SAFE AND EFFICIENT IN-CONTEXT LEARNING VIA RISK CONTROL

Anonymous authors
Paper under double-blind review

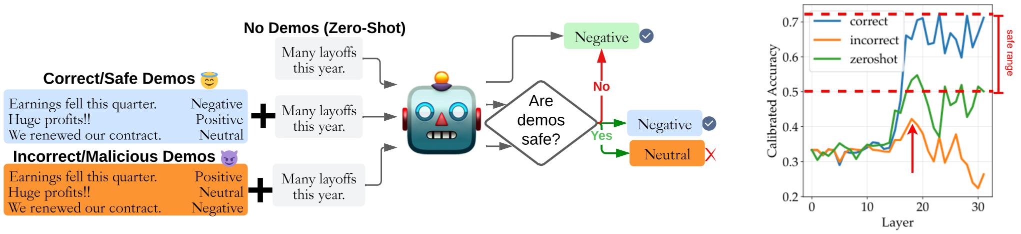
ABSTRACT

Large language models (LLMs) demonstrate a remarkable ability to learn new tasks from a few in-context examples. However, this flexibility introduces safety concerns: LLMs can be influenced by incorrect or malicious demonstrations – for example, if an adversary tampers with or injects harmful examples without a human supervisor noticing. This motivates principled designs in which the system itself includes built-in mechanisms to guard against such attacks. We propose a novel approach to limit the degree to which harmful demonstrations can degrade model performance. First, we define a baseline "safe" behavior for the model – the model's performance given no in-context demonstrations (zero-shot). Next, we apply distribution-free risk control (DFRC) to control the extent to which in-context samples can decay performance below zero-shot. We achieve this by leveraging dynamic early exit prediction, ignoring later attention heads that attend the most to the unsafe inputs. Finally, we propose modifications to DFRC that allow it to both control risk for harmful inputs *and* leverage performance and efficiency gains on helpful inputs. We present both theoretical and empirical results showing that our approach can effectively control risk for harmful in-context demonstrations while simultaneously achieving substantial computational efficiency gains with helpful demonstrations.

1 INTRODUCTION

Large language models (LLMs) have shown an impressive ability to be adapted to a wide variety of tasks through methods such as prompt tuning and in-context learning, many of which require only minimal data and do not require expensive fine-tuning. Yet this adaptability introduces safety concerns: incorrect, adversarial, or otherwise harmful demonstrations can degrade performance or elicit unsafe outputs. Imagine an LLM deployed for a specific use case and adapted through in-context demonstrations; these demonstrations may be misleading for a number of reasons – such as unintended user error or intentional tampering by an adversary (such as many-shot jailbreaking (Anil et al., 2024) and prompt injections (Liu et al., 2024; Das et al., 2024)). For instance, consider a LLM coding agent prompted with example API calls. A developer may mistakenly provide hard-coded credentials, accidentally teaching the model to replicate insecure code; an adversary could insert a demonstration where user validation is bypassed, introducing security vulnerabilities. Such vulnerabilities could escape the notice of a human system designer, motivating us to develop built-in safeguards so that the LLM defaults to disregarding these compromised demonstrations.

In this paper, we apply distribution-free risk control (DFRC) to mitigate the influence of corrupted in-context examples by comparing the loss of the adapted model to the default zero-shot model. LLMs' zero-shot performance on a wide variety of tasks has become quite strong in recent years and continues to improve (Kojima et al., 2023). This makes zero-shot LLMs comparatively well-understood and predictable, whereas ICL models on arbitrary user-supplied demonstrations may reflect uncontrolled or adversarial distribution shifts. Using the zero-shot model as a baseline also anchors risk control in a setting that has undergone extensive pre-deployment safety testing (Zhang et al., 2024; Yuan et al., 2025), unlike the highly variable in-context examples that arise during deployment. Recent work has shown that depth controls how much an LLM can learn from in-context examples; LLMs "overthink" on harmful examples, meaning their performance peaks at some intermediate layer and drops in deeper (later) layers (Halawi et al., 2024). We replicate these results on our tasks; an example is shown in Fig. 1(b), where the model does well given correct demonstra-



(a)

(b)

Figure 1: (a) A LLM is given in-context demonstrations of unknown quality (helpful or harmful). The model needs to infer whether to rely on the answer it obtains using the given demonstrations. If not, it falls back to the answer it would give without seeing any demonstrations at all (zero-shot). (b) When given incorrect demonstrations, it is better to either early-exit (at the layer indicated by the red arrow) or simply not use the given demonstrations than to use the model's final prediction – staying in the "safe" performance range between zero-shot and correct demonstrations. Details on early-exit LLMs are provided in §2.1.

tions, reasonably well given no demonstrations (zero-shot), and very poorly given incorrect demonstrations. Inspired by this, we implement early-exiting as a natural mechanism for applying this risk control.

In summary, our contributions are as follows: We propose three important contributions to enable safety for ICL: a novel formulation of early-exit models for safety using the safe zero-shot baseline (§3.1), a novel ICL loss designed to measure overthinking (§3.3), and a simple adaptation of the Learn-then-Test (LTT) risk control framework that balances safety with efficiency gains (§3.4). We show, via extensive experiments, that our approach can effectively prevent overthinking while still allowing the LLM to benefit from helpful demonstrations, providing robust safety guarantees even with mixed-quality inputs (§4.1) and enabling major computational efficiency gains compared to prior approaches (§4.2). Experiments across 8 diverse benchmark tasks and 4 distinct models show that our framework is able to guarantee safety on model outputs relative to zero-shot, while simultaneously achieving a greater than 50% speedup in comparison to previous approaches. Overall, to the best of our knowledge, this is the first work to establish a principled framework that controls the risk of harmful in-context demonstrations, while simultaneously leveraging dynamic early exit mechanisms to achieve performance and computational efficiency gains with helpful demonstrations.

2 Preliminaries

Data Let $x \in \mathcal{X}$ denote an input text (e.g., a question), and let $y \in \mathcal{Y}$ be its associated label (e.g., a choice from a predetermined set of possible answers). Given our focus on classification, the label space is defined as $\mathcal{Y} = \{1, \ldots, K\}$, with K representing the number of possible answers. Moreover, we denote a *context* set of N_c demonstrations as $c = \{(x_i, y_i)\}_{i=1}^{N_c}$. Lastly, a data-generating distribution over $\mathcal{X} \times \mathcal{Y}$ is denoted with $\mathcal{P}$.

Model We denote by $p(\cdot|x, c)$ an LLM that takes a given input x together with the context set c, and outputs a probability distribution over possible labels $k \in \mathcal{Y}$. We sample both the input x and the context c from the same dataset, but explicitly prevent any overlap between x and the elements of c for a given prompt. By excluding x from c, we can prevent the model "copying" answers from the context rather than demonstrating meaningful in-context learning behavior, while still drawing both from the same distribution to avoid any bias in context/input dataset selection.

2.1 Early-Exit Language Models for Classification Tasks

Traditionally, LLMs pass through all L layers of the model before making a prediction. In contrast, early-exit LLMs (Elbayad et al., 2020; Schuster et al., 2022) offer the option to yield a prediction after each layer. This is achieved by passing the current hidden representation through an "unembed-

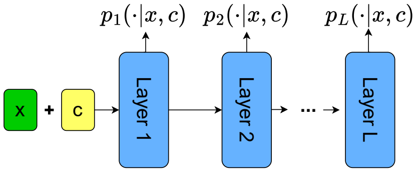
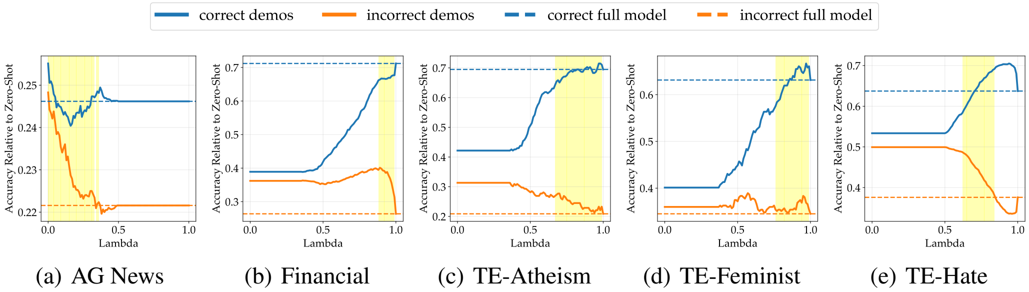


Figure 2: With an early-exit LLM, we execute every layer until our confidence exceeds the λ threshold, after which we directly make a prediction from the intermediate layer.

$$\hat{y}_\lambda := \begin{cases} \arg\max_{k\in\mathcal{Y}} p_1\left(k|x,c\right) & \text{if } \mathfrak{C}_1 \geq \lambda, \\ \arg\max_{k\in\mathcal{Y}} p_2\left(k|x,c\right) & \text{else if } \mathfrak{C}_2 \geq \lambda, \\ \vdots & \vdots \\ \arg\max_{k\in\mathcal{Y}} p_L\left(k|x,c\right) & \text{otherwise.} \end{cases} \tag{1}$$



(a) AG News (b) Financial (c) TE-Atheism (d) TE-Feminist (e) TE-Hate

Figure 3: Some choices of λ thresholds can both attain performance gains from correct demonstrations *and* control overthinking from incorrect demonstrations. The highlighted regions show where λ values exist such that **we lose no more than 5% of the accuracy gains from correct demonstrations while still doing better than the full model given incorrect demonstrations**.

ding” matrix that maps the hidden state to the vocabulary, pruned to the finite set of possible labels for a particular task. Specifically, for each layer $l \in \{1, \dots, L\}$ where L represents the total number of layers in the model, a confidence score $\mathfrak{C}_l \in [0, 1]$ and an exit threshold $\lambda \in [0, 1]$ are defined. An early prediction is returned as soon as the confidence at the current layer exceeds the threshold, with each condition evaluated in the order of the layers in the original model. This is outlined in Eq. 1 and illustrated in Fig. 2.

Here, p_l denotes the LLM's predictive distribution at the l-th layer. While various choices of confidence scores are possible, we use a simple one derived from the maximum class probability: $\mathfrak{C}_l := \max_{k\in\mathcal{Y}} p_l(k \mid x, c)$. This choice of confidence measure is common in prior work (Schuster et al., 2022), and we provide additional justification of this choice (compared to alternative confidence scores) through ablation studies in §J.2.

