Towards effective disambiguation for Machine Translation with Large Language Models

WMT 2023

Vivek Iyer, Patrick Chen and Alexandra Birch

Background

- **WSD:** “The problem of multiple meanings” in MT (Weaver, 1947)
- Recently, we’ve seen modern NMT systems can struggle with WSD, especially with polysemous or rare word senses [2].
- **DiBiMT ambiguity benchmark** [2]: OPUS, mBART50, M2M100 show <30% accuracy for ambiguous word translation; Google and DeepL perform better at 50-60%.

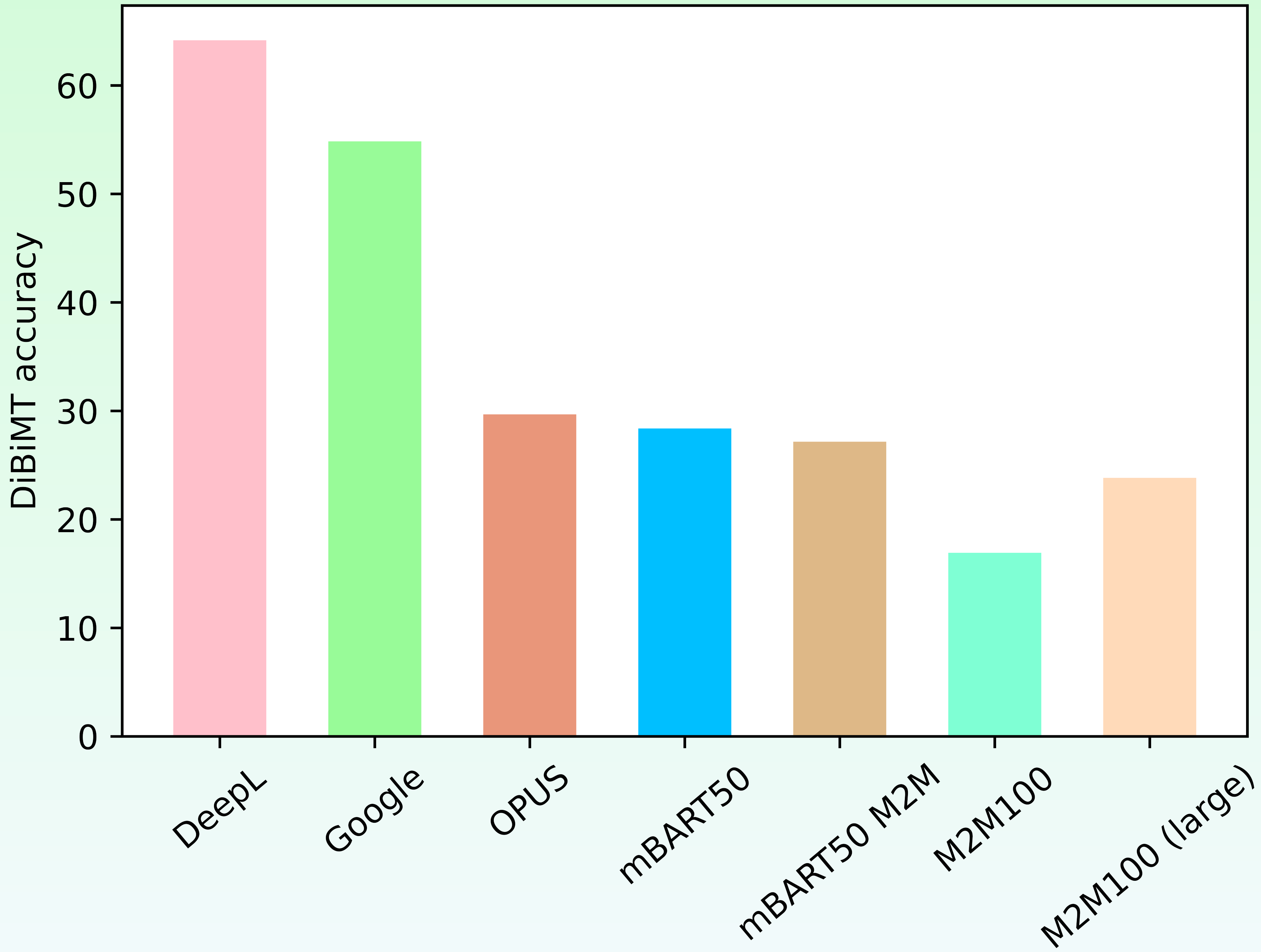


Figure 1: Disambiguation accuracy of some well-known MT systems [2]

Can LLMs bridge this gap? 👀

- **Challenge:** NMT systems, trained on narrow domain parallel text, can struggle with rarer word senses.
- **Advantage of LLMs:** More data, more contexts!
- **Disadvantage of LLMs:** Might also prefer fluency over accuracy ⚠️
- **Our Goal:** Detailed analysis of effectiveness of LLMs in **translation of ambiguous sentences.**

Source	The horse had a blaze between its eyes.
DeepL	那匹马的两眼之间有一团火焰。 (There is a flame between the horse's eyes.)
BLOOMZ (176B)	这匹马的眼睛之间有一道白线。 (There is a white line between the horse's eyes.)

