BERT Lost Patience: Won't Be Robust to Adversarial Slowdown

Zachary Coalson, Gabriel Ritter, Rakesh Bobba, Sanghyun Hong

Oregon State University

Oregon State University

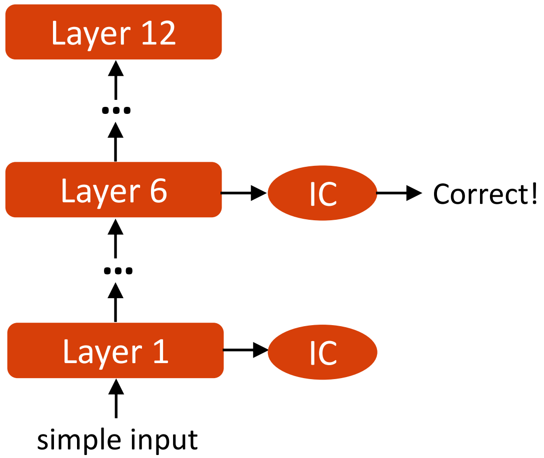
SAIL
Secure AI Systems Lab

Language Models Are Becoming Computationally Demanding

- BERT (2018)
  - 71.2% on QQP
  - 110M parameters
  - 0.05 s / 100 tokens
  - ≈ $0.52 / 1M queries
- XLM-RoBERTa XXL (2021)
  - 92.6% on QQP
  - 10.7B parameters
  - 6.71 s / 100 tokens
  - ≈ $70 / 1M queries

MULTI-EXIT LANGUAGE MODELS

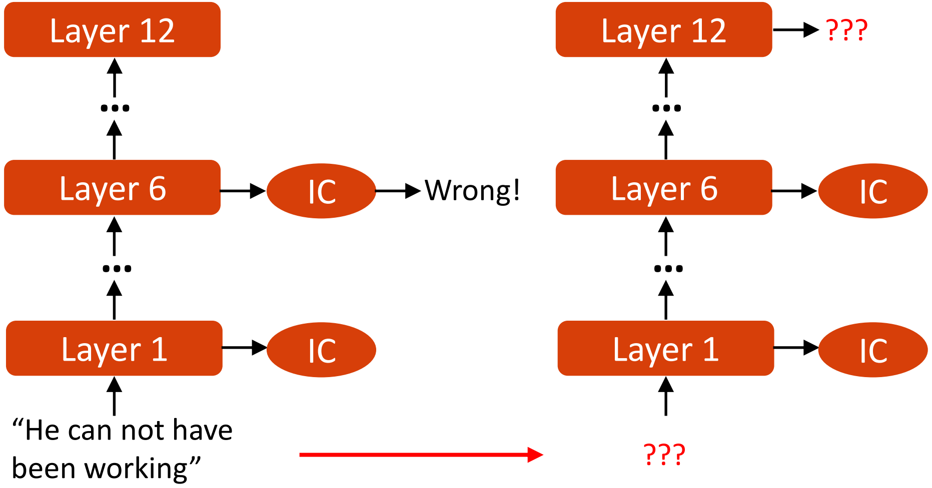
- Language models *overthink*[1]
- Multi-exit language models[2,3,4]
  - Introduce *internal classifiers (ICs)* or *early-exits* to layers
  - Enable *input-adaptive* inference
  - Provide 2-3x computational savings without accuracy loss



[1]Kaya et al., Shallow-Deep Network: Understanding and Mitigating Network Overthinking, ICML 2019
[2]Zhou et al., *BERT Loses Patience: Fast and Robust Inference with Early Exit*, NeurIPS 2020
[3]Xin et al., *DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference*, ACL 2020
[4]Liao et al., *A Global Past-Future Early Exit Method for Accelerating Inference of Pre-trained Language Models*, ACL 2021

