Exploring Explanations Improves the Robustness of In-Context Learning

Ukyo Honda
CyberAgent, Tokyo, Japan
honda_ukyo@cyberagent.co.jp

Tatsushi Oka
Keio University, Tokyo, Japan
tatsushi.oka@keio.jp

Abstract

In-context learning (ICL) has emerged as a successful paradigm for leveraging large language models (LLMs). However, it often struggles to generalize beyond the distribution of the provided demonstrations. A recent advancement in enhancing robustness is ICL with *explanations* (X-ICL), which improves prediction reliability by guiding LLMs to understand and articulate the reasoning behind correct labels. Building on this approach, we introduce an advanced framework that extends X-ICL by systematically *exploring explanations* for all possible labels (X^2-ICL), thereby enabling more comprehensive and robust decision-making. Experimental results on multiple natural language understanding datasets validate the effectiveness of X^2-ICL, demonstrating significantly improved robustness to out-of-distribution data compared to the existing ICL approaches.[1]

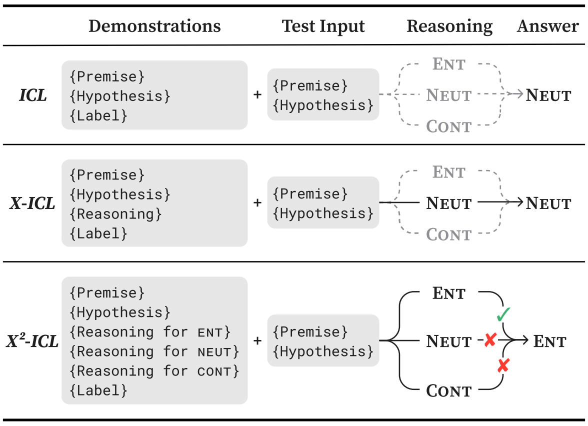

Figure 1: Overview of ICL, X-ICL, and our proposed X^2-ICL. In this example, the task is natural language inference, where ENT, NEUT, and CONT represent entailment, neutral, and contradiction labels, respectively. Solid lines denote outputs generated by LLMs during inference, while dotted gray lines represent paths not explicitly produced by LLMs. ICL produces only the final label without providing reasoning, whereas X-ICL generates a single reasoning path before selecting a label. In contrast, X^2-ICL first generates reasoning for all possible labels and then chooses the label supported by the most valid reasoning.

1 Introduction

In-context learning (**ICL**; Brown et al., 2020; Radford et al., 2019) has demonstrated remarkable success with the advancement of large language models (LLMs). In ICL, LLMs make inferences given a limited number of *demonstrations*, i.e., labeled examples provided as context, without any parameter updates. This efficiency is a key advantage over the predominant paradigm of fine-tuning, which requires a large amount of task-specific training data and parameter updates. Not only efficient, but ICL has also been shown to be effective across a variety of tasks, making it a new paradigm for leveraging LLMs (Dong et al., 2024).

However, ICL has been reported to be restricted in its ability to generalize beyond the given demonstrations. ICL often exhibits degraded performance when evaluated on *out-of-distribution* (OOD) data (Tang et al., 2023; Wang et al., 2023b,a; Mueller et al., 2024; Zhou et al., 2024; Siska et al., 2024; Yuan et al., 2024). Following the previous studies, we specifically refer to OOD data as data drawn from distributions that have been *adversarially shifted* from the distribution of the given demonstrations. Recent studies have addressed the robustness challenge in ICL (Li et al., 2023; Sun et al., 2024; Jang et al., 2024; Zhou et al., 2024; He et al., 2024).

A promising advancement in addressing this challenge is ICL with *explanations* (**X-ICL**; He et al., 2024). In X-ICL, LLMs are provided with labeled examples along with explanations that justify the assigned labels. The objective is to enable the model to learn human-like *reasoning*[2] rather than

[1]The code is publicly available at: https://github.com/CyberAgentAILab/x2-icl.

[2]The explanation that justifies an assigned label represents the reasoning process from the perspective of a model that

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23693–23714

relying on superficial patterns within the provided demonstrations. This approach is conceptually similar to few-shot chain-of-thought (CoT) reasoning, which provides examples of human-written step-by-step reasoning as context (Wei et al., 2022).[3] However, a technical distinction of X-ICL is that the provided in-context explanations are generated by LLMs, rather than being manually crafted. This machine-generated nature makes X-ICL scalable, and it has demonstrated strong performance across various OOD evaluation tasks in natural language understanding (NLU).

In this study, we extend X-ICL to enhance robustness by further leveraging its scalability. In OOD data, the reasoning learned from demonstrations is not always reliable. As the data distribution is adversarially shifted from that of the demonstrations, models must carefully analyze inputs from multiple aspects to ensure accurate predictions. Our idea is that *exploring explanations* behind all available labels in demonstrations (**X^2-ICL**) will encourage the model to more thoroughly consider the alternative reasoning paths. To achieve this, we augment the demonstrations by providing diverse explanations, each generated by conditioning LLMs on a different possible label. Figure 1 provides an overview of our method.

Our extensive experiments demonstrate the enhanced OOD robustness of X^2-ICL across various datasets in NLU. Specifically, X^2-ICL outperforms both ICL and X-ICL on six to eight out of eight OOD datasets for each of five different LLMs. We expect that the effectiveness of our approach will offer important insights into the robustness of ICL, emphasizing the value of systematically and comprehensively exploring various reasoning paths.

2 Preliminaries

To formally introduce existing methods and our proposed method, this section first establishes the notation and problem setup for classification tasks. Subsequently, we describe how latent variables arise in classification tasks.

2.1 Notation and Problem Setup

Consider a classification problem with random variables (x, y), where x comprises the predictive features and $y \in \mathcal{Y}$ denotes the class label from a finite set $\mathcal{Y} := \{1, \dots, L\}$. Let $p(x, y)$ and $p(y|x)$ denote the joint and conditional probability distributions, respectively.

For this problem, we define a decision rule $\delta : \mathcal{X} \to \mathcal{Y}$ in the functional space Δ and evaluate it using the binary (0-1) loss function $\mathcal{L} : \mathcal{Y} \times \Delta \to \{0, 1\}$, defined as

$$\mathcal{L}\big(y, \delta(x)\big) := \mathbb{1}\big\{y \neq \delta(x)\big\}, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ represents the indicator function, taking 1 when the condition specified in the braces is satisfied and 0 otherwise.

Our goal is to find a decision rule $\delta \in \Delta$ that minimizes the the expected loss function, which is written as the misclassification probability:

$$\mathbb{E}\big[\mathcal{L}\big(y, \delta(x)\big)\big] = \Pr\{y \neq \delta(x)\}. \tag{2}$$

Furthermore, the misclassification probability can be expressed through the conditional distribution:

$$\Pr\{y \neq \delta(x)\} = \int_{\mathcal{X}} \Pr\{y{\neq}\delta(x)|x\} p(x) dx. \tag{3}$$

This formulation shows that minimizing the overall misclassification error is equivalent to minimizing $\Pr\{y{\neq}\delta(x)|x\}$ for each $x \in \mathcal{X}$. Since $\Pr\{y{\neq}\delta(x)|x\} = 1 - \Pr\{y{=}\delta(x)|x\}$, the optimal decision rule $\delta^* \in \Delta$ takes the form:

$$\delta^*(x) := \operatorname*{arg\,max}_{y \in \mathcal{Y}} p(y|x). \tag{4}$$

This decision rule, known as the Bayes optimal classifier, predicts the label y from features x by selecting the one with the highest conditional probability among all labels.

2.2 Latent-Variable Modeling

The classification framework can be extended to incorporate latent variables that model the underlying structure of the data. Let r be a possibly high-dimensional latent variable that resides in the space $\mathcal{R}$. The likelihood of observed data is represented as:

$$p(y|x) = \int_{\mathcal{R}} p(y|r, x) p(r|x) dr, \tag{5}$$

where $p(y|r, x)$ describes the enriched conditional probability incorporating the latent information, and $p(r|x)$ characterizes the latent probability conditional on observed features.

learns from such explanations. Thus, we use the terms *explanation* and *reasoning* interchangeably, as in He et al. (2024).

[3] While there is a zero-shot CoT approach, which does not provide reasoning examples in the context but only adds an instruction to perform such reasoning (Kojima et al., 2022), X-ICL builds on the few-shot CoT approach.

This latent variable r encodes additional structure about the relationship between features and labels, such as underlying reasons or mechanisms. This approach builds on fundamental work in classification analysis (Thurstone, 1927; Luce, 1959; Marschak, 1960; McFadden, 1973; Dempster et al., 1977), and has since become a cornerstone of modern classification methods. Importantly, latent variables exist for all possible labels, with the observed label emerging through their interaction. Formally, we define

$$\boldsymbol{r} := (r_1, \dots, r_L), \tag{6}$$

where r_ℓ denotes the latent element associated with label $\ell \in \mathcal{Y}$.

3 Methods

In this section, we first review ICL and X-ICL under the framework introduced in Section 2, and then we introduce X^2-ICL, our proposed method. Figure 1 presents an overview and highlights the differences among these methods. Hereafter, we refer to a set of *demonstrations* as a training sample $\mathcal{D}_n := \{(x_i, y_i)\}_{i=1}^n$ of n observations independently drawn from the joint distribution $p(x, y)$. Our task is to predict the label y' for a new feature variable x' in the test set.

3.1 ICL

ICL emerges as an important paradigm where LLMs learn to perform a task by observing demonstrations within its context window, without updating their parameters (Brown et al., 2020; Radford et al., 2019). The demonstrations $\mathcal{D}_n$ are concatenated into a prompt sequence that LLMs process, along with a new query point x', to predict the corresponding label y'.

Formally, we can explain the ICL decision rule in the likelihood framework as described in Section 2.1. LLMs identify and follow patterns in the demonstrations, which corresponds to obtaining a model trained for the task represented by the demonstrations. We denote the model implicitly learned through ICL as $\hat{p}(y|x)$, which is defined as follows:

$$\hat{p}(y|x) := p(y|x, \mathcal{D}_n). \tag{7}$$

Here, we omit the parameters of LLMs for brevity. The model predicts the most likely label for a new input x' through

$$\delta^{\mathrm{ICL}}(x') := \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|x'). \tag{8}$$

Algorithm 2 in Appendix A summarizes the above procedure.

The decision rule δ^{ICL} can be regarded as a map from the query input to a label prediction, conditioned on the provided set of demonstrations. The key distinction is that δ^{ICL} learns through pattern recognition within the demonstrations rather than parameter updates.

3.2 X-ICL: ICL with Explanations

X-ICL incorporates explanations into the demonstrations, thereby guiding LLMs to learn the reasoning process that connects features to labels (He et al., 2024). This approach can be interpreted as a form of latent-variable modeling, where the explanation serves as the latent variable r that capture the underlying structure governing the relationship between features and labels.

To augment the demonstrations with such explanations, X-ICL leverages LLMs to generate reasoning for each example through few-shot ICL. For this purpose, X-ICL additionally requires *meta-prompt*, a set of examples annotated with explanations: $\mathcal{S}_m := \{(x_j, r_j, y_j)\}_{j=1}^m$. Importantly, the meta-prompt can be quite small; in the tasks examined, a single example per label was sufficient (He et al., 2024). Therefore, the manual effort required to construct the meta-prompt $\mathcal{S}_m$ is significantly lower than that needed to manually create explanations for all examples in the demonstrations $\mathcal{D}_n$.

During preprocessing, for each data point in $\mathcal{D}_n$, X-ICL uses $\mathcal{S}_m$ to collect the reasoning r_y that justifies the assignment of label y to x. Here, r_y can be considered as the y-th element of $\boldsymbol{r}$ in Eq. (6). Formally, X-ICL approximates the conditional probability distribution of explanations as follows:

$$\tilde{p}(r|y, x) := p(r|y, x, \mathcal{S}_m), \tag{9}$$

and then samples reasoning according to:

$$r_y \sim \tilde{p}(r|y, x). \tag{10}$$

As a result of this augmentation process, the demonstration set in X-ICL becomes $\{(x_i, r_{y_i}, y_i)\}_{i=1}^n$.

In probabilistic terms, we express the joint distribution of label and its specific latent variable $p(y, r_y|x)$ as

$$p(y, r_y|x) = p(y|r_y, x)p(r_y|x). \tag{11}$$

During ICL with the augmented demonstrations, LLMs learn these two components: the latent variable generation process $\hat{p}(r_y|x)$ and the label prediction mechanism $\hat{p}(y|r_y, x)$. At inference time, given a test feature $x' \in \mathcal{X}$, X-ICL first samples a reasoning r'_y from $\hat{p}(r_y|x)$ and then determine the classification rule:

$$\delta^{\text{X-ICL}}(x') := \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|r'_y, x'). \tag{12}$$

Algorithm 3 in Appendix A summarizes the above procedure.

From the perspective of latent variable modeling, the problem of X-ICL is that it imposes restrictive constraints on the latent space, binding it tightly to realized values in the demonstrations. This constraint naturally leads to strong in-sample performance but significantly limits generalizability by creating a shallow latent structure that only spans a confined region of the possible latent space.

3.3 X^2-ICL: ICL with Exploration of Explanations

To mitigate the constraints imposed in X-ICL, we propose X^2-ICL, a method designed to effectively explore the latent reasoning space. Unlike conventional approaches that focus solely on the observed label, X^2-ICL examines the underlying rationale that could justify each possible label for the given input features. By doing so, our approach preserves the dimensionality of the latent reasoning space, allowing for a more comprehensive exploration of potential reasoning paths.

Intuitively, this approach maintains a broader consideration of plausible reasoning paths. This is particularly crucial in OOD scenarios, where the reasoning learned from demonstrations may not always be reliable, as it is constrained by the observed input–label pairs. In such cases, alternative reasoning processes, such as those based on different aspects of the input features, may still be valid. Our method aims to preserve these diverse reasoning paths, thereby enhancing robustness to the distribution shifts.

We employ the same meta-prompt to augment explanations within demonstrations. However, we extend this approach by augmenting not only the explanations for the observed labels but also those for unrealized labels. This provides a more comprehensive understanding of the reasoning space. We draw r_ℓ separately for each $\ell \in \mathcal{Y}$ as follows:

$$r_\ell \sim \tilde{p}(r_\ell|y = \ell, x), \tag{13}$$

In this step, instead of using only the observed label y, we consider what would be reasoning r if the label is ℓ. Considering all potential labels, we can uncover the collection of reasoning r in Eq. (6).

Algorithm 1 X^2-ICL

1: **Input:** $\mathcal{D}_n$ and $\mathcal{S}_m$
2: **for** $i = 1, \dots, n$ **do**
3:   **for** each label $\ell = 1, \dots, L$ **do**
4:     Generate a latent variable $r_{i,\ell}$ for (x_i, ℓ): $r_{i,\ell} \sim \tilde{p}(r_\ell|y_i = \ell, x_i)$
5:   **end for**
6:   Obtain a collection of latent variables: $\boldsymbol{r}_i = (r_{i,1}, \dots, r_{i,L})$
7: **end for**
8: **Estimation:**
9:   Augumented data: $\{(x_i, \boldsymbol{r}_i, y_i)\}_{i=1}^n$
10:   Model: $p(y, \boldsymbol{r}|x)$
11:   Estimator: $\hat{p}(y, \boldsymbol{r}|x) = \hat{p}(y|\boldsymbol{r}, x)\hat{p}(\boldsymbol{r}|x)$
12: **Inference:**
13:   Test input x'
14:   Draw $\boldsymbol{r}' := (r'_1, \dots, r'_L) \sim \hat{p}(\boldsymbol{r}|x')$
15:   Compute $\{\hat{p}(y'|\boldsymbol{r}', x') : y' \in \mathcal{Y}\}$
16: **Output:** Classification
17:   $y^* = \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|\boldsymbol{r}', x')$

By applying the above algorithm to the demonstrations, we obtain an augmented data $\{(x_i, \boldsymbol{r}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{r}_i := (r_{i,1}, \dots, r_{i,L})$. Note that $\boldsymbol{r}_i$ represents a set of reasoning paths, with each reasoning path $r_{i,\ell}$ corresponding to a specific label ℓ. The joint distribution of label and latent variable $p(y, \boldsymbol{r}|x)$ is expressed as

$$p(y, \boldsymbol{r}|x) = p(y|\boldsymbol{r}, x)p(\boldsymbol{r}|x). \tag{14}$$

Based on the augmented demonstrations, LLMs learn the label prediction mechanism $\hat{p}(y|\boldsymbol{r}, x)$ as well as the joint distribution of latent variables $\hat{p}(\boldsymbol{r}|x)$. For inference, our model takes a test input $x' \in \mathcal{X}$ and then draws a collection of latent variables $\boldsymbol{r}'$ from $\hat{p}(\boldsymbol{r}|x')$. Then, we obtain the prediction probability of each label by $\hat{p}(y'|\boldsymbol{r}', x')$ for every $y' \in \mathcal{Y}$. Finally, the classification rule is given by

$$\delta^{\text{X}^2\text{-ICL}}(x') := \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|\boldsymbol{r}', x'). \tag{15}$$

Algorithm 1 describes this estimation and inference procedure.

The three methods above represent distinct theoretical principles in their treatment of latent structures. ICL operates primarily through pattern recognition within observed samples, while X-ICL extends this framework by incorporating reasoning mechanisms for observed labels. Our approach diverges from these frameworks by adopting a more comprehensive treatment: it systematically uncovers latent reasoning structures for all possible labels, irrespective of whether they are explicitly realized in the observed data.

4 Experiments

In this section, we evaluate the OOD robustness of X^2-ICL, primarily through comparison with ICL and X-ICL across diverse NLU datasets.

4.1 Experimental Setup

Tasks. We employed two NLU tasks: natural language inference (NLI) and paraphrase identification. NLI is the task of classifying the semantic relationship between a hypothesis and a premise into one of three categories: *entailment*, *neutral*, or *contradiction*. Paraphrase identification is the task of determining whether a given pair of sentences has the same meaning, classifying them as either *yes* or *no*. Examples for each label are shown in Figures 4 and 5 in Appendix F.

Datasets. We used OOD evaluation datasets for NLI and paraphrase identification. To ensure the correctness of the labels, we selected datasets that were either manually curated and thoroughly verified or generated based on well-defined rules. In particular, for NLI, we employed HANS (McCoy et al., 2019), NAN (Truong et al., 2022), PISP (Liu et al., 2020a,b), ST (Naik et al., 2018), and ANLI$_{R1–R3}$ (Nie et al., 2020a), and for paraphrase identification, we used PAWS (Zhang et al., 2019). We used SNLI training data (Bowman et al., 2015) as demonstrations for evaluation on the NLI test sets, ANLI$_{R1}$ training data for the ANLI test sets, and QQP training data (Iyer et al., 2017) for PAWS, following X-ICL (He et al., 2024).[4] Although ICL does not involve updating the parameters of LLMs, it enables LLMs to learn patterns from demonstrations provided in the context. The evaluation data above are designed to reflect adversarial distribution shifts relative to the training data, making them suitable benchmarks for evaluating the OOD robustness of ICL. All the datasets are in English and are publicly available for model evaluation. See Appendix B for more details of the datasets.

Evaluation. We followed the evaluation protocol of X-ICL (He et al., 2024). The number of in-context demonstrations was set to eight, which is commonly referred to as the 8-shot ICL setting. We used four different random seeds to select both demonstrations and test samples, sampling up to 500 test examples from each test set (Jang et al., 2024). The reported scores correspond to the accuracy averaged over the four random seeds.

Models. We used three cutting-edge LLMs: GPT-4o (`gpt-4o-2024-08-06`), Gemini-1.5-Pro (`gemini-1.5-pro-002`), and Gemini-2.0-Flash (`gemini-2.0-flash-001`). To assess the broader applicability of our approach, we additionally employed more computationally efficient open-source models: Phi-4-14B[5] (Abdin et al., 2024) and DeepSeek-R1-8B[6] (Guo et al., 2025). Appendix C provides details on decoding configurations.

Explanations. To generate explanations for incorporation into demonstrations, we employed the few-shot meta-prompts introduced by X-ICL (He et al., 2024). The exact meta-prompts used in our experiments are provided in Figures 4 and 5 in Appendix F. These meta-prompts include only a single explanation per label, which minimizes the manual effort required to create explanation examples in advance. In all experimental settings, we used GPT-4o to generate the explanations incorporated into the demonstrations. Examples of X^2-ICL's prompts are presented in Figures 2 and 3 in Appendix F.

4.2 Main Results

Table 1 shows that X-ICL outperforms ICL on OOD datasets, and X^2-ICL further enhances OOD performance beyond X-ICL overall. The improvement is more pronounced when employing the higher-performing LLMs. This suggests that X^2-ICL relies on strong reasoning capabilities to be effective. Unlike X-ICL, which requires only a single reasoning path, X^2-ICL explores multiple

[4] Additional reason for using ANLI trainig data is that ANLI premises are generally much longer than those in other NLI datasets. This clear difference could interfere with evaluating the robustness that ANLI is intended to measure, namely, robustness to subtle changes that challenge models but are less noticeable to humans. To accurately assess the distribution shift targeted by ANLI, we avoid using demonstrations from other NLI datasets. Note that the test sets of ANLI are considered OOD even with respect to ANLI training data, due to differences in the rounds of dataset creation and in the annotators, which are known sources of distribution shifts (Nie et al., 2020a; Gururangan et al., 2018; Geva et al., 2019).

[5] `https://huggingface.co/unsloth/phi-4`

[6] `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B`

	Natural Language Inference								*Paraphrase*	
	SNLI	**HANS**	**NAN**	**PISP**	**ST**	**ANLI$_{R1}$**	**ANLI$_{R2}$**	**ANLI$_{R3}$**	QQP	**PAWS**
GPT-4o										
ICL	**90.95**$_{\pm 1.20}$	88.05$_{\pm 2.52}$	75.97$_{\pm 0.84}$	77.90$_{\pm 2.09}$	78.25$_{\pm 3.64}$	70.67$_{\pm 2.60}$	61.05$_{\pm 2.23}$	61.58$_{\pm 2.32}$	**83.65**$_{\pm 2.45}$	65.15$_{\pm 1.10}$
X-ICL	90.00$_{\pm 0.85}$	86.35$_{\pm 1.11}$	78.29$_{\pm 1.23}$	81.40$_{\pm 1.70}$	81.50$_{\pm 2.75}$	75.58$_{\pm 0.92}$	63.87$_{\pm 0.73}$	65.07$_{\pm 2.23}$	82.75$_{\pm 3.19}$	63.80$_{\pm 3.42}$
X^2-ICL	90.25$_{\pm 0.50}$	**88.85**$_{\pm 1.04}$	**78.78**$_{\pm 0.37}$	**83.76**$_{\pm 0.60}$	**82.35**$_{\pm 2.24}$	**77.40**$_{\pm 0.48}$	**67.61**$_{\pm 1.89}$	**67.70**$_{\pm 1.63}$	78.85$_{\pm 3.53}$	**70.85**$_{\pm 3.88}$
Gemini-1.5-Pro										
ICL	**89.80**$_{\pm 1.07}$	86.85$_{\pm 1.37}$	75.39$_{\pm 1.47}$	**80.66**$_{\pm 0.97}$	84.30$_{\pm 0.96}$	74.04$_{\pm 1.84}$	64.48$_{\pm 2.30}$	64.08$_{\pm 2.20}$	**83.00**$_{\pm 2.69}$	65.20$_{\pm 2.87}$
X-ICL	85.70$_{\pm 1.75}$	87.80$_{\pm 1.55}$	**77.23**$_{\pm 1.28}$	78.23$_{\pm 1.38}$	83.35$_{\pm 1.96}$	73.56$_{\pm 1.52}$	64.33$_{\pm 0.96}$	65.59$_{\pm 0.95}$	82.25$_{\pm 2.65}$	64.60$_{\pm 2.81}$
X^2-ICL	82.70$_{\pm 1.96}$	**88.40**$_{\pm 0.94}$	76.55$_{\pm 1.03}$	80.32$_{\pm 2.05}$	**85.60**$_{\pm 1.36}$	**75.48**$_{\pm 1.01}$	**66.77**$_{\pm 2.02}$	**67.11**$_{\pm 1.31}$	80.65$_{\pm 2.68}$	**69.80**$_{\pm 4.46}$
Gemini-2.0-Flash										
ICL	**87.15**$_{\pm 1.43}$	73.05$_{\pm 2.37}$	74.32$_{\pm 0.97}$	72.91$_{\pm 3.87}$	80.45$_{\pm 1.89}$	69.23$_{\pm 2.24}$	61.74$_{\pm 1.23}$	59.54$_{\pm 1.93}$	81.05$_{\pm 3.24}$	58.10$_{\pm 3.67}$
X-ICL	84.65$_{\pm 1.28}$	78.25$_{\pm 3.09}$	75.97$_{\pm 0.32}$	75.27$_{\pm 1.86}$	80.75$_{\pm 0.64}$	**75.29**$_{\pm 2.58}$	63.72$_{\pm 1.11}$	65.79$_{\pm 1.34}$	**82.85**$_{\pm 2.37}$	56.25$_{\pm 7.46}$
X^2-ICL	84.00$_{\pm 2.55}$	**85.40**$_{\pm 2.29}$	**76.07**$_{\pm 1.07}$	**83.69**$_{\pm 1.04}$	**82.50**$_{\pm 1.06}$	74.90$_{\pm 1.31}$	**65.93**$_{\pm 1.97}$	**67.37**$_{\pm 1.38}$	81.55$_{\pm 2.24}$	**66.40**$_{\pm 4.17}$
Phi-4-14B										
ICL	85.45$_{\pm 2.82}$	80.95$_{\pm 2.39}$	74.22$_{\pm 0.50}$	71.63$_{\pm 2.22}$	74.30$_{\pm 1.78}$	70.29$_{\pm 1.61}$	60.29$_{\pm 0.29}$	61.25$_{\pm 1.12}$	**81.10**$_{\pm 2.11}$	69.45$_{\pm 4.98}$
X-ICL	**90.00**$_{\pm 1.99}$	84.05$_{\pm 1.90}$	**78.49**$_{\pm 1.21}$	76.08$_{\pm 1.89}$	78.15$_{\pm 1.43}$	73.65$_{\pm 2.44}$	60.29$_{\pm 1.23}$	62.37$_{\pm 1.89}$	80.60$_{\pm 3.06}$	68.20$_{\pm 5.82}$
X^2-ICL	80.45$_{\pm 2.24}$	**84.45**$_{\pm 1.06}$	77.71$_{\pm 1.43}$	**78.37**$_{\pm 1.23}$	**80.15**$_{\pm 1.72}$	**75.38**$_{\pm 1.40}$	**63.34**$_{\pm 1.46}$	**63.42**$_{\pm 1.23}$	70.75$_{\pm 4.37}$	**75.90**$_{\pm 3.61}$
DeepSeek-R1-8B										
ICL	59.55$_{\pm 1.53}$	74.65$_{\pm 3.64}$	64.73$_{\pm 3.20}$	62.74$_{\pm 3.59}$	66.00$_{\pm 1.98}$	55.87$_{\pm 1.70}$	49.09$_{\pm 2.89}$	49.87$_{\pm 1.85}$	74.25$_{\pm 2.18}$	**48.55**$_{\pm 2.30}$
X-ICL	63.20$_{\pm 1.18}$	78.45$_{\pm 1.98}$	69.38$_{\pm 1.64}$	67.25$_{\pm 1.35}$	67.40$_{\pm 1.07}$	60.87$_{\pm 2.58}$	50.53$_{\pm 3.98}$	**50.59**$_{\pm 1.64}$	**76.60**$_{\pm 1.41}$	46.60$_{\pm 3.72}$
X^2-ICL	**66.10**$_{\pm 2.08}$	**81.65**$_{\pm 2.77}$	**70.45**$_{\pm 1.11}$	**68.73**$_{\pm 1.96}$	**68.10**$_{\pm 1.55}$	**63.56**$_{\pm 2.95}$	**52.13**$_{\pm 2.30}$	49.67$_{\pm 1.46}$	74.15$_{\pm 0.72}$	41.95$_{\pm 2.80}$

Table 1: Main results. The scores are the mean and standard deviation of accuracy computed across four different sets of demonstrations. The highest mean score for each model is highlighted in bold. SNLI and QQP share the same data distribution as the demonstrations, whereas the other datasets are OOD.

reasoning paths. Consequently, a higher reasoning capability is essential for its effectiveness.

In contrast, X-ICL and X^2-ICL exhibit degraded performance on SNLI and QQP, which share the same underlying distribution as the demonstrations. Due to the identical underlying distribution, simply following the patterns present in the demonstrations results in strong performance. However, the augmented explanations are drawn from LLMs and may not follow the exact distribution of the original input–label pairs. This deviation could lead to the degradation of the explanation-based methods in performance on the in-distribution data. This trade-off between in-distribution and OOD performance remains a challenge to be addressed in future work. However, given the consistent effectiveness of our approach across a diverse range of OOD datasets, we believe that our method makes a significant contribution to enhancing the robustness of ICL.

4.3 Comparison with Retrieval-Based ICL

While our method explores the reasoning space to achieve robust predictions, one might wonder whether exploring the demonstration space could achieve similar results. In ICL, selecting demonstrations that are semantically similar to a test input has been shown to perform well (Liu et al., 2022). Subsequent research has extended the retrieval-based approach to enhance the diversity of selected demonstrations (Levy et al., 2023; Ye et al., 2023; Gupta et al., 2023). To assess the effectiveness of exploring the demonstration space in OOD settings, we conducted a comparison with diversity-aware retrieval-based ICL.

We selected Set-BSR (Gupta et al., 2023) for comparison. Set-BSR uses BERTScore (Zhang et al., 2020) to measure semantic similarity and retrieves a diverse set of demonstrations that collectively cover various aspects of similarity to the test inputs. We employed one model each from the closed-source and open-source groups. In Table 2, we observed that Set-BSR improves performance in some cases, but overall, its performance improvements on OOD datasets are notably smaller than those achieved by the explanation-based ICL methods. This aligns with the original findings in the X-ICL paper (He et al., 2024), which also reported stronger OOD robustness for explanation-based ICL compared to retrieval-based ICL. These results suggest that diversity in the reasoning space contributes significantly more to OOD robustness than diversity in the demonstration space.[7]

Additionally, we emphasize that explanation-based ICL and retrieval-based ICL are designed for different scenarios. Retrieval-based ICL typi-

[7]These two types of diversity are not necessarily equivalent. For instance, even if the surface forms or topics of two premise–hypothesis pairs differ greatly, the underlying reasoning can still be identical, e.g., "the hypothesis contains information that is not explicitly mentioned in the premise, so the label is neutral."

	Natural Language Inference								Paraphrase	
	SNLI	HANS	NAN	PISP	ST	ANLI$_{R1}$	ANLI$_{R2}$	ANLI$_{R3}$	QQP	PAWS
GPT-4o										
Set-BSR	**90.60**$_{\pm 0.83}$	85.40$_{\pm 2.05}$	77.62$_{\pm 0.66}$	79.99$_{\pm 0.92}$	77.95$_{\pm 1.08}$	74.42$_{\pm 0.80}$	58.69$_{\pm 0.18}$	61.18$_{\pm 0.46}$	**85.20**$_{\pm 1.40}$	**72.25**$_{\pm 2.11}$
X-ICL	90.00$_{\pm 0.85}$	86.35$_{\pm 1.11}$	78.29$_{\pm 1.23}$	81.40$_{\pm 1.70}$	81.50$_{\pm 2.75}$	75.58$_{\pm 0.92}$	63.87$_{\pm 0.73}$	65.07$_{\pm 2.23}$	82.75$_{\pm 3.19}$	63.80$_{\pm 3.42}$
X^2-ICL	90.25$_{\pm 0.50}$	**88.85**$_{\pm 1.04}$	**78.78**$_{\pm 0.37}$	**83.76**$_{\pm 0.60}$	**82.35**$_{\pm 2.24}$	**77.40**$_{\pm 0.48}$	**67.61**$_{\pm 1.89}$	**67.70**$_{\pm 1.63}$	78.85$_{\pm 3.53}$	70.85$_{\pm 3.88}$
Phi-4-14B										
Set-BSR	87.65$_{\pm 1.64}$	83.35$_{\pm 2.22}$	76.94$_{\pm 1.50}$	71.50$_{\pm 1.53}$	76.30$_{\pm 1.28}$	**76.54**$_{\pm 0.00}$	56.40$_{\pm 0.00}$	58.42$_{\pm 0.00}$	78.00$_{\pm 2.23}$	75.07$_{\pm 1.76}$
X-ICL	**90.00**$_{\pm 1.99}$	84.05$_{\pm 1.90}$	**78.49**$_{\pm 1.21}$	76.08$_{\pm 1.89}$	78.15$_{\pm 1.43}$	73.65$_{\pm 2.44}$	60.29$_{\pm 1.23}$	62.37$_{\pm 1.89}$	**80.60**$_{\pm 3.06}$	68.20$_{\pm 5.82}$
X^2-ICL	80.45$_{\pm 2.24}$	**84.45**$_{\pm 1.06}$	77.71$_{\pm 1.43}$	**78.37**$_{\pm 1.23}$	**80.15**$_{\pm 1.72}$	75.38$_{\pm 1.40}$	**63.34**$_{\pm 1.46}$	**63.42**$_{\pm 1.23}$	70.75$_{\pm 4.37}$	**75.90**$_{\pm 3.61}$

Table 2: Comparison between retrieval-based ICL and explanation-based ICL. The top side of the dotted line corresponds to retrieval-based ICL, and the bottom side to explanation-based ICL. The notation follows that of Table 1.

cally requires access to large numbers of labeled examples to select effective demonstrations, while explanation-based ICL needs far fewer labeled examples and explanation annotations.[8] Thus, retrieval-based ICL is suitable when there is plenty of labeled data, whereas explanation-based ICL is effective when only minimal annotation is feasible.

4.4 Ablation Study

In this ablation study, we investigate the key factors contributing to the improved OOD performance of X^2-ICL. Specifically, we used GPT-4o to evaluate performance on the ANLI test sets, which are the most challenging among the OOD datasets in our main experiments. This level of difficulty is well-suited for detailed evaluation of the robustness of high-performing LLMs, as less challenging datasets may not reveal subtle differences.

(i) Systematic Exploration of Reasoning. Self-consistency (SC; Wang et al., 2023c), which stochastically samples multiple outputs from models, shares similarities with X^2-ICL in its ability to explore multiple reasoning paths. However, the key distinction lies in the structured nature of X^2-ICL's exploration: X^2-ICL systematically considers reasoning paths for all possible labels, whereas SC is constrained by the reasoning paths present between the given inputs and observed labels. To assess the impact of this systematic exploration, we compare SC with X^2-ICL. For a fair comparison, we align the number of reasoning paths in SC with the number of labels (three in NLI) and set the sampling temperature to 0.7, following Wang et al. (2023c). As shown in Table 3 (i), X^2-ICL consistently outperforms SC across all the ANLI test sets, indicating that the structured exploration is crucial in improving OOD robustness.

[8]For Set-BSR, we followed the setup described by Gupta et al. (2023) and used up to 44,000 demonstrations. For X-ICL and X^2-ICL, we used only eight demonstrations, in addition to a single explanation-annotated demonstration for each label, resulting in a total of 8 + 3 demonstrations in the case of NLI.

	ANLI$_{R1}$	ANLI$_{R2}$	ANLI$_{R3}$
X-ICL	75.58$_{\pm 0.92}$	63.87$_{\pm 0.73}$	65.07$_{\pm 2.23}$
Ablation			
(i) X-ICL + SC	75.96$_{\pm 0.80}$	65.62$_{\pm 1.33}$	65.53$_{\pm 2.38}$
(ii) X-ICL + Instruction	74.52$_{\pm 1.38}$	64.33$_{\pm 1.55}$	65.39$_{\pm 3.16}$
(iii-a) ICL + CoT	74.90$_{\pm 4.37}$	65.55$_{\pm 1.67}$	65.99$_{\pm 0.84}$
(iii-b) ICL + Explore CoT	74.42$_{\pm 3.21}$	65.55$_{\pm 3.07}$	65.66$_{\pm 2.60}$
X^2-ICL	**77.40**$_{\pm 0.48}$	**67.61**$_{\pm 1.89}$	**67.70**$_{\pm 1.63}$

Table 3: Ablation study examining the contribution of systematic reasoning exploration. The results for GPT-4o are reported. The notation follows that of Table 1.

(ii) Instruction for Exploration. X^2-ICL incorporates an instruction designed to guide reasoning exploration:

> *Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.*

See Figures 2 and 3 in Appendix F for examples. This instruction serves to clarify the purpose of providing multiple reasoning paths, and additionally, implicitly conveys that the entire label space is covered within these reasoning paths. A potential concern is that this instruction might offer an unfair advantage over X-ICL, as previous research reported that presenting information about the label space enhanced ICL performance (Min et al., 2022). To isolate the effect of this implicit label-space information from our overall method, we conducted a comparison with X-ICL augmented with a comparable instruction:

> *Instruction: Explain the reasoning and then select the correct label from entailment, neutral, or contradiction.*

As shown in Table 3 (ii), adding this instruction to X-ICL does not lead to significant performance improvements. The results indicate that the performance gains in X^2-ICL arise from the reasoning exploration encouraged by the instruction, rather than from the implicit label-space information.

(iii) Explanations within Demonstrations. We evaluated the necessity of including explanations within the demonstrations. This analysis is not only for X^2-ICL but also for X-ICL. LLMs possess strong reasoning capabilities, allowing them to engage in reasoning processes even without explicit demonstrations of how to do so within the given context (Kojima et al., 2022). To assess the benefit of X-ICL and X^2-ICL over this zero-shot reasoning, we appended the CoT prompt to the end of the input, without providing any explanations within the demonstrations (Kojima et al., 2022). Additionally, we included a sentence to ensure adherence to the specified output format. The full CoT prompt is as follows:

> *Let's think step by step. Ensure that your response ends with Label: followed by your final answer.*

The results in Table 3 (iii-a) indicate that the performance of this zero-shot CoT is comparable to that of X-ICL, suggesting that presenting a single reasoning path per demonstration does not provide substantial additional information to the cutting-edge LLM. However, X^2-ICL still outperforms this baseline, demonstrating that explicitly guiding LLMs to explore multiple reasoning paths in a systematic manner remains beneficial for enhancing the OOD robustness.

Next, we investigated whether reasoning paths can be systematically explored without providing demonstrations of such reasoning. To this end, we modified the above CoT prompt as follows:

> *Let's think step by step, exploring the reasons why each label could be correct. Ensure that your response ends with Label: and your final answer.*

The results in Table 3 (iii-b) show no significant performance difference compared to (iii-a). Upon inspection, we found that the model often failed to provide label-conditioned reasoning and instead produced generic step-by-step explanations. This shows that even advanced LLMs like GPT-4o struggle to generate comprehensive reasoning without demonstrations. Explanations in demonstrations are essential to guide LLMs in performing structured, comprehensive reasoning.

The results in (i)–(iii) show that the robustness of X^2-ICL arises from the systematic exploration of explanations learned from demonstrations.

4.5 Qualitative Analysis

We analyzed both the successful and failure cases of X^2-ICL to gain a deeper understanding of its behavior. Specifically, we examined the outputs generated by GPT-4o on the ANLI_{R1} test set. Due to space limitations, the outputs analyzed in this section are provided in Appendix G.

Successful Cases. In summary, X^2-ICL tends to guide the model to carefully examine the details of the input, thereby facilitating reasoning grounded in confidently inferable information. In contrast, X-ICL often draws conclusions from the absence of explicit information in the input. For instance, as illustrated in Figure 6, X^2-ICL deduced that a person whose name is associated with an equation is likely influential, even though this is not explicitly stated in the premise. Similarly, in Figure 7, X^2-ICL correctly inferred the end date of an American football player's career by performing a calculation based on details provided in the input, despite the end date not being explicitly mentioned. In both examples, X-ICL incorrectly assigned the neutral label based on the absence of explicit mention of the above information. In Addition, X^2-ICL was also effective in preventing the model from making hasty inferences by highlighting the absence of decisive information within the given context, as seen in Figures 8 and 9.

Among the 42 test instances where the answers of X-ICL and X^2-ICL differed, there were 24 cases where X-ICL predicted neutral, but X^2-ICL predicted entailment or contradiction. In contrast, only in 14 instances, X-ICL predicted a label other than neutral, but X^2-ICL predicted neutral. This suggests that X^2-ICL has a greater tendency to analyze the input more closely and draw on inferable information when making decisions.

Failure Cases. While generally beneficial, such in-depth consideration of details does not always lead to correct answers. For example, as seen in Figure 10, X^2-ICL incorrectly inferred a driver's race participation history based on his current career. In some cases, the reasoning appears plausible but does not match the gold label. In Figure 11, X^2-ICL concluded that the language of a film was not English because it was awarded in the Academy Award for Best Foreign Language Film category.

While this reasoning seems reasonable, it contradicts the gold label. Additionally, X^2-ICL sometimes relies on encyclopedic knowledge beyond the given premise, which is not intended for this task, leading to incorrect predictions. Figure 12 shows an example where X^2-ICL made inappropriate use of encyclopedic knowledge about a film director.

In 7 out of 100 test instances, X-ICL made correct predictions while X^2-ICL failed. Of these, three involved X^2-ICL using encyclopedic knowledge and two involved inferred information. In contrast, the remaining two were answered based on the absence of information, as shown in Figure 13.

The in-depth reasoning that incorporates more details is a key characteristic of X^2-ICL. Overall, as the experimental results demonstrate, this reasoning strategy enhances OOD robustness.

5 Related Work

OOD Robustness in ICL. Prior to X-ICL (He et al., 2024), several approaches were explored that did not incorporate explanations. Sun et al. (2024) introduced a method using random labels and minimal task descriptions to mitigate spurious correlations between labels and tasks. In Appendix E, we show that explanation-based approaches substantially outperform this approach. Jang et al. (2024) calibrated model output probabilities to reduce reliance on semantic priors within demonstrations. However, this calibration requires access to output probabilities, which critically limits its applicability to cutting-edge LLMs, which are often closed-source. Zhou et al. (2024) proposed balancing the concepts within demonstrations, but this balancing requires prior identification of potentially harmful concepts, which is an impractical requirement. This requirement has been acknowledged in the literature on OOD robustness as a significant practical limitation (Clark et al., 2019; Liu et al., 2021; Creager et al., 2021; Yang et al., 2023; Honda et al., 2024). Li et al. (2023) applied a prompt optimization method in OOD settings. However, their approach presupposes that the target distribution has already been identified and that a batch of unlabeled examples from that same distribution is available. Moreover, the optimization process involves repeated sampling from LLMs, which can be computationally expensive.

Unlike the above approaches, X-ICL and X^2-ICL achieve strong OOD performance without requiring knowledge of the underlying distribution or optimization, working effectively for both open-source and closed-source models.

ICL with Explanations. CoT reasoning has been shown to significantly improve ICL in both few-shot (Wei et al., 2022) and zero-shot settings (Kojima et al., 2022). Self-consistency (SC; Wang et al., 2023c) extends this approach by sampling multiple reasoning paths and aggregating their outcomes to improve overall performance. Our X^2-ICL also explores multiple reasoning paths; however, it differs from SC in that it systematically traverses these paths rather than relying on stochastic sampling. In Section 4.4 (i), we have demonstrated that this structured exploration leads to improved OOD generalization compared to SC.

OOD Robustness prior to ICL. OOD robustness has been one of the major challenges in machine learning, with extensive research conducted within the fine-tuning paradigm before the rise of ICL. Principal existing approaches include reweighting (Clark et al., 2019; He et al., 2019; Karimi Mahabadi et al., 2020), distributionally robust optimization (Sagawa et al., 2020), and invariant risk minimization (Arjovsky et al., 2019). The approach most similar to X-ICL is the one that leverages human-annotated rationales (Chen et al., 2022; Stacey et al., 2022; Kavumba et al., 2023; Ludan et al., 2023). Analogous to X^2-ICL, Honda et al. (2024) modeled various latent features and utilized them for prediction. Although these fine-tuning methods share conceptual similarities with approaches for improving ICL robustness, their reliance on extensive training data and parameter updates makes them unsuitable for direct use in ICL.

6 Conclusion

In this study, we introduced X^2-ICL, a novel framework designed to improve the robustness of ICL on OOD data. Our method extends X-ICL by incorporating a reasoning mechanism that systematically explores explanations for all possible labels, enabling a more comprehensive exploration of the reasoning space. Through extensive experiments, we have demonstrated that X^2-ICL exhibits strong robustness compared to both standard ICL and X-ICL, particularly in handling OOD scenarios. We believe that our findings, which emphasize the importance of explicit and comprehensive exploration of reasoning, represent a significant step toward developing robust ICL methods.

Limitations

X^2-ICL relies on a fixed label space to systematically explore the reasoning underlying the decision-making process. Thus, it is not directly applicable to open-ended tasks and may be inefficient when handling tasks with an excessively large label set.

As observed in Section 4.2, X^2-ICL exhibits lower performance on in-distribution datasets while demonstrating strong robustness to OOD datasets. Addressing this performance trade-off remains an open challenge. For example, investigating the adaptive use of X^2-ICL depending on the inputs to mitigate the trade-off would be an interesting direction for future work. Overall, we emphasize that this study represents a significant step toward improving OOD robustness in ICL.

One possible reason for the trade-off is the inherent label ambiguity in standard NLI datasets (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b; Jiang and Marneffe, 2022). The practice of single-label assignment in these datasets, without consideration of the ambiguity, may amplify distributional differences between datasets. Future work includes investigating additional-label or multi-label prediction approaches (Jiang and Marneffe, 2022) on ambiguity-aware datasets such as ChaosNLI (Nie et al., 2020b), which may help disentangle the impact of label ambiguity from the trade-off.

X^2-ICL requires models with advanced reasoning capabilities to function effectively, as discussed in Section 4.2. This necessity arises from the fact that our approach involves generating and evaluating multiple reasoning paths that are consistent with the given input. While this reliance on high-level reasoning presents a limitation, we believe that the rapid advancements in LLMs will progressively alleviate this issue over time.

Due to the comprehensive reasoning during inference, X^2-ICL incurs higher costs than ICL and X-ICL. Appendix D shows the comparison of the costs. This is another limitation of our approach and reducing the costs will be our future work. However, we emphasize that the primary focus and contribution of this study lies in proposing the effective ICL method for addressing OOD data, rather than improving computational efficiency.

Ethical Considerations

LLMs are known to exhibit biases toward certain social groups (Santurkar et al., 2023), which can inadvertently affect their outputs. However, we do not anticipate that our proposed method will reinforce such biases. X^2-ICL facilitates the exploration of diverse reasoning paths, which may contribute to mitigating inherent biases.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905v1*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893v3*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. 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.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948v1*.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.

Ukyo Honda, Tatsushi Oka, Peinan Zhang, and Masato Mita. 2024. Not eliminate but aggregate: Post-hoc control over mixture-of-experts to address shortcut shifts in natural language understanding. *Transactions of the Association for Computational Linguistics*, 12:1268–1289.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs. Accessed on Feb. 9, 2025.

Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. 2024. Rectifying demonstration shortcut in in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4294–4321, Mexico City, Mexico. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? {yes} it is {true} because {it weakens superficial cues}. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180, Dubrovnik, Croatia. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023. Robust prompt optimization for large language models against distribution shifts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1539–1554, Singapore. Association for Computational Linguistics.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Tianyu Liu, Zheng Xin, Baobao Chang, and Zhifang Sui. 2020a. HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6852–6860, Marseille, France. European Language Resources Association.

Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020b. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608, Online. Association for Computational Linguistics.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York. Reprinted by Dover Publications, 2005.

Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023. Explanation-based finetuning makes models more robust to spurious cues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4420–4441, Toronto, Canada. Association for Computational Linguistics.

Jacob Marschak. 1960. Binary choice constraints and random utility indicators. *Stanford Research Memorandum*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Daniel McFadden. 1973. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

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.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. 2024. In-context learning generalizes, but not always robustly: The case of syntax. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4761–4779, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10421, Bangkok, Thailand. Association for Computational Linguistics.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11349–11357.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.

L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Jindong Wang, Xixu HU, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2023a. On the robustness of chatGPT: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Zhuojun Jiang, Zhaoheng Zheng, Zhuofeng Wu, Muhao Chen, and Chaowei Xiao. 2023b. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950v2*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39584–39622. PMLR.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492, Bangkok, Thailand. Association for Computational Linguistics.

Algorithm 2 ICL

1: **Input:** $\mathcal{D}_n$
2: **Estimation:**
3: Data: $\mathcal{D}_n$
4: Model: $p(y|x)$
5: Estimator: $\hat{p}(y|x)$
6: **Inference:**
7: Test input: x'
8: Compute $\{\hat{p}(y'|x') : y' \in \mathcal{Y}\}$
9: **Output:** Classification
10: $y^* = \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|x')$

Algorithm 3 X-ICL

1: **Input:** $\mathcal{D}_n$ and $\mathcal{S}_m$
2: **for** $i = 1, \ldots, n$ **do**
3: Generate a latent variable r_{y_i} for x_i:
$$r_{y_i} \sim \tilde{p}(r_{y_i}|y_i, x_i)$$
4: **end for**
5: **Estimation:**
6: Augumented data: $\{(x_i, r_{y_i}, y_i)\}_{i=1}^n$
7: Model: $p(y, r_y|x)$
8: Estimator: $\hat{p}(y, r_y|x) = \hat{p}(y|r_y, x)\hat{p}(r_y|x)$
9: **Inference:**
10: Test input: x'
11: Draw $r'_y \sim \hat{p}(r_y|x')$
12: Compute $\{\hat{p}(y'|r'_y, x') : y' \in \mathcal{Y}\}$
13: **Output:** Classification
14: $y^* = \arg\max_{y' \in \mathcal{Y}} \hat{p}(y'|r'_y, x')$

A Algorithms

We present the algorithms for ICL and X-ICL in Algorithms 2 and 3, respectively.

B Dataset Details

HANS is characterized by a shift in the correlation between the labels and the word overlap of the premise and the hypothesis. In HANS, predictions based solely on word overlap can achieve only chance-level accuracy. NAN is constructed by combining multiple linguistic rules related to negation, which gives rise to rarely observed patterns in the standard NLI datasets. PISP identifies specific patterns occurring in the hypotheses of each label in NLI datasets and constructs a dataset consisting of examples that do not adhere to the observed pattern–label co-occurrence. ST synthesizes adversarial examples targeting multiple patterns that correlate with labels, including word overlap, negation, sentence length, etc. ANLI consists of three distinct splits, R1, R2, and R3, each generated through a different round of adversarial data collection conducted by human annotators. In our study, we specifically used the hard subsets,[9] in which the assigned labels were verified by a third human annotator to ensure reliability.[10] PAWS is a dataset in which the correlation between the labels and word overlap is adversarially shifted from QQP.

C Decoding Configurations

We used Azure OpenAI API to run GPT-4o and Gemini API for the Gemini models. The Azure content filters were set to the least restrictive setting. For the open-source models, we used Hugging Face Transformers library (Wolf et al., 2020).[11] The models were loaded with the default configurations and run on a single NVIDIA A100 GPU with 40GB of memory.

To maximize reproducibility, we set the decoding temperature to zero for the closed-source models: GPT-4o, Gemini-1.5-Pro, and Gemini-2.0-Flash. Note that deterministic decoding methods are not available for these models. For the open-source models, Phi-4 and DeepSeek-R1-8B, we employed greedy decoding.

During preliminary experiments, we observed that the open-source models exhibited weaknesses in adhering to the expected output format, often generating responses that were difficult to parse. To mitigate this issue, we incorporated the following system prompt to explicitly enforce adherence to the required output format:

> *Answer the question by following the provided examples. Ensure that your response ends with Label: and your final answer.*

D Comparison of Costs

We present the average costs per instance, including demonstrations prepended to each instance, based on experiments conducted with GPT-4o on the SNLI dataset using the same seed. Here, an "instance" refers to a single data point; in NLI tasks, this is a pair consisting of a premise and a hypothesis. We used OpenAI tiktoken with the

[9] https://github.com/xlhex/acl2024_xicl/tree/main/data/testset

[10] https://github.com/facebookresearch/anli/blob/main/mds/verifier_labels.md

[11] https://github.com/huggingface/transformers

	Number of Tokens		Wall-Clock Time	Expense
	Input	Output	Seconds	USD
ICL	366.3	3.3	0.38	0.0009
X-ICL	539.3	36.5	0.79	0.0017
X^2-ICL	1,240.3	120.3	1.83	0.0043

Table 4: Average costs per instance. Wall-clock time is measured in seconds, and expense in US dollars.

`o200k_base` encoding to count the tokens.[12] The expenses are calculated based on the number of tokens and the GPT-4o (`2024-08-06 Global`) pricing for the Azure OpenAI API as of May 2025.

Table 4 shows the results. We observed that X^2-ICL incurs higher computational costs than ICL and X-ICL due to its more comprehensive reasoning process. This is a limitation of our method, and reducing the costs remains future work. However, we emphasize that cost efficiency is not the primary focus of our work. Our main objective is to improve OOD robustness, and apart from the computational costs, X^2-ICL does not require any additional resources compared to X-ICL.

E Comparison with Other Non-Explanation-Based Methods

Sun et al. (2024) proposed simple methods to enhance the robustness of ICL. Their first finding is that increasing the number of demonstrations improves robustness. Additionally, they introduced a technique in which the original labels are replaced with random symbols, accompanied by a minimal description that maps these symbols to their corresponding labels. This approach aims to mitigate spurious correlations between labels and target tasks that LLMs may learn. Unlike other prior approaches, these methods offer a notable advantage in terms of ease of implementation: they only need to increase the number of demonstrations or replace the original labels with random symbols, without requiring any other modifications.

We evaluated the effectiveness of these methods in comparison to explanation-based approaches: X-ICL and X^2-ICL. For the implementation of the first method, we concatenated all four sets of demonstrations utilized in the main experiment, resulting in a single run of 32-shot ICL. For the second method, referred to as *mixed prompts*, we replaced the original labels with random symbols and provided a minimal task description, following the format of Sun et al. (2024). The task description we employed is as follows:

> *Given a premise and a hypothesis, if the premise entails the hypothesis, the answer is A4; if the hypothesis contradicts the premise, the answer is B6; and if the premise and hypothesis are unrelated, the answer is 7X.*

The results are reported in Table 5. In our experiments using GPT-4o on the ANLI test sets, we did not observe any improvement over standard ICL when applying these methods. This suggests that the benefits of these techniques may not consistently generalize across different datasets and model architectures. In contrast, explanation-based approaches have demonstrated effectiveness in a broader range of settings (see also Table 1), highlighting their robustness.

	ANLI$_{R1}$	ANLI$_{R2}$	ANLI$_{R3}$
ICL (8-shot)	$70.67_{\pm 2.60}$	$61.05_{\pm 2.23}$	$61.58_{\pm 2.32}$
Sun et al. (2024)			
ICL (32-shot)	70.00	60.06	63.16
ICL + mixed prompts	$70.10_{\pm 1.15}$	$57.93_{\pm 2.32}$	$57.83_{\pm 2.27}$
X-ICL (8-shot)	$75.58_{\pm 0.92}$	$63.87_{\pm 0.73}$	$65.07_{\pm 2.23}$
X^2-ICL (8-shot)	$\mathbf{77.40}_{\pm 0.48}$	$\mathbf{67.61}_{\pm 1.89}$	$\mathbf{67.70}_{\pm 1.63}$

Table 5: Comparison between explanation-based methods and those that do not incorporate explanations. The results for GPT-4o are reported. The notation follows that of Table 1.

F Prompt Examples

Figures 2 and 3 present examples of the X^2-ICL prompt. Figures 4 and 5 show the exact meta-prompts we used to generate explanations that were incorporated into the demonstrations.

G Output Examples

Figures 6–13 illustrate the output examples of X^2-ICL. See Section 4.5 for a detailed discussion and analysis of these examples.

[12] https://github.com/openai/tiktoken

X^2-ICL prompt for NLI

Premise: The Florida Scholastic Press Association (FSPA) is a not-for-profit organization founded in 1946. It is the scholastic press association for the state of Florida and its members consist of more than 300 student publications, online media teams and broadcast programs from the state. The main mission of the organization is to educate, train and support scholastic journalists and their advisers.
Hypothesis: The FSPA has over 300 members
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.
Possible Reasoning for entailment: The premise states that FSPA's members consist of more than 300 student publications, online media teams, and broadcast programs, which entails that FSPA has over 300 members.
Possible Reasoning for neutral: The premise states that the FSPA's members consist of more than 300 student publications, online media teams, and broadcast programs, but it does not specify the exact number of members, so the hypothesis that the FSPA has over 300 members is not definitively supported or contradicted.
Possible Reasoning for contradiction: The premise states that the FSPA's members consist of more than 300 student publications, online media teams, and broadcast programs, not individual members, so the hypothesis incorrectly interprets the premise.
Label: entailment
###
Premise: We Have Always Lived in the Castle is an upcoming American mystery thriller film directed by Stacie Passon and written by Passon and Mark Kruger, based on the 1962 novel of the same name by Shirley Jackson. The film stars Taissa Farmiga, Alexandra Daddario, Crispin Glover, and Sebastian Stan.
Hypothesis: We Have Always Lived in the Castle came out in 2018.
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.
Possible Reasoning for entailment: The premise states that "We Have Always Lived in the Castle" is an upcoming film, which implies it was released after the premise was written, supporting the hypothesis that it came out in 2018.
Possible Reasoning for neutral: The premise states that the film is upcoming, but it does not provide information about its release year, so it is unclear if it came out in 2018.
Possible Reasoning for contradiction: The premise states that the film is upcoming, which contradicts the hypothesis that it was released in 2018.
Label: contradiction
###
...
###
Premise: `{Test Premise}`
Hypothesis: `{Test Hypothesis}`
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.

Figure 2: An example of X^2-ICL prompt for NLI.

X^2-ICL prompt for paraphrase identification

Question 1: How can I be like Sheldon Cooper?
Question 2: How do I become like Sheldon Cooper?
Question: Do both questions ask the same thing?
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.
Possible Reasoning for no: Although both questions seek advice about becoming like Sheldon Cooper, "can I be" asks about the possibility, while "how do I become" asks about the specific steps or process, making them technically different questions.
Possible Reasoning for yes: Both questions ask about ways to become/emulate the character Sheldon Cooper, using different but synonymous phrasing ("be like" vs. "become like").
Label: yes
###
Question 1: Why can't some people think for themselves?
Question 2: Why don't people think for themselves?
Question: Do both questions ask the same thing?
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.
Possible Reasoning for no: While similar, the first question asks about "some people" who are unable to think for themselves, while the second question asks about people in general who choose not to think for themselves. The scope and implication of inability versus choice are different.
Possible Reasoning for yes: Both questions ask about why people don't/can't engage in independent thinking, with slightly different wording but the same core meaning.
Label: no
###
...
###
Question 1: `{Test Question 1}`
Question 2: `{Test Question 2}`
Question: Do both questions ask the same thing?
Instruction: Explore the reasoning behind all the labels. Then, select the label that has the most valid reasoning.

Figure 3: An example of X^2-ICL prompt for paraphrase identification.

Meta-prompt for NLI

Assume that you're an expert working on natural language inference tasks. Given a premise, a hypothesis, and the corresponding label, please write a concise and precise reason to explain why the label is assigned to the example by following the provided examples:

Premise: A boy peers out of an open window.
Hypothesis: The boy looks out the window.
Label: entailment
Reason: The boy peers out of a window, so the boy looks out the window.
###
Premise: A kid doing a trick on a skateboard.
Hypothesis: The kid eating lunch inside the cafeteria.
Label: contradiction
Reason: The kid cannot be doing a trick and eating lunch at the same time
###
Premise: A man jumps off of his skateboard on the top of a cement ramp.
Hypothesis: a man jumps off a skateboard at the top of a ramp.
Label: neutral
Reason: A man can jump off a skateboard without being at the top of a ramp.
###
Premise: `{Demonstration Premise}`
Hypothesis: `{Demonstration Hypothesis}`
Label: `{Demonstration Label}`

Figure 4: Meta-prompt for generating explanations to be included in NLI demonstrations.

Meta-prompt for paraphrase identification

Assume that you're an expert working on paraphrase identification tasks. Given questions 1 and 2 and the corresponding label, please write a concise and precise reason to explain why the label is assigned to the example by following the provided examples:

Question 1: Does life get harder as you get older?
Question 2: Does life really get harder as you get older?
Question: Do both questions ask the same thing?
Label: yes
Reason: Both questions ask whether life does get harder as you get older.
###
Question 1: What is the National nanotechnology initiative?
Question 2: What is the lead time for SSN4EGS411 board?
Question: Do both questions ask the same thing?
Label: no
Reason: completely different questions
###
Question 1: `{Demonstration Question 1}`
Question 2: `{Demonstration Question 2}`
Question: Do both questions ask the same thing?
Label: `{Demonstration Label}`

Figure 5: Meta-prompt for generating explanations to be included in paraphrase-identification demonstrations.

Successful case of X^2-ICL

Premise: In astrophysics, the Chandrasekhar equation is a dimensionless form of the Poisson equation for the density distribution of a spherically symmetric isothermal gas sphere subjected to its own gravitaional force, named after the Indian American astrophysicist Subrahmanyan Chandrasekhar. The equation reads
Hypothesis: Subrahmanyan Chandrasekhar was a very influential scientist.
Label: entailment

Output of X-ICL
Reason: The premise mentions that the Chandrasekhar equation is named after Subrahmanyan Chandrasekhar, suggesting his influence in the field of astrophysics. However, it does not provide enough information to definitively conclude the extent of his influence.
Label: neutral

Output of X^2-ICL
Possible Reasoning for entailment: The premise mentions that the Chandrasekhar equation is named after Subrahmanyan Chandrasekhar, which suggests he was influential enough to have an equation named after him.
Possible Reasoning for neutral: The premise provides information about the Chandrasekhar equation but does not explicitly state the level of influence Subrahmanyan Chandrasekhar had in the scientific community.
Possible Reasoning for contradiction: The premise does not provide any information that contradicts the hypothesis about Subrahmanyan Chandrasekhar's influence.
Label: entailment

Figure 6: An example in which X^2-ICL provides the correct answer, whereas X-ICL fails. X^2-ICL enabled the model to carefully examine the details of the input and to reason based on confidently inferred information.

Successful case of X^2-ICL

Premise: David Carter (born November 27, 1953 in Vincennes, Indiana) is a former American football offensive guard and center who played 9 seasons in the National Football League for the Houston Oilers and New Orleans Saints. He played college football at Western Kentucky University.
Hypothesis: David Carter played professional football in the NFL until he was 52 years old.
Label: contradiction

Output of X-ICL
Reason: The premise does not provide information about the specific years David Carter played in the NFL, only that he played for 9 seasons. Without knowing the years he played, we cannot determine if he played until he was 52 years old.
Label: neutral

Output of X^2-ICL
Possible Reasoning for entailment: The premise provides information about David Carter's birthdate and his NFL career, but it does not specify the years he played, so it cannot confirm or deny the hypothesis.
Possible Reasoning for neutral: The premise states that David Carter played 9 seasons in the NFL, but it does not provide the specific years of his career, leaving his age at retirement uncertain.
Possible Reasoning for contradiction: Given that David Carter was born in 1953, if he played until he was 52 years old, he would have been playing in 2005. This is unlikely given the typical career length of NFL players and the fact that he played 9 seasons, suggesting he did not play until he was 52.
Label: contradiction

Figure 7: An example in which X^2-ICL provides the correct answer, whereas X-ICL fails. X^2-ICL enabled the model to carefully examine the details of the input and to reason based on confidently inferred information.

Successful case of X^2-ICL

Premise: Thank You Happy Birthday is the second studio album by American rock band Cage the Elephant. It was released on January 11, 2011 to positive critical reception. The album was produced by Jay Joyce, who worked in the same capacity on the band's eponymous debut album.
Hypothesis: Cage the Elephant had positive reviews thanks to the work of Jay Joyce who produced it.
Label: neutral

Output of X-ICL
Reason: The premise states that "Thank You Happy Birthday" received positive critical reception and was produced by Jay Joyce. This supports the hypothesis that the positive reviews were due, at least in part, to Jay Joyce's production work.
Label: entailment

Output of X^2-ICL
Possible Reasoning for entailment: The premise states that "Thank You Happy Birthday" received positive critical reception and was produced by Jay Joyce, suggesting that his production work contributed to the positive reviews.
Possible Reasoning for neutral: The premise states that the album received positive reviews and was produced by Jay Joyce, but it does not explicitly link the positive reviews directly to his work, leaving room for other factors to have contributed to the reception.
Possible Reasoning for contradiction: The premise does not provide any information that contradicts the hypothesis, as it acknowledges both the positive reviews and Jay Joyce's role as producer.
Label: neutral

Figure 8: An example in which X^2-ICL provides the correct answer, whereas X-ICL fails. X^2-ICL prevented the model from making hasty inferences by drawing attention to the lack of decisive information in the given context.

Successful case of X^2-ICL

Premise: The Best of David Bowie 1974/1979 is a compilation album by David Bowie released in 1998 (see 1998 in music). It follows "The Best of David Bowie 1969/1974" (1997) and includes material released between 1974–1979. This album was also included as the second disc of the compilation "The Platinum Collection" (2005/2006).
Hypothesis: David Bowie didn't only release an album in 1998 but also in 1979.
Label: neutral

Output of X-ICL
Reason: The premise mentions that "The Best of David Bowie 1974/1979" is a compilation album released in 1998, which includes material from 1974 to 1979. This implies that David Bowie released music during the years 1974 to 1979, including 1979, in addition to the compilation album released in 1998.
Label: entailment

Output of X^2-ICL
Possible Reasoning for entailment: The premise mentions that "The Best of David Bowie 1974/1979" is a compilation album released in 1998, which includes material from 1974 to 1979. This implies that David Bowie released music during those years, including 1979.
Possible Reasoning for neutral: The premise confirms the release of a compilation album in 1998 and includes material from 1974 to 1979, but it does not explicitly state that an album was released in 1979.
Possible Reasoning for contradiction: The premise does not explicitly state that David Bowie released an album in 1979, only that the compilation includes material from that year, so it does not directly support the hypothesis.
Label: neutral

Figure 9: An example in which X^2-ICL provides the correct answer, whereas X-ICL fails. X^2-ICL prevented the model from making hasty inferences by drawing attention to the lack of decisive information in the given context.

Failure case of X^2-ICL

Premise: Garrett Smithley (born April 27, 1992) is an American professional stock car racing driver. He currently competes full-time in the NASCAR Xfinity Series, driving the No. 0 Chevrolet Camaro for JD Motorsports and the No. 40 Toyota Camry for MBM Motorsports. Smithley has also competed in the Camping World Truck Series and ARCA Racing Series.
Hypothesis: Garrett Smithley used to compete in the NASCAR Xfinity Series.
Label: contradiction

Output of X-ICL
Reason: The premise states that Garrett Smithley currently competes full-time in the NASCAR Xfinity Series, indicating that he is still active in the series, not that he used to compete in it.
Label: contradiction

Output of X^2-ICL
Possible Reasoning for entailment: The premise states that Garrett Smithley currently competes full-time in the NASCAR Xfinity Series, which implies that he used to compete in it as well.
Possible Reasoning for neutral: The premise provides information about Garrett Smithley's current participation in the NASCAR Xfinity Series, but it does not explicitly state his past involvement, leaving room for interpretation.
Possible Reasoning for contradiction: The premise states that Garrett Smithley currently competes in the NASCAR Xfinity Series, which does not contradict the hypothesis that he used to compete in it.
Label: entailment

Figure 10: An example in which X-ICL provides the correct answer, whereas X^2-ICL fails. X^2-ICL caused the model to reason using information that was weakly inferable yet incorrect.

Failure case of X^2-ICL

Premise: Gyula Trebitsch (3 November 1914 - 12 December 2005) was a German film producer born in Budapest, Hungary. He was nominated in 1956 for the Academy Award for Best Foreign Language Film along with Walter Koppel for their film "The Captain of Kopenick".
Hypothesis: Gyula Trebitsch produced films in english.
Label: neutral

Output of X-ICL
Reason: The premise does not provide information about the language of the films produced by Gyula Trebitsch. It only mentions that he was nominated for an Academy Award for a film, but does not specify the language of that film or any other films he produced.
Label: neutral

Output of X^2-ICL
Possible Reasoning for entailment: The premise mentions that Gyula Trebitsch was nominated for an Academy Award for Best Foreign Language Film, which implies that he produced films in a language other than English, not in English.
Possible Reasoning for neutral: The premise provides information about Gyula Trebitsch's nomination for a foreign language film, but it does not specify whether he produced films in English or not, leaving the possibility open.
Possible Reasoning for contradiction: The premise indicates that Gyula Trebitsch was nominated for a Best Foreign Language Film, suggesting that his work was not in English, which contradicts the hypothesis.
Label: contradiction

Figure 11: An example in which X-ICL provides the correct answer, whereas X^2-ICL fails. X^2-ICL caused the model to reason using information that was inferable yet incorrect with respect to the gold label assigned.

Failure case of X^2-ICL

Premise: Seven Seas to Calais (Italian: Il dominatore dei sette mari) is a 1962 Italian adventure film by Eastmancolor in CinemaScope, directed by Rudolph Maté (his final film) and Primo Zeglio. It stars Rod Taylor, Keith Michell and Edy Vessel. The film depicts the career of Sir Francis Drake.
Hypothesis: Rudolph Maté began directing films in 1960.
Label: neutral

Output of X-ICL
Reason: The premise states that "Seven Seas to Calais" was directed by Rudolph Maté in 1962, but it does not provide information about when he began directing films. Therefore, the hypothesis cannot be confirmed or denied based on the given information.
Label: neutral

Output of X^2-ICL
Possible Reasoning for entailment: The premise states that "Seven Seas to Calais" was directed by Rudolph Maté and released in 1962, which could imply that he began directing films around 1960.
Possible Reasoning for neutral: The premise provides information about the film "Seven Seas to Calais" and its directors but does not provide any information about when Rudolph Maté began his directing career.
Possible Reasoning for contradiction: The premise does not provide any information about when Rudolph Maté began directing films, so it cannot be concluded that he started in 1960. Additionally, Rudolph Maté had a long career in film before 1960, which contradicts the hypothesis.
Label: contradiction

Figure 12: An example in which X-ICL provides the correct answer, whereas X^2-ICL fails. X^2-ICL caused the model to rely on encyclopedic knowledge beyond the given context, which is not appropriate for the NLI task.

Failure case of X^2-ICL

Premise: Oksa is a village in Jędrzejów County, Świętokrzyskie Voivodeship, in south-central Poland. It is the seat of the gmina (administrative district) called Gmina Oksa. It lies approximately 18 km north-west of Jędrzejów and 41 km south-west of the regional capital Kielce. The village has an approximate population of 890.
Hypothesis: The population of Oksa is large.
Label: contradiction

Output of X-ICL
Reason: The premise states that the population of Oksa is approximately 890, which is not considered large for a village. The term "large" is subjective, but typically a population of 890 would be considered small or average for a village.
Label: contradiction

Output of X^2-ICL
Possible Reasoning for entailment: The premise states that Oksa has an approximate population of 890, which could be considered large for a village.
Possible Reasoning for neutral: The premise provides the population of Oksa, but whether 890 is considered large is subjective and depends on the context or comparison to other villages.
Possible Reasoning for contradiction: The premise states that Oksa has an approximate population of 890, which is generally not considered large, especially when compared to cities or larger towns.
Label: neutral

Figure 13: An example in which X-ICL provides the correct answer, whereas X^2-ICL fails. In contrast to the other examples, X^2-ICL failed to utilize the inferable information in this case, which led the model to respond based on the absence of explicit information in the given context.