Fine-tuned Language Models are Continual Learners

Tuhin Chakrabarty*

tuhin.chakr@cs.columbia.edu

Thomas Scialom *

tscialom@fb.com

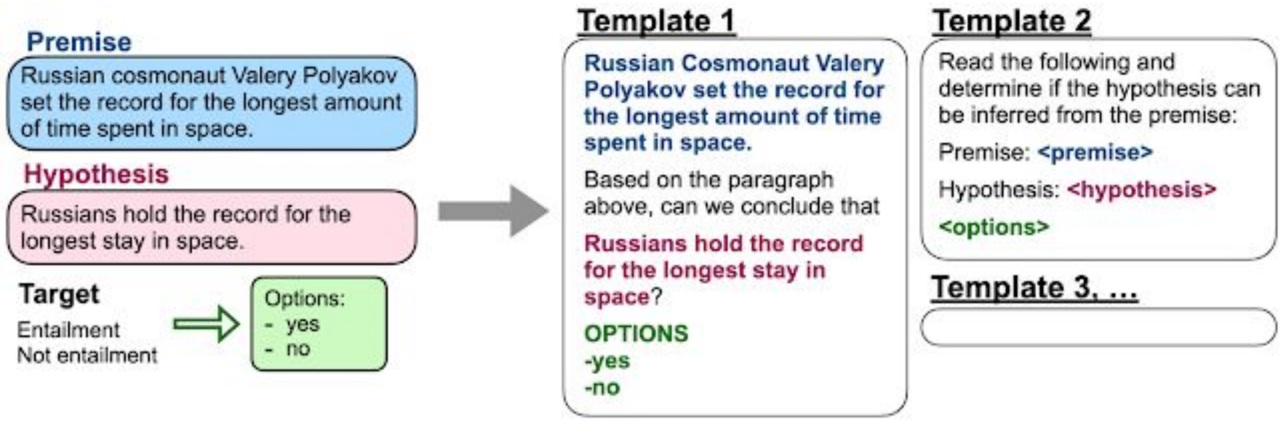
Smaranda Muresan

smara@cs.columbia.edu

Meta AI

Natural Language Processing
Columbia University

Tasks as Instructions



Fine-tuned Language Models are Zero-Shot Learners Wei et al (2022)

Tasks as Instructions

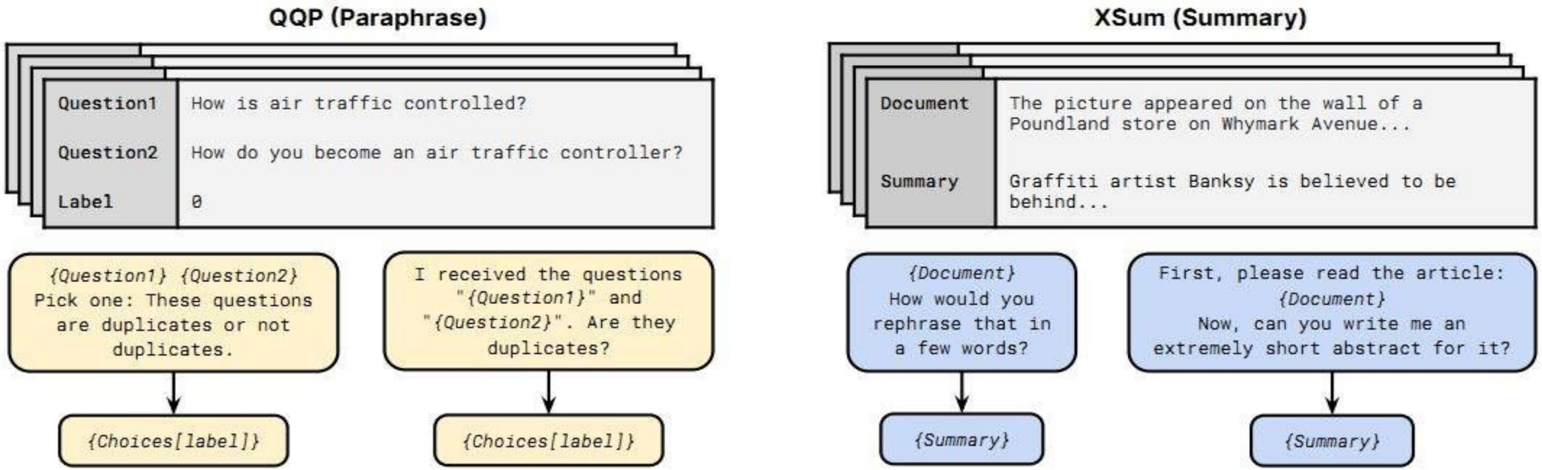
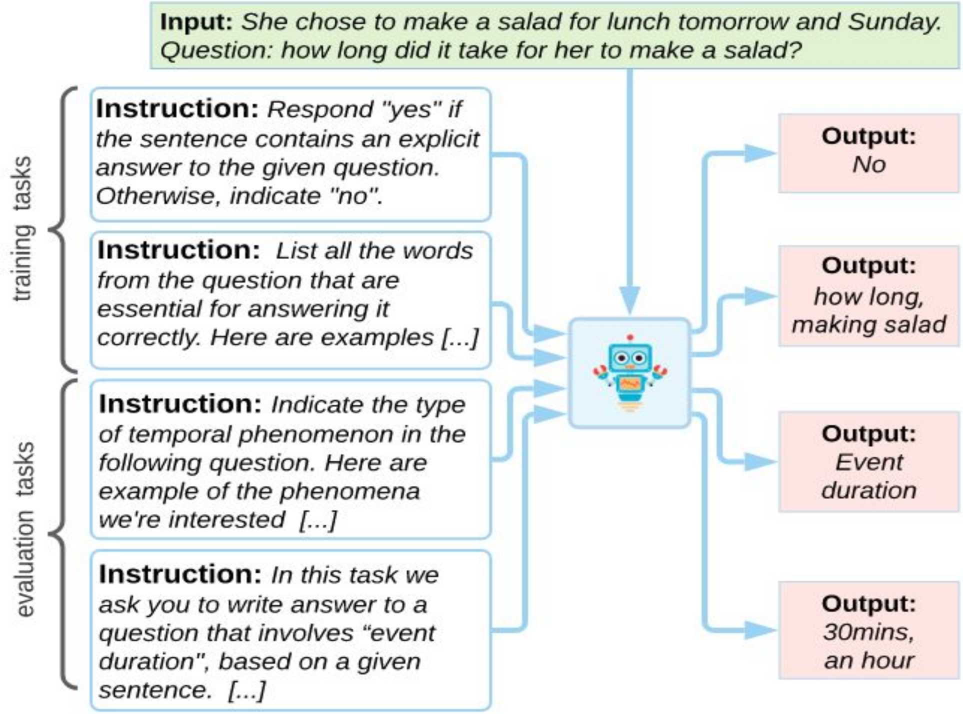