2.2 Controlling Predictive Risk via Distribution-Free Risk Control

Risk control frameworks (Angelopoulos et al., 2021; Bates et al., 2021a) enable principled selection of thresholds $\lambda \in \Lambda$ across various machine learning problems, ranging from conformal prediction (Angelopoulos et al., 2023) to adaptive inference (Schuster et al., 2022; Jazbec et al., 2024). Concretely, first a problem-specific *loss* function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is defined. The *risk* associated with a candidate threshold λ is then defined as the expected loss

$$R(\lambda) := \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell(p_\lambda(x), y)\right]$$

where p_λ denotes a threshold-dependent predictor (e.g., an early-exit LLM, see Eq. 1). The goal is to leverage a calibration dataset $\mathcal{D}_{\text{cal}} \sim \mathcal{P}^{N_{\text{cal}}}$ to find a threshold $\hat{\lambda}$ such that the risk is guaranteed to be small – i.e., $R(\hat{\lambda}) \leq \epsilon$ for some $\epsilon > 0$ – on new test points, which are assumed to be independently and identically distributed, or *iid*, with the samples from $\mathcal{D}_{\text{cal}}$.

3 In-Context Learning Risk Control via Early-Exit

In this section, we detail our approach to mitigating overthinking for in-context learning, i.e., preventing the LLM from picking up on harmful demonstrations by combining early-exiting with risk control. We note that across all our tasks and datasets, correct in-context demonstrations improve performance above zero-shot, and incorrect demonstrations harm performance to worse than zero-shot; so we use the terms “correct” vs “helpful” and “incorrect” vs “harmful” interchangeably.

We begin by introducing a novel formulation of early-exit models for safety by using the safe zero-shot baseline (§3.1). We then show that applying our early-exit approach can effectively prevent overthinking while still allowing the LLM to benefit from helpful demonstrations, provided an appropriate exit threshold λ is chosen (§3.2). Next, we propose a novel ICL loss designed to measure overthinking on which we can apply risk control to ensure safety on new test points (§3.3). Finally, we introduce a simple adaptation of the Learn-then-Test (LTT) framework to accommodate losses that may take on negative values, as is often the case in ICL (§3.4).

3.1 Safe In-Context Learning Predictor

We begin with our base in-context learning model p given input x and in-context demonstrations c without early-exit, $p(\cdot|x,c)$. This model can “overthink” (overfit to potentially harmful demonstrations c), significantly degrading the output over the zero-shot model, $p(\cdot|x)$. We propose to define a new safe in-context learning model $\bar{p}_\lambda(\cdot|x,c)$ by augmenting this base model for in-context learning in the following two ways: (i) enable the model to make predictions from intermediate exits given a confidence threshold and (ii) if at no exit the confidence exceeds the threshold, ignore the context c and use the zero-shot prediction. With these augmentations, the safe ICL model can be defined as:

$$\bar{y}_\lambda := \begin{cases} \arg\max_{k\in\mathcal{Y}} p_1\left(k|x,c\right) & \text{if } \mathfrak{C}_1 \geq \lambda, \\ \vdots & \vdots \\ \arg\max_{k\in\mathcal{Y}} p_L\left(k|x,c\right) & \text{else if } \mathfrak{C}_L \geq \lambda, \\ \arg\max_{\mathbf{k}\in\mathcal{Y}} \mathbf{p_L}\left(\mathbf{k}|\mathbf{x}\right) & \textbf{otherwise.} \end{cases} \tag{2}$$

Note that such a predictor enables us to leverage early-exit to gain efficiency and performance on helpful demonstrations, while also intervening early to avoid harmful demonstrations before the model fully processes them. When early-exit alone is insufficient to guarantee safety, *the zero-shot model serves as a reliable baseline* (see the last condition in Eq.2).

3.2 Early Exit Reduces Overthinking

As observed by Halawi et al. (2024), overthinking primarily arises in the deeper (i.e., later) layers of models, where it can override the strong inductive biases or correct predictions established in earlier layers. This phenomenon can result in degraded or unsafe outputs, as illustrated in Fig. 1. While prior work proposes pruning certain attention heads in the later layers as a preventive measure (Halawi et al., 2024), we instead advocate for leveraging dynamic early exiting (Teerapittayanon et al., 2016) which has previously been employed to mitigate overthinking in settings outside the scope of LLMs and in-context learning (Kaya et al., 2019; Jazbec et al., 2023). Notably, early exiting offers a natural solution to the overthinking problem: by terminating inference at an intermediate layer, it prevents the model from fully processing potentially misleading context. Thus, stopping early increases accuracy when the context is harmful. We show that overthinking behavior occurs on all our tasks in Fig. 17. Although not our primary goal, one nice side benefit of early-exiting is efficiency gains, as not all layers are processed before outputting a prediction.

We demonstrate the effectiveness of early exiting in mitigating ICL overthinking in Fig.3. Across most of the datasets considered, and for a broad range of exit thresholds λ, an early-exit LLM (p_λ) outperforms the full model (p_L) on inputs with incorrect demonstrations (§I). Importantly, for certain thresholds – the yellow highlighted area in Fig.3 – early termination also does not significantly degrade performance on samples with helpful demonstrations. These results underscore the potential of dynamic inference to prevent LLMs from picking up on harmful demonstrations.

3.3 In-Context Learning Risk

After describing how early exiting can be used to prevent overthinking, we now turn to the problem of selecting an appropriate early-exit strategy using risk control, as introduced in §2.2. As a first step, we propose a novel in-context learning loss:

$$\ell_{\text{ICL}}(\lambda; x, y, c) := \ell(\bar{y}_\lambda(x, c), y) - \ell(\hat{y}(x), y) \; , \tag{3}$$

where $\bar{y}_\lambda(x, c) := \arg\max_k \bar{p}_\lambda(k, x, c)$ and $\hat{y}(x) := \arg\max_k p_L(k|x)$ denote the predictions of the safe ICL model with demonstrations c (see Eq. 2) and the full zero-shot model, respectively. Crucially, both predictions are produced by the same underlying LLM. The ℓ_{ICL} loss compares the performance of the early-exit model with demonstrations to that of the model without demonstrations. This formulation makes it well-suited for measuring overthinking: if the demonstrations are harmful, the loss will be positive; if they are helpful, the loss will be negative. In contrast, prior early-exiting work (Schuster et al., 2022; Jazbec et al., 2024) has focused exclusively on loss definitions that compare early-exit outputs to final outputs given the same input. Such losses are less effective for identifying and addressing overthinking, as they do not account for the drop in performance in later layers due to harmful in-context demonstrations. Our model thus allows for robust measurement of risk considering *both correct and incorrect demonstrations*, enabling effective risk control with mixed-quality context.

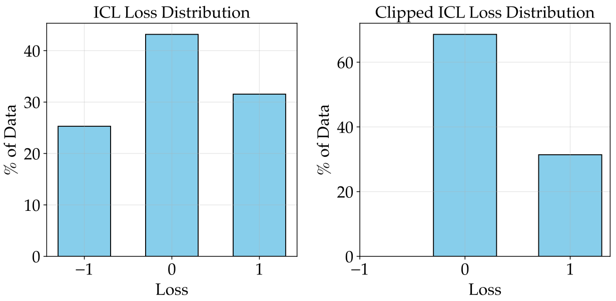
Having defined the overthinking loss, we now turn to identifying an appropriate threshold $\hat{\lambda}$ using a suitable risk control framework and a calibration dataset $\mathcal{D}_{\text{cal}} = \{(x_i, y_i, c_i)\}_{i=1}^{N_{\text{cal}}}$. Our goal is to find a threshold for which the overthinking risk is small, i.e.,

$$R_{\text{ICL}}(\hat{\lambda}) = \mathbb{E}_{(x,y,c)\sim\mathcal{P}}[\ell_{\text{ICL}}(\hat{\lambda}; x, y, c)] \leq \epsilon \; ,$$

where $\epsilon > 0$ is a user-specified tolerance level representing the acceptable degree of overthinking. Naturally, smaller values of ϵ impose stricter control, prioritizing thresholds that suppress overthinking more aggressively. However, this may come at the cost of reduced performance on helpful demonstrations—a tradeoff we explore in §H. Since we observe that risks computed using the ICL loss ℓ_{ICL} are not monotonic with respect to λ (Fig. 7), the Learn-then-Test (LTT) framework (Angelopoulos et al., 2021) is the only viable option; hence, we use LTT as our risk-control approach for selecting our exit threshold $\hat{\lambda}$.

3.4 Domain-Preserving Risk Transformation

While LTT (Angelopoulos et al., 2022) supports non-monotonic losses/risks, it requires the loss to be bounded, $\ell \in [0, 1]$, due to its reliance on the Hoeffding-Bentkus bound (Bentkus, 2004). Schuster et al. (2022) circumvented this by clipping all negative loss values to zero. However, for our in-context learning risk, negative losses are important: they correspond to helpful demonstrations from which we want to leverage performance gains over zero-shot. These negative losses are also quite common on our tasks, and clipping them means we lose a lot of information about the true underlying loss distribution (Fig. 4). By clipping these losses to zero, the risk control procedure cannot distinguish between performing *at* or *better than* the baseline (in our case, zero-shot). This introduces substantially more conservative early-exiting, which can make LTT impractical when we want to favor correct demonstrations for performance improvement and greater efficiency gains.


ICL Loss Distribution

Clipped ICL Loss Distribution

Figure 4: We show the distribution of our ICL loss on the TweetEval-Hate dataset with a 50% mix of correct and incorrect demonstrations. There are a significant number of negative loss values, which the loss-clipping approach sets to 0. Our risk transformation approach enables us to preserve the original underlying loss distribution.

We propose a novel risk transformation approach to overcome this limitation on LTT. In particular, given a risk level ϵ and a bounded loss $\ell(\lambda; x, y, c) \in [a, b]$ for any $a, b \in \mathbb{R}$, we can execute the following procedure:

1. Compute $\epsilon' = \frac{\epsilon - a}{b - a}$ and $\ell'(\lambda; x, y, c) = \frac{\ell_{\text{ICL}}(\lambda; x, y, c) - a}{b - a}$.
2. Define ϵ' as the new risk level and $\ell'(\lambda; x, y, c)$ as the new loss. $\ell'(\lambda; x, y, c) \in [0, 1]$ and $\epsilon' \in [0, 1]$, so the Hoeffding-Bentkus bound is satisfied.
3. Apply LTT to select $\hat{\lambda}$ with the risk level ϵ' and the risk $R(\ell'(\hat{\lambda}; x, y, c))$. $\hat{\lambda}$ also controls risk $R(\ell(\hat{\lambda}; x, y, c))$ at level ϵ, as we prove in §C.

The key insight for this approach to work is that controlling the risk $R(\lambda) := \mathbb{E}_{x,y,c}[\ell]$ at level ϵ is equivalent to controlling the risk $R'(\lambda) := \mathbb{E}_{x,y,c}[\ell']$ at level ϵ'. Intuitively, this is because we are applying the same invertible transformation to both the loss and the risk level, and we can simply reverse the transformation after applying LTT to return to the domain of the original loss and risk level. The full proof can be found in §C.

