Empirical Analysis of PAC Learnability in In-Context Learning

Yutong He
Heidelberg University, Computational Linguistics
yutong.he@stud.uni-heidelberg.de

Abstract

This project empirically investigates the theoretical framework of PAC learnability in in-context learning (ICL). Using the TREC question classification benchmark, we estimate the independence constant (c_1), prompt likelihood constant (c_2), and evaluate the concentration property (Lemma 1) and margin inequality (Theorem 1) under varying numbers of in-context examples k. Results show that with increasing k, contextual independence strengthens, prompt likelihood decreases but remains positive, and task representations become more identifiable, confirming key theoretical predictions. Project code available at GitHub.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in performing novel tasks simply by conditioning on a few examples within their input, a phenomenon known as In-Context Learning (ICL). Despite its empirical success, the theoretical understanding of ICL, especially the reason and the conditions of its feasibility, remains an open question.

The recent paper “`The Learnability of In-Context Learning`”(Wies et al., 2023) introduces a formal PAC (Probably Approximately Correct) learnability framework for analyzing ICL. Within this framework, ICL is viewed as an implicit learning process where a pre-trained model approximates a hypothesis class among all latent downstream tasks via conditioning, rather than gradient-based updates(von Oswald et al., 2023). The paper provides several key theoretical results, including Lemma 1 and Theorem 1, which establish the conditions under which in-context learning can generalize from few-shot prompts to unseen examples.

However, the paper primarily focuses on the theoretical guarantees and leaves open the question of how these assumptions and theorems manifest in real models. For example, whether the independence and concentration assumptions hold in practical LLMs, and how the generalization bound in Theorem 1 relates to the behavior of modern transformer-based architectures, remains underexplored.

This project aims to bridge this gap by providing empirical evidence supporting the theoretical claims in the paper. Specifically, we focus on experimentally validating Lemma 1 2, which concerns the concentration of task likelihoods under increasing context size), as well as Theorem 1 2, which relates the model's context representations to task identifiability and margin-based generalization. Using Phi-3(Microsoft / Hugging Face, 2024)[2] as the pre-trained model and the TREC question classification dataset(Li and Roth, 2002), we systematically test the assumptions, estimate key constants (c_1 as the independence between few-shot samples in a prompt, c_2 as the probability of prompts produced by the pre-trained model), and visualize how the concentration of task likelihoods, task identifiability, and margin success evolve with the number of in-context examples k.

By connecting the theoretical PAC learnability framework with empirical results, this work aims to provide a clearer picture of how in-context learning emerges and generalizes in real LLMs.

2 Theoretical Foundations

To establish the connection between the theoretical framework and our empirical experiments, we first restate the core assumptions and results from the paper(Wies et al., 2023). These definitions form the foundation of our empirical validation of Lemma 1 and Theorem 1.

Assumption 2 (Approximate Independence). There exists a constant $0 < c_1 \leq 1$ such that for

[1] `https://github.com/WitchPuff/ICL_PAC_Learnability`

[2] `https://huggingface.co/microsoft/Phi-3-mini-4k-instruct`

any two strings $s_1, s_2 \in \Sigma^\star$ and any concept ϕ, the following holds:

$$c_1 \leq \frac{P_\phi(s_1 \oplus \text{``}\backslash n'') \cdot P_\phi(s_2)}{P_\phi(s_1 \oplus \text{``}\backslash n'' \oplus s_2)} \leq 1.$$

Intuitively, this assumption requires that two successive strings concatenated by a delimiter token ("\n") are approximately independent according to the distribution P_ϕ. When s_1 and s_2 are exactly independent, the ratio equals 1. The constant c_1 thus quantifies the deviation from perfect independence. This assumption allows the use of concentration inequalities on the likelihood of in-context prompts and ensures that the prompt acts as a meaningful reweighting of the prior mixture components rather than being ignored.

