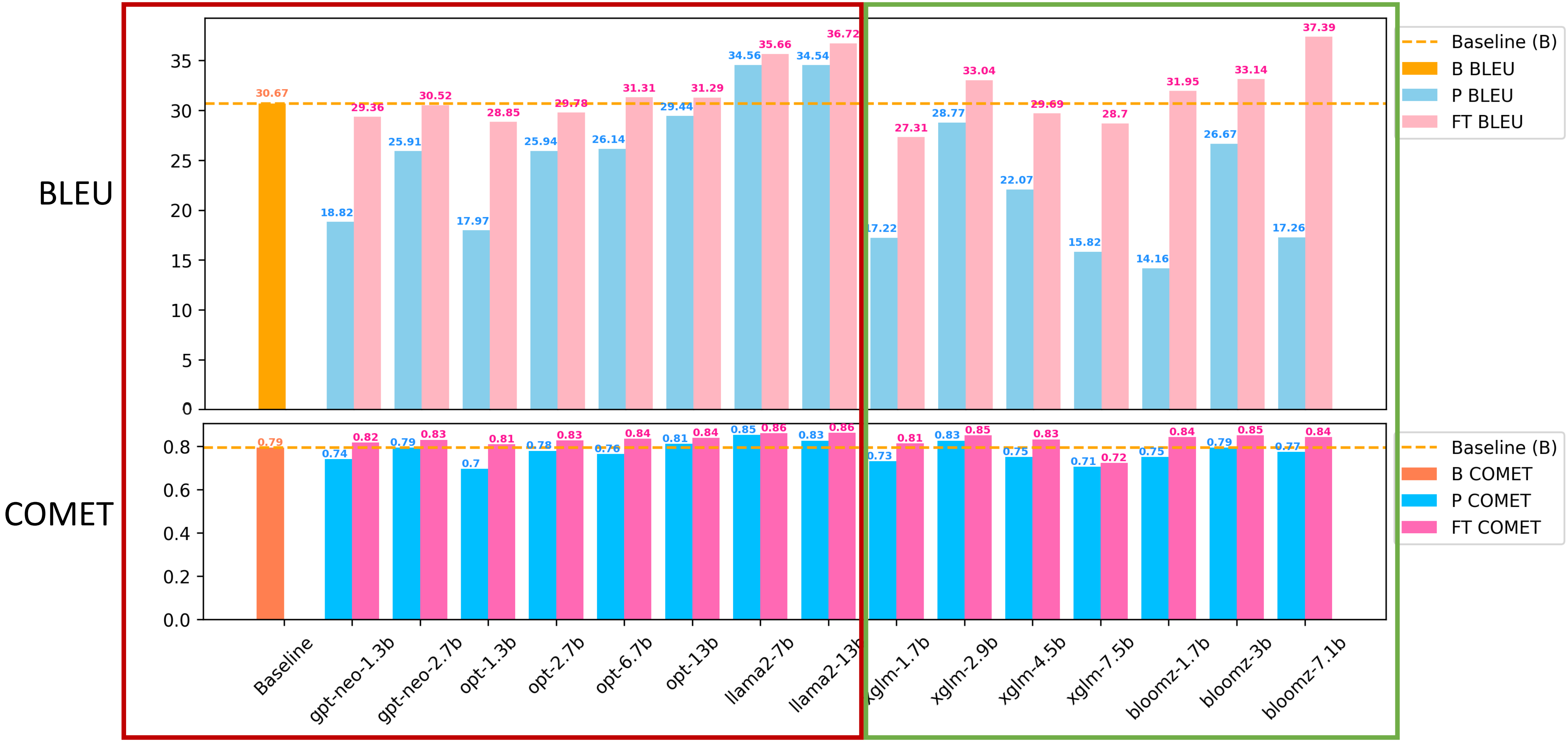
BLEU

COMET



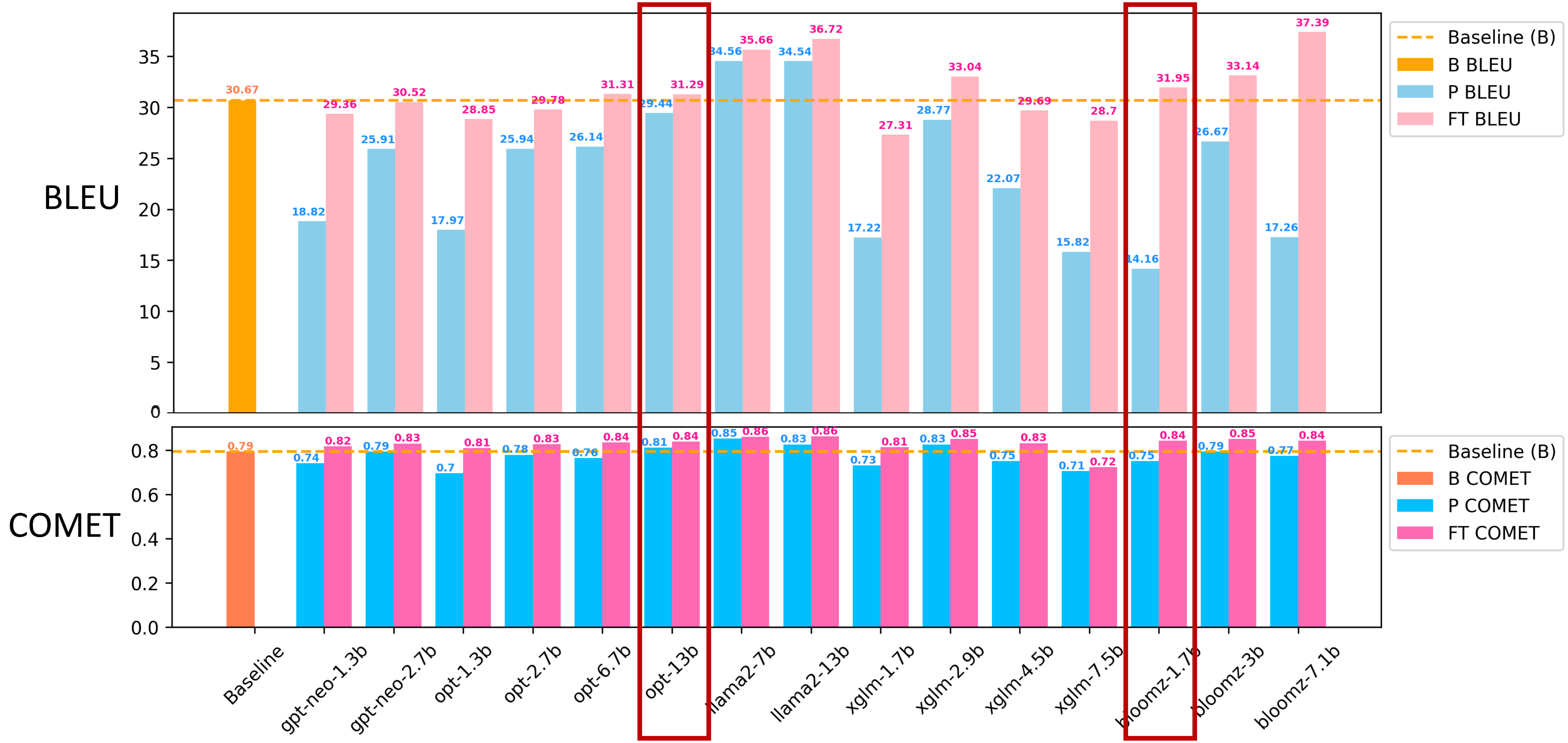
Prompting vs. Fine-tuning LLMs

- No clear advantage is discerned comparing English-centric and multilingual LLMs

