# Using Natural Language Explanations to Improve Robustness of In-context Learning

**Xuanli He**[♣] **Yuxiang Wu**[♥] **Oana-Maria Camburu**[♣]
**Pasquale Minervini**[♠] **Pontus Stenetorp**[♣]
[♣]University College London [♥]Weco AI [♠]University of Edinburgh
z.xuanli.he@gmail.com yuxiang@weco.ai p.minervini@ed.ac.uk
{o.camburu, p.stenetorp}@ucl.ac.uk

## Abstract

Recent studies demonstrated that large language models (LLMs) can excel in many tasks via in-context learning (ICL). However, recent works show that ICL-prompted models tend to produce inaccurate results when presented with adversarial inputs. In this work, we investigate whether augmenting ICL with natural language explanations (NLEs) improves the robustness of LLMs on adversarial datasets covering natural language inference and paraphrasing identification. We prompt LLMs with a small set of human-generated NLEs to produce further NLEs, yielding more accurate results than both a zero-shot-ICL setting and using only human-generated NLEs. Our results on five popular LLMs (GPT3.5-turbo, Llama2, Vicuna, Zephyr, and Mistral) show that our approach yields over 6% improvement over baseline approaches for eight adversarial datasets: HANS, ISCS, NaN, ST, PICD, PISP, ANLI, and PAWS. Furthermore, previous studies have demonstrated that prompt selection strategies significantly enhance ICL on in-distribution test sets. However, our findings reveal that these strategies do not match the efficacy of our approach for robustness evaluations, resulting in an accuracy drop of 8% compared to the proposed approach.[1]

## 1 Introduction

The landscape of AI has recently undergone a significant transformation with the advent of large language models (LLMs). These models can produce accurate predictions on unseen data after observing a small number of demonstrations. Remarkably, they can achieve this based on examples provided directly in their inputs, without explicit retraining or fine-tuning – this learning paradigm is referred to as *in-context learning* (ICL, Brown et al., 2020; Rae et al., 2021). However, ICL struggles to execute complex tasks, such as arithmetic, commonsense, and symbolic reasoning (Rae et al., 2021). To improve the effectiveness of ICL in solving tasks requiring complex reasoning, Wei et al. (2022b) drew inspiration from natural language explanations (NLEs) to introduce a method denoted as the Chain-of-Thought (CoT) prompting. CoT prompting involves prompting a model with a sequence of intermediate steps or reasoning processes to guide it towards generating more accurate answers.[2] In this work, we denote ICL equipped with NLEs as *X-ICL*. Despite its simplicity, X-ICL has advanced the performance of ICL across a broad range of complex reasoning tasks (Wei et al., 2022b; Wang et al., 2023b).

Similarly to supervised learning, ICL tends to be vulnerable to adversarial examples (Wang et al., 2023a). Previous research shows that improving the robustness of fine-tuned models against such adversarial datasets is possible by fine-tuning with task-relevant NLEs (Chen et al., 2022; Ludan et al., 2023). Inspired by this, we hypothesize that incorporating NLEs into ICL could also improve the robustness of LLMs against adversarial examples. To this end, we evaluate the robustness of X-ICL on eight adversarial datasets: HANS, ISCS, NaN, ST, PICD, PISP, ANLI, and PAWS.

Moreover, the effectiveness of X-ICL so far relies on the availability of human-written NLEs (Wei et al., 2022b), which usually require domain-specific knowledge, making them hard to collect. However, the advent of LLMs uncovered a range of possibilities where LLMs can assist human annotators (Bang et al., 2023; Guo et al., 2023). Motivated by this development, we investigate using three LLMs, namely GPT3.5-turbo, Llama2, and Vicuna,

[1]Code and datasets are accessible at: `https://github.com/xlhex/acl2024_xicl`

[2]CoTs and NLEs are similar concepts, as they both describe the reasoning process behind a decision in natural language; as NLEs were introduced before CoTs (Camburu et al., 2018; Hendricks et al., 2018), we use the former term.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499
August 11-16, 2024 ©2024 Association for Computational Linguistics

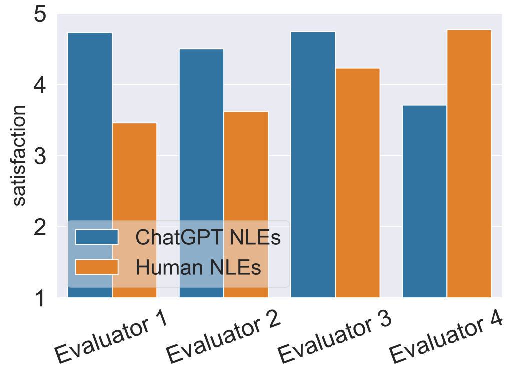


Figure 1: Human evaluation on 100 NLEs generated by GPT3.5-turbo (labeled as *ChatGPT NLEs*) and 100 NLEs generated by human annotators (labeled as *Human NLEs*). The satisfaction scores span from 1 (extremely dissatisfied) to 5 (extremely satisfied).