Assumption 3 (Positive Token Likelihood). There exists a constant $c_2 > 0$ such that for any string $s \in \Sigma^\star$, any token $\sigma \in \Sigma$, and any concept ϕ,

$$P_\phi(\sigma \mid s) > c_2.$$

This lower bound prevents zero-likelihood issues caused by unnatural concatenations of input–output pairs in prompts. Without this assumption, the likelihood of a prompt p could be zero, rendering in-context predictions meaningless. We further assume that the prior distribution over mixture components is strictly positive, that every component has non-zero probability of appearing in the pretraining distribution.

Lemma 1 (Concentration of Task Likelihood Ratios). Let $\mathcal{D}$ be a pretraining distribution satisfying Assumptions 2 and 3, and let $\phi^\star$ be a downstream task mixture component such that the minimal Kullback–Leibler divergence Δ_{KL} between $\phi^\star$ and all other components satisfies:

$$\Delta_{KL} > 8 \log \frac{1}{c_1 c_2}.$$

Then there exists a function $m_{\tilde{\mathcal{D}}} : (0,1)^2 \to N$ such that for any $\varepsilon, \delta > 0$ and any $\phi \neq \phi^\star$, if the number of in-context examples $k \geq m_{\tilde{\mathcal{D}}}(\varepsilon, \delta)$, we have:

$$\Pr_{p \sim (P_{\phi^\star})^k} [\frac{P_\phi(p)}{P_{\phi^\star}(p)} < \varepsilon] \geq 1 - \delta,$$

where the probability is taken over the random sampling of the k in-context examples. In other words, as k increases, the likelihood ratio between incorrect and correct components concentrates exponentially towards zero, at a rate depending on Δ_{KL}. The bound also holds when the prompt labels are randomly flipped, and $m_{\tilde{\mathcal{D}}}$ can be chosen to be polynomial in $\log(1/\delta)$, $\log(1/\varepsilon)$, $\log(1/(c_1c_2))$, $1/\Delta_{KL}$, and the sequence length T.

Theorem 1 (Margin-Based Generalization in In-Context Learning). Let $\mathcal{D}$ and $\tilde{\mathcal{D}}$ be a pair of pretraining and downstream task distributions satisfying Assumption 4 and all assumptions of Lemma 1. Then there exists $m_{\tilde{\mathcal{D}}} : (0,1)^2 \to N$ such that for every test example x and two label candidates $y, \tilde{y}$ with positive margin $\Delta(x, y, \tilde{y}) > 0$, and for any $\delta > 0$, if the number of in-context examples $k \geq m_{\tilde{\mathcal{D}}}(\Delta(x, y, \tilde{y})/2, \delta)$, we have:

$$\Pr\left[\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + c_1^2 - 1\right] \geq 1-\delta.$$

This theorem states that, given enough in-context examples, the model's predictive margin for the correct label improves beyond half of the ground-truth margin, up to an additive constant depending on c_1. It also holds under random label flipping, and $m_{\tilde{\mathcal{D}}}$ grows polynomially in $\log(1/\delta)$, $\log(1/\varepsilon)$, $\log(1/(c_1c_2c_3))$, $1/\Delta_{KL}$, and T.

Together, they form the theoretical foundation of our experimental study: Lemma 1 motivates our concentration and task identification experiments, while Theorem 1 connects the learnability of context representations with margin-based generalization in real models.

3 Empirical Pipeline

To connect the theoretical results of Lemma 1 and Theorem 1 with measurable model behavior, we design a set of controlled experiments that investigate how the number of in-context examples k influences several key metrics derived from the PAC learnability framework for in-context learning (ICL). Rather than attempting to compute the exact theoretical constants and precise constraints, our aim is to observe whether the predicted relationships among k, independence (c_1), prompt likelihood (c_2), concentration $\frac{P_\phi(p)}{P_{\phi^\star}(p)}$, margin$\Delta(p, x, y, \tilde{y})$, and task identifiability emerge empirically in a real pretrained language model.

