# Smaller, Faster, Smarter? Reproducing and Evaluating DistilBERT

**Daniel Botros, Jenny Chen, Jessica Cho, Eric Gao, Xinyu Ge**

**Cornell University**

## Problem

LLMs deliver strong performance but are too slow and resource-intensive for real-time or low-power applications, creating a need for smaller, faster alternatives.


Fig. 1. LLM parameter sizes over time

## Goal

Can we shrink BERT without sacrificing performance on real tasks?

Using **knowledge distillation** we aim to train a smaller mode that is:

- lighter
- faster
- retains most of performance of larger model

Students learn the soft probabilities produced by the teacher, not just the labels which helps generalization.

**"Somewhat okay, but could be better."**
(Sample from IMDb review)

Teacher soft outputs / probability:
**Positive: 0.55** **Negative: 0.45**

Student mimics this behavior instead of learning a single label.

## Methods

**Teacher:** Google BERT base (uncased)
**Student:** Paper's DistilBERT base (uncased)
→ Remove half of BERT layers
→ Remove token-type embeddings / poolers
→ Copy every other BERT layer (initialization)
→ Pretrain for 10 epochs on wikitext (3 hours)

Student learns to imitate teacher behavior using knowledge distillation during masked language model (MLM) pretraining.

**Triple Loss (Distillation) Objective**

Combination of:

1. **MLM Loss**: learn to predict masked token
   - core language understanding
2. **KL-Div Loss**: mimic teacher soft probabilities
   - generalization
3. **Cosine Embedding Loss**: align hidden states
   - mimic internal behavior of teacher

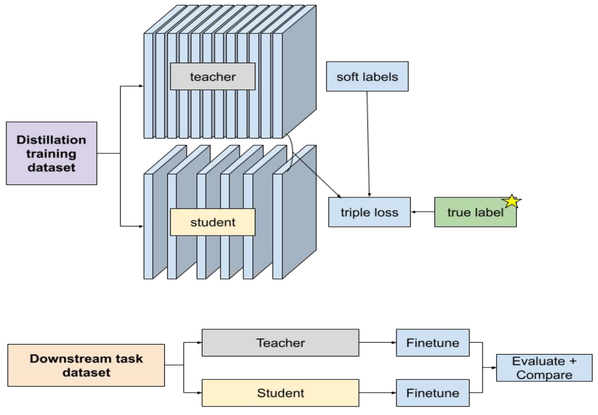


Fig. 2. Methodology Diagram

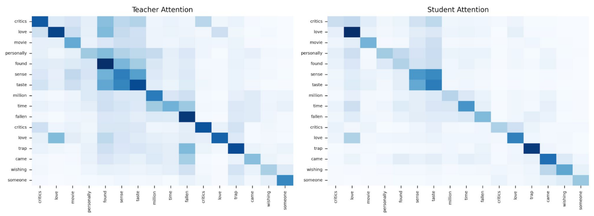


Fig. 3. Attention map for student IMDb mistake

## Results and Evaluation

Both student and teacher are fine-tuned and evaluated on the following downstream tasks:

- IMDb reviews – sentiment classification
- GLUE Benchmark tasks
  - CoLA – is a sentence grammatically valid?
  - MRPC – do two sentences mean the same thing?

These benchmarks test sentiment, grammar, and semantics reflecting core NLU capabilities.

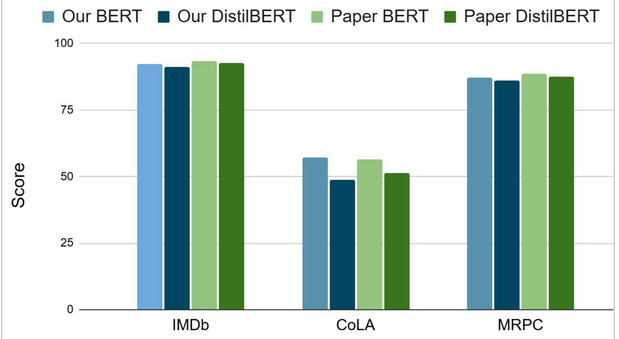


Fig. 4. Our and paper model performance across tasks

DistilBERT retained **94.35%** of the teacher's performance with nearly **half** the number of parameters (~60%) at **2.61x** the speed!

**Table 1: Benchmark Tasks**

| Model | IMDb (Accuracy) | CoLA (Matthews corr) | MRPC (Acc. / F1) |
|---|---|---|---|
| **Teacher** | **92.4** | **57.2** | **87.2** |
| | **93.5** | **56.3** | **88.6** |
| **Student** | **91.2** | **48.9** | **86.2** |
| | **92.8** | **51.3** | **87.5** |

*Blue - our results* *Green - paper results*

*94.35% Performance Retained*
*97% Performance Retained*

**Table 2: Speed Comparison**
**Average Across Test Dataset**

| Model | IMDb (ms) | CoLA (ms) | MRPc (ms) |
|---|---|---|---|
| **Teacher** | **9.5** | **9.2** | **60.1** |
| **Student** | **4.8** | **4.7** | **15.6** |

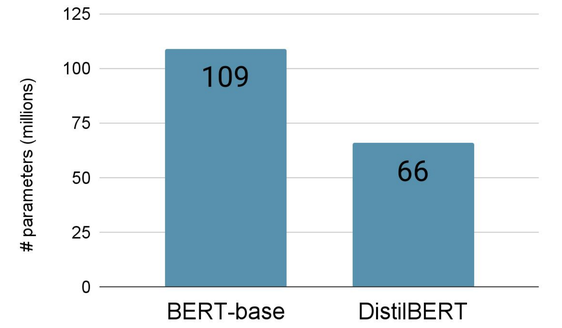


Fig. 5. Student and teacher model parameters

## Conclusions

Knowledge-distilled models are more accessible, environmentally sustainable, resource efficient, and equally as effective as large models.

## Future Work

- Experiment with removing different layers to investigate potential trade-offs and impact on performance compared to model size
- Train student using ensemble of teachers

## References

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv.org. https://arxiv.org/abs/1910.01108

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ArXiv:1804.07461 [Cs]. https://arxiv.org/abs/1804.07461