Figure 3: Prompt templates from the P3 prompt collection. Each dataset has multiple prompt templates consisting of an input and target template. Italics indicate the formatting instructions. These use the raw fields of the example as well as template metadata. For example, the paraphrase prompts use *Choices*, a template-level variable consisting of `Not duplicates`, `Duplicates` for the first prompt and `No`, `Yes`. These templates are materialized to produce the prompted instance shown in Figure 1. The complete set of prompt templates used in T0 is given in Appendix G.

Multitask Prompted Training Enables Zero-Shot Task Generalization Sanh et al (2022)

Tasks as Instructions



Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks (Wang et al 2022)

Context

- **Instruction Tuning:**

Pretrained LM (T5) — Self-supervision

Instruction Tuning (T0) — Massively trained to perform many tasks via natural language instructions

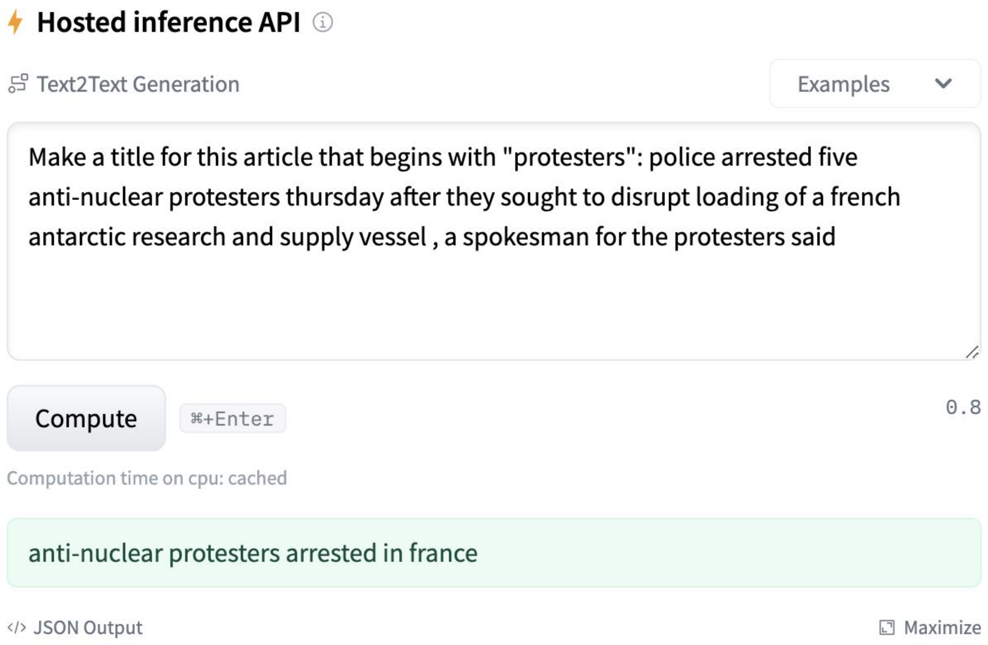
ZeroShot Inference — Inference on unseen tasks

Courtesy to
https://ai.googleblog.com/2021/10/introducing-flan-more-generalizable.html

➢ Impressive zero-shot performance

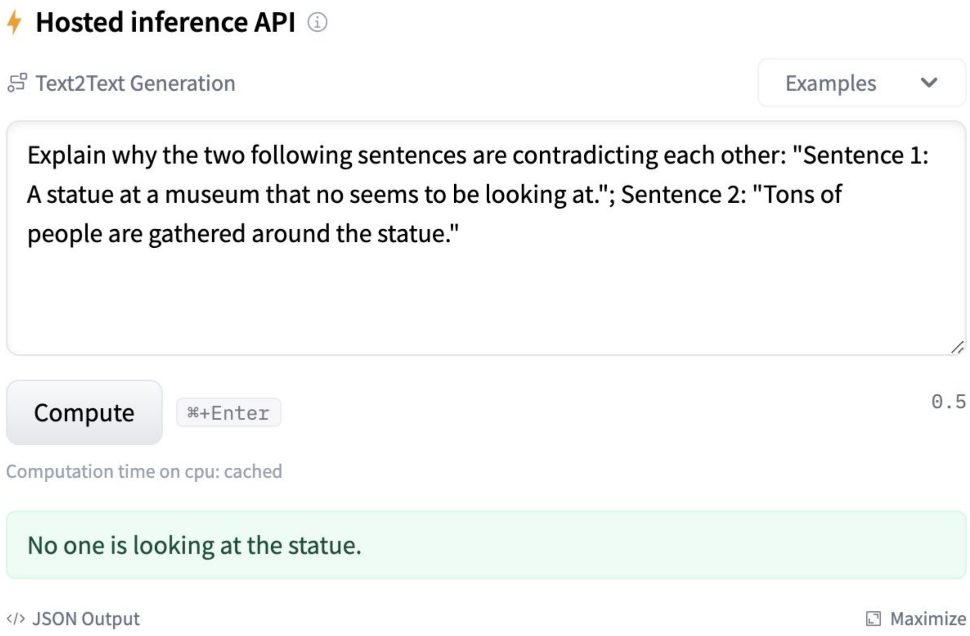
Context

- **Instruction Tuning:**
  - Still performs poorly on a wide range of tasks in a zero-shot setting:



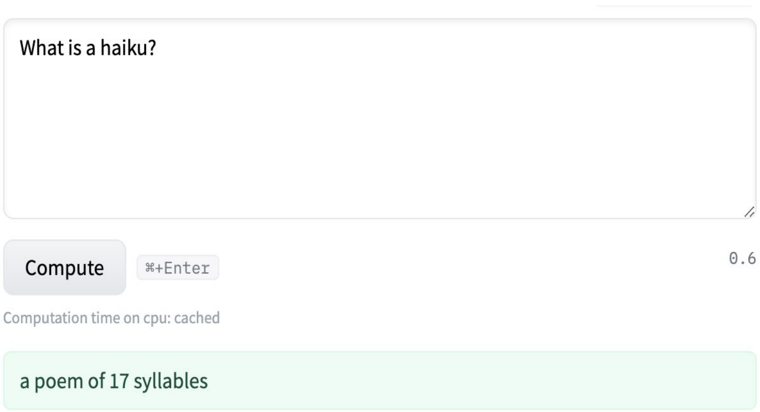
Context

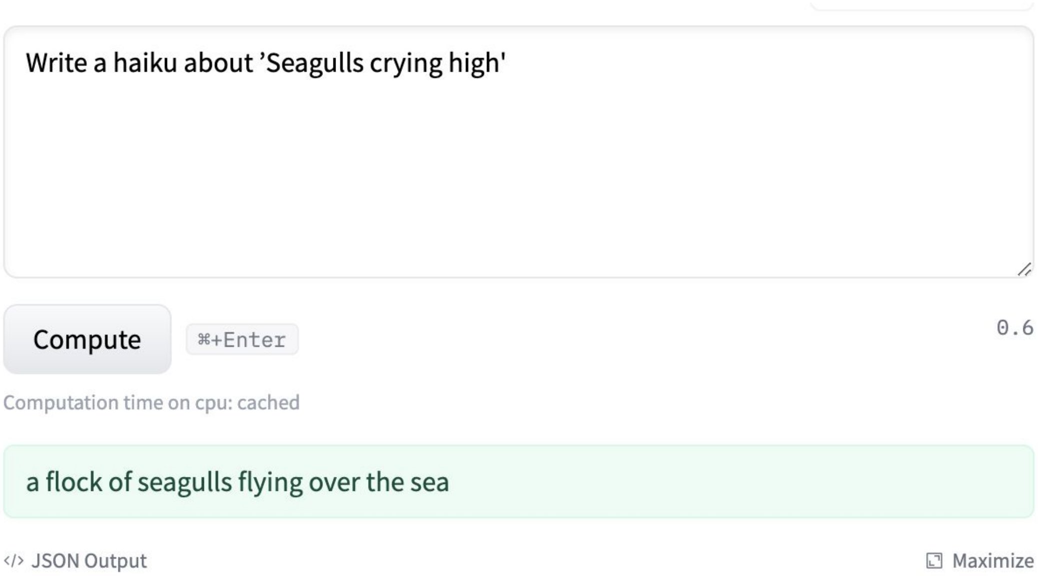
- **Instruction Tuning:**
  - Still performs poorly on a wide range of tasks in a zero-shot setting:



Context

- **Instruction Tuning:**
  - Still performs poorly on a wide range of tasks in a zero-shot setting:



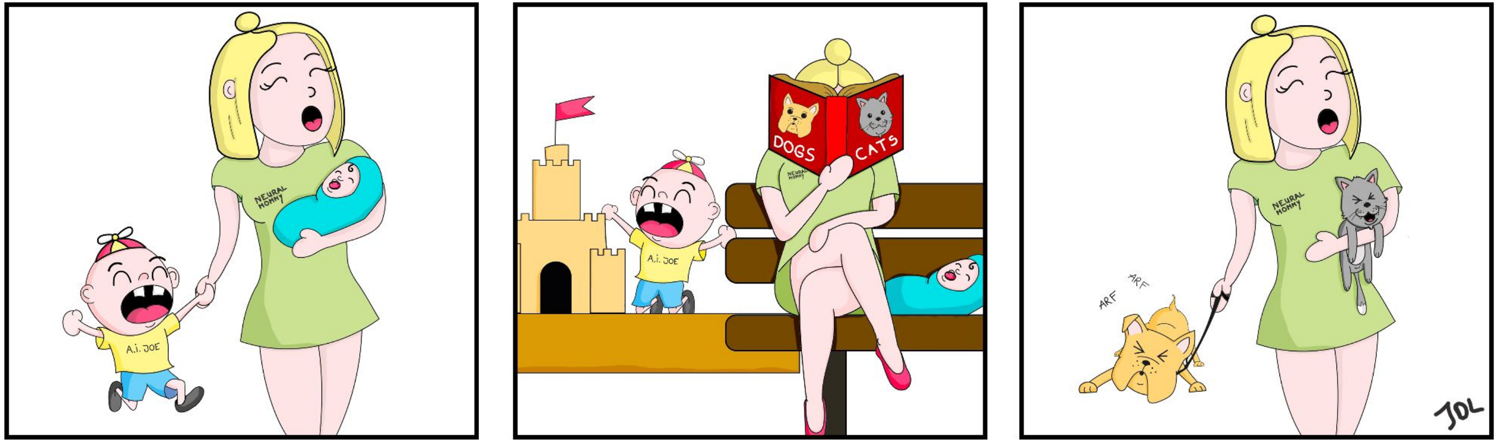


Context

- **Instruction Tuning:**
  - Still performs poorly on a wide range of tasks: To improve their ability on new and diverse tasks, one needs to fine-tune these models again

CATASTROPHIC FORGETTING

Why neural networks make terrible mothers...