Application to In-Context Learning. Note that due to our focus on classifcation, our ICL loss is bounded $\ell_{\text{ICL}}(\lambda) \in [-1, 1]$; hence we can plug in $a = -1, b = 1$ when performing loss scaling in our setting. We contrast our approach with the previous approach of clipping negative losses to zero $\ell_{\text{clip}}(\lambda) := \max\{0, \ell_{\text{ICL}}(\lambda)\}$ and empirically verify that our approach both satisfies the same risk control guarantees and achieves much greater efficiency gains (see Fig. 5 and 6).

4 Experiments

Tasks We use a total of 8 tasks for our work, spanning three diverse domains (sentiment analysis, hate speech detection, and semantic classification). Our tasks are the Stanford Sentiment Treebank (Socher et al., 2013), FinancialPhrasebank (Malo et al., 2013), TweetEval-Hate, -Atheism, and -Feminist (Barbieri et al., 2020), AG News (Del corso et al., 2005), Text REtrieval Conference (TREC) (Li & Roth, 2002), and Unnatural (Halawi et al., 2024). A detailed description of these datasets and the domains they cover can be found in §B.

Models We compare four models – two regular LLaMA models (Llama-3-8B (et al., 2024) and Llama-2-7B (et al., 2023)) and two LayerSkip LLaMA models (layerskip-llama3-8B and layerskip-llama2-7B (Elhoushi et al., 2024)). The LayerSkip models are additionally pre-trained to encourage the production of higher-quality intermediate representations, which provide a helpful point of comparison with the models that are not explicitly pretrained as early-exit models.

Experimental Design In our experiments, we use the following settings:

- *Selecting Calibration Data:* From each dataset, we randomly draw 50% for our calibration dataset (on which we compute $\hat{\lambda}$) and the remaining 50% is our test data on which we present results.
- *Label Transformation:* Existing datasets are often memorized during model pre-training (Li et al., 2024). Thus, we transform the tasks into a format that is equivalent to, but distinct from, their original form to mitigate these linguistic biases, mirroring the approach taken in prior work (Fang et al., 2025; Pan et al., 2023). We show that dataset memorization happens, and that this label transformation approach mitigates this effect, in §J.4.
- *In-Context Demonstrations:* During the risk control calibration step, we compute a single $\hat{\lambda}$ for risk control on a 50-50 mix of *both incorrect and correct demonstrations*. Incorrect demonstrations are obtained by permuting the labels, as in Halawi et al. (2024).
- *Contextual Calibration:* Since our focus is on classification tasks, we examine how frequently the model assigns a higher probability to the correct label than to any alternative. However, model outputs can be highly sensitive to minor changes in the prompt (Gao et al., 2021), an observation which we verified through additional experiments, detailed in §J.4. To address this instability, we apply contextual calibration (Zhao et al., 2021) to balance the label probabilities.

- *Evaluation Metrics:* We evaluate our models primarily on the in-context learning risk (as defined in §3.3), demonstrating that with our approach, risk always remains below the user-defined ϵ threshold.

4.1 Empirical Verification of Risk Control Guarantees

We empirically verify that our approach always controls the risk across all models and datasets, even when given a mix of correct and incorrect demonstrations in the prompts. Fig.5 shows that our approach satisfies the DFRC guarantees on risk: the in-context learning risk is controlled across all models, tasks, and risk levels ϵ. We provide additional results in the appendix (§E) comparing these results with the loss-clipping approach, showing that our risk transformation approach is consistently less conservative and better matches the user-defined risk level ϵ than loss-clipping approach. This highlights both the validity of our theoretical results as well as applicability to real-world tasks. We additionally demonstrate that our risk-control guarantees hold regardless of the distribution of correct vs incorrect demonstrations in the data distribution (§G).

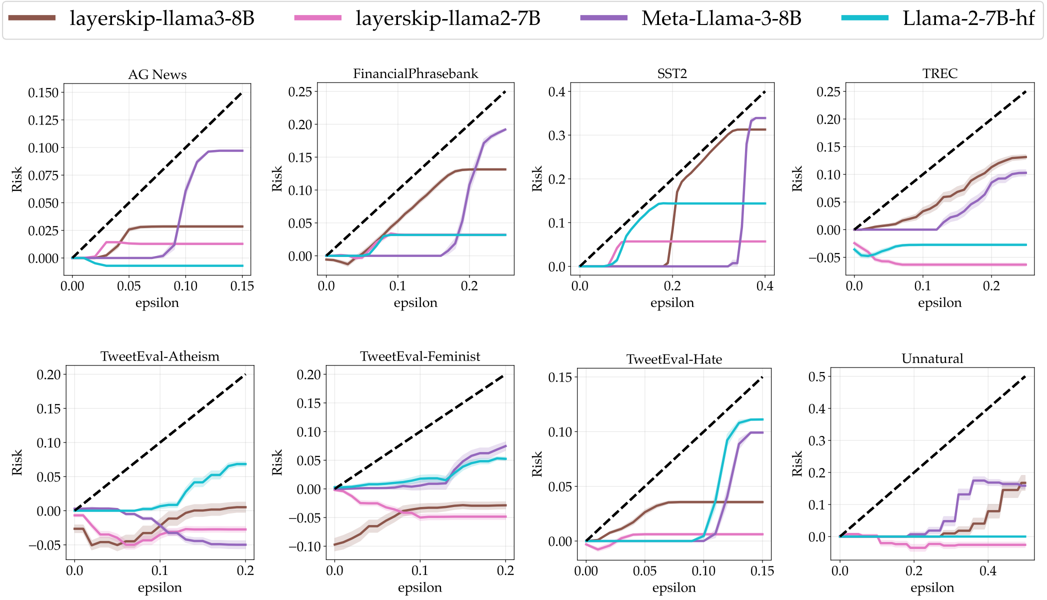


Figure 5: Empirical risk vs the user-specified risk level ϵ using our safe ICL model and ℓ_{ICL} loss over a set of mixed correct and incorrect demonstrations. Aligning with the theoretical guarantees, the risk is controlled across all models and tasks. Shaded regions correspond to one standard error over 100 experiments and are included on all plots.

4.2 Comparison of Efficiency Gains with Loss Clipping

Our adaptation of the loss-bounding approach outperforms prior LTT approaches (Jazbec et al., 2024; Schuster et al., 2022) by increasing efficiency gains while preserving the same risk control assurances. At all ϵ levels, our efficiency gains – the number of layers of computation we save by applying our approach – are strictly greater than when using clipped risk (Fig. 6), and are often significantly greater. For example, controlling the prediction gap risk with our approach at $\epsilon = 0.05$ results in an average of 53% less layers evaluated over all datasets and models compared to loss-clipping, while still satisfying the same rigorous risk control guarantees. Though not a primary goal of our work, it is a notable side benefit of our approach.

4.3 Class-Conditional Effects for Correct and Incorrect Demonstrations

We examine the effects of our chosen $\hat{\lambda}$ on different sub-populations of our data, corresponding to whether the model is given correct or incorrect demonstrations. We find that where there exists a $\hat{\lambda}$

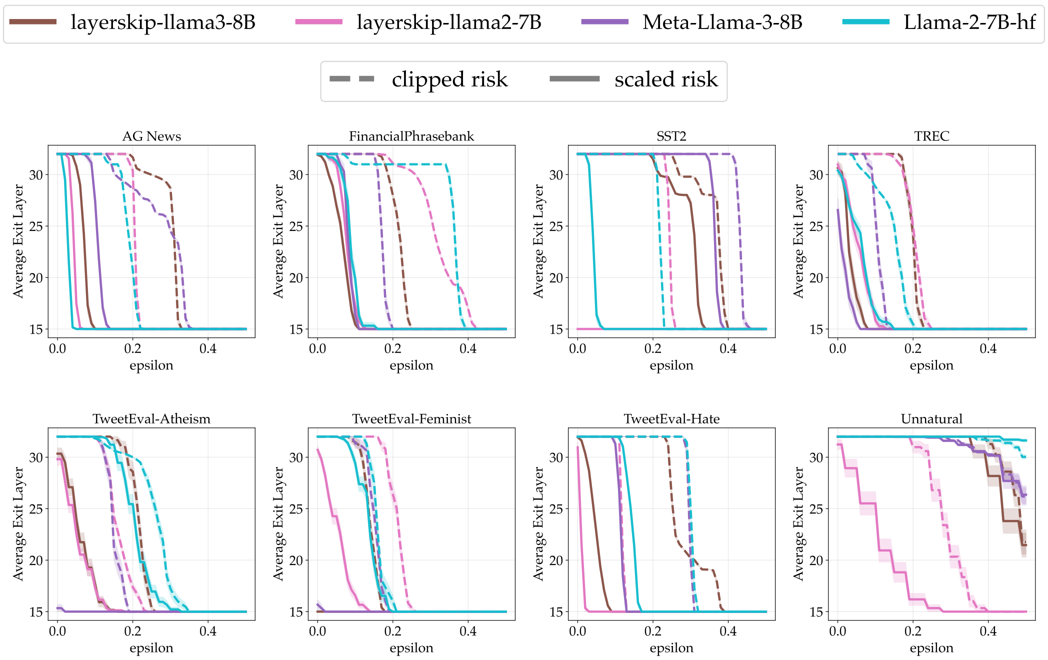


Figure 6: We demonstrate that our risk transformation approach enables much greater efficiency gains than the loss-clipping approach by leveraging performance gains from correct demonstrations.

that both controls overthinking risk and preserves the accuracy gains from correct demonstrations, our approach is able to find it. In other cases, we find that there can be a direct tradeoff between taking advantage of performance gains given correct demonstrations and controlling overthinking risk given incorrect demonstrations, depending on the choice of $\hat{\lambda}$; results are shown in the appendix in Fig.16.

Though the marginal guarantees from risk control do not extend in theory to class-conditional risk control (how well our risk control works within each subgroup of our data, i.e. correct vs incorrect demonstrations) when using a mixed calibration dataset with both incorrect and correct in-context demonstrations, our approach still shows promise for finding an appropriate $\hat{\lambda}$ when one exists that *can* perform well for both correct and incorrect demonstrations. In practice, for applications where controlling risk given unsafe prompts is more important than performance gains from helpful prompts or achieving greater efficiency gains, we can instead use the loss-clipping approach, which is much more conservative and therefore may better control risk on unsafe prompts alone, as shown in Fig.14 and 15. However, even with loss clipping, there are no guarantees for risk control on incorrect demonstrations alone for the same reason; as shown in Fig.14 and 15, loss clipping can still violate the risk-control on the subclass of only incorrect demonstrations.

Ideally, we would like to condition on these sub-populations separately, but we cannot know ahead of time whether we are given correct or incorrect demonstrations. Future work should investigate how we can integrate additional control mechanisms with our approach to ensure safe behavior in *all* subgroups of safe and unsafe prompts, potentially by borrowing ideas from class-conditional conformal prediction literature (Ding et al., 2023).

