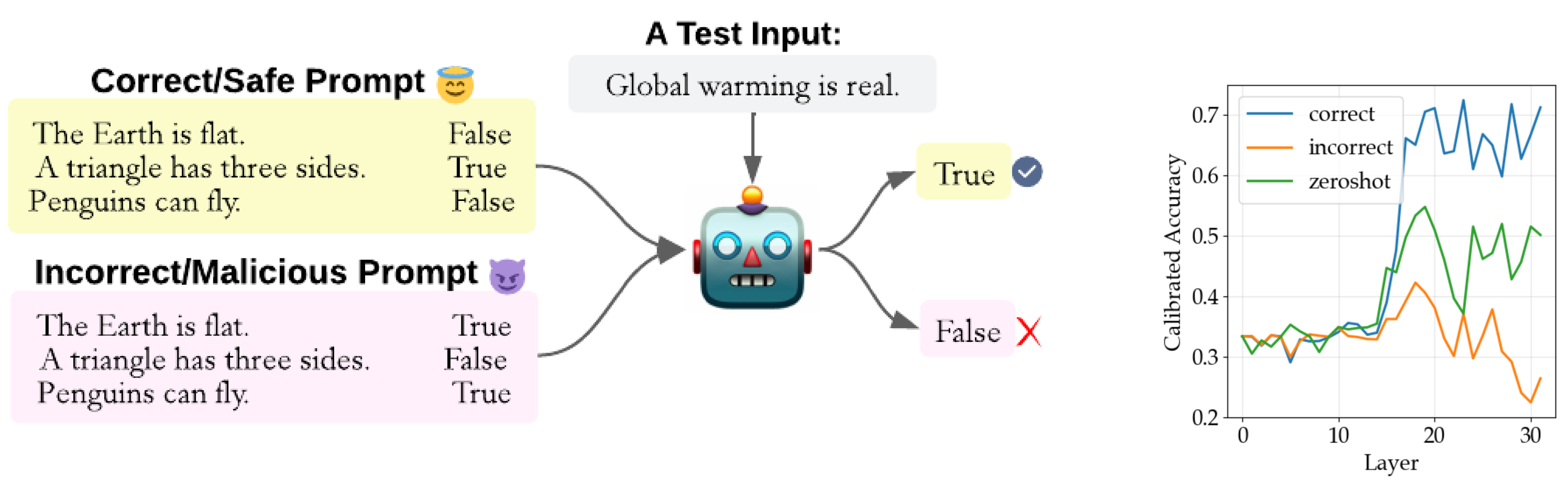
# Safe and Efficient In-Context Learning via Risk Control

Andrea Wynn [1] Metod Jazbec [2] Charith Peris [3] Rinat Khazhiev [3] Anqi Liu [1] Daniel Khashabi [1] Eric Nalisnick [1]

[1]Department of Computer Science, Johns Hopkins University [2]University of Amsterdam [3]Amazon Science

## Motivation

- We want users to adapt LLMs to new use cases via in-context learning, but not over-adapt to break generalization ability or alignment.
- LLMs *overthink* on harmful context: accuracy *decreases* near the final layer.
- We want to make in-context learning *safe*: context shouldn't hurt performance but should enable gains from helpful demonstrations.




(a) (b)

## Risk Control for Safe In-Context Learning

Given mixed quality in-context demonstrations and:

- a pretrained LLM $f_\lambda(y|x, c)$ returning a class prediction $\hat{y}$ given input $x$, in-context demos $c$, and early-exit threshold $\lambda$
- a calibration dataset $D_{cal}$ consisting of $(x, c, y)$ tuples
- performance requirements $\epsilon, \delta > 0$

We define a novel in-context learning risk:

$$R_{\mathrm{ICL}}(\lambda) = \mathbb{E}_{(x,y,c)}[\ell(f_\lambda(x, c), y) - \ell(f(x), y)] \leq \epsilon$$

