The Power of Scale for Prompt Tuning

Brian Lester, Rami Al-Rfou, Noah Constant

Presented by Brian Lester at EMNLP 2021

Google Research

01

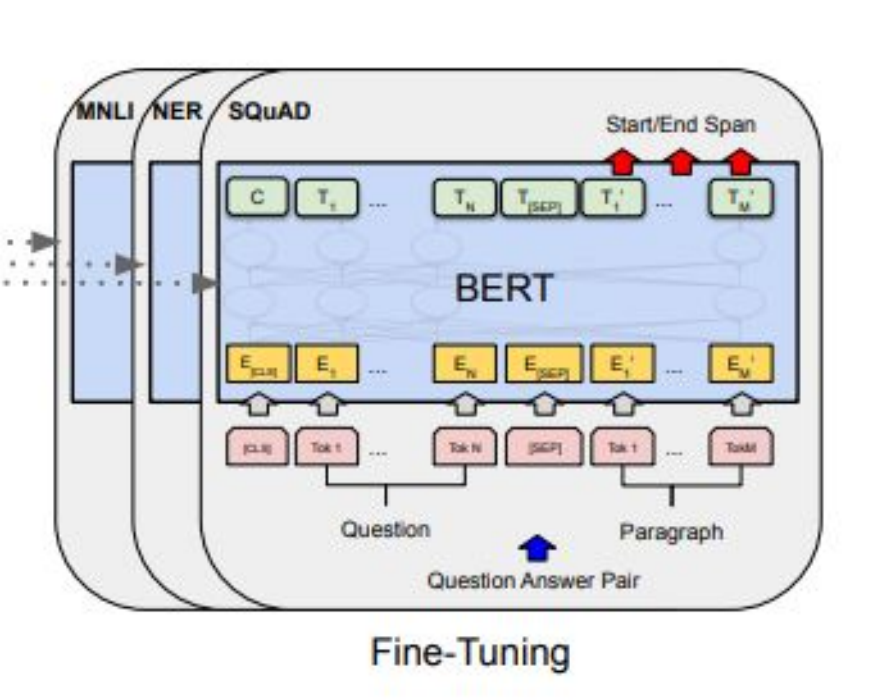
Background: Fine Tuning

Google Research

Fine Tuning

- Given:
  - A pretrained model
  - A labeled dataset
- Update weights of pretrained model by supervised learning on labeled dataset

Strong performance on many tasks. Starting point of most SotA methods today.

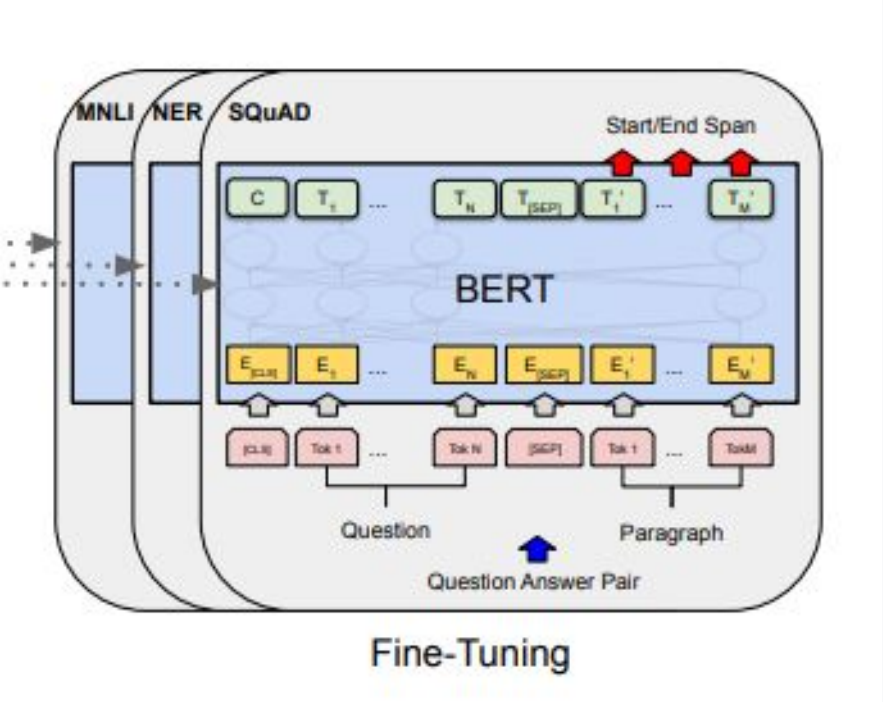


Google Research

Fine Tuning

Some sharp edges:

- Each model you train is a **fork**.
- These **models are so big** even fine-tuning often takes complex SPMD programming and large compute platform.
- Serving can be difficult
  - A model has to be served for each task.
  - Need enough requests to keep model saturated.
  - Swapping models into memory can take a long time.



Google Research

02

Background: Prompt Design

Google Research

Prompt Design

Can we **add text to our model** to trick a language model into doing what we want?

- Knowledge Extraction from LMs
- GPT-3

We don't have to copy our model but there are still problems:

- Human effort to get prompts
- Prompts are specific to models
- Limit to the number of supervised examples you can fit in the input
- Poor performance

Google Research

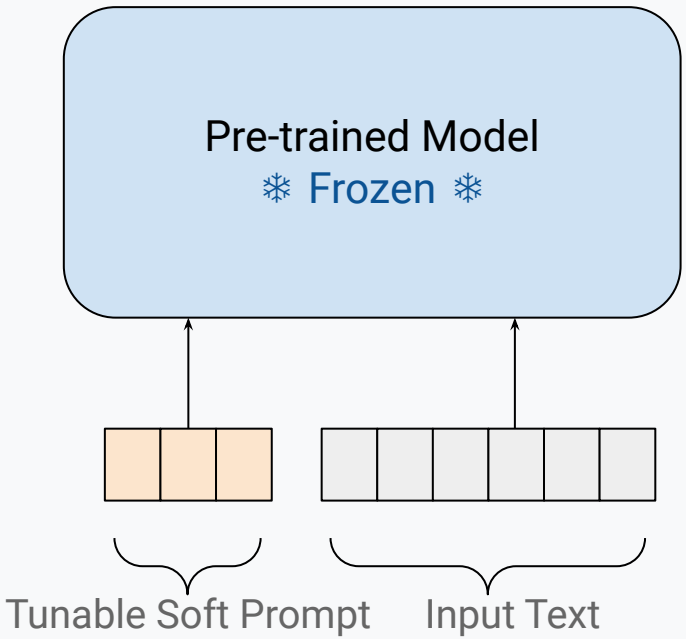
03

Our Approach: Prompt Tuning

Google Research

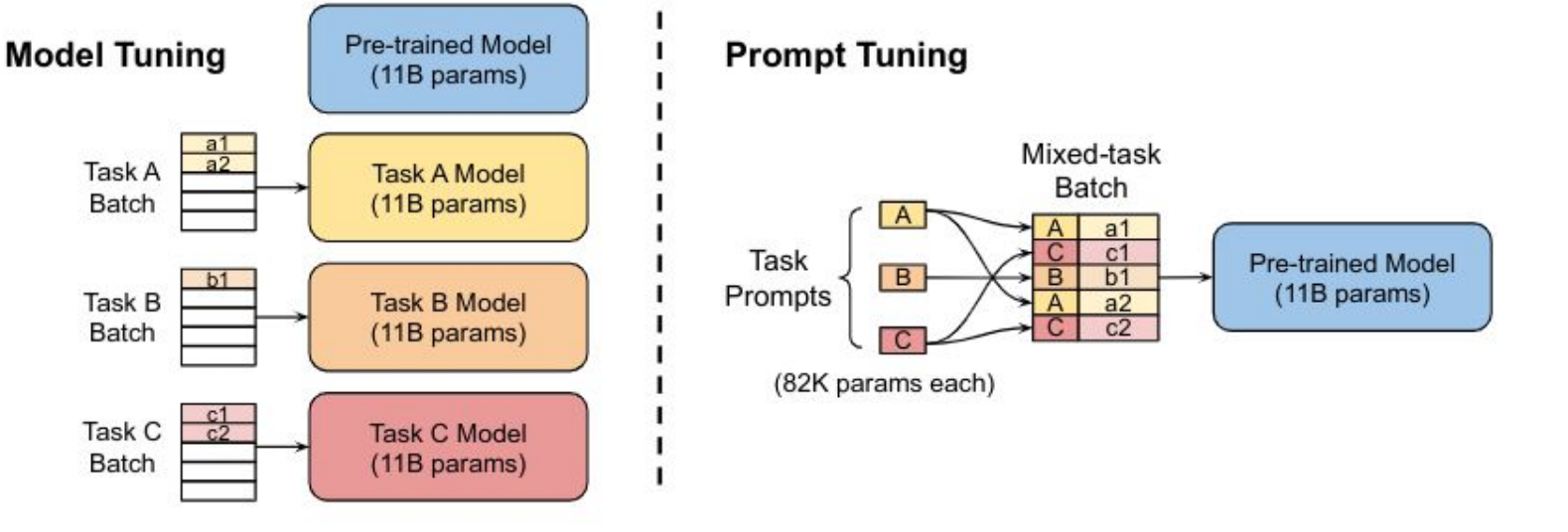
Prompt Tuning

- Prepend **virtual tokens** to input.
- Prompt and input representations flow through model like normal.
- **Learn embeddings of only these special tokens**, via backprop. Keep rest of model **fixed**.
- Advantages:
  - Lets us use whole training dataset.
  - Lets us automatically learn a new prompt for a new model.
  - Lets us keep model frozen.
  - Prompts are much smaller.



Google Research

Prompt Tuning



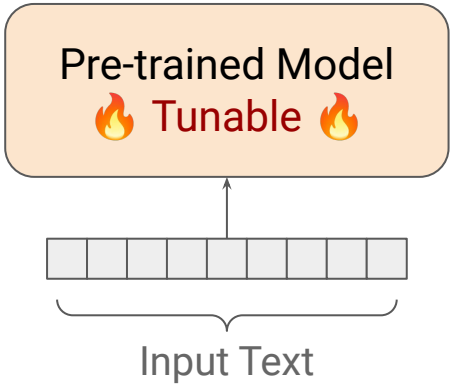
Google Research

Prompt Tuning

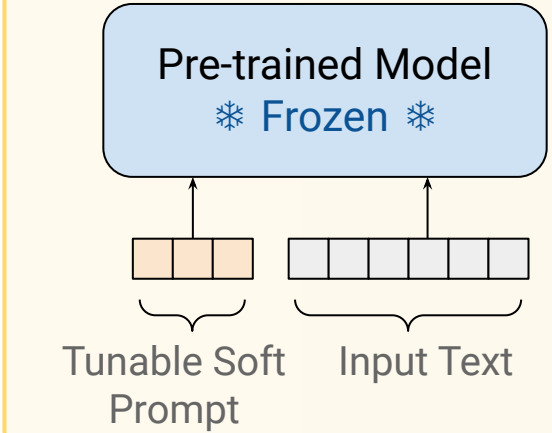
Strong Task Performance

Efficient Multitask Serving

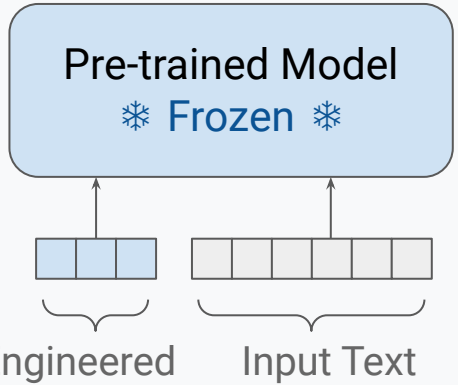
Model Tuning
(a.k.a. “Fine-Tuning”)



Prompt Tuning
(Ours)



Prompt Design
(e.g. GPT-3)

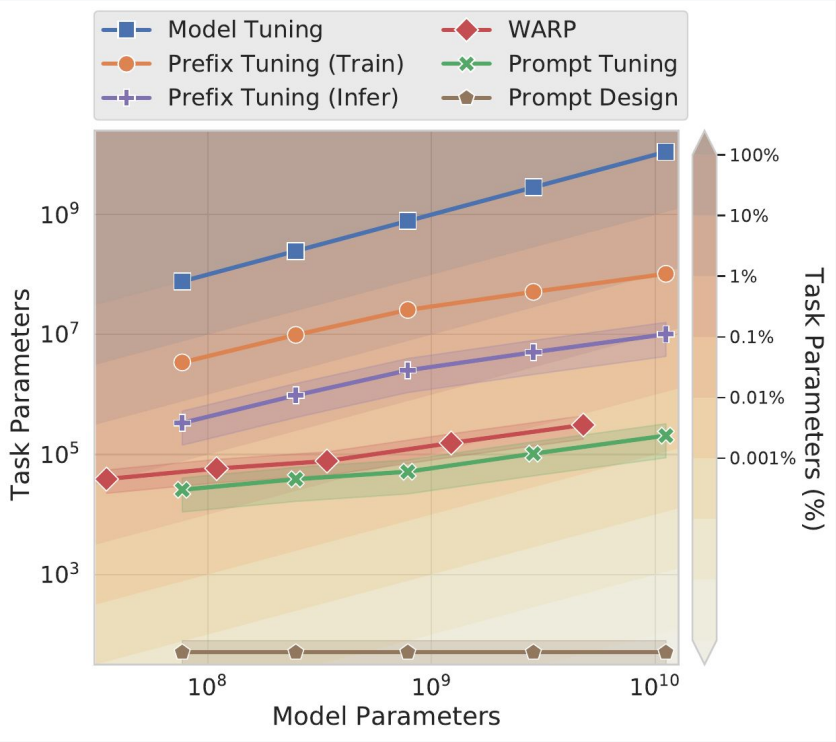


Google Research

Related Work

While we were working on this project a few other papers doing similar work appeared on arXiv

- Li and Liang, 2021
- Hambardzumyan, et. al, 2021
- Liu, et. al. 2021
- Qin and Eisner, 2021
- Logeswaran, et. al., 2020



Prompt Tuning has the smallest parameter count compared to other approaches that focus on learning a continuous representation of prompt tokens.

Google Research

04

Prompt Tuning Results

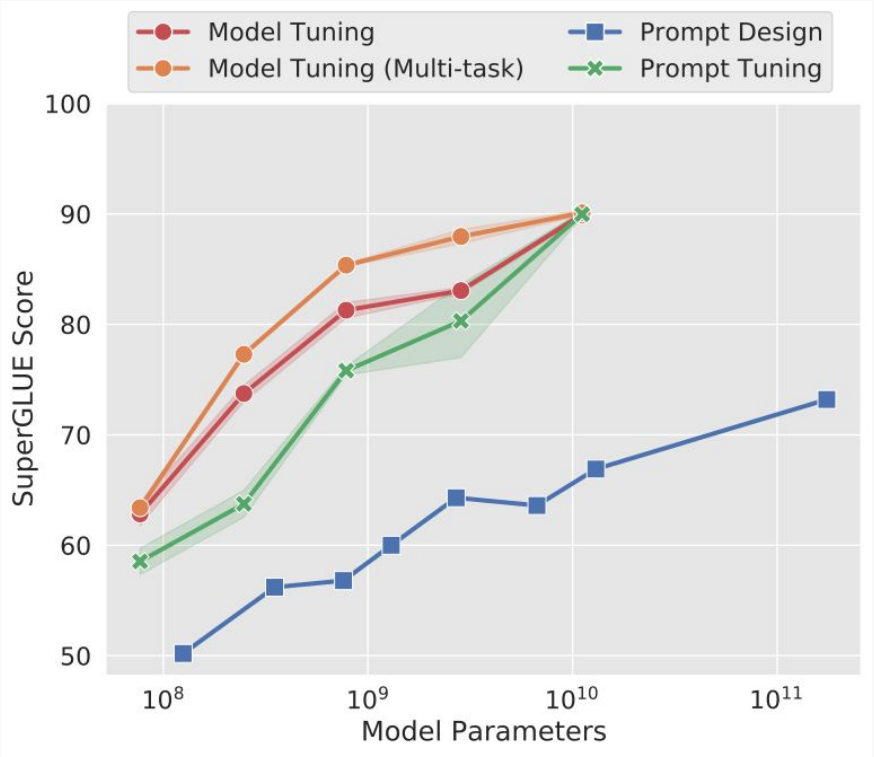
Google Research

Prompt Tuning Results

Prompt Tuning performs very well on SuperGLUE.

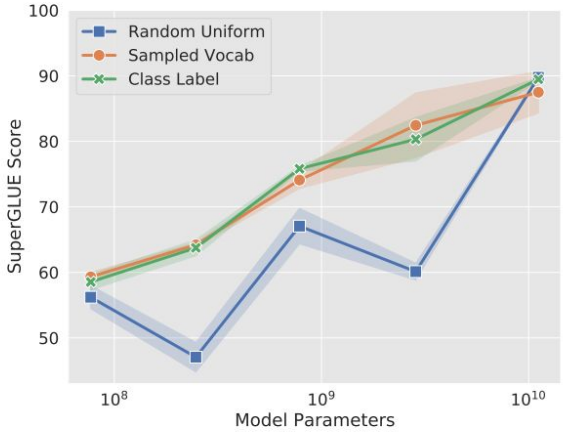
Across all model sizes, Prompt Tuning is far ahead of Prompt Design.

As the model size grows, Prompt Tuning **closes the gap** with Model Tuning.

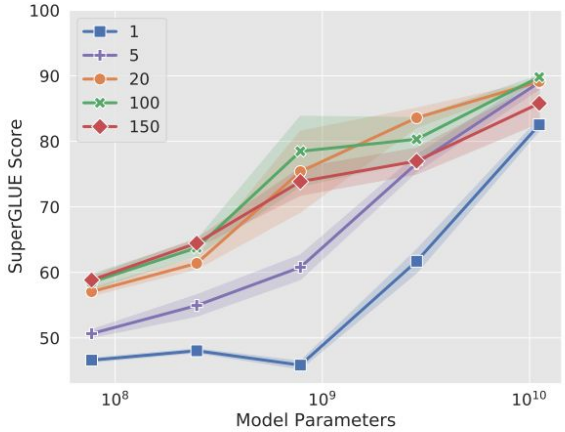


We see that Prompt Tuning is far stronger than prompt design for models of comparable sizes, and that as the size of the LM grows Prompt Tuning catches up to Model Tuning

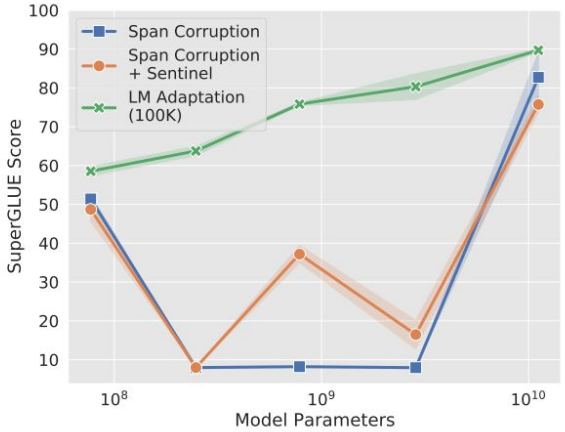
Google Research



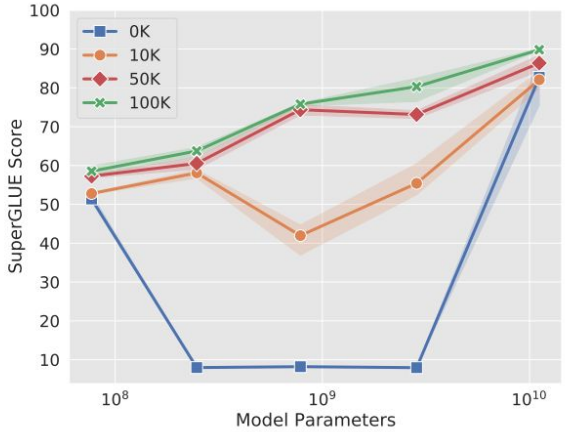
(a) Prompt initialization scheme



(b) Prompt length



(c) Pre-training method



(d) LM adaptation steps

Ablations:

Each plot is SuperGLUE score vs the number of parameters in the frozen model for “Something we ablated”.

- Prompt initialization
- Prompt length
- Pre-trained objective
- LM objective train time

As the scale of the frozen model grows, a lot of these design decisions matter a lot less.

Google Research

Domain Shift - Question Answering

We also look at zero-shot in the context of question answering. We used the MRQA 2019 shared task on generalization of question answering.

Prompt Tuning has stronger zero-shot performance on datasets with larger domain shifts, including a remarkable +12.5 for TextbookQA.

Dataset	Domain	Model	Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8

Prompt Tuning has stronger zero-shot performance on datasets with larger domain shifts, including a remarkable +12.5 for TextbookQA.

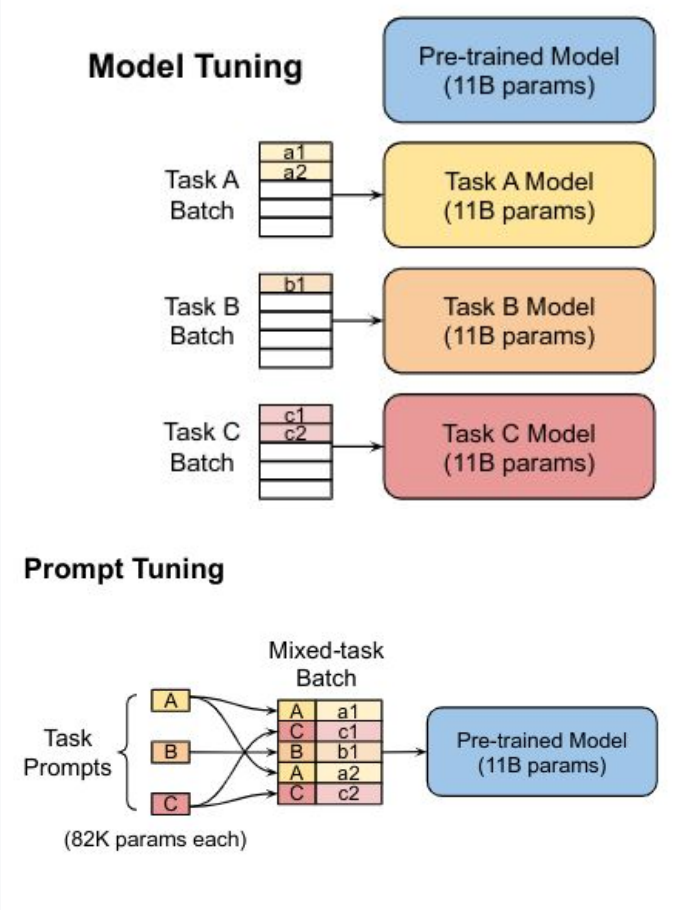
Google Research

Prompt Tuning Ensembles

Prompt Tuning allows for efficient ensembling of large models.

Instead of training (and storing) N copies of a large model, **train N prompts** which are much smaller.

Instead of performing N forward passes through N models, prompt tuning lets us replicate the input, prepend different prompts, and perform a single forward pass with a batch of size N.



Google Research

Prompt Tuning Ensembles

Prompt Tuning allows for efficient ensembling of large models.

Instead of training (and storing) N copies of a large model, **train N prompts** which are much smaller.

Instead of performing N forward passes through N models, prompt tuning lets us replicate the input, prepend different prompts, and perform a single forward pass with a batch of size N.

Ensembling prompts produces stronger results.

Dataset	Metric	Average	Best	Ensemble
BoolQ	acc.	91.1	91.3	**91.7**
CB	acc.	99.3	100.0	**100.0**
	F1	99.0	100.0	**100.0**
COPA	acc.	98.8	100.0	**100.0**
MultiRC	EM	65.7	66.3	**67.1**
	$F1_a$	88.7	89.0	**89.4**
ReCoRD	EM/F1	92.7	92.9	**93.2**
	F1	99.4	93.5	**93.9**
RTE	acc.	92.6	**93.5**	**93.5**
WiC	acc.	76.2	76.6	**77.4**
WSC	acc.	95.8	**96.2**	**96.2**
SuperGLUE (dev)		90.5	91.0	**91.3**

Ensembling 5 XXL prompts using simple majority voting yields higher SuperGLUE scores than taking the best model in the ensemble. RTE and WSC are the only datasets where the ensemble is equivalent to the best model.

Google Research

05

Interpretability

Google Research

Interpretability - Nearest Neighbors

Our “soft prompts” are learned in embedding space, so we need to convert back to tokens. We use cosine distance to find the top-5 nearest neighbors to each token in the prompt.

We see strong semantic clusters in the top-5 neighbors. Some are lexically similar but other more diverse.

We see class labels in neighbors of prompts. They persist in the “class-label” setting and are learned in the other initialization methods.

- Lexically Similar Clusters:
  - *Technology*
  - *technology*
  - *Technologies*
  - *technological*
  - *technologies*
- More Diverse Clusters:
  - *entirely*
  - *completely*
  - *totally*
  - *altogether*
  - *100%*

Google Research

Thank You! Questions?

Brian Lester
Google AI Resident
- Twitter: @blester125
- Site: https://blester125.com

Code, Checkpoints, Prompts, and Examples
- https://github.com/google-research/prompt-tuning

Google Research