3.1 Estimates of Assumptions

We begin by estimating the two key constants introduced in the theoretical framework: c_1 from

Assumption 2 and c_2 from Assumption 3. Assumption 4, which requires a positive prior probability for each downstream concept, is not directly estimated but empirically ensured by selecting test samples with positive zero-shot margins, indicating that the corresponding task has non-negligible prior support under the pretraining distribution.

- **Assumption 2 (Independence constant c_1).** We empirically approximate c_1 by measuring how close two concatenated text segments are to being independent under the model's likelihood. For sampled text pairs (s_1, s_2), we compute:

  $$\log c_1 = \log P(s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”}) + \log P(s_2) - \log P(s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”} \oplus s_2),$$

  and take its average as the estimate of $\log c_1$.

  During experiments, we observed that the independence among concatenated samples decreases substantially as both the number of few-shot examples and total text length increase. This degradation mainly stems from the highly similar structural patterns across examples. For instance, each sample shares identical prompt templates such as “`Q:`” and “`A:`”, leading to repetitive surface forms and stronger dependencies within prompt.

  To mitigate this effect, we introduce random separator text between examples when constructing prompts.

  Additionally, we report a per-token normalized version to control for prompt length effects:

  $$\log c_1^{(\texttt{per-token})} = \frac{\log P(s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”})}{|s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”}|} + \frac{\log P(s_2)}{|s_2|} - \frac{\log P(s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”} \oplus s_2)}{|s_1 \oplus \text{“}\texttt{\textbackslash n}\text{”} \oplus s_2|},$$

  where $|s|$ denotes the number of tokens in the text s.

- **Assumption 3 (Minimum token likelihood c_2).** We estimate c_2 by measuring the average log-likelihood assigned by the model to few-shot prompts sampled from the downstream task $\phi^\star$. For each k-shot prompt p, we compute its total log-likelihood:

  $$\log c_2 = \log P_{\phi^\star}(p),$$

  Moreoever, for reference, we also report a normalized per-token estimate:

  $$\log c_2^{(\texttt{per-token})} = \frac{1}{|p|} \log P_{\phi^\star}(p),$$

  where $|p|$ denotes the number of tokens in the prompt.

- **Assumption 4 (Positive prior probability c_3).** Assumption 4 requires that the prior probability of each downstream concept under the pretraining distribution satisfies $P_D(\phi) > c_3 > 0$. In practice, we ensure this condition by selecting only test samples whose zero-shot margin is positive:

  $$\Delta(x, y, \tilde{y}) = \log P(y|x) - \log P(\tilde{y}|x) > 0.$$

  A positive zero-shot margin indicates that the model already assigns higher prior probability to the correct label before observing any in-context examples, implying that the corresponding concept $\phi^\star$ has non-negligible prior support under $P_D(\phi)$.

3.2 Experimental Design

The empirical pipeline is designed to mirror the logical structure of the theoretical framework. Each theoretical component (Assumptions 2–4, Lemma 1, and Theorem 1) is instantiated through corresponding measurable quantities and experiments.

1. **Estimating constants c_1 and c_2.** We begin by empirically approximating the independence constant c_1 and the minimum token likelihood c_2 under Assumptions 2–3. Specifically, c_1 measures how independent two concatenated text segments are under the model's likelihood, while c_2 quantifies how far the model assigns non-zero probability to naturally constructed few-shot prompts. Both quantities are computed across multiple k-shot settings to observe how model stability changes with prompt length.

2. **Evaluating Lemma 1 (Concentration and Task Identifiability).** We empirically test the concentration property predicted by Lemma 1: as the number of in-context examples k increases, the likelihood ratio $\frac{P_\phi(p)}{P_{\phi^\star}(p)}$ for incorrect task components $\phi \neq \phi^\star$ should fall below ε with high probability. For each downstream task $\phi^\star$, we compute the conditional log-likelihoods $\log P_{\phi^\star}(p)$ and $\log P_\phi(p)$ for all other candidate tasks ϕ on the same k-shot prompt p. A trial is counted as “concentrated” if $(\log P_\phi(p) - \log P_{\phi^\star}(p)) < \log \varepsilon$ for all

$\phi \neq \phi^\star$. The proportion of concentrated trials over R repetitions is reported as the empirical concentration probability.