5 Related Work

In-context learning (ICL). In-context learning, also known as few-shot learning, enables LLMs to perform new tasks by conditioning on a limited number of input-output examples within the prompt without requiring gradient-based fine-tuning (Dong et al., 2024; Radford et al., 2019; Brown et al., 2020; Srivastava et al., 2023). This allows LLMs to adapt to novel tasks by generalizing from examples in the prompt, but this flexibility also makes LLMs vulnerable. For example, novice users who provide incorrect examples may make an LLM perform worse on a task than it would have without the user's input (Halawi et al., 2024), or adversarial users can design prompts to make LLMs bypass their safeguards (Xu & Wang, 2024). In our work, we consider how the *quality* of in-context demonstrations can impact the safety of the model's outputs.

Evolution of representations through LLM layers. Recent studies have begun mapping how LLM hidden representations evolve with depth, revealing structured processing phases. Cheng et al. (2025); Cheng & Antonello (2024); Cheng et al. (2023) identify a pronounced intermediate-layer "abstraction phase" in which hidden states predict brain responses to language stimuli, showing that LLMs compress inputs into low-dimensional manifolds early in processing. Layer-wise analysis has also been applied to tasks such as ciphers (Fang et al., 2025), long-context failures (Lu et al., 2024), and multilingual representations (Bafna et al., 2025; Muller et al., 2021; Wendler et al., 2024; Schut et al., 2025) in LLMs. We leverage these intermediate representations to determine when to safely make a prediction on a task.

Distribution-free risk control (DFRC). Risk control is a statistical framework for controlling various measures of risk in machine learning systems. Given a trained model, a finite set of calibration data, and a loss function reflecting the chosen measure of safety, DFRC bounds the expected loss, i.e., *risk*, as a function of some low-dimensional parameter λ (Angelopoulos et al., 2023; Bates et al., 2021b). Among existing frameworks, Learn Then Test (LTT) (Angelopoulos et al., 2022) is widely used, as it is the only method that provides guarantees without requiring monotonicity of the loss or risk. However, LTT assumes the loss is bounded within $[0, 1]$, which can be restrictive in certain scenarios (see §3.4 for a discussion on the implications of bounded loss in the ICL setting and our approach to addressing this limitation). Notable examples of using risk control in the context of LLMs include controlling performance degradation due to accelerated inference (Schuster et al., 2022; Jazbec et al., 2024) and mitigating prompt-induced variability (Zollo et al., 2024). In contrast, our work is the first to leverage risk control for managing the impact of harmful demonstrations on the downstream performance of LLMs.

Early Exiting. Early exiting in deep neural networks enhances computational efficiency by terminating inference at intermediate layers for simpler inputs, thereby reducing resource consumption with minimal performance degradation (Teerapittayanon et al., 2016). While early exiting has been widely adopted to accelerate inference (Huang et al., 2018; Zhou et al., 2020; Elbayad et al., 2020; Han et al., 2021; Schuster et al., 2022), our work introduces a novel application of early exit architectures: mitigating the influence of incorrect demonstrations in in-context learning. Although the use of early exiting to prevent ICL overthinking has also been discussed in Halawi et al. (2024), their approach relies on static layer pruning. In contrast, our method employs dynamic, per-sample early exiting based on confidence thresholding, which enables us to make robust guarantees for safety via risk control.

6 Conclusion and Discussion

Our work introduces a novel risk-controlled early-exit framework for safe in-context learning (ICL) that robustly handles demonstrations of mixed quality. We present three important contributions to enable safety for ICL: a novel early-exit model formulation using a zero-shot baseline; a novel ICL loss designed to measure overthinking; and an adaptation of the LTT risk control framework to work for our setting. This integrated approach improves the safety, reliability, and computational efficiency of ICL under mixed-quality demonstrations.

A potential limitation of our work – and of prior work on risk control – is that we do not make conditional guarantees on risk control for correct vs incorrect demonstrations when a model is presented with demonstrations of mixed quality. Future work could investigate class-conditional risk control to provide more robust safety assurances under mixed-quality prompts. Additionally, future approaches could be applied where both correct and incorrect demonstrations are provided within the same prompt; this may reflect many real-world use cases, as a human user may have inconsistent performance even when constructing the same prompt.

Ethics Statement

This work contributes to safer deployment of LLMs by introducing a principled method to detect and mitigate "overthinking" by accounting for potential user error. Our approach additionally improves computational efficiency, making LLM inference more environmentally and economically sustainable. However, though our approach improves safety in the average case, it does not provide strong guarantees for specific subpopulations or worst-case prompts, potentially leaving some harmful scenarios unmitigated. Additionally, the increased reliance on automated risk control mechanisms may give users a false sense of security, especially if deployed without proper monitoring or human oversight.

Reproducibility Statement

Detailed descriptions of the experiment setup, including datasets, methods, and implementation details, are referenced in §4. The exact prompt formats used for all tasks are presented in §L and data processing steps are outlined in §4, with justification provided through ablation studies in §J. We provide a proof in §C showing the validity of our risk transformation approach (§3.4). Additionally, source code for all experiments is provided along with this paper submission and a GitHub link will be included in the final camera-ready paper.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv Preprint (arXiv:2110.01052)*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022. URL `https://arxiv.org/abs/2110.01052`.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2023. URL `https://arxiv.org/abs/2208.02814`.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Fracesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanhan, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 129696–129742. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/ea456e232efb72d261715e33ce25f208-Paper-Conference.pdf`.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *arXiv preprint arXiv:2506.22724*, 2025. URL `https://arxiv.org/abs/2506.22724`.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020. URL `https://arxiv.org/abs/2010.12421`.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 2021a.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6), September 2021b. ISSN 0004-5411. doi: 10.1145/3478535. URL `https://doi.org/10.1145/3478535`.

Vidmantas Bentkus. On hoeffding's inequalities. *Annals of probability*, 2004.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Emily Cheng and Richard J. Antonello. Evidence from fmri supports a two-phase abstraction process in language models, 2024. URL https://arxiv.org/abs/2409.05771.

Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models, 2023. URL https://arxiv.org/abs/2310.13620.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers, 2025. URL https://arxiv.org/abs/2405.15471.

Nilanjana Das, Edward Raff, and Manas Gaur. Human-interpretable adversarial prompt attack on large language models with situational context, 2024. URL https://arxiv.org/abs/2407.14644.

Gianna Del corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. pp. 97–106, 01 2005. doi: 10.1145/1060745.1060764.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer, 2020. URL https://arxiv.org/abs/1910.10073.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.681. URL http://dx.doi.org/10.18653/v1/2024.acl-long.681.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Zhouxiang Fang, Aayush Mishra, Muhan Gao, Anqi Liu, and Daniel Khashabi. ICL Ciphers: Quantifying "Learning" in In-Context Learning via Substitution Ciphers. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2025. URL https://arxiv.org/abs/2504.19395.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295/.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations, 2024. URL https://arxiv.org/abs/2307.09476.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7436–7456, 2021.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. *International Conference on Learning Representations*, 2018.

Metod Jazbec, James Allingham, Dan Zhang, and Eric Nalisnick. Towards anytime classification in early-exit architectures by enforcing conditional monotonicity. *Advances in Neural Information Processing Systems*, 36:56138–56168, 2023.

Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian Andersson Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. *Advances in Neural Information Processing Systems*, 37:129825–129854, 2024.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *International conference on machine learning*, pp. 3301–3310. PMLR, 2019.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Sudhanshu Kumar, Partha Pratim Roy, Debi Prosad Dogra, and Byung-Gyu Kim. A comprehensive review on sentiment analysis: Tasks, approaches and applications, 2023. URL https://arxiv.org/abs/2311.11250.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

Yucheng Li, Frank Guerin, and Chenghua Lin. An open source data contamination report for large language models, 2024. URL https://arxiv.org/abs/2310.17589.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models, 2024. URL https://arxiv.org/abs/2403.04957.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into llm long-context failures: When transformers know but don't tell. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP) - *Findings*, 2024. URL https://arxiv.org/abs/2406.14673.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013. URL https://arxiv.org/abs/1307.5336.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Understanding the cross-lingual ability of multilingual bert, 2021. URL https://arxiv.org/abs/2101.11109.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning ”learns” in-context: Disentangling task recognition and task learning, 2023. URL https://arxiv.org/abs/2305.09731.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. URL https://openai.com/blog/better-language-models/.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling, 2022. URL https://arxiv.org/abs/2207.07061.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english?, 2025. URL https://arxiv.org/abs/2502.15603.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna

Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (TMLR), 2023. URL https://arxiv.org/abs/2206.04615.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pp. 2464–2469. IEEE, 2016.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers, 2024. URL https://arxiv.org/abs/2402.10588.

Yue Xu and Wenjie Wang. LinkPrompt: Natural and universal adversarial attacks on prompt-based language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6473–6486, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.360. URL https://aclanthology.org/2024.naacl-long.360/.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Jialuo Chen, Hui Xue, Xiaoxia Liu, Wenhai Wang, Kui Ren, and Jingyi Wang. S-eval: Towards automated and comprehensive safety evaluation for large language models, 2025. URL https://arxiv.org/abs/2405.14191.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models, 2024. URL https://arxiv.org/abs/2309.07045.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021. URL `https://arxiv.org/abs/2102.09690`.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems*, 2020.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models, 2024. URL `https://arxiv.org/abs/2311.13628`.

A Appendix

B Detailed Description of Tasks

Sentiment analysis. Sentiment analysis refers to the computational study of opinions, emotions, and attitudes expressed in text (Kumar et al., 2023), which requires inferring polarity or stance from often subtle or domain-specific cues. We use three sentiment analysis datasets: Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) involves binary sentiment classification of movie reviews, Financial Phrasebank (Malo et al., 2013) extends this task to the financial domain, and TweetEval-Feminist (Barbieri et al., 2020) centers on sentiment detection toward feminism in social media posts.

Hate Speech Detection. Hate speech detection involves identifying language that expresses hatred, discrimination, or hostility toward individuals or groups, often within socially sensitive contexts. We examine two datasets from the TweetEval benchmark (Barbieri et al., 2020) that address this problem in distinct but related ways. The TweetEval-Hate dataset consists of tweets directly annotated for the presence or absence of hate speech, whereas the TweetEval-Atheism dataset is used in studying hate speech due to its focus on religion-related discourse, where antagonistic or prejudiced language is common.