Table 1: An example of English-to-Chinese translation involving an ambiguous term "blaze". For BLOOMZ, we use 1-shot prompting to obtain the translation.

Contributions

- Compare LLMs vs NMT systems on “ambiguous translation” of 5 languages
  - 12 NMT models: Commercial & Open-Source MT systems
  - 7 LLMs: Base & instruction-tuned models of varying {multilinguality, size}
- Adapt LLMs for disambiguation
  - ICL with *similar ambiguous contexts*
  - LoRA FT on curated *ambiguous corpora*
- Evaluate on FLORES-200 [7] to confirm gains in overall MT quality

Definitions

i.e. Being clear about what I mean (MUST prevent irony)

- **Word Senses**: contextualized meaning of a word
- **Ambiguous MT**: translating lexically "ambiguous" words in a sentence
    - Rare senses (low "sense frequency")
    - Polysemous senses (high "polysemy degree")

Evaluation Setup

BASELINES We select the leading NMT systems and most widely used LLMs[1] for evaluation.

Table 1a) NMT Systems

Category	System	# Params
Commercial	Google Translate[1]	Unknown
	DeepL[2]	
Open-source	OPUS [8]	74M
	mBART50 [9]	611M
	M2M100 [10]	418M
		1.2B
	NLLB-200 [7]	0.6B
		1.3B
		3.3B
		54B

Table 1b) LLMs

Category	System	# Params
BLOOM family	BLOOM [11]	7B
		176B
	BLOOMZ [12]	7B
		176B
LLaMa family	LLaMa [13]	7B
		65B
	Alpaca [14]	7B

LANGUAGES

DiBiMT pairs	En→Es	En→It	En→Zh	En→De	En→Ru

[1]at the time of experiment formulation

The DiBiMT Benchmark

- DiBiMT:
  - 500 ambiguous sentences
  - 1 ambiguous word per sentence
  - Human-curated and verified “***Good”*** + “***Bad”*** translations of this word
- Given an ambiguous word in a sentence:
  - Accuracy = %Good / (%Good + %Bad)
  - MISS cases: Neither “Good”, Nor “Bad”. Unknown!

Naive setting: k-shot prompting

- We choose demonstrations randomly from the dev set.

Translate the following sentence from {src_lang} to {tgt_lang}: {src_demo$_1$}
The translation in {tgt_lang} is: {tgt_demo$_1$}

⋮ k demonstrations

Translate the following sentence from {src_lang} to {tgt_lang}: {src_demo$_k$}
The translation in {tgt_lang} is: {tgt_demo$_k$}

[FOUNDATION LLM]

Translate the following sentence from {src_lang} to {tgt_lang}: {src_test}
The translation in {tgt_lang} is:

OR

[INSTRUCTION-TUNED LLM]