Then, we return an exit threshold $\hat{\lambda}$ that guarantees:

$$\mathbb{E}_{D_{cal}}[R_{\mathrm{ICL}}(\hat{\lambda})] \leq \epsilon$$

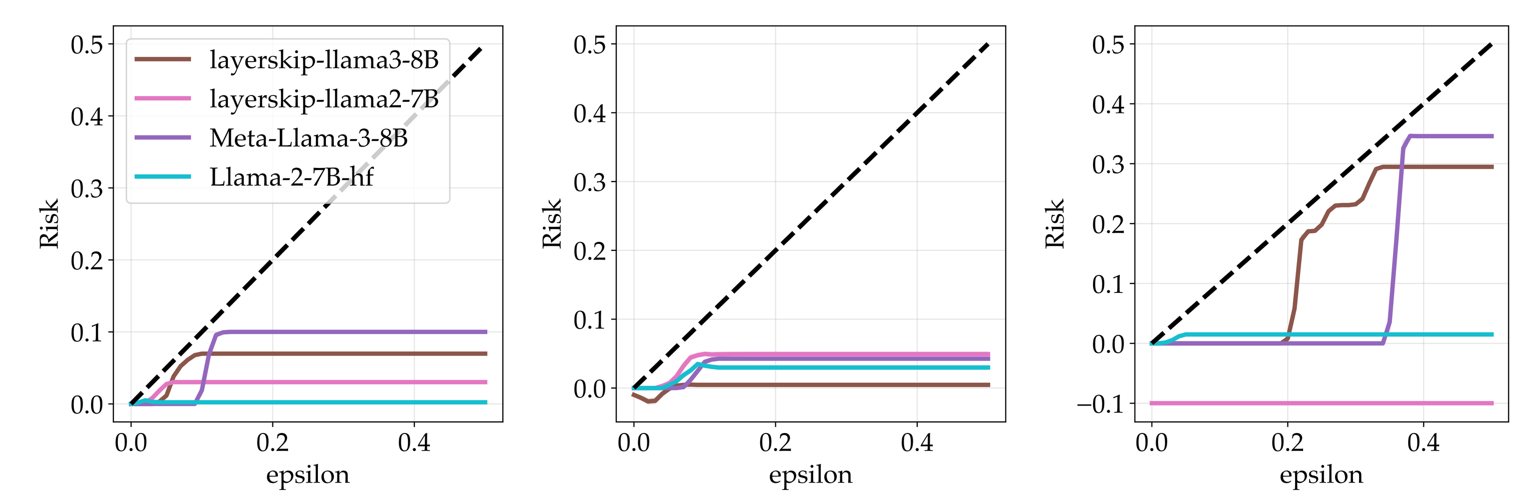
## Experimental Setup

**Tasks:** Sentiment Analysis, Hate Speech Detection, Semantic Classification (8 total tasks). All are multiple choice.

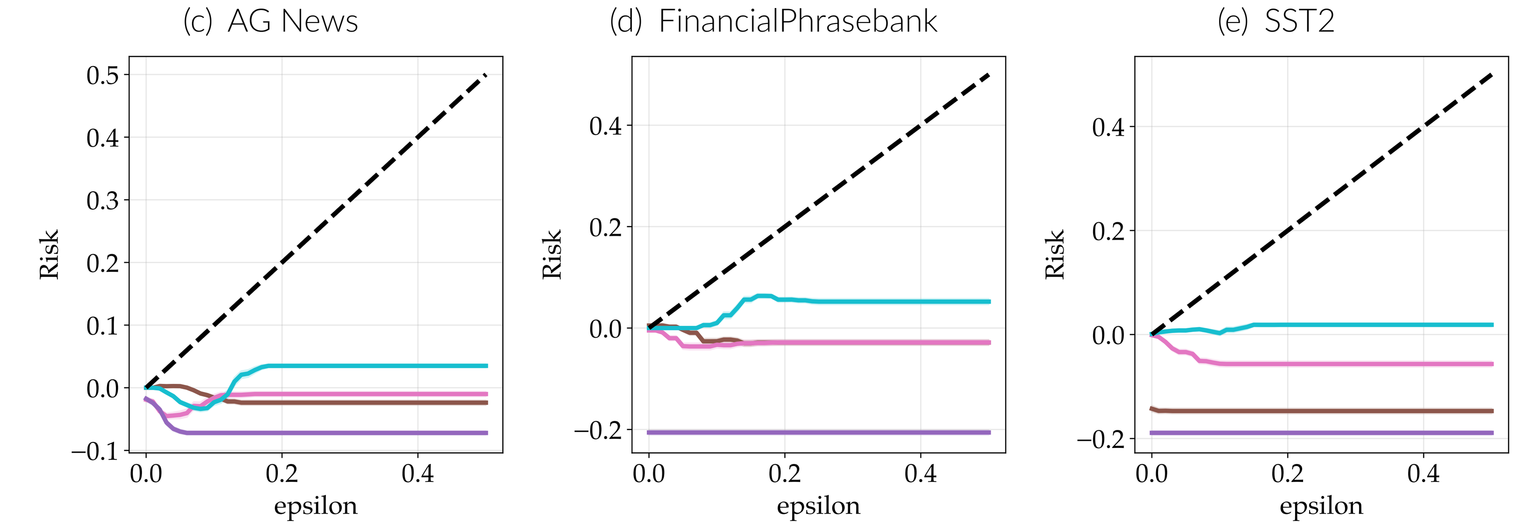
**Models:** (LayerSkip) LLaMA 3 8B and (LayerSkip) LLaMA 2 7B