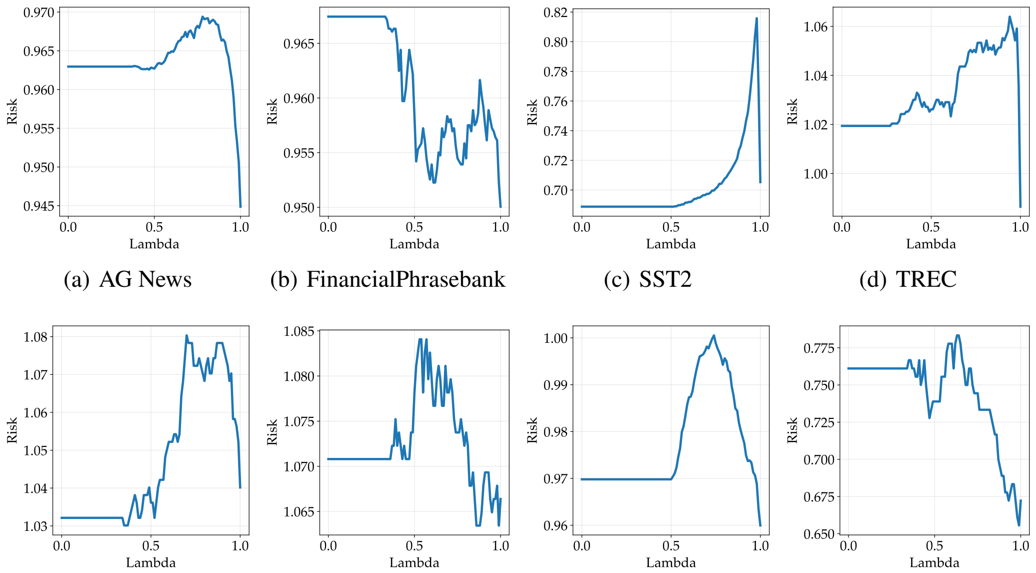
Semantic Classification. Semantic classification tasks involve assigning text to high-level conceptual categories based on meaning, structure, or subject matter, focusing on identifying the core informational content of input text. We examine three semantic classification tasks: AG News (Del corso et al., 2005) involves classifying news headlines into four broad areas – World, Sports, Business, and Science/Technology. The Text REtrieval Conference (TREC) dataset (Li & Roth, 2002) involves classifying open-domain questions into six semantic types (e.g., entity, location, numeric). The Unnatural dataset, a toy dataset constructed by (Halawi et al., 2024), assigns short text descriptions to one of three semantic categories: sports, animals, or plants.

C Proof of Risk Transformation Approach

Here, we prove that our risk transformation approach chooses an appropriate $\hat{\lambda}$ that controls our ICL risk.

Proof. Let $\ell(\lambda, x, c, y) \in [a, b]$, and choose a risk level $\epsilon \in (a, b)$ and some probability δ. Compute $\epsilon' = \frac{\epsilon - a}{b - a}$ and $\ell' = \frac{\ell(\lambda,x,c,y) - a}{b - a}$. Apply the Learn Then Test procedure to ϵ', ℓ' to select $\hat{\lambda}$. Formally, as shown in (Angelopoulos et al., 2021), this guarantees that $\mathbb{P}(R(\ell') \leq \epsilon') \geq 1 - \delta$ for a fixed δ. So, with probability at least $1 - \delta$, we have that $R(\ell') \leq \epsilon'$.

Towards contradiction, assume that $\hat{\lambda}$ does not control the original risk $R(\ell(\lambda, x, c, y))$ at level ϵ. Thus, $R(\ell(\hat{\lambda}, x, c, y)) = \mathbb{E}_{x,y,c}[\ell(\hat{\lambda}, x, c, y)] > \epsilon$. By definition of ℓ', we have that $\mathbb{E}_{x,y,c}[\ell(\lambda, x, c, y)] = \mathbb{E}_{\ell'}(b - a) + a = R(\ell')(b - a) + a$. Similarly, by definition of ϵ', we have that $\epsilon = \epsilon'(b - a) + a$. So we can show the following:



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

Figure 7: We show that across many of our models and datasets, the risk is non-monotonic in λ.

$$
\begin{aligned}
R(\ell(\hat{\lambda}, x, c, y)) &> \epsilon \\
R(\ell')(b-a) + a &> \epsilon'(b-a) + a \\
R(\ell')(b-a) &> \epsilon'(b-a) \\
R(\ell') &> \epsilon'
\end{aligned}
$$

However, we know that $\hat{\lambda}$ controls $R(\ell')$ at level ϵ'. This is a contradiction. This proves that $\hat{\lambda}$ must also control $R(\ell(\hat{\lambda}, x, c, y))$ at level ϵ.

D Non-Monotonicity of Risk

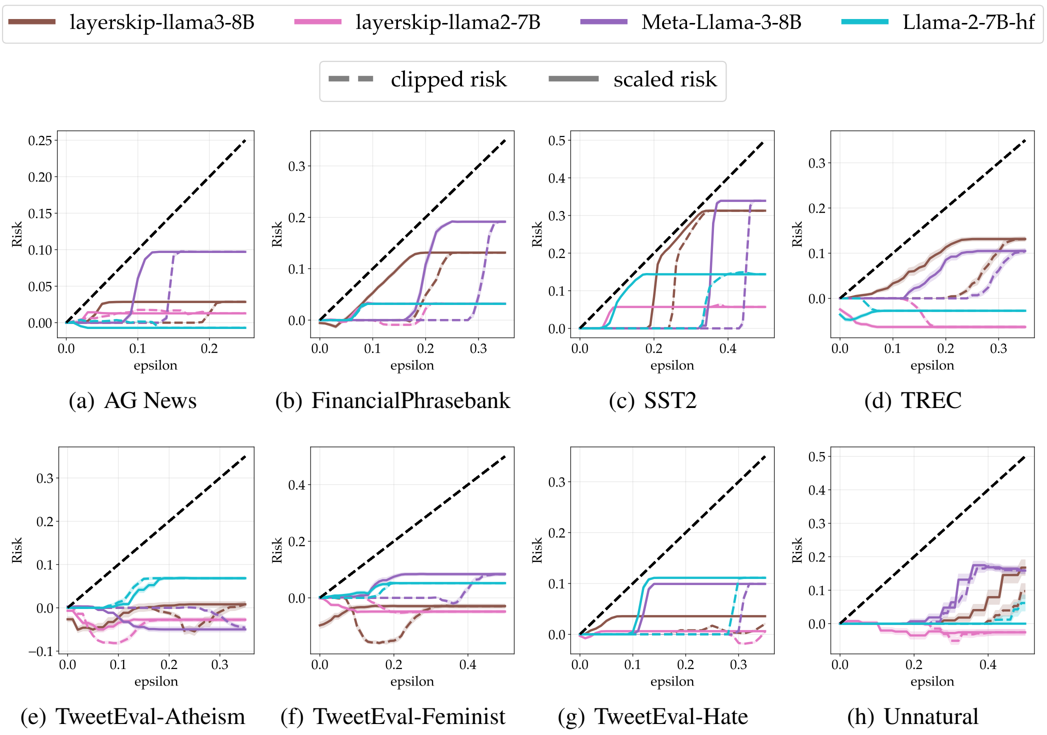
We show that our risk is non-monotonic in λ across many of our models and datasets in Fig.7. This indicates that we cannot use many of the existing methods in the conformal risk control literature, because they require an assumption of monotonicity (Jazbec et al., 2024), and motivates our choice of Learn Then Test in our work as it does not require this assumption.

E Comparison of Risk Control Approaches

Across all tasks and models, we find that our risk transformation approach is less conservative and better matches the user-defined risk level ϵ than the loss-clipping approach. A direct comparison is illustrated in Fig. 8 with a mix of 50% correct and 50% incorrect demonstrations.

F Demonstrating Robustness of Risk Control with Variation in Threshold λ

We provide a plot, Figure 9, with confidence intervals added to our results in Fig. 3 by bootstrapping samples from the datasets. We then only highlight λ values for which no point within our confidence interval loses more than 5% of the accuracy gains from correct demonstrations while still doing better than the full model given incorrect demonstrations. We also include error bars for the accuracy



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

Figure 8: We show that across all tasks and models, our risk transformation approach is less conservative and better matches the user-defined risk level ϵ than the loss-clipping approach. These experiments used 50% correct and 50% incorrect demonstrations.

of correct and incorrect demonstrations. Notably, even with these confidence bounds, there remain regions where the yellow band persists—demonstrating that even when accounting for the variance in samples, there still exist values of λ that satisfy our desiderata. This provides clearer evidence for the robustness of our results.

G Risk Control at Any Proportion of Correct vs Incorrect Demos

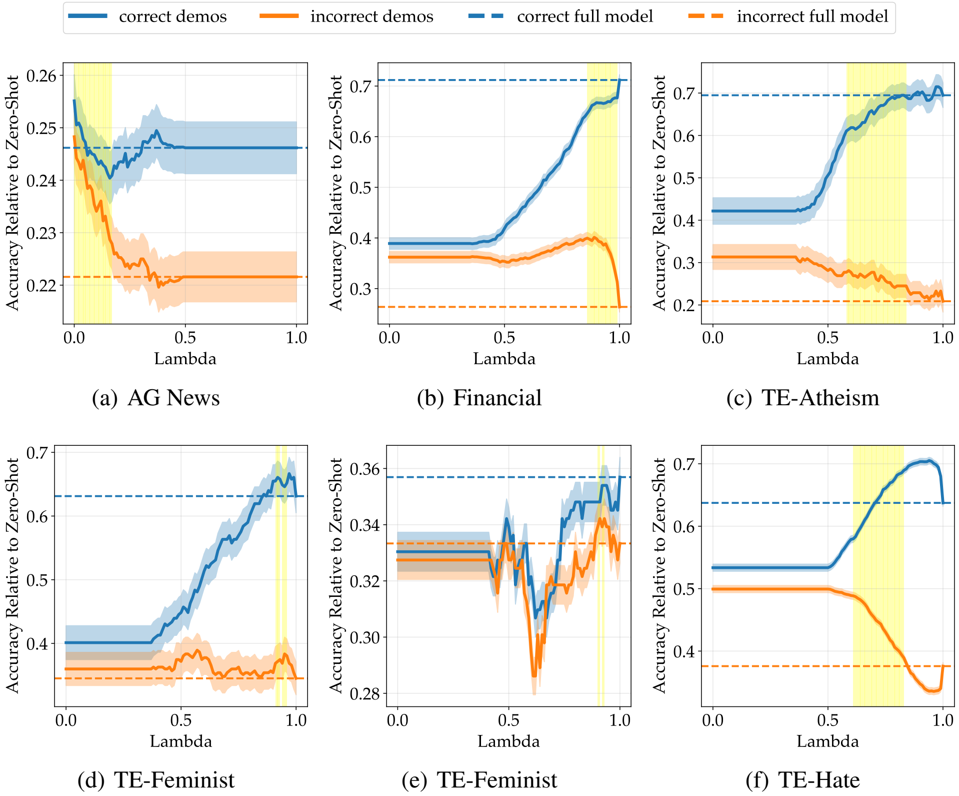
Here, we show results demonstrating that regardless of the proportion of correct vs incorrect demos in the calibration data, we are still able to control the combined risk over a test set drawn i.i.d. from the same distribution. Results shown in Figures 10, 11 and 12 for cases when there are more correct than incorrect demonstrations (a scenario that is likely in real-world applications), but we also show in Fig.13 that even when we have many more incorrect than correct demonstrations, our risk-control guarantees still hold.