To quantify the divergence between the true task distribution $P_{\phi^\star}$ and alternative hypotheses P_ϕ, we also compute the average Kullback–Leibler (KL) divergence:

$$D_{\mathrm{KL}}(P_{\phi^\star} \parallel P_\phi) = E_{p \sim P_{\phi^\star}}[\log P_{\phi^\star}(p) - \log P_\phi(p)],$$

which measures how distinguishable the true task $\phi^\star$ is from competing tasks in the model's induced likelihood space.

In addition, we analyze the task identifiability of the model's hidden representation $\phi(context)$ by probe analysis, where we train a linear probe on the hidden states of different task prompts. This probe assesses how separable the task representations become as k increases, bridging the theoretical notion of identifiability with observable representational structure in the model.

3. **Evaluating Theorem 1 (Margin Inequality).** We empirically verify the margin improvement property predicted by Theorem 1. In the theoretical formulation, the margin is defined in the probability space as:

$$\Delta(p, x, y, \tilde{y}) = P(y \mid p, x) - P(\tilde{y} \mid p, x),$$

and the guarantee states that:

$$\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + c_1^2 - 1$$

with a probability $> 1 - \delta$, where c_1 are the lower-bound constants from Assumptions 2.

In practice, we operate in the log-likelihood space since language models directly output $\log P(\cdot)$ rather than $P(\cdot)$. Accordingly, the contextual margin is computed as:

$$\Delta(p, x, y, \tilde{y}) = \log P(y \mid p, x) - \log P(\tilde{y} \mid p, x),$$

and the zero-shot margin as:

$$\Delta(x, y, \tilde{y}) = \log P(y \mid x) - \log P(\tilde{y} \mid x).$$

To remain consistent with the theoretical inequality, we translate the multiplicative constants into additive log form and test:

$$\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + 2\log(c_1),$$

The constant $\theta = 2\log(c_1)$ therefore serves as the empirical log-space equivalent of the theoretical multiplicative term $(c_1^2 - 1)$ in probability space. In practice, we adopt the length-normalized estimate $\theta = 2\log c_1^{(per-token)} \times |p|$, where $|p|$ denotes the number of tokens in the current prompt, to account for prompt-length scaling.

For each k-shot configuration, we report the proportion of test instances for which the above inequality holds with probability greater than $(1 - \delta)$ (set to $\delta = 0.01$ in our experiments), serving as the empirical success probability of Theorem 1.

Consisting of these steps, the comprehensive empirical pipeline should ground the PAC learnability framework of in-context learning in observable model behavior.

3.3 Key metrics.

We track the following quantities as the number of in-context examples k increases:

- $\log c_1$ **and** $\log c_2$**:** empirical estimates of the independence constant and the minimum prompt likelihood bound, corresponding to Assumptions 2–3.
- **Concentration probability:** $\Pr\left[\frac{P_\phi(p)}{P_{\phi^\star}(p)} < \varepsilon\right]$, measuring how likely incorrect task hypotheses $\phi \neq \phi^\star$ become less probable than a threshold ε as k grows.
- **KL divergence:** $D_{\mathrm{KL}}(P_{\phi^\star} \parallel P_\phi) = E_{p \sim P_{\phi^\star}}[\log P_{\phi^\star}(p) - \log P_\phi(p)]$, quantifying the separability between the true task distribution and alternative task likelihoods.
- **Probe accuracy:** the classification accuracy of a linear probe trained on hidden representations, serving as an indicator of task identifiability and representational separability.
- **Margin success rate:** $\Pr[\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + \theta]$, where $\theta = 2\log(c_1)$, indicating how often the contextual margin exceeds the theoretical lower bound predicted by Theorem 1.

