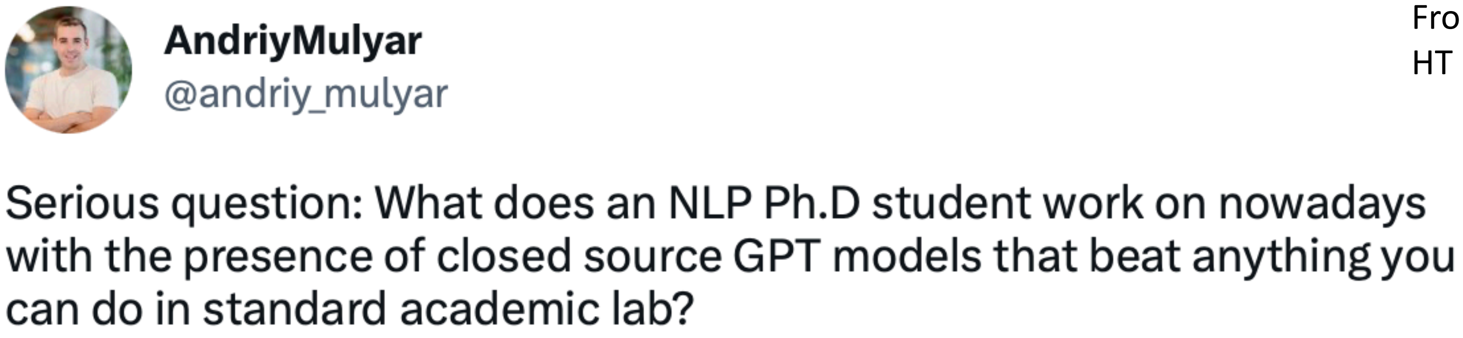
Large Language Models (LLMs) for Machine Translation Adaptation

Suzanna Sia and Kevin Duh

Johns Hopkins University

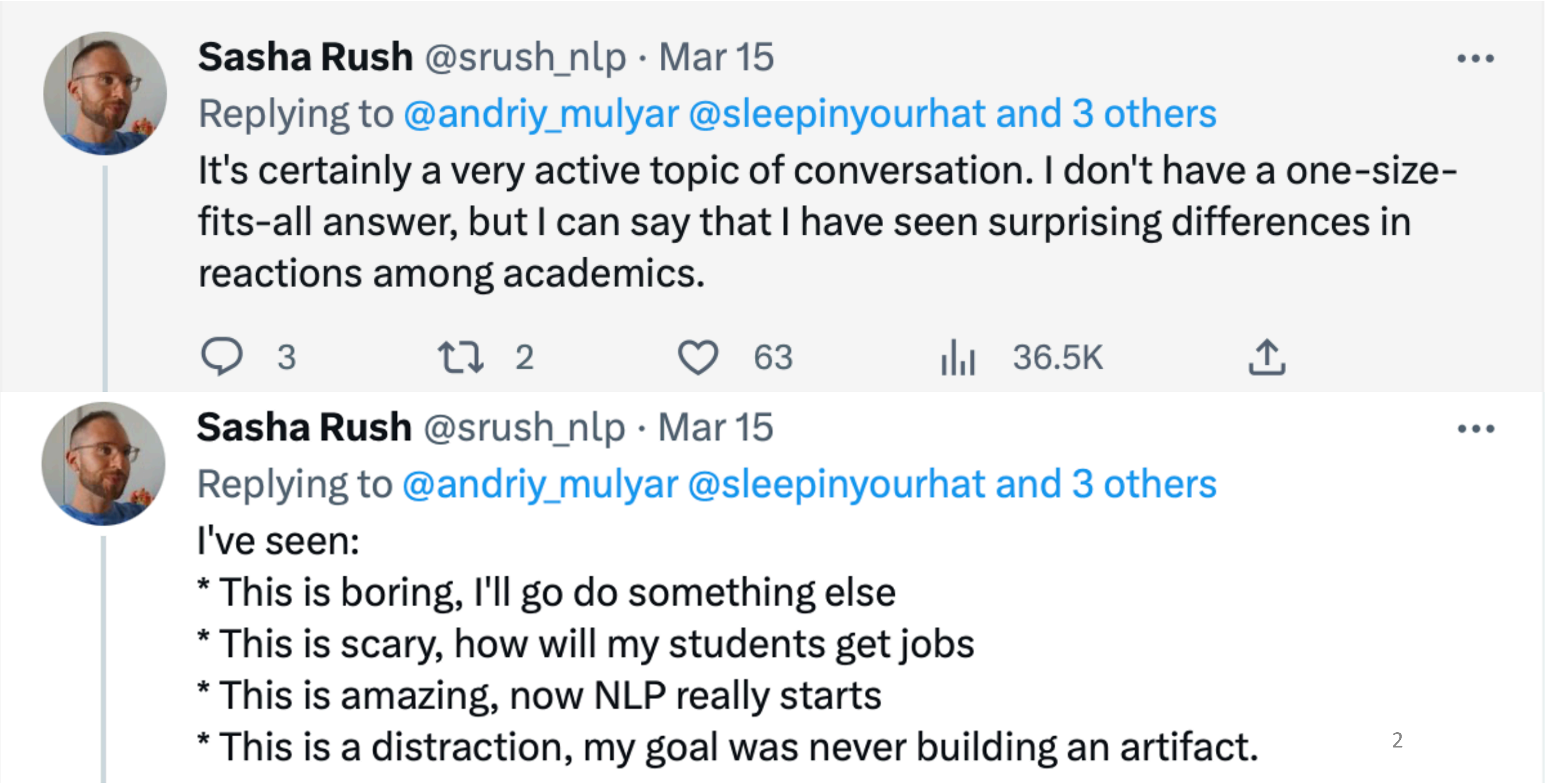
Or: Trying to make sense of all the talk around ChatGPT

From: https://twitter.com/andriy_mulyar/status/1636139257805828096
HT @mdredze



AndriyMulyar
@andriy_mulyar

Serious question: What does an NLP Ph.D student work on nowadays with the presence of closed source GPT models that beat anything you can do in standard academic lab?



Sasha Rush @srush_nlp · Mar 15

Replying to @andriy_mulyar @sleepinyourhat and 3 others

It's certainly a very active topic of conversation. I don't have a one-size-fits-all answer, but I can say that I have seen surprising differences in reactions among academics.