H Class-Conditional Risk Control

We provide all results from our experiments investigating the class-conditional risk levels over all datasets and models using a 50-50 split of correct and incorrect demonstrations. Results are shown in Figures 16 and 15.

I Overthinking Across Datasets

We show that overthinking occurs across all of our datasets in Fig.17. This provides additional motivation for using early-exiting as a natural approach to control risk on all tasks.



(a) AG News

(b) Financial

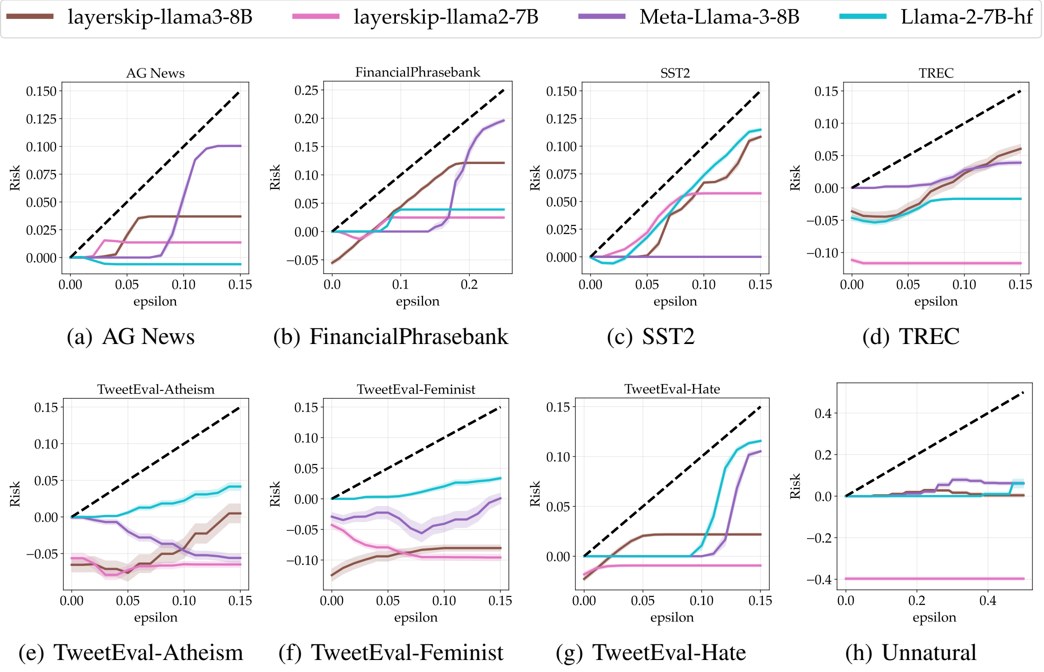
(c) TE-Atheism

(d) TE-Feminist

(e) TE-Feminist

(f) TE-Hate

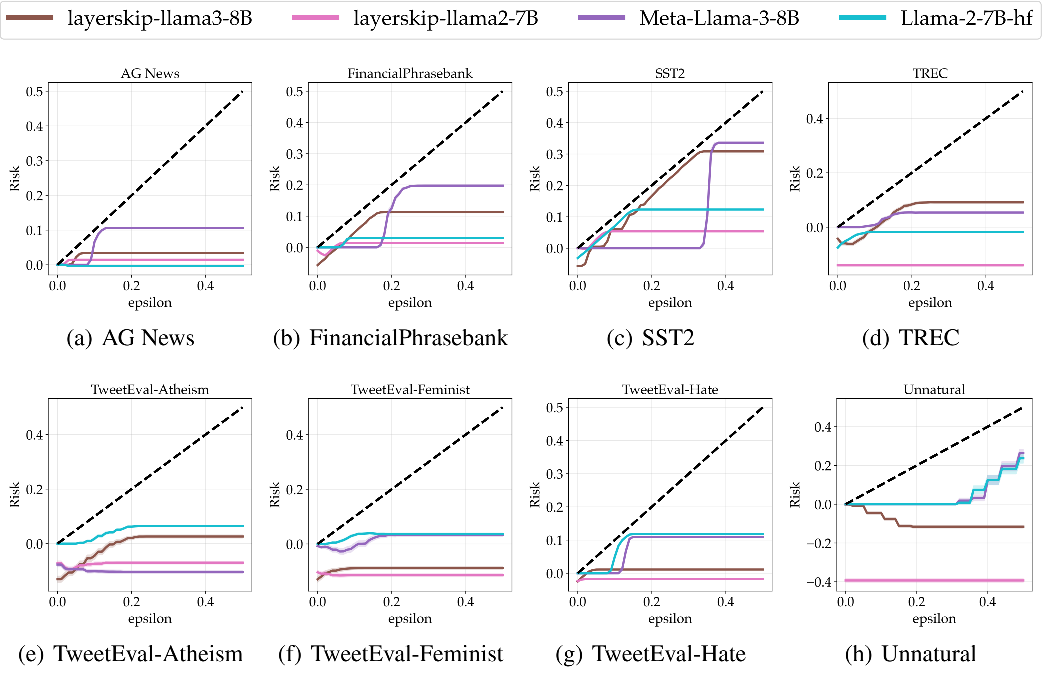
Figure 9: Some choices of λ thresholds can both attain performance gains from correct demonstrations *and* control overthinking from incorrect demonstrations. We show the robustness of our approach to different selections of λ by adding error bars that further restrict the choices of λ. Collecting more i.i.d. samples from the dataset will simply yield narrower error bars and thus more choices of λ.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

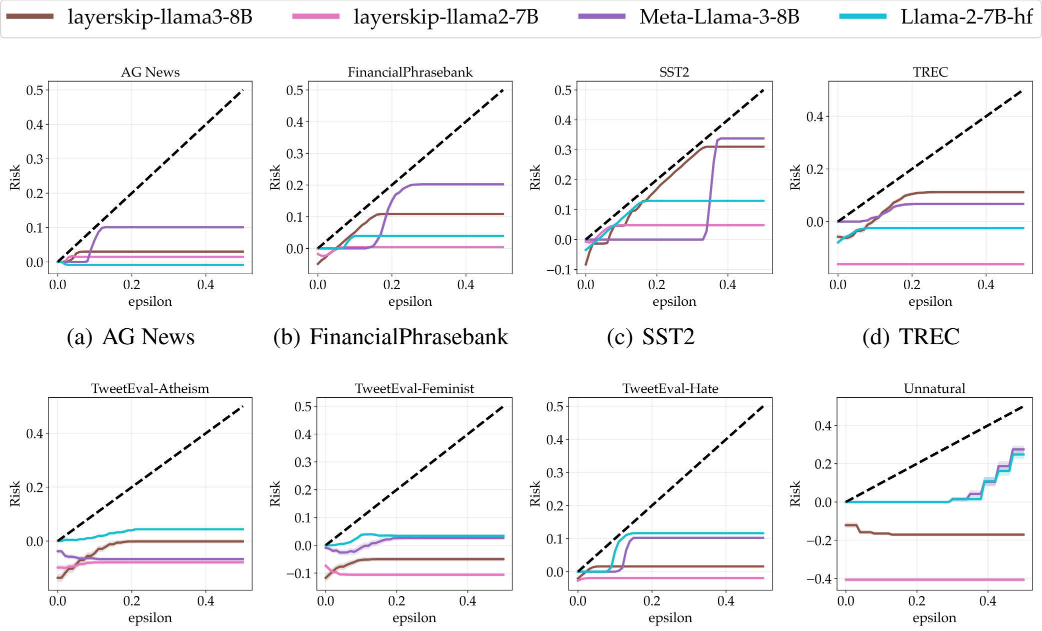
Figure 10: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of 75% correct and 25% incorrect demonstrations.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

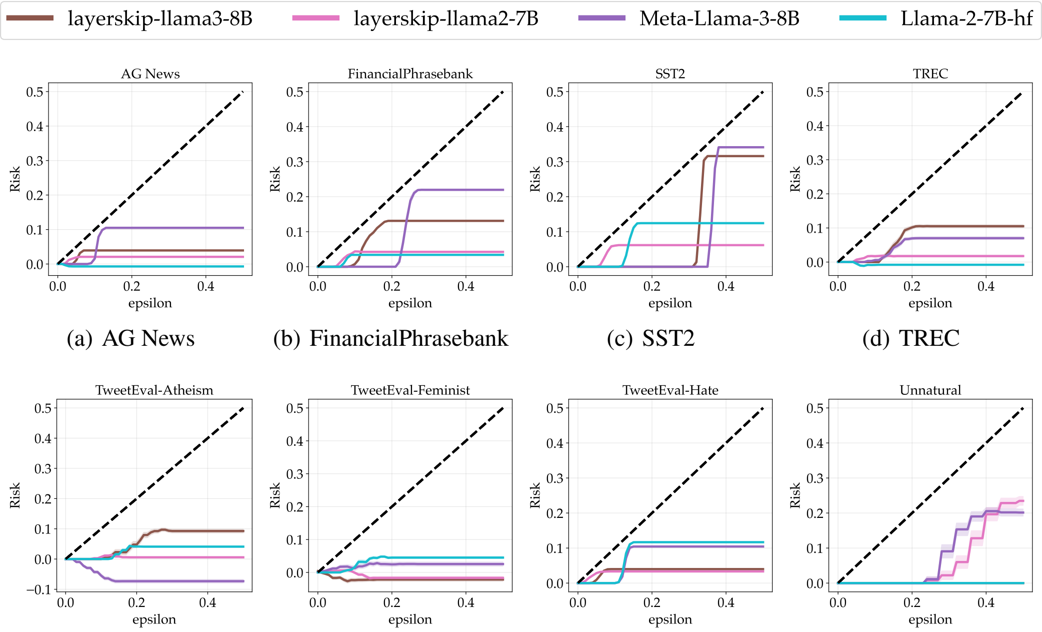
Figure 11: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of 90% correct and 10% incorrect demonstrations.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