to generate NLEs for ICL. We then use human annotators to assess the quality of 200 human-written and LLM-generated NLEs. As shown in Figure 1, most annotators (3 out of 4) prefer NLEs produced by ChatGPT (GPT3.5-turbo) over those crafted by humans.[3] This observation further motivates us to evaluate models prompted with LLM-generated NLEs.

We then evaluate the improvement in the robustness of X-ICL in three settings – in two of the settings, an LLM is prompted with LLM-generated NLEs (generated in zero-shot-ICL and few-shot-ICL settings, and in the last setting, the LLM is prompted with human-generated NLEs. In the evaluation, we consider five popular LLMs (*i.e.,* Mistral (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), Vicuna (Chiang et al., 2023), Llama2 (Touvron et al., 2023) and GPT3.5-turbo) on eight adversarial datasets. Our experimental results suggest that X-ICL produces more accurate results than ICL and, moreover, that NLEs generated by ChatGPT in a few-shot-ICL setting (by prompting ChatGPT with human-generated NLEs) significantly improve over the ICL baseline (+6%) for the majority of the considered datasets and LLMs. Thus, our findings suggest that an integrated approach, combining human inputs with LLMs, can provide a more effective solution than utilizing either human annotators or LLMs in isolation. Finally, we show that while prompt selection strategies (*i.e.,* retrieving relevant training examples) can significantly improve the accuracy of ICL on in-distribution test sets (Gupta et al., 2023; Levy et al., 2023; Ye et al., 2023), they are less effective on adversarial datasets when compared to X-ICL methods, with our approach (few-shot-ICL) outperforming them by more than 8% in accuracy.

## 2 Related Work

**Learning with Explanations.** There has been a surge of work on explaining predictions of neural NLP systems, from highlighting decision words (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017; Serrano and Smith, 2019) to generating NLEs (Camburu et al., 2018; Narang et al., 2020; Wiegreffe and Marasovic, 2021). Our work concentrates on the latter category, namely, the self-generation of NLEs for justifying model predictions. Rajani et al. (2019) propose a two-stage training process to improve the prediction performance for commonsense reasoning tasks. In their work, the first stage revolves around generating NLEs, which are then used to inform the label prediction training process in the second stage. Alternatively, one can leverage a multi-task framework to generate NLEs and labels simultaneously (Hase et al., 2020). Li et al. (2022) propose advancing the reasoning abilities of smaller LMs by leveraging NLEs generated by GPT-3 (Brown et al., 2020). NLEs have also vastly been employed beyond NLP, such as in computer vision (Hendricks et al., 2018; Zellers et al., 2019; Majumder et al., 2022), in the medical domain (Kayser et al., 2022), and for self-driving cars (Kim et al., 2018), with some works showing improved task performance when training with NLEs (Kayser et al., 2021). However, these studies primarily concentrate on supervised fine-tuning approaches, which is different from the focus of this work, *i.e.,* ICL.

**Prompting with NLEs.** Despite its remarkable performance on several downstream tasks (Brown et al., 2020), ICL can still produce inaccurate results in tasks requiring reasoning abilities, such as arithmetic, logical, and commonsense reasoning tasks (Rae et al., 2021; Srivastava et al., 2022). To improve the reasoning abilities of LLMs, Wei et al. (2022b) introduced CoT prompting. This technique prompts an LM to generate a sequence of concise sentences that imitate the reasoning process an individual might undergo to solve a task before providing the ultimate answer, essentially to provide an NLE/CoT before generating the final answer. Furthermore, Wang et al. (2023b) propose to improve CoT prompting by combining multiple diverse reasoning paths generated by LLMs, en-

[3]More details are available in Appendix D.1.

hancing the accuracy of a greedy CoT prompting approach. However, these aforementioned methods need human-written NLEs as CoT in the prompts. Instead, our LLM-based zero-shot-ICL regime harnesses the power of an LLM to synthesize NLEs without human-written NLEs.