Sasha Rush @srush_nlp · Mar 15

Replying to @andriy_mulyar @sleepinyourhat and 3 others

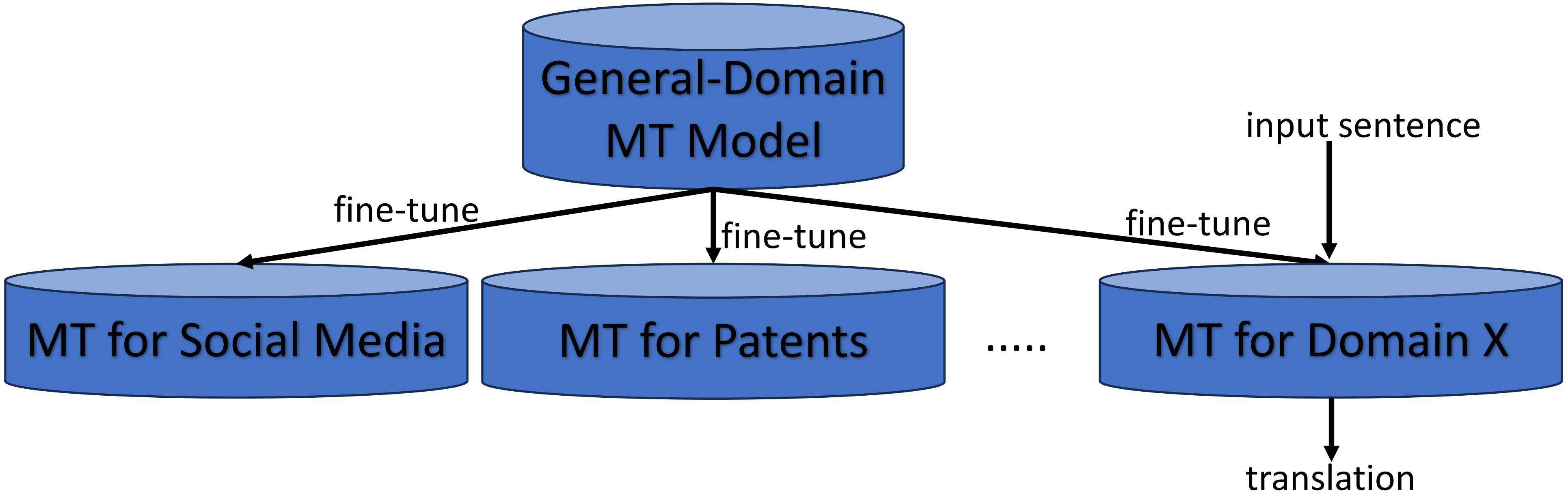
I've seen:
* This is boring, I'll go do something else
* This is scary, how will my students get jobs
* This is amazing, now NLP really starts
* This is a distraction, my goal was never building an artifact.

My reaction to LLMs (currently)

- This is both exciting and risky. I’m hedging my bets (25% effort)
- Will off-the-shelf LLM replace dedicated Machine Translation (MT)?
  - I think not
  - LLM BLEU scores are surprisingly good, but dedicated MT trained in a supervised way on bitext still better in terms of accuracy and cost.
- Does LLM enable new applications in translation?
  - Yes.
  - Focus of this talk: Explore On-the-Fly adaptation

Domain Adaptation

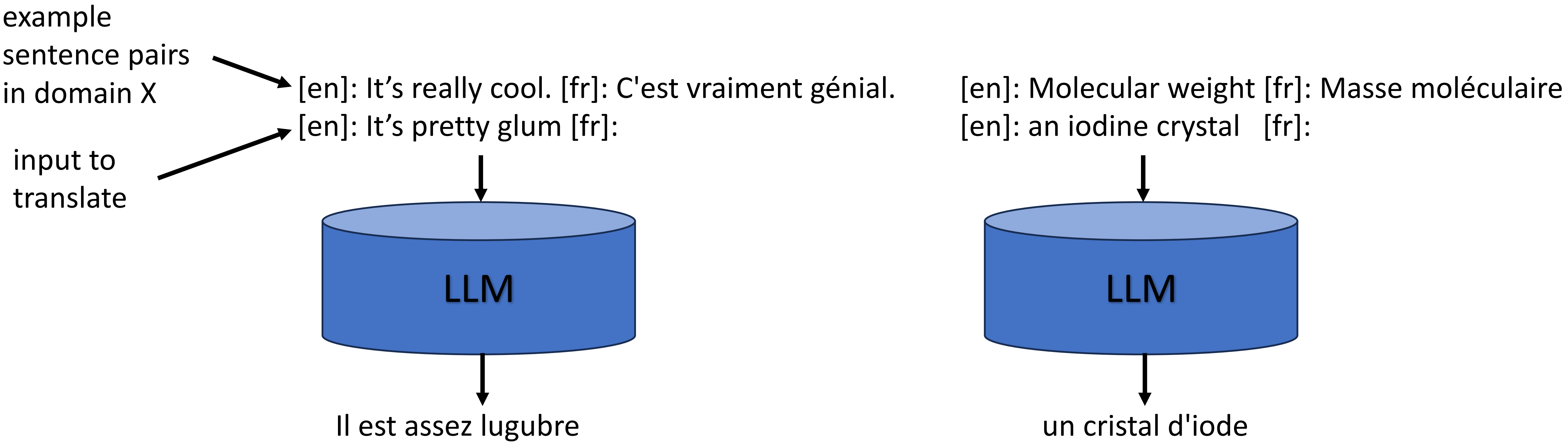
- A well-established method for domain adaptation is fine-tuning, but:
  - Requires pre-specified definition of domain
  - Generates a new model per domain



- This is static; can't prepare too many adapted models in practice.

The Promise of On-the-Fly Adaptation via In-Context Learning

- “In-Context Learning” is the ability of LLM to perform new tasks when given a few examples in the prompt.
- Only one model. Enables fine-grained domains and personalization



example sentence pairs in domain X

input to translate

[en]: It’s really cool. [fr]: C'est vraiment génial.
[en]: It’s pretty glum [fr]:

LLM

Il est assez lugubre

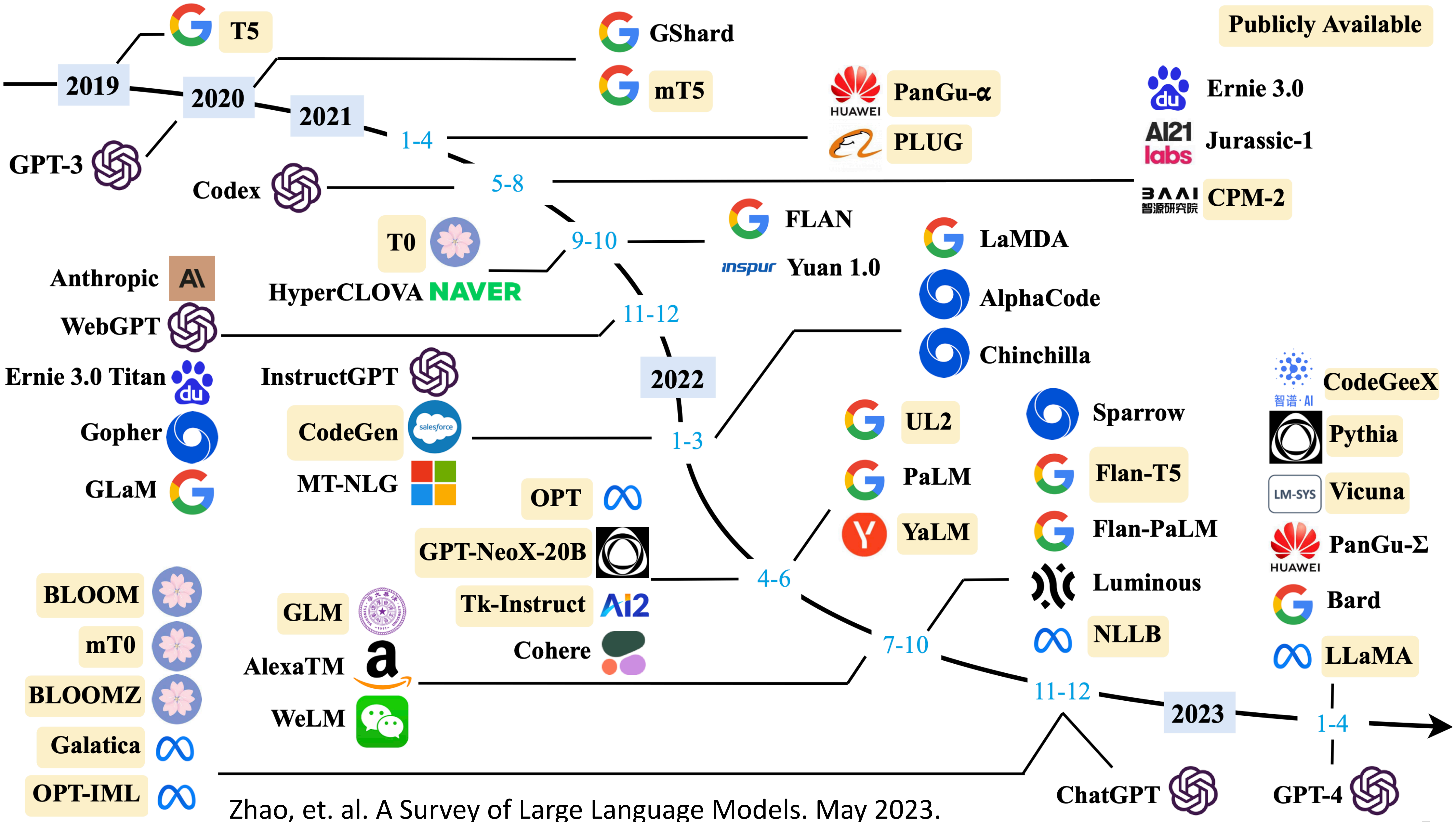
[en]: Molecular weight [fr]: Masse moléculaire
[en]: an iodine crystal [fr]:

LLM

un cristal d'iode

Outline

1. Brief background on LLMs
2. Two experiments in on-the-fly adaptation
   - Domain adaptation
   - Document-level adaptation
3. Improving LLMs for MT
   - Soft prompt tuning
4. Discussion



Publicly Available

2019

T5

GPT-3

2020

GShard

mT5

2021

1-4

PanGu-α

PLUG

Codex

5-8

Ernie 3.0

Jurassic-1

CPM-2

T0

HyperCLOVA NAVER

9-10

FLAN

Yuan 1.0

Anthropic

WebGPT

11-12

Ernie 3.0 Titan

Gopher

GLaM

LaMDA

AlphaCode

Chinchilla

InstructGPT

CodeGen

MT-NLG

2022

1-3

UL2

PaLM

YaLM

OPT

GPT-NeoX-20B

4-6

Tk-Instruct

Cohere

GLM

AlexaTM

WeLM

7-10

Sparrow

Flan-T5

Flan-PaLM

Luminous

NLLB

BLOOM

mT0

BLOOMZ

Galatica

OPT-IML

11-12

ChatGPT

2023

CodeGeeX

Pythia

Vicuna

PanGu-Σ

Bard

LLaMA

1-4

GPT-4

Zhao, et. al. A Survey of Large Language Models. May 2023. https://arxiv.org/pdf/2303.18223.pdf

BERT vs GPT-3

- Both trained as “Language Models” but something seems fundamentally different

Pre-train on large data
+ Fine-tune on Task A
= Great performance on Task A

BERT

Pre-train on large data
+ Scale Up
= Emergent ability on many tasks (AGI?)

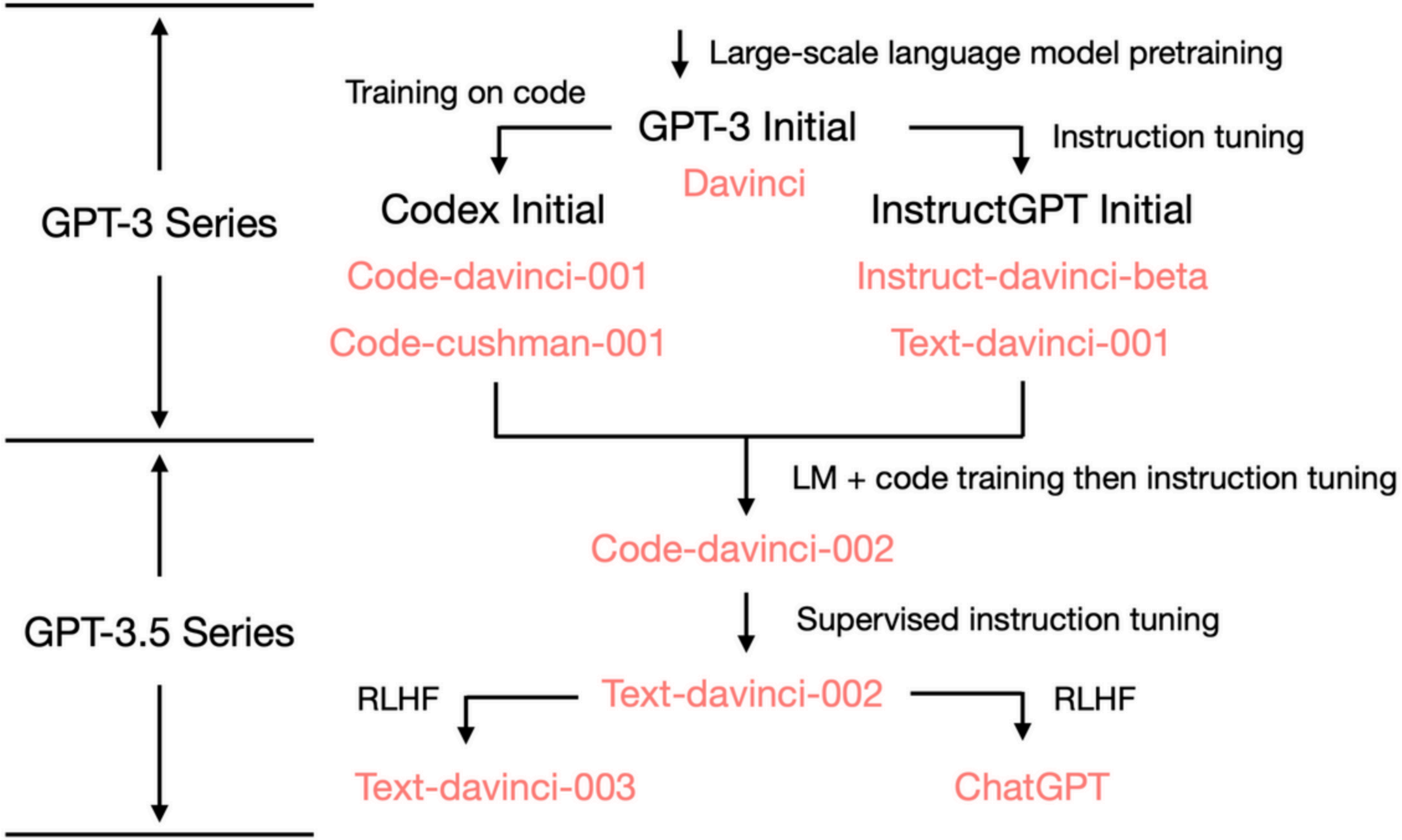
GPT-3

“There’s this idea of emergence that caught me and also, I think, many researchers, by surprise – that you can just train a language model, predict the next token on a tons of raw text, and then it can answer questions, it can summarize documents, have dialogue, translate, classify text, learn all sorts of different kind of pattern manipulation, format dates, and so on. It was just really eye-opening …”

– Percy Liang (Stanford), 2022/06
https://web.stanford.edu/class/cs224u/podcast/liang/

Why do these abilities emerge? I dunno

- Large scale?
- Overparameterization?
- Instruction tuning?
- Training on code?
- RLHF?
- Magic?



