Training In-Context and In-Weights Mixtures Via Contrastive Context Sampling

Deeptanshu Malu
deeptanshumalu@cse.iitb.ac.in
IIT Bombay

Deevyanshu Malu
deevyanshumalu@cse.iitb.ac.in
IIT Bombay

Aditya Nemiwal
adityanemiwal@gmail.com
IIT Bombay

Sunita Sarawagi
sunita@iitb.ac.in
IIT Bombay

Abstract

We investigate training strategies that co-develop in-context learning (ICL) and in-weights learning (IWL), and the ability to switch between them based on context relevance. Although current LLMs exhibit both modes, standard task-specific fine-tuning often erodes ICL, motivating IC-Train— fine-tuning with in-context examples. Prior work has shown, largely via synthetic experiments, that emergence of ICL after IC-Train depends on factors such as task diversity and training duration.

In this paper, our focus is fine-tuning for continuous adaptation with ICL on a fixed task, where task diversity cannot be synthetically controlled. In such scenarios, we show that the similarity structure between target inputs and context examples plays an important role. Random context leads to loss of ICL and IWL dominance, while only similar examples in context causes ICL to degenerate to copying labels without regard to relevance. To address this, we propose a simple Contrastive-Context which enforces two types of contrasts: (1) mix of similar and random examples within a context to evolve a correct form of ICL, and (2) varying grades of similarity across contexts to evolve ICL-IWL mixtures. We present insights on the importance of such contrast with theoretical analysis of a minimal model. We validate with extensive empirical evaluation on four LLMs and several tasks. Diagnostic probes confirm that contrasted contexts yield stable ICL-IWL mixtures, avoiding collapse into pure ICL, IWL, or uncontextualized blind copying.

1. Introduction

Our goal is to train models on tasks with limited labeled data so they can continuously incorporate feedback in the form of new examples at test time, without further training. Such adaptation should boost accuracy on inputs resembling the new examples, while retaining generalization to inputs without close neighbors. In-context learning (ICL) offers a natural mechanism for flexibly absorbing labeled data by placing test inputs alongside labeled examples. Pre-trained LLMs (Brown et al., 2020) show strong ICL ability, yet prior studies (Alves et al., 2023; Duan et al., 2024; Wang et al., 2024; Goyal et al., 2025), as well as our experiments, find that fine-tuning often diminishes it. A natural alternative is to fine-tune in ICL mode (IC-Train) (Min et al., 2022; Alves et al., 2023; Bornschein et al., 2025), but this is challenging due to competition with in-weights learning (IWL). Moreover, ICL emergence is transient, shaped by factors like training steps and task diversity (Chan et al., 2024; Singh et al., 2025; Nguyen & Reddy, 2025; Wurgaft et al., 2025; Ku et al., 2025). Most studies explore synthetic tasks where diversity can be controlled. In contrast, we study a stable real task, where ICL is needed not to learn a new task but to flexibly and continually absorb examples. Here, models must operate as ICL–IWL mixtures—adapting through ICL when context is relevant, relying on IWL for examples without close neighbors, and deciding at inference which to use.

In this paper, we show that a critical but overlooked factor in training viable ICL-IWL mixtures is the similarity structure

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

between the target and in-context examples. Two standard defaults, viz, random sampling and selecting the k nearest examples, push models to opposite extremes:

1. With random examples as context, IWL dominates and prior ICL ability is lost.
2. With overly similar examples, IWL is diminished and the model learns to copy labels, ignoring input relevance, and engaging in a degenerate form of ICL.

To address this, we propose a training strategy Contrastive-Context that samples examples for the contexts so as to (i) span multiple similarity levels to the target and (ii) contrast in similarity among themselves. When highly similar examples are unavailable to create the contrast, we generate synthetic in-context examples via small perturbations of the target.

We evaluate Contrastive-Context on several tasks and models and show that Contrastive-Context consistently strengthens ICL while preserving IWL, outperforming both random and nearest-neighbor sampling under in-domain and out-of-domain evaluation. On the entire spectrum of target-context relatedness, Contrastive-Context provides gains over standard zero-shot fine-tuning, whereas other forms of IC-Train perform worse than zero-shot in at least one end of the context-target similarity spectrum.

We theoretically analyze a minimal two-layer transformer, and show that a stationary point of the self-attention layer is an optimal mixture of ICL and IWL when trained with contrasted contexts. To extend the understanding to LLMs, we design probes to disentangle ICL, IWL, and blind-copying, and observe the emergence of ICL-IWL mixtures when fine-tuning LLMs with Contrastive-Context, while settling for the corner points of ICL, IWL, or blind-copying for other context regimes. Our study establishes inter-example and example-target similarity as a key driver of whether fine-tuning enhances, erodes, or deforms ICL capabilities and mixes with in-weights learning.

Our contributions. (1) We identify inter-example similarity as a critical but underexplored factor shaping the emergence (or erosion) of ICL during fine-tuning with in–context examples. (2) We propose Contrastive-Context, a training strategy that samples examples across similarity levels within and across contexts and injects synthetic perturbations when needed. (3) We empirically evaluate the methods on four 1B–8B models over four machine translation tasks, eleven Text-to-SQL task, three multilingual semantic parsing tasks, and two synthetic alignment reasoning tasks. Our experiments show that Contrastive–Context consistently improves accuracy across diverse in–context configurations and domains. (4) We theoretically analyze the three context regimes on a minimal two-layer transformer to provide insights on the critical role of both inter and intra context contrasts for evolving ICL-IWL mixtures. (5) We empirically study emergence of different forms of learning on real models and tasks with three well-designed probes, showing how Contrastive-Context enables ICL–IWL mixtures without collapsing into one of pure IWL, pure ICL, or blind copying.

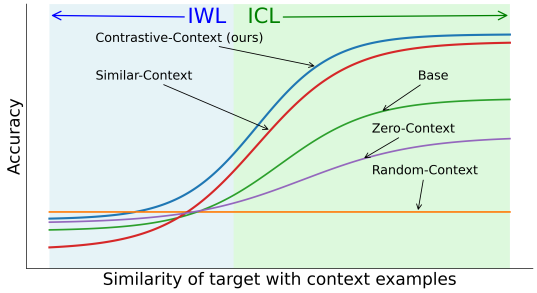


Figure 1. Visual summary of main findings of the paper towards our goal of in-weights learning (IWL) a task while retaining ICL for continuous adaptation with new examples. On the X-axis is target to context similarity — IWL important on left side and ICL on right side. Standard fine-tuning with zero in-context examples causes drop in ICL compared to base model. Fine-tuning with in-context example (IC-Train) sensitive to target-context similarity: random context leads to sharp drop in ICL, similar context does not develop IWL and instead is prone to blind copying. Our method Contrastive-Context retains both IWL and ICL and teaches model to switch between them.

2. IC-Train with Varying Target-Context Relatedness

Let P_θ denote a model to be trained on a task (e.g. low-resource translation). Typically, P_θ will be a pre-trained LLM. We are given a labeled dataset D of N pairs of inputs-outputs $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ drawn from the underlying joint distribution $P(X, Y)$ of this task. Our goal is to train P_θ using D so that its performance remains robust under these deployment settings: (1) As more labeled pairs are appended to D (e.g., via user feedback), the model's accuracy should improve on test inputs with highly similar cases in D, *without requiring further parameter updates.* (2) For test inputs lacking similar examples in D, accuracy should be no worse than that of a model trained in the standard zero-shot setting. A natural candidate to meet these goals is IC-Train. In standard fine-tuning, we maximize likelihood

on each example independently as $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D} \log P_\theta(\mathbf{y}|\mathbf{x})$. IC-Train samples $k+1$ labeled examples from D, place the first k pairs as in-context examples while maximizing likelihood of target as:

$$\mathbb{E}_{(\mathbf{x}_*,\mathbf{y}_*)\sim D}\, \mathbb{E}_{C=\{(\mathbf{x}_i,\mathbf{y}_i):1:k\}\sim D} \log P_\theta(\mathbf{y}_*|C,\mathbf{x}_*) \tag{1}$$

The above training explicitly trains the model to leverage in-context examples, thereby better preparing it to absorb any labeled data during deployment. This paper shows that a crucial but often overlooked factor in IC-Train is the inter-example similarity among the $k+1$ samples. To make this explicit, we rewrite the IC-Train objective in terms of the strategy $\texttt{Choose}(D,\mathbf{x})$, for choosing the k examples from D to accompany a target input $\mathbf{x}_*$:

$$\mathbb{E}_{(\mathbf{x}_*,\mathbf{y}_*)}\, \mathbb{E}_{C\sim \texttt{Choose}(D,\mathbf{x}_*)\}} \log P_\theta(\mathbf{y}_*|C,\mathbf{x}_*) \tag{2}$$

As we will show, $\texttt{Choose}(\mathbf{x}_*, D)$ critically influences whether IC-Train strengthens or erodes our desired robustness properties. Most prior work studied IC-Train with k context examples chosen at random independent of $\mathbf{x}_*$. We refer this method as **Random-Context**. A second strategy is to select as C the top-k examples from D most similar to $\mathbf{x}_*$, which we call **Similar-Context**. As we show in Section 4, both these strategies fail to meet the robustness goals outlined earlier, albeit for different reasons. Under Random-Context, the model relies heavily on in-weights learning and fails to benefit from new related in-context examples. Under Similar-Context, the model learns to exploit labels in context without adequately judging whether the corresponding input $\mathbf{x}_i$s are relevant to $\mathbf{x}_*$. This causes accuracy to suffer on test examples without close neighbors, and in extreme cases can cause the model to blindly copy labels from context. We therefore propose an alternative Contrastive-Context strategy.

Contrastive-Context: The key idea is to pair target instances with contexts that create contrast both within examples in a context and across contexts. We create contrast across batches by choosing a fraction $1-p$ with random context, and fraction $\frac{p}{2}$ with similar context. We use the remaining $\frac{p}{2}$ to create contrast within a context as follows: we sample one example weighed by similarity to the target, and the remaining randomly. When the labeled pool is small, there may not be examples close enough to the target during fine-tuning, causing the model to loose the capability of harnessing highly related context examples. We therefore augment ϵ fraction of the training instances with synthetic highly similar example by small perturbation of $\mathbf{x}_*$. For NLP tasks, the perturbation is obtained by getting a paraphrase of $\mathbf{x}_*$.

Context Type	Prompt Structure	Prob.
Random-Context	$x_1y_1\ x_2y_2\ \cdots x_ky_k\ x_*y_*$	$1-p$
Similar-Context	$x_1y_1\ x_2y_2\ \cdots x_ky_k\ x_*y_*$	$p/2$
Contrastive-Context	$x_1y_1\ x_2y_2\ \cdots x_ky_k\ x_*y_*$	$p/2$

Table 1. Here, x_*y_* denote the target input and output. x_iy_i are ICL examples where x_i is similar to x_*. A small fraction (ϵ) of the Random-Context and Similar-Context examples are augmented with highly similar perturbations of the the target example. Other colored pairs indicate randomly sampled examples.

We will show that a similar example juxtaposed with unrelated ones in a context forces the model to harness in-context labels only after establishing similarity to the target. And varying similarity levels across batches fosters balance of in-weights with in-context learning.

3. Theoretically Analyzing the Impact of Target-Context Relatedness

We provide insights for why different contexts types lead to different generalization across diverse example-context relatedness. Specifically, we will show that IC-Train with Random-Context reduces to IWL, with Similar-Context reduces to ICL or blind-copying, and with Contrastive-Context learns to switch between ICL and IWL based on context-target similarity. We design a minimal model of a two-layer transformer and theoretically analyze its stationary points under various training regimes on a simple regression task. In Section 4.1 we support the insights via empirical evidence on real tasks over pre-trained LLMs.

Let input covariates $\mathbf{x} \in \mathbf{R}^d$ be sampled from a distribution $P(X)$ where $\|\mathbf{x}\|_2 = 1$, with labels given by $y = f(\mathbf{x})$, where $f(\mathbf{x})$ denotes the function to be learned. Our goal is to learn f using a Transformer trained using IC-Train under the three context sampling regimes.

Figure 2 provides an overview. The transformer takes $k+2$ inputs: first k examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^k$ form the context C, an end-of-context marker EOC, and then the target input $\mathbf{x}_*$. We design a minimal two-layer Transformer that performs two steps. We defer detailed realization of the parameters of such a transformer in Appendix A.1. First, an attention block of the first layer aggregates the context examples at the $k+1$th position to obtain $\sum_{j=1}^k (\mathbf{x}_j, \mathbf{y}_j, 1)$. The feed-forward block of the first layer computes an in-weights estimate $\hat{f}(\mathbf{x}_*)$ from the query token's initial features. The second attention layer performs a selective computation at the query position governed by three learnable scalar parameters, θ_1, θ_2 and θ_3, which control the trade-off between proper ICL, blind context Copy, and IWL. The steps in this layer are:

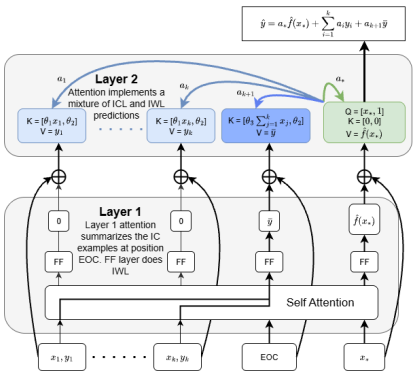


Figure 2. Schematic of a minimal two-layer transformer with a summarizer and in-weights learner ($\hat{f}$) in layer-1 and a three parameter second layer that implements the ICL-IWL mixtures.

1. Compute attention scores for context examples $(a_1, \ldots, a_k)$. The unscaled attention score s_i for each token $i \in [k]$ is determined by the similarity $\theta_1 \mathbf{x}_i^\top \mathbf{x}_*$ plus a global bias θ_2. The value vector for position i is the token's label, $\mathbf{y}_i$.
2. Compute a *EOC-attention* score (a_{k+1}) from the query to the EOC token. The unscaled attention score s_{k+1} is $\theta_3 \sum_{j=1}^k \mathbf{x}_j^\top \mathbf{x}_*$ plus θ_2. The value vector for position i is $\bar{\mathbf{y}} = \frac{1}{k}\sum_{j=1}^k \mathbf{y}_j$.
3. Compute a *self-attention* score (a_*) from the query to itself, with a fixed logit of 0. The value vector for the last position is the in-weights prediction, $\hat{f}(\mathbf{x}_*)$.
4. Compute the output of the attention layer at the query position. The prediction for a target $\mathbf{x}_*$ is thus:

$$\hat{\mathbf{y}} \;=\; a_* \hat{f}(\mathbf{x}_*) \;+\; \sum_{i=1}^k a_i \mathbf{y}_i \;+\; a_{k+1}\bar{\mathbf{y}}, \quad \text{where}$$

$$s_i = \exp\left(\theta_1 \mathbf{x}_i^\top \mathbf{x}_* + \theta_2\right) \qquad \forall i \in [k]$$

$$s_{k+1} = \exp\left(\theta_3 \sum_{j=1}^k \mathbf{x}_j^\top \mathbf{x}_* + \theta_2\right)$$

$$Z = 1 + \sum_{i=1}^{k+1} s_i$$

$$a_i = \frac{s_i}{Z} \quad \forall i \in [k+1], a_* = \frac{1}{Z}$$

We analyze the stationary points of the squared loss under different context selection strategies.

$$\mathcal{L}(\theta_1, \theta_2, \theta_3, \hat{f}) = \mathbb{E}_{\mathbf{x}_* \sim D, C \sim \texttt{Choose}(D, \mathbf{x}_*)}\left[(f(\mathbf{x}_*) - \hat{\mathbf{y}})^2\right]$$

Assumptions for Analysis Our analysis relies on the following assumptions.

- **[Lipschitz]** The ground-truth function f is L-Lipschitz: for all $\mathbf{x}, \mathbf{x}'$, $|f(\mathbf{x}) - f(\mathbf{x}')| \le L\|\mathbf{x} - \mathbf{x}'\|_2$.
- **[Context Regimes]** Fix small parameters $0 \le \delta_1, \delta_2 \ll 1$. For any target $\mathbf{x}_*$, the context selection procedure $\texttt{Choose}(D, \mathbf{x}_*)$ yields one of the following regimes:
  - (i) *Random-Context:* $\forall i \in [1, k],\ \mathbf{x}_i^\top \mathbf{x}_* \le \delta_1$.
  - (ii) *Similar-Context:* $\forall i,\ \mathbf{x}_i^\top \mathbf{x}_* \ge 1 - \delta_2$.
  - (iii) *One-Near-Context:* For one $j^\star \in [1, k]$ $\mathbf{x}_{j^\star}^\top \mathbf{x}_* \ge 1 - \delta_2$ and for all $i \ne j^\star$, $\mathbf{x}_i^\top \mathbf{x}_* \le \delta_1$.
  - (iv) *Random-Context + Similar-Context:* With probability 0.5, instance has a Random-Context, and with probability 0.5, it has a Similar-Context.
  - (v) *Contrastive-Context:* With probability p, a training instance has a Random-Context, and with probability $1-p$, it has a One-Near-Context.
- **[In-weights MSE Comparison]** Let $E = \mathbb{E}_D[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2]$ be the population MSE of the in-weights estimator. We assume that due to a limited training budget, this estimator is outperformed by ICL from a very similar example

but is better than ICL from a random example. That is, E is bounded by $L^2\|\mathbf{x}_i - \mathbf{x}_*\|_2^2$. For a similar point, $\|\mathbf{x}_i - \mathbf{x}_*\|_2^2 = 2(1 - \mathbf{x}_i^\top \mathbf{x}_*) \leq 2\delta_2$. For a random point, $\|\mathbf{x}_i - \mathbf{x}_*\|_2^2 \geq 2(1 - \delta_1)$. Overall, we get these dataset-dependent bounds:

$$2L^2\delta_2 \;\leq\; E \;\leq\; 2L^2(1 - \delta_1) \tag{A}$$

Optimal Parameters for Different Regimes We now analyze the stationary points[1] of the loss $\mathcal{L}$ for each context regime. Detailed proofs of stationarity and optimality are in Appendix A.5.

Case 1: Random-Context. When all context examples are dissimilar to the target, the optimal strategy is to ignore the context and rely on in-weights learning.