**Learning Robust Models.** Several works show that NLP models are prone to performance degradation when presented with adversarial examples, a consequence of inherent artifacts or biases within the annotation of the training dataset (Naik et al., 2018; McCoy et al., 2019; Nie et al., 2020; Liu et al., 2020b). Various strategies have been proposed to mitigate biases within NLP models, *e.g.,* initially training a weak model to recognize superficial features, subsequently enforcing a target model to learn more robust and generalizable characteristics (He et al., 2019; Clark et al., 2019; Karimi Mahabadi et al., 2020; Yaghoobzadeh et al., 2021; Korakakis and Vlachos, 2023). Additionally, data augmentation presents another viable option (Minervini and Riedel, 2018; Wu et al., 2021, 2022). Moreover, studies have shown that supervised finetuning of models using rationales or human-written NLEs can significantly enhance the models' resilience against adversarial datasets (Chen et al., 2022; Stacey et al., 2022; Kavumba et al., 2023; Ludan et al., 2023). Unlike them, our research examines the robustness of X-ICL across eight adversarial datasets, highlighting a novel finding: NLEs generated by LLMs surpass those produced by human annotators in enhancing model robustness. In addition, unlike human-written NLEs, those produced by LLMs exhibit greater scalability and adaptability across diverse tasks.

# 3 Methodology

This section first outlines the workflow of X-ICL. Then, the focus shifts to detailing how an LLM can generate an NLE for a labeled instance.

## 3.1 ICL with NLEs (X-ICL)

LLMs can provide significantly more accurate predictions across various reasoning tasks when supplied with human-written NLEs (Wei et al., 2022b,a).

In X-ICL, given an instance, the task is to generate the most likely prediction and NLE for that instance. More formally, in X-ICL, given an unlabeled instance $\boldsymbol{x}' \in \mathcal{X}$ and a set of training examples $(\boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i \in \mathcal{X}$ is an instance, $\boldsymbol{y}_i \in \mathcal{Y}$ is its label, and $\boldsymbol{r}_i \in \mathcal{E}$ is the corresponding explanation, the task is to identify the most likely label and explanation for $\boldsymbol{x}'$:

$$\arg\max_{(\boldsymbol{r}',\boldsymbol{y}')\in\mathcal{E}\times\mathcal{Y}} P_\theta\left((\boldsymbol{r}',\boldsymbol{y}') \mid (\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{y}_i)_{i=1}^{k},(\boldsymbol{x}')\right),$$

where $\theta$ denotes the model parameters, and $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{E}$ are the sets of all possible instances, labels, and explanations, respectively.

The objective is to generate the most likely combination of label $\boldsymbol{y}'$ and explanation $\boldsymbol{r}'$ from an LLM, after prompting it with the demonstration examples, including labeled instances and NLEs $(\boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{y}_i)_{i=1}^{k}$, as well as the unlabeled instance $\boldsymbol{x}'$.

## 3.2 Generating NLEs with LLMs

In existing X-ICL works, human-written NLEs $\boldsymbol{r}$ were used for the instances within the demonstration set. Instead, in this work, we opt for the NLEs synthesized via LLMs. This preference is driven by noting that NLEs produced by LLMs tend to receive higher approval ratings from human evaluators, as indicated in Figure 1. We argue that this preference will boost the performance of X-ICL. The methods utilized for the generation of NLEs are outlined below.

**Few-shot prompting for NLEs** Our methodology, also shown in Figure 2, initiates by leveraging a set of labeled instances, each accompanied by a human-crafted NLE, to prompt LLMs. The primary aim is to encourage the LLMs to generate a correct NLE (*i.e.,* the ground-truth arguments) for the correctly predicted answer for a test instance. The most likely NLE is then generated as follows:

$$\arg\max_{\boldsymbol{r}'\in\mathcal{E}} P_\theta(\boldsymbol{r}' \mid \boldsymbol{s}, (\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{r}_j)_{j=1}^{m}, (\boldsymbol{x}',\boldsymbol{y}')), \quad (1)$$

where $\boldsymbol{s}$ denotes a meta-prompt representing the task. More details on the meta-prompt and demonstration sets are available in Appendix B.

**Zero-shot prompting for NLEs** We further extend our approach to situations where human-written NLEs are absent, which is generally more prevalent across most datasets. In this context, LLMs are prompted to generate an NLE for a labeled instance devoid of any pre-existing examples with NLEs. The objective bears a resemblance to Equation (1), albeit without the inclusion of the demonstration set $(\boldsymbol{x}_j, \boldsymbol{y}_j, \boldsymbol{r}_j)_{j=1}^{m}$.