3.4 Expected outcome.

If the theoretical framework accurately describes in-context learning dynamics, we expect the following empirical trends to emerge:

1. **Assumption 2 (Independence).** As k increases, the estimated $\log c_1$ and $\log c_1^{(per-token)}$ rises and approaches 0, while its per-token normalized form becomes more stable, indicating that concatenated few-shot examples become increasingly independent under the model's likelihood.

2. **Assumption 3 (Non-zero likelihood).** The prompt likelihood constant c_2 is strictly positive, ensuring that every token in the prompt has a non-zero conditional probability. In log space, $\log c_2$ is finite (typically negative), preventing degenerate cases where the prompt probability vanishes.

3. **Lemma 1 (Concentration).** The concentration probability $\Pr[\frac{P_\phi(p)}{P_{\phi^\star}(p)} < \varepsilon]$ should monotonically increase with k, reflecting the exponential convergence of incorrect task likelihoods toward zero. Meanwhile, the average KL divergence $D_{\mathrm{KL}}(P_{\phi^\star} \parallel P_\phi)$ is expected to grow, confirming improved task separability.

4. **Task identifiability.** The probe accuracy on hidden representations should improve as k grows, suggesting that the model forms more distinct and linearly separable representations of task identity $\phi(context)$.

5. **Theorem 1 (Margin improvement).** The contextual margin success probability $\Pr[\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + 2\log(c_1)]$ should follow a similar increasing trend, indicating that larger context size enhances decision confidence consistent with the theoretical lower bound.

4 Experiments

4.1 Setup

Model. We use the `microsoft/Phi-3-mini-4k-instruct` model as the representative large language model, chosen for its strong few-shot reasoning ability and efficient 4k-token context window. All computations are performed in inference mode with half-precision (FP16) to ensure numerical stability and reproducibility.

Dataset. We adopt the **TREC** question classification benchmark as the downstream dataset. The benchmark consists of open-domain questions labeled with one of six coarse-grained semantic categories: `ABBR` (abbreviation), `DESC` (description or definition), `ENTY` (entity or object), `HUM` (human or person), `LOC` (location or place), and `NUM` (numerical expression). The original task requires predicting the semantic type of a given natural-language question, such as determining whether "`Who wrote Hamlet?`" belongs to the `HUM` category.

We prepare six binary classification downstream subtasks from the TREC benchmark: `HUM vs. LOC`, `NUM vs. DESC`, `HUM vs. ENTY`, `HUM vs. NUM`, `LOC vs. NUM`, and `ABBR vs. ENTY`. Each subtask represents a distinct latent concept $\phi^\star$ sampled from the pretraining distribution mixture.

Each subtask is represented as a `Task` object containing: (i) a list of `(question, label)` pairs, (ii) a verbalizer mapping label IDs to natural language labels (e.g., `1` → "`human`", `0` → "`location`").

For each binary task, k-shot samples are drawn with balanced class sampling and optional random label flipping.

Prompt construction. Each few-shot example is formatted into a unified textual prompt combining an instruction prefix and a sequence of question–answer pairs. For every downstream task, we initialize the prompt with a general instruction of all tasks that omits the label options:

```
"Decide the category of each question.
      Answer with one of the provided
            examples if exist.",
```

so that the model must infer the task structure from the examples themselves.

Each example in the prompt follows a `question-answer` format:

```
Q: <question>  A: <label>.
```

During concatenation, we prepend each example with a random separator (e.g., `### SEP_XQZT ###`) and an `Example` header:

```
### SEP_XQZT ### Example:### SEP_XQZT ###
```

to mimic natural discourse boundaries and mitigate over-structuring caused by repetitive prefixes (`Q:`, `A:`). This stochastic segmentation helps preserve approximate independence between few-shot samples while maintaining consistent syntactic framing across all prompts.