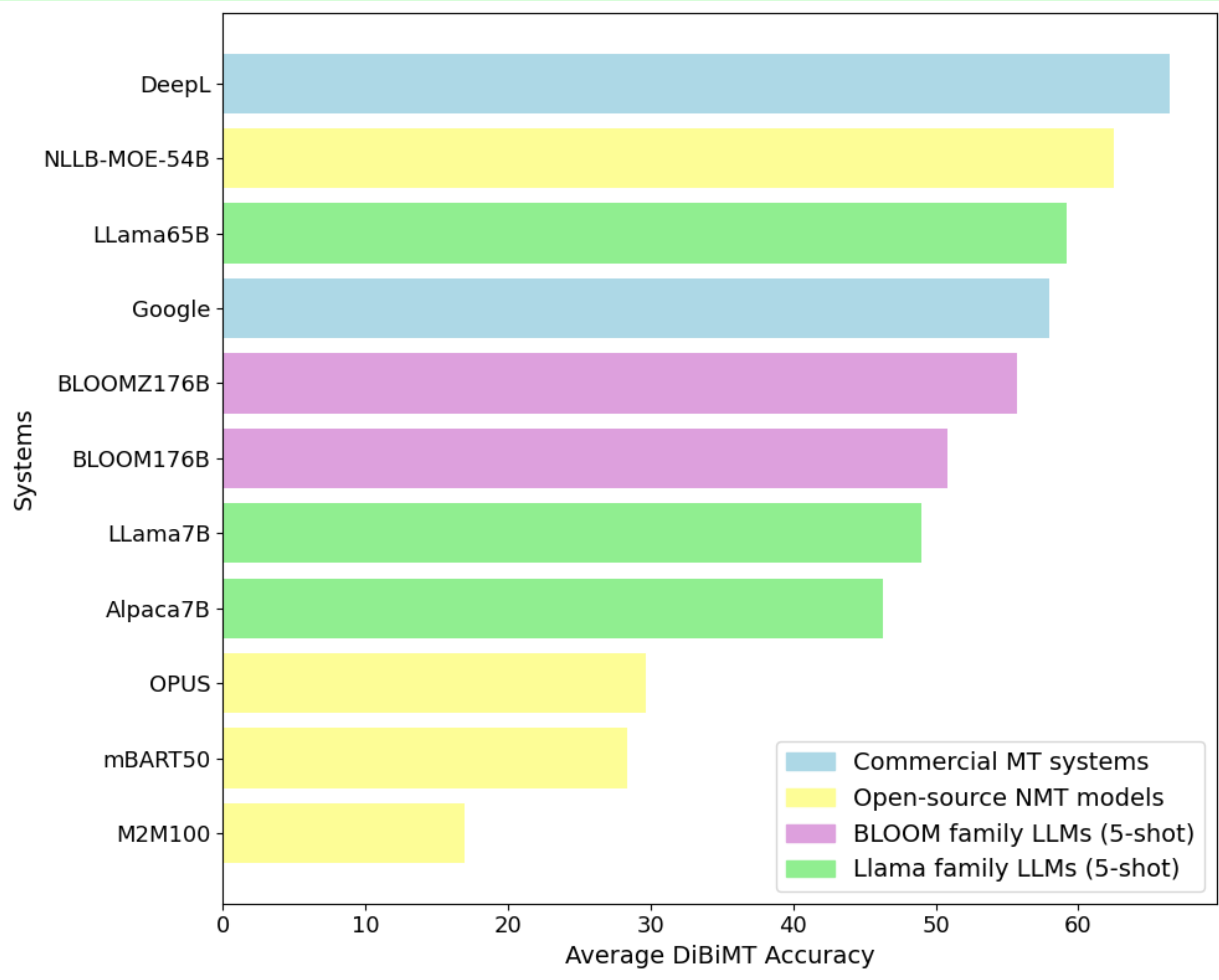
Translate the following sentence from {src_lang} to {tgt_lang}: {src_test}
Can you translate the input sentence to {tgt_lang}?

Figure 2: Template used for k-shot prompting

Results (random k-shot prompting)

On average:

- Naive 5-shot prompting of int8 quantised LLMs outperforms many open-source NMT models
- Matches Google Translate
- Slightly underperforms SOTA systems (DeepL and NLLB) on the DiBiMT benchmark.



More nuanced results

Table 2: DiBiMT accuracy for languages a) seen and b) unseen[1] by LLMs during pretraining.

System	En-Es	En-It	En-Zh	En-Ru	En-De
DeepL	63.91	65.47	58.42	67.53	76.64
Google	54.73	53.59	52.09	62.03	67.35
NLLB 54B	61.33	67.19	48.02	67.88	67.97
LLaMA 7B	56.33	48.66	27.92	56.83	55.26
LLaMA 65B	60.78	63.47	42.49	66.31	62.98
BLOOM 176B	65.53	45.99	61.73	42.92	38.06
BLOOMZ 176B	68.55	49.22	63.36	52.6	44.94

Seen langs

Unseen langs

Orange: Best NMT score

Violet: Best LLM score

Underline: Best Overall score

- **Seen languages:** BLOOMZ leading in 2 languages (En-Es and En-Zh)
- **Unseen languages:** Worse than NMT systems; hallucination
- Foundation LLMs << Instruction-tuned LLMs
- Scale ↑, Performance ↑

[1]Not intentionally included in the pretraining set

So, what do LLMs get wrong?