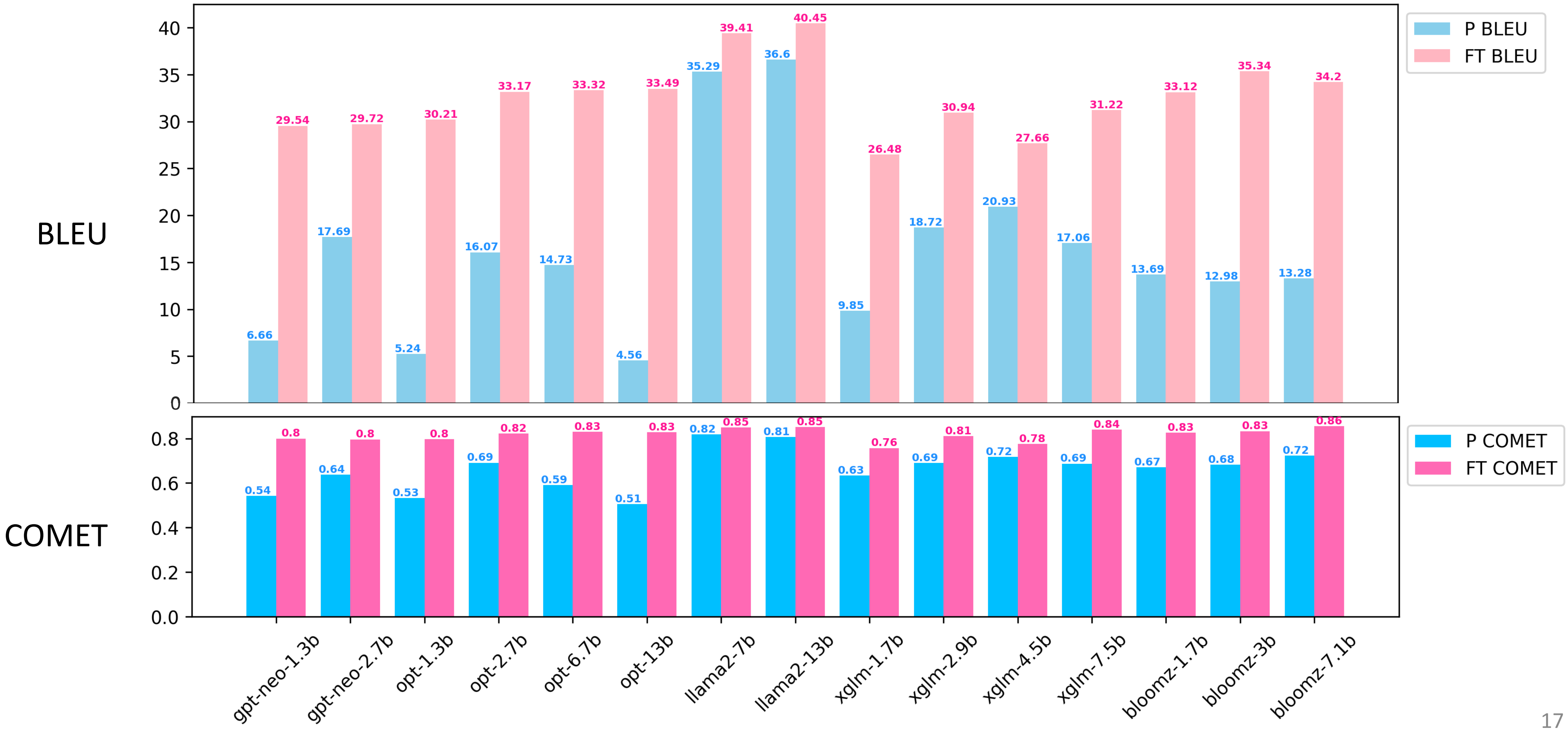


Prompting vs. Fine-tuning LLMs

- Bigger models do not necessarily outperform smaller ones: fine-tuned BLOOMZ-1.7B outperforms OPT-13B