All experiments are conducted using the `microsoft/Phi-3-mini-4k-instruct` language model, evaluated under controlled few-shot configurations. Inference is performed with half-precision (`torch.float16`) on a single NVIDIA RTX 4090 GPU.

Few-shot configuration. We evaluate across a range of context sizes $k \in \{0, 1, 2, 3, 4, 6, 10, 14, 18\}$. For each k-shot setting, we sample balanced examples from both classes and construct prompts using the question–answer format described in Section 4.1. To test robustness, we optionally introduce random label flipping with probability `FLIP` $\in \{0, 0.5\}$.

Experimental repetitions. For stable estimates of statistical quantities, we repeat the sampling process multiple times:

- **Independence estimation (Assumption 2):** 100 random trials per k with instruction-blind prompts; the mean log ratio is reported as $\log c_1$.
- **Concentration experiment (Lemma 1):** For each k-shot configuration, the concentration probability is estimated by averaging over $R = 100$ independently sampled few-shot prompts, following a Monte Carlo approximation scheme. The concentration threshold is set to $\varepsilon = 10^{-2}$.
- **Margin inequality (Theorem 1):** Each test runs with $R = 40$ random draws and up to 80 test instances satisfying the positive zero-shot margin condition $\Delta(x, y, \tilde{y}) > 0$.

Probing setup. To examine the identifiability of the in-context task representation $\phi(context)$, we conduct a linear probing analysis on the hidden states extracted from the model. For each k-shot configuration, we sample 100 prompts per downstream task, encode them using the model, and obtain hidden representations from the final layer (mean-pooled across tokens).

A logistic regression classifier (`max_iter=2000`, `C=0.1`) is trained to predict the task identity among all candidate subtasks, and evaluated on a held-out split (80/20 train–test ratio). The resulting probe accuracy quantifies how linearly separable the task representations are as k increases. Additionally, we compute the Fisher ratio between within-class and between-class variance of representations as a complementary measure of class separability.

All experiments are logged and visualized via `Weights & Biases`, tracking $\log c_1$, $\log c_2$, KL divergence, concentration probability, margin success rate, and probe accuracy across increasing k.

4.2 Results

Effect of random label flipping. Throughout all experiments, we compare two conditions: `qa_0` (no label flipping) and `qa_0.5` (randomly flipping labels with probability 0.5). This manipulation tests the robustness of the model's in-context learning dynamics under label noise and measures how much the few-shot inference depends on the internal consistency of examples.

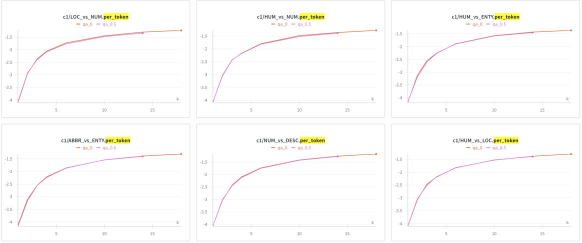

Figure 1: Per-token $\log c_1$ estimates as a function of k.

Independence constant c_1 (Assumption 2). Per-token $\log c_1$ steadily increases from approximately -4 toward -1.5 as k grows, for both `qa_0` and `qa_0.5` (Fig. 1). This suggests that concatenated segments become more independent on average when more diverse few-shot contexts are observed. The two curves nearly overlap, implying that random label flipping does not strongly affect independence, as it mainly alters label semantics rather than textual structure.

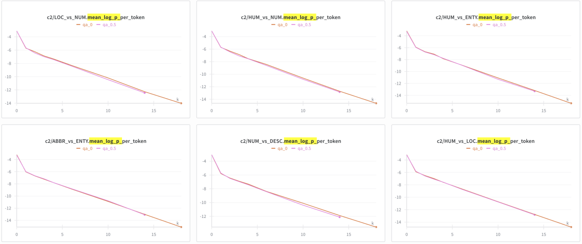

Figure 2: Per-token $\log c_2$ (average prompt likelihood) as a function of k.