A qualitative comparison

- 20 predictions for En-Zh DiBiMT: **DeepL** and **BLOOMZ 176B (1-shot)**
  - DeepL: Better overall at MT but can be too literal
  - BLOOMZ: Contextual translations but can omit details

#	Source	BLOOMZ	DeepL
1	He's not in my line of business.	他不是我的生意。 He is not my business. (did not translate “line")	他不在我的业务范围内 . He is out of my business (area).
2	He waited impatiently in the blind.	他焦急地等待着。 He waited anxiously (did not translate "in the blind")	他在盲人区等得不耐烦 . He waited impatiently in the area designated to be used by blind people
3	How much head do you have at the Glens Falls feeder dam?	你有多少头牛在格伦瀑布的蓄水池里? How many cows do you have in the reservoir/cistern at Glen Falls?	格伦瀑布支坝的水头有多大? How big is the water head at the Glen Falls branch dam?
4	The mechanic bled the engine.	机械师在引擎上流血。 the mechanic is bleeding on the engine	机械师给发动机放气 . the mechanic is getting rid of air from the engine

Table 11: BLOOMZ ERROR cases on DiBiMT

Improvement 1: In-context Learning with similar ambiguous contexts

- Demonstrations = other “same-sense” occurrences of the ambiguous word in dev corpus
- Larger LLMs gain more
- More examples, more gains!