Context

- **Instruction Tuning:**
  - Still performs poorly on a wide range of tasks:

➢ How to keep learning new tasks without forgetting existing skills?
Continual Learning

Continual Learning

- A widely studied topic
  - > 20 papers accepted at NeurIPS 2022
  - A new dedicated team at DeepMind

The Continual Learning Team, led by Marc'Aurelio Ranzato

- Several complex algorithms with poor results so far

- An open research question

> *"The fact that catastrophic forgetting is a thing, indicates that something is fundamentally wrong with our approaches"*
>
> Nicholas Roy (MIT) at the CVPR 2022 pan*thing*el on Embodied AI

Why T0 to study Continual Learning?

Clearly there has been great progress with these models, however

FLAN (Wei et al 2022) is not publicly available and given its size (137B) it probably is highly resource intensive and perhaps difficult to use it in an academic setting

Conversely **T0** (Sanh et al 2022) is publicly available and orders of magnitude smaller and hence we resort to working with **T0**.

Continual Learning Strategies

1. **Architectural Strategies**: specific architectures, layers, activation functions, and/or weight-freezing strategies are used to mitigate forgetting.

2. **Regularization Strategies**: the loss function is extended with terms promoting selective consolidation of the weights which are important to retain past memories. Include regularization techniques such as weight sparsification, dropout, early stopping.

3. **Rehearsal Strategies**: past information is periodically replayed to the model, to strengthen connections for memories it has already learned. A simple approach is storing part of the previous training data and interleaving them with new patterns for future training. A more challenging approach is pseudo-rehearsal with generative models.

Continual Learning via Rehearsal

We define the tasks to be learned as a task sequence $T = (T_1, T_2, \cdots, T_N)$ of N tasks. D_i is the corresponding dataset for task T_i. Formally, the training data augmented with **rehearsal** D^r_i is defined as:

$$D_i^r = D_i \bigcup \sum_{j=1}^{i-1} (rD_j)$$

where **r** is the rehearsal hyper-parameter that controls the percentage of examples sampled from previous tasks $T_1, T_2, \cdots, T_{i-1}$. We note that **r=0** corresponds to no memory, and **r=1** is equivalent to a multi-task setup using all the previous examples.

T0 Training tasks

Multiple-Choice QA: CommonsenseQA, DREAM, QUAIL, QuaRTz, Social IQA, WiQA, Cosmos, QASC, Quarel, SciQ, Wiki Hop , ARC, OpenBook QA, PiQA, RACE, BoolQ,

Extractive QA: Adversarial QA, Quoref, DuoRC, ROPES, Squad v2

Closed-Book QA: Hotpot QA*, Wiki QA, Trivia QA, Web Questions

Structure-To-Text: Common Gen, Wiki Bio

Sentiment: Amazon, App Reviews, IMDB, Rotten Tomatoes, Yelp

Summarization: CNN Daily Mail, Gigaword, MultiNews, SamSum, XSum

Topic Classification: AG News, DBPedia, TREC

Paraphrase Identification: MRPC, PAWS, QQP

T0 Zero-Shot Evaluation tasks

Natural language inference: ANLI, CB, RTE

Coreference resolution: WSC, Winogrande

Word sense disambiguation: WiC

Sentence completion: COPA, HellaSwag, Story Cloze

New tasks to be learned

Text Smpfl(Simp)	Instruction	Make this text simpler: "A Georgian inscription around the drum attests his name."	ASSET, Wiki Auto
	Output	*A Georgian writing on the drum is his name.*	
Headline Generation (HGen)	Instruction	Make a title for this article that begins with "protesters": police arrested five anti-nuclear protesters thursday after they sought to disrupt loading of a french antarctic research and supply vessel , a spokesman for the protesters said .	Gigaword
	Output	*protesters target french research ship*	
Haiku Gen (Haiku)	Instruction	Generate a haiku about 'Seagulls crying high'	r/Haiku
	Output	*Seagulls crying high / the air smelling of sea salt / Or is it my tears?*	
Covid QA (CQA)	Instruction	In the context of the COVID pandemic,who is at greater risk of dying from COVID19?	Covid QA (Moller et al 2020)
	Output	*patients with underlying medical conditions and the elderly*	
Inquisitive Question Gen(InqQG)	Instruction	Given the following text, write the possible curious question it answers: "Positrons do not travel backwards in time. Positron-electron annihilation .... So, we know they collide frequently enough to light up the galaxy in that part of the spectrum	ELI5
	Output	*How often do electrons and positrons collide in nature?*	
Empathetic Dialog Generation (EmDg)	Instruction	The associated emotion is "disappointed". Now what would be your response, given the following dialogue context:=== - I had to cancel our family vacation coming up next month.	Empathetic Dialog (Rashkin et al 2019)
	Output	*I am really sorry to hear that. I hope everything is alright.*	
Explanation Generation (Exp)	Instruction	Explain why the two following sentences are contradicting each other: "Sentence 1: A statue at a museum that no seems to be looking at."; Sentence 2: "Tons of people are gathered around the statue."	eSNLI (Camburu et al 2018)
	Output	*If tons of people are gathered around the statue, it is not possible that no one seems to be looking at it.*	
Twitter Stylometry (TwSt)	Instruction	Write a tweet about #WelcomeToNewYork, in the style of taylorswift13	Tareaf et al (2017)
	Output	*GUYS. #WelcomeToNewYork will be up on iTunes any minute now. This is not a drill!! GO GO GO*	

Evaluation for new tasks

Text Simplification	BLEU-4 & SARI
Headline Generation	ROUGE-1 & Constraint Satisfiability
r/Haiku	{ 0.33*(0.5*(#$syll_{pred}$ - #$syll_{gold}$) + 0.5*(#$numlines_{pred}$ - #$numlines_{gold}$)) + 0.33*BLEU-4(pred,gold) + 0.33*Constraint(0/1) }
Covid QA (Moller et al 2020)	BertScore (deberta-mnli)
ELI5	1Tok/BertScore (deberta-mnli)
Empathetic Dialog (Rashkin et al 2019)	BertScore (deberta-mnli)
eSNLI (Camburu et al 2018)	BertScore (deberta-mnli)
Tareaf et al (2017)	Style Classifier & BertScore (deberta-mnli)

Continual T0 with rehearsal

We used the same hyper-parameters as the ones reported in T0 paper. The only new hyper-parameter introduced in our paper is the ***rehearsal proportion*** ***r***. We explore $r \in [0\%, 0.25\%, 1\%]$

For each of T0 training tasks, we consider 100,000 examples for training, such that 1% rehearsal corresponds to 1,000 examples that will be used as the memory buffer for rehearsal. Thus, for datasets with fewer training examples, we upsample them and conversely for largest datasets like Gigaword or Simplification, we limit to 100,000 examples.