GPT-3 Series

GPT-3.5 Series

Large-scale language model pretraining

GPT-3 Initial

Davinci

Training on code

Instruction tuning

Codex Initial

Code-davinci-001

Code-cushman-001

InstructGPT Initial

Instruct-davinci-beta

Text-davinci-001

LM + code training then instruction tuning

Code-davinci-002

Supervised instruction tuning

Text-davinci-002

RLHF

RLHF

Text-davinci-003

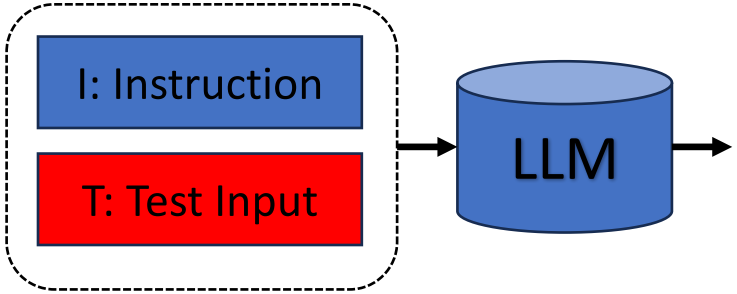
ChatGPT

Recommended read: Blogpost by Yao Fu, Hao Peng, Tushar Khot. May 2023.
How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources

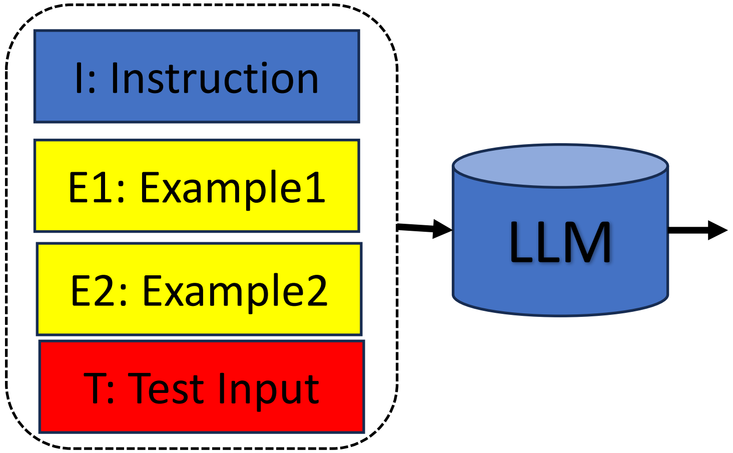
In-Context Learning

I: Instruction	Translate English to French	
E1: Example1	[en]: A discomfort which lasts.	[fr]: Un malaise qui dure
E2: Example2	[en]: HTML is a language for formatting	[fr]: HTML est un langage de formatage
T: Test Input	[en]: After you become comfortable with formatting [fr]:	

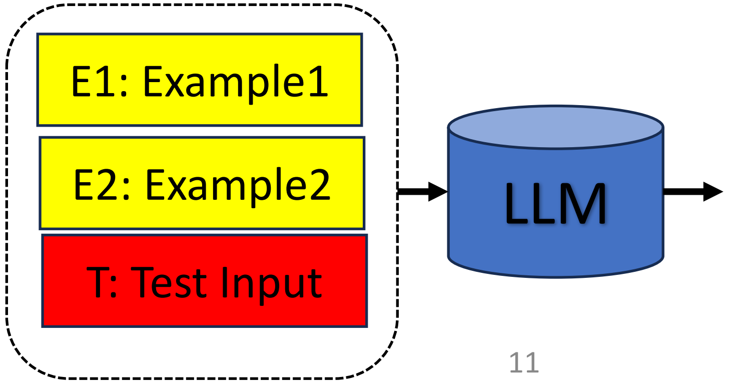
Zero Shot

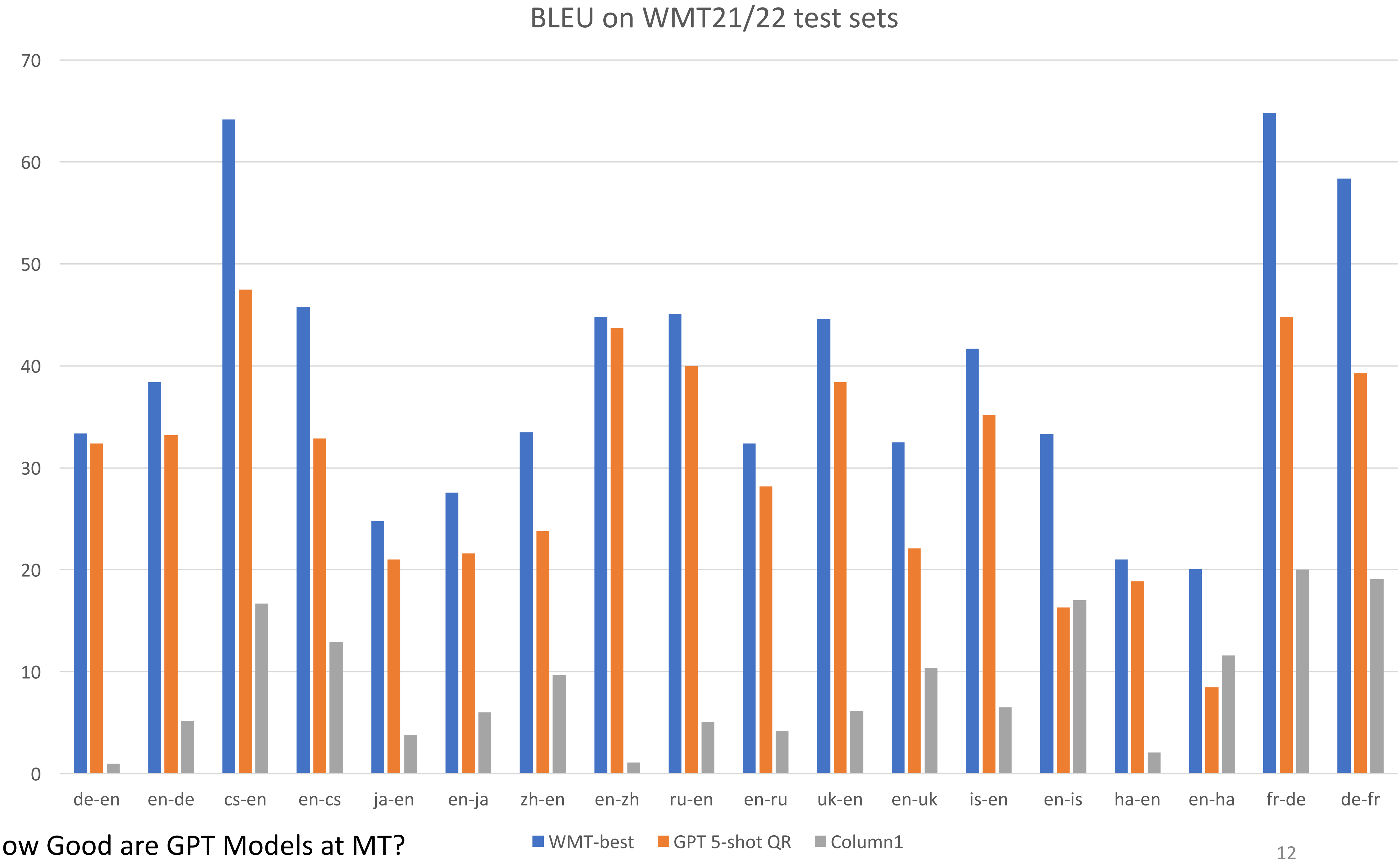


Few Shot (w/ Instruction)