- **Optimal Parameters:** The limit $\theta_2 \to -\infty$ is a stationary point. This forces all attention scores $s_i \to 0$, causing the attention weight on the in-weights prediction to dominate ($a_* \to 1$).
- **Resulting Loss:** The prediction becomes $\hat{\mathbf{y}} \to \hat{f}(\mathbf{x}_*)$, and the loss is $\mathcal{L}_{\text{param}} = E$.
- **Brittleness:** With a *One-Near-Context* context, it would fail to use the highly relevant example achieving a sub-optimal loss of E instead of the ICL loss $\leq 2L^2\delta_2$, which is lower as per Eqn A.

Case 2: Similar-Context. When all context examples are highly similar to the target, the model should perform ICL by averaging the context labels.

- **Optimal Parameters:** The limit $\theta_3 + \theta_2 \to \infty$ while $\theta_3 \gg \theta_1$ is a stationary point. This drives the score $s_{k+1} \to \infty$, causing the weight on the in-weights prediction to vanish ($a_* \to 0$). The individual ICL attention values also vanish ($a_j \to 0\ \forall i \in [k]$).
- **Resulting Loss:** The prediction becomes an average of context labels, $\hat{\mathbf{y}} \to \bar{\mathbf{y}}$, with a low loss of $\mathcal{L}_{\text{icl}} \leq 2L^2\delta_2$.
- **Brittleness:** This model may learn to always trust the context, if it reaches $\theta_3 + \theta_2 \to \infty$ using $\theta_2 \to \infty$. When given a *Random-Context*, it would still average the random labels: $a_* = 0, a_i = 0\ \forall i \in [k], a_{k+1} \to 1 \implies \hat{\mathbf{y}} = \sum_i \mathbf{y}_i/k$. This leads to a high error, close to $2L^2(1 - \delta_1)$ worse than IWL (Eqn A).

Case 3: One-Near-Context. Here, the optimal strategy is to copy label of the near example.

- **Optimal Parameters:** The limit $\theta_1 \to \infty$ while $\theta_1 + \theta_2 \to \infty$, $\theta_1 \gg \theta_3$ is a stationary point. This makes the score $s_{j^\star}$ for the near point dominate all others, so that $a_{j^\star} \to 1$.
- **Resulting Loss:** The prediction converges to $\hat{\mathbf{y}} \to \mathbf{y}_{j^\star}$, with loss $\mathcal{L}_{\text{icl}} \leq 2L^2\delta_2$.
- **Brittleness:** This model learns a "copy-the-best" heuristic. In a *Random-Context*, the large θ_1 amplifies small differences in $\mathbf{x}_i^\top \mathbf{x}_*$, causing it to copy the closest label, which is still random.

Case 4: Random-Context + Similar-Context. Here the model should learn an IC Copy-IWL mixture that can switch between the two based on the context.

- **Optimal Parameters:** The limit (a) $\theta_3 + \theta_2 \to \infty$, (b) $\theta_2 \to -\infty$ where $\theta_3 \gg \theta_1$
- **Adaptive Behavior:** This parameter setting produces optimal behavior in Random-Context and Similar-Context regimes, but the model remains brittle in the One-Near-Context regime. • Under *Random-Context*, condition (b) forces $a_* \to 1$, correctly defaulting to the in-weights prediction $\hat{\mathbf{y}} \to \hat{f}(\mathbf{x}_*)$. • Under *Similar-Context*, it behaves like the model trained on *Similar-Context* due to condition (a), thus doing a label average.
- **Brittleness:** With a *One-Near-Context* context, it would fail to use the highly relevant example. Due to condition (a), $a_* = 0, a_i = 0\ \forall i \in [k], a_{k+1} \to 1 \implies \hat{\mathbf{y}} = \sum_i \mathbf{y}_i/k$. This leads to a high error, close to $2L^2(1 - \delta_1)$ for large k worse than IWL (Eqn A).

Case 5: Contrastive-Context. Here the model should learn an ideal ICL-IWL mixture that can switch between the two based on the context.

- **Optimal Parameters:** The limit (a) $\theta_1 + \theta_2 \to \infty$, (b) $\theta_2 \to -\infty$ where $\theta_1 \gg \theta_3$
- **Adaptive Behavior:** This parameter setting produces optimal behavior in all regimes, overcoming the brittleness of specialized models. • Under *Random-Context*, condition (b) forces $a_* \to 1$, correctly defaulting to the in-weights prediction $\hat{\mathbf{y}} \to \hat{f}(\mathbf{x}_*)$. • Under *One-Near-Context* context, conditions (a) and (b) force $a_{j^\star} \to 1$, correctly switching

[1] Note that while our analysis uses the theoretical limits of $\theta_i \to \pm\infty$, these represent optimization directions. In practice, due to the exponential scaling in the softmax, the desired behavior of attention weights saturating at 0 or 1 is achieved once the parameters θ_i attain a sufficiently large finite magnitude.

to the ICL prediction $\hat{\mathbf{y}} \to \mathbf{y}_{j^*}$. • Under *Similar-Context*, $a_* \to 0$ and $a_{k+1} \to 0$ thus yielding a weighted average of all the labels.

The above theoretical analysis reveals the importance of creating contrast both within examples in a context and across contexts to ensure that the model learns to harness context only based on similarity to the target, and to rely on IWL when context examples are not similar enough.

4. Empirical Study

We empirically compare IC-Train trained under the three types of context and standard Zero-Context training. We fine-tune four open source models on four machine translation tasks, eleven Text-to-SQL task, three multilingual semantic parsing tasks, and in Appendix C.4 also include two synthetic alignment reasoning tasks.

Evaluation Setup. In order to study the effect of different grades of relatedness of the target test input to the in-context examples, each test example in D is evaluated under three different contexts: (1) Randomly selected k examples from D, (2) Select k examples most similar to $\mathbf{x}_*$ using BM25, (3) $k-1$ random examples and 1 example most similar to $\mathbf{x}_*$ using BM25, (4) $k-1$ random examples and one closely related example obtained to be a paraphrase of $\mathbf{x}_*$ with $\mathbf{y}_i = \mathbf{y}_*$. This mode checks the scenario of whether a model can harness closely related feedback from users after training. Over the union of all (context,test) instances, we plot accuracy as a function of the maximum similarity of the input $\mathbf{x}_*$ to an example in its context.

Models. We experiment with the following open-source LLMs: 1. Llama 3.2 1B base, 2. Llama 3.1 8B base (Grattafiori et al., 2024) 3. Qwen 2.5 7B (Yang et al., 2024). 4. Mistral 7B v0.3 (Jiang et al., 2023)

Machine Translation (MT) Tasks. We translate from English to four different languages: Lithuanian, Tamil, Hindi, and German that cover a wide spectrum of an LLM's exposure to these languages. We train on standard benchmarks, and evaluate under two test-sets, for a total of 32 (model,task,test-set) configurations. For testing, we consider two types: a generic dataset from Flores (ID), and to evaluate the capability of continuous adaptation, an out-of-distribution dataset (OOD) from specialized domain such as judiciary and religion. Appendix Section C.2 provides more information. The prompt is given in Appendix B.1. For all training setups (Random-Context, Similar-Context, Contrastive-Context) we choose k randomly from $[0 \ldots 5]$. We use COMET-22 (Rei et al., 2020) to measure accuracy.

In Figure 3 we plot accuracy against the maximum target-context similarity on combinations of four models, four tasks, two test settings, and five training methods: Zero-Context, IC-Train with Random-Context, Similar-Context, Contrastive-Context, and the original untuned base model. For clarity we group similarity values into three bins of Low, Medium, and High similarity. The error bars shows the 95% confidence interval with 1000 bootstrap resampling of the test data. Raw plots and over all 32 task-model combinations appear in Appendix Figure E. Based on these plots, we make the following observations:

1. Zero-Context (**orange** bar) boosts accuracy of the base model (**blue** bar) for low to medium target-context similarity. But it causes loss of ICL ability of the base model as seen by the drop in accuracy for high target-context similarity cases (right side of each plot) in almost all 32 model-task-test settings. This is in agreement with the conclusions of earlier studies (Alves et al., 2023; Duan et al., 2024; Wang et al., 2024) and our analysis.
2. IC-Train with Random-Context (**green** bar) provides the least accuracy with increased target-context similarity across all settings. Its performance is worse than even Zero-Context when presented with similar examples in context! Thus, Random-Context cannot harness highly related examples in-context.
3. IC-Train with Similar-Context (**red** bar) suffers in the low similarity region, and is worse than all other forms of fine-tuning in that region. It provides decent accuracy in medium similarity range, but its capability to harness highly related examples in context is worse than baseline's for almost all model language combinations in both ID and OOD settings.
4. Contrastive-Context (**violet** bar), is among the highest performing methods. Compared to Similar-Context, the second best performer, Contrastive-Context scores when highly related examples are present in context, while being competitive in low similarity ranges.

Text to SQL Task. We experiment on Text-to-SQL as an instance of a Text-to-code generation task where the need for online adaptation to private databases is compelling. We use the BIRD dataset (Li et al., 2024b), with its official train split for fine-tuning and dev split spanning eleven distinct databases. More details in Appendix C.3. The results shown

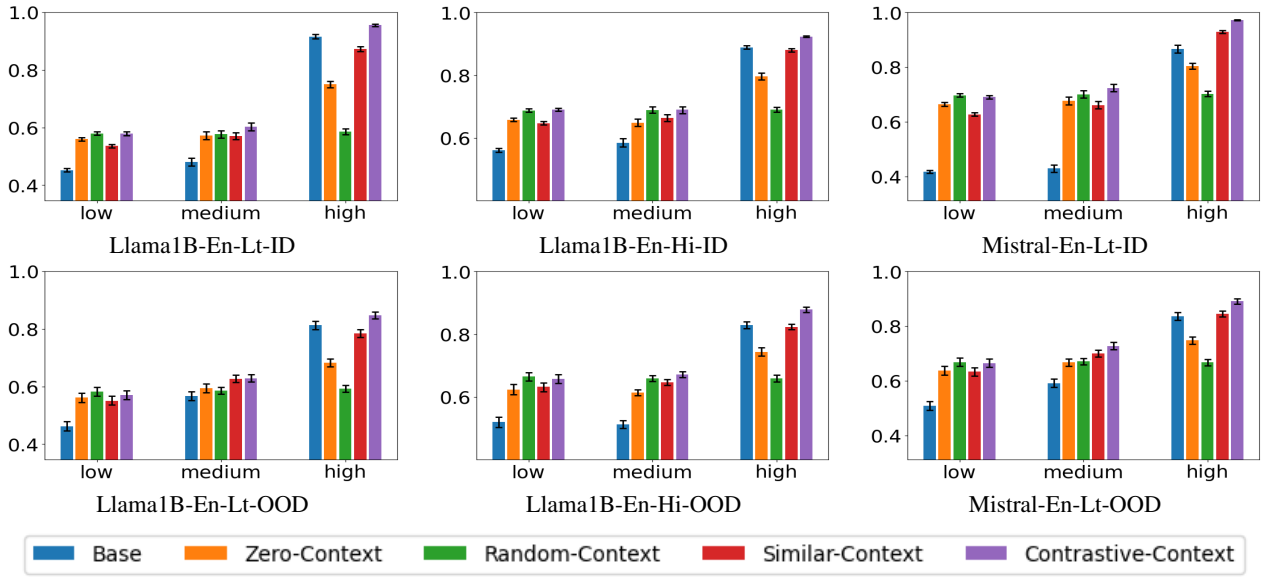


Figure 3. Effect of fine-tuning a base model with different strategies (Zero-Context, and IC-Train under Random-Context, Similar-Context, and Contrastive-Context) on accuracy over varying grades of similarity to in-context examples for 32 different models, language-pairs, and test-sets. Remaining plots in Appendix Figure E. X-axis: Level of maximum similarity of target with in-context examples. The similarity ranges here are - Low: $0-0.33$, Medium: $0.33-0.67$, High: $0.67-1$. Y-axis: Accuracy (COMET score). Main observations: **Contrastive-Context is among the most accurate across the entire spectrum of target-context relatedness. On targets with high context similarity, model fine-tuned with Zero-Context is worse than baseline, Random-Context even worse than Zero-Context. On targets with low context similarity, Similar-Context is worse than Zero-Context and Random-Context.**

in Figure 6(a) further add evidence to the robustness of Contrastive-Context to handling contexts at varying similarity levels. Observe how IC-Train with Random-Context is worse than even Zero-Context in the high similarity range, and with Contrastive-Context we obtain the best accuracy across all levels. In this task Similar-Context does not suffer in the low-similarity range because the labeled pool is small, and Similar-Context almost reduces to random for many instances.

Multilingual Semantic Parsing task. We use the MTOP dataset from XSemPLR (Zhang et al., 2023b), a unified benchmark for cross-lingual semantic parsing. We experiment on three languages — Spanish, German, and French. The results in Appendix Figure 6(b) show that Similar-Context suffers in the low and medium similarity regions compared to Random-Context and Contrastive-Context. In the high similarity region, Random-Context suffers compared to Similar-Context and Contrastive-Context. In this task too Contrastive-Context performs competitively or better across all similarity levels.

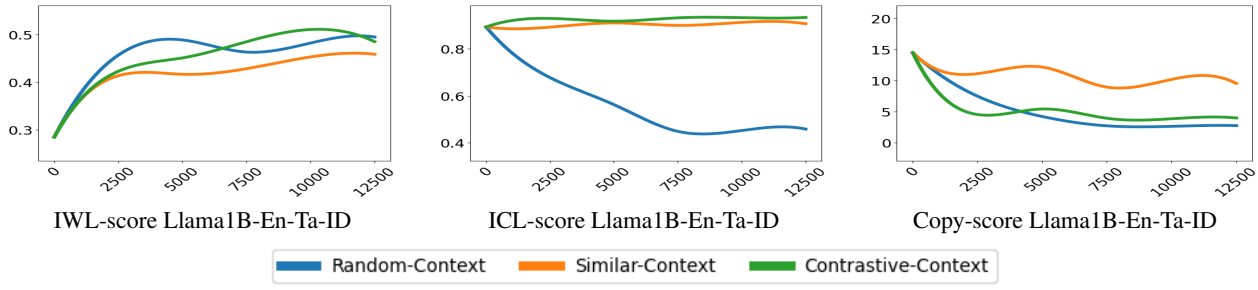
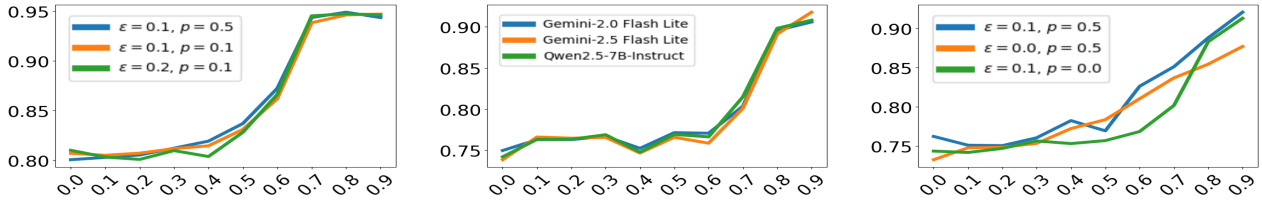


Figure 4. Emergence of different forms of learning in three training methods: Random-Context, Similar-Context, and Contrastive-Context. X-axis is training steps and Y-axis denotes scores of one of the three probes. Results of other model-task and datasets in Appendix Figure F. IWL-score of Similar-Context is lowest, ICL-score of Random-Context diminishes fast with training, Copy-score of Similar-Context is higher. Contrastive-Context provides best retention of ICL and IWL capabilities without resorting to copying.



(a) Sensitivity to ϵ, p values Llama1B-En-De-ID

(b) Quality of paraphrasing Llama1B-En-De-OOD

(c) Role of different types of contrasts Llama8B-En-Lt-OOD

Figure 5. Ablations to show (a) the robustness of Contrastive-Context to variant non-zero ϵ, p values, (b) various paraphrasing models, (c) importance of various levels of similarity in the training data. X-axis is target-context similarity and Y-axis accuracy.

4.1. Emergence of different forms of learning on LLMs

In Section 3 we analyzed a minimal model, to understand the effect of target-context similarity on the emergence of different forms of learning. To extend this understanding to large models and real tasks, we design three probes and analyze these as training progresses.

PROBES TO DETECT DIFFERENT FORMS OF LEARNING

Let $\hat{\mathbf{y}}_C = \text{argmax}_{\mathbf{y}} P_\theta(\mathbf{y}|C, \mathbf{x}_*)$ denote the predicted output under set $C = [\mathbf{x}_1, \mathbf{y}_1, \ldots \mathbf{x}_k, \mathbf{y}_k]$ and $\hat{\mathbf{y}}_\phi$ denote zero-shot prediction. Let $\mathcal{K}(\mathbf{x}, \mathbf{x}')$ denote similarity between two inputs and $\text{sim}(\mathbf{y}, \mathbf{y}')$ between outputs.

In-Weights Learning Probe. We quantify the in-weights learning as the similarity between predictions under random and empty context. $\text{IWL-score}(P_\theta) = \mathbb{E}_{\mathbf{x}_*}\, \mathbb{E}_{C\sim\text{Random}(D,k)}\, \text{sim}(\hat{\mathbf{y}}_C, \hat{\mathbf{y}}_\phi)$.