Note that here, while we used *rehearsal for the **training** data of T0 training tasks*, *we never used any data from T0 zero-shot tasks, so it remains completely zero-shot*. **It is important to highlight that rehearsal is the standard for CL, and a zero-shot set up with no rehearsal has never been explored yet to the best of our knowledge.**

Continual T0 with rehearsal

First, we test CLR independently on three tasks (Headline Generation with Constraint, Simplification, and Haiku Generation), by varying the rehearsal hyper-parameter between 0%, 0.25% and 1%, respectively

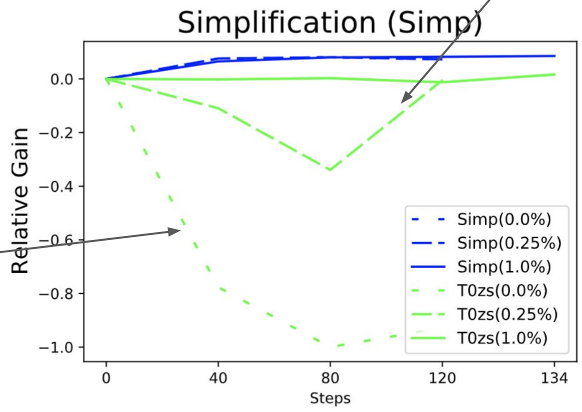
We observe that for the three tasks, the rehearsal value does not affect the task result: Conversely, the rehearsal value has a dramatic impact on the T0 zero-shot results

Continual T0 with rehearsal

The results are normalised in % such that -1 corresponds to 100% decrease and +1 means +100% increase w.r.t. the initial performance

- Starting from a single new task to learn
  - Blue: performance on the new task
  - Green: ZeroShot performance

0.25% or 1% rehearsal: No forgetting



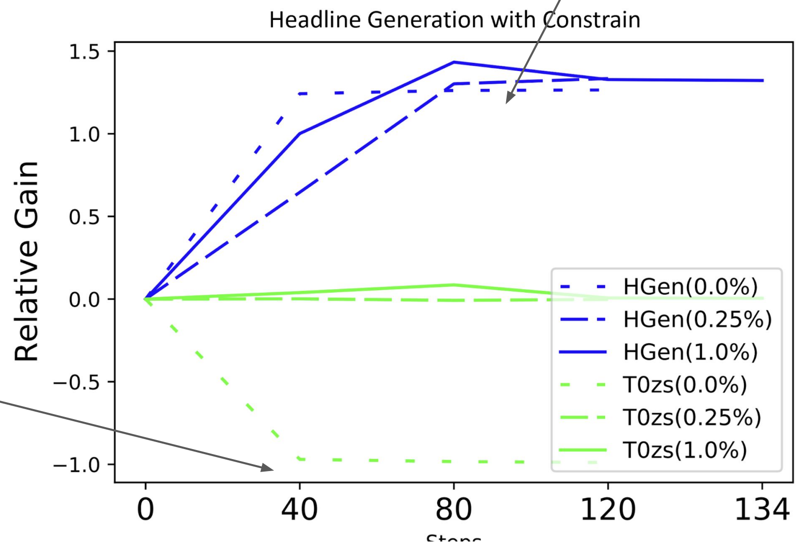
No rehearsal: catastrophic forgetting

Continual T0 with rehearsal

The results are normalised in % such that -1 corresponds to 100% decrease and +1 means +100% increase w.r.t. the initial performance

- Starting from a single new task to learn
  - Blue: performance on the new task
  - Green: ZeroShot performance

0.25% or 1% rehearsal:
No forgetting



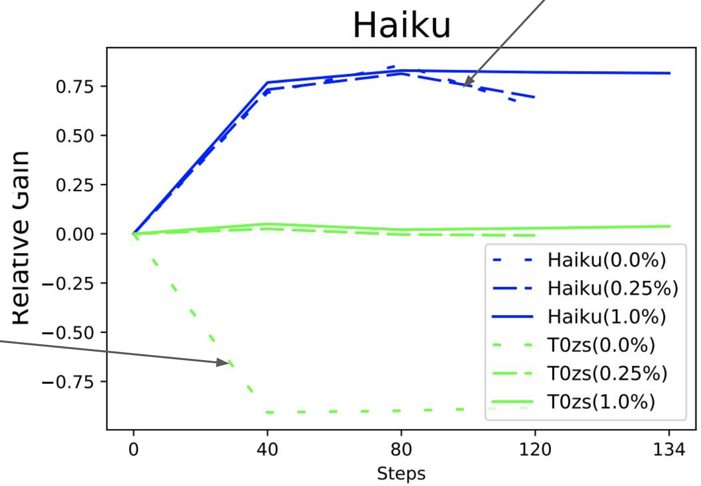
No rehearsal:
catastrophic forgetting

Continual T0 with rehearsal

The results are normalised in % such that -1 corresponds to 100% decrease and +1 means +100% increase w.r.t. the initial performance

- Starting from a single new task to learn
  - Blue: performance on the new task
  - Green: ZeroShot performance

0.25% or 1% rehearsal: No forgetting



No rehearsal: catastrophic forgetting

Continual T0 with rehearsal

As observed from our previous experiments using Continual Learning via rehearsal we can learn a new task at any time without catastrophic forgetting, with just a very little rehearsal percentage.

