RESEARCHERS AUDIT THE ROBUSTNESS OF MULTI-EXIT MODELS TO ADVERSARIAL SLOWDOWN

- Neural network language models "**overthink**[1]": they use more layers than necessary for a correct classification. **Multi-exit** language models counteract overthinking by introducing internal classifiers that allow the model to stop inference early if it is confident in its answer.
- An increasing amount of research has explored multi-exit mechanisms for large language models[2,3,4]. Prior work has found that multi-exit mechanisms can provide **2-3x speed-up with no accuracy loss**.
- With the introduction of these computational savings, a new threat arises—**adversarial slowdown**, which involves **perturbing** (changing the words of) a model input with the intent of slowing down a multi-exit model. This threat is analogous to a **denial-of-service-attack**, as an attacker would be able to greatly reduce the availability of the model and increase the costs associated with deploying it.
- In this work, we answer the following research questions:

1. How robust are the computational savings of multi-exit models to adversarial input perturbations?
2. What factors contribute to this vulnerability?
3. How can we defend these models against adversarial slowdown?

- Our results suggest that future work is necessary for developing efficient yet robust multi-exit models.

Oregon State University

LANGUAGE MODELS ARE VULNERABLE TO SLOWDOWN

This research explores the robustness of multi-exit language models to adversarial slowdown.

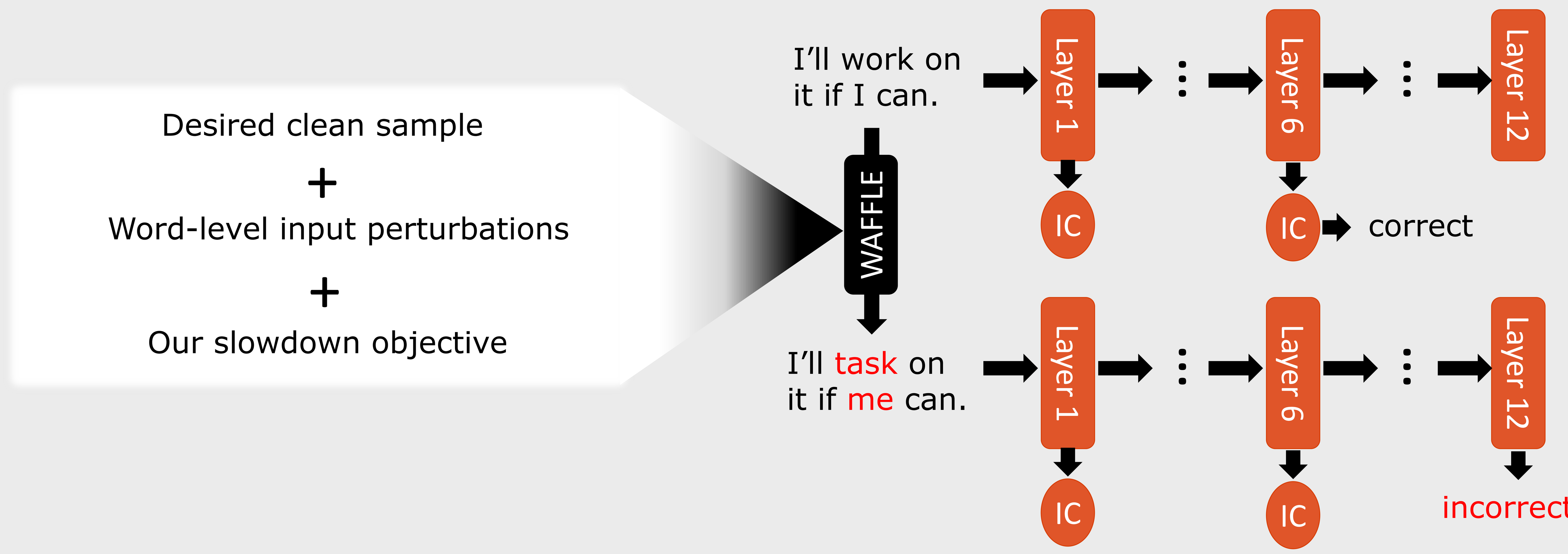

Figure 1. Overview of our attack.

ATTACK RESULTS

- The table below shows slowdown results for the strongest multi-exit mechanism we tested (PastFuture[4]) on two datasets. Acc. is accuracy and Eff. is efficacy, a metric proportional to speed-up. TF[6] refers to the attack algorithm we used.

Attack	RTE		MRPC	
	Acc.	Eff.	Acc.	Eff.
Clean	71%	0.52	88%	0.50
TF[6] (Base)	41%	0.46	36%	0.24
TF[6] (Ours)	51%	0.17	42%	0.15

Table 1. Slowdown results.

- On the GLUE[5] benchmark, our attack **reduces average efficacy (speed-up) by 70%** on three multi-exit models.
- **More complex mechanisms are more vulnerable**, meaning this problem cannot be solved by making better models.
- Our attack is transferable, meaning we can craft adversarial examples on one model and use them on another. This makes it a **practical threat for deployed multi-exit models**.

LINGUISTIC ANALYSIS

- Perturbation count is **not correlated** to magnitude of slowdown.
- There is a high prevalence of **subject-predicate disagreement** and **changed named entities**, which suggests that language models can be "confused" in the same way humans are.

POTENTIAL COUNTERMEASURES

- **Adversarial training**, a common defense, **negates computational savings** on clean samples and **recovers no efficacy** on perturbed samples.
- **Input sanitization** via large language models (e.g. ChatGPT[7]) **greatly recovers accuracy and efficacy** and is a potential future direction.

COLLABORATORS

This work was done with fellow Oregon State University researchers.

Zachary Coalson

coalsonz@oregonstate.edu

Gabriel Ritter

Dr. Rakesh Bobba

Dr. Sanghyun Hong

REFERENCES

1. Hong et al., A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference, ICLR 2021
2. Xin et al., DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference, ACL 2020
3. Zhou et al., BERT Loses Patience: Fast and Robust Inference with Early Exit, NeurIPS 2020
4. Liao et al., A Global Past-Future Early Exit Method for Accelerating Inference of Pre-trained Language Models, ACL 2021
5. Wang et al., GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, ACL 2018
6. Jin et al., Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment, AAAI 2020
7. https://chat.openai.com/