In-Context Learning Probe. We quantify ICL capability by comparing prediction $\hat{\mathbf{y}}_C$ under a one-similar context C to the labels in context weighted by their similarity to $\mathbf{x}_*$. One-similar context is obtained by injecting a paraphrase for $\mathbf{x}_*$ with gold $\mathbf{y}_*$ among remaining $k-1$ random examples, so $C = [\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_{i^*} = \text{paraphrase}(\mathbf{x}_*), \mathbf{y}_{i^*} = \mathbf{y}_*, \ldots \mathbf{x}_k, \mathbf{y}_k]$ $\text{ICL-score}(P_\theta) = \mathbb{E}_{\mathbf{x}_*}\, \mathbb{E}_{C\sim\text{One-similar}(D,\mathbf{x}_*,k)} \sum_{j=1}^{k} \frac{\mathcal{K}(\mathbf{x}_*,\mathbf{x}_j)\text{sim}(\hat{\mathbf{y}}_C,\mathbf{y}_j)}{\sum_i \mathcal{K}(\mathbf{x}_*,\mathbf{x}_i)}$

Blind Copy Probe. A model with propensity to copy indiscriminately from the context would output a $\hat{\mathbf{y}}$ similar to a label $\mathbf{y}_i$ in-context irrespective of $\mathbf{x}_*$'s similarity to $\mathbf{x}_i$. We quantify this by shuffling the one-similar context (above) so that no $\mathbf{x}_i$ and $\mathbf{y}_i$ are correctly matched. An example for $k = 3$ is: $C = [\mathbf{x}_1, \mathbf{y}_*, \mathbf{x}_2 = \text{paraphrase}(\mathbf{x}_*), \mathbf{y}_3, \mathbf{x}_3, \mathbf{y}_1]$. Pure ICL would output $\mathbf{y}_3$ based on similarity of $\mathbf{x}_2$ with $\mathbf{x}_*$ whereas a model that copies independent of $\mathbf{x}$ could output one of the other labels. Thus, we define copy score as maximum similarity to a label other than the one attached to $\mathbf{x}_{i^*}$ where i^* is the position of $\mathbf{x}_*$'s paraphrase. $\text{Copy-score}(P_\theta) = \mathbb{E}_{\mathbf{x}_*}\, \mathbb{E}_{C\sim\text{Shuffle}(\text{One-similar}(D,\mathbf{x}_*,k))}\, \max_{i\in C, i\neq i^*} \text{sim}(\hat{\mathbf{y}}_C, \mathbf{y}_i)$ We used cosine similarity of sentence embeddings for $\mathcal{K}$ and COMET for sim for the first two probes but BLEU for the copy score to measure the model's propensity to blindly copy the lexical tokens.

Main observations. In Figure 4 we show the emergence of different forms of learning as measured by these three probes for various model, language pair, and dataset combinations. More combinations appear in the Appendix Figure F. We can make the following observations from these graphs: (1) The IWL-score (first column) of Random-Context, and Contrastive-Context steadily increases with training steps with Contrastive-Context lagging only slightly behind. However, for Similar-Context the IWL-score is distinctly lower, sometimes by a significant margin. This, coupled with the observation that Similar-Context provides lower accuracy in Figure 3 when target-context similarity is low, shows that Similar-Context is less effective to learn the task in-weights, and/or is distracted by irreleveant examples. (2) The ICL-score (second column) shows that for Random-Context ICL capability steadily drops with training, explaining why it provides almost the same accuracy across different grades of target-context similarity. For Contrastive-Context and Similar-Context, the ICL capability stays almost the same or increases. (3) The copy score (last column) decreases with training using Random-Context, and Contrastive-Context and gets substituted by either in-weights on in-context learning. In contrast, for Similar-Context the copy score is higher than for the other models. On real datasets, Similar-Context training may not consistently show increased copy tendencies because not every target will find all similar top-k examples, and the training data may naturally contain a mix of random and similar contexts. But in spite of the natural mixing, we observe Similar-Context to result in reduced IWL and more blind copying compared to Contrastive-Context.

4.2. Ablations

We present brief ablations on our method here. Appendix D has more details. **Sensitivity to ϵ, p values.** Contrastive-Context creates contrast using parameter p to control the split of random Vs similar contexts, and control the fraction of highly similar synthetic examples (ϵ). In Figure 5(a) we show the method is robust to alternative values ϵ, p between 0.1 and 0.5. **Quality of paraphrasing.** To study impact paraphrase quality, we evaluated with two alternative models `gemini-2.5-flash-lite` a strong model, and `Qwen2.5-7B-Instruct` as an open source, possibly weaker model compared to `gemini-2.0-flash-lite`. Fig 5(b) shows that Contrastive-Context remains stable across these three paraphrasing models. **Role of different types of contrasts.** We perform ablations to evaluate the role of different types of contrasts that Contrastive-Context creates in Fig 5(c). We set $\epsilon = 0$ to suppress generation of synthetic highly similar examples and observe that test instances with high target- context similarity suffer. We set $p = 0$ to suppress the generation of similar real examples which causes medium scale accuracy to drop.

5. Related Work

Starting from the landmark study of (Brown et al., 2020) showing the emergence of ICL in pre-trained LLMs, ICL has been extensively studied along various aspects including understanding ICL emergence (Garg et al., 2022; Zhang et al., 2023a; Olsson et al., 2022; Shi et al., 2024; Agarwal & Sarawagi, 2025), evaluating ICL on real tasks (Kossen et al., 2024; Bertsch et al., 2024), and instance selection during deployment (Levy et al., 2023; Kothyari et al., 2025). Our focus is on understanding the interplay between ICL and IWL during task-specific training with in-context examples. Prior work studied how ICL-IWL emergence is influenced by two kinds target-context relatedness.

Relatedness of the $\mathbf{x}$ to $\mathbf{y}$ mapping function. Many prior work study a setup where training is over a mixture of T tasks sampled from a task family (say linear regression with task-specific weights). For the k in-context examples, the output $\mathbf{y}_i$ is determined by the same task $f_\tau(\mathbf{x})$. During IC-Train with a task mixture, a model is said to develop ICL if it uses the in-context examples to infer the f_τ, and IWL if the T tasks are learned in parameters. Several studies show that as task diversity (T) increases, models prefer ICL over IWL (Reddy, 2024; Singh et al., 2024; Akyürek et al., 2024; Edelman et al., 2024; Park et al., 2025; Nguyen & Reddy, 2025; Wurgaft et al., 2025; Ku et al., 2025; Kim et al., 2025; Singh et al., 2025; Fu et al., 2024). The tasks considered are synthetic regression, classification, or sequence completion (Park et al., 2025; Akyürek et al., 2024; Rajaraman et al., 2024; Edelman et al., 2024). All these works use random selection of in-context examples from $P(X)$. Our work departs by studying IC-Train on a single real sequence to sequence task and isolating how inter-example relatedness shapes the ICL–IWL tradeoff.

Relatedness of $\mathbf{x}$ tokens. (Chan et al., 2022) show that when the data distribution is bursty, causing related examples to appear in the context, the model develops ICL capabilities, whereas for non-bursy distributions IWL emerges — ICL and IWL develop together when the data follows a Zipfian distribution. (Singh et al., 2023) further observe that ICL is a transient phenomenon, and asymptotically could reduce to IWL. (Zucchet et al., 2025; Bratulić et al., 2025) discuss how in-context repetition promotes ICL. (Gao & Das, 2024) propose a prompting-based method using contrastive examples to improve ICL, but here they define contrast using the correctness of the context. We present a finer-grained analysis in terms of similarity that generalizes repetition. Further, we stress the importance of contrast within the in-context examples to promote ICL over blind copying. In machine translation, Alves et al. (2023) show that fine-tuning LLMs in ICL mode preserves ICL ability, often lost during standard zero-shot fine-tuning. We show that poorly chosen context can harm ICL more than zero-shot.

6. Conclusion

We studied how to train models to balance in-context learning (ICL) with in-weights learning (IWL) and switch between them based on context relevance. Our analysis shows this balance is fragile, and strongly influenced by the target-context similarity patterns: random contexts lead to IWL dominance, while overly similar ones reduce ICL to degenerate copying. We introduced Contrastive-Context, a simple strategy of creating contrast both within examples in a context and across contexts. A theoretical analysis with a minimal transformer provide insights on why such contrasts are essential. Experiments over 32 model–task–test settings for low-resource MT, eleven Text-2-SQL tasks, three multilingual semantic parsing tasks, and a synthetic alignment task demonstrate that contrasted contexts preserve IWL while sustaining robust ICL. Probes on large LLMs further confirm similarity structure as a decisive factor in avoiding collapse into IWL, ICL, or blind copying.

References

Agarwal, H. and Sarawagi, S. The missing alignment link of in-context learning on sequences. 2025.

Akyürek, E., Wang, B., Kim, Y., and Andreas, J. In-context language learning: Architectures and algorithms. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3Z9CRr5srL.

Alves, D., Guerreiro, N., Alves, J., Pombal, J., Rei, R., de Souza, J., Colombo, P., and Martins, A. Steering large language models for machine translation with finetuning and in-context learning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11127–11148, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.744. URL https://aclanthology.org/2023.findings-emnlp.744/.

Bertsch, A., Ivgi, M., Alon, U., Berant, J., Gormley, M. R., and Neubig, G. In-context learning with long-context models: An in-depth exploration, 2024.

Bornschein, J., Lyle, C., Li, Y., Rannen-Triki, A., He, X. O., and Pascanu, R. Fine-tuned in-context learners for efficient adaptation, 2025. URL https://arxiv.org/abs/2512.19879.

Bratulić, J., Mittal, S., Hoffmann, D. T., Böhm, S., Schirrmeister, R. T., Ball, T., Rupprecht, C., and Brox, T. Unlocking in-context learning for natural datasets beyond language modelling, 2025. URL https://arxiv.org/abs/2501.06256.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chan, B., Chen, X., Gyorgy, A., and Schuurmans, D. Toward understanding in-context vs. in-weight learning. 2024. URL https://api.semanticscholar.org/CorpusID:273695891.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., McClelland, J. L., and Hill, F. Data distributional properties drive emergent in-context learning in transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Duan, H., Tang, Y., Yang, Y., Abbasi, A., and Tam, K. Y. Exploring the relationship between in-context learning and instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3197–3210, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.182. URL https://aclanthology.org/2024.findings-emnlp.182/.

Edelman, E., Tsilivis, N., Edelman, B. L., eran malach, and Goel, S. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qaRT6QTIqJ.

Fu, J., Yang, T., Wang, Y., Lu, Y., and Zheng, N. Breaking through the learning plateaus of in-context learning in transformer. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=2K87GFLYWz.

Gao, X. and Das, K. Customizing language model responses with contrastive in-context learning, 2024. URL https://arxiv.org/abs/2401.17390.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=flNZJ2eOet.

Goyal, S., Baek, C., Kolter, J. Z., and Raghunathan, A. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SPS6HzVzyt.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kim, J., Kwon, S., Choi, J. Y., Park, J., Cho, J., Lee, J. D., and Ryu, E. K. Task diversity shortens the icl plateau, 2025. URL https://arxiv.org/abs/2410.05448.

Kossen, J., Gal, Y., and Rainforth, T. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YPIA7bgd5y.

Kothyari, M., Sarawagi, S., Chakrabarti, S., Arora, G., and Merugu, S. Diverse in-context example selection after decomposing programs and aligned utterances improves semantic parsing. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, April 2025.

Ku, A. Y., Griffiths, T. L., and Chan, S. C. Y. Predictability shapes adaptation: An evolutionary perspective on modes of learning in transformers, 2025. URL https://arxiv.org/abs/2505.09855.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1548.

Levy, I., Bogin, B., and Berant, J. Diverse demonstrations improve in-context compositional generalization. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 1401–1422. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.78. URL https://doi.org/10.18653/v1/2023.acl-long.78.

Li, H., Zhang, J., Li, C., and Chen, H. Resdsql: Decoupling schema linking and skeleton parsing for text-to-sql. In *AAAI*, 2023.

Li, H., Zhang, J., Liu, H., Fan, J., Zhang, X., Zhu, J., Wei, R., Pan, H., Li, C., and Chen, H. Codes: Towards building open-source language models for text-to-sql. In *SIGMOD*, 2024a.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36, 2024b.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context, 2022. URL https://arxiv.org/abs/2110.15943.

Nguyen, A. and Reddy, G. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=INyi7qUdjZ.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. Scaling neural machine translation to 200 languages. *Nature*, 630(8018): 841–846, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., Dassarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T. B., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *ArXiv*, abs/2209.11895, 2022. URL https://api.semanticscholar.org/CorpusID:252532078.

Park, C. F., Lubana, E. S., and Tanaka, H. Competition dynamics shape algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XgH1wfHSX8.

Rajaraman, N., Bondaschi, M., Makkuva, A. V., Ramchandran, K., and Gastpar, M. Transformers on markov data: Constant depth suffices. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=jnCM5EHd2H.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 02 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00452. URL https://doi.org/10.1162/tacl_a_00452.

Reddy, G. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025, 2020. URL https://arxiv.org/abs/2009.09025.

Shi, Z., Wei, J., Xu, Z., and Liang, Y. Why larger language models do in-context learning differently? In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=WOa96EG26M.

Singh, A., Chan, S., Moskovitz, T., Grant, E., Saxe, A., and Hill, F. The transient nature of emergent in-context learning in transformers. *Advances in neural information processing systems*, 36:27801–27819, 2023.

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C., and Saxe, A. M. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=O8rrXl71D5.

Singh, A. K., Moskovitz, T., Dragutinovic, S., Hill, F., Chan, S. C. Y., and Saxe, A. M. Strategy coopetition explains the emergence and transience of in-context learning, 2025. URL https://arxiv.org/abs/2503.05631.

Tiedemann, J. Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Wang, Y., Bai, A., Peng, N., and Hsieh, C.-J. On the loss of context-awareness in general instruction fine-tuning, 2024. URL https://arxiv.org/abs/2411.02688.

Wurgaft, D., Lubana, E. S., Park, C. F., Tanaka, H., Reddy, G., and Goodman, N. D. In-context learning strategies emerge rationally, 2025. URL https://arxiv.org/abs/2506.17859.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *ArXiv*, abs/2306.09927, 2023a. URL https://api.semanticscholar.org/CorpusID:259187776.

Zhang, Y., Wang, J., Wang, Z., and Zhang, R. Xsemplr: Cross-lingual semantic parsing in multiple natural languages and meaning representations. In *ACL*, 2023b.

Zucchet, N., D'Angelo, F., Lampinen, A. K., and Chan, S. C. The emergence of sparse attention: impact of data distribution and benefits of repetition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jMhRbV47pS.

A. Additional Details for Theoretical Analysis

A.1. Architecture of a Transformer that implements the Minimal Model

Here, we detail the construction of a Transformer whose final prediction matches the minimal model analyzed in Section 3: $\hat{\mathbf{y}} = a_* \hat{f}(\mathbf{x}_*) + \sum_{i=1}^{k} a_i \mathbf{y}_i + a_{k+1}\bar{\mathbf{y}}$. The final prediction is taken from the output embedding of the last token in the sequence (the query token).

A.2. Token Representations and Processing

Let the model's internal embedding dimension be $d_{\text{model}} = 3d + 3$. For clarity, I_n is the $n \times n$ identity matrix, $\mathbf{0}_{m\times n}$ is an $m \times n$ zero matrix, and $\mathbf{0}_n$ is an n-dimensional zero vector.

1. Initial Embeddings The input sequence consists of k context tokens, one `EOC` token and one query token. Their initial embeddings are structured to segregate features.

- **Context Token** i: $\mathbf{h}_i^{\text{initial}} = [\mathbf{x}_i; 1; \mathbf{0}_d; \mathbf{0}_d; 0; \mathbf{y}_i] \in \mathbf{R}^{3d+3}$
- **`EOC` Token** $k+1$: $\mathbf{h}_{k+1}^{\text{initial}} = [\mathbf{0}_d; 1; \sum_{j=1}^{k} \mathbf{x}_j; \mathbf{0}_d; 0; \bar{y}] \in \mathbf{R}^{3d+3}$
- **Query Token** $k+2$: $\mathbf{h}_{k+2}^{\text{initial}} = [\mathbf{0}_d; 0; \mathbf{0}_d; \mathbf{x}_*; 1; 0] \in \mathbf{R}^{3d+3}$

2. Position-wise Feed-Forward Network (FFN) A position-wise FFN, acting as the in-weights estimator $\hat{f}$, is applied across all token positions. Its computation is conditional on the $(3d+2)$-th dimension of the input embedding, which serves as a gate: it is '1' for the query token and '0' for all context tokens. Consequently, the FFN's output is non-zero only for the query token, where it produces the scalar estimate $\hat{f}(\mathbf{x}_*)$. This output is then added into the final dimension of the embedding via a residual connection:

$$\hat{f}(\mathbf{x}_*) = \text{FFN}(\mathbf{h}_{k+2}^{\text{initial}})$$
$$\mathbf{h}_{k+2} = \mathbf{h}_{k+2}^{\text{initial}} + [\mathbf{0}_{3d+2}; \hat{f}(\mathbf{x}_*)]$$

The embeddings of context tokens are unchanged as their gate value is zero, effectively nullifying the FFN's contribution for those positions.

3. Final Embeddings (Input to Attention) The embeddings entering the final attention layer are:

- **Context Token** i: $\mathbf{h}_i^{\text{initial}} = [\mathbf{x}_i; 1; \mathbf{0}_d; \mathbf{0}_d; 0; \mathbf{y}_i]$
- **`EOC` Token** $k+1$: $\mathbf{h}_{k+1}^{\text{initial}} = [\mathbf{0}_d; 1; \sum_{j=1}^{k} \mathbf{x}_j; \mathbf{0}_d; 0; \bar{y}]$
- **Query Token** $k+2$: $\mathbf{h}_{k+2}^{\text{initial}} = [\mathbf{0}_d; 0; \mathbf{0}_d; \mathbf{x}_*; 1; \hat{f}(\mathbf{x}_*)]$.

A.3. Projection Matrices

We design the weight matrices to project these embeddings into query, key, and value spaces. Let $d_q = d_k = d + 1$ and $d_v = 1$.

Query Matrix $W_Q \in \mathbf{R}^{(d+1)\times(3d+3)}$ W_Q is a block matrix that isolates the feature vector $[\mathbf{x}_*; 1]$ from the query token.

$$W_Q = \begin{bmatrix} \mathbf{0}_{(d+1)\times(2d+1)} & I_{d+1} & \mathbf{0}_{(d+1)\times 1} \end{bmatrix}$$

The query vector from the final position is:

$$Q_{k+2} = \mathbf{h}_{k+2} W_Q^T = [\mathbf{x}_*; 1] \in \mathbf{R}^{d+1}.$$

Queries from context positions are not used for the final prediction.