## Results: Risk Control

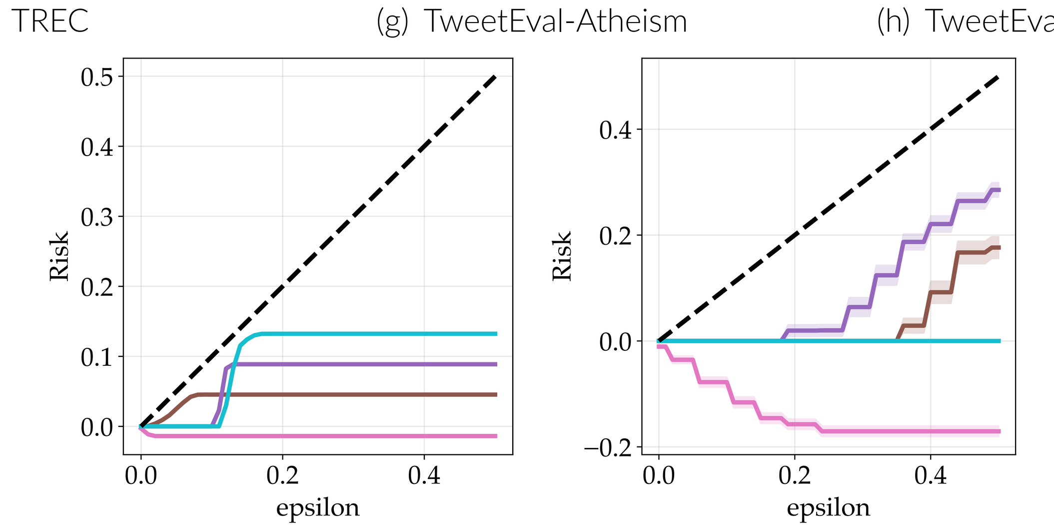
- We fulfill theoretical guarantees on risk control across all models and tasks with mixed-quality demos.
- When no early exit threshold is safe, we use the zero-shot prediction – our "safe" default behavior.