Prompt likelihood constant c_2 (Assumption 3). The mean per-token log-likelihood $\log c_2$ decreases almost linearly with k, from around -4 to -14 across tasks (Fig. 2). This shows that as prompts become longer and more varied, their overall probability naturally decreases. Both `qa_0` and `qa_0.5` exhibit similar trajectories, indicating that label noise does not change the overall scaling of model likelihoods across prompt lengths. Despite this decay, the total prompt likelihood c_2 remains strictly positive, thus satisfying Assumption 3 and indicating that the current prompts are still plausible under the model's distribution.

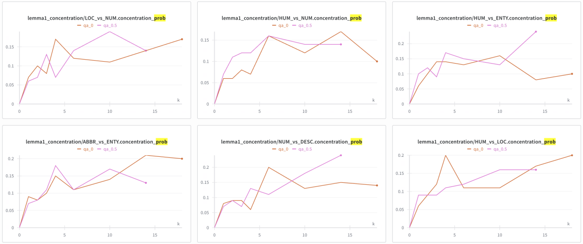

Figure 3: Concentration probability $\Pr[\frac{P_\phi(p)}{P_{\phi^\star}(p)} < \varepsilon]$ across subtasks as k increases, under different label-flip conditions.

Concentration probability (Lemma 1). Across all subtasks, the concentration probability rises gradually with k and remains within 0.05$\sim$0.20 (Fig. 3). Random label flipping (p=0.5) slightly reduces the overall concentration rate and increases fluctuations, especially for `HUM vs. LOC` and `HUM vs. NUM`, indicating that inconsistent labels introduce variance in the model's implicit task inference. Nevertheless, the general upward trend with k is preserved, supporting the theoretical prediction that $\frac{P_\phi(p)}{P_{\phi^\star}(p)} \to 0$ as more in-context evidence accumulates.

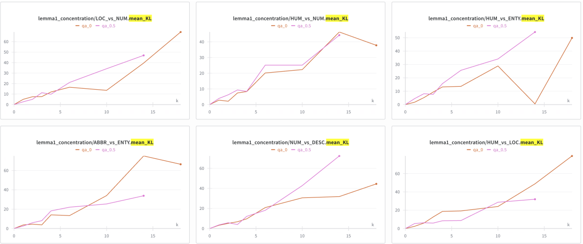

Figure 4: Mean log-likelihood difference $E[\log P_{\phi^\star}(p) - \log P_\phi(p)]$ across k.

Log-likelihood divergence. The mean log-likelihood difference $E[\log P_{\phi^\star}(p) - \log P_\phi(p)]$ monotonically increases with k across all subtasks (reaching up to $\sim$60 on average), indicating a growing divergence between the true and false task-induced distributions. We refer to this quantity as a KL-like divergence, since it captures a directional separation in log-probability space similar with the Kullback–Leibler divergence, though computed directly from empirical prompt likelihoods. Label flipping weakens this trend, showing that random label noise reduces the model's ability to distinguish between tasks. Notably, the `HUM vs. NUM` and `HUM vs. ENTY` subtasks show relatively lower divergence values. This can be due to their higher semantic overlap, as both involve human-related or categorical entities, making their prompts less distinguishable in likelihood space.

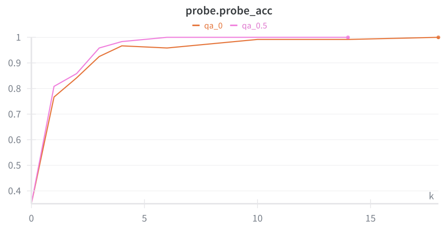

Figure 5: Linear probe accuracy of task representations across k.

Task identifiability. Linear probing on hidden states shows a strong monotonic improvement from 0.35 at k=0 to nearly 1.0 at k≥10, demonstrating that the model's internal representations of task identity become almost perfectly linearly separable as context grows. The `qa_0.5` condition slightly lags behind `qa_0` for small k, but converges at large k, suggesting that label noise primarily delays, but does not prevent the emergence of distinct task representations.