Key Matrix $W_K \in \mathbf{R}^{(d+1)\times(3d+3)}$ W_K extracts context features and applies the learnable parameters θ_1, θ_2 and θ_3.

$$W_K = \begin{bmatrix} \theta_1 I_d & \mathbf{0}_{d\times 1} & \theta_3 I_d & \mathbf{0}_{d\times(d+2)} \\ \mathbf{0}_{1\times d} & \theta_2 & \mathbf{0}_{1\times d} & \mathbf{0}_{1\times(d+2)} \end{bmatrix}$$

The key vectors relative to the query Q_{k+2} are:

- For a context token: $K_i = \mathbf{h}_i W_K^T = [\theta_1 \mathbf{x}_i; \theta_2] \in \mathbf{R}^{d+1}$.
- For the EOC token: $K_{k+1} = \mathbf{h}_{k+1} W_K^T = [\theta_3 \sum_{j=1}^k \mathbf{x}_j; \theta_2] \in \mathbf{R}^{d+1}$.
- For the context token: $K_{k+2} = \mathbf{h}_{k+2} W_K^T = [\mathbf{0}_d; 0] \in \mathbf{R}^{d+1}$.

Value Matrix $W_V \in \mathbf{R}^{1\times(3d+3)}$ W_V is a simple selector for the last dimension of the embeddings, which holds the label $\mathbf{y}_i$ for context tokens and the in-weights prediction $\hat{f}(\mathbf{x}_*)$ for the query token.

$$W_V = \begin{bmatrix} \mathbf{0}_{1\times(3d+2)} & 1 \end{bmatrix}$$

The resulting scalar value vectors are:

- For a context token: $V_i = \mathbf{h}_i W_V^T = \mathbf{y}_i \in \mathbf{R}$.
- For the EOC token: $V_{k+1} = \mathbf{h}_{k+1} W_V^T = \bar{y} \in \mathbf{R}$.
- For the query token: $V_{k+2} = \mathbf{h}_{k+2} W_V^T = \hat{f}(\mathbf{x}_*) \in \mathbf{R}$.

A.4. Attention and Final Output

The raw attention logits from the query Q_{k+2} to all other tokens are:

$$\begin{aligned} \text{logit}(k+2, i) &= Q_{k+1} \cdot K_i = \theta_1 \mathbf{x}_*^\top \mathbf{x}_i + \theta_2 \text{ (for } i = 1, \ldots, k). \\ \text{logit}(k+2, k+1) &= Q_{k+1} \cdot K_i = \theta_3 \sum_{j=1}^k \mathbf{x}_*^\top \mathbf{x}_j + \theta_2 \text{ (for } i = k+1). \\ \text{logit}(k+2, k+2) &= Q_{k+2} \cdot K_{k+2} = 0. \end{aligned}$$

As we use very sparse Q, K matrices and the inputs are unit norm, the logits are not likely to blow up in magnitude. Thus we omit the standard scaling of logits by $1/\sqrt{d_k}$ for analytical clarity.

After applying the softmax, the attention weights a_i (for $i = 1..k+1$) and $a_* \equiv a_{k+2}$ are exactly as defined in the main text. The output of the attention mechanism at the query position is a scalar, computed as the weighted sum of the scalar value vectors:

$$\begin{aligned} \hat{\mathbf{y}} &= \sum_{j=1}^{k+1} a_j V_j = a_{k+2} V_{k+2} + a_{k+1} V_{k+1} + \sum_{i=1}^k a_i V_i \\ &= a_* \hat{f}(\mathbf{x}_*) + a_{k+1} \bar{y} + \sum_{i=1}^k a_i \mathbf{y}_i. \end{aligned}$$

This completes the construction.

A.5. Derivations, Stationarity, and Optimality

Here we provide the detailed derivations for the stationarity and optimality of the parameter limits discussed in the main text.

A.5.1. Common Algebra

For a fixed target $\mathbf{x}_*$ and context C, the squared loss is

$$\ell = \Big(f(\mathbf{x}_*) - \hat{\mathbf{y}} \Big)^2$$

where $s_i = \exp\left(\theta_1 \left(\mathbf{x}_i\right)^\top \mathbf{x}_* + \theta_2\right),\ \forall i \in [k]$, $s_{k+1} = \exp\left(\theta_3 \sum_{j=1}^k (\mathbf{x}_j)^\top \mathbf{x}_* + \theta_2\right)$ and $S = \sum_{i=1}^k s_i$. The derivative

of ℓ with respect to a generic score s_j is:

$$\begin{aligned}\frac{\partial \ell}{\partial s_j} &= -2(f(\mathbf{x}_*) - \hat{\mathbf{y}}) \cdot \frac{\partial \hat{\mathbf{y}}}{\partial s_j} \\ &= \frac{2(f(\mathbf{x}_*) - \hat{\mathbf{y}})(\hat{\mathbf{y}} - \mathbf{y}_j)}{(1 + S + s_{k+1})} \;\; \forall j \in [k]\end{aligned}$$

$$\frac{\partial \ell}{\partial s_{k+1}} = \frac{2(f(\mathbf{x}_*) - \hat{\mathbf{y}})(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{(1 + S + s_{k+1})}$$

Using the chain rule with

$$\begin{aligned}\frac{\partial s_j}{\partial \theta_1} &= s_j((\mathbf{x}_j)^\top \mathbf{x}_*),\ j \in [k] \\ &= 0,\ j = k+1\end{aligned}$$

$$\frac{\partial s_j}{\partial \theta_2} = s_j,\ j \in [k+1]$$

$$\begin{aligned}\frac{\partial s_j}{\partial \theta_3} &= 0,\ j \in [k] \\ &= s_{k+1} \sum_{i=1}^{k} ((\mathbf{x}_i)^\top \mathbf{x}_*),\ j = k+1\end{aligned}$$

the gradients of the loss with respect to the attention parameters are:

$$\frac{\partial \ell}{\partial \theta_1} = \sum_{j=1}^{k} \frac{\partial \ell}{\partial s_j}\, s_j\, ((\mathbf{x}_j)^\top \mathbf{x}_*)$$

$$\frac{\partial \ell}{\partial \theta_2} = \sum_{j=1}^{k+1} \frac{\partial \ell}{\partial s_j}\, s_j.$$

$$\frac{\partial \ell}{\partial \theta_3} = \frac{\partial \ell}{\partial s_{k+1}}\, s_{k+1} \sum_{i=1}^{k} ((\mathbf{x}_i)^\top \mathbf{x}_*)$$

The gradient of the population loss $\mathcal{L}$ is the expectation of these quantities.

A.5.2. Proof of Stationarity for Each Regime

We establish that the limits described are stationary points by showing the pointwise gradients vanish. By the dominated convergence theorem (assuming bounded moments), this implies the gradient of the population loss also vanishes.

Case 1: Random-Context. We know $\forall\ j \in [k],\ (\mathbf{x}_j)^\top \mathbf{x}_* \to 0$. Thus in the limit $\theta_2 \to -\infty$, every score $s_j \to 0\ \forall\ j \in [k+1]$ and $s_{k+1} = \exp\left(\theta_3 \sum_{j=1}^{k} (\mathbf{x}_j)^\top \mathbf{x}_* + \theta_2\right) \to 0$. The gradients $\partial\ell/\partial\theta_1$, $\partial\ell/\partial\theta_2$ and $\partial\ell/\partial\theta_3$ contain a factor of s_j, causing the full gradient to vanish pointwise.

Case 2: Similar-Context. We know $\forall\ j \in [k],\ (\mathbf{x}_j)^\top \mathbf{x}_* \to 1$, which implies $s_j \to \exp(\theta_1 + \theta_2)\ \forall\ j \in [k]$ and $s_{k+1} \to \exp(k\theta_3 + \theta_2)$.

Case 2a: Thus, in the limit $\theta_3 + \theta_2 \to \infty$ with $\theta_3 \gg \theta_1$, the score $s_{k+1} \gg s_i\ \forall\ i\ \in [k], s_{k+1} \gg 1$, and $S \to \infty$. This causes $a_i \to 0\ \forall\ i\ \in [k]$ and $a_* \to 0$, hence $\hat{\mathbf{y}} \to \bar{\mathbf{y}}$

Every term of the gradient $\partial l/\partial\theta_1$ has a factor $s_j/(1+S+s_{k+1}) \to 0$. The gradient $\partial l/\partial\theta_3$ has a factor $(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \to 0$. Finally, the gradient $\partial l/\partial\theta_2$ contains both the terms mentioned before: in the first k terms and the last term. Hence all the gradients vanish.

Case 2b: Thus, in the limit $\theta_1 + \theta_2 \to \infty$ with $\theta_1 \gg \theta_3$, all scores $s_j \to \infty, s_j \gg s_{k+1}\ \forall j \in [k]$, and $S \to \infty$. This causes $a_{k+1} \to 0$ and $a_* \to 0$, hence $\hat{\mathbf{y}} \to \sum s_j \mathbf{y}_j / S$

Applying $\forall j \in [k]\ (\mathbf{x}_j)^\top \mathbf{x}_* \to 1$

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_1} &= \frac{2(f(\mathbf{x}_*) - \hat{\mathbf{y}})}{(1+S+s_{k+1})} \sum_j s_j(\hat{\mathbf{y}} - y_j) \\ &= \frac{2(f(\mathbf{x}_*) - \hat{\mathbf{y}})}{(1+S+s_{k+1})}(S\hat{\mathbf{y}} - \sum_j s_j y_j) \\ &\to 0\end{aligned}$$

The gradient $\partial l/\partial\theta_3$ contains the factor $(s_{k+1}/(1+S+s_{k+1})) \to 0$. In the gradient $\partial l/\partial\theta_2$ the first k terms $\to 0$ as shown above and the last term is similar to that of $\partial l/\partial\theta_3$. Hence all the gradients vanish.

Case 3: One-Near-Context. We know $(\mathbf{x}_{j^*})^\top \mathbf{x}_* \to 1$ and $(\mathbf{x}_j)^\top \mathbf{x}_* \to 0,\ \forall j \in [k+1] \setminus j^*$. This causes $a_i \to 0\ \forall\ i\ \in [k+1] \setminus j^*$, $a_* \to 0$ and $a_{j^*} \to 1$, hence $\hat{\mathbf{y}} \to \mathbf{y}_j$ Thus in the limit $\theta_1 + \theta_2 \to \infty,\ \theta_1 \to \infty$, $\theta_1 \gg \theta_3$, $s_{j^*} \to \infty$, and $S \to \infty$. In the gradient $\partial l/\partial\theta_1$ and $\partial l/\partial\theta_2$, terms where $j \neq j^*$ tend to 0 due to the factor $(s_j/(1+S+s_{k+1})) \to 0$ and the term where $j = j^*$ tend to 0 due to the factor $(\hat{\mathbf{y}} - \mathbf{y}_{j^*}) \to 0$. Finally, the gradient $\partial l/\partial\theta_3$ contains the factor $(s_{k+1}/(1+S+s_{k+1})) \to 0$. Hence all the gradients vanish.

Case 4: Random-Context + Similar-Context. In the limit $\theta_3 + \theta_2 \to \infty,\ \theta_2 \to -\infty, \theta_3 \gg \theta_1$, irrespective of the type of context (Random or Similar), the gradients vanish.

- If the context is *Random-Context*, all scores $s_j \to 0\ \forall j \in [k]$. This matches the analysis for Case 1, and the gradient vanishes.
- If the context is *Similar-Context*, $s_{k+1} \to \infty$. This matches the analysis for Case 2a, and the gradient vanishes.

Case 5: Contrastive-Context (Random-Context + One-Near-Context). In the limit $\theta_1 + \theta_2 \to \infty,\ \theta_2 \to -\infty$, $\theta_1 \gg \theta_3$, for both types of context the gradient vanishes.

- If the context is *Random-Context*, all scores $s_j \to 0\ \forall j \in [k+1]$. This matches the analysis for Case 1, and the gradient vanishes.
- If the context is *One-Near-Context*, $s_{j^\star} \to \infty$. This matches the analysis for Case 3, and the gradient vanishes.

A.5.3. Optimality Analysis

Optimality in Random-Context. At the stationary point, the prediction is $\hat{\mathbf{y}} \to \hat{f}(\mathbf{x}_*)$ and the population loss is $\mathcal{L}_{\text{param}} = E$. The alternative, a fully ICL prediction, would be an average of context labels where all context points are far from the target. Since $\|\mathbf{x}_i - \mathbf{x}_*\|_2^2 \geq 2(1-\delta_1)$, the Lipschitz assumption implies an ICL-induced error of at least $2L^2(1-\delta_1)$. Bound in (A) states $E \leq 2L^2(1-\delta_1)$, so the parametric extreme is optimal.

Optimality in Similar-Context. The loss at this ICL stationary point is determined by the prediction $\hat{\mathbf{y}} \to \sum_i w_i \mathbf{y}_i$. Since all context points are near the target ($\|\mathbf{x}_i - \mathbf{x}_*\|_2^2 \leq 2\delta_2$), the Lipschitz property implies $(f(\mathbf{x}_*) - \mathbf{y}_i)^2 \leq 2L^2\delta_2$. By Jensen's inequality, $\mathcal{L}_{\text{icl}} = \mathbb{E}[(f - \hat{\mathbf{y}})^2] \leq \mathbb{E}[\sum_i w_i (f - \mathbf{y}_i)^2] \leq 2L^2\delta_2$. The parametric loss is $\mathcal{L}_{\text{param}} = E$. Bound (A) states $E \geq 2L^2\delta_2$. Thus, $\mathcal{L}_{\text{icl}} \leq \mathcal{L}_{\text{param}}$, making the ICL extreme optimal.

Optimality in One-Near-Context. At this ICL stationary point, the prediction becomes $\hat{\mathbf{y}} \to y_{j^\star}$. The loss is $\mathbb{E}[(f(\mathbf{x}_*) - y_{j^\star})^2]$. Since $\mathbf{x}_{j^\star}$ is near $\mathbf{x}_*$, $\|\mathbf{x}_{j^\star} - \mathbf{x}_*\|_2^2 \leq 2\delta_2$, so by Lipschitz continuity, $\mathcal{L}_{\text{icl}} \leq 2L^2\delta_2$. The parametric loss is $\mathcal{L}_{\text{param}} = E$. Bound (A) again states $E \geq 2L^2\delta_2$, so $\mathcal{L}_{\text{icl}} \leq \mathcal{L}_{\text{param}}$, making this strategy optimal.

Optimality in Random-Context + Similar-Context. This stationary point is optimal for the mixed distribution because it dynamically selects the best strategy for each scenario. It defaults to the parametric model for Random-Context (achieving the optimal loss E) and switches to ICL for Similar-Context (achieving the optimal loss $\leq 2L^2\delta_2$). As it achieves the minimum possible loss for any draw from the distribution, it minimizes the expected loss over the entire distribution.

Optimality in Contrastive-Context. This stationary point is optimal for the mixed distribution because it dynamically selects the best strategy for each scenario. It defaults to the parametric model for Random-Context (achieving the optimal loss E) and switches to ICL for One-Near-Context (achieving the optimal loss $\leq 2L^2\delta_2$). As it achieves the minimum possible loss for any draw from the distribution, it minimizes the expected loss over the entire distribution.

A.5.4. Learning Dynamics of the In-Weights Estimator $\hat{f}$

We now analyze the learning signal for the in-weights estimator $\hat{f}$ during training. The prediction error can be decomposed as:

$$\hat{\mathbf{y}} - f(\mathbf{x}_*) = a_*(\hat{f}(\mathbf{x}_*) - f(\mathbf{x}_*)) + (1 - a_*)(\mathbf{y}_{\text{wavg}} - f(\mathbf{x}_*)),$$

where $\mathbf{y}_{\text{wavg}} = \sum_{i=1}^{k}(\frac{a_i + a_{k+1}/k}{1-a_*})\mathbf{y}_i$ is the weighted average of context labels. The gradient of the instantaneous loss $\ell = (\hat{\mathbf{y}} - f(\mathbf{x}_*))^2$ with respect to the output $\hat{f}(\mathbf{x}_*)$ is:

$$\frac{\partial \ell}{\partial \hat{f}(\mathbf{x}_*)} = 2(\hat{\mathbf{y}} - f(\mathbf{x}_*))\frac{\partial \hat{\mathbf{y}}}{\partial \hat{f}(\mathbf{x}_*)} = 2a_*(\hat{\mathbf{y}} - f(\mathbf{x}_*)).$$

Substituting the decomposed error, the gradient expression guiding the learning of $\hat{f}$ becomes:

$$\frac{\partial \ell}{\partial \hat{f}(\mathbf{x}_*)} = 2\big[a_*^2(\hat{f}(\mathbf{x}_*) - f(\mathbf{x}_*)) + a_*(1 - a_*)(\mathbf{y}_{\text{wavg}} - f(\mathbf{x}_*))\big].$$

This gradient reveals a tension between learning from the parametric path (first term) and being influenced by the ICL path (second term). We analyze how this dynamic plays out as the attention parameters converge in each regime.

1. **Random-Context:** In this regime, $\mathbf{y}_{\text{wavg}}$ is an average of labels from dissimilar examples, making it a poor estimator of $f(\mathbf{x}_*)$. The term $(\mathbf{y}_{\text{wavg}} - f(\mathbf{x}_*))$ is therefore large and noisy. Thus the initial updates for $\hat{f}$ may not be in the right direction. However, as training progresses and a_* approaches 1, the factor $a_*(1 - a_*)$ in the second term vanishes. This silences the "polluting" influence of the noisy context. The gradient becomes dominated by the first term, $2a_*^2(\hat{f} - f) \approx 2(\hat{f} - f)$, which is precisely the gradient for a standard MSE objective. Thus, the model's learning shifts entirely to improving $\hat{f}$.
2. **Similar-Context:** Here, the context examples are highly relevant, so $\mathbf{y}_{\text{wavg}}$ is an excellent estimator of $f(\mathbf{x}_*)$ from the beginning. The term $(\mathbf{y}_{\text{wavg}} - f(\mathbf{x}_*))$ is small. So even though the initial updates for $\hat{f}$ may be good, the model can achieve low loss quickly by relying on ICL, which it does by learning to drive $a_* \to 0$. As a_* decreases, both the a_*^2 and $a_*(1 - a_*)$ factors in the gradient shrink towards zero. The learning signal for $\hat{f}$ is rapidly suppressed, effectively halting its training as the model commits to its ICL strategy.
3. **One-Near-Context:** $\mathbf{y}_{\text{wavg}}$ may initially be imperfect. And as the model learns to increase θ_1, the weights concentrate on the single near example $j^\star$, and $\mathbf{y}_{\text{wavg}}$ rapidly converges to $y_{j^\star} \approx f(\mathbf{x}_*)$. Concurrently, the model drives $a_* \to 0$. As in the previous case, this causes both terms in the gradient to vanish, suppressing updates to $\hat{f}$ once the model learns to identify and copy the relevant context.
4. **Random-Context + Similar-Context:** The training process for $\hat{f}$ becomes selective and efficient. The model is exposed to a mix of *Random-Context* and *Similar-Context*.
   - When presented with an *Random-Context*, the dynamics from point (1) apply. The attention system learns to set $a_* \approx 1$, providing a strong, clean gradient signal to train $\hat{f}$.
   - When presented with a *Similar-Context*, the dynamics from point (2) apply. The attention system learns to set $a_* \approx 0$, and the gradient for $\hat{f}$ is suppressed.