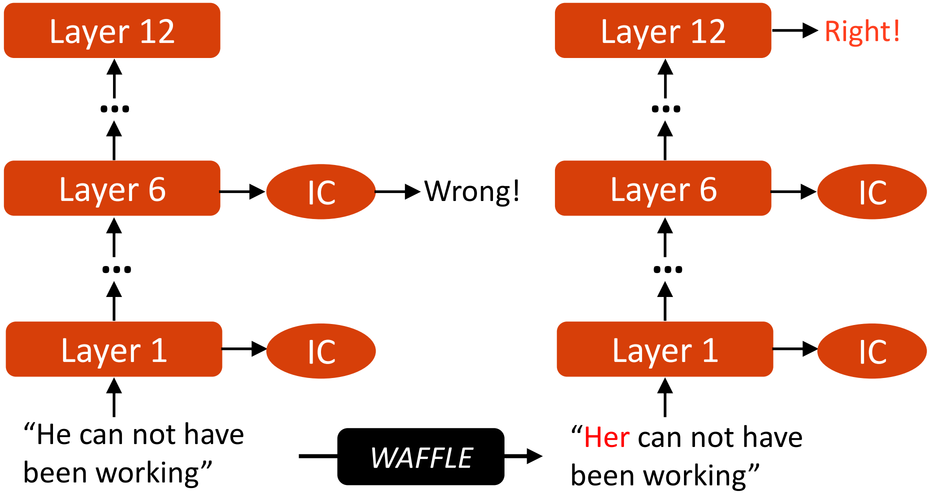
Adversarial Implications of Multi-exit Language Models

- Research questions:
  - How robust are the computational savings to adversarial input perturbations?
  - What factors attributed to the vulnerability?
  - How can we defend against adversarial slowdown?



Our Method for Auditing the Vulnerability to Slowdown

- WAFFLE attack
  - Performs word-level input perturbations
  - Our slowdown objective
    - Pushes IC outputs toward uniform distribution
  - Implemented on existing adversarial text attack framework[1,2]



[1]Jin et al., Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment, AAAI 2020
[2]Yoo et al., Towards Improving Adversarial Training of NLP Models, EMNLP 2021

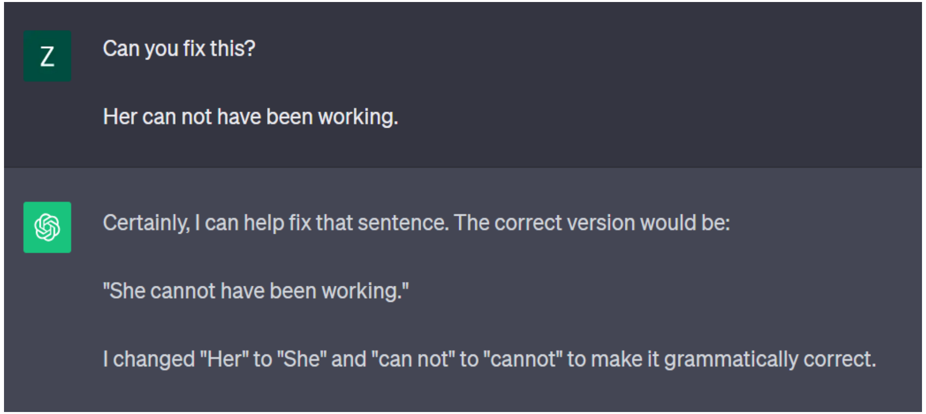
Multi-Exit Models are Not Robust to Adversarial Slowdown

- Able to induce high slowdown in three multi-exit language models[1,2,3]
  - GLUE benchmark
  - 70% average *efficacy* reduction
- More complex mechanisms are more vulnerable
- High transferability
  - Cross-seed: 33% efficacy reduction
  - Cross-mechanism: 21% efficacy reduction
- Linguistic analysis:
  - High perturbation count ≠ effective
  - Subject-predicate disagreement and changed named entities

[1]Zhou et al., *BERT Loses Patience: Fast and Robust Inference with Early Exit*, NeurIPS 2020
[2]Xin et al., *DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference*, ACL 2020
[3]Liao et al., *A Global Past-Future Early Exit Method for Accelerating Inference of Pre-trained Language Models*, ACL 2021

Input Sanitization is an Effective Countermeasure

- Adversarial training is not a countermeasure
  - No exit layer reduction or accuracy recovery
- LLMs can be used as input sanitizers
  - 91% and 375% increase in efficacy
  - 12% and 24% points of accuracy recovered
  - Computationally intensive



THANK YOU!

Zachary Coalson
E-mail: coalsonz@oregonstate.edu
Code: github.com/ztcoalson/WAFFLE

See You All at Our Poster Session!

Great Hall & Hall B1+B2 #1703 @ 3PM Wed

Oregon State University

SAIL
Secure AI Systems Lab