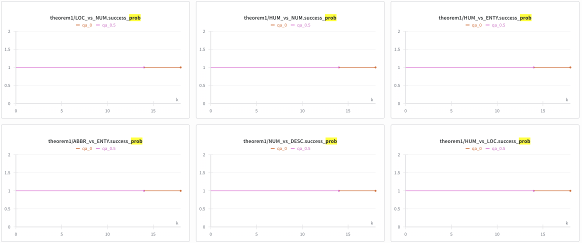

Figure 6: Margin success probability (Theorem 1) ($\delta = 0.01$).

Margin success probability (Theorem 1). Across all subtasks, the measured success probability $\Pr[success] > 1 - \delta$ ($\delta = 0.01$) remains consistently at 1.0 for every k value. This means that the empirical inequality

$$\Delta(p, x, y, \tilde{y}) > \frac{1}{2}\Delta(x, y, \tilde{y}) + \theta$$

is always satisfied given $\delta = 0.01$. This confirms that our global $\theta = 2 \log c_1^{(per-token)} \times |p|$, where

$|p|$ denotes the number of tokens in the current prompt, acts as a highly conservative lower bound. Label flipping has no visible effect because the threshold is already too loose.

5 Limitations

While the proposed empirical framework successfully reproduces the theoretical tendencies predicted by the PAC learnability analysis, several limitations remain:

- **Approximate bounds.** Our current estimates of c_1, c_2, and θ capture the overall trends rather than the exact theoretical bounds. The concentration property (Lemma 1) is empirically validated, but the margin inequality (Theorem 1) yields uniformly high success rates ($\sim$1.0), suggesting that our current θ is overly loose. Tighter or adaptive bounds (e.g., per-sequence or quantile-based) are required for finer evaluation.

- **Computational constraints.** The number of in-context examples k is limited by GPU memory and the model's context window, which prevents testing larger k where theoretical guarantees would be more evident.

- **Parameter sensitivity.** The experiments are conducted under fixed ε and margin thresholds (ε=10^{-2}, $\delta = 10^{-2}$). A more systematic sweep across different thresholds could provide a clearer view of convergence behavior and stability.

- **Prompt configuration.** We use a consistent question–answer format and random separators, but have not yet examined other prompting styles, instructions, or domain-specific phrasing. Future work may study how prompt design affects independence, likelihood concentration, and task identifiability.

6 Conclusion

This work provides an empirical bridge between the PAC learnability framework of in-context learning and the observable behavior of large language models. By estimating the theoretical constants (c_1, c_2, θ) and testing the concentration and margin inequalities (Lemma 1 and Theorem 1), we demonstrate that core theoretical predictions can indeed be observed in practice.

Our results confirm that the independence and likelihood assumptions (Assumptions 2–3) hold approximately in real models: as the number of in-context examples k increases, textual independence strengthens, task likelihood concentration improves, and task representations become linearly separable. These trends collectively validate the theoretical view that in-context learning operates as implicit Bayesian inference over latent task representations.

However, the uniformly high margin success rates indicate that our current bounds are overly conservative. Future work should refine the empirical estimation of θ (e.g., using sequence-specific or quantile-based scaling), explore larger k values under extended context windows, and diversify prompt formats to test generality across domains.

Overall, this study shows that even with approximate assumptions and bounded contexts, language models exhibit measurable statistical structures that align with the PAC learnability theory, offering a first step toward connecting theoretical guarantees with practical ICL phenomena.

References

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1–7, Taipei, Taiwan. Association for Computational Linguistics.

Microsoft / Hugging Face. 2024. Phi-3 mini 4k instruct. https://huggingface.co/microsoft/Phi-3-mini-4k-instruct. Accessed: YYYY-MM-DD.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 36637–36651. Presented at NeurIPS 2023.