   This demonstrates a sophisticated division of labor: the in-weights estimator $\hat{f}$ is trained almost exclusively on the subset of data where the context is uninformative and its parametric knowledge is actually needed.
5. **Contrastive-Context:** The training process for $\hat{f}$ becomes selective and efficient. The model is exposed to a mix of *Random-Context* and *One-Near-Context*.

- When presented with an *Random-Context*, the dynamics from point (1) apply. The attention system learns to set $a_* \approx 1$, providing a strong, clean gradient signal to train $\hat{f}$.
- When presented with a *One-Near-Context*, the dynamics from point (3) apply. The attention system learns to set $a_* \approx 0$, and the gradient for $\hat{f}$ is suppressed.

This demonstrates a sophisticated division of labor: the in-weights estimator $\hat{f}$ is trained almost exclusively on the subset of data where the context is uninformative and its parametric knowledge is actually needed.

B. Prompts for various tasks

B.1. Prompts for obtaining translations with in-context examples

```
Translate the source text from English to German.
Source: Redistribution is only possible if there is actually something
 ↪  produced to be redistributed.
Target:
```

```
Consider the following 4 translations from English to German.
Example 1:
Source: This is now to be given concrete shape in the proposal to give EU
 ↪  citizens the right of free movement and residence.
Target: Jetzt soll sie in dem Vorschlag über das Recht der Unionsbürger
 ↪  auf Freizügigkeit und Aufenthalt konkret ausgestaltet werden.

Example 2:
Source: These data can only be seen as striking, alerting us to the urgent
 ↪  need for humanitarian aid to these countries to be properly directed
 ↪  to the area of health and the provision of basic medical care to highly
 ↪  deprived communities.
Target: Diese Zahlen sind erschreckend und müssen für uns ein Alarmsignal
 ↪  sein, dass es dringend notwendig ist, die humanitäre Hilfe für diese
 ↪  Länder in angemessener Weise in den Bereich des Gesundheitswesens und
 ↪  der Bereitstellung von medizinischer Grundversorgung für die am
 ↪  meisten benachteiligten Gemeinschaften zu lenken.

Example 3:
Source: There are a number of actions we need to take in future in the
 ↪  area of budget control within the agencies.
Target: Es gibt eine Reihe von Maßnahmen, die wir in Zukunft im Bereich der
 ↪  Haushaltskontrolle in den Agenturen ergreifen müssen.

Example 4:
Source: It is precisely in the area of the environment that people's
 ↪  dissatisfaction has been most deeply felt.
Target: Gerade im Umweltbereich spürt man die Unzufriedenheit der
 ↪  Bevölkerung am deutlichsten.

Translate the source text from English to German.
Source: It is precisely in the area of budget policy that these three
 ↪  basic principles need to be given concrete expression.
Target:
```

B.2. Prompt to generate paraphrases

```
Generate 5 diverse paraphrases for the given input English sentence.
Before paraphrasing, carefully analyze the original sentences and keywords.
```

```
Ensure significant variation in phrasing, structure, and wording while
↪  preserving meaning.
The paraphrases should be such that their translation into another language
↪  could still be the same.
Only modify the grammar and sentence structure such that all the paraphrases
↪  should be translatable to a common sentence in another language.

### Inputs:
Original Sentences:
{sentences}

### Example:
1.
#### Input:
Original Sentences:
[
    "Sentence 1": "In New York City, the streets were alive with the sound of
    ↪  honking taxis and chatter from the crowded sidewalks.",
    "Sentence 2": "\"I can't believe it's already time for the annual
    ↪  reunion,\" Emily said, her voice filled with excitement.",
    "Sentence 3": "The Amazon rainforest is home to countless species of
    ↪  plants and animals, many of which have yet to be discovered.",
]

#### Example Paraphrases:
{{
    "Sentence 1": [
        "The streets of New York City buzzed with the noise of honking taxis
        ↪  and voices from the bustling sidewalks.",
        "In the heart of New York City, honking taxis and lively chatter
        ↪  filled the streets.",
        "New York City's streets were vibrant with the sounds of honking taxis
        ↪  and the chatter of crowded sidewalks.",
        "In the bustling streets of New York City, honking taxis and lively
        ↪  conversations created a lively atmosphere.",
        "The sound of honking taxis and the chatter from crowded sidewalks
        ↪  made the streets of New York City come alive."
    ],
    "Sentence 2": [
        "Emily exclaimed, her excitement evident, \"I can't believe it's time
        ↪  for the annual reunion already.\"",
        "With excitement in her voice, Emily said, \"I can't believe the
        ↪  annual reunion is here already.\"",
        "Emily's voice was filled with excitement as she said, \"I can't
        ↪  believe it's already time for the annual reunion.\"",
        "With a voice full of excitement, Emily remarked, \"I can't believe
        ↪  the annual reunion is already upon us.\"",
        "Filled with excitement, Emily exclaimed, \"I can't believe it's
        ↪  already time for the annual reunion!\""
    ],
    "Sentence 3": [
        "The Amazon rainforest shelters an untold number of plant and animal
        ↪  species, many still waiting to be discovered.",
        "Countless species of plants and animals call the Amazon rainforest
        ↪  home, many of which remain undiscovered.",
```

```
        "Home to countless species of plants and animals, the Amazon
        ↪  rainforest holds many that are yet to be discovered.",
        "The Amazon rainforest is a habitat for innumerable species of plants
        ↪  and animals, many of which are still unknown.",
        "Many species of plants and animals, yet to be discovered, thrive in
        ↪  the Amazon rainforest."
    ],
}}

### Output Instructions:
- Generate exactly 5 paraphrases.
- Ensure correct JSON output syntax like the example
- Always escape quotes within a paraphrase (like in example 2)
```

C. Datasets and Additional Tasks

C.1. Additional Setup Details

For training, we adopted LoRA with a rank of 16, a scaling factor of $\alpha = 32$, and a dropout rate of 0.05. Only the attention projection matrices (except the output projection matrix) were trained. We also trained all the attention projection matrices along the MLP layer matrices. The results for this ablation can be found in Appendix D.6. Training was performed with a batch size of 2 using Adam with a learning rate of 2×10^{-4} and a linear decay learning rate scheduler. To obtain paraphrases for Contrastive-Context, we employed the `gemini-2.0-flash-lite` model. For Similar-Context, we used the `all-MiniLM-L6-v2` model to further filter the training instances, retaining only those with an average similarity of the few-shot examples with $\mathbf{x}_*$ greater than 0.5.

C.2. Low Resource Translation

	Train		ID Test		OOD Test	
LP	Source	Size	Source	Size	Source	Size
En-De	Europarl	50000	Flores+	1012	EMEA	500
En-Hi	Samanantar	50000	Flores+	1012	Judicial	500
En-Ta	Samanantar	50000	Flores+	1012	Tanzil	500
En-Lt	Europarl	50000	Flores+	1012	EMEA	500

Table 2. Datasets Used for MT. Flores+ refers to Flores-200. **Sources:** Europarl, Tanzil, and EMEA from Tiedemann (2012); Samanantar from Ramesh et al. (2022); Flores+ from NLLB Team et al. (2024); Judicial from Kunchukuttan et al. (2018).

Dataset	Link
Europarl	`Helsinki-NLP/europarl`
Tanzil	`Helsinki-NLP/tanzil`
EMEA	`Helsinki-NLP/emea`
Samanantar	`ai4bharat/samanantar`
Flores+	`openlanguagedata/flores_plus`
Judicial	`cfilt/iitb-english-hindi`

Table 3. Links to datasets used

C.3. Text-to-SQL

The paraphrases for this task are obtained using OpenAI-O3-mini. Given an NL query, SQL pair $(\mathbf{x}, \mathbf{y})$, we prompted the LLM to generate first a paraphrased SQL $\tilde{\mathbf{y}}$ that differs from $\mathbf{y}$ only in the mention of literal constants, and then generating the corresponding NL question $\tilde{\mathbf{x}}$ describing the SQL $\tilde{\mathbf{y}}$. An example paraphrased pair appears below:

$\mathbf{x}$: Is molecule TR183 known to be a carcinogen?
$\mathbf{y}$: SELECT T.label FROM molecule AS T WHERE T.molecule_id = 'TR183'

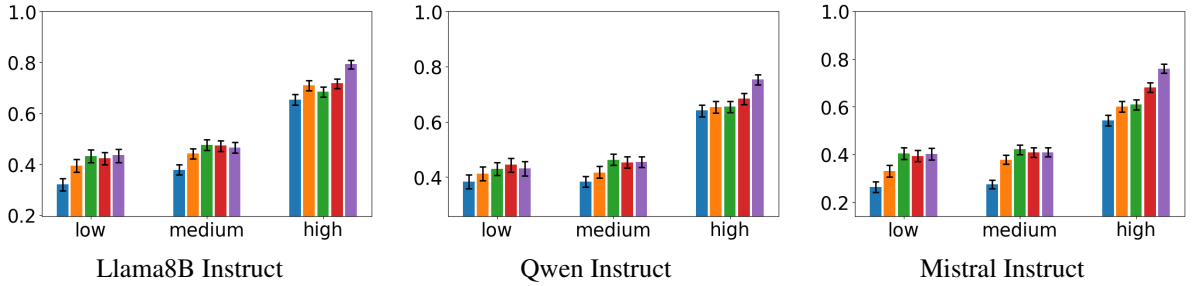
Prompt	Prompt string: $\mathbf{x}^1 : \mathbf{y}^1 \ \mathbf{x}^2 : \mathbf{y}^2 \ \mathbf{x}^3 : \mathbf{y}^3 \ \mathbf{x}_*$
Example $m = 3, c = 2$	`ACB: rijjpr CAB: jjriwp ABC: rtprjh BCA:`
Example $m = 2, c = 4$	`AC: ririjhjh CA: jjhhriir AB: rttrprrp BA:`

.

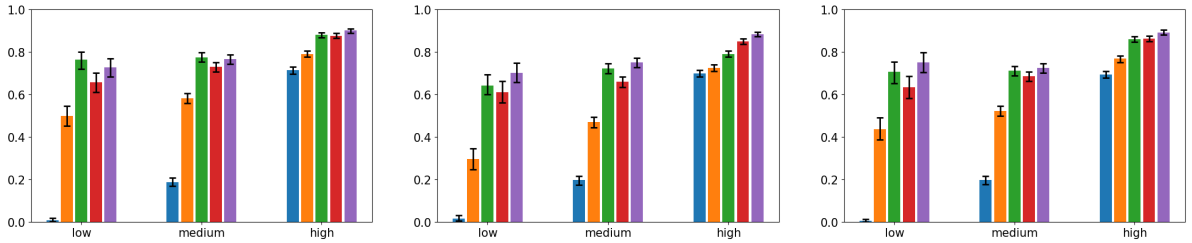
$\tilde{\mathbf{x}}$: Molecule KAC16 is carcinogenic. Yes or No?

$\tilde{\mathbf{y}}$: SELECT T.label FROM molecule AS T WHERE T.molecule_id = 'KAC16'

The prompt to the LLM comprises of generic natural language instructions, followed by a description of the schema and metadata of the database queried, followed by the in-context examples, and then the current test question $\mathbf{x}_*$. Following standard practice, instead of providing the entire database metadata, we filter a subset of schema information (relevant table and column names) using an open source schema filtering tool [2] (Li et al., 2023), (Li et al., 2024a).



(a) Execution accuracy for the Text-to-SQL task.



(b) Exact-Match accuracy for semantic parsing on MTOP

Figure 6. Accuracy (Y-axis) against three levels of similarity (X-axis). The similarity ranges here are - (a) Low: $0-33$, Medium: $33-67$, High: $67-100$, (b) Low: $0-0.33$, Medium: $0.33-0.67$, High: $0.67-1$. Error bars show 95% confidence intervals.

C.4. Synthetic Alignment Reasoning Task

We use the dataset as discussed by (Agarwal & Sarawagi, 2025) for reasoning about the alignment between two synthetic sequences. This task first defines a vocabulary $\mathcal{V}$ of symbols, each symbol $\sigma \in \mathcal{V}$ is associated with a probabilistic finite state automata (PFA) to generate sequences of length c over elements over its own vocabulary. The input sequence $\mathbf{x}$ contains m symbols chosen from $\mathcal{V}$, the output sequence $\mathbf{y}$ is obtained by a fixed (unknown) permutation of $\mathbf{x}$, and then sampling a length c sequence from the FSA for each element in the permuted $\mathbf{x}$. More details can be found in (Agarwal & Sarawagi, 2025). In order to correctly generate the output for a sequence, the model needs to reason about the input-output alignment and then the PFA of each token in $\mathcal{V}$. Examples appear below:

For this task we choose $|\mathcal{V}| = 25$ and ran under two settings of sequence lengths ($m = 4,\ c = 4$) and ($m = 6,\ c = 6$) with the alignment as a permutation. The training is done on 200 instances for 5 epochs. In the original paper, each

[2] https://github.com/RUCKBReasoning/text2sql-schema-filter

instance sampled its own PFA to simulate an infinite mixture. In this paper, since our goal is to develop ICL-IWL mixture for a fixed task, the PFAs per symbol were fixed throughout the entire process. We sampled in-context examples at different similarity levels as follows: the Random-Context data is sampled using a vocabulary of size 25, and the Similar-Context data is created by using a vocabulary of size 12. For the Contrastive-Context data the paraphrases are generated by taking the vocabulary as the set of letter used in the target example, $\mathbf{x}_*$. Hence, the paraphrase is simple a shuffling of the letters in $\mathbf{x}_*$.

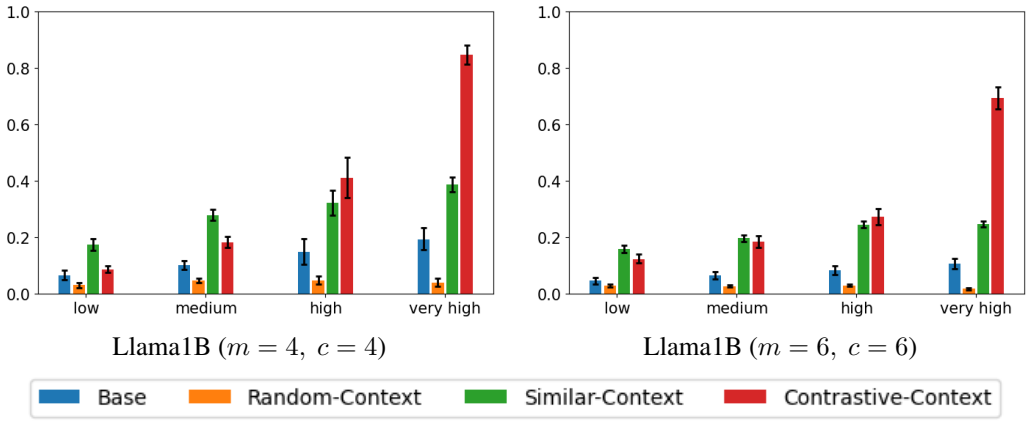


Figure 7. X-axis: Level of maximum Jaccard similarity between the sets of letters used in the target and the context examples. Y-axis: Average maximum accepted length of respective subsequence in y for the PFAs of each letter in $\mathbf{x}_*$. Random-Context performs worse than the baseline irrespective of the similarity level, probably because it has learnt the PFAs before learning the alignment, and thus it chooses a PFA at random for every letter in $\mathbf{x}_*$. On the other hand, the baseline model performs better because it blindly copies an ICL target example sequence corresponding to an ICL context example having some symbols in the same position as $\mathbf{x}_*$. Although Contrastive-Context performs second only to Similar-Context in the low and medium similarity regions, Contrastive-Context outperforms all the models in the high and very high similarity regions.

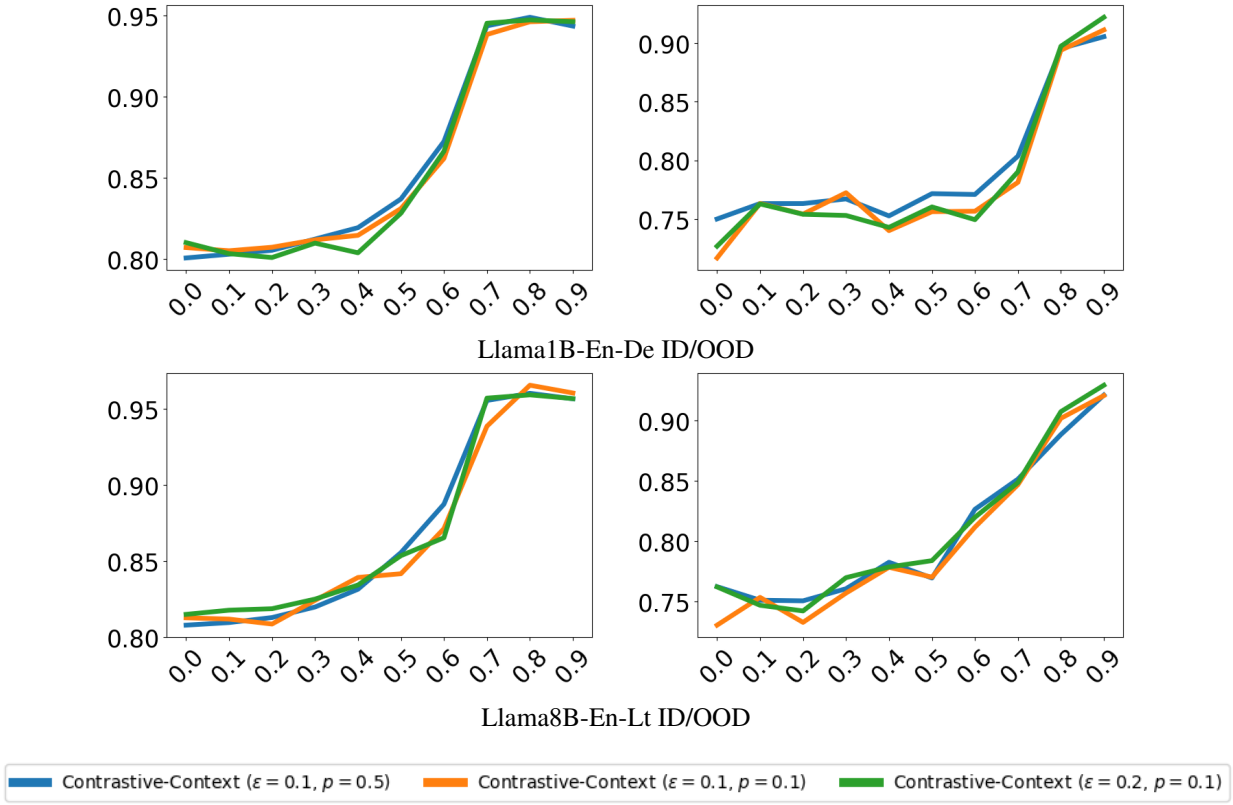


Figure 8. Contrastive-Context is robust to specific choices of ϵ and p as long as sufficient contrast is maintained.