As a next step, we propose to measure whether language models can progressively learn more tasks without catastrophic forgetting. This is an important direction as it would allow the models to continually increase their knowledge and capabilities without forgetting the knowledge already acquired.

Continual T0 with rehearsal

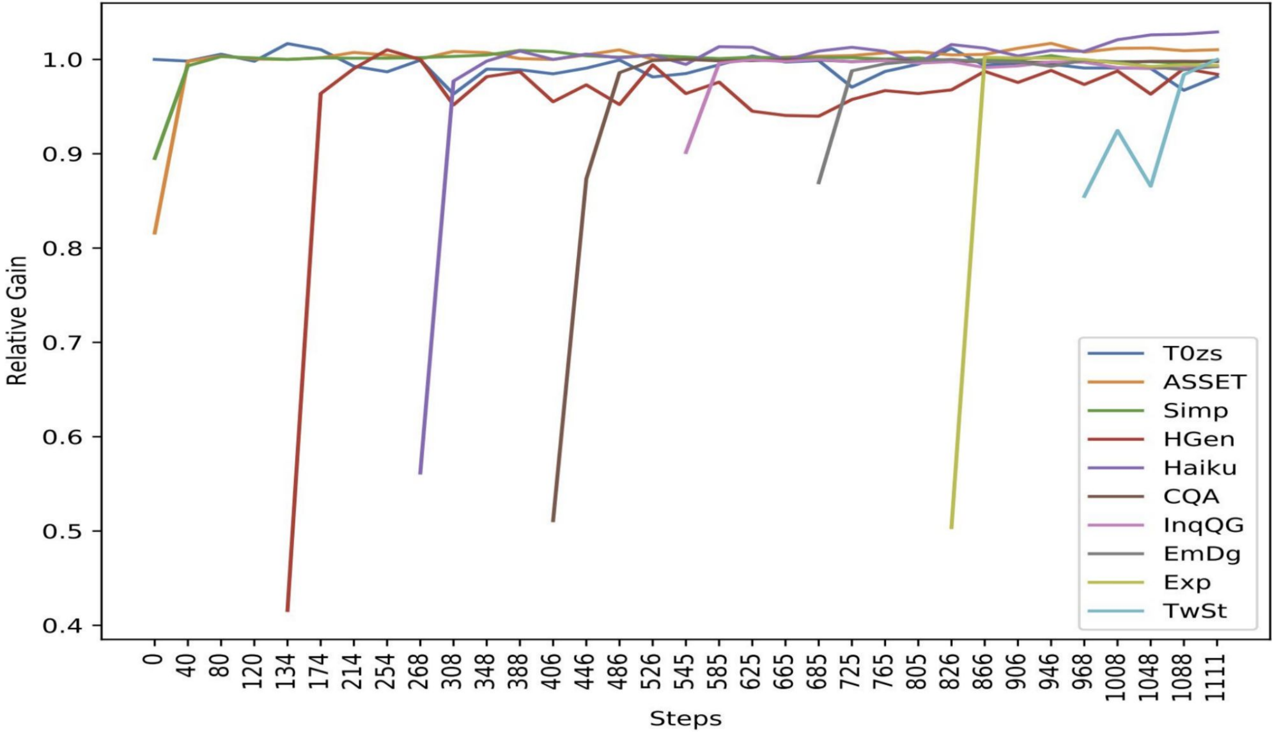
To measure the actual success for CL on a sequence of N tasks, we introduce the notion of *Upper Bound (UB)*. UB corresponds to the maximum performance achieved by the model, when fine-tuned only on a specific task, T_n.

Arguably, the model succeeds in CL, if it maintains a performance close to UB, while learning new tasks.

The normalised results, i.e , *Relative Gain* for a given task T_n , correspond to the actual scores s divided by their task T_n UB, s_{Tn}/UB_{Tn} .Hence, 1 corresponds to performing similar to the UB for any task. The model is expected to start below 1 before step ***n*** since it has not been trained yet on T_n , while for the latest steps t with t > n , results below 1 indicate task forgetting.

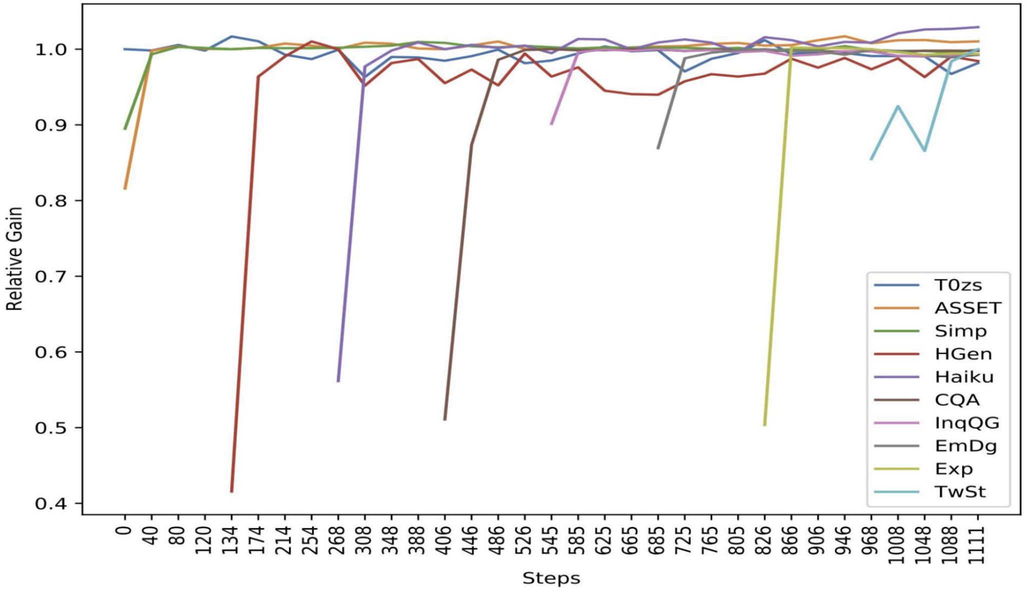
Continual T0 with rehearsal

- Applying the 1% rehearsal to learn progressively 8 new tasks



Continual T0 with rehearsal

- Applying the 1% rehearsal to learn progressively 8 new tasks



- The 8 tasks are learned and not forgotten
- Performance is maintain for T0 evaluation set (12 ZS datasets + 50 seen datasets)