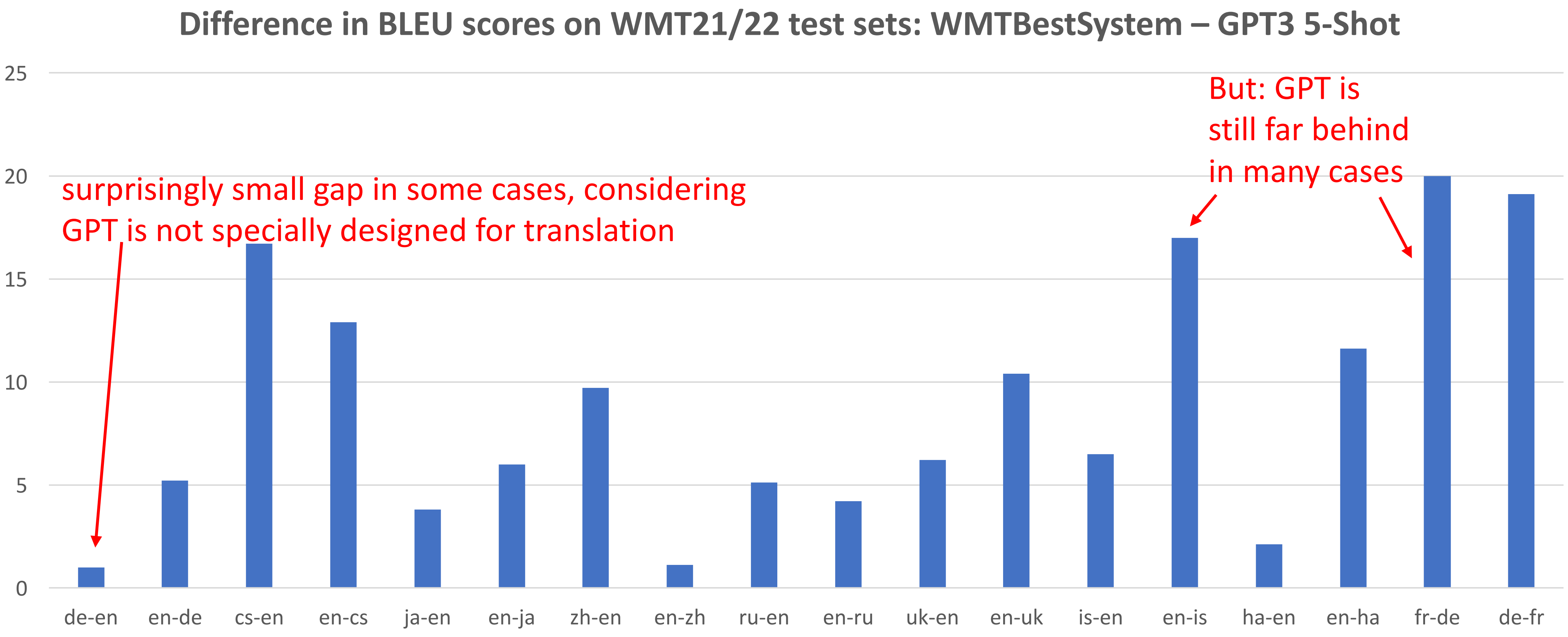
Few Shot (Example only)





Hendy, et. al. How Good are GPT Models at MT?
A Comprehensive Evaluation. https://arxiv.org/pdf/2302.09210.pdf

Evaluation: dedicated MT vs off-the-shelf LLM



surprisingly small gap in some cases, considering GPT is not specially designed for translation

But: GPT is still far behind in many cases

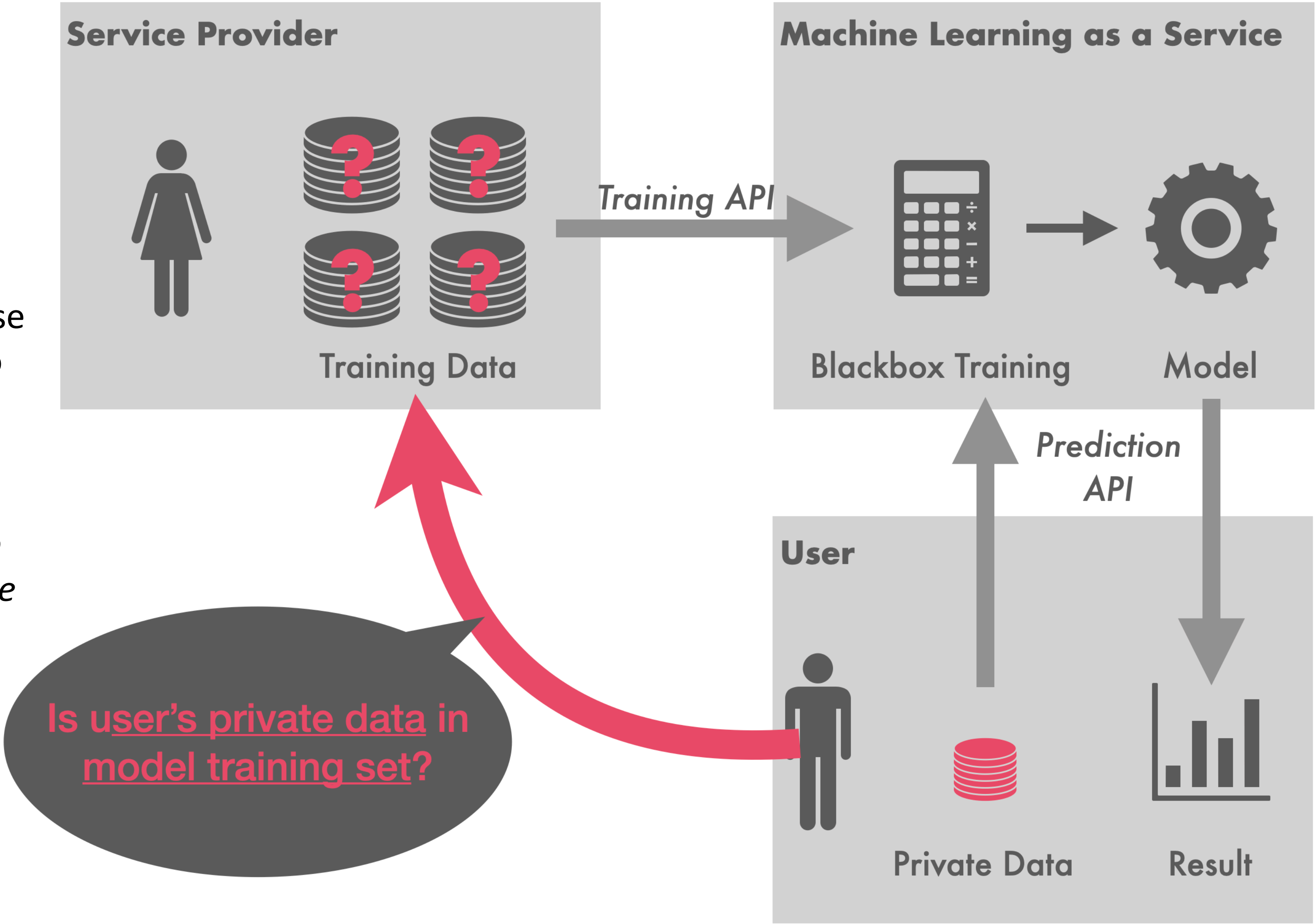
Hendy, et. al. How Good are GPT Models at MT? A Comprehensive Evaluation. https://arxiv.org/pdf/2302.09210.pdf

GPT3: text-davinci-003, with in-context examples from WMT training data with high LaBSE scores

Aside:
Test set leakage

LLM evaluation may be challenging because they are trained on so much data. Need to check for (partial) overlap with test

Potential solution: Membership Inference e.g. See *Hisamoto, Duh, Post. Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your MT System? (TACL 2020)*



Outline

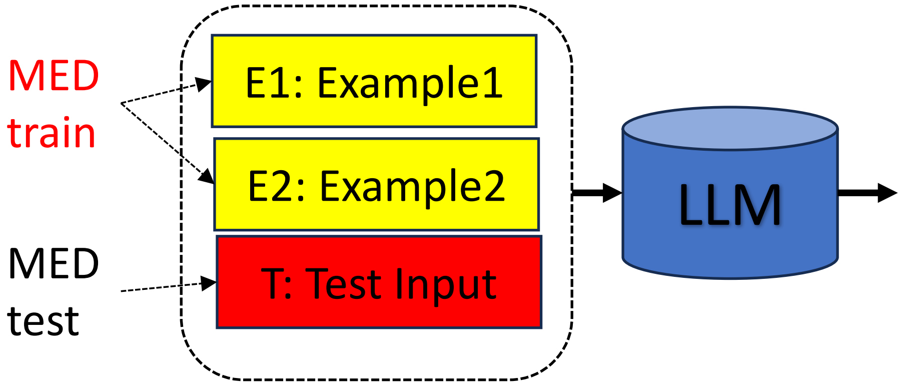
1. Brief background on LLMs
2. Two experiments in on-the-fly adaptation
    - Domain adaptation
    - Document-level adaptation
3. Improving LLMs for MT
    - Soft prompt tuning
4. Discussion

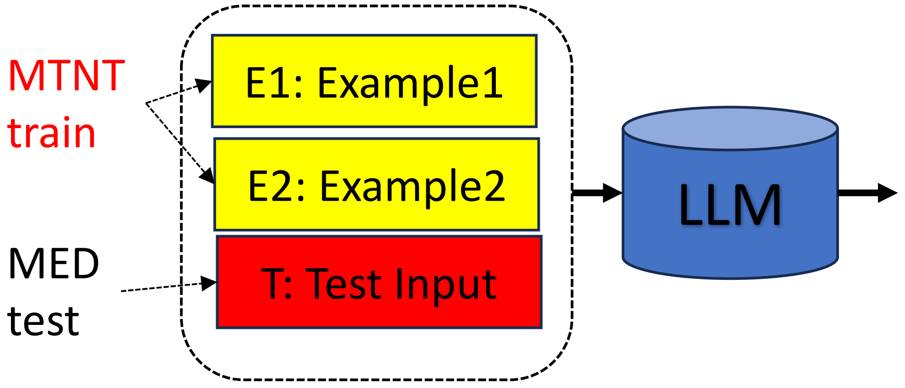
My current thinking

- Will off-the-shelf LLM replace dedicated Machine Translation (MT)?
  - No
  - Even if accuracy gap reduces, development & deployment cost is still big
  - Is academia best situated to tackle these challenges?
- Does LLM enable new applications in translation?
  - Yes
  - We'll explore: On-the-fly adaptation.
  - Goal: no need to pre-specify domains, one model but personalization for all

Domain Adaptation: Experiment Setup

- 4 test domains: FLORES, MED, MTNT, TED
- Vary the examples given to LLM
- Research questions:
  - Do in-domain translation examples improve improve results?
  - Do results vary by the LLM implementation?





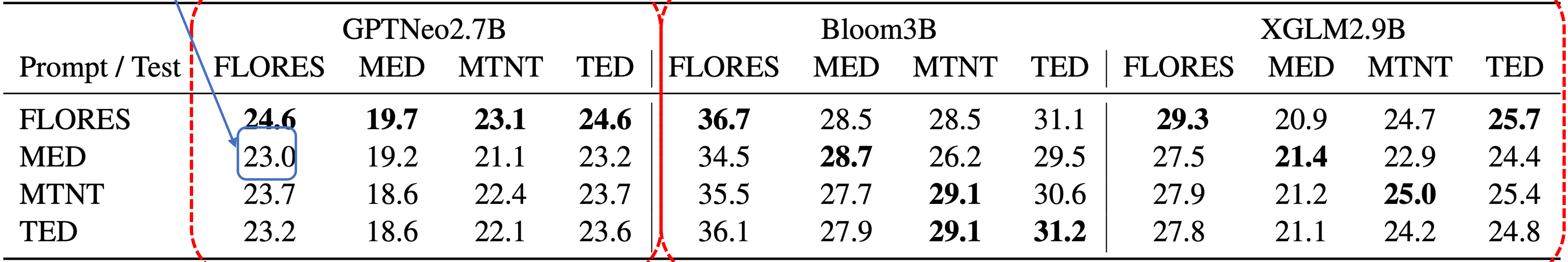
Open-access LLMs used in experiments

	GPTNeo 2.7B	XGLM 2.9B	BLOOM 3B
Citation	Black et al., 2021 (EleutherAI)	Lin et al. 2021 (Meta)	Scao et al. 2022 (BigScience)
Reason	Meant to replicate GPT3 2.7B	Advertised as multilingual	Advertised as multilingual
Architecture	32 layers, 2560 hidden dim	38 layers, 2048 hidden dim	30 layers, 2560 hidden dim
Data & Languages	Mainly English: 97% estimated. 420B tokens	30 languages from CC100XL. English 49%, Russian 9%, Chinese 8%, German 5%, Spanish 5%, French 4%. 500B tokens	45 natural languages, 12 programing languages. English 31%, Chinese 18%, Code 13%, French 13%, Spanish 11%, Portuguese 5%. 350B tokens
Compute (not comparable, just reference)	?	The XGLM 7.5B model was trained on 256 A100 GPUs for about 3 weeks	18 weeks of 52 nodes of 8 80GB A100 GPUs (for all models?)

GPT-3 175B (96 layers, 12k hidden dim) is trained on ~300B tokens.

En-Fr MT BLEU results

e.g. Prompt with 5 MED examples, test on FLORES



	GPTNeo2.7B				Bloom3B				XGLM2.9B			
Prompt / Test	FLORES	MED	MTNT	TED	FLORES	MED	MTNT	TED	FLORES	MED	MTNT	TED
FLORES	**24.6**	**19.7**	**23.1**	**24.6**	**36.7**	28.5	28.5	31.1	**29.3**	20.9	24.7	**25.7**
MED	23.0	19.2	21.1	23.2	34.5	**28.7**	26.2	29.5	27.5	**21.4**	22.9	24.4
MTNT	23.7	18.6	22.4	23.7	35.5	27.7	**29.1**	30.6	27.9	21.2	**25.0**	25.4
TED	23.2	18.6	22.1	23.6	36.1	27.9	**29.1**	**31.2**	27.8	21.1	24.2	24.8

Results make less sense here.

Diagonal tend to be best: in-domain examples improve translation quality (1-2 BLEU)

Document-level Adaptation: Motivation

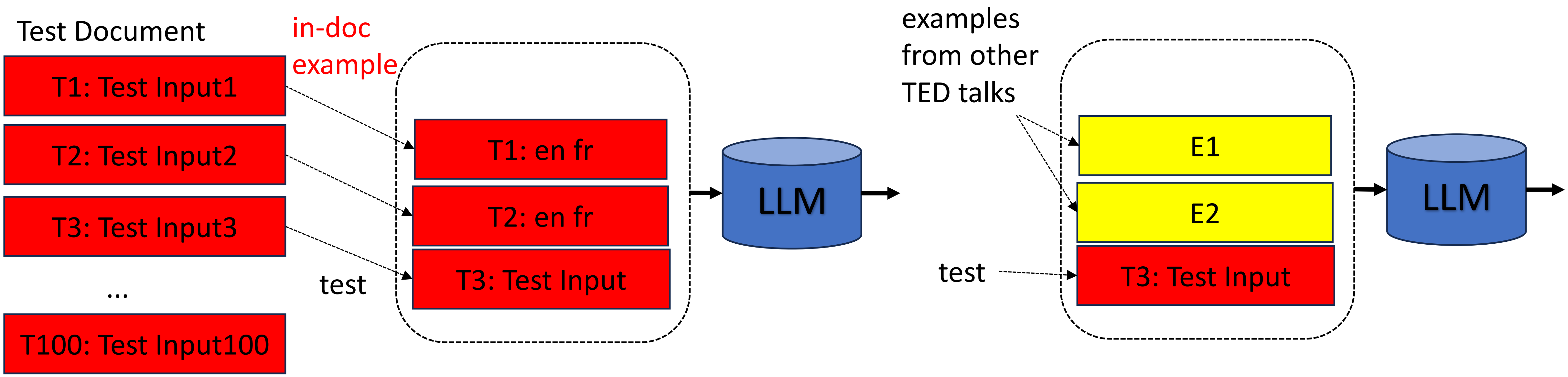
- Domain boundaries are fuzzy and coarse-grained
- Document is a more clear-cut unit
  - Internally coherent, e.g. one sense per discourse
  - Many practical translation tasks are document-based

Each TED talk is translated as a whole

source: www.ted.com

Document-level Adaptation: Experiment Setup

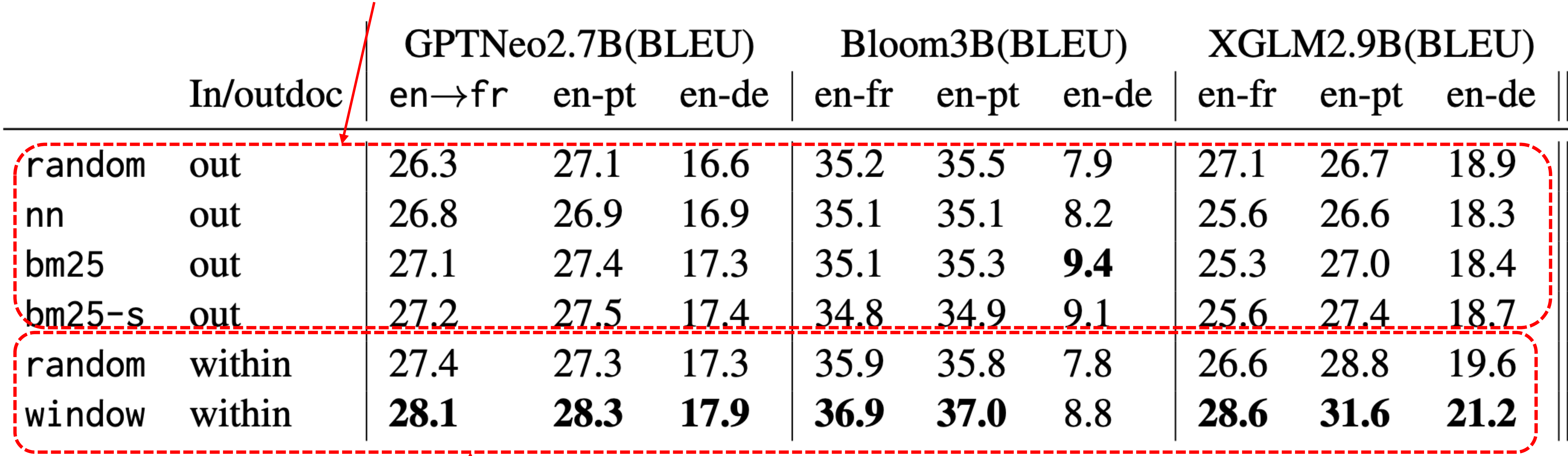
- Re-split TED talks data:
  - 120 test documents, 100-120 sentence each → in-doc examples
  - The rest are used “out-doc” prompt examples
- Research question: are in-doc examples better out-doc examples?



Document Adaptation Results (3 language pairs)

For out-doc examples, little difference among:

- random sampling
- sentence retrieval by nearest neighbor embedding (nn) or BM25
- submodular optimization (with BM25 like objective)



		GPTNeo2.7B(BLEU)			Bloom3B(BLEU)			XGLM2.9B(BLEU)		
	In/outdoc	en→fr	en-pt	en-de	en-fr	en-pt	en-de	en-fr	en-pt	en-de
random	out	26.3	27.1	16.6	35.2	35.5	7.9	27.1	26.7	18.9
nn	out	26.8	26.9	16.9	35.1	35.1	8.2	25.6	26.6	18.3
bm25	out	27.1	27.4	17.3	35.1	35.3	**9.4**	25.3	27.0	18.4
bm25-s	out	27.2	27.5	17.4	34.8	34.9	9.1	25.6	27.4	18.7
random	within	27.4	27.3	17.3	35.9	35.8	7.8	26.6	28.8	19.6
window	within	**28.1**	**28.3**	**17.9**	**36.9**	**37.0**	8.8	**28.6**	**31.6**	**21.2**

Random In-doc examples and especially Sliding Window of previous in-doc examples perform well

Preliminary Conclusions

- Domain Adaptation
  - On-the-fly adaptation to domain is possible, but results are mixed
- Document-level Adaptation
  - Good fit for LLMs
  - In-doc examples are better than out-doc examples
  - Next step: use predicted translations (not oracle) in prompts

Outline

1. Brief background on LLMs
2. Two experiments in on-the-fly adaptation
    - Domain adaptation
    - Document-level adaptation
3. Improving LLMs for MT
    - Soft prompt tuning
4. Discussion

Motivation: Are there general techniques to improve the translation capability of LLMs?

- During LLM construction:
  - Better selection of multilingual corpora
  - Fine-tuning during the Instruction Tuning / RLHF phase
- After LLM construction (cannot rely on too much data):
  - Adaptors
  - Prompt Engineering
  - Soft Prompt Tuning or Prefix Tuning
  - Etc

Soft Prompt Tuning

Train embeddings for this token

1. We can do train this just on monolingual French:

[fr] f1 f2 f3

[fr] f4 f5

by fixing all LLM parameter except [fr] token

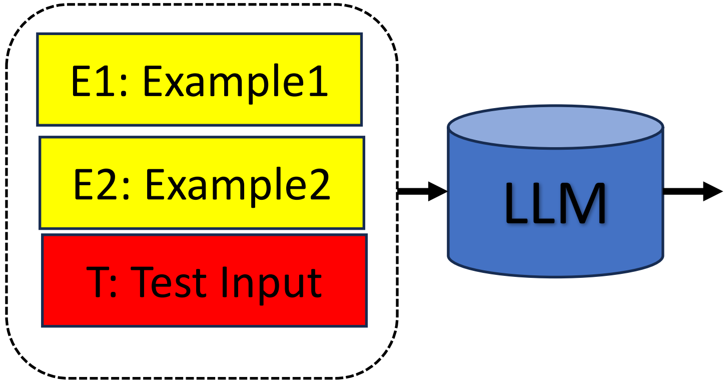
2. We can also then fine-tune on a bit of bitext

[en] e1 e2 e3 [fr] f1 f2 f3

[en] e4 e5 [fr] f4 f5

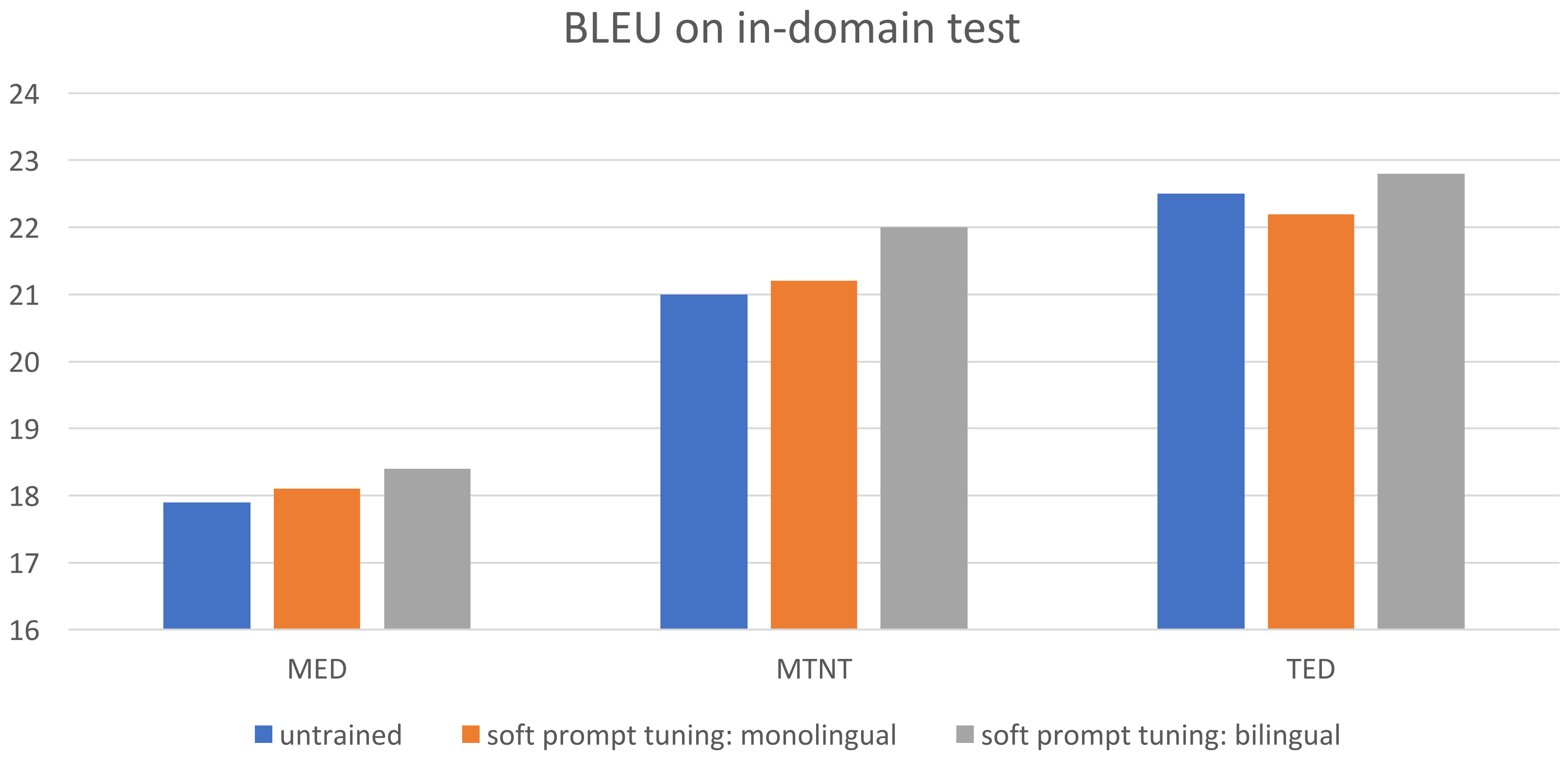
E1: Example1	[en]: A discomfort which lasts.	[fr]: Un malaise qui dure
E2: Example2	[en]: HTML is a language for formatting	[fr]: HTML est un langage de formatage
T: Test Input	[en]: After you become comfortable with formatting [fr]:	

Few Shot Setup



Soft prompt tuning results

- Setup: GPTNeo-2.7B, similar to domain adaptation experiments previously but use 20 in-domain examples as prompts
- We observe slight improvements (esp bilingual soft prompt tuning)



Outline

1. Brief background on LLMs
2. Two experiments in on-the-fly adaptation
   - Domain adaptation
   - Document-level adaptation
3. Improving LLMs for MT
   - Soft prompt tuning
4. Discussion

For more information, refer to:

- Sia & Duh (2023). In-context Learning as Maintaining Coherency: A Study of On-the-fly MT Using LLMs
- Sia & Duh (2022). Prefix Embeddings for In-Context MT

Summary

- My current bet: LLMs won’t replace dedicated MT but they will enable new niche applications
- Example: On-the-fly adaptation, esp. at document-level
- From academia, we can’t easily build or work with the largest LLM, but there are some knobs we can tune (e.g. soft prompt tuning)
- The challenge is to figure out which niche applications + knobs

Discussion questions

- Do you think LLMs will become a true “foundation model”, part of every NLP system?
- Are there new applications of in-context learning for translation?
- What should we work on, if multilingual LLM is our ultimate goal?
- How much effort would you put into LLM vs non-LLM research?