For example, in Appendix D.2, what does it mean if the performance is low in the medium similarity range but high in small and large context-target similarity, or in Sec 3.2, why does random-context has a different behavior.

D. Ablations

D.1. Experiments with non-zero ϵ and p

We present robustness of Contrastive-Context to alternative choice of ϵ and p parameters in Figure 8.

D.2. Importance of training with different similarity levels

Contrastive-Context includes in-context examples at three levels of similarity — randomly selected examples which are of low similarity, naturally occurring top-k most similar examples from the training pool which are typically of medium similarity levels, and highly similar examples obtained by paraphrasing the target. We establish the necessity of employing these multiple similarity levels during training by creating two ablations: first, we remove the highly similar paraphrased examples by setting $\epsilon = 0$, and second, we remove the natural similar examples by setting $p = 0$.

Figure 9 shows evaluation on test instances created with mixed similarity levels as described in Section 4 (Evaluation Setup). We observe that test instances with high target-context similarity suffer on models trained with $\epsilon = 0$. This is evidenced by the dip in accuracy of the **orange** line compared to the original Contrastive-Context (blue line) in the right quarter of the X-axis). Likewise, the model trained without naturally occurring Top-K pairs ($p = 0$), performs poorly on test instances where IC examples are at medium levels of similarity with the target. This is seen by the dip in accuracy of the **green** line in the middle part of the X axis.

Overall, we observe the trend that it is important for IC-Train to be trained at multiple similarity levels in order to provide the best accuracy under all levels of relatedness of the test example with the context.

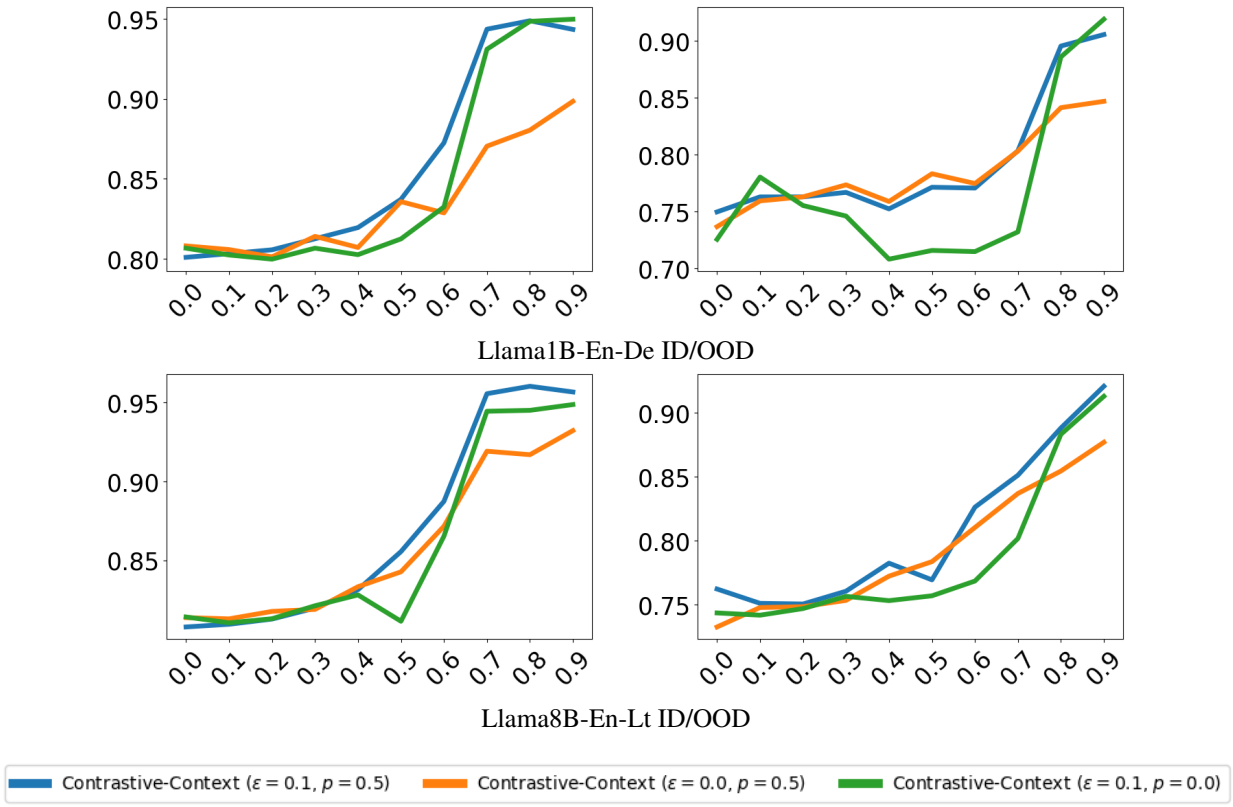


Figure 9. Ablations on Contrastive-Context to establish the importance of training with different similarity levels: X-axis is target-example similarity and Y-axis is accuracy. Removing highly similar examples obtained via paraphrasing ($\epsilon = 0$) causes test accuracy in the high similarity range to suffer. Removing natural similar examples ($p = 0$) causes accuracy in the medium range to suffer.

D.3. Intra-Context Vs Inter-Context Contrast

Contrastive-Context creates contrasts in similarity levels both amongst examples within a context, and across contexts. Here we present ablations to understand the importance of contrasts within a context by comparing with a baseline that mixes Random-Context and Similar-Context to preserve only inter context contrast. We present results in Figure 10. On these tasks, the Random-Context and Similar-Context mixture (**green** bar) are as good as Contrastive-Context (**blue** bar) in the low and medium similarity regime. In the high similarity regime, because of the absence of paraphrases, this mixture looses out. When we remove the paraphrases from Contrastive-Context (**orange** bar) the two methods are equivalent on this task. However, conceptually, a model that mixes Random-Context and Similar-Context could learn to

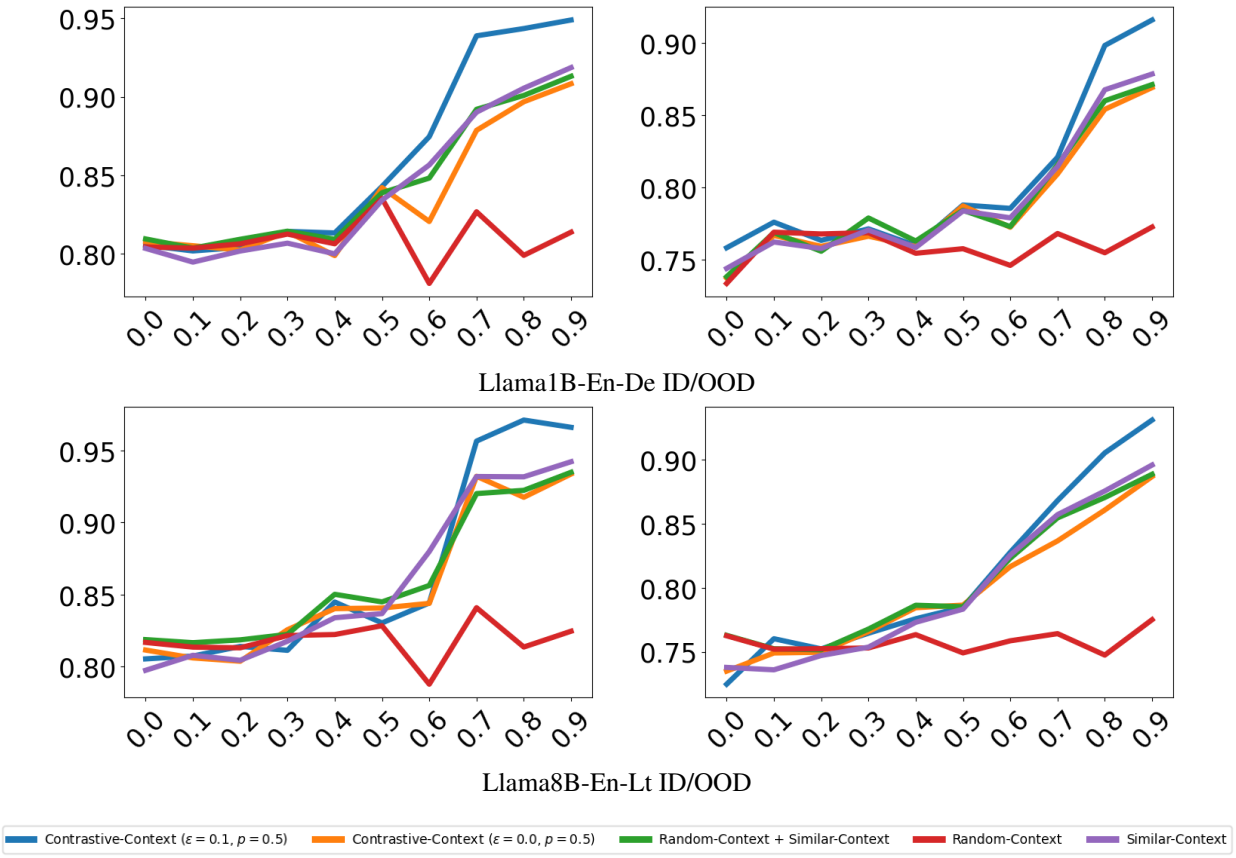


Figure 10. Comparing a Random-Context+Similar-Context mixture with Contrastive-Context that additionally also creates contrasts within a context, and other baselines Random-Context and Similar-Context.

swing between ignoring the context (IWL) and blindly copying from the context based on aggregated similarity with the context as we show in our theoretical analysis. This forms a IWL+Copy mixture, which might not perform well when tested with only a subset of context examples as relevant. Contrast within a context promotes true in-context learning where similarity between the x-s determines which y-s are copied. In real-life limited data regimes, the top-K similar examples may differ in their similarity to the target, and thus Random-Context+Similar-context may behave like Contrastive-Context like we observed above.

D.4. Synthetic Paraphrases for Data Augmentation Vs Contrast

An interesting question is whether the observed accuracy boost with synthetic paraphrased examples is due to increased contrast, or vanilla data augmentation. To answer this question, we added the synthetic paraphrases for all target training examples to the training pool as augmented data. We then invoked the Similar-Context method on this augmented dataset. A comparison appears in Figure 11. We observe that with such data augmentation Similar-Context does perform well on test data in the high similarity range but shows a huge drop in accuracy in the low similarity range. This is because Similar-Context goes for the Top-K most similar examples, and all training instance contain a paraphrase in

the context causing no incentive for IWL to develop. Contrastive-Context's use of paraphrases in a contrastive setting is important to develop IWL+ICL mixtures.

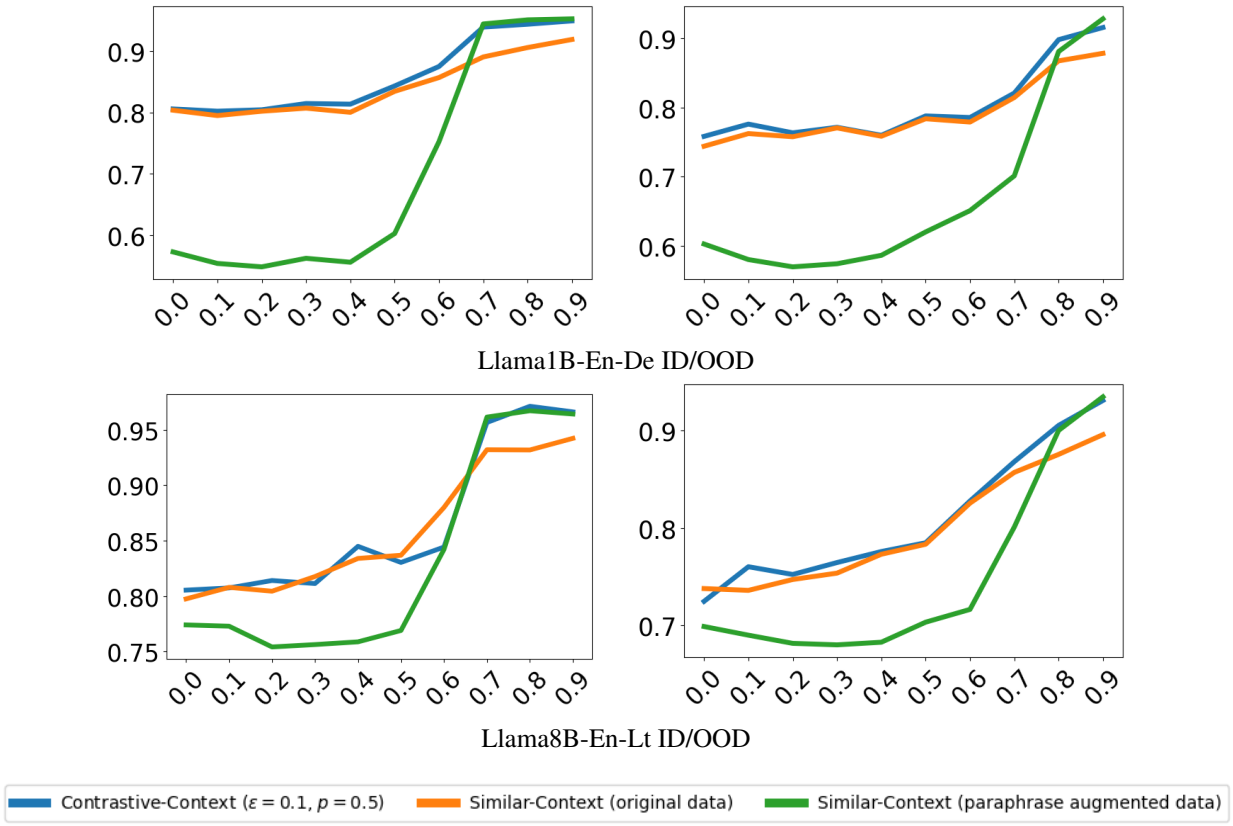


Figure 11. Comparing Similar-Context on original data and on data augmented with paraphrases. X-axis is similarity of test target instance with in-context examples, and Y axis is accuracy. Contrastive-Context uses paraphrases selectively to create contrast, whereas Similar-Context greedily selects Top-K most similar examples. While performance of Similar-Context improves in the high similarity range, it suffers in the low similarity range because IWL does not develop with highly similar context.

D.5. Experiments with various paraphrasing models

Although our proposed method relies on an external paraphrasing model (specifically, `gemini-2.0-flash-lite`), we demonstrate that the choice of the paraphraser has negligible impact on overall performance. To validate this, we conduct additional experiments using two alternative models `gemini-2.5-flash-lite` as an example of a strong model, and `Qwen2.5-7B-Instruct` as an open source, possibly weaker model. The results in Figure 12 shows that Contrastive-Context remains stable regardless of whether the paraphrasing model is open-source or proprietary.

D.6. Effect of LoRA Target Modules

We investigate the effect of applying LoRA to different subsets of transformer projections. In our main experiments, we keep the target modules as only the attention projection matrices (except the output projection matrix), but as an ablation, we also experiment by keeping all the attention projection matrices and the MLP layer matrices trainable. The comparison is shown in Figure 13. We observe that with additional trainable parameters, the overall performance increases, but the general trend of results across training regimes remains unchanged.

E. Additional Experiments Comparing Different Training Strategies

We present results of more model-language pair combinations that could not be fit in Figure 3 of the main paper in Figure 14.

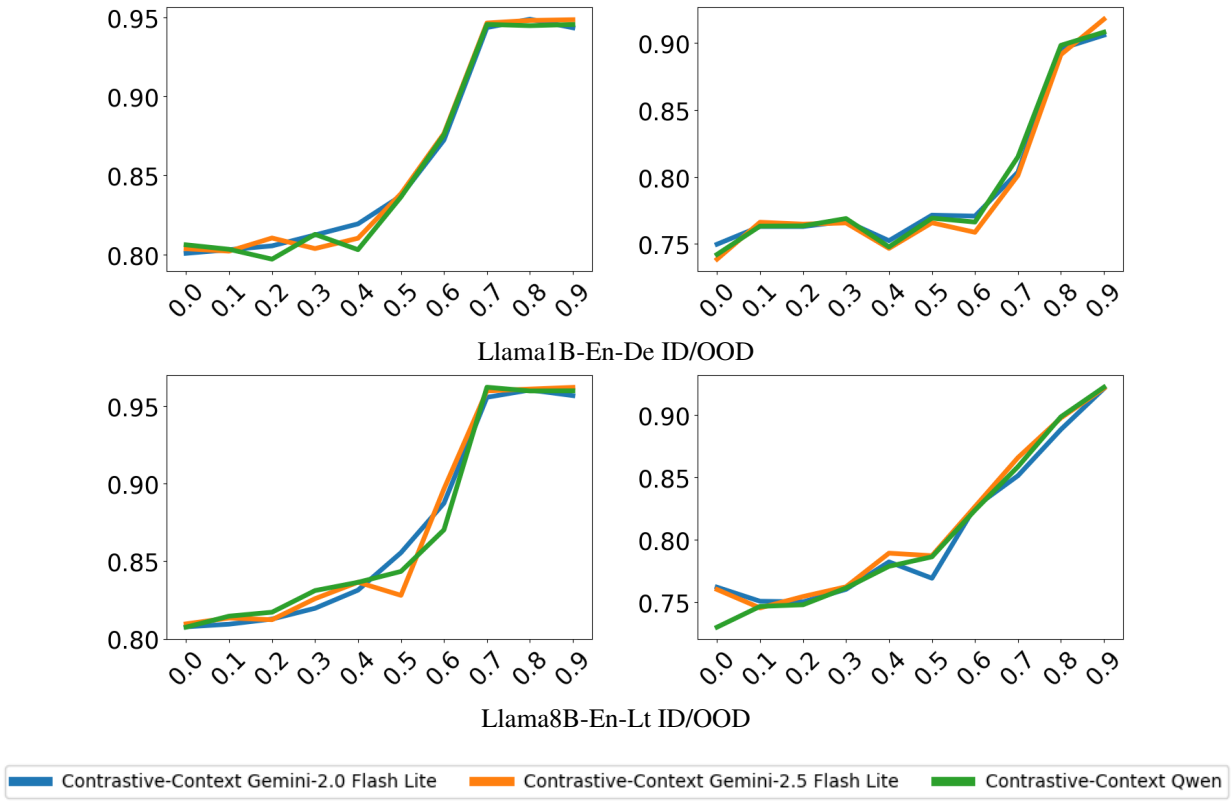


Figure 12. Comparing Contrastive-Context with highly similar examples created with various paraphrasing models. X-axis is target-example similarity and Y-axis is accuracy on the test data. Contrastive-Context is robust to the choice of the paraphrasing model.