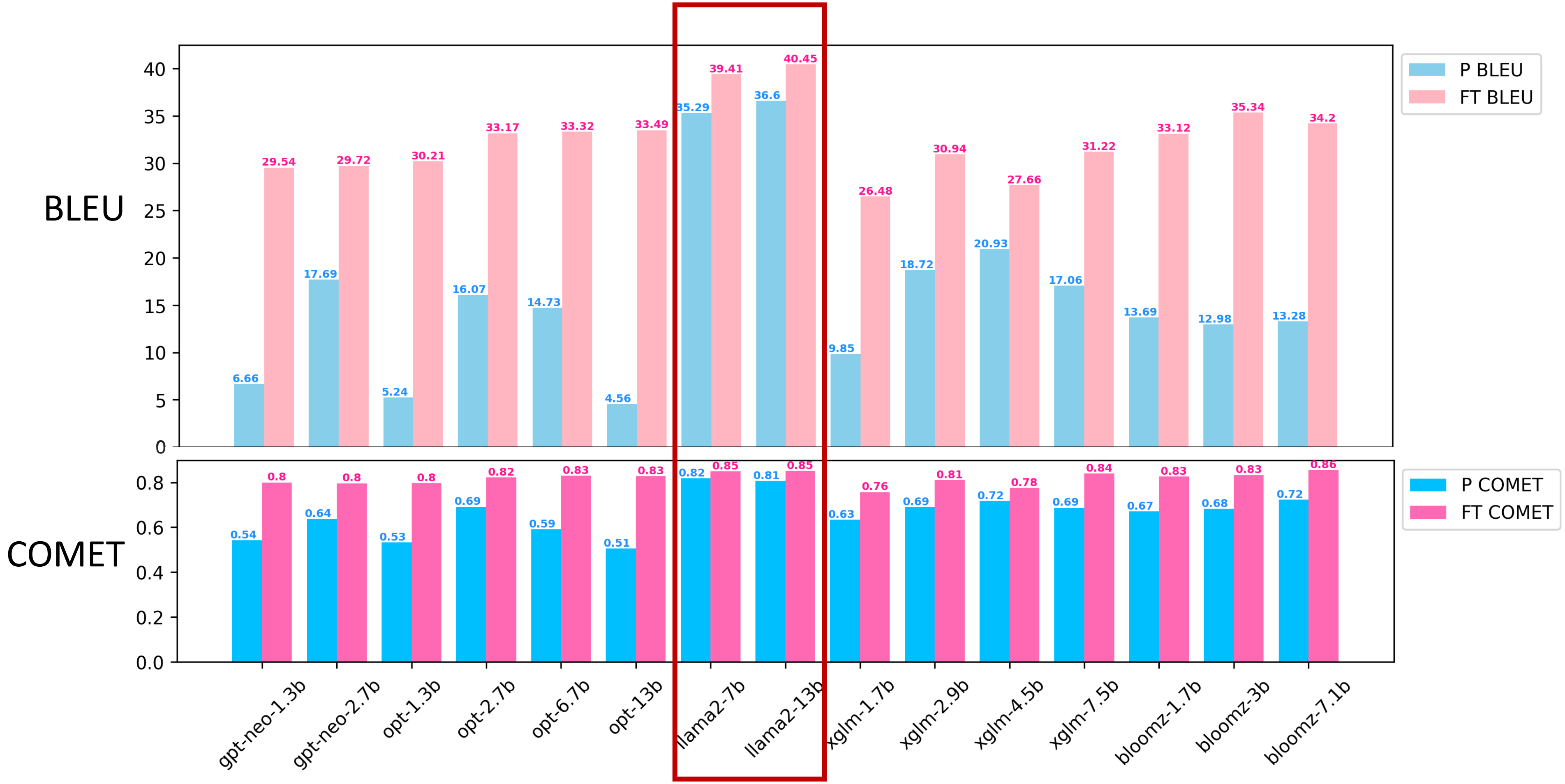


Document-level Translation



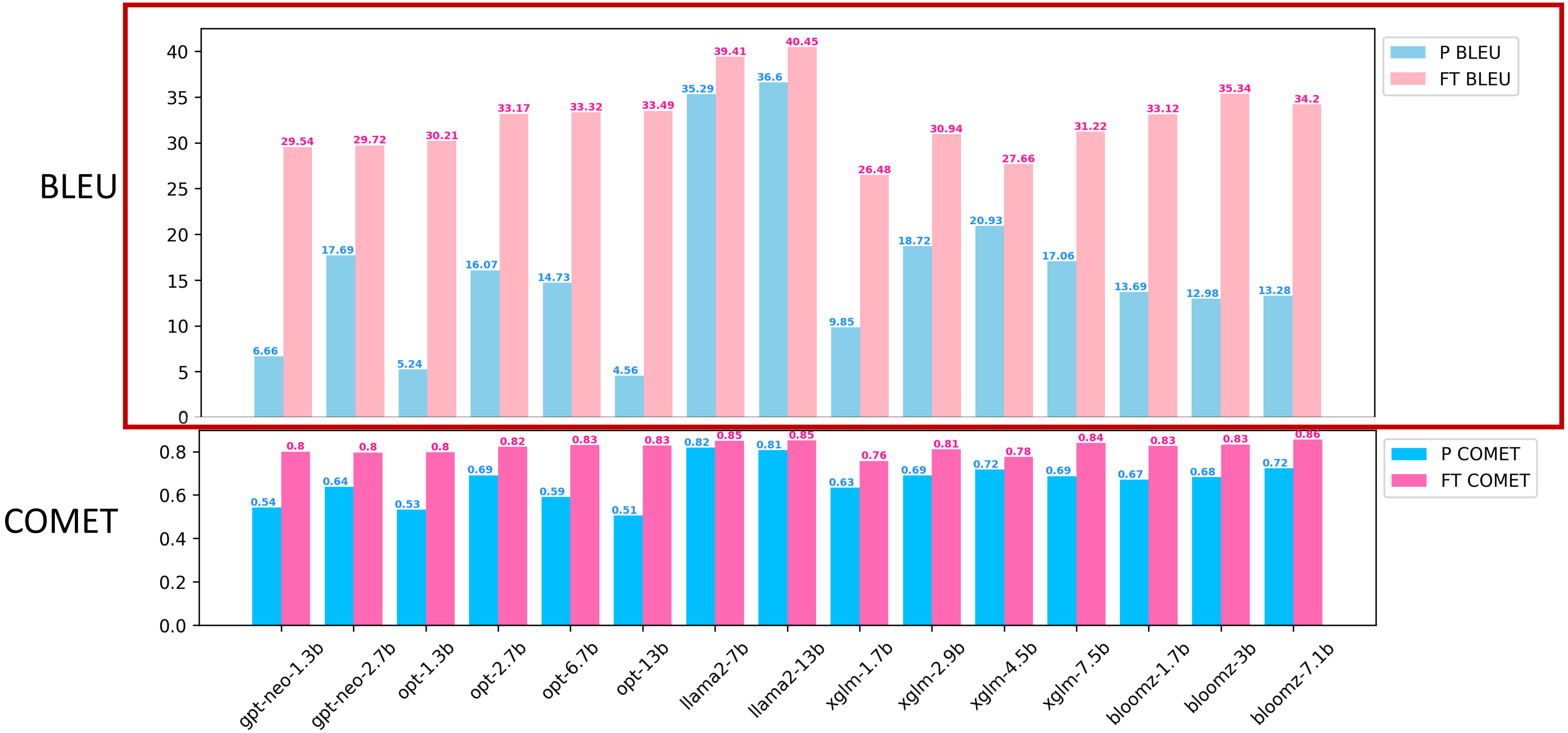
Document-level Translation

- Most LLMs struggle at document translation