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

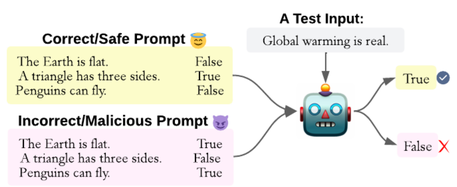
Models sometimes receive harmful inputs during in-context learning, which can make them *overthink*: performing worse than if they stopped early or ignored the examples.

We want to ensure in-context learning is safe: examples shouldn't hurt zero-shot performance but should enable gains from helpful demonstrations.

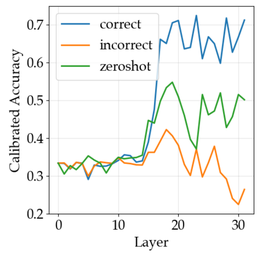
We can do this via early-exit and get major computational savings as a side benefit.

Problem: Overthinking

- LLMs *overthink* on incorrect in-context demonstrations: after some intermediate layer, the accuracy *decreases* as we progress to the final layer.
- Previous risk control approaches assume the context is helpful; thus, these approaches cannot handle overthinking.



(a)



(b)

Solution: Risk Control for Safe In-Context Learning

We adapt previous risk control methods for in-context learning with mixed quality demonstrations. Given:

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

We define a novel in-context learning risk:

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

Then, we return an exit threshold $\hat{\lambda}$ with the following guarantee:

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

Experimental Setup

Tasks:

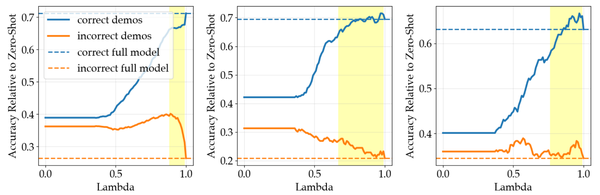
- **Sentiment Analysis:** SST2, FinancialPhrasebank, TweetEval-Feminist
- **Hate Speech Detection:** TweetEval-Hate, TweetEval-Atheism
- **Semantic Classification:** AG News, Text REtrieval Conference (TREC), Unnatural

Models: LLaMA 3 8B and LLaMA 2 7B; LayerSkip LLaMA 3 8B and LLaMA 2 7B

In-Context Demos: Mix of between 5%-95% correct demos, with the rest incorrect.

Selecting an Appropriate Threshold

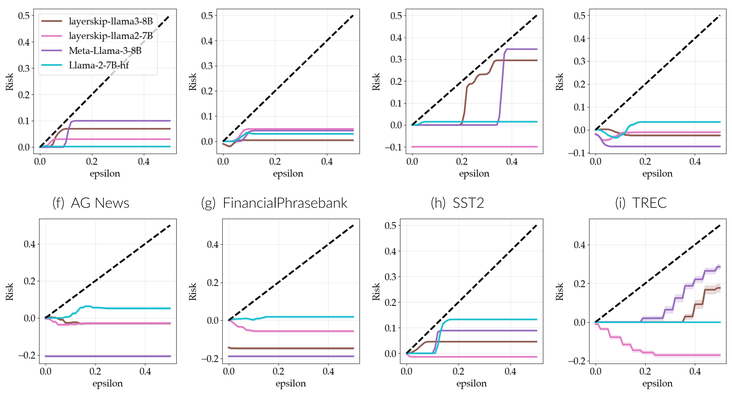
- When we have mixed-quality demonstrations, can we control risk for harmful demonstrations while still getting performance gains from helpful demonstrations?
- Sometimes, yes! Highlighted regions show where **we lose $\leq 5\%$ of accuracy from correct demos while outperforming the full model given incorrect demos.**



(c) Financial (d) TE-Atheism (e) TE-Feminist

Results: Risk Control

When risk cannot be controlled using our early-exit risk control alone, we replace the model's output with the zero-shot full model - our "safe" default behavior. Aligning with the theory, our approach controls risk across all models and tasks.

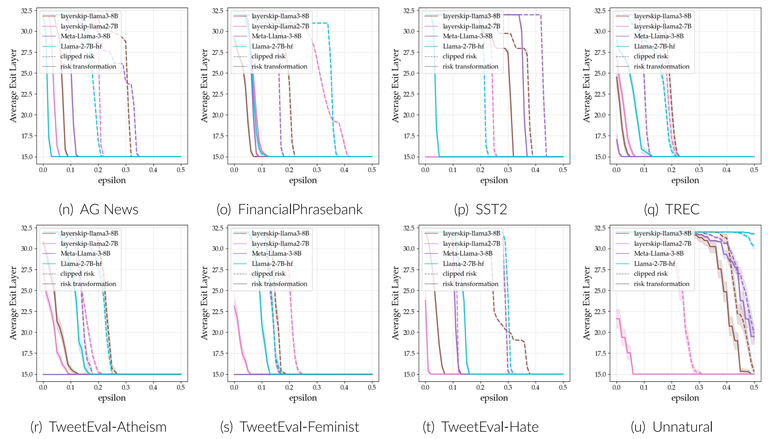


(f) AG News (g) FinancialPhrasebank (h) SST2 (i) TREC

Results: Efficiency Gains

We achieve much greater efficiency gains than the previous "clipped loss" approach from *Fast yet Safe* [2] by leveraging performance gains from correct demonstrations.

Dotted lines correspond to the "clipped loss" approach; solid lines correspond to our approach. Efficiency gain = number of layers skipped at test time.



(n) AG News (o) FinancialPhrasebank (p) SST2 (q) TREC

(r) TweetEval-Atheism (s) TweetEval-Feminist (t) TweetEval-Hate (u) Unnatural

Discussion

- We introduce a novel approach with formal safety guarantees *even when the model sometimes receives harmful inputs*, enabling us to robustly control the degree to which users can adapt model behavior, on average.
- We introduce a novel in-context learning (ICL) loss, enable ignoring harmful context, and propose a domain scaling trick to control risk relative to the *zero-shot* model, rather than the full model (which may be unsafe).
- We perform a total of 128 experiments across various models and datasets demonstrating that we achieve robust safety guarantees with better performance and efficiency than prior approaches.

References

[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.

Acknowledgments

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