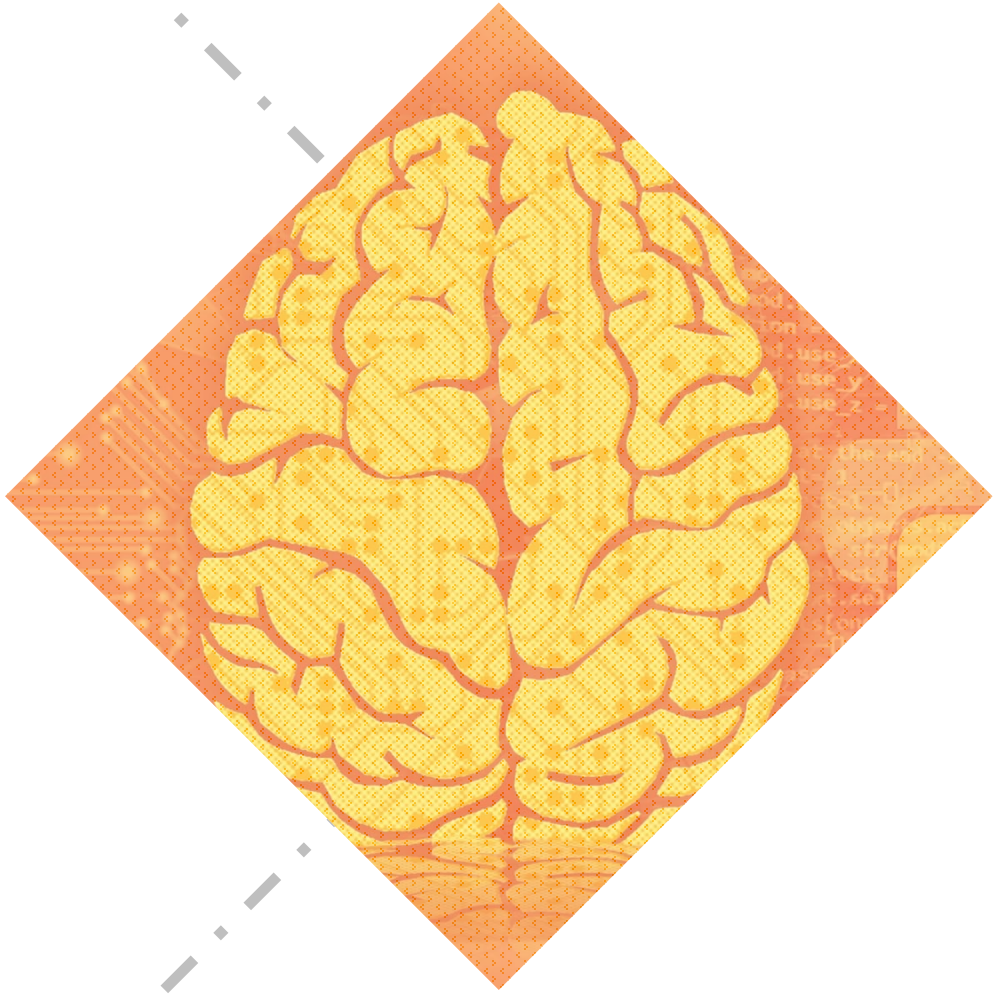
BERTVision

Improving performance on a wide range of **N**atural **L**anguage **P**rocessing tasks using parameter-efficient model architectures training on BERT's hidden state activations

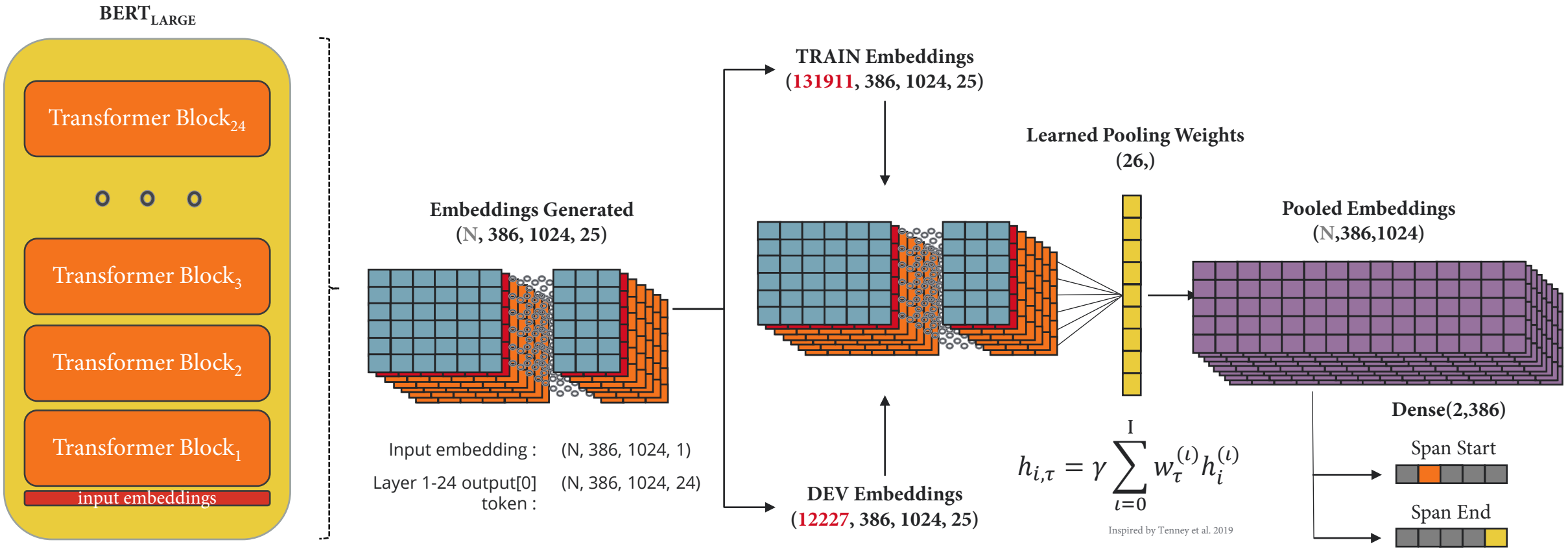
Siduo Stone Jiang *and* **William Casey King** *and* **Andrew Fogarty** *and* **Cristopher Benge**



