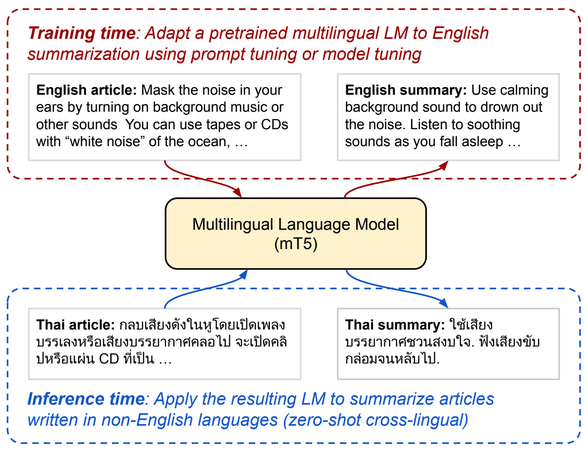
Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation

Tu Vu[1,2], Aditya Barua[1], Brian Lester[1], Daniel Cer[1], Mohit Iyyer[2], Noah Constant[1]

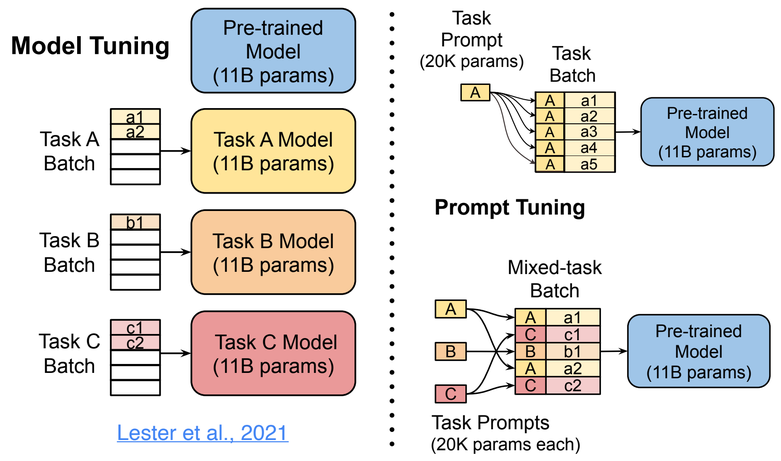
Google Research[1] UMass Amherst[2]

Zero-shot cross-lingual generation (XGen)



A model is required to learn a generative task from labeled data in one language (i.e., English), and then perform the equivalent task in another language at inference time.

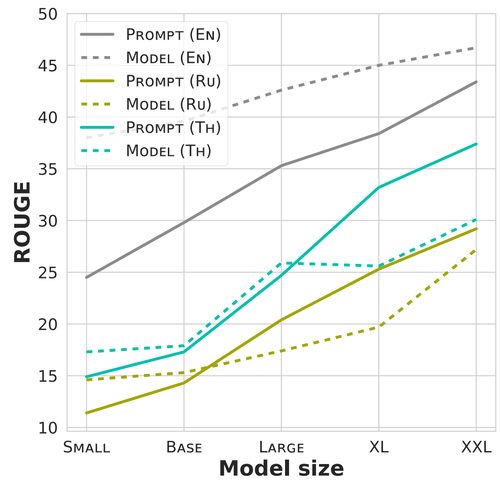
Model Tuning vs. Prompt Tuning



Lester et al., 2021

Model Tuning: *fine-tunes the entire model on each task*
Prompt Tuning: *learns only a small amount of additional parameters while keeping the entire model frozen*

Prompt Tuning is better than Model Tuning on larger language shifts!

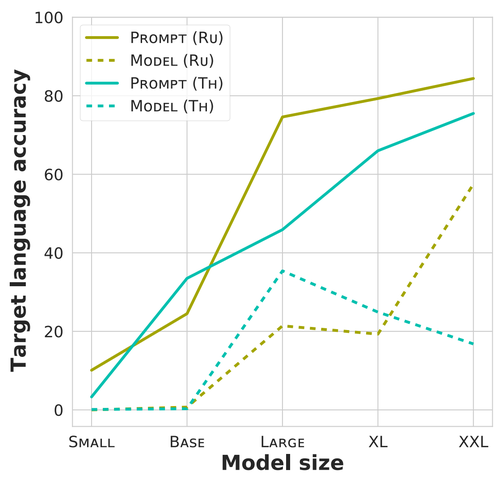


Zero-shot inference on other languages than English is challenging for both methods. Interestingly, Prompt Tuning can provide large gains over Model Tuning.

Model Tuning suffers more from catastrophic forgetting

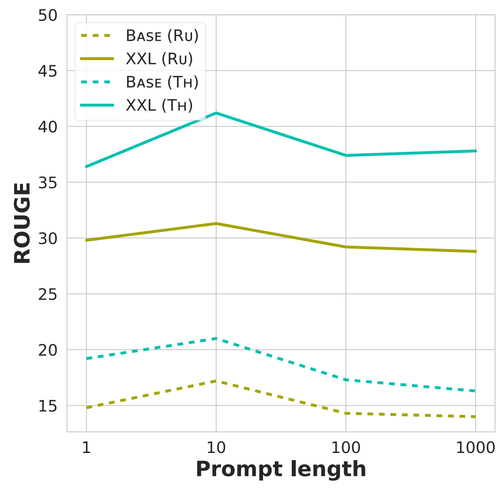
# train steps	5K	10K	30K	50K
Prompt	กำหนดเป้าหมายของคุณ. แบ่งเป้าหมายของคุณออกเป็นเป้าหมายย่อยๆ. ตั้งเป้าหมายในทุกๆ วัน. ตั้งเป้าหมายในระยะยาว.	กำหนดเป้าหมายของคุณ. แบ่งเป้าหมายของคุณเป็นเป้าหมายย่อยๆ. ติดตามความก้าวหน้าของคุณ. ถามตัวเองว่า "ฉันสามารถทำอะไรในวันนี้เพื่อบรรลุเป้าหมายของฉัน? ฉันสามารถทำอะไรในอีก 3 สัปดาห์ข้างหน้าเพื่อบรรลุเป	Define your goal. Divide your goal into smaller goals. Keep track of your progress. Ask yourself what you can do today. Ask yourself what you can do in 3 weeks. Ask yourself what you can do in a year.	Establish a time frame for your goals. Divide your goals into smaller steps. Ask yourself what you can do today. Ask yourself what you can do in the next 3 weeks. Ask yourself what you can do in the n
Model	Set a time limit for your goal	Set a time limit for your goal	Be specific about your goals. Set benchmarks for yourself. Keep track of your progress. Ask yourself questions.	Be specific with your goals. Set benchmarks and routines to help you achieve your goals. Keep track of your progress. Ask yourself questions to help you stay on track