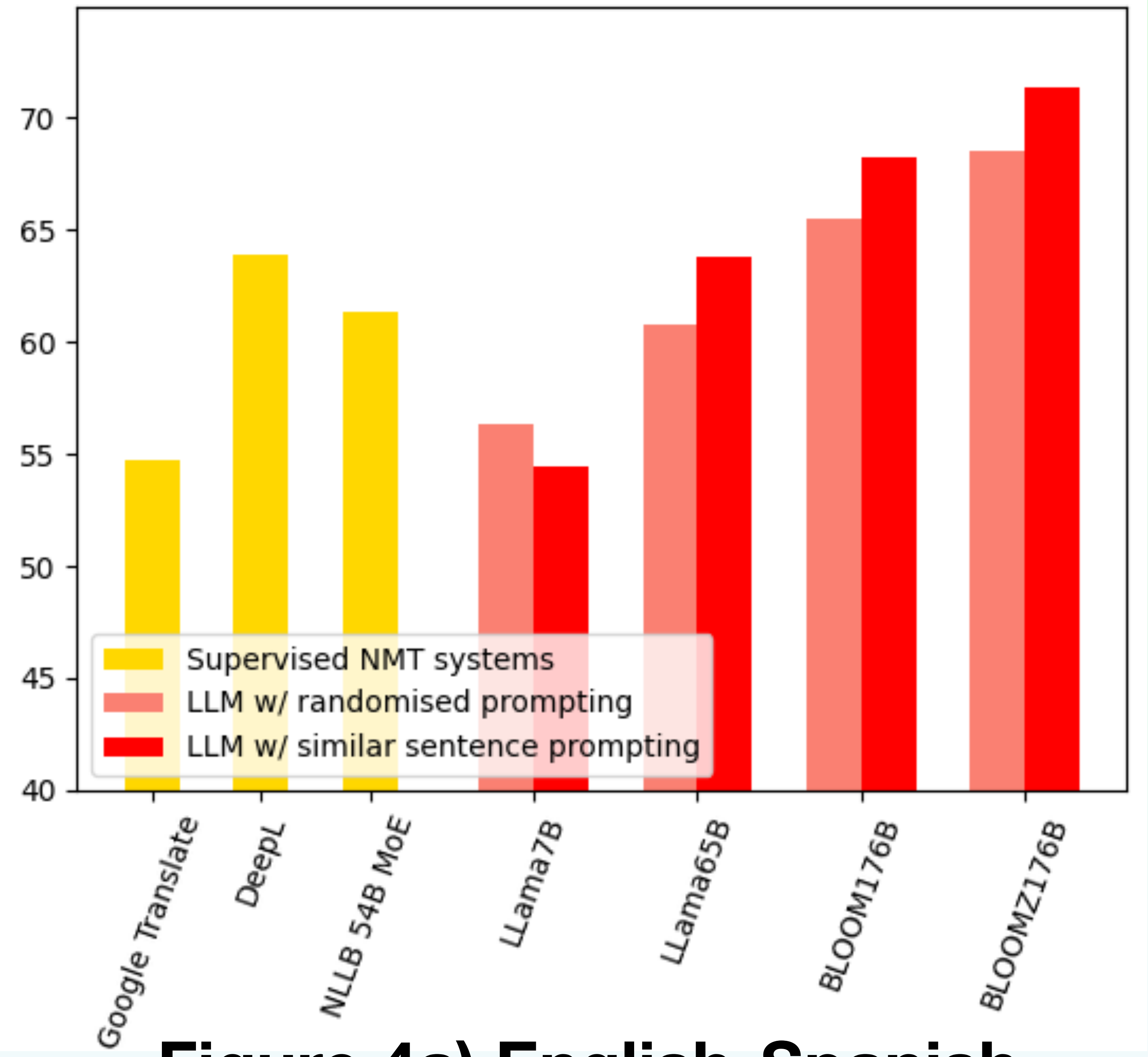


Figure 4a) English-Spanish

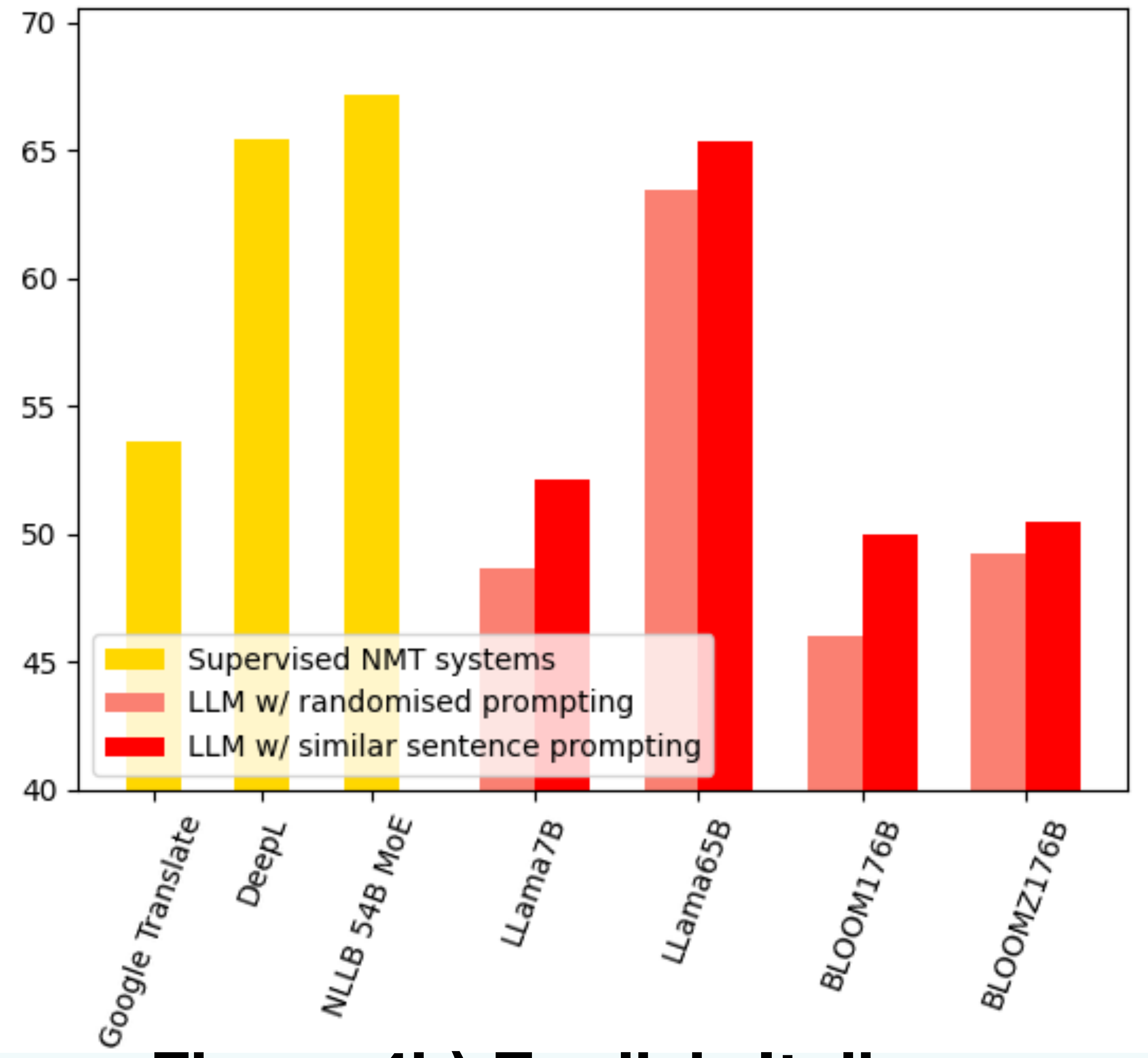


Figure 4b) English-Italian

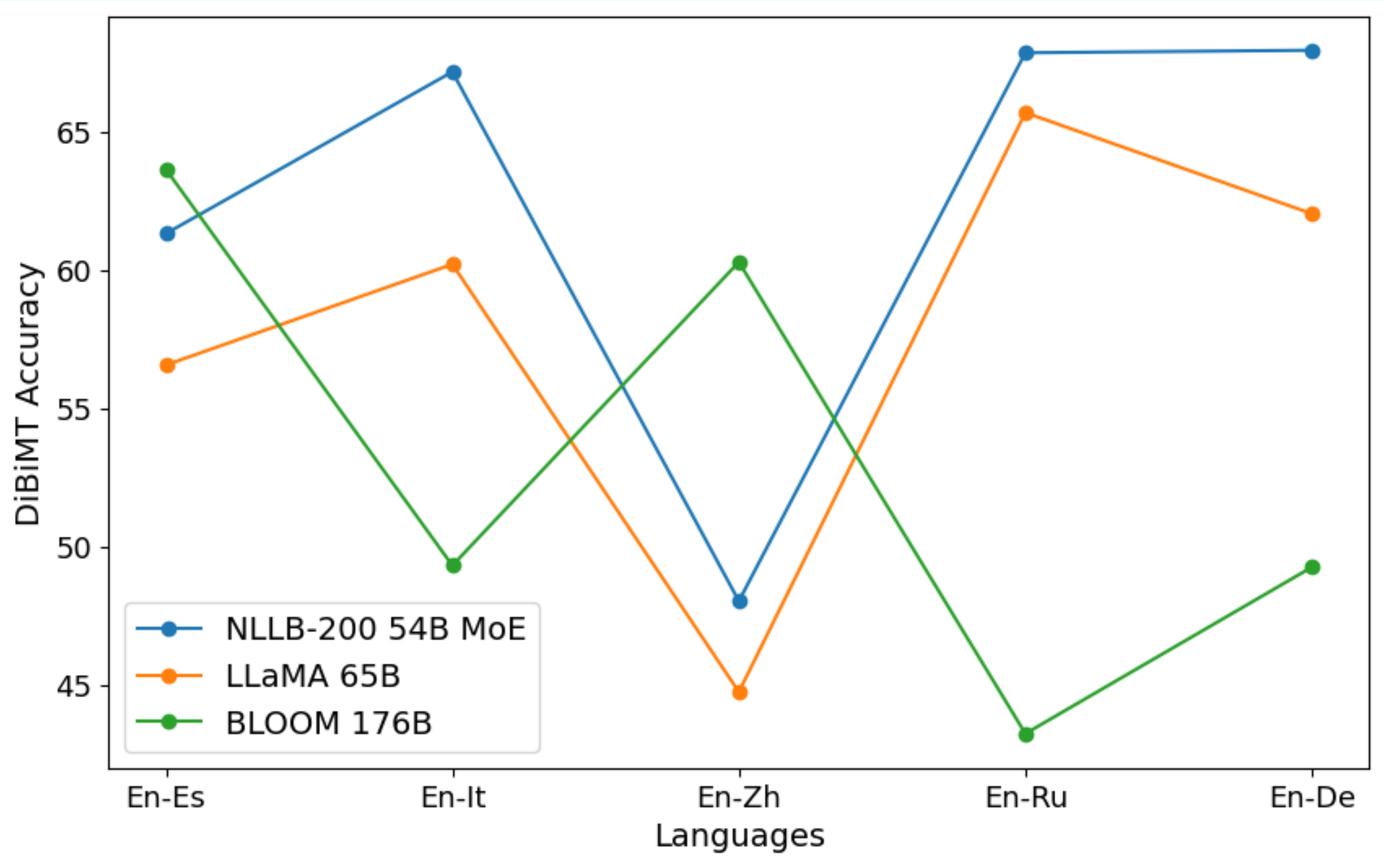
Improvement 2: LoRA Fine-Tuning

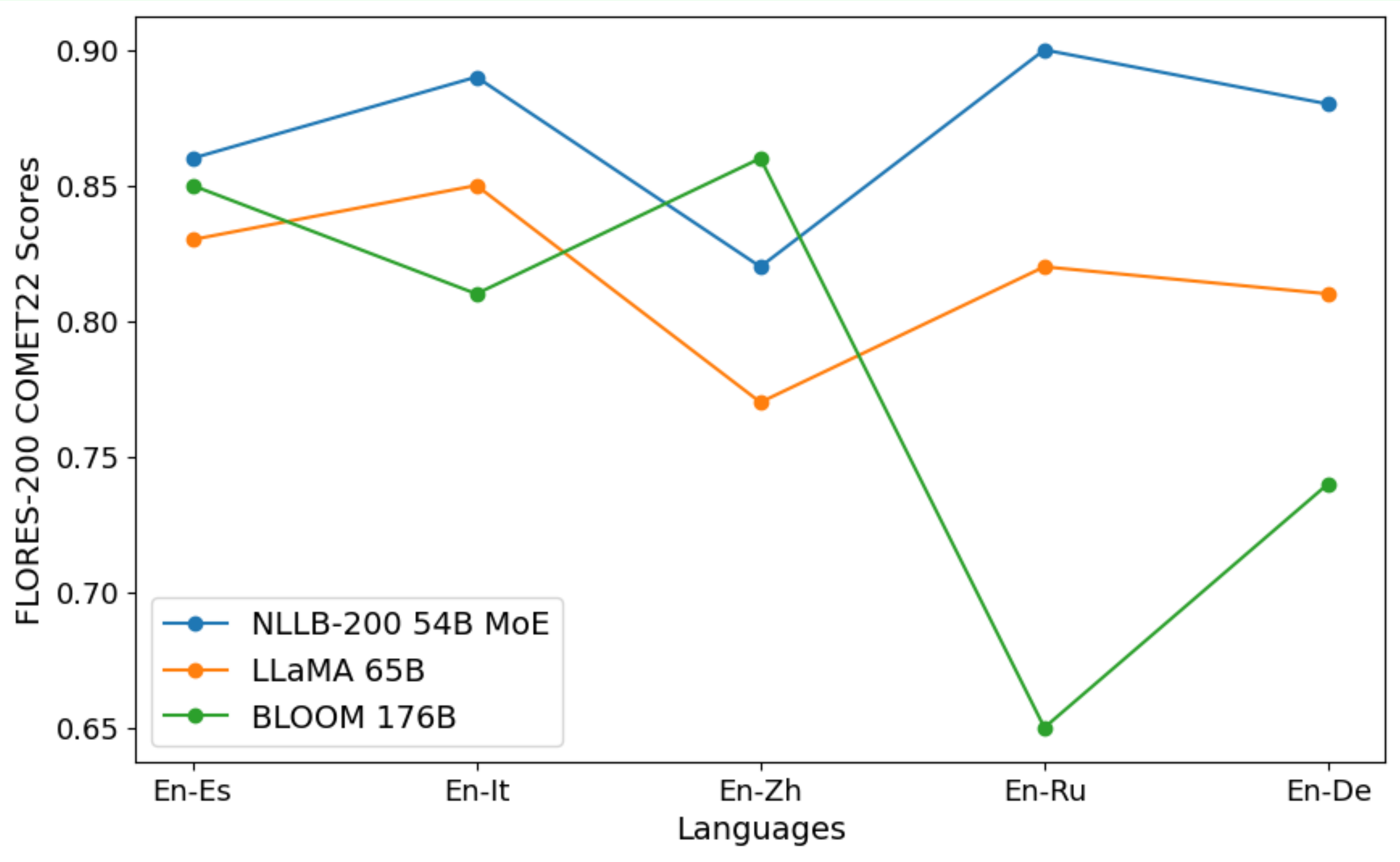
- We curate **Ambiguous Europarl** (https://data.statmt.org/ambiguous-europarl) by filtering out most ambiguous sentences from Europarl
- LoRA FT: Alpaca, BLOOM and BLOOMZ 7B
- Improves Accuracy.
    - 2 epochs with ~65K sentences are sufficient

System	Alpaca 7B	BLOOM 7B	BLOOMZ 7B	Alpaca 7B	BLOOM 7B	BLOOMZ 7B
w/o FT	49.75	55.69	60.87	45.24	28.79	40.68
FT	63.27	57.86	60.39	59.62	37.72	39.73

FLORES 200 Evaluation

Q1. How do these trends extend to overall MT quality?