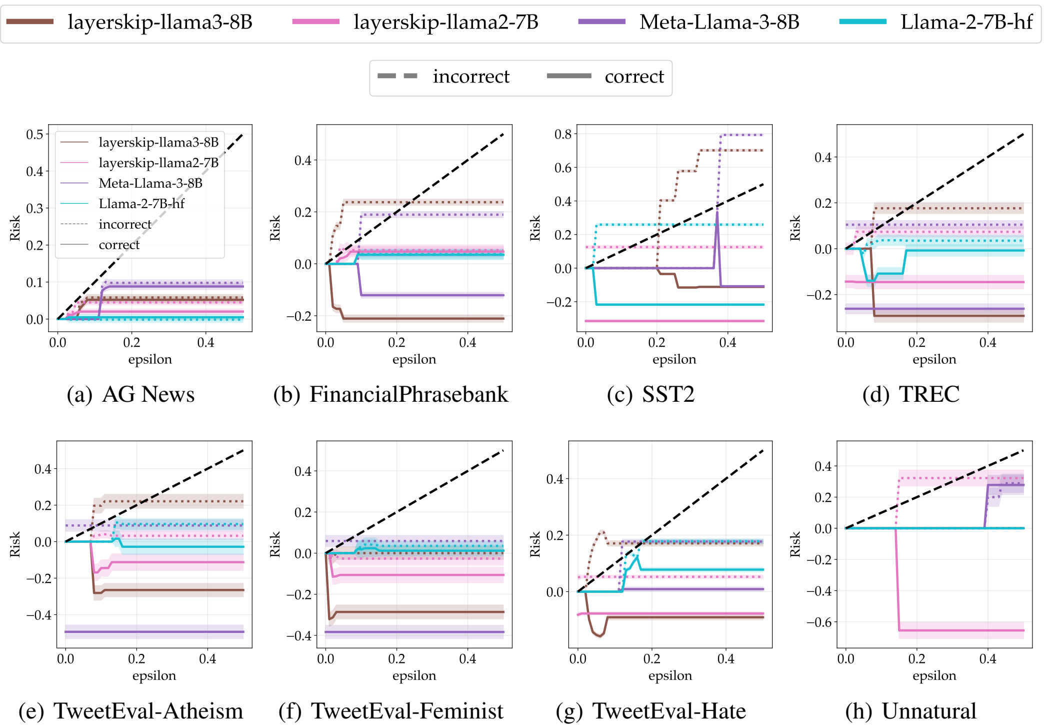
Figure 12: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of 95% correct and 5% incorrect demonstrations.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

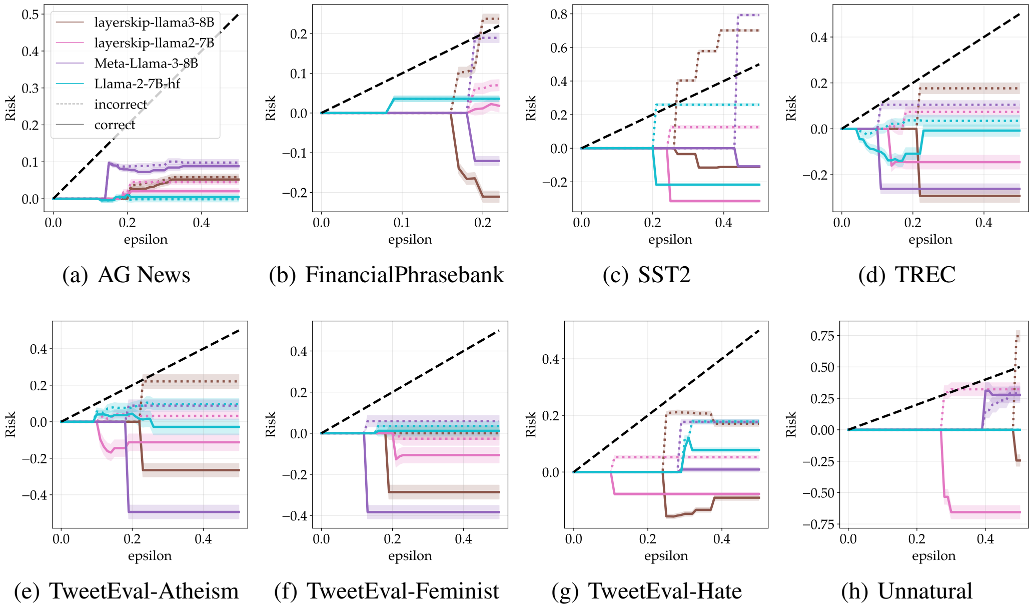
Figure 13: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of 10% correct and 90% incorrect demonstrations.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

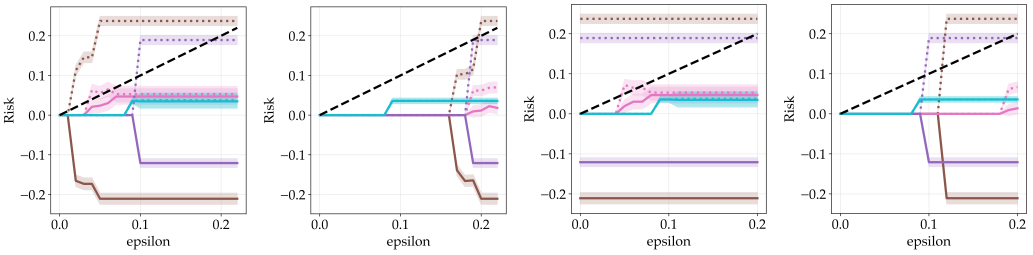
Figure 14: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of 50% correct and 50% incorrect demonstrations. We examine the class-conditional risks for correct and incorrect demonstrations respectively. Shaded regions correspond to one standard error computed over 100 experiments.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

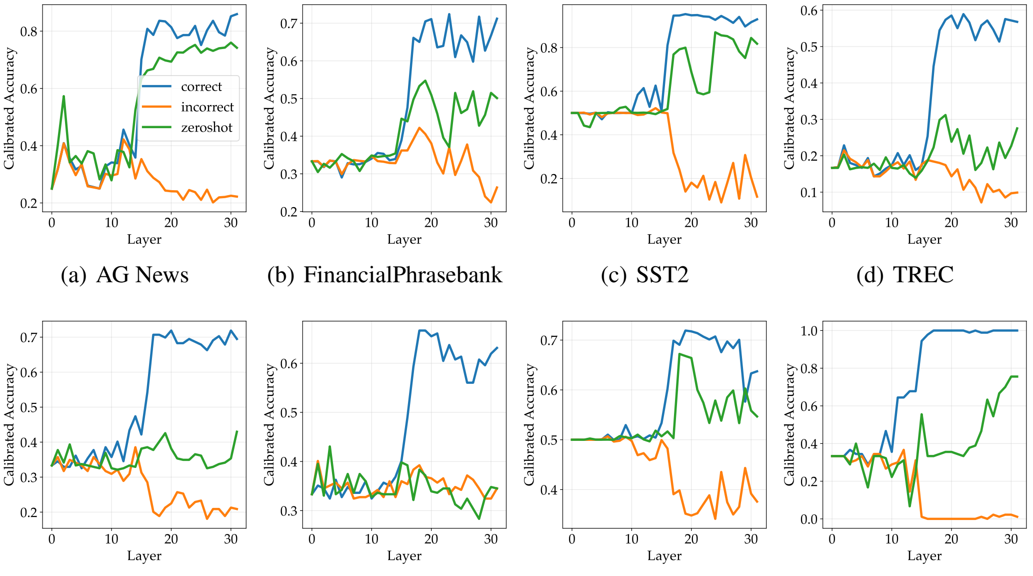
(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

Figure 15: Empirical risk vs the user-specified risk level ϵ using the clipped-loss approach over a set of some correct and some incorrect demonstrations. This approach prioritizes controlling risk for incorrect demonstrations but can reduce performance gains when given correct demonstrations.



(a) Scaling, 50-50 (b) Clipped, 50-50 (c) Scaling, 10-90 (d) Clipped, 10-90

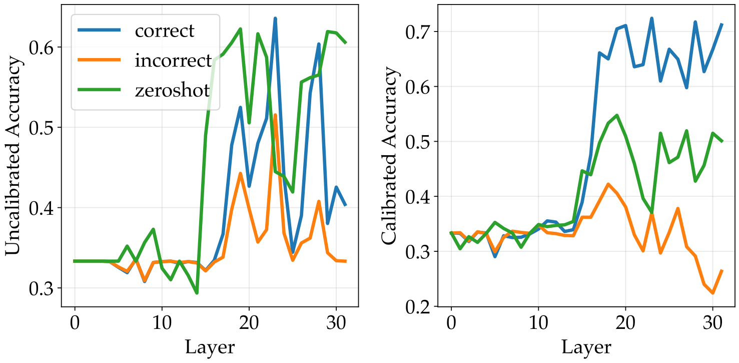
Figure 16: Empirical risk vs the user-specified risk level ϵ using our risk transformation approach over a set of some correct and some incorrect demonstrations. We examine the class-conditional risks for correct and incorrect demonstrations respectively on the FinancialPhrasebank dataset, where we either have a 50-50 balanced split of correct vs incorrect demos, or 10% incorrect and 90% correct demos. We find that our approach defaults to the zero-shot behavior much less often than the loss-clipping approach regardless of the proportion of correct demonstrations (as seen by the risk 0 regions for small ϵ). Shaded regions correspond to one standard error computed over 100 experiments.



(a) AG News (b) FinancialPhrasebank (c) SST2 (d) TREC

(e) TweetEval-Atheism (f) TweetEval-Feminist (g) TweetEval-Hate (h) Unnatural

Figure 17: Overthinking occurs across widely varying datasets - demonstrating that, if given incorrect in-context demonstrations, we should either early-exit or default to zero-shot behavior to ensure safety. However, if we are given correct demonstrations, we would like to both take advantage of the performance benefit (relative to zero-shot) and early-exit when we do not need all layers of the model to arrive at the correct answer. All plots are generated using the LayerSkip LLaMA-3 8B model.



(a) No Contextual Calibration (b) With Contextual Calibration

Figure 18: We find that calibration is necessary to stabilize accuracy and confidence across layers of the model. This enables effective early-exit for risk control. An example is shown here for FinancialPhrasebank.

J Ablations

We performed many ablation studies to arrive at the setup of our experiments in this paper. Details of these ablations are provided here, as well as the particular settings under which we ran our experiments.

J.1 Contextual Calibration

Contextual calibration (Zhao et al., 2021) reduces instability arising from the specific choice of prompt format and the choice and ordering of in-context examples; it has been widely applied in recent work, including in (Halawi et al., 2024). We performed ablations with and without contextual calibration, and found that contextual calibration was necessary to stabilize accuracy and confidence across the layers of the model. A plot comparing an experiment with and without contextual calibration is shown in Fig.18.

