Google DeepMind

Low-Rank Adaptation for Multilingual Summarization: An Empirical Study

Chenxi Whitehouse
University of Cambridge

NAACL 2024

Chenxi Whitehouse

University of Cambridge

Fantine Huot

Google DeepMind

Jasmijn Bastings

Google DeepMind

Mostafa Dehghani

Google DeepMind

Chu-Cheng Lin

Google Cloud

Mirella Lapata

Google DeepMind

Today's Agenda

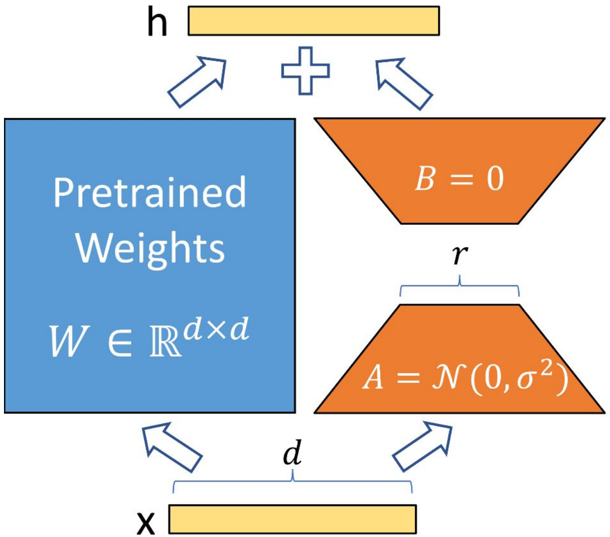
- Research Questions and Motivation
- Low-Rank Adaptation (LoRA)
- LoRA vs Full Fine-tuning for Multilingual Generation
  - High/Full Data
  - Low Data
  - Cross-lingual Transfer
- Conclusions

Background & Motivation

- Large Language Models (LLMs) are becoming increasingly powerful, but their growing size also makes training less practical
- **Parameter-efficient fine-tuning** (PEFT) approaches are desirable, especially for tasks requiring extensive memory, e.g., with long input
- Existing PEFT approaches include
    - Adapters
    - Prefix tuning
    - **LoRA** (Low-Rank Adaptation)
- We focus on a challenging task with long input: **Multilingual summarization**, where LoRA is under-explored

What is LoRA?

[Hu et al. 2021]


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

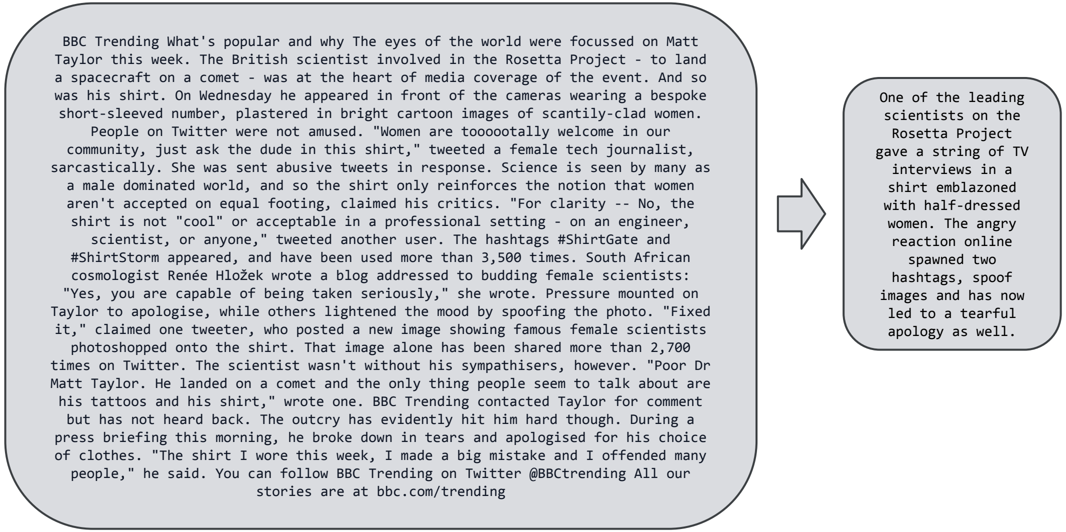
LoRA: Low-Rank Adaptation [of LLMs]

- **Freezes** the pre-trained model weights (***W***) and **adds** trainable low-rank matrices (***A*** & ***B***) into the Transformer architecture
- **No extra cost or latency** at inference time
  - Can merge LoRA with the frozen parameters
- **Up to 10,000 times fewer trainable parameters**
  - Up to 3 times less GPU memory (GPT-3)
- **Competitive performance** vs full fine-tuning (on classification or monolingual generation tasks)

Why do we study LoRA for Multilingual Summarization?

Multilingual Summarization is Complex

- Models are expected to fluently generate in **many languages**
- High/low resource: not all languages have (sufficient) data
- Long input and output (e.g. XLSum):



BBC Trending What's popular and why The eyes of the world were focussed on Matt Taylor this week. The British scientist involved in the Rosetta Project - to land a spacecraft on a comet - was at the heart of media coverage of the event. And so was his shirt. On Wednesday he appeared in front of the cameras wearing a bespoke short-sleeved number, plastered in bright cartoon images of scantily-clad women. People on Twitter were not amused. "Women are toooootally welcome in our community, just ask the dude in this shirt," tweeted a female tech journalist, sarcastically. She was sent abusive tweets in response. Science is seen by many as a male dominated world, and so the shirt only reinforces the notion that women aren't accepted on equal footing, claimed his critics. "For clarity -- No, the shirt is not "cool" or acceptable in a professional setting - on an engineer, scientist, or anyone," tweeted another user. The hashtags #ShirtGate and #ShirtStorm appeared, and have been used more than 3,500 times. South African cosmologist Renée Hložek wrote a blog addressed to budding female scientists: "Yes, you are capable of being taken seriously," she wrote. Pressure mounted on Taylor to apologise, while others lightened the mood by spoofing the photo. "Fixed it," claimed one tweeter, who posted a new image showing famous female scientists photoshopped onto the shirt. That image alone has been shared more than 2,700 times on Twitter. The scientist wasn't without his sympathisers, however. "Poor Dr Matt Taylor. He landed on a comet and the only thing people seem to talk about are his tattoos and his shirt," wrote one. BBC Trending contacted Taylor for comment but has not heard back. The outcry has evidently hit him hard though. During a press briefing this morning, he broke down in tears and apologised for his choice of clothes. "The shirt I wore this week, I made a big mistake and I offended many people," he said. You can follow BBC Trending on Twitter @BBCtrending All our stories are at bbc.com/trending

One of the leading scientists on the Rosetta Project gave a string of TV interviews in a shirt emblazoned with half-dressed women. The angry reaction online spawned two hashtags, spoof images and has now led to a tearful apology as well.

Research Questions

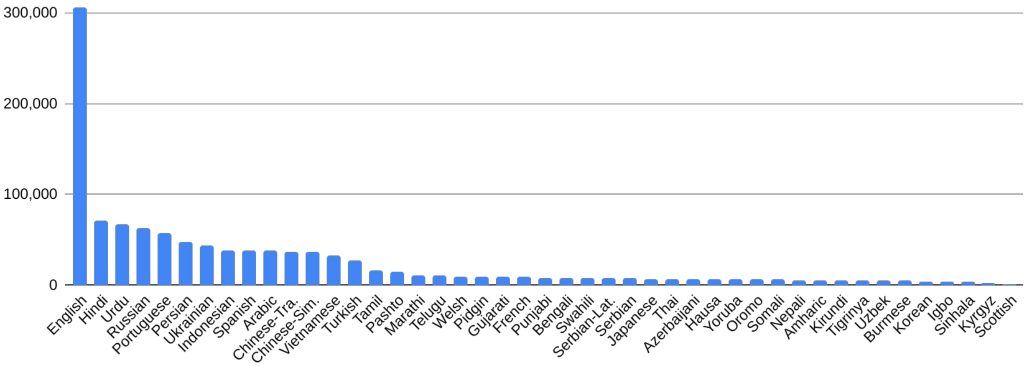
- **How does LoRA perform vs full fine-tuning (FT) under different scenarios for multilingual summarization?**
- We consider the following scenarios:
  - **Scenario 1: High-data**:
    Languages with sufficient training data
    (Automatic data collection, crowdsourcing)
  - **Scenario 2: Low-data:**
    A dozen or a few hundred examples available
    Low-resource language, annotated data
  - **Cross-lingual transfer on unseen languages:**
    Scenario 3: Only English Data Available
    Scenario 4: Multiple Languages Data Available
    Scenario 5: Some Examples in Target Language Available
- Does the model setup, including LoRA rank and model size, impact the performance?

Models and Datasets

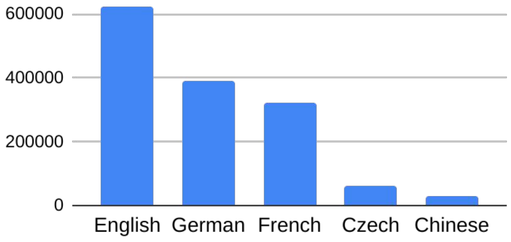
- 2 PaLM-2 models (**XXS**, S)
- 2 multilingual summarization datasets: XLSum, XWikis
- Metrics: Summary **Relevance** (Rouge-L), **Faithfulness** (NLI), and **Conciseness** (Seahorse [Clark et al, 2023])

	Task	Source	#Train/Val/ Test Data	#Languages
XLSum	Summarisation	BBC News	1.12M / 114K / 114K	45
XWikis	Summarisation (multi-sent.)	Wikipedia	1.43M / 40K / 35K	5

Number of Training Examples per Language in XLSum



Number of Training Examples in XWikis



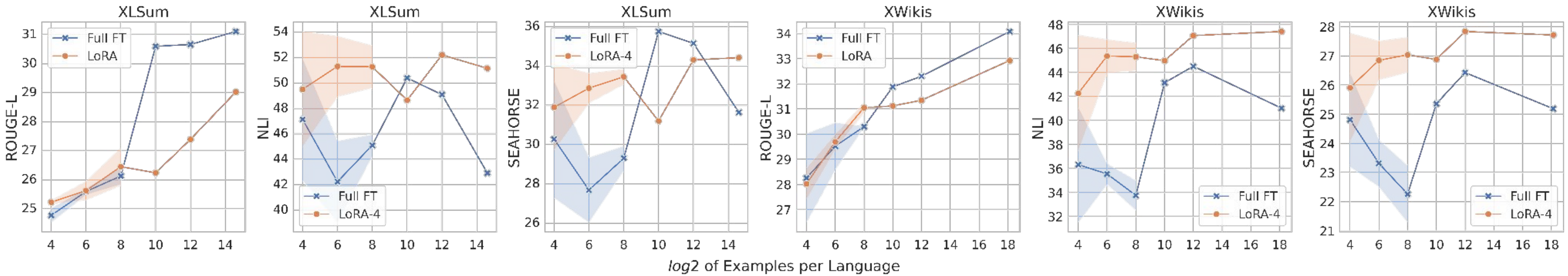
Scenario 1: High-Data Regime

- Train on all the data available for each language
- Full FT outperforms LoRA on summary relevance (R-L). LoRA with higher ranks enhances Rouge-L scores
- LoRA is superior on summary faithfulness (NLI) & conciseness (SH), lower ranks see better scores

	XLSum			**XWikis**		
	R-L	NLI	SH	R-L	NLI	SH
Full-FT	**31.11**	42.93	31.64	**34.08**	41.04	25.19
LoRA-64	29.79	45.51	31.80	34.04	45.34	27.02
LoRA-16	29.77	48.48	33.25	33.80	46.10	27.42
LoRA-4	29.03	**51.16**	**34.42**	32.92	**47.43**	**27.72**

Scenario 2: Low-Data Regime

- We randomly select 16, 64, 256, 1024, and 4096 examples per language and train together (balanced data)
- LoRA achieves overall better faithfulness (NLI) and conciseness (Seahorse) than Full FT
- For ROUGE-L, Full FT outperforms LoRA when provided > 1K training examples
- Low-data training on LoRA is more stable (Full FT more sensitive for the selection of checkpoints)



Cross-lingual Transfer

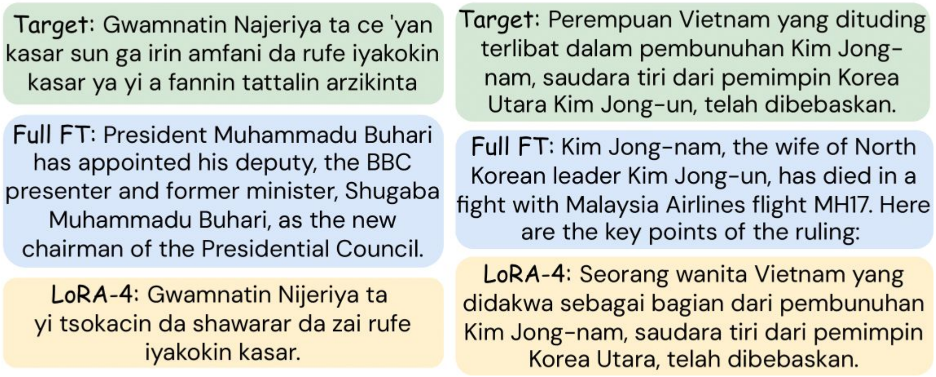
Scenario 3: Only English Data Available

- Train and validate on English data only
- Full FT does not transfer to unseen languages: output in English rather than target languages

	XLSum			XWikis		
	R-L	NLI	SH	R-L	NLI	SH
Full-FT	5.20	4.49	6.88	17.51	35.95	22.43
LoRA-4	**21.13**	**39.07**	**23.08**	**23.86**	**45.54**	**25.96**

Hausa

Indonesian



Target: Gwamnatin Najeriya ta ce 'yan kasar sun ga irin amfani da rufe iyakokin kasar ya yi a fannin tattalin arzikinta

Full FT: President Muhammadu Buhari has appointed his deputy, the BBC presenter and former minister, Shugaba Muhammadu Buhari, as the new chairman of the Presidential Council.

LoRA-4: Gwamnatin Nijeriya ta yi tsokacin da shawarar da zai rufe iyakokin kasar.

Target: Perempuan Vietnam yang dituding terlibat dalam pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara Kim Jong-un, telah dibebaskan.

Full FT: Kim Jong-nam, the wife of North Korean leader Kim Jong-un, has died in a fight with Malaysia Airlines flight MH17. Here are the key points of the ruling:

LoRA-4: Seorang wanita Vietnam yang didakwa sebagai bagian dari pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara, telah dibebaskan.

Scenario 4: Multiple Available Languages

More available training languages

- XLSum: Train on 10 source languages (high resource), test on 10 target languages (10 different language families)
- XWikis: **leave-one-out** cross validation

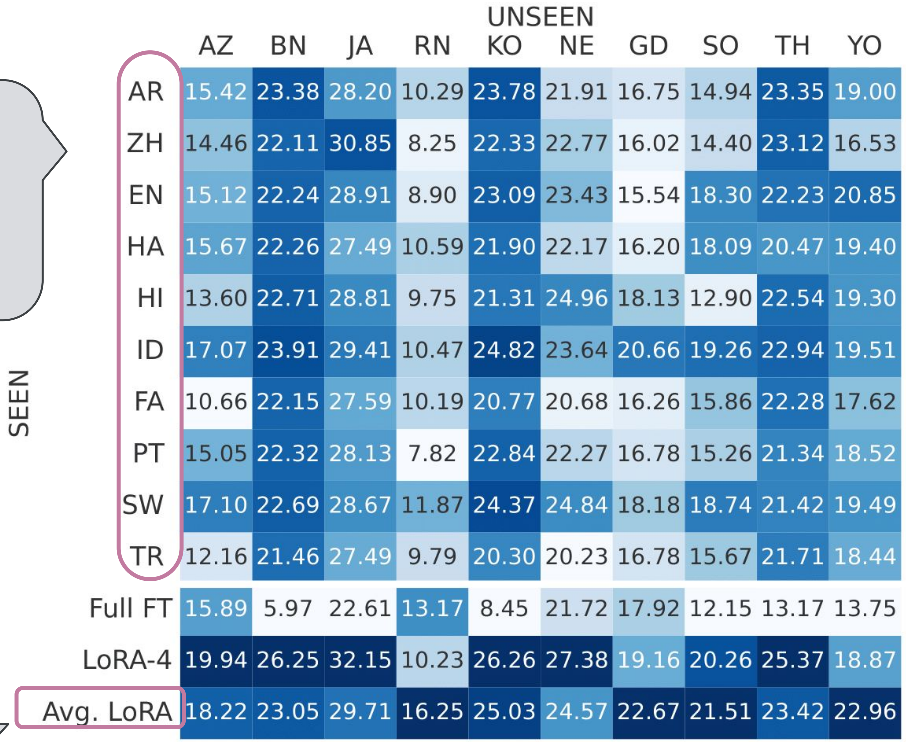
Training options

- Full FT or LoRA on available languages
- We further explore LoRA weight composition
  - LoRA on individual languages + **composition** (weighted average) of LoRA modules for target languages
  - **Benefit**: flexibly adding more available modules

Scenario 4: Multiple Available Languages

Language-specific LoRA of the 10 training languages

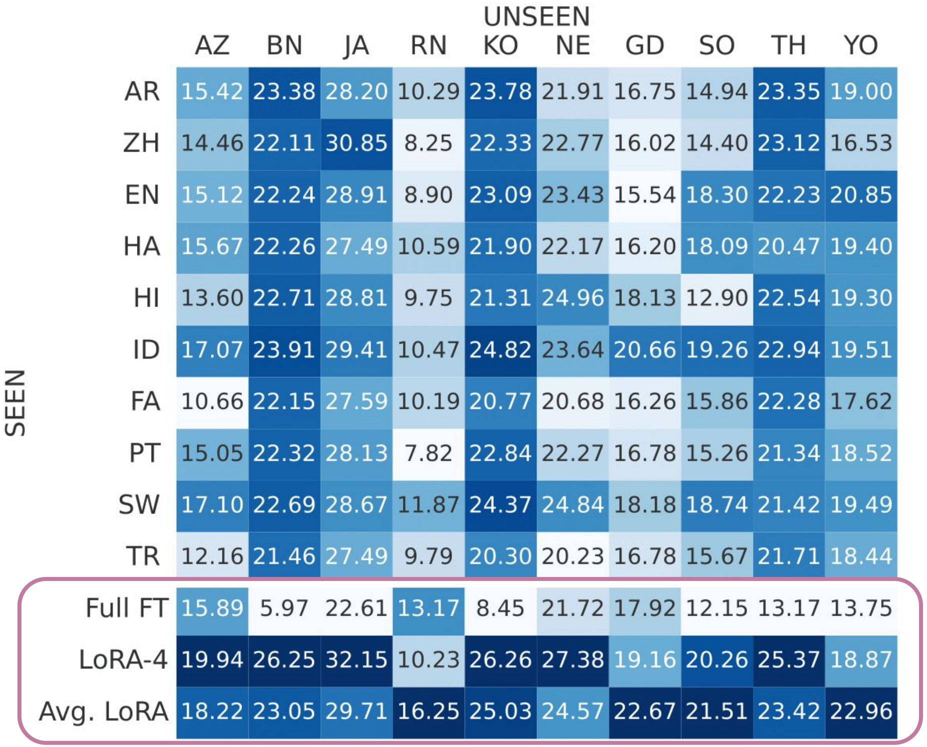
Weighted average of the 10 LoRA modules above



		UNSEEN									
		AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
SEEN	AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
	ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
	EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
	HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
	HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
	ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
	FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
	PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
	SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
	TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
	Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
	LoRA-4	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
	Avg. LoRA	18.22	23.05	29.71	16.25	25.03	24.57	22.67	21.51	23.42	22.96

ROUGE-L scores for 10 test languages on XLSum

Scenario 4: Multiple Available Languages

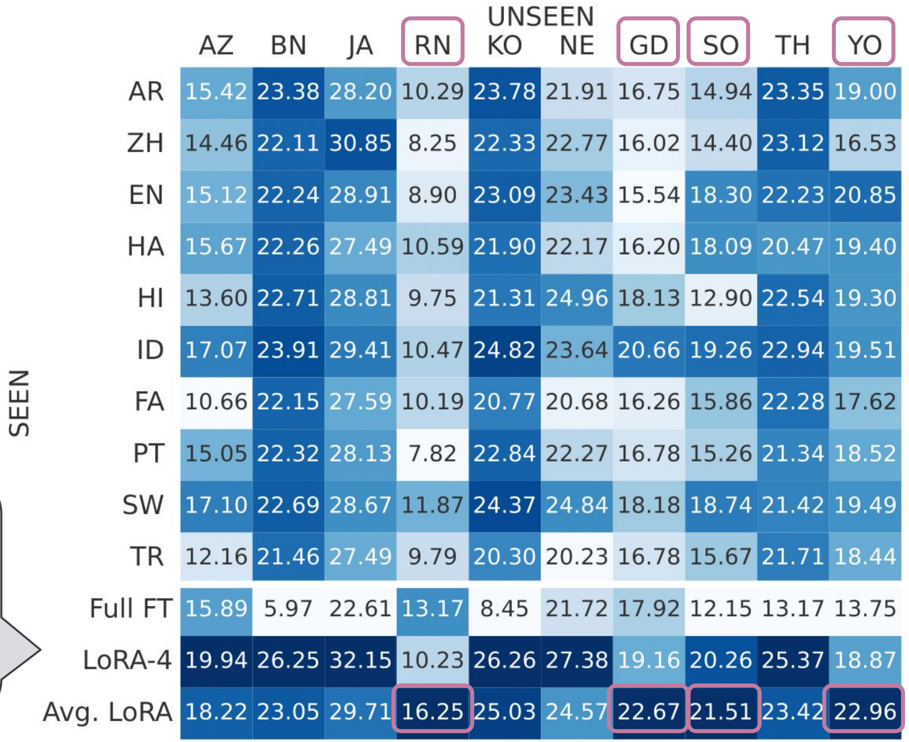


		UNSEEN									
		AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
SEEN	AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
	ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
	EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
	HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
	HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
	ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
	FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
	PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
	SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
	TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
	Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
	LoRA-4	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
	Avg. LoRA	18.22	23.05	29.71	16.25	25.03	24.57	22.67	21.51	23.42	22.96

ROUGE-L scores for 10 test languages on XLSum

With 10 languages trained together, Full Fine-tuning still lags behind in cross-lingual transferability

Scenario 4: Multiple Available Languages



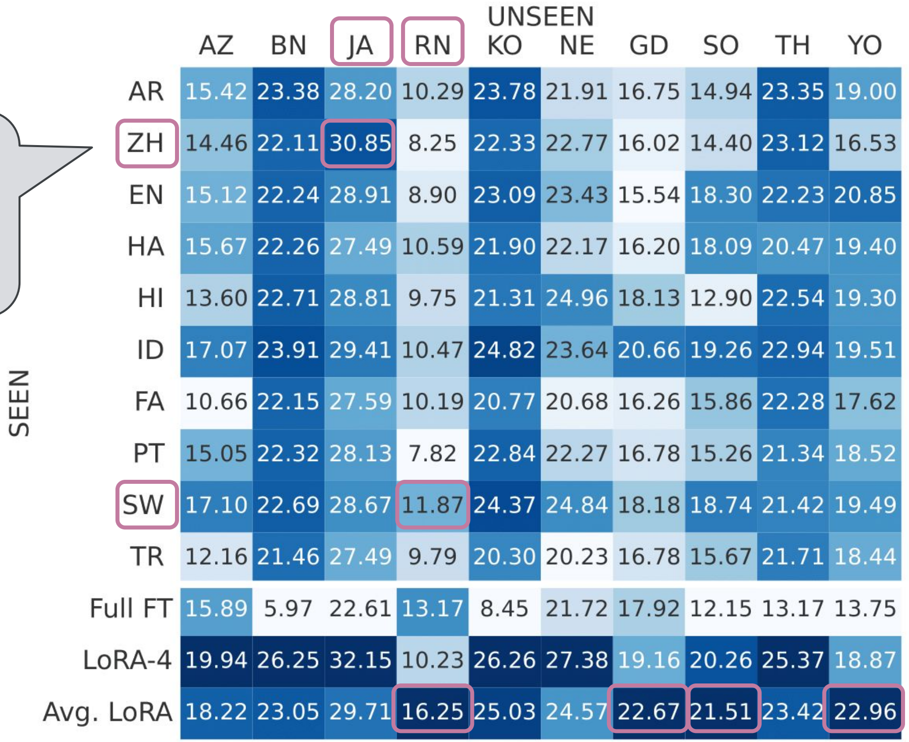
SEEN \ UNSEEN	AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
LoRA-4	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
Avg. LoRA	18.22	23.05	29.71	16.25	25.03	24.57	22.67	21.51	23.42	22.96

ROUGE-L scores for 10 test languages on XLSum

Lower resource languages (Kirundi - RN, Scottish - GD, Somali - SO, Yoruba - YO) work best with individual LoRA training

Scenario 4: Multiple Available Languages

Similar languages may transfer better



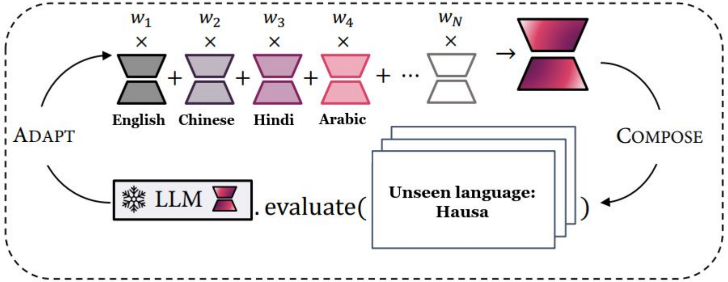
		UNSEEN									
		AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
SEEN	AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
	ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
	EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
	HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
	HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
	ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
	FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
	PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
	SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
	TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
	Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
	LoRA-4	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
	Avg. LoRA	18.22	23.05	29.71	16.25	25.03	24.57	22.67	21.51	23.42	22.96

ROUGE-L scores for 10 test languages on XLSum

Scenario 5: Some Available Examples in Target Language

How can we make use of the target examples?

- Continue full/LoRA fine-tuning on available examples
- In-context-learning:
  - infeasible for multilingual summarization: input too long
- Composing LoRA modules with LoraHub [Huang et al. 2023]
  - **gradient-free** optimization to find the optimal **weighted sum** of the available LoRA modules, based on the score on samples from the target task



LoRAHub Learning for Unseen Languages

Scenario 5: Some Available Examples in Target Language

- Assume a handful target examples available (16), compare LoRA continued learning (CL) and LoraHub
- LoRA continued learning superior performance
- A few examples significantly improves Full FT compared to the zero-shot results

Zero-shot				16-shot			
	R-L	NLI	SH		R-L	NLI	SH
Full FT	14.48	28.87	13.71	Full FT	22.31	30.15	18.79
LoRA	22.59	37.39	24.21	LoRA (CL)	**24.71**	**41.12**	**26.47**
Avg.LoRA	**22.74**	**49.14**	**32.44**	LoraHub	23.37	38.95	26.07

Zero-and 16-shot scores for average of 10 test languages on XLSum

Conclusions

LoRA vs Full FT for Multilingual Summarization:

- LoRA achieves **superior performance** vs Full FT:
  - Zero-shot and few-shot cross-lingual transfer
  - Low-data regime (< 1K examples)
  - Summary faithfulness and conciseness
  - In addition, LoRA continued learning outperforms LoraHub under few-shot settings
- LoRA achieves **on-par performance** vs Full FT in larger models (see paper)
- LoRA achieves **worse performance** vs Full FT:
  - Smaller models
  - High-data regime, particularly for summary relevance

Thank you.

Google DeepMind