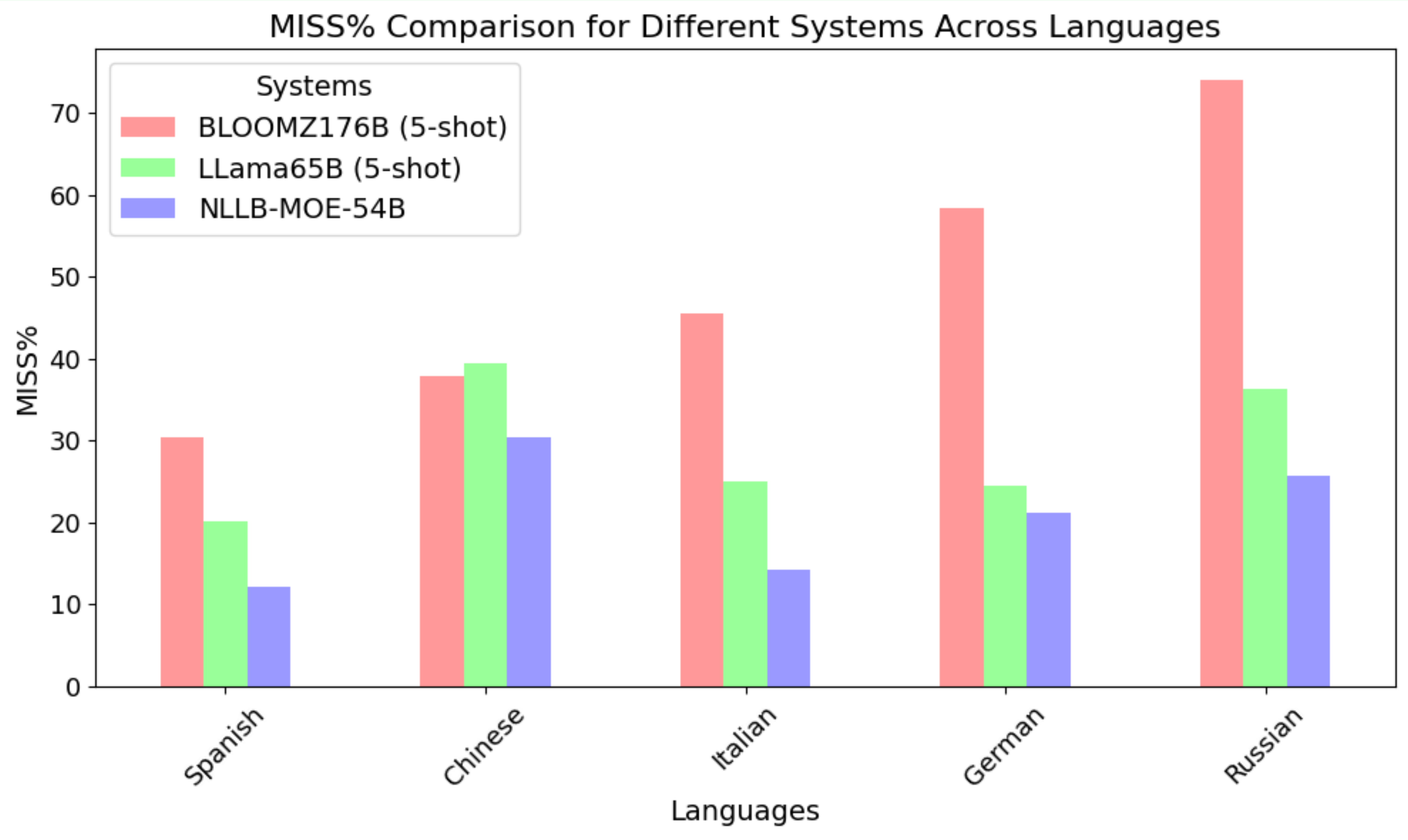




- Similar trends; COMET22 is **less drastic** than DiBiMT accuracy
    - **NLLB-200 54B MoE >>** 1-shot LLaMa 65B
    - BLOOM ~ NLLB on **seen languages** (En-Es & En-Zh).

Wait…. we MISSed something

- **What about MISS%?** Translations that are neither Good/Bad - unknown category!
- High MISS% for BLOOMZ on "unseen" languages! Less for Llama



Conclusion

- Open-source LLMs[1] are competitive, but do not consistently beat NMT models like NLLB-200/DeepL. Reasons:
    - Lack of multilinguality
    - Lack of instruction-tuning at scale
    - Hallucination (omission, wrong language etc.)
- But, still pretty darn promising!
    - More flexible and adaptable than MT systems
    - Can tune for WSD with a) ICL w/ similar contexts, and b) LoRA FT on curated corpora
    - Disambiguation tuning improves overall MT quality too

[1] The ones we tested (<Aug 2023)

THANK YOU!

Questions are unambiguously welcome :)

References

[1] Weaver, W. (1952). Translation. In Proceedings of the Conference on Mechanical Translation.

[2] Campolungo, N., Martelli, F., Saina, F., & Navigli, R. (2022, May). DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4331-4352).

[3] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

[4] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research 2022 (pp. 2835-8856)*.

[5] Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2022). Prompting palm for translation: Assessing strategies and performance. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

[6] Zhang, B., Haddow, B., & Birch, A. (2023). Prompting large language model for machine translation: A case study.In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 1722, 41092–41110.

[7] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

[8] Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT--Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

References

[9] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics

[10] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, *22*(1), 4839-4886.

[11] Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... & Bari, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

[12] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., ... & Raffel, C. (2022). Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

[13] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[14] Taori, R., Gulrajani I., Zhang T., Dubois Y., Li X., Guestrin C., Liang P., and Hashimoto T.. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca

[15] Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., & Cecconi, F. (2021, August). Ten Years of BabelNet: A Survey. In IJCAI (pp. 4559-4567)

[16] Barba, E., Pasini, T., & Navigli, R. (2021, June). ESC: Redesigning WSD with extractive sense comprehension. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4661-4672).