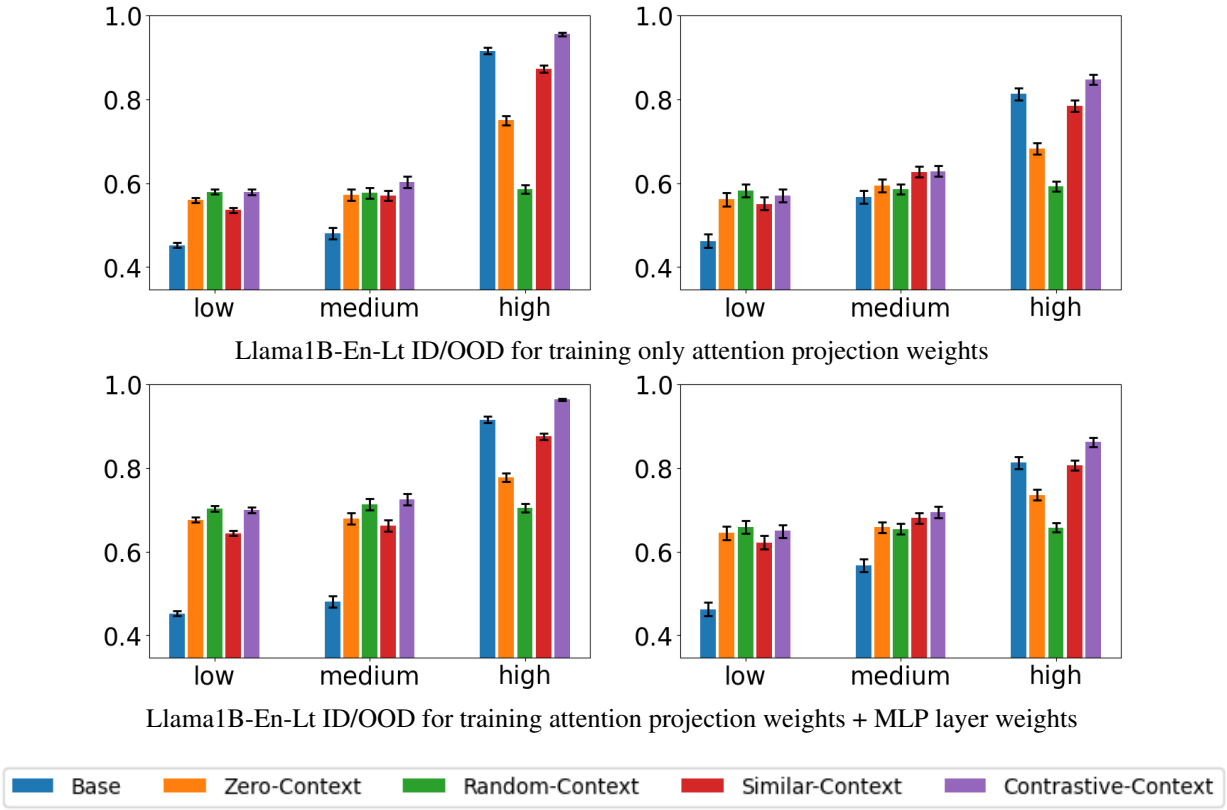
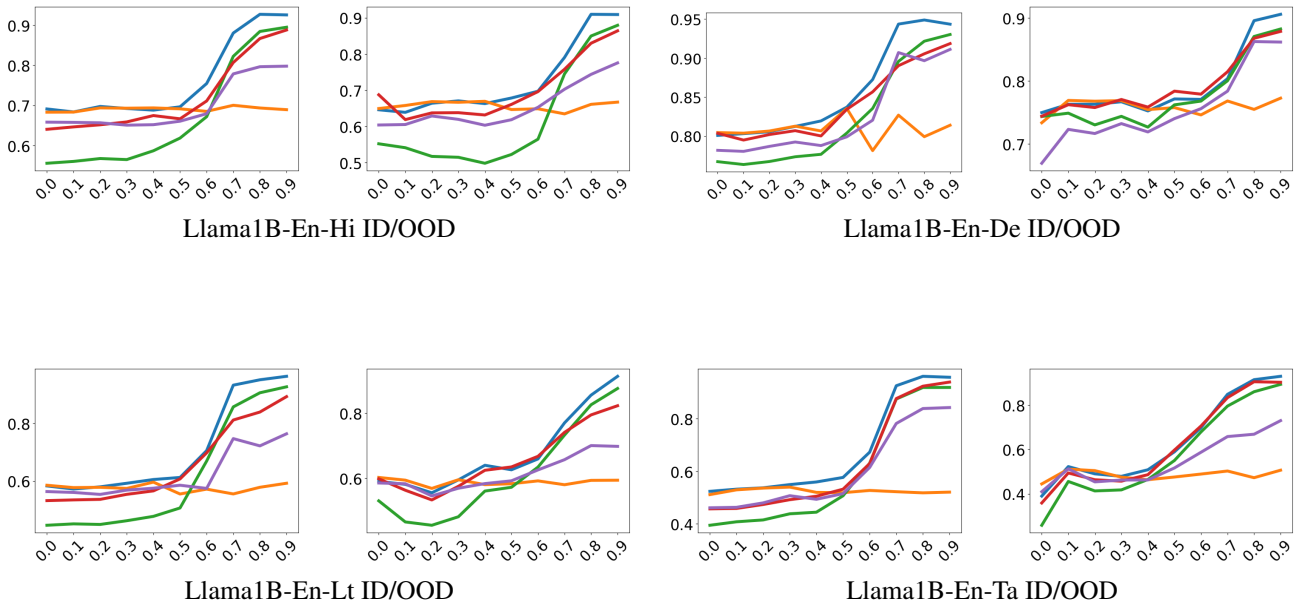
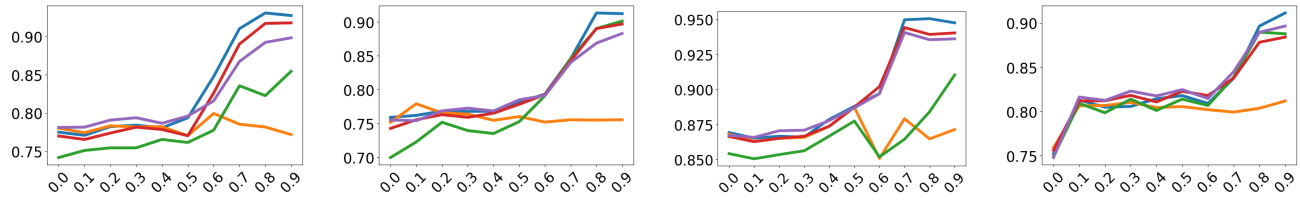


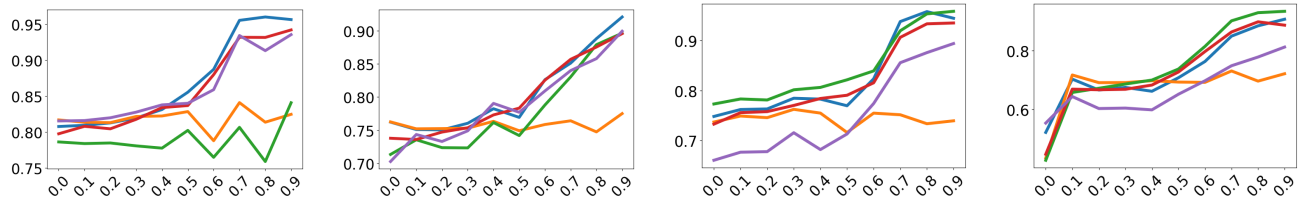
Figure 13. Comparing training only attention projection weights (except output projection) with training attention projection weights + MLP layer weights. X-axis: Level of maximum similarity of target with in-context examples. The similarity ranges here are - Low: $0 - 0.33$, Medium: $0.33 - 0.67$, High: $0.67 - 1$. Y-axis: Accuracy (COMET score).





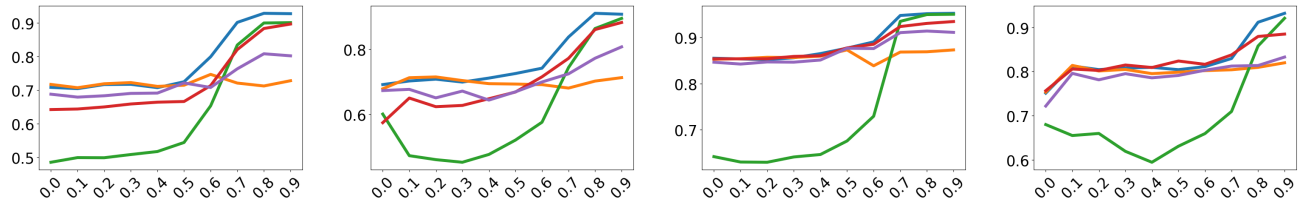
Llama8B-En-Hi ID/OOD

Llama8B-En-De ID/OOD



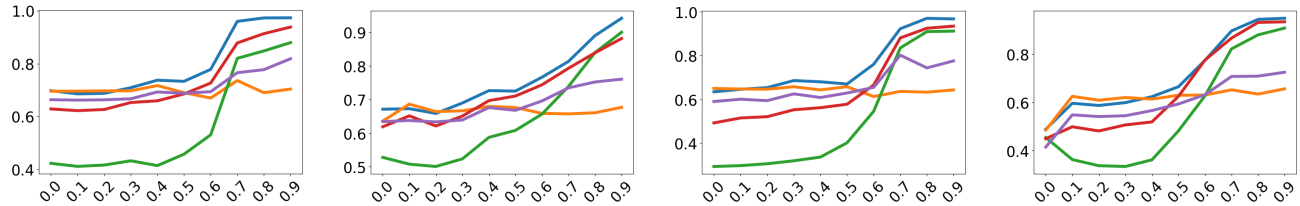
Llama8B-En-Lt ID/OOD

Llama8B-En-Ta ID/OOD



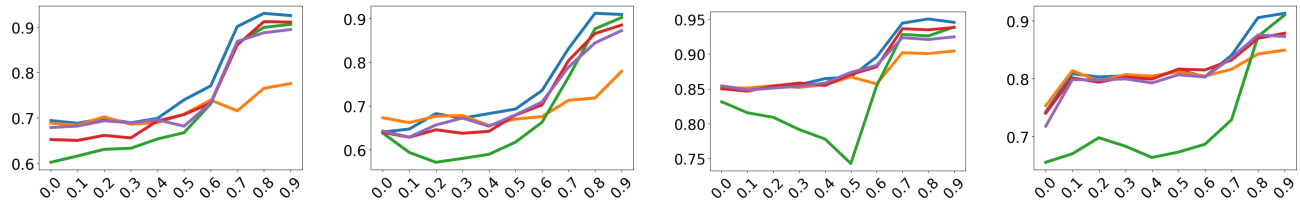
Mistral-En-Hi ID/OOD

Mistral-En-De ID/OOD



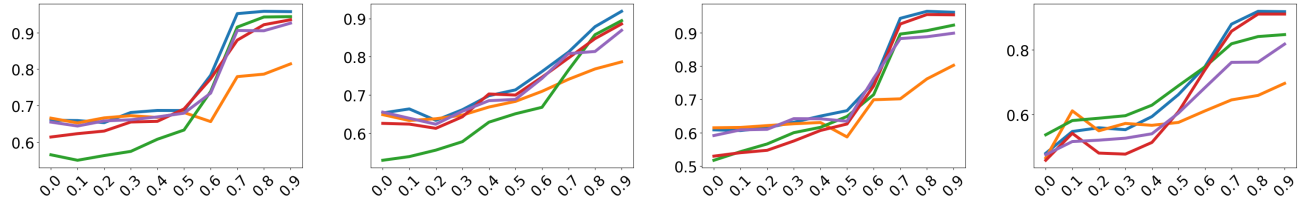
Mistral-En-Lt ID/OOD

Mistral-En-Ta ID/OOD



Qwen-En-Hi ID/OOD

Qwen-En-De ID/OOD



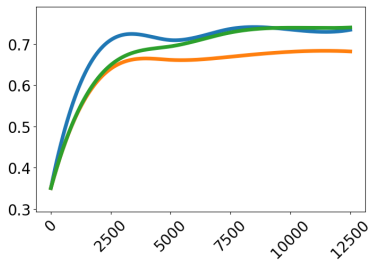
Qwen-En-Lt ID/OOD

Qwen-En-Ta ID/OOD

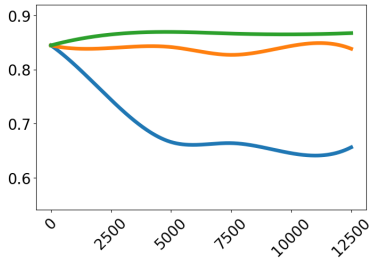
Figure 14. All 32 plots for four different models tested on four different language pairs with ID and OOD distribution. X-axis is target-example similarity and Y-axis is accuracy.

F. Additional Experiments on Learning Dynamics

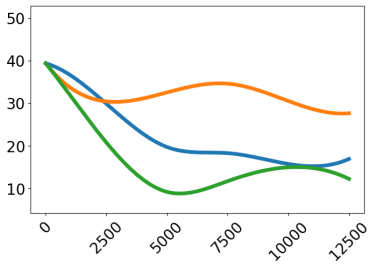
We present results of more model-language pair combinations that could not be fit in Figure 4 of the main paper in Figure 15.