with prompting, except for LLaMA2



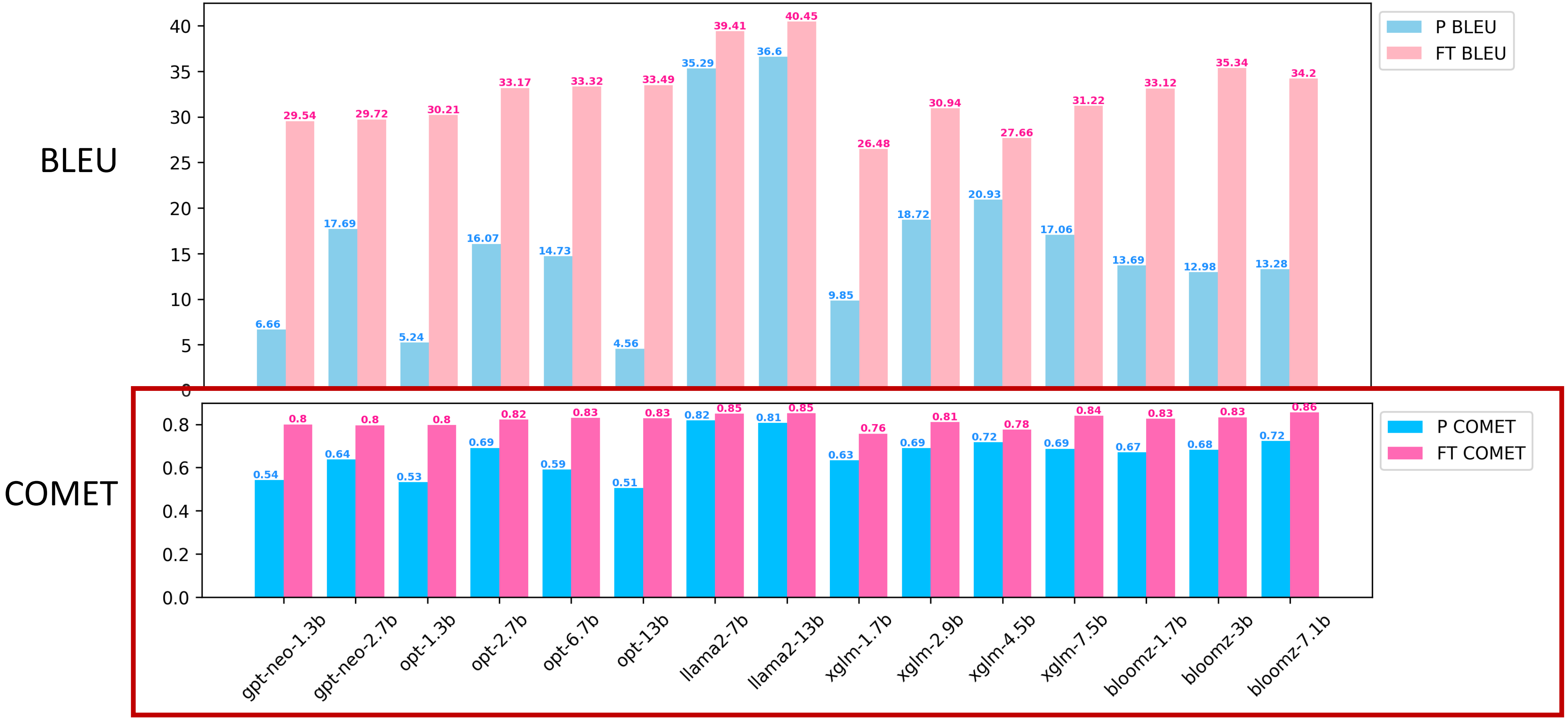
Document-level Translation

- 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.

	params(%)	#GPUs	time(hrs)
No QLoRA	27.40	4	52
QLoRA	0.02	1	10

Table 3: Fine-tuning *xglm-2.9b* with and without QLoRA to achieve the BLEU score of 30.05.[4] 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.