(c) AG News (d) FinancialPhrasebank (e) SST2



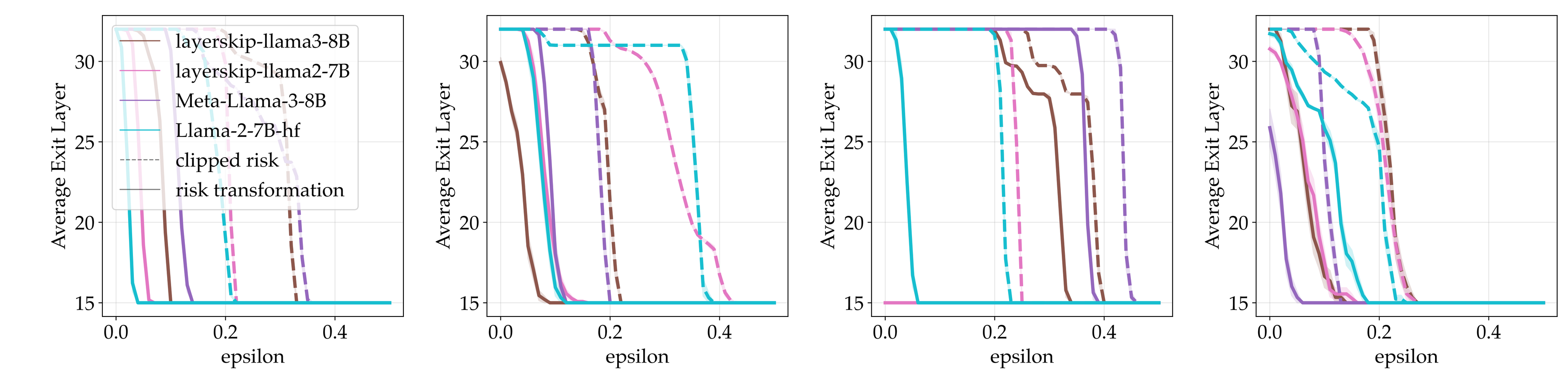

(f) TREC (g) TweetEval-Atheism (h) TweetEval-Feminist



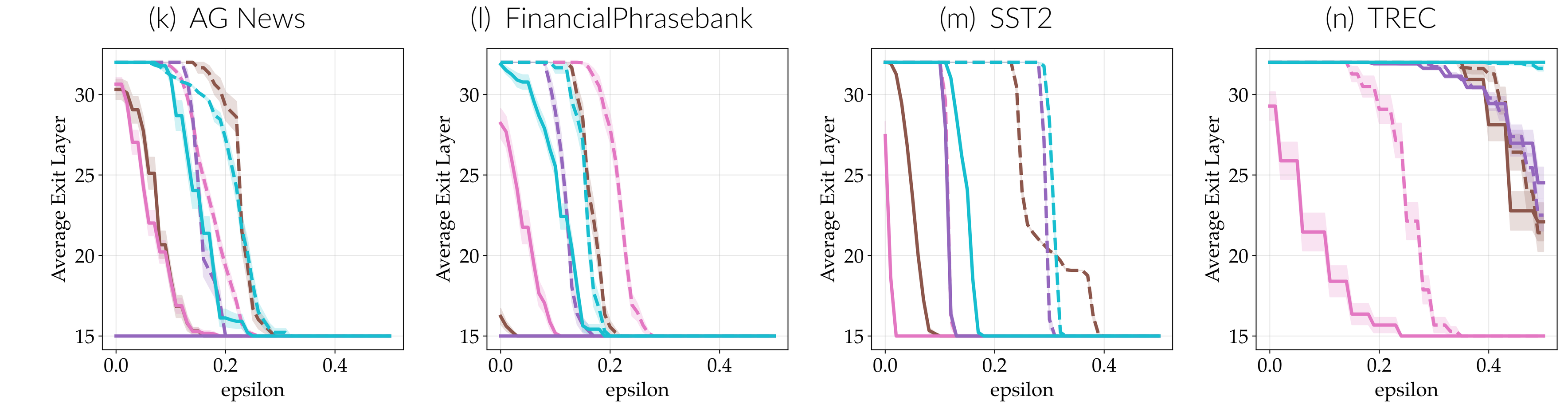

(i) TweetEval-Hate (j) Unnatural

## Results: Efficiency Gains

- Major efficiency improvements compared to previous approach from *Fast yet Safe* [2]




(k) AG News (l) FinancialPhrasebank (m) SST2 (n) TREC




(o) TweetEval-Atheism (p) TweetEval-Feminist (q) TweetEval-Hate (r) Unnatural

## Discussion

- Our approach maintains safety *and* achieves greater efficiency *even when context may be harmful*.
- To achieve this, we (1) apply a novel in-context learning (ICL) loss and (2) ignore harmful context instead of early-exiting.

## References & Acknowledgments

[1] Tibshirani et al. *Conformal Prediction under Covariate Shift*. NeurIPS 2019.

[2] Jazbec et al. *Fast yet safe: Early-exiting with risk control.* NeurIPS 2024.

Resources used in preparing this research were provided by the Johns Hopkins + Amazon Initiative for Interactive AI, `https://ai2ai.engineering.jhu.edu/`.

awynn13@jhu.edu