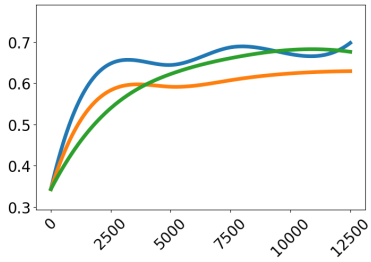
IWL score: 1B-En-Hi-ID



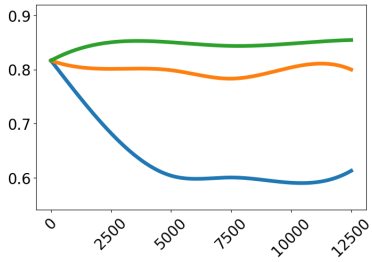
ICL score: 1B-En-Hi-ID



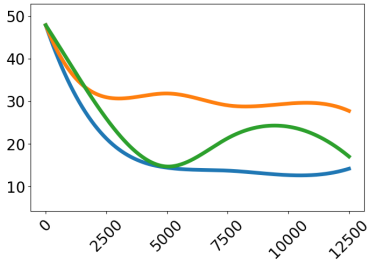
Copy score: 1B-En-Hi-ID



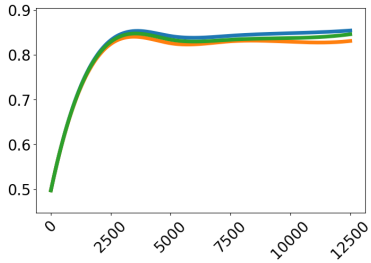
IWL score: 1B-En-Hi-OOD



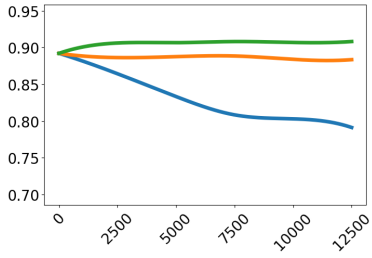
ICL score: 1B-En-Hi-OOD



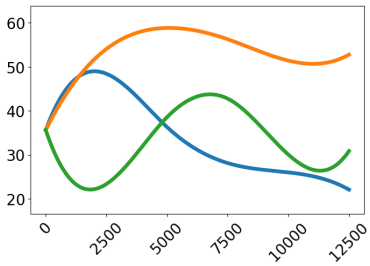
Copy score: 1B-En-Hi-OOD



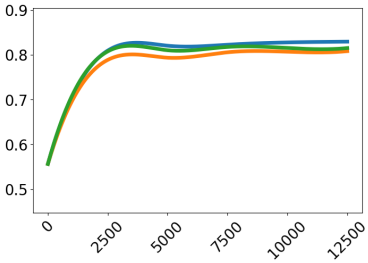
IWL score: 1B-En-De-ID



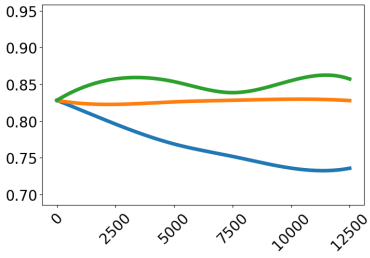
ICL score: 1B-En-De-ID



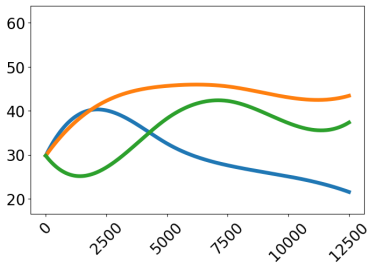
Copy score: 1B-En-De-ID



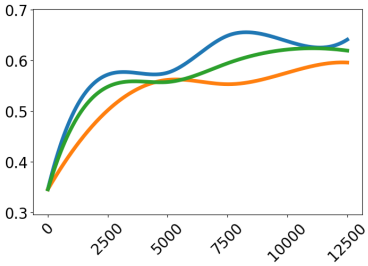
IWL score: 1B-En-De-OOD



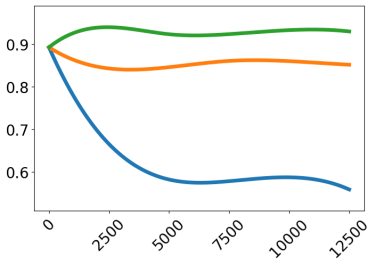
ICL score: 1B-En-De-OOD



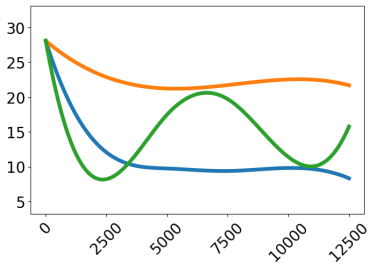
Copy score: 1B-En-De-OOD



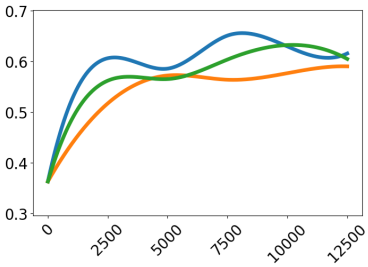
IWL score: 1B-En-Lt-ID



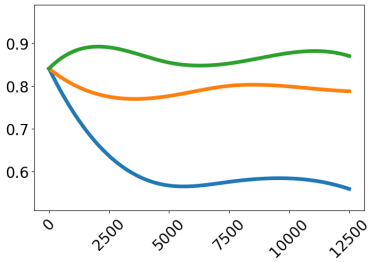
ICL score: 1B-En-Lt-ID



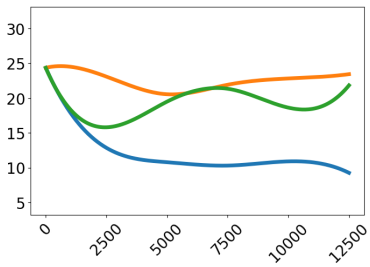
Copy score: 1B-En-Lt-ID



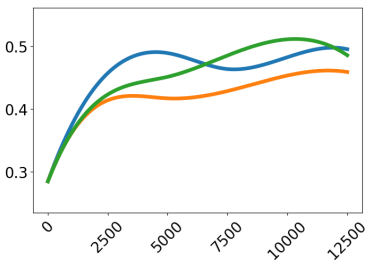
IWL score: 1B-En-Lt-OOD



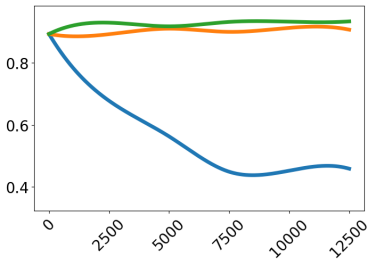
ICL score: 1B-En-Lt-OOD



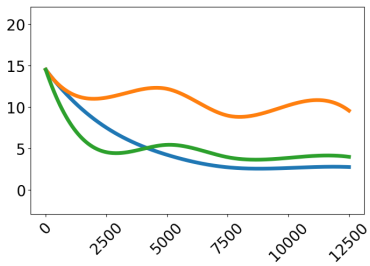
Copy score: 1B-En-Lt-OOD



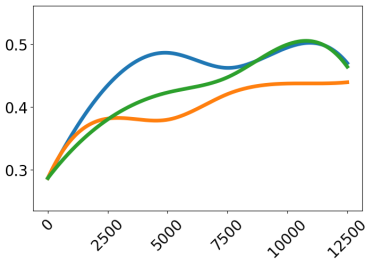
IWL score: 1B-En-Ta-ID



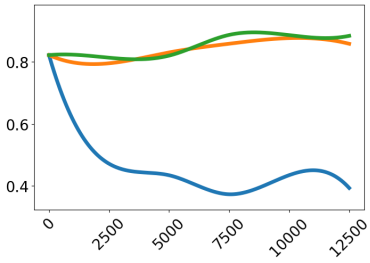
ICL score: 1B-En-Ta-ID



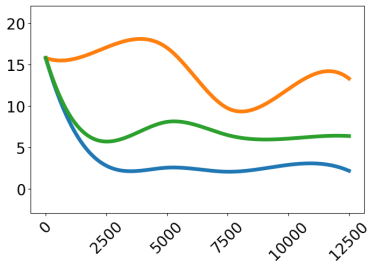
Copy score: 1B-En-Ta-ID



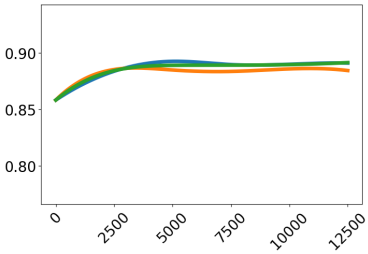
IWL score: 1B-En-Ta-OOD



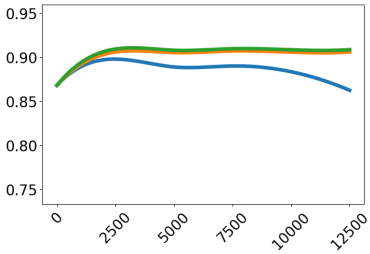
ICL score: 1B-En-Ta-OOD



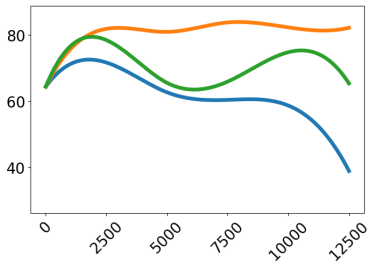
Copy score: 1B-En-Ta-OOD



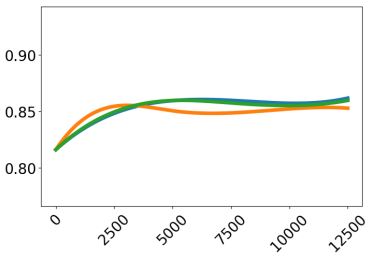
IWL score: 8B-En-De-ID



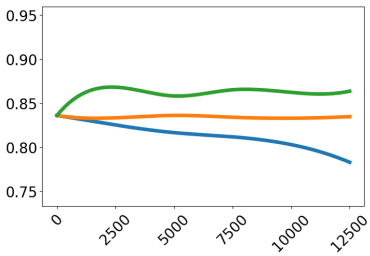
ICL score: 8B-En-De-ID



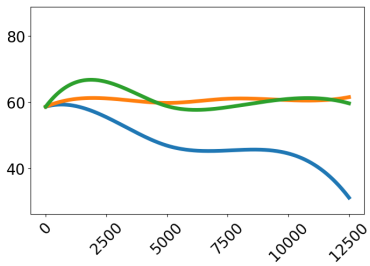
Copy score: 8B-En-De-ID



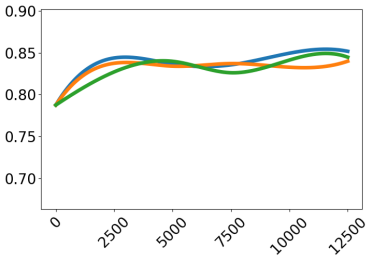
IWL score: 8B-En-De-OOD



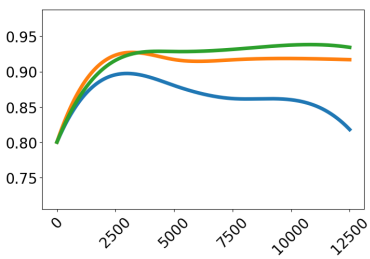
ICL score: 8B-En-De-OOD



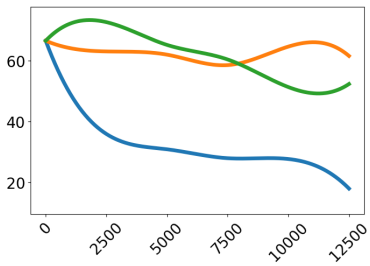
Copy score: 8B-En-De-OOD



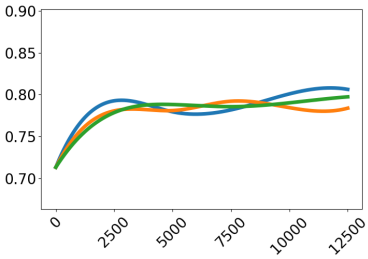
IWL score: 8B-En-Lt-ID



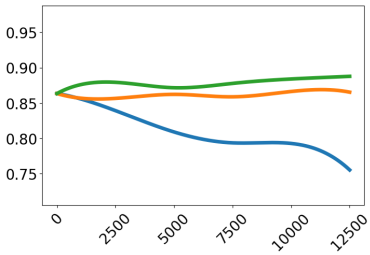
ICL score: 8B-En-Lt-ID



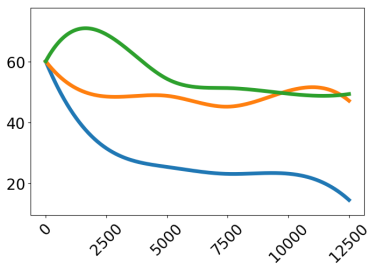
Copy score: 8B-En-Lt-ID



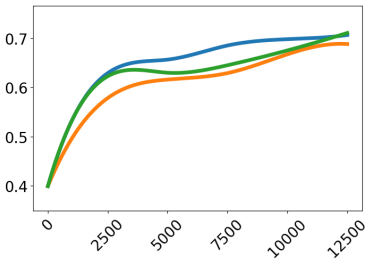
IWL score: 8B-En-Lt-OOD



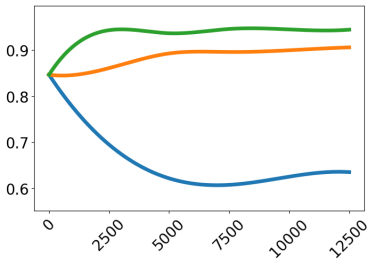
ICL score: 8B-En-Lt-OOD



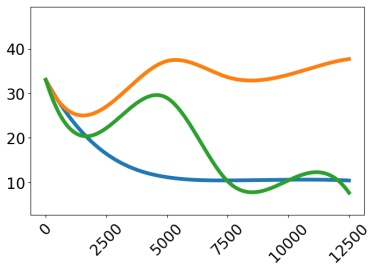
Copy score: 8B-En-Lt-OOD



IWL score: Mistral-En-Lt-ID



ICL score: Mistral-En-Lt-ID



Copy score: Mistral-En-Lt-ID

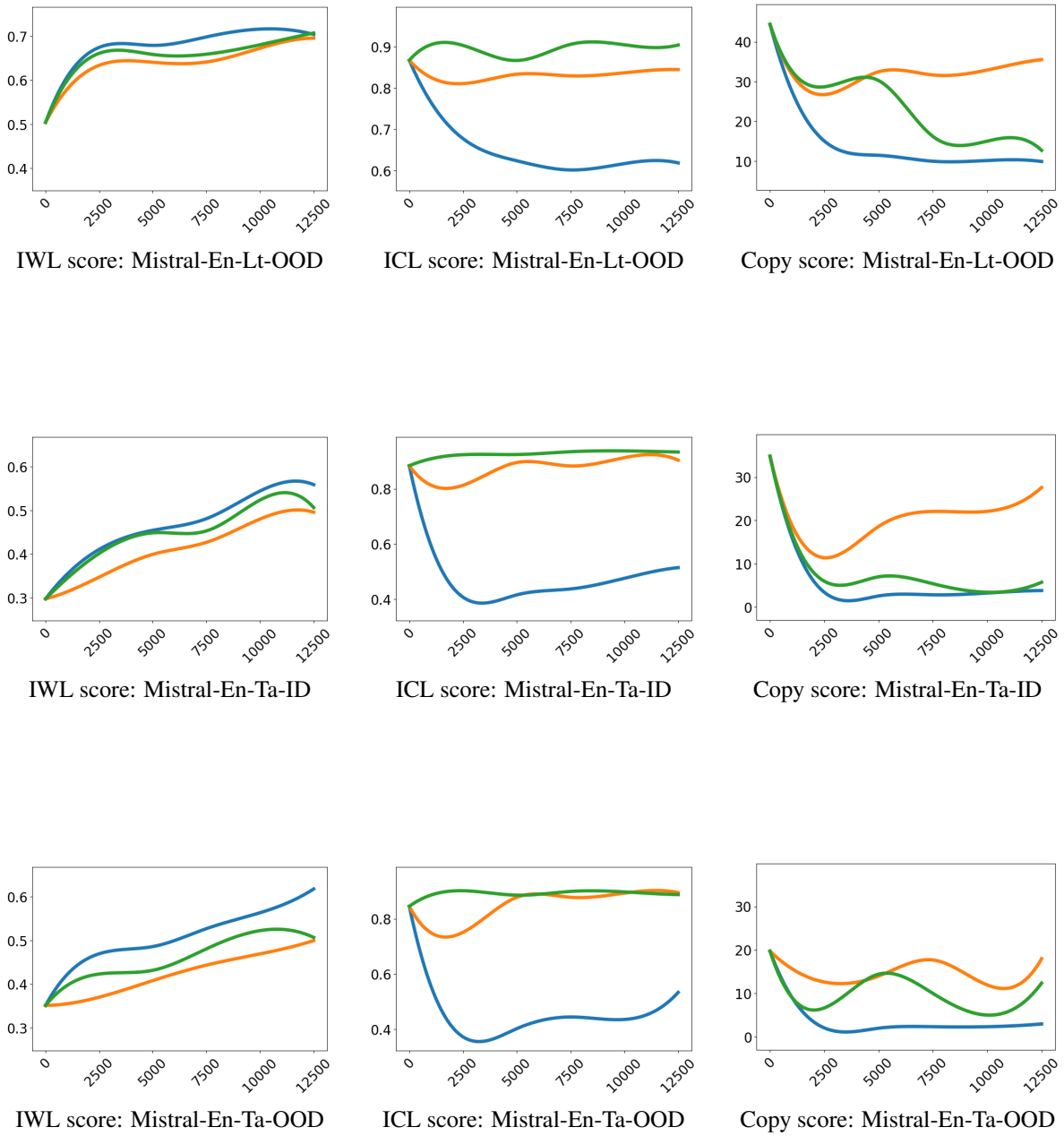


Figure 15. Learning dynamics plots for other models and datasets. X-axis is training steps and Y-axis denotes scores of one of the three probes.

F.1. Learning Dynamics (Unsmoothed)

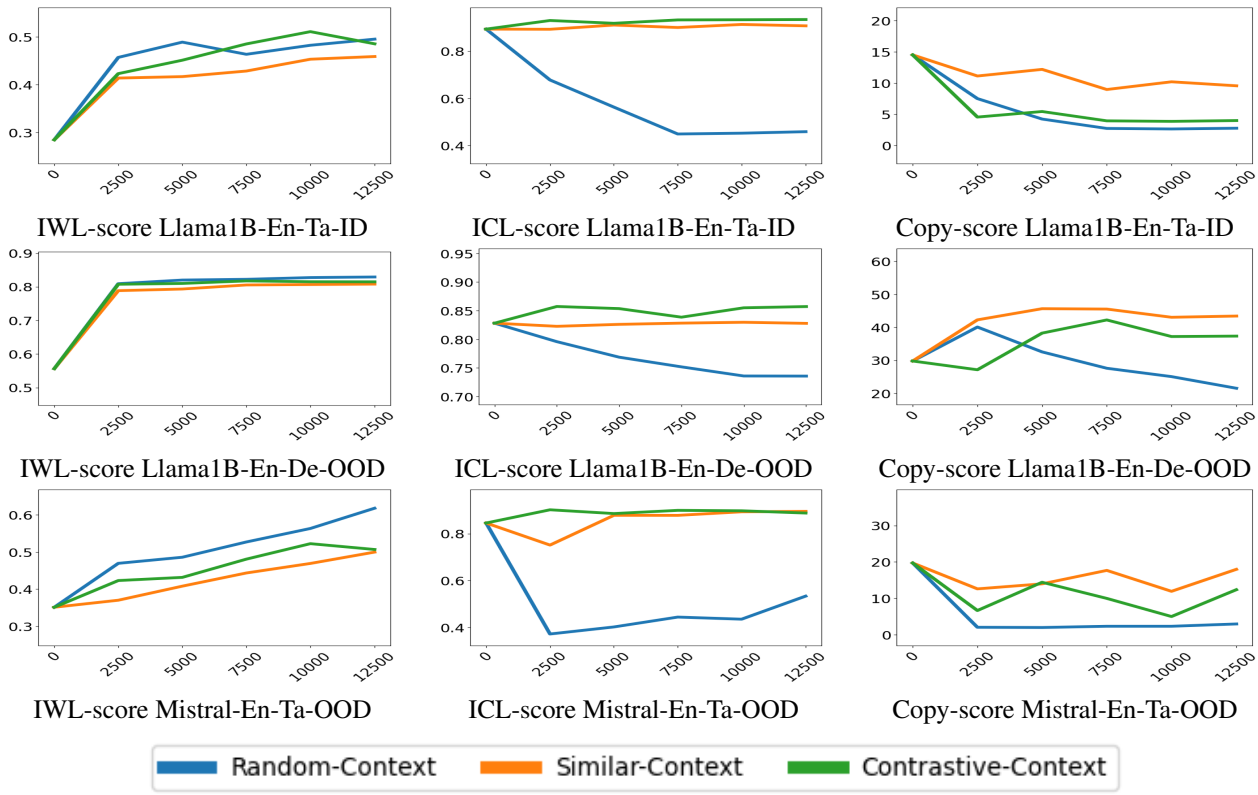


Figure 16. Unsmoothed learning dynamics plots for a few models and datasets. X-axis is training steps and Y-axis denotes scores of one of the three probes.