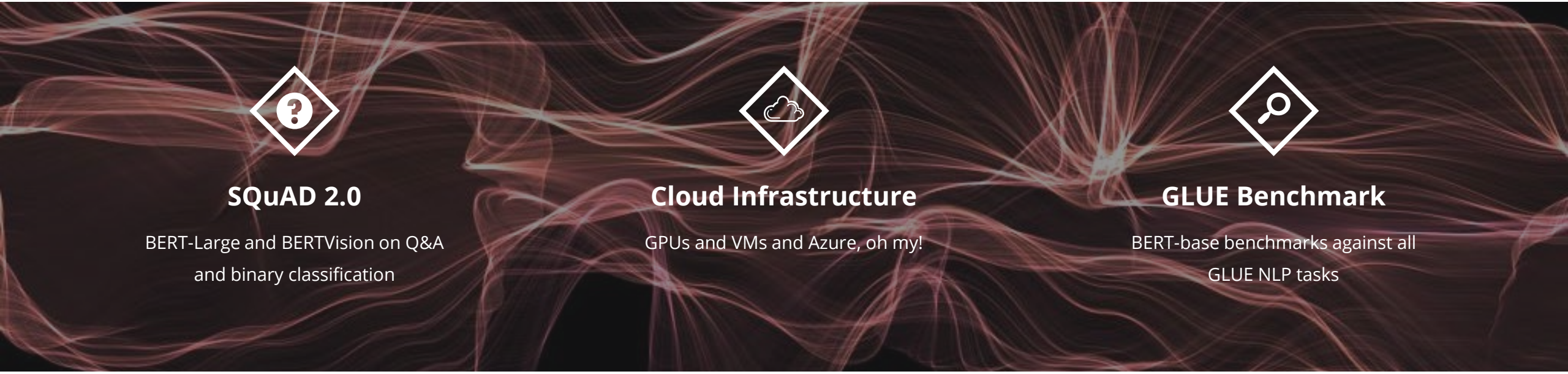
Impact & Feasibility

- Transformer models are powerful, but have a high parameter count and computational footprint.
- Other compression techniques prune parameters or reduce numeric precision, but at a cost.
- SOTA progress in NLP with transformers is on a collision course with absurdity (1T+ parameters).
- Our current method still requires minor fine-tuning of BERT, but we're evaluating transfer learning with the smaller models in this project.
- If our solution ultimately requires BERT weights, we could offer inferencing as-a-service through REST API.

Introducing: BERTVision



Preliminary Work



SQuAD 2.0

BERT-Large and BERTVision on Q&A and binary classification

Cloud Infrastructure

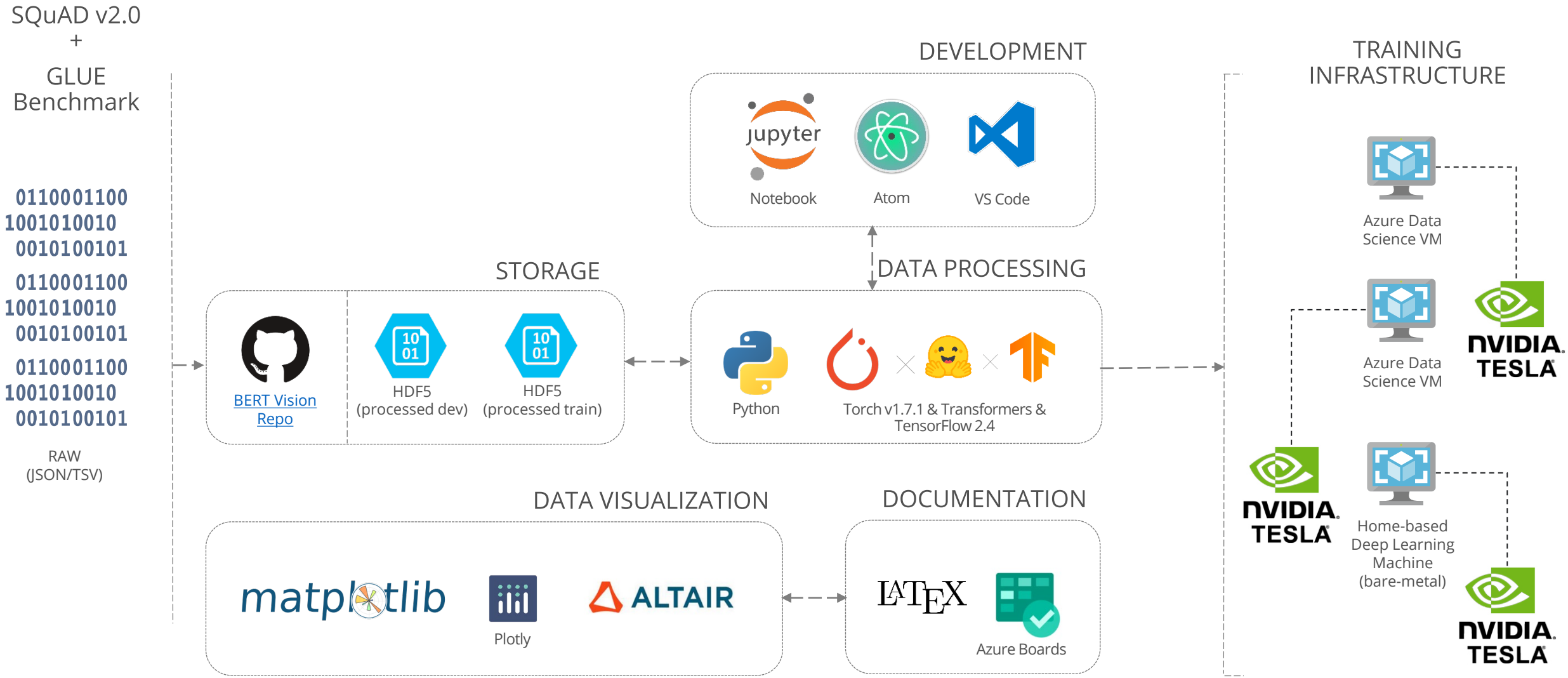
GPUs and VMs and Azure, oh my!

GLUE Benchmark

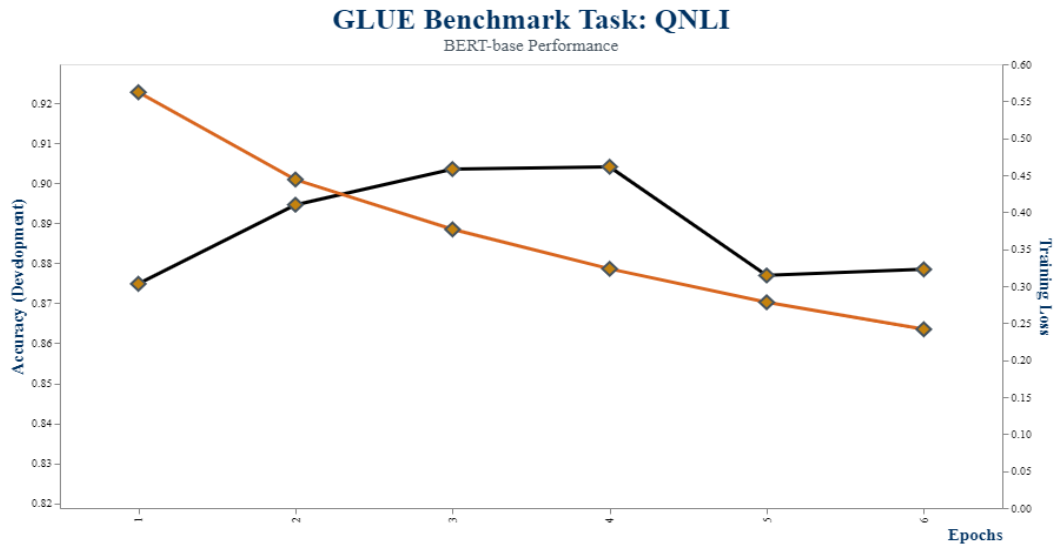
BERT-base benchmarks against all GLUE NLP tasks

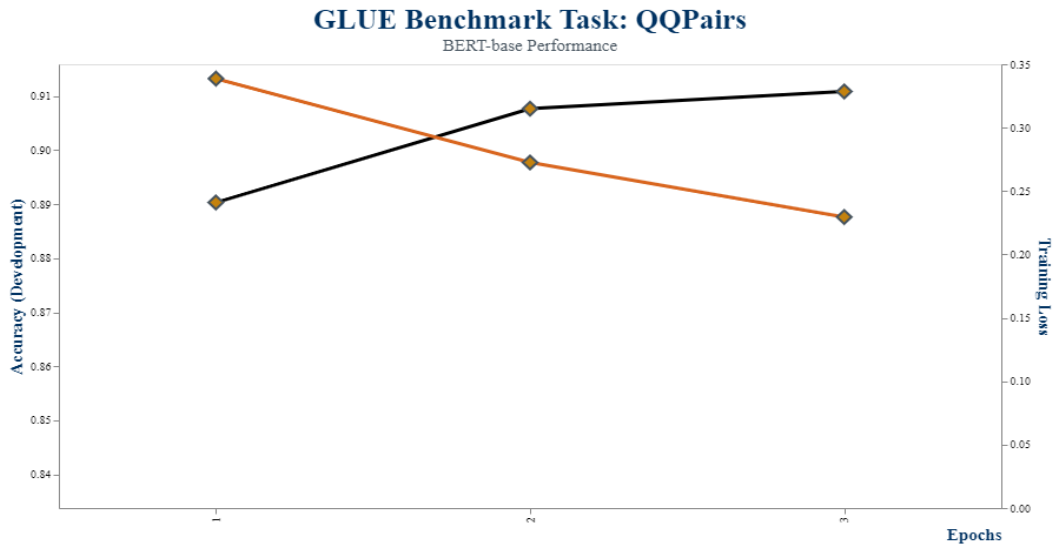
We have a robust pipeline of work completed to-date, including most of our infrastructure needs, a preliminary paper in Overleaf from our results in W266 against SQuAD 2.0, and up-to-date benchmark data on all GLUE NLP tasks for the BERT-base (110M parameter) transformer model

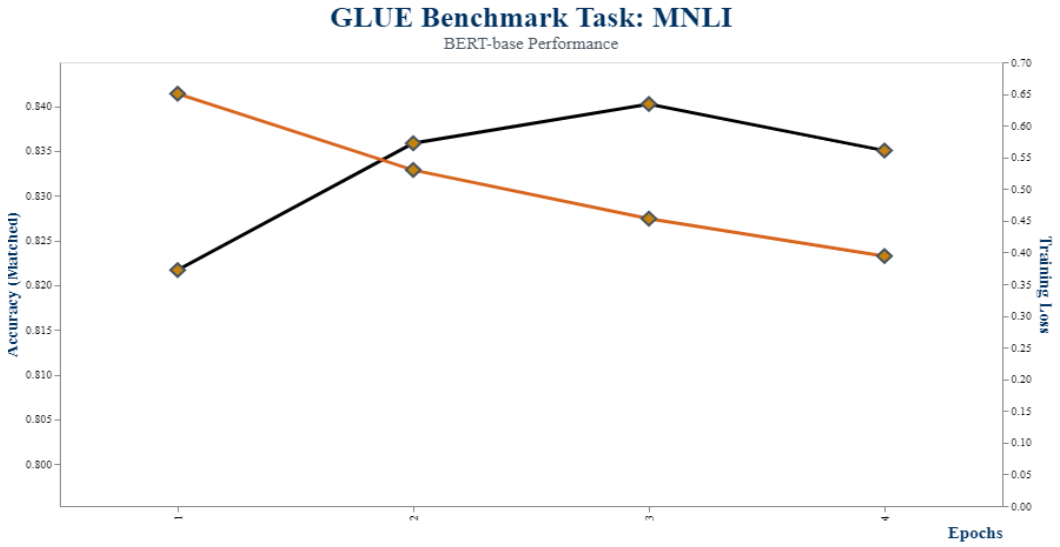
BERTVision Development Pipeline

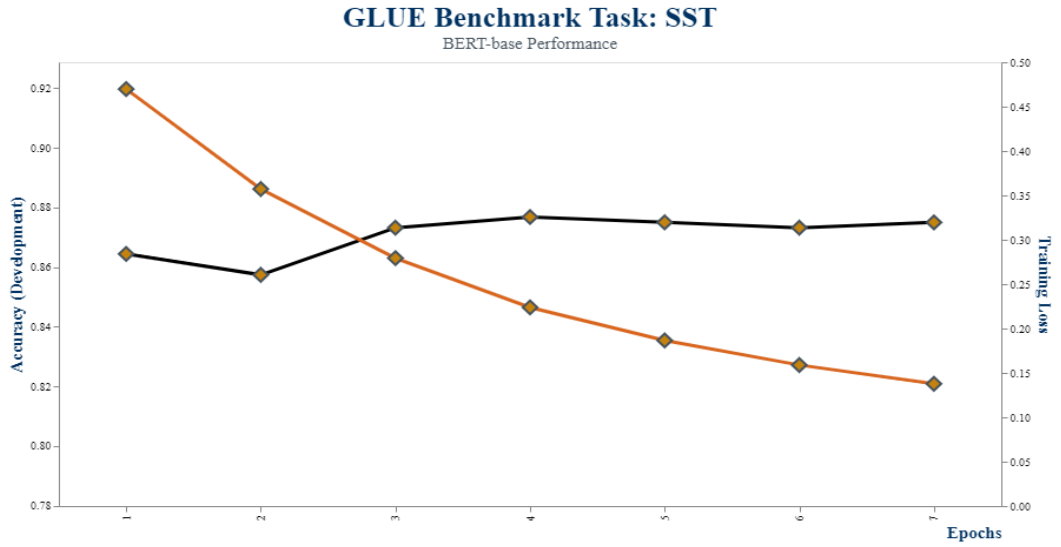


BERTVision GLUE Benchmarks*





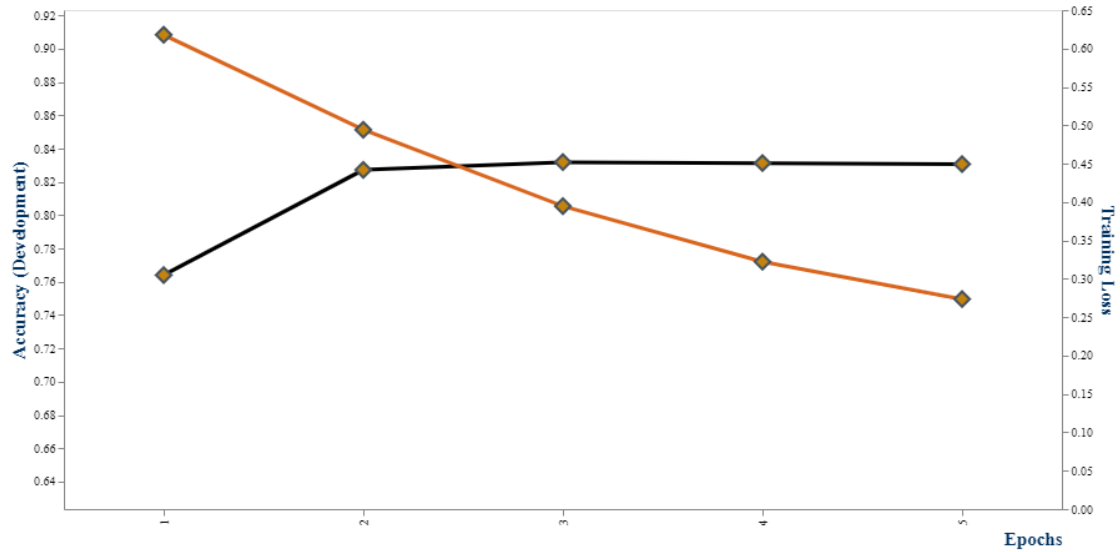




*BERT-base performance

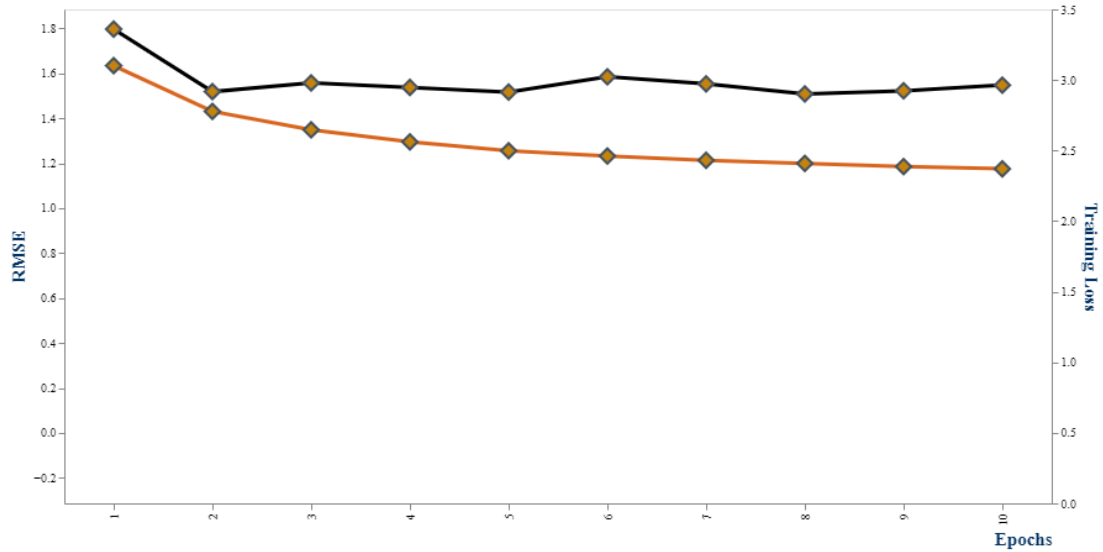
GLUE Benchmark Task: MSR

BERT-base Performance



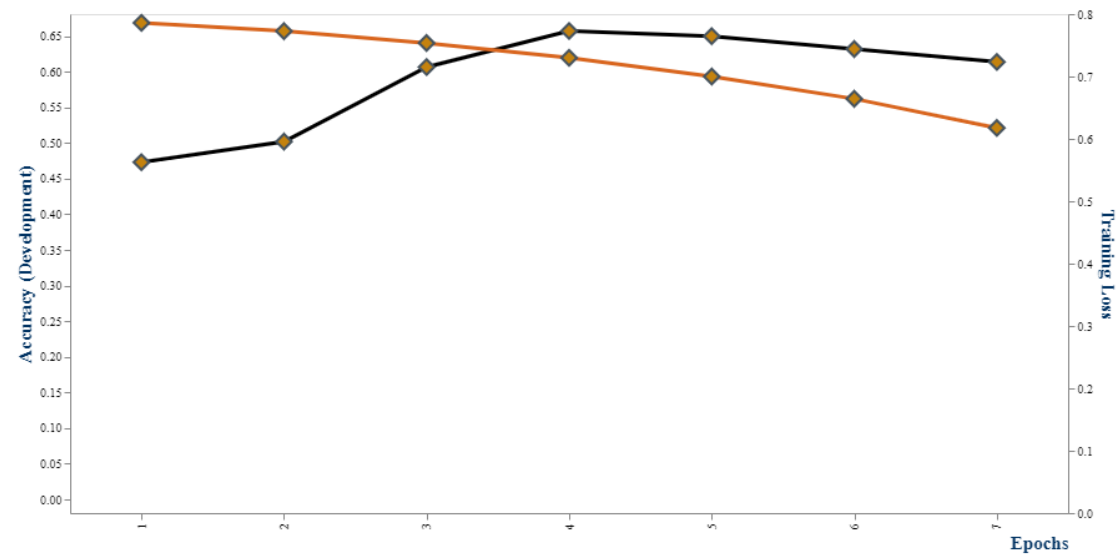
GLUE Benchmark Task: STS-B

BERT-base Performance



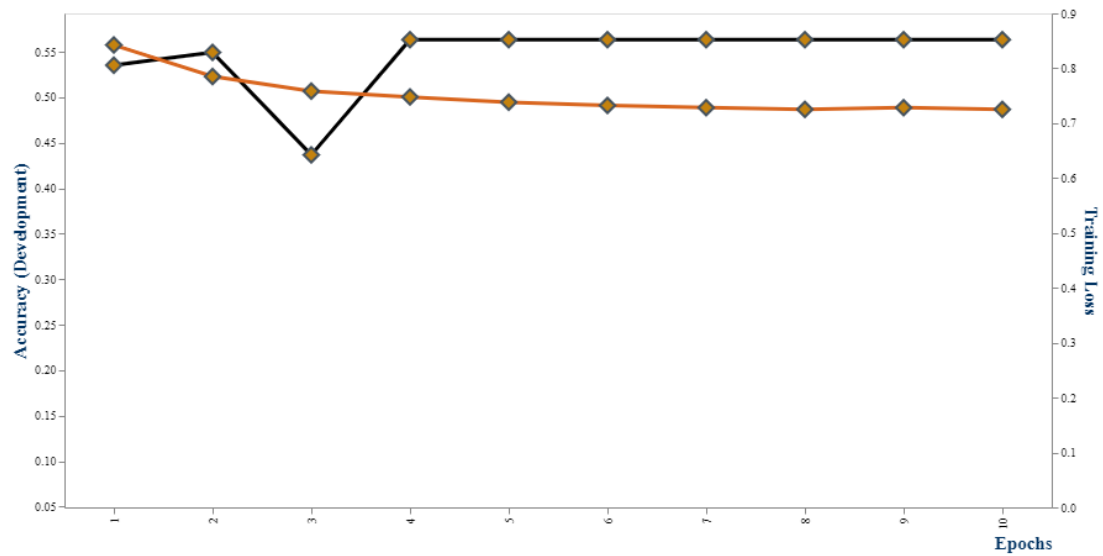
GLUE Benchmark Task: RTE

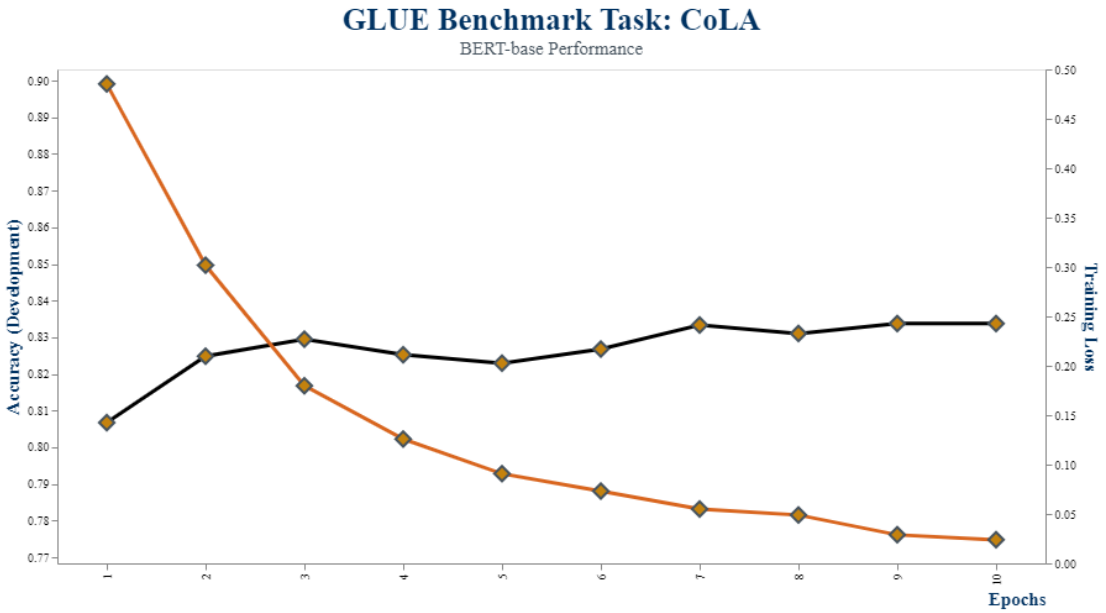
BERT-base Performance



GLUE Benchmark Task: WNLI

BERT-base Performance





	sentence_source	label	label_notes	sentence
1942	r-67	0	*	Every student who ever goes to Europe ever has...
7611	sks13	1	NaN	Is there anything to do today?
7685	sks13	0	*	John convinced the rice to be cooked by Bill.
3923	ks08	1	NaN	The birds devour the worm.
722	bc01	0	*	Home was gone by John.
130	cj99	0	*	The more John eats, the tighter keep your mout...
5387	b_73	0	*	She has problem enough as it is.
4566	ks08	0	*	The roof is leaked.
4480	ks08	0	*	John did not leaving here.
5548	b_73	1	NaN	Mary has more than two friends.

The **C**orus **o**f **L**inguistic **A**cceptability is a collection of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors. The challenge is to predict a binary classification for each sentence, identifying those that are grammatically correct (TRUE) and those that are not grammatically correct (FALSE).

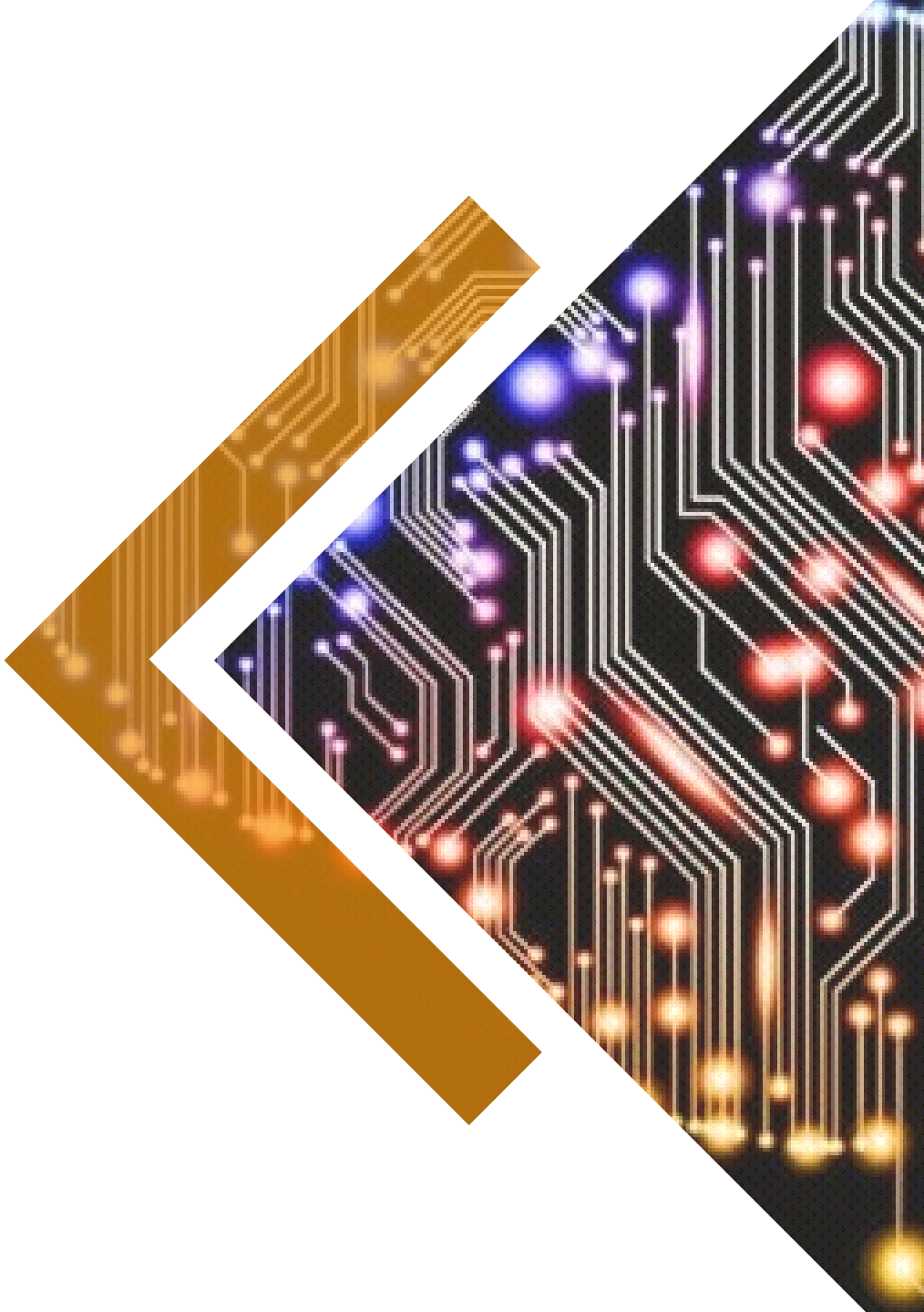
Semester Plan

We have essentially 9 weeks to complete baseline of dozens of large (340M+ parameter) models, evaluate transfer learning tasks between datasets and between task categories for BERTVision models, interpret results, and publish a high-quality paper aimed at the EMNLP 2021 conference [held in the sunny Dominican Republic on November 7th – 11th].

"The future is uncertain, and the end is always near." –

Jim Morrison, Roadhouse Blues (1970)

Our aim is to test a novel idea; the future of which is uncertain. if transfer learning is not as successful as we hope, we will need to shift focus to the value of extending BERT performance cheaply with our already demonstrated technique of embedding extraction and linear learning.

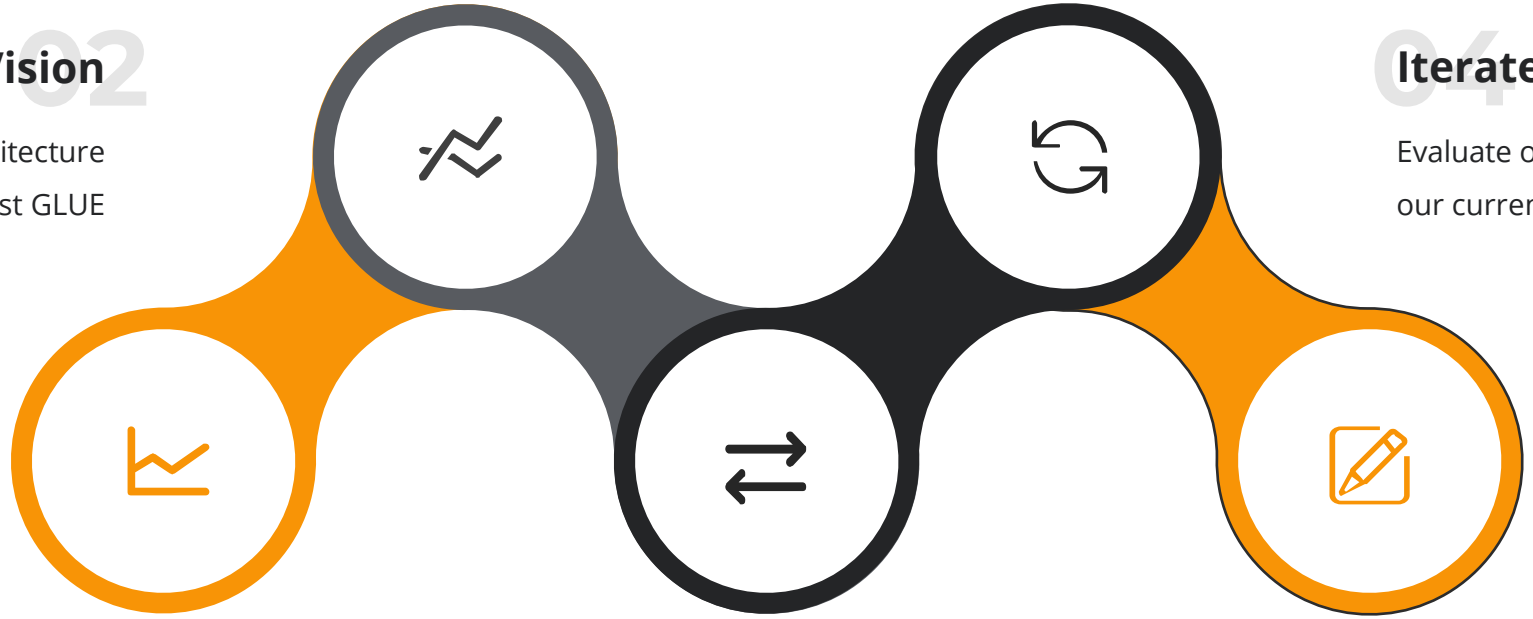


Remaining Research

02

Baseline BERTVision

Baseline the current architecture of BERTVision against GLUE



04

Iterate on Model Design

Evaluate other approaches beyond our current adapter-linear design

01

Baseline BERT-large

Capture BERT-large baseline performance against GLUE tasks

03

Transfer Learning

Evaluate BERTVision mini-models on transfer learning tasks

05

Publish

Work with Alberto, Puya, and Daniel to author a journal quality paper (targeting EMNLP 2021)

Success Evaluation

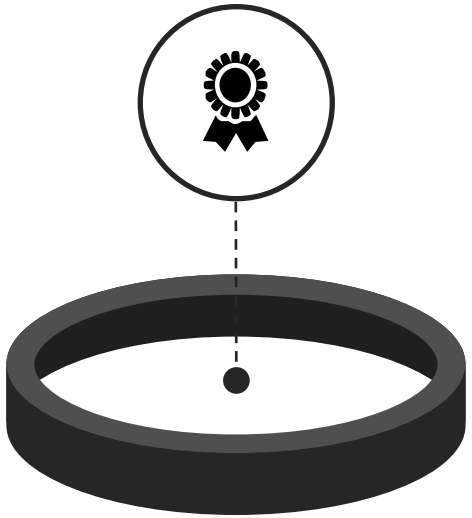
Success will depend greatly on our ability to live up to the vision of the project: to provide a novel approach to NLP tasks that requires a fraction of BERT's computation cost and time by learning from the dormant hidden state activations within.

BERTVision Models 91%

BERT-Large Models 71%

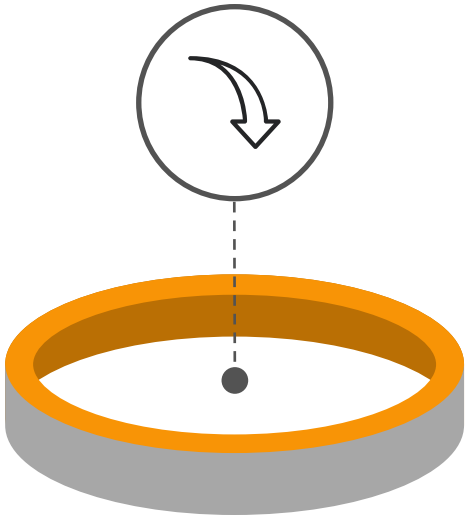
* A fictionalized representation of our goal: to beat BERT-large across a spectrum of NLP tasks at a fraction of the size and cost.

Metrics for Success



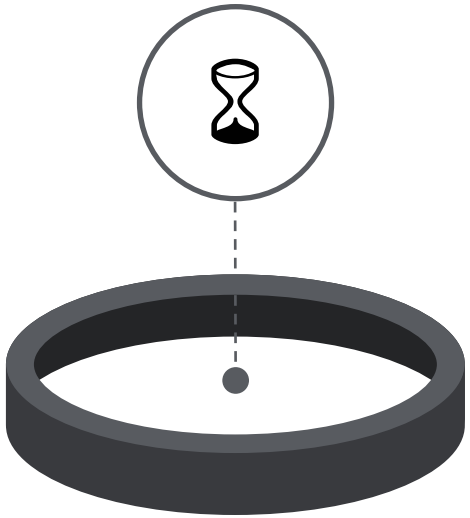
BERT v BERTVision

We will evaluate our models with BERT-large for the NLP tasks



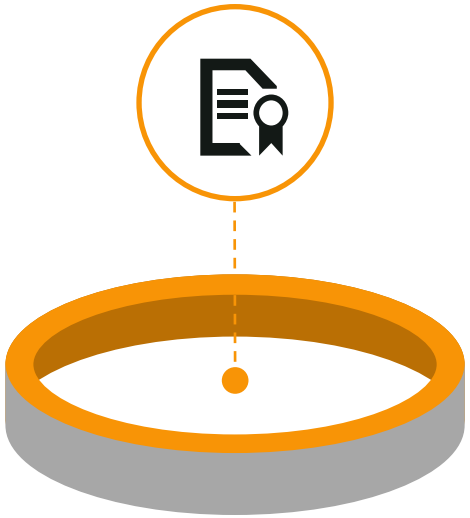
Think Small

Our success depends on shrinking the footprint of BERT



Reduced Time

Time to fine-tune and get results is a critical part of our value



Academic Contribution

Contribute to the NLP research community? Check!

THANKS

Questions?