Bigger models are less prone to forget



Moving to larger model sizes mitigates catastrophic forgetting to a large extent.

Too much prompt capacity is harmful



Paradox of capacity: *On the one hand, greater capacity helps to better learn the summarization task. On the other hand, the greater the capacity to learn from English data, the more the model forgets other languages.*

Mitigating catastrophic forgetting

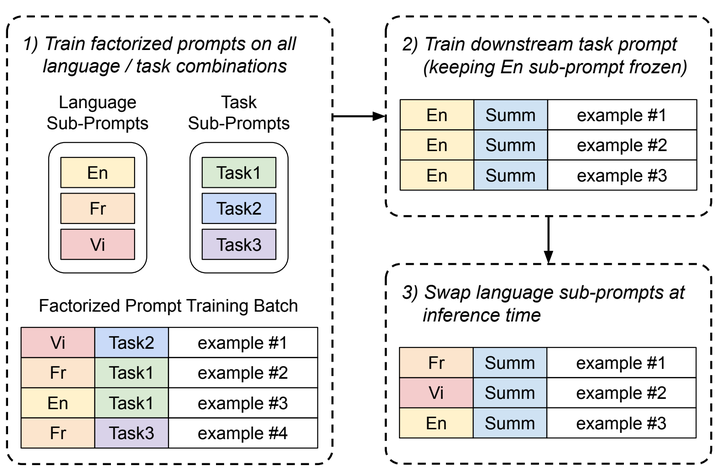
Mixing in unlabeled multilingual training data (MIX-UNSUP)

- **1%** unsupervised training task (i.e., span corruption) either from the target language
- **99%** WikiLingua-0

Factorized prompts (FP)

- explicitly factoring soft prompts into "task" and "language" components that can be recombined in novel pairings during inference

Factorized prompts



We learn recomposable language and task sub-prompts by training on all language / task combinations from a set of unsupervised tasks covering all languages.

Qualitative Analysis

Prompt Tuning	Model Tuning
वयस्क व्यक्ति का दांत निकलवाने के लिए डेंटिस्ट के पास जाएँ. वयस्क व्यक्ति का दांत खुद न निकालें.	Go to a डेंटिस्ट. Do not try to loose the दांत on your own.
Giảm độ ẩm trong nhà. Pha loãng giấm với nước. Xịt hỗn hợp lên thảm. Rắc muối nở lên mặt thảm. Làm khô thảm. Nhờ chuyên gia xử lý.	Lower the humidity. Mix giấm với nước. Apply giấm mixture lên thảm. Sprinkle muối nở lên thảm. Allow thảm to dry. Use quạt to làm khô thảm. Consider xử lý thảm bị hư hại

Sample Hindi (top) and Vietnamese (bottom) predictions of our XXL model tuned with Prompt and Model Tuning. While the summaries are all understandable to a bilingual speaker, Prompt Tuning tends to stay within the target language, whereas Model Tuning is more prone to code switching between English (red) and the target language.

Mixing in unsupervised multilingual data prevents catastrophic forgetting

Size	Method	Thai	
		ROUGE	Lang. Acc.
Base	Prompt	17.3	33.5
	Prompt + MIX-UNSUP	20.9	**76.9**
	Model	17.9	0.3
	Model + MIX-UNSUP	**25.2**	56.8
XXL	Prompt	**37.4**	**75.5**
	Prompt + MIX-UNSUP	**37.4**	74.0
	Model	30.1	16.8
	Model + MIX-UNSUP	32.4	32.4

MIX-UNSUP improves XGen capacities for Model Tuning. For Prompt Tuning, it provides a benefit where catastrophic forgetting is more severe.

Factorized prompts are helpful for overcoming severe catastrophic forgetting

Size	Method	Thai	
		ROUGE	Lang. Acc.
Base	Prompt	17.3	33.5
	Prompt + MIX-UNSUP	20.9	**76.9**
	Prompt + FP	17.9	0.3
XXL	Prompt	**37.4**	**75.5**
	Prompt + MIX-UNSUP	**37.4**	74.0
	Prompt + FP	36.9	**80.8**

FP are successful at improving target language accuracy in all conditions. However, this does not always translate to higher ROUGE. In settings where Prompt Tuning shows the most severe forgetting (e.g., at BASE size), FP provide large gains.

Key take-aways

- ✦ a challenging benchmark for zero-shot cross-lingual generation (XGen)
- ✦ increasing model scale and decreasing tunable parameter capacity are key for overcoming catastrophic forgetting
- ✦ methods for further mitigating catastrophic forgetting, including mixing in unlabeled multilingual data and factorized prompts
- ✦ LM-Adapted mT5 checkpoints

References

✦ Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In EMNLP 2021, pages 3045–3059.