J.2 Confidence Measures

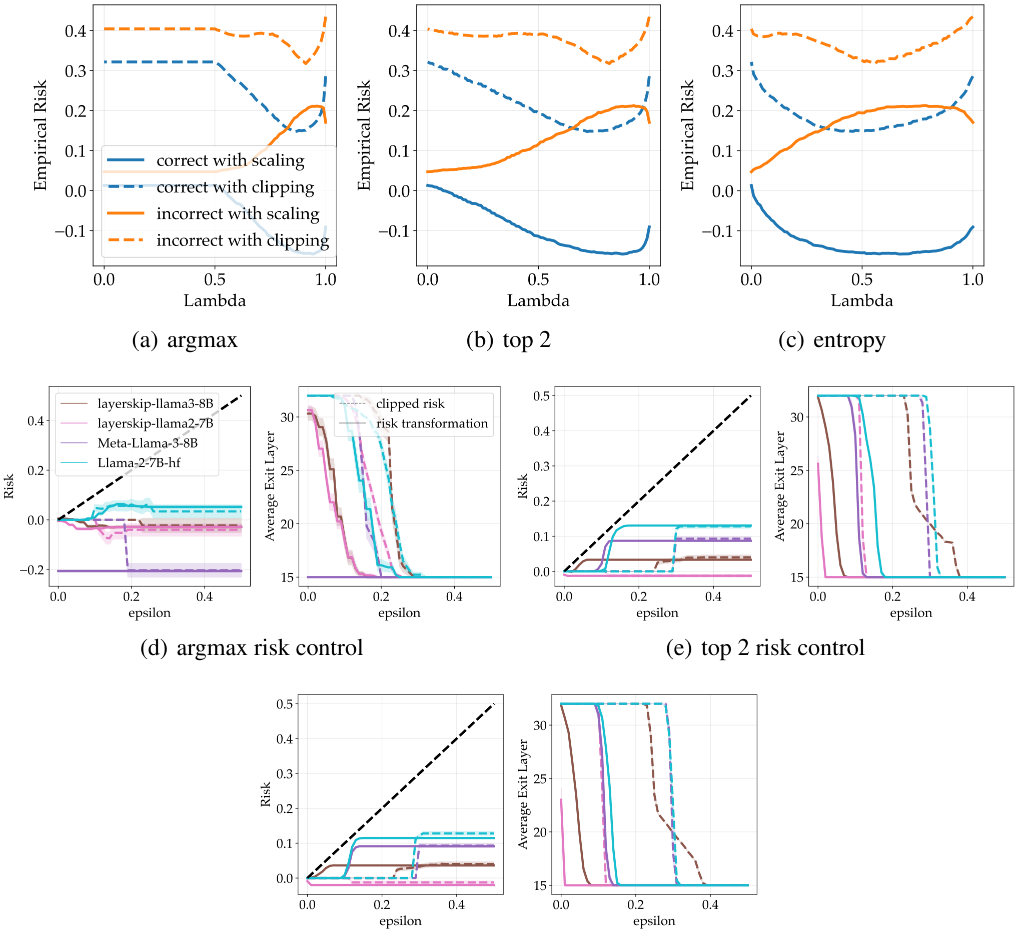
We test different ways of measuring confidence in the model's prediction to evaluate whether this impacts our risk control approach. The three measures we use are as follows:

- *argmax*: Taking the simple argmax of the logits after applying softmax.
- *top 2*: Take the difference between the top 2 largest values of the logits after applying softmax.
- *entropy*: Compute the entropy over the logits post-softmax.

We present results on the TweetEval Hate dataset in Fig.19. We find that though the choice of confidence measure will affect the level of risk for specific λ values, there is no significant impact on our risk-control approach, as it works under all scenarios. We choose *argmax* as our confidence measure for all experiments in the paper, as this is the most common approach taken in other work.

J.3 First Exit

We find that the models are frequently overconfident in the wrong answers in earlier layers. Through detailed examination of the models' generated text from intermediate layers, we also find that the quality in very early layers is extremely low and gradually improves through the layers. This means that risk control based on model confidence will provide trivial results when we early-exit from anywhere in the model; we cannot have a confidence-based λ threshold which allows us to early-exit while preserving performance. We address this by applying our risk-control approach only on



(a) argmax

(b) top 2

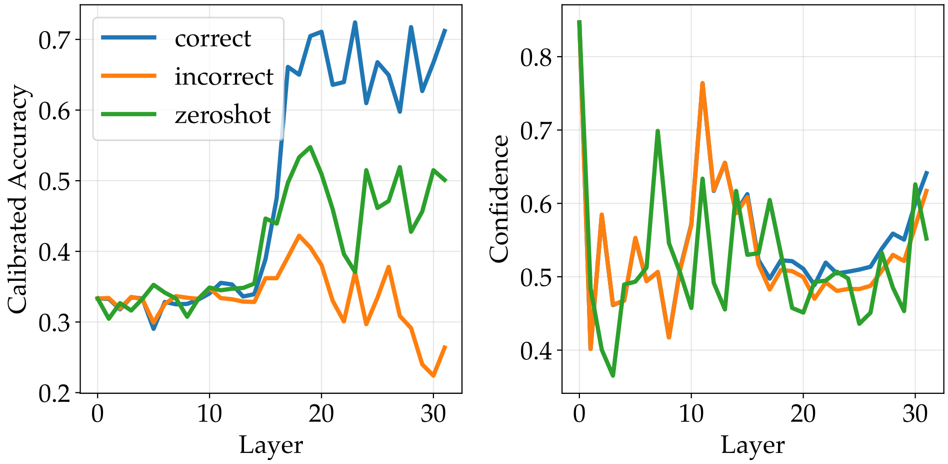
(c) entropy

(d) argmax risk control

(e) top 2 risk control

(f) entropy risk control

Figure 19: The top row shows λ vs risk for different measures of confidence for the TweetEval-Hate dataset with the LayerSkip LLaMA 3 model, showing that different measures of confidence can impact the way that λ-thresholds on confidence affect risk (both with loss-clipping and true relative loss). The risk control plots for all models on TweetEval-Hate are shown in the last three plots; there is no significant impact on the choice of confidence measure on our risk-control approach, which works equally well on all three, without any significant differences in efficiency gains or risk level across models.



(a) Accuracy (b) Confidence

Figure 20: We empirically find that the models become less overconfident and more accurate in the last half of the layers (from roughly layer 16 of 32), a finding which is consistent across all four of our models and all eight datasets. This motivates our choice to apply our early-exit risk control procedure on only the *last half* of the layers in the model. Results are shown here for FinancialPhrasebank; we found similar results across all models and datasets.

the last half of the layers, meaning that the earliest possible exit for our 32-layer models is layer 16. Empirical results justifying this choice are shown in Fig.20.

J.4 True Dataset Labels

We show that the model has already memorized existing datasets during pre-training; results displayed in Fig.21. These results are also confirmed in prior work (Pan et al., 2023; Fang et al., 2025). This motivates our approach of transforming the task to a format that is equivalent to, but distinct from, their original form by assining arbitrary “dummy” labels for each label of the dataset.

K Compute Resources

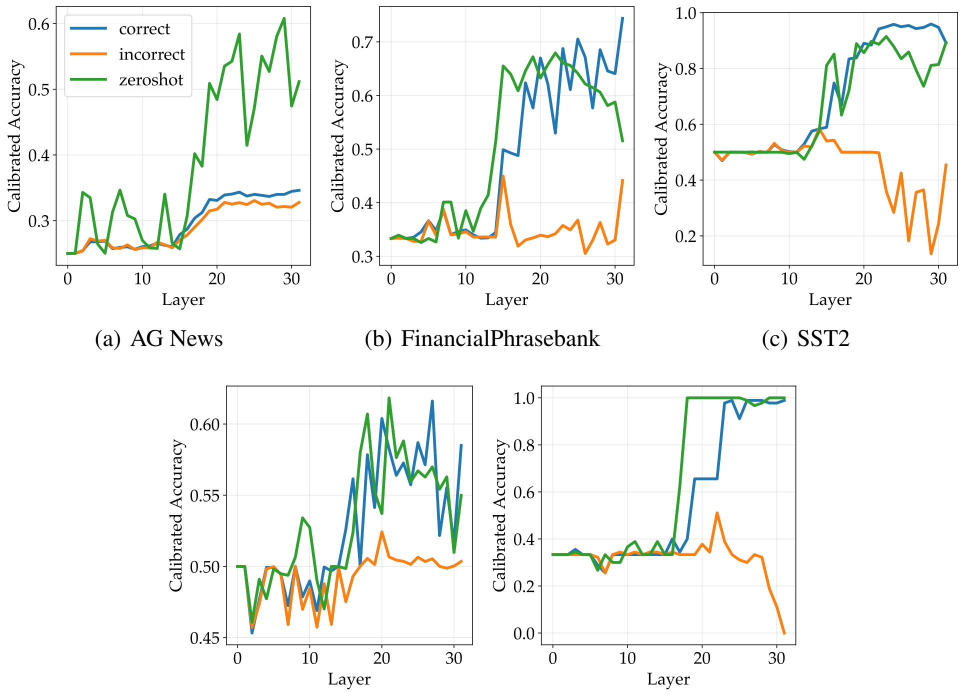
To run all experiments with language models, we used 4 A100 GPUs on the Johns Hopkins DSAI compute cluster. Plots and risk-control were executed locally.

L Prompt Format

The prompt format is presented below for the AG News dataset. The same format is used across all datasets, with the only difference being the list of possible labels. {text} indicates the input on which the model is asked to make a prediction. {demo i} and {label i} indicate the text-label pairs that constitute the in-context examples (where the label either corresponds to the true label or the substituted “incorrect” label).

List of “dummy” labels. We define a fixed substitution between the true labels of each dataset and the “dummy” labels we use in our prompts. In particular, for each true label, we substitute each label with one of the following words: river, stone, cloud, chair, table, grass. We find that there is not a significant effect of using any particular substitution, so we simply randomly select label-substitute pairs.

Zero-Shot Prompt. Your job is to classify the topic of a news article given a description of the article. The possible topics are: world, sports, business, science/technology. Output only the topic



(a) AG News

(b) FinancialPhrasebank

(c) SST2

(d) TweetEval-Hate

(e) Unnatural

Figure 21: Here, we show the accuracy of predictions from each layer of the model. We show that our models have already memorized many of our datasets during pre-training, as shown by the fact that zero-shot will often do as good or better than the model even given correct in-context demonstrations.

of the news article and nothing else. Do not provide chain of thought reasoning before your answer. Description: {text} Topic:

In-Context Demonstrations. Your job is to classify the topic of a news article given a description of the article. The possible topics are: world, sports, business, science/technology. Output only the topic of the news article and nothing else. Do not provide chain of thought reasoning before your answer. Below are a few examples of description-topic pairs. Description: {demo 1} Topic: {label 1} Description: {demo 2} Topic: {label 2} ... Description: {text} Topic:

Dummy Labels - Zero-Shot Prompt. Your job is to classify the topic of a news article given a description of the article. Output river if the topic is world, stone if the topic is sports, cloud if the topic is business, and chair if the topic is science/technology. Output only the topic of the news article and nothing else. Do not provide chain of thought reasoning before your answer. Description: {text} Topic:

Dummy Labels - In-Context Demonstrations. Your job is to classify the topic of a news article given a description of the article. Output river if the topic is world, stone if the topic is sports, cloud if the topic is business, and chair if the topic is science/technology. Output only the topic of the news article and nothing else. Do not provide chain of thought reasoning before your answer. Below are a few examples of description-topic pairs. Description: {demo 1} Topic: {label 1} Description: {demo 2} Topic: {label 2} ... Description: {text} Topic: