On the Relationship between the Choice of Representation and In-context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) allows large language models to adapt to new tasks via prompt demonstrations without parameter updates. While ICL mirrors several aspects of supervised learning, it requires labels to be instantiated as vocabulary tokens, making label representation a critical hyperparameter. We propose that label representation acts as an initialization for ICL. To test this hypothesis, we develop an algorithm that systematically enumerates label sets with varying degrees of semantic relevance based on the model's priors. We then evaluate ICL trajectories across multiple classification tasks, model families, scales, and demonstration counts. Our analysis reveals that label representations influence learning dynamics: the performance ranking of representations remains constant regardless of demonstration count, and the efficiency of learning depends on the interplay between the initial representation and model scale. Our study demonstrates that ICL should be viewed as a unified learning system whose behavior is jointly driven by representation, language model, and demonstrations.

1 Introduction

Large language models can adapt to new tasks from just a few examples provided directly in the prompt without any parameter updates. This capability, known as in-context learning (ICL), has proven remarkably effective: given several input-output demonstrations, a model can infer the underlying task and generalize to new inputs (Brown et al., 2020).

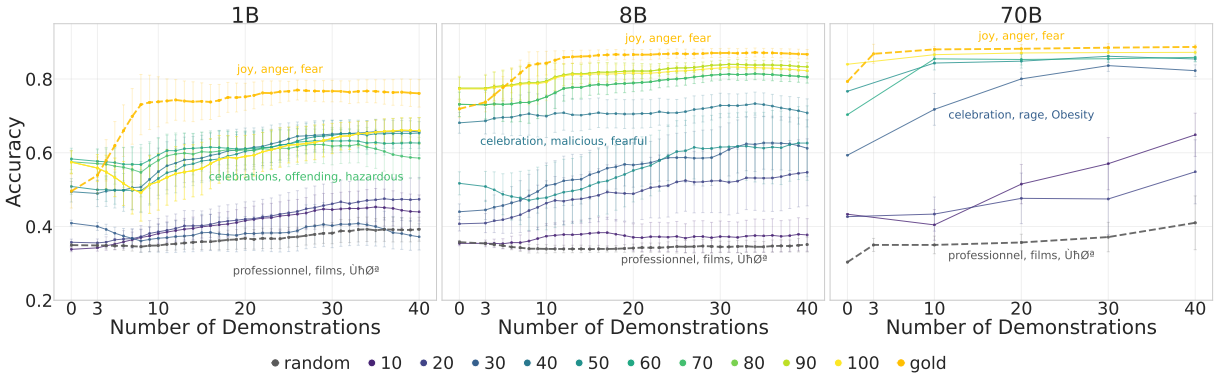
The behavior of ICL bears a striking resemblance to that of classical machine learning systems. There, performance is jointly determined by the learning algorithm, the training examples, and inductive biases introduced through architecture, initialization, and hyperparameters. These components appear to have natural analogues in ICL, where the language model plays the role of the learning algorithm, the demonstrations act as training examples (Chen et al., 2023; Agarwal et al., 2024; Bertsch et al., 2025), and the prompt representation introduces inductive biases analogous to hyperparameters. In traditional machine learning, labels are typically represented as one-hot vectors, whereas in ICL they must be instantiated as tokens in the vocabulary. As a result, label representation becomes part of the learning problem itself as a hyperparameter.

We propose that label representation acts as an initialization for in-context learning. To investigate this hypothesis, we study how the semantic representation of labels, the demonstrations, and the underlying language model jointly influence ICL trajectories. We develop an algorithm that systematically enumerates label sets with varying degrees of semantic relevance to a task. Using these label sets, we then evaluate ICL performance across differing numbers of demonstrations, model families, and sizes, and classification tasks.

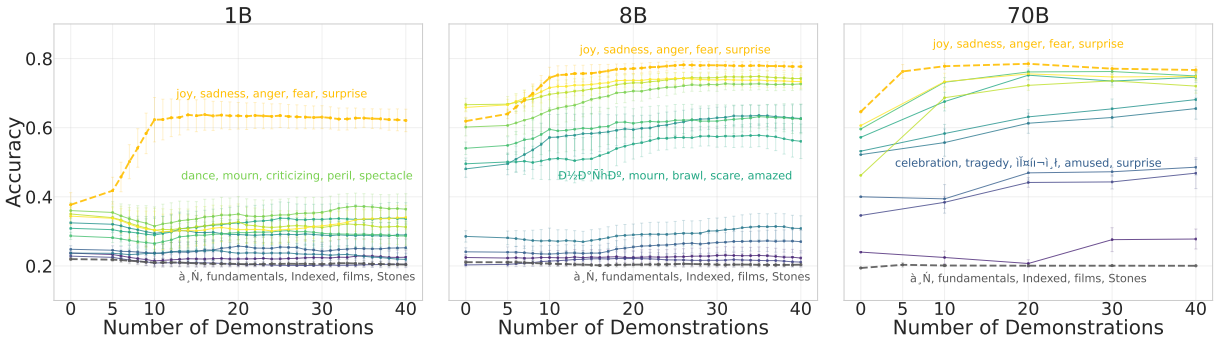
Our main result is shown in Figure 1. We find that (1) the *ranking* of representations in terms of accuracy is constant across different numbers of demonstrations, following the initial order (without any demonstrations); (2) the *range* of accuracies attainable is largely constrained by representation: the representations at both ends of the spectrum do not benefit much from demonstrations, but the representations in the middle do; (3) the *efficiency* of improvement due to demonstrations depends strongly on model scale. Our study encourages viewing of ICL as a learning algorithm whose behavior jointly depends on representation, language model, and demonstrations.

2 Related work

There have been a flurry of academic studies on ICL that have revealed its properties and characterized



(a) Sentiment Analysis 3-way classification. 1B: $K \in \{80, 90, 100\}$ examples resulted in the same set; 8B: $K \in \{60, 70\}$ and $K \in \{80, 90\}$ same set; 70B: $K \in \{40, 50\}$ and $K \in \{70, 80, 90, 100\}$ same set



(b) Sentiment Analysis 5-way classification. 70B: $K \in \{70, 80\}$ same set

Figure 1: Sentiment Analysis accuracy vs. number of demonstrations with different label representations across Llama model sizes for (a) 3-class and (b) 5-class settings. The curves are smoothed with a window size of 10, with error bars showing 95% CI over 10 runs. Full raw curves with up to 100 demonstrations are shown in Appendix A.3. The legend shows the number of labeling examples K used to fit the label set. Different K values may result in the same label sets. For these sets, the color shown is that of the higher K.

ICL as a new class of learning. In this section, we list some of these studies that have shed light on the ICL mechanism over the past few years.

Learning mechanisms. Theoretical work has explained ICL as implicit Bayesian inference by training language models from scratch on controlled synthetic data (Xie et al., 2022; Wies et al., 2023; Panwar et al., 2024; Jiang, 2023). Arora et al. (2025) have shown that Bayesian scaling laws are a good fit for the ICL curve. Another line of studies has interpreted ICL as implicitly performing gradient descent (Von Oswald et al., 2023; Ahn et al., 2023) and/or other types of learning algorithms (Akyürek et al., 2023; Garg et al., 2022; Bai et al., 2023; Li et al., 2023). All these mathematical observations encourage us to view ICL as a real learning algorithm and to perform careful empirical investigations to study its properties in real-world settings.

Content effects. The performance of LLMs in various tasks depends on how the task is presented and on how specific variables are named. McCoy et al. (2024) found that LLM accuracy is influenced by the probability of the task to be performed, the probability of the target output, and the probability of the provided input. Razeghi et al. (2022) found that few-shot reasoning is correlated with term frequencies in pretraining data. Lampinen et al. (2024) found that LLMs are not fully-abstract reasoners, that is, they do not always learn a function which they can apply to an arbitrary input. Instead, these models show content effects similar to those of humans who reason more accurately about familiar or grounded situations, compared to unfamiliar or abstract ones. These findings suggest that ICL from demonstrations might also depend on the model's prior probabilities. We expect the content of a task to impact the language model's perceived difficulty of it and thus its few-shot performance.

Prompt representation and model scale. Many studies observed that the success of ICL depends on properties of the demonstrations (Perez et al., 2021), such as the distribution of input text, the label space (Min et al., 2022), the order of examples (Lu et al., 2021; Liu et al., 2024), and the overall format of the sequence (Zhao et al., 2021). Different kinds of in-context learning were observed depending on model scale and the choice of how labels are represented. In particular, two types of labeling schemes have been studied extensively: gold (or semantically-meaningful) labeling and abstract (or semantically-void) labeling. Wei et al. (2023) and Pan et al. (2023) observed that gold labels outperform abstract labels and the accuracy gap is bigger for smaller models. They characterized true "task learning" as an emergent capability associated with scale. More recently, Kirsanov et al. (2025) noted that learning from abstract labels required more demonstrations than from gold labels.

Prompt optimization. A common approach to improving LLMs' performance without any extra weight update is via "prompt engineering," that is, by crafting prompts manually. Recent studies introduce prompt optimizers that search over strings to identify high-performing prompts (Yuksekgonul et al., 2025; Zhou et al., 2023; Yang et al., 2024b; Guo et al., 2024; Agrawal et al., 2025). The choice of label representation can be viewed as a prompt optimization problem.

3 Method

Throughout this work, we use the term "learning" to refer to improvements in task accuracy as the number of in-context demonstrations increases. We do not make claims about the specific underlying mechanism responsible for these improvements. We formulate the goal of an ICL task as solving

$$\arg\max_{y \in \mathcal{C}} p_\theta(\tau(y)|x, D_\tau), \tag{1}$$

where $D_\tau = \{(x_n, \tau(y_n))\}_{n=1}^N$ refers to a (small) number of input-output pairs. $\tau(y)$ defines a label set or how we represent each class $y \in \{1, 2, \ldots, \mathcal{C}\}$ as a token in a predefined vocabulary, i.e., $\tau : \{1, 2, \ldots, \mathcal{C}\} \to V$, where V is a vocabulary of unique tokens. D_τ refers to presenting the dataset D using τ to encode the classes. By properly formatting D, x and $\tau(y)$, LLMs have been found to be able to implicitly learn to predict the correct label associated with a new instance x.

Intuitively, we separate ICL behavior into a task-learning component driven by demonstrations and a prior over label tokens induced by pretraining. This decomposition is conceptual rather than exact, but it provides a useful framework for reasoning about how representations influence ICL trajectories. We propose that the ICL predictive probability $p_\theta(\tau(y) \mid x, D_\tau)$ depends on two components: $q_\theta(\tau(y) \mid x, D_\tau)$ and $p_\theta(\tau(y) \mid x)$).

The first term $q_\theta(\tau(y)|x, D)$ corresponds to *learning*, and the second term $p(\tau(y)|x)$ corresponds to *prior* knowledge learned by the language model during pretraining. We assume that the first component, *learning*, is generally invariant to how we represent the classes. In other words,

$$q_\theta(\tau(y)|x, D_\tau) \approx q_\theta(\tau'(y)|x, D_{\tau'}). \tag{2}$$

q_θ describes the learning ability of a language model θ and is influenced by its parameter count.

On the other hand, prior knowledge must be sensitive to the choice of τ, as it lacks the context which is presented in the form of in-context demonstrations. Unless $\tau(y)$ is *meaningful* under the pretraining corpus, the language model cannot work with an arbitrary representation of a class *a priori*. That is, it is almost certain that

$$p_\theta(\tau(y)|x) \neq p_\theta(\tau'(y)|x), \tag{3}$$

for $\tau \neq \tau'$. p_θ describes a language model's prior and is influenced by the pretraining data.

In this work, we study the interaction between label representation τ, language model θ and number of demonstrations N. While we can easily vary θ and N, we need a method to systematically enumerate τ from semantically void to meaningful representations. Below, we introduce an algorithm to do so.

3.1 Class representation optimization

We assume access to a set of K examples, which we refer to as a labeling set, and knowledge of the class that each example belongs to (how the examples are clustered). The goal is to find, for each class, a name that is represented by a single token in the vocabulary, that is meaningful under the pretraining corpus. To name $\mathcal{C}$ classes, we want to choose a set of $\mathcal{C}$ tokens from $|V|$ possible tokens in a given vocabulary, $\tau = (l_1, l_2, ...l_\mathcal{C}) \in V^\mathcal{C}$. A good representation map τ should maximize the probability assigned to the correct class $y^\star$, when

represented as $\tau(y^\star)$. We can write this directly as an objective function:

$$\max_{(l_1,l_2,\ldots,l_\mathcal{C})\in V^\mathcal{C}} \sum_{k=1}^{K} \log \frac{\exp\left(f_\theta(x_k, l_{y_k})\right)}{\sum_{c=1}^{\mathcal{C}} \exp\left(f_\theta(x_k, l_c)\right)} \quad (4)$$

where x_k are the input examples, $y_k \in \{1, 2, ...\mathcal{C}\}$ are the classes they belong to, $l_{y_k} = \tau(y_k)$ is the label assigned to class y_k, and f_θ is the language model's logit. When K is small, the optimization can exploit spurious correlations specific to a few examples. As K increases, only semantically meaningful labels generalize consistently across examples, forcing the optimization toward labels aligned with the task semantics.

We optimize this objective via hill climbing, shown in Algorithm 1: we start with an initial random token assignment for each class and iterate the following until no improvements can be made: (1) for each class, try all possible alternative tokens while keeping the rest of class names fixed, (2) evaluate the objective under the current assignment, (3) pick the best token if it improves the overall objective, (4) if there is an improvement, repeat. We run this algorithm ten times while varying random seeds and pick the assignment out of up to ten that maximizes the objective in Equation 4.

We obtain label maps τ of different quality by varying the number of labeling examples K. As K increases it becomes harder to find an assignment for which the labels have high probability for many inputs. To maximize the objective, that assignment should be generalizable: class names should be meaningful for other possible inputs. Thus, as K increases, we expect the semantics of the labels to be closer to those of gold labels.

Algorithm 1 Hill Climbing for Token Assignment Optimization

```
Input: Initial token assignment for each class
Input: Set of candidate tokens, training sentences with labels
Output: Optimized token assignments
assignments ← initial_assignments
objective ← CalculateObjective(assignments)
repeat
    improved ← false
    for each class in classes do
        candidates ← all tokens except current for class
        for each token in candidates do
            Compute objective value assigning this token to class using Eq. 4
        end for
        best_token ← token with highest objective
        if best_token improves objective then
            assignments[class] ← best_token
            Update objective
            improved ← true
            break
        end if
    end for
until not improved or max iterations reached
return assignments, objective
```

4 Experimental setup

Tasks. We experiment on four NLP tasks: Sentiment Analysis (Kirsanov et al., 2025), Question Classification (Hovy et al., 2001; Li and Roth, 2002, TREC), Textual Entailment Recognition (Williams et al., 2018, MultiNLI), Multiple Choice Question Answering (Hendrycks et al., 2021a,b, MMLU).

Models. The models we use in these experiments are: Llama 3.2 1B, Llama 3.1 8B, Llama 3.1 70B Instruct (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Qwen2.5-7B (Yang et al., 2024a).

We describe the evaluation pipeline for the Sentiment Analysis task. The other tasks follow the same general steps.

Data and prompting. Each dataset is split into a labeling set, a demonstration set, and a test set, which are disjoint. The labeling set is used to enumerate class name assignments, the demonstration set is used for the support examples for ICL, and the test set is used for the query inputs in ICL. For each task, the task is presented in a minimal format[1] with no explicit instructions, only N demonstrations and a query.

Label sets. We obtain up to ten diverse label sets for each task by optimizing Equation 4 while varying K the number of labeling examples based on task.

Let us consider the label sets obtained with the 70B model for the Sentiment Analysis 3-way classification task with gold labels {*fear, anger, joy*}. Naturally, using a small $K = 10$ leads to overfitting on labels that have a high zero-shot probability only for those labeling examples. This yielded random words as labels such as {*biomedical, malware, cloudy*}. With a small K, we cannot find label sets that appear relevant for a sentiment classification task. For a medium value of $K = 40$, the labels obtained are more general {*panic, rage, Cheers*}, a much better fit for the task. While these labels are clearly descriptive, they are slightly odd choices for class names. Finally, using a large $K = 70$

[1]We experimented with three minimal prompt formats, which performed comparably. The results are shown in Appendix F.

leads to natural category names for a sentiment classification task such as {*fear, angry, happy*}.

The label sets obtained with the same K value vary with different models. For instance, on the 5-way classification task, for $K = 100$, the 1B model found {*spectacle, dance, condolences, peril, pissed*}, the 8B model found {*surprising, joyful, sorrow, fears, anger*}, and the 70B model found {*surprise, happy, sad, anxious, ang*}.

In-context learning. We sample up to 100 examples from the demonstration set and name them according to one of the label sets previously obtained. We create demonstrations with a given label set by using that set to label the inputs in a context, and preserve the original class clustering from the dataset. These input-output pairs are concatenated, and, together with a query, are given as a prompt to a model. The model then predicts the class for a novel input selected from the test set, which has not been shown in any of the demonstrations and was not used to compute the label sets. We report the average accuracy for the test set, over 10 runs, in which the inputs of the demonstrations are resampled every time.

5 Results

5.1 Sentiment Analysis

Figure 1 shows the accuracy vs. number of demonstrations in ICL tasks with different label sets for the Llama 1B, 8B, and 70B models, for Sentiment Analysis 3-way and 5-way classification. In Appendix A.4 we show results with Mistral-7B-v0.3 (Jiang et al., 2023) and Qwen2.5-7B (Yang et al., 2024a; Team, 2024). Across all experimental conditions, we observe that the accuracy is generally increasing with the number of demonstrations. The zero-shot accuracies span a wide range from chance to ceiling. The ICL behavior across demonstrations exhibits a strong dependency on the zero-shot accuracy for its label representation. More specifically, we find that the zero-shot accuracy associated with a representation largely determines the subsequent learning trajectory. The initial representation predicts the ordering of trajectories across demonstrations, constrains the attainable gains from additional examples, and interacts with model scale to determine how quickly performance improves.

Consistent label set ranking. We first study whether the initial zero-shot performance obtained with a label representation representation predicts its behavior under additional demonstrations. We observe that for each N-shot classification task, the accuracies for ICL with different label sets remain ordered according to their initial zero-shot accuracy. We compute the ranking correlation between the zero-shot accuracies and the N-shot accuracies. We find that the correlations are indeed high across all model sizes (typically $\rho > 0.8$), although there is a lot of variance for the 1B model (see Appendix B Figure 9). These results suggest that the initialization induced by the representation is highly predictive of the entire learning trajectory.

Representation constrains attainable accuracies. The chosen label set largely determines the range of accuracies attainable under that representation. In other words, the initialization established by the representation constrains how much improvement can be achieved through additional demonstrations. While most of the learning curves are typically monotonically increasing, there is a lot of variance in how much ICL improves with more demonstrations. The increase between the minimum accuracy (zero-shot) and the maximum accuracy (40-shot) ranges from 0% up to 25%. We observe that the representations fall into three categories: small, medium, and high zero-shot accuracy.

A *low zero-shot* (<40%) accuracy indicates that the chosen representation is poorly aligned with the task. In such cases, the model may first need to infer the underlying task semantics before it can effectively use the demonstrations. Typically for representations like this, the models start with near-chance zero-shot accuracy, and the accuracy increases only very little or not at all regardless of how many demonstrations are presented (e.g. Figure 1b 8B model, $K \in \{10, 20\}$). The representations with a *medium zero-shot accuracy* (40%–60%) benefit the most from demonstrations. They can get 15%–25% improvement from the baseline by seeing demonstrations. These labels are sufficiently suggestive of the task {*medically, offending, celebrations*} that the model can eventually determine the mapping. The last group of representations consists of the *high zero-shot accuracy* (>60%) representations, those that match or are very close to gold labels. These label sets are already close to the ceiling accuracy and obtain only a small increase from the baseline before the curve plateaus.

Most of the curves appear to increase more

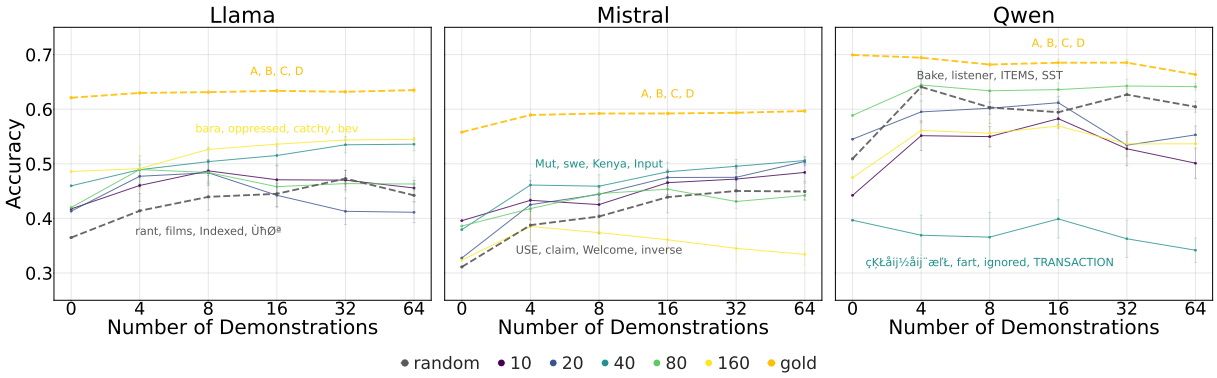


Figure 2: MMLU accuracy vs. number of demonstrations with different representations for the option-line prefixes.

slowly around 40 demonstrations, indicating a possible upper bound. We confirm this is indeed an upper bound by running ICL with up to 100 demonstrations in Appendix A.3. However, there are exceptions where the accuracy has not yet plateaued with 40 demonstrations (see Figure 1a 70B model, $K = 10$), suggesting that it is possible to overcome the limits of the representation with a large number of demonstrations and a larger model.

While representation determines the overall trajectory and attainable accuracy range, model scale influences how efficiently the model improves within those constraints.

Model size influences the learning rate. Given a fixed representation, the rate at which accuracy improves with additional demonstrations depends strongly on model scale. In Appendix B Figure 10 we show that the slope depends on model size and zero-shot accuracy. The larger 70B model is more efficient; it makes more use of fewer examples and thus exhibits steeper curves (such as Figure 1a, 70B model, $K = 30$). The N-shot accuracies for this curve are highly correlated with N (see Figure 10a orange curve, zero-shot accuracy 59%). With representations of a similar zero-shot accuracy (40%-60% range), the smaller models can also learn, but their curves increase more slowly (and thus have a lower correlation between N and N-shot accuracy), suggesting that it would take many more demonstrations to attain the same accuracy that the 70B model achieves with 20-30 demonstrations.

These observations are based on distinct label sets (of similar zero-shot accuracy) for each of the models. We confirm that this is also the case when the same label sets are used. We fix the label sets found with the Llama 1B model and run ICL with all three Llama models. We repeat this experiment with the labels found with Llama 70B. The results are shown in Appendix H and confirm the effect of model size, which has also been observed by Wei et al. (2023).

Overall, these results suggest that representation determines the initialization and broad shape of the learning trajectory, while model scale primarily affects the efficiency with which models improve from demonstrations.

5.2 Multiple-Choice Question Answering

MMLU (Massive Multitask Language Understanding) offers a different task format: the input in the ICL demonstrations consists of a question and four options and the output is one of those options. Each option is presented with a prefix (A, B, C, D) and the output is one of these tokens. The representation we vary is the choice of option-line prefixes.

Unlike sentiment classification, the representation now appears both in the answer token and in the option formatting itself. On emotion or NLI tasks, τ maps each class label to a single token that appears after a fixed prompt suffix (e.g. "Label:"). On MMLU, τ plays two roles: (1) option-line prefixes: each line is rewritten so the letter c is replaced by $\tau(c)$ (e.g. 'A: London' $\rightarrow$ 'Compilation: London') and (2) answer token: demonstrations and the query end with 'Answer: ' followed by the token for the correct option. Since the input is no longer fixed and also depends on the choice representation, we cannot apply Algorithm 1 directly. We instead modify it to an EM style algorithm, where the E-step precomputes logits under a fixed τ, and the M-step maximizes the hill climbing objective. Details and pseudocode for this algorithm are shown in Appendix I, Algorithm 2. The label sets are shown in Table 11.

The learning curves with up to 64 demonstrations for Llama 8B, Mistral 7B and Qwen 7B are

shown in Figure 2. We find that, generally, the trends from Sentiment Analysis hold. The representations that the models can work best with (highest zero-shot accuracy) are the usual ones (A, B, C, D), and the improvement from the baseline with more demonstrations is small for this category of labels. All models can learn to use medium zero-shot accuracy representations and answer accordingly. However, there are some representations that the models cannot work with (e.g., Qwen $K = 40$ curve). We observe that, for Qwen, ICL is not complete: the accuracy goes down after some number of demonstrations, regardless of the choice of representation. The highest accuracy attainable with the representations we test is achieved by gold labels, zero-shot. We leave it for future work to investigate why this is the case.

5.3 Question Classification and NLI

The Question Classification (TREC) and Natural Language Inference (MultiNLI) tasks follow a very similar setup to the Sentiment Analysis task. TREC is a 5-way classification task, where the inputs are questions and the gold labels are "Entity, Description, Human, Location, Numeric." The optimized label representations and learning curves are shown in Appendix C. MultiNLI is a natural language inference task where the inputs consist of premise-hypothesis pairs and the gold labels are "entailment, neutral, contradiction". The optimized labels and results are shown in Appendix E. In both of these tasks, the learning curves exhibit the previously observed behavior of ICL, confirming the generality of the findings.

6 Discussion

6.1 Variance under semantic equivalence

While human-defined gold labels are often chosen based on their semantic fitness, there are typically more label choices with the same meaning. This experiment allows us to formally quantify the variance from choosing different vocabulary tokens to represent the same concept. By benchmarking the gold labels against multiple sets of synonyms, we isolate token-level variance from semantic meaning. We find that even though all synonym sets share similar meanings, changing the specific tokens introduces a performance variance of up to 25% in the zero-shot setting (3-class) and a persistent 15% variance even after 100 demonstrations (5-class), as seen in Figure 3. The best label set

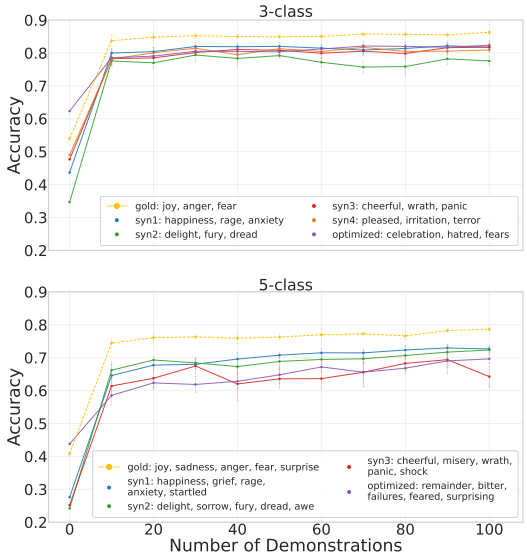


Figure 3: Learning curves using synonyms of gold labels.

discovered by our algorithm performs within this semantic variance range, suggesting that it is indeed comparable to the gold label.

6.2 Imbalanced datasets

All previous experiments used balanced classes in all labeling, demonstrations, and test sets. To test the sensitivity of our findings to class imbalance, we ran an experiment with Mistral where both label set enumeration procedure and the ICL demonstrations use the imbalanced ratios of classes, while the test set remains balanced. We split the classes for sentiment analysis into Joy 40%, Sadness and Anger 20%, Fear and Surprise 10%. Using the predictions from Qwen, we report the F1 scores for classes Joy and Surprise using label sets τ found by Algorithm 1 in Figure 4 [2]. Our main finding holds: the choice of label influences the learning trajectory despite the two classes being of different proportions.

6.3 Effect of prior

Prior work on shuffled-label settings has examined whether ICL reflects true task learning or reliance on pretrained priors (Min et al., 2022; Pan et al., 2023; Wei et al., 2023). We revisit these results through representation quality, distinguishing (1) shuffled demonstrations with original targets, and (2) shuffled demonstrations with shuffled targets. The results are shown in Appendix G.

[2] $\tau(\text{Joy})$ = celebration for several K values. The learning curves vary for different K since $\tau(\text{other classes})$ differ.

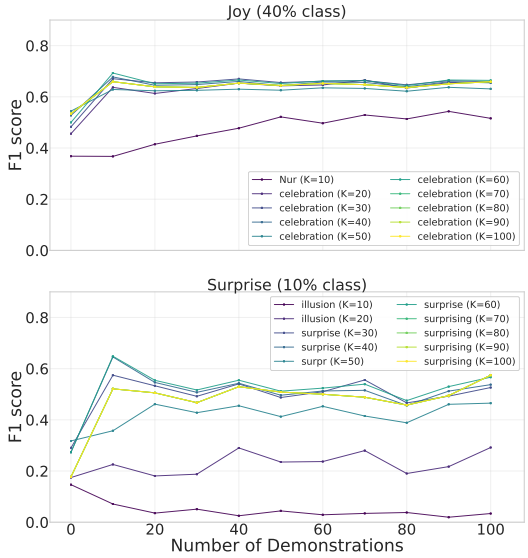


Figure 4: F1 scores for tokens optimized using different K for the examples from classes Joy 40% and Surprise 10%.

Shuffled demonstrations only. When labels are shuffled only in demonstrations while targets remain unchanged, zero-shot accuracy remains identical to the original setting and the few-shot accuracy does not degrade substantially, despite the corrupted demonstrations. This helps explain why shuffled demonstrations can still yield strong performance, which has been interpreted as evidence against genuine learning in ICL. Our results suggest instead that high zero-shot accuracy causes the small accuracy decrease.

Shuffled demonstrations and targets. When both demonstrations and targets are shuffled, label assignments become inconsistent with class structure, leading to low zero-shot accuracy across representations. In this setting, accuracies (Figure 16c) are lower than in the shuffled-demos, original-targets case (Figure 16b) when few demonstrations are used. Pan et al. (2023) similarly observed that abstract label sets underperform random allocations for small models, attributing this to reliance on priors rather than learning. We instead find that abstract and shuffled label sets begin with lower zero-shot accuracy, so improvements from demonstrations are insufficient to overcome this disadvantage. In contrast, gold-aligned labels provide a stronger starting point even when randomly permuted.

Overall, representation quality strongly shapes interpretations of shuffled-label experiments. The observed behavior does not imply absence of learning, but reflects constraints imposed by the initial representation on the learning trajectory.

7 Conclusion

Previous work has questioned whether ICL reflects true learning or reliance on pretrained priors (Min et al., 2022). We argued that representation choice shapes ICL behavior and developed a method to construct label representations varying in semantic relevance. We find that representation largely determines baseline performance, and that relative ordering of label sets is consistent across demonstrations. While more demonstrations and larger models can partially overcome weak representations, label quality remains an important factor throughout learning. Learning efficiency depends on both representation and model scale: medium-quality representations often benefit most from additional demonstrations, and larger models learn faster.

Our findings on the interaction between learning and representation in LLMs closely reflect what we know of more conventional neural network learning. The search for high-performing prompts for LLMs is in spirit similar to hyperparameter search (Bengio and LeCun, 2007; Liu et al., 2019) for neural network classifiers that learn via backpropagation. Perez et al. (2021) found that good prompts are effective because they are chosen using large validation sets. Prompts influence the model behavior similarly to how a choice of initialization influences neural network training. In particular, the choice of label representation in ICL is analogous to the hyperparameter selection for a neural network classifier. The different choices of representation determine the learning trajectory in both cases: for LLMs, a high quality representation leads to a high zero-shot accuracy and faster convergence; for neural network classifiers, a good initialization can lead to efficient learning (LeCun et al., 2012).

We conclude that ICL should be understood not as a purely emergent reasoning capability nor as simple retrieval from pretrained priors, but as a learning system whose behavior arises from the interaction between model, representation, and demonstrations. More broadly, our work identifies label representation as a distinct inference-time hyperparameter unique to language-based learning systems, and highlights the importance of studying representation and learning jointly when analyzing the capabilities of large language models.

Limitations

Our framework has a few limitations. First, in the tasks we considered, the labels are limited to one token, which might not be the case in more complex tasks. We chose to only explore this option since the search space for label sets with our algorithm grows exponentially with the number of tokens. While our proposed decomposition does not change when $\tau : \{1, 2..\mathcal{C}\} \rightarrow V^M$, the optimization problem becomes a search in V^M, where M is the maximum number of tokens for the label representation. We expect our findings to hold, but formally quantifying label representations consisting of multiple tokens or free form, including tasks beyond classification, is left for future work.

Second, we focus only on how labels are represented. Variations in input representations may also influence learning. For instance, choosing inputs with high priors under the model could improve efficiency (Ceballos-Arroyo et al., 2024). Although we treated inputs as fixed, real-world datasets often allow for choosing among many possible inputs. Moreover, a more complex prompt format could also influence performance. Thus, future work should examine how the inputs and prompt structure affect learning.

Third, the in-context learning setting we studied is restricted to one round of prompting. It would be interesting to study how multi-turn interactions influence ICL in cases where the demonstrations are presented in an online manner, possibly together with extra information in each round (Lee et al., 2023).

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *Preprint*, arXiv:2507.19457.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2023. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 45614–45650. Curran Associates, Inc.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah Goodman. 2025. Bayesian scaling laws for in-context learning. In *Second Conference on Language Modeling*.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pages 57125–57211. Curran Associates, Inc.

Yoshua Bengio and Yann LeCun. 2007. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *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)*, pages 12119–12149, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *Preprint*, arXiv:2304.09960.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. 2025. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1855–1888, Albuquerque, New Mexico. Association for Computational Linguistics.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 2012. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain QA. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544, Singapore. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. 2024. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *Preprint*, arXiv:2402.10738.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Annual Meeting of the Association for Computational Linguistics*.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. 2024. In-context learning through the bayesian prism. In *The Twelfth International Conference on Learning Representations*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *Preprint*, arXiv:2303.03846.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 36637–36651. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024b. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. 2025. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639:609–616.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Sentiment Analysis Details and Additional Experiments

A.1 Dataset

We use a synthetic sentiment classification dataset from Kirsanov et al. (2025), which contains 1,000 sentences split equally among 5 classes for 5-way classification. We also use a subset of 600 sentences covering only 3 of the classes for 3-way classification. We split the dataset into a labeling set (25%), a demonstration set (25%), and a test set (50%).

Prompt Example

Text: The underdog candidate's supporters are over the moon as he clinches the nomination against all odds.
Category: joy

Text: Oh, I'm sure the new bill they've proposed will solve all our problems instantly.
Category: anger

Text: The corridors of power will become a labyrinth of whispers, where truth is a forgotten echo.
Category: fear

Text: "GENIUS INVENTS REVOLUTIONARY APP THAT ADDS ZERO VALUE TO USERS' LIVES"
Category:

Table 1: Prompt used for sentiment classification.

A.2 List of label sets

We show the label sets obtained from Algorithm 1.

A.3 Full learning curves

We show the learning curves with up to 100 demonstrations for the Sentiment Analysis task: 3-way classification in Figure 5 and 5-way classification in Figure 6.

K	1B	8B	70B
10	Nutrition Giz Legends	Gluten Laptop clouds	biomedical malware cloudy
20	diabetes Hacker Presbyterian	Diabetes Revenge spirit	fitness computer joyful
30	overweight annoy scholarships	FDA console celebration	Obesity rage celebration
40	medically offending celebrating	fearful malicious celebration	panic rage Cheers
50	medically offending celebrations	digestive insulting accomplishments	panic rage Cheers
60	panicked offending celebrations	fears insults joyful	worry complain celebration
70	hazardous offending celebrations	fears insults joyful	fear angry happy
80	à¤kà¤¤à¤° (*danger*) offending celebrations	fears complaints joyful	fear angry happy
90	à¤kà¤¤à¤° (*danger*) offending celebrations	fears complaints joyful	fear angry happy
100	à¤kà¤¤à¤° (*danger*) offending celebrations	fears complaint joyful	fear angry happy

Table 2: Label sets obtained from running Algorithm 1 on K labeling examples for 3-way classification. The gold labels are "fear, anger, joy."

K	1B	8B	70B
10	movie Musik Causes Roller NRL	theater COLOR HEALTH ride Offensive	Marvel MUSIC HEALTH roller veh
20	witches audition bere adip Messi	cinema Broadcasting Deng nut Rugby	Magical positive loss Dietary Baseball
30	trick Dresses bere hysteria Messi	surprising MÃ©d (*Med*) ÑġD¾D½ (*dream*) snack soccer	surprise celebration tragedy amused ìĬ¤í1¬ì¸ł (*sports*)
40	puzzle Ventures à¤h (*A*) carniv Penalty	surprising ÑĤDµD» (*tel*) resignation aliment soccer	surprising positives heartbreaking D¿D¸Ñī (*shout*) frustrated
50	spectacle talent mourn endanger offense	surprised baÅŁarÄ± (*success*) sadness xen Rage	surprising positives heartbreaking Brussels frustrated
60	spectacle production mourn peril offense	amazed D½D°ÑĥDº (*science*) mourn scare brawl	surprising positives heartbreaking nerv agg
70	spectacle productions mourn peril racket	surprising íh (*?*) sorrow terror hostile	surprise pleasant sorrow fears rage
80	spectacle dance mourn peril criticizing	surprising celebrates condolences terror rage	surprise pleasant sorrow fears rage
90	magician dancer mourning risking wrath	surprising joyful sorrow fears rage	surprise Lift broken fears rage
100	spectacle dance condolences peril pissed	surprising joyful sorrow fears anger	surprise happy sad anxious ang

Table 3: Label sets obtained from from running Algorithm 1 on K labeling examples for 5-way classification. The gold labels are "surprise, joy, sadness, fear, anger."

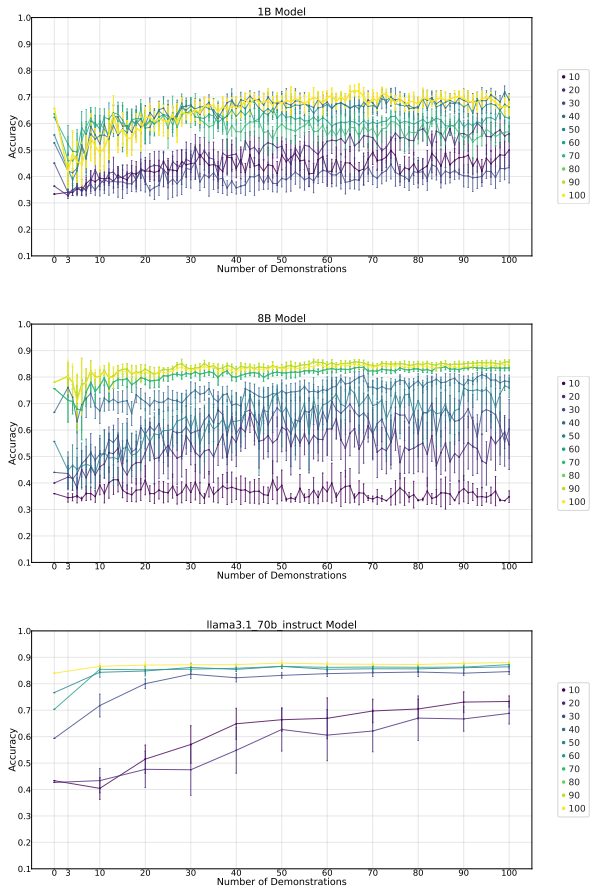


Figure 5: Full raw (unsmoothed) learning curves for up to 100 demonstrations for Llama models for 3-way classification for the sentiment analysis task.

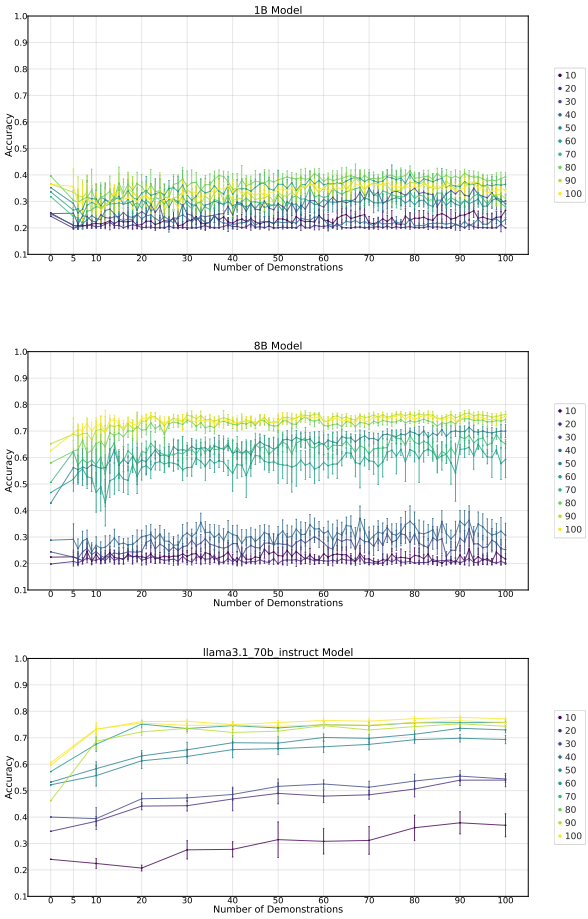


Figure 6: Full raw (unsmoothed) learning curves for up to 100 demonstrations for Llama models for 5-way classification for the sentiment analysis task.

A.4 Mistral-7B-v0.3 and Qwen2.5-7B

We show the label sets and learning curves for the sentiment analysis task with Mistral-7B-v0.3 and Qwen2.5-7B for 3-way and 5-way classification.

K	Mistral	Qwen
10	diet, hack wholes	Sick, offensive Ath
20	nutrition, hack excitement	JsonResult, parliamentary applause
30	protein, complaint excited	weighing, warn congrat
40	fear, angry insp	ArgumentError, bitch Wonderful
50	fear, angry joy	fears, hatred celebration
60	fear, angry joy	noir, bitch Wonderful
70	panic, rage smiles	fears, abusive congrat
80	fear, anger joy	nightmares, offender luz
90	fear, angry joy	terror, offending brag
100	fear, angry joy	terror, offending brag

Table 4: Label sets obtained from running Algorithm 1 on K examples for 3-way classification

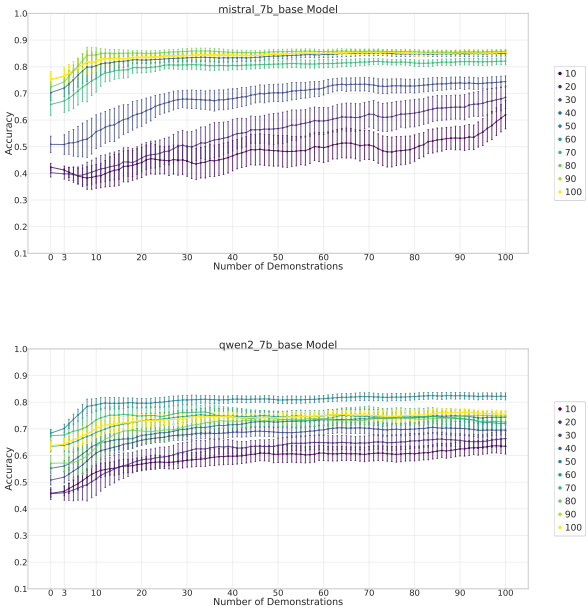


Figure 7: Mistral and Qwen 3-way classification for the sentiment analysis task

K	Mistral	Qwen
10	photography, festival AU, roll Bull	Dialogue, SUM clinical, Roller negatives
20	aston, rehe diseases, stomach managers	Ey, academy diagnosed, kaufen threats
30	amazing, dress defeat, monster hockey	surprising, moda diagnoses, fart unlawful
40	surprise, competition loss, monster soccer	surprising, RoundedRectangle privacy, immature mud
50	surprise, dress grief, monster angry	surprising, RoundedRectangle failures, Moo FUCK
60	amazing, invent depress, fright piss	exploding, ExecutionContext Ø§ÙHØ·, onOptionsItemSelected misogyn
70	amazing, next depress, terror angry	astonishing, Validates distressed, nightmares mud
80	aston, inspire despair, fears rage	astonishing, remains distressed, feared abuses
90	aston, pose unhappy, afraid complaint	surprising, remainder failures, feared bitter
100	aston, inspire despair, fears rage	surprising, remainder failures, feared bitter

Table 5: Label sets obtained from running Algorithm 1 on K examples for 5-way classification

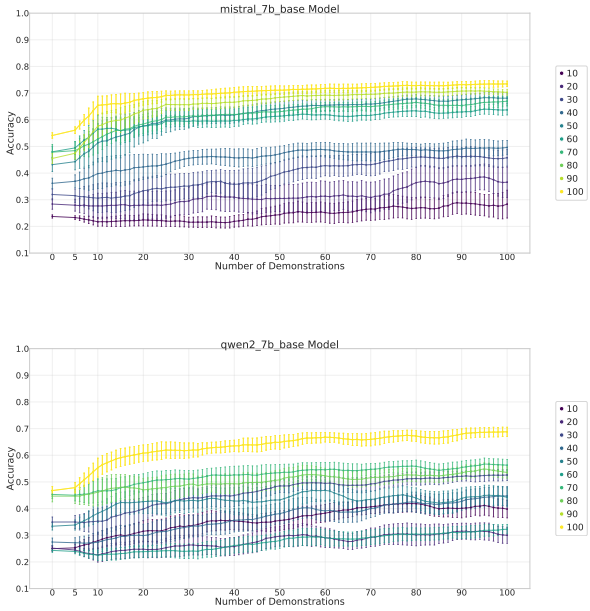
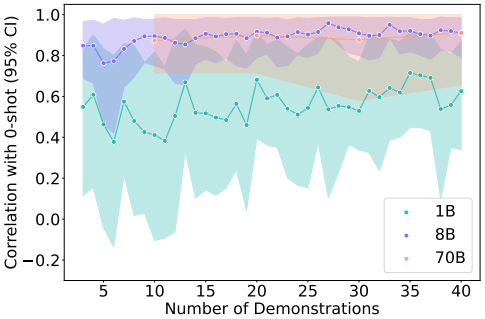
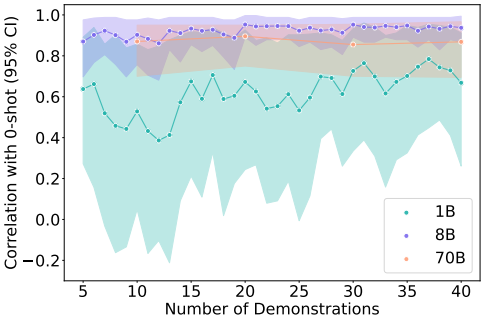


Figure 8: Mistral and Qwen 5-way classification for the sentiment analysis task

B Ranking correlations

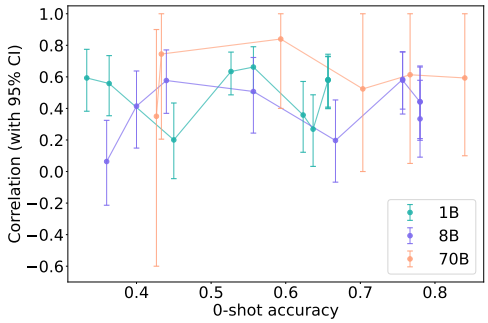


(a) 3-way classification

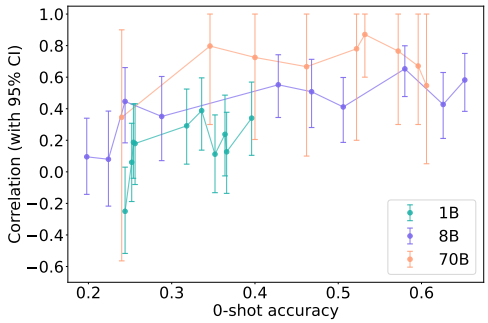


(b) 5-way classification

Figure 9: Ranking correlation coefficient between the zero-shot accuracy and the N-shot accuracy vs. N number of demonstrations. $N \in \{\text{num classes}, ...40\}$ for 1B and 8B models, $N \in \{10, 20, 30, 40\}$ for 70B model. The CI are computed over 1000 bootstrapping samples from 10 runs per N-shot accuracy. **The order of label sets in terms of quality stays consistent across N-shot experiments.**



(a) 3-way classification



(b) 5-way classification

Figure 10: **Larger models exhibit steeper curves.** Evaluation of learning curves for label sets obtained with different K labeling examples. Ranking correlation coefficient between N and N-shot accuracy vs. zero-shot accuracy for each curve. $N \in \{\text{num classes}, ...40\}$ for 1B and 8B models, $N \in \{10, 20, 30, 40\}$ for 70B model. Higher correlation indicates that the accuracy for that curve is often strictly increasing with N (steeper curve), while lower accuracy indicates that the accuracy can be plateauing or decreasing on some intervals (flatter curve). The CI are computed over 1000 bootstrapping samples from 10 runs per N-shot accuracy.

C Question Classification: TREC

We conducted additional experiments using the TREC dataset (Li and Roth, 2002; Hovy et al., 2001) for question classification. The gold labels are "Entity, Description, Human, Location, Numeric." The input sentences in this dataset are questions. This makes applying our framework to this task challenging since high probability next tokens following a question would be answers rather than class names. Despite this fact, our findings from sentiment analysis hold true on the TREC dataset. An example prompt is shown in Table 6.

Prompt Example

Text: Which of the following TV newsmen was a Rhodes scholar?
Category: human

Text: How many cullions does a male have?
Category: numeric

Text: What strait separates North America from Asia?
Category:

Table 6: Prompt used for TREC question classification.

C.1 Llama 3.1 8B

K	Llama 8B
10	gods, Derm Philippe, Helsinki, Wade
20	Judaism, Skin Christians, Ukraine, Watts
30	easiest, Carb Isaiah, Antarctica, dollars
40	GENERIC, SCC quienes, Antarctica, dollars
50	which, Explain who, Nations, Rate
60	Taste, explain Malcolm, maps, Timing
70	Taste, explain qui, maps, Timing
80	taste, development personalities, whereabouts DºD¾D»D¸ÑïDµÑġÑĤD²D¾
90	taste, development quienes, destinations, timed
100	Taste, explanations qui, locations, amount

Table 7: Label sets obtained from running Algorithm 1 on K examples for 5-way classification

C.2 Mistral 7B

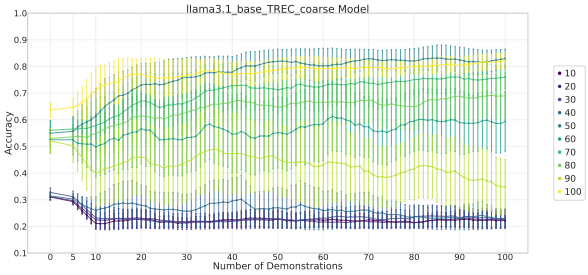


Figure 11: Llama 8B TREC

K	Mistral 7B
10	universe, Fer artists, Tokyo, gover
20	Jung, advice Leonard, Finland, college
30	matching, explaining kings, continent, accounting
40	filling, explaining athletes, UEFA, amount
50	night, explan whom, UEFA, amount
60	conj, explan whom, locations, amount
70	eating, explan whom, locations, amount
80	drinking, explan whom, locations, amount
90	eating, explan whom, locations, amount
100	cul, explan whom, locations, amount

Table 8: Label sets obtained from running Algorithm 1 on K examples for 5-way classification

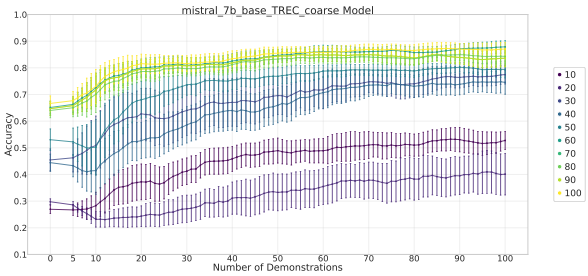


Figure 12: Mistral 7B TREC

C.3 Qwen 7B

K	Qwen 7B
10	—-, Bronx Kaz, PhoneNumber, PRES
20	comparator, Chung Boris, wherever, æķ°
30	ConfigurationManager, Geh quoted, wherever, amounts
40	meinen, NDEBUG Santa, wherever, amounts
50	me, Technologies Vi, geographical, Num
60	animal, Anatomy Vi, geographical, Num
70	animal, Lecture Vi, geographical, Num
80	mt, Lecture Vi, WHERE, Num
90	serif, ×¢ entreprise, gÃ©, numb
100	serif, ×¢ entreprise, gÃ©, numb

Table 9: Label sets obtained from running Algorithm 1 on K examples for 5-way classification

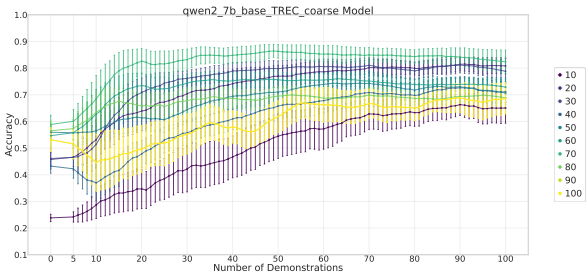


Figure 13: Qwen 7B TREC

D Multiple Choice: MMLU

MMLU dataset consists of a train split of 1,792 questions and a test split of 12,888 questions. We use the half of the train split for the labeling set and half for demonstrations. We run Algorithm 2 with up to $K = 160$ labeling examples to obtain label sets, see Table 11. We then run ICL by sampling up to 64 demonstrations per run and evaluate on the entire test set. An example prompt is shown in Table 10.

Prompt Example

Question: The seat of higher cortical functioning is typically associated with the _____.
A: Temporal lobe
B: Frontal lobe
C: Occipital lobe
D: Hippocampus
Answer: B

Question: A person who received a Level 4 rating on the Rancho Los Amigos Scale of Cognitive Functioning...
A: is nonresponsive to all stimuli
B: is confused and incoherent, may exhibit bizarre behavior
C: is alert and oriented but unable to integrate tasks...
D: is functioning normally but has some impairment in judgment
Answer: B

Question: The terms "crisis of the 1970s" is used to refer to:
A: declining profits and rising unemployment
B: the eradication of the welfare state
C: rising divorce rates and the decline of the traditional family
D: an unfortunate twist in fashion sensibility
Answer:

Table 10: Prompt used for MMLU multiple-choice classification.

K	Llama	Mistral	Qwen
10	Compilation Consumption VAR redistribute	fus Hamburg VM diagn	Coupe incomes exh verify
20	contr Kostenlos Millennium sure	sear Manchester FIFA Springs	isa Franco crumbling Nets
40	McGr sculpture IDS EXTRA	Mut swe Kenya Input	çĶŁåij½åij¨æľŁ fart ignored TRANSACTION
80	Khu aVar slime LOGGER	naz coins G shortly	AE BE CG Devon
160	bara oppressed catchy bev	Mol manus Oscar gewann	AFL Bridge teaser CircularProgress

Table 11: Optimized MMLU relabelings τ (classes A–D) by model and relabeling set size K.

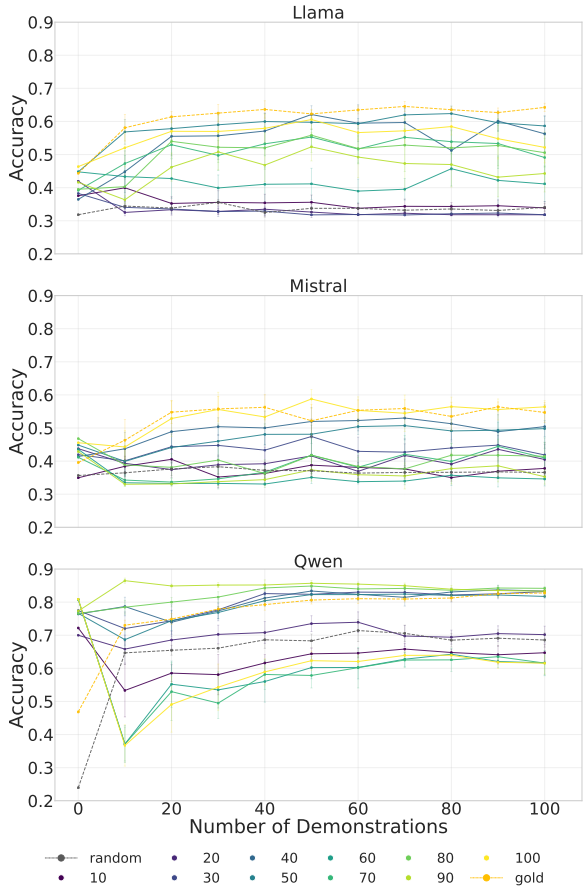


Figure 14: MultiNLI: Accuracy vs. number of demonstrations

E Textual Entailment Recognition: MultiNLI

MultiNLI dataset consists of a train split of 392,702 premise-hypothesis pairs and a test split of 9,815 examples (matched partition). We use 100 examples from train for relabeling and run Algorithm 1 with up to $K = 100$ labeling examples to obtain label sets, see Table 13. We then run ICL by sampling up to 100 demonstrations per run and evaluate on the entire test set. The results for 7/8B models are show in Figure 14. An example prompt is shown in 12.

Prompt Example

Input:
Premise: I'll hurry over that part.
Hypothesis: "I'll be quick with that part."
Label: entailment

Input:
Premise: Overall, though, the agency representatives questioned the need for a standardized approach...
Hypothesis: The agency representatives are extremely clever
Label: neutral

Input:
Premise: The new rights are nice enough
Hypothesis: Everyone really likes the newest benefits
Label:

Table 12: Prompt used for MultiNLI natural language inference classification.

K	Llama	Mistral	Qwen
10	Visitors hometown unreasonable	HP England reaction	Leslie uber backwards
20	yÃ¶ Oakland doomed	body assumptions reject	FactoryBot gaz backwards
30	AydÄ±n ä¸Ĭæµ· ÙĨØ¬	phrases arts ridic	decom Neb !==
40	ÑĥÑĤ fidelity declined	cir wisdom nit	decom unc contradictions
50	identifies wisdom loser	verb AG error	decom Neb Contrast
60	Unter wisdom disqualified	SY AG ERROR	reordered viewWillAppear thai
70	Hind Beauty ERROR	Chain Wonder ERROR	reordered viewWillAppear thai
80	ClassName fond declined	prefix Friends error	Compound Unc contradiction
90	Sext inspiration loser	Chain Meeting ERROR	ALLOW Unc contradiction
100	Encode inspiration loser	singular happiness impossible	reordered viewWillAppear thai

Table 13: Optimized MultiNLI relabelings τ (entailment, neutral, contradiction) by model and K.

F Effect of prompt format

In addition to our original prompt format which (Text/Category), we tried two other prompts: Sentence/Label and Arrow (Input → Output) with Qwen 7B on the Sentiment Analysis Task. The results are comparable, as illustrated in Figure 15. showing rank preservation and confirming robustness to minor prompt changes.

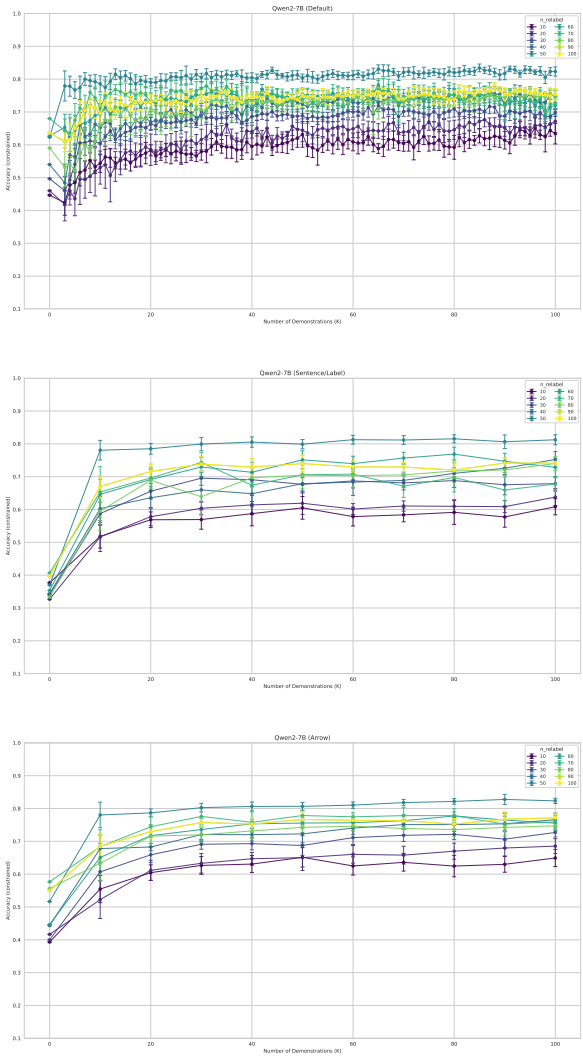


Figure 15: Learning curves for up to 100 demonstrations with Qwen 7B for 3-way classification for the sentiment analysis with three different prompt formats.

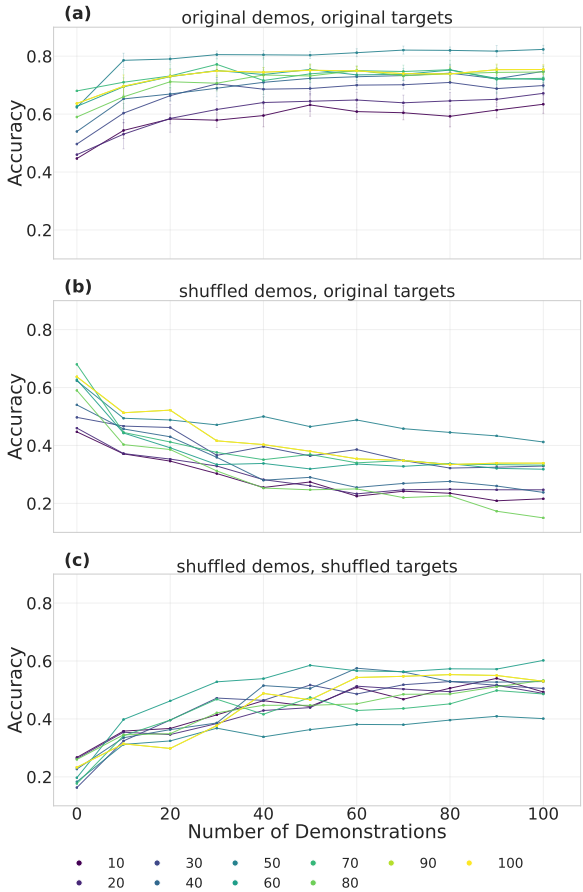


Figure 16: Effect of semantic priors with shuffled labels.

G Corrupted labels

Details for Section 6.3.

H Transfer of labels across models

In order to strictly isolate the effect of model size from the choice of representation, we evaluated different models on the exact same label sets. We fixed the optimized label sets discovered by the Llama 1B model and the Llama 70B model, and then evaluated both of these sets across all three model sizes (1B, 8B, and 70B) on the 3-way Sentiment Analysis tasks. We show the results with the labels found by Llama 70B in Figure 17 and Figure 18 shows the results with the labels from Llama 1B. By holding the representation constant, we can now cleanly observe the isolated effect of model scale. We found that (1) when prompted with the exact same label set, the larger 70B model exhibits a steeper learning curve than the 1B and 8B models and (2) the effect holds regardless of which model generated the labels.

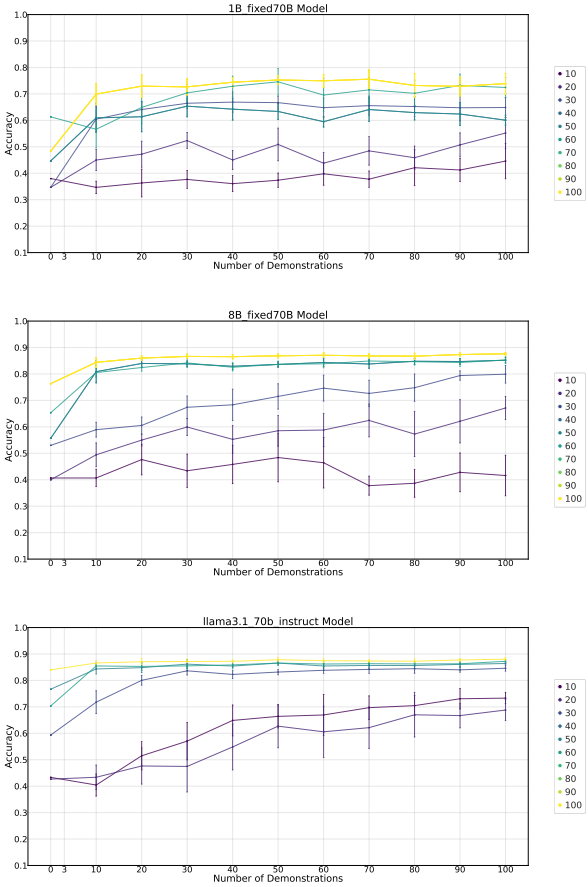


Figure 17: Learning curves for up to 100 demonstrations for Llama models for 3-way classification for the sentiment analysis task using label sets found with Llama 70B.

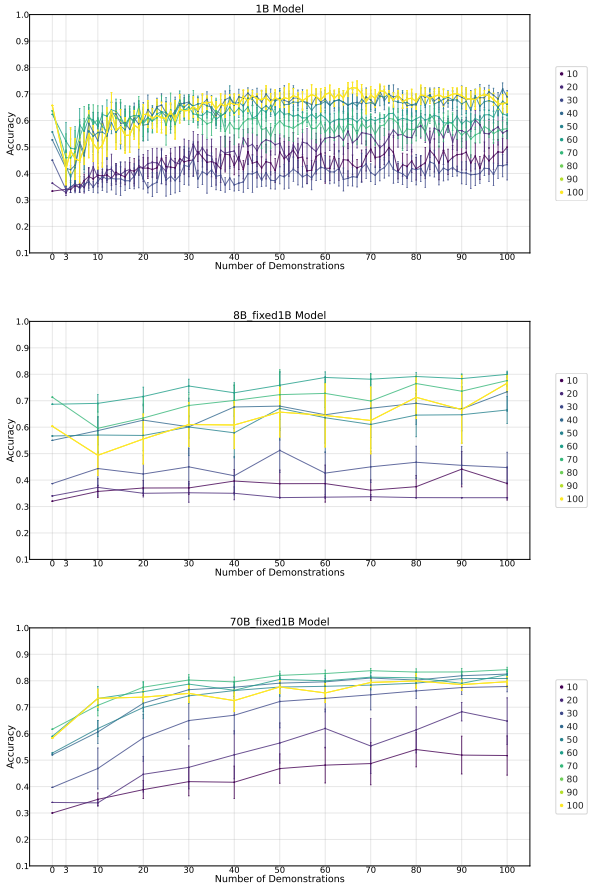


Figure 18: Learning curves for up to 100 demonstrations for Llama models for 3-way classification for the sentiment analysis task using label sets found with Llama 1B.

I Implementation details

I.1 Algorithm 1

We first precompute the next token logits for all inputs with only one forward pass through the relabeling sentences in the dataset. During optimization, for each τ we perform a simple lookup for the logits of the K sentences corresponding to the tokens in τ and normalize. The algorithm finds "good enough" solutions in the sense that it allows us to enumerate labels sets covering a wide spectrum zero-shot accuracies. Moreover, the label sets found with large K are close (or identical) in meaning with the gold labels. These label sets already have a zero-shot accuracy close (within 3-5%) of the ceiling accuracy obtained after seeing demonstrations, further confirming their quality. While our algorithm is not guaranteed to find the global optimum, it is extremely simple and efficient, making it suitable for studying ICL from demonstrations or for applications which require high-quality labels.

I.2 Algorithm 1 EM Variant

Algorithm 2 EM Optimization of MMLU Option-Line Relabeling τ

Input: Training questions $\{x_i\}$ with gold labels $\{y_i\}$, $y_i \in \{A, B, C, D\}$
Input: Candidate token vocabulary $\mathcal{V}$ (whole-word tokens only)
Input: Language model f, tokenizer, max EM iterations T_{EM}, hill-climb restarts R
Output: Optimized relabeling $\tau^\star\colon \{A, B, C, D\} \to \mathcal{V}$
$\tau \leftarrow \tau_{\text{gold}}$ ▷ identity mapping; not random
$prev_\tau \leftarrow \emptyset$
for $t = 1$ **to** T_{EM} **do**
  for each training example (x_i, y_i) **do**
    $\tilde{x}_i \leftarrow \textsc{RewriteOptions}(x_i, \tau)$
    $z_i \leftarrow f(\tilde{x}_i \,\|\, \texttt{\textbackslash nAnswer:})$ ▷ last-position logit vector
  end for
  $\tau_{\text{new}}, \mathcal{O} \leftarrow \textsc{HillClimbTokens}(\{z_i\}, \{y_i\}, \mathcal{V}, R)$ ▷ Alg. 1; random restart inits; objective Eq. 4
  if $\tau_{\text{new}} = \tau$ **or** $\tau_{\text{new}} = prev_\tau$ **then**
    $\tau \leftarrow \tau_{\text{new}}$; **break**
  end if
  $prev_\tau \leftarrow \tau$; $\tau \leftarrow \tau_{\text{new}}$
end for
return $\tau, \mathcal{O}$

I.3 In-context learning

For ICL, we sample 10 different sets of N demonstrations and query the entire test set. We use prefix KV caching for demonstrations, so that during testing we only need to compute the attention between the query sentence and the demonstrations.

Runtime. We report approximate wall clock times for any fixed label set with N demonstrations for 300-500 test sentences, for 10 runs on an 80GB A100 or H100 GPU: 5-7 minutes for a 7B or 8B model and 20-24h for the 70B model.