➢ No more catastrophic forgetting, **LM are Continual Learners.**

Continual T0 with rehearsal

	T0tr R1	T0zs Acc	ASSET B4/SARI	Simp B4/SARI	HGen R1/Cons	Haiku H_{cust}	CQA BS	InqQG 1Tok/BS	EmDg BS	Exp BS	TwSt Clf/BS
T0_3B	49.8	48.2	70.1/41.0	12.8/41.1	33.6/32.2	34.2	47.6	2.1/58.7	48.6	32.7	54.4/38.0
T0pp	54.2	65.6	56.5/37.7	11.7/40.1	34.9/35.9	31.6	46.0	2.4/59.8	49.7	37.2	66.4/45.1
UB_3B	49.8	48.2	79.9/45.2	13.8/44.6	39.7/81.0	62.6	90.0	5.3/63.3	55.7	71.8	74.8/56.5
UB_pp	54.2	65.6	85.3/46.1	15.0/44.8	41.9/86.9	63.9	90.0	4.9/65.7	56.6	73.5	74.4/57.9
Lamol	32.6	33.6	37.3/12.6	8.4/21.4	22.9/33.5	25.8	46.6	1.8/47.9	45.1	27.6	50.1/35.2
CT03B	47.9	46.6	78.0/44.5	14.6/43.7	37.3/77.5	60.4	86.8	5.2/61.9	55.3	72.4	74.8/56.5
CT0pp	**53.7**	**64.4**	**85.9/46.6**	**14.6/44.7**	**40.7/85.5**	**65.8**	**89.8**	**4.8/65.2**	**56.2**	**73.0**	**74.4/57.9**
revfinal	48.1	48.8	83.3/45.4	14.6/43.9	39.0/81.6	61.2	88.6	4.4/61.9	55.0	72.4	73.2/57.3

LAMOL : SOTA with pseudo generative replay *
revfinal : CTO3B in reverse order

Our work

- Some outputs:

Instr	*Write a haiku about* 'mountain winds haunt'
CT0	**mountain winds haunt**, the hollow of the stones, voices echo there.
T0pp	a lone tree in the mountains is haunted by the wind

Instr	*Explain why the two following sentences do not entail each other:* "Sentence 1: A woman with a green headscarf, blue shirt and a very big grin."; Sentence 2:"The woman has been shot."
CT0	A woman cannot be smiling if she has been shot.
T0pp	No

Compositional Instructions

- On model's compositionality:

We explore how our model succeeds in understanding constraint instructions beyond the one it was exposed during training.

Our model was trained on Headline Generation with Constraint (HGen) instructions with only one match, such as *Make a title for this article containing ``X"*. To test generalization, we prompt our CT0 model with unseen instructions with 2 and 3 matches, such as *Make a title for this article containing ``X" and ``Y"*, or Make a title for this article containing ``X" and ``Y" and ``Z".

We also compose instructions from constraint and Twitter Stylometry resulting in instructions such as Write a tweet about X, in the style of Y, containing Z

	HGen			TwSt
# Cons	1	2	3	1
CT0	77.0	56.4	39.5	46.4
$CT0_{NoCons}$	33.6	15.4	8.1	10.7

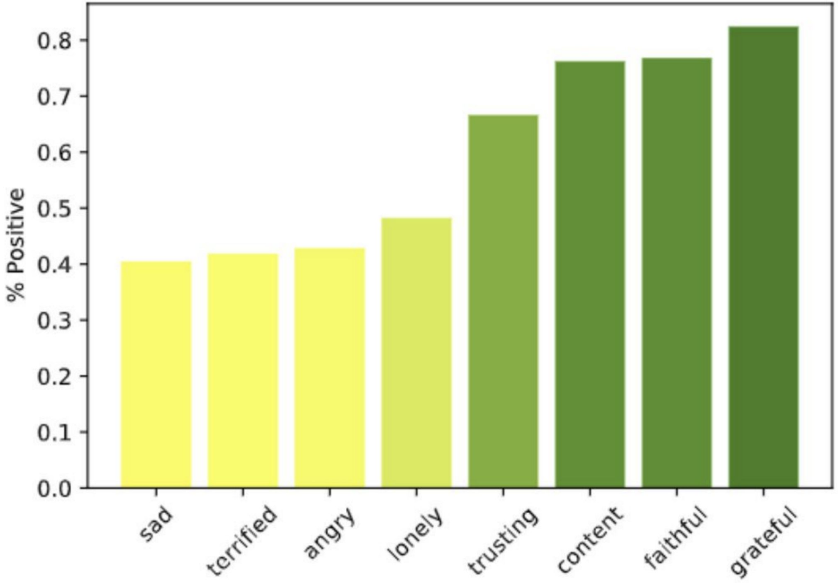
Table 4: Table showing Constraint generalisation i.e % of instructions completely respected, when providing constraints for unseen prompts. $CT0_{NoCons}$ corresponds to providing the same input without constrain.

Compositional Instructions

- On model's compositionality:

Generate a haiku about ``held my hand”. The associated emotion is ``faithful”:

He held my hand through thick and thin/ Through sickness and health/ Through life and death



Success for Continual Learning

- Why does Continual Learning work so well?
  - Instruction tuning or Multi-task training?

    Is it because T0 is a large instruction tuned model trained on a multitask fashion?

    Does it work for T5 as well ?
  - Scaling?

    Are the continual learning capabilities emerging from bigger models ?

    Will it work for a smaller model like T5-small ?

Success for Continual Learning

	T0tr R1	T0zs Acc	ASSET B4/SARI	Simp B4/SARI	HGen R1/Cons	Haiku H_{cust}	CQA BS	InqQG 1Tok/BS	EmDg BS	Exp BS	TwSt Clf/BS
UB_rand	N/A	N/A	0.5/24.3	0.0/29.6	1.5/0.1	9.6	25.2	1.2/25.4	36.3	33.1	24.7
UB_T5small	N/A	N/A	87.8/45.9	15.6/43.2	35.3/67.8	53.4	54.1	3.4/57.0	51.3	33.8	52.4/54.6
UB_T53b	N/A	N/A	87.0/45.6	15.4/43.7	33.0/89.4	63.0	89.9	2.92/61.5	55.3	71.6	75.6/55.4
UB_T0	49.8	48.2	79.9/45.2	13.8/44.6	39.7/81.0	62.6	90.0	5.3/63.3	55.7	71.8	74.8/56.5
CTrand	N/A	N/A	0.0/22.9	0.0/28.5	0.2/0.0	9.6	25.2	1.2/27.9	28.1	30.7	24.7
CT5small	N/A	N/A	85.5/45.8	15.0/42.8	34.6/64.8	51.8	49.5	3.3/56.0	51.2	32.3	52.4/54.6
CT53B	N/A	N/A	84.6/45.8	14.8/44.0	38.3/88.3	62.3	85.8	4.64/62.1	55.5	73.1	75.6/55.4
CT03B	47.9	46.6	78.0/44.5	14.6/43.7	37.3/77.5	60.4	86.8	5.2/61.9	55.3	72.4	74.8/56.5

CT53B is trained in similar way as CT03B just with T5-3B as an initial checkpoint instead of T0-3B
CT5rand is a 3B Transformer randomly initialised.
CT5small is a T5small model trained on only tasks (without converting them to instructions)

Success for Continual Learning

Instruction tuning / Multi-task training / Scaling

Self Supervision (a.k.a Intensive pretraining)

- ➢ **Self-supervision is enough to unlock Continual Learning**

Open Questions

- How to enable Continual Learning without rehearsal?
- How does self-supervision enable Continual Learning?
- Would Continual Learning break for 100 tasks? 1000 tasks?
- Multimodal Continual Learning?

Thanks

tuhin.chakr@cs.columbia.edu

	T0zs Acc	ASSET B4/SARI	Simp B4/SARI	HGen R1/Cons	Haiku H_{cust}	CQA BS	InqQG 1Tok/BS	EmDg BS	Exp BS	TwSt Clf/BS
T0_3B	48.2	70.1/41.0	12.8/41.1	33.6/32.2	34.2	47.6	2.1/58.7	48.6	32.7	54.4/38.0
T0pp (11B)	65.6	56.5/37.7	11.7/40.1	34.9/35.9	31.6	46.0	2.4/59.8	49.7	37.2	66.4/45.1
+Simp 3B	<u>48.9</u>	79.9/<u>45.2</u>	13.8/<u>44.6</u>	30.3/31.0	30.9	43.9	2.0/56.1	40.2	34.9	50.8/42.5
+Simp 11B	**66.7**	85.3/46.1	15.0/44.8	34.9/36.1	33.0	47.2	2.1/59.0	48.1	39.2	68.8/47.6
+HGen 3B	46.9	81.4/44.9	14.1/43.9	<u>39.7/81.0</u>	33.7	44.2	2.5/55.9	45.9	55.2	19.6/37.3
+HGen 11B	65.5	84.5/46.1	**15.3**/44.8	**41.9/86.9**	35.9	46.6	2.9/59.7	48.9	36.4	69.6/48.1
+Haiku 3B	48.8	<u>81.6</u>/45.0	14.6/43.9	39.0/78.2	62.6	43.0	2.3/54.9	47.2	39.0	65.6/44.5
+Haiku 11B	64.6	83.5/46.1	14.9/45.1	41.1/83.0	63.9	46.0	2.9/59.9	48.9	37.5	66.4/46.2
+CQA 3B	48.5	79.7/44.4	14.0/43.8	37.6/75.4	62.2	<u>90.0</u>	2.0/54.4	42.5	38.7	66.4/45.3
+CQA 11B	64.6	84.3/46.1	14.5/**44.9**	40.9/83.7	63.6	**90.0**	2.9/59.2	48.5	42.7	67.2/47.3
+InqQG 3B	47.4	65.2/41.2	14.6/43.8	37.9/77.7	60.4	89.6	<u>5.3/63.3</u>	46.8	34.2	59.2/45.4
+InqQG 11B	65.5	85.5/46.3	14.9/44.8	40.6/81.7	64.5	89.9	4.9/**65.7**	49.2	47.7	61.2/45.9
+EmDg 3B	48.6	73.9/43.8	<u>15.0</u>/43.7	38.0/77.7	<u>62.9</u>	88.6	4.7/62.7	<u>55.7</u>	35.2	53.6/42.7
+EmDg 11B	66.4	85.3/46.3	15.1/44.7	40.9/84.1	65.0	89.9	**5.3**/65.5	**56.6**	37.0	61.6/45.8
+Exp 3B	47.4	74.6/44.0	14.2/43.5	37.9/80.9	60.9	86.5	4.9/62.3	55.2	71.8	54.8/43.4
+Exp 11B	65.0	85.6/46.5	14.9/44.7	40.7/84.6	64.5	89.8	4.8/65.5	56.5	**73.5**	63.6/46.3
+TwSt 3B	46.6	78.0/44.5	14.6/43.7	37.3/77.5	60.4	86.8	5.2/61.9	55.3	<u>72.4</u>	<u>74.8/56.5</u>
+TwSt 11B	64.4	**85.9/46.6**	14.6/44.7	40.7/85.5	**65.8**	89.8	4.8/65.2	56.2	73.0	**74.4/57.9**
rev_final	48.8	83.3/45.4	14.6/43.9	39.0/81.6	61.2	88.6	4.4/61.9	55.0	72.4	73.2/57.3