Notably, the NLEs generated by the aforementioned approaches can be seamlessly integrated

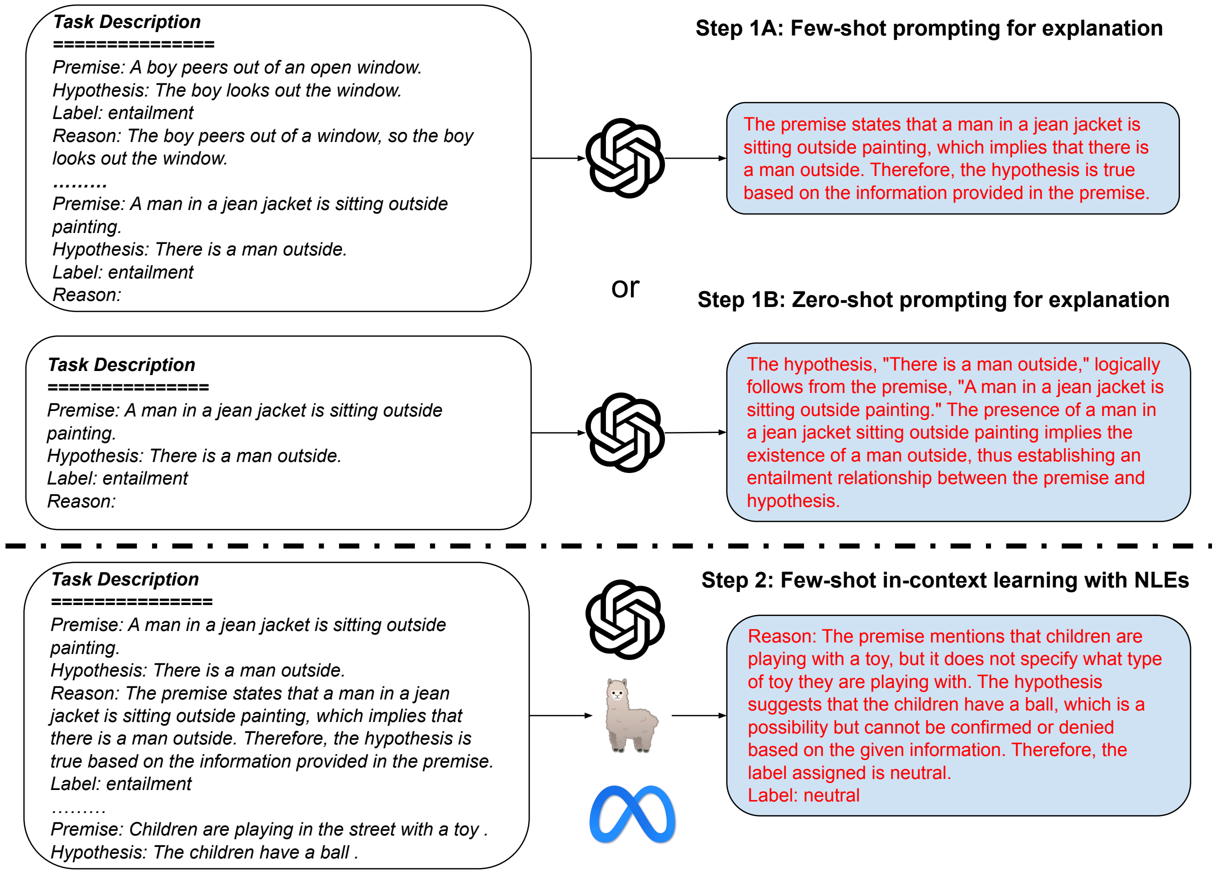


Figure 2: Illustrction of using LLM-generated NLEs for ICL: (1) prompt an LLM in a few-shot or zero-shot manner to generate NLEs for new instances; (2) prompt LLMs using ICL with the NLEs generated in step 1.

into the existing X-ICL framework as delineated in Section 3.1. We primarily focus on using GPT-3.5 (more specifically, GPT3.5-turbo-0613 – we will refer to this model as ChatGPT) to synthesize NLEs. Given that LLMs, such as ChatGPT, may have been trained on datasets incorporating NLEs, it challenges the assumption of genuine zero- or few-shot learning scenarios. To clarify terminology and avoid confusion, we redefine 'zero-shot learning' as the absence of demonstration sets, and 'few-shot ICL' as learning that utilizes a demonstration set. Thus, we denote the aforementioned two approaches as zs-X-ICL (ChatGPT) and fs-X-ICL (ChatGPT), respectively. In addition, we explore the application of two other widely used open-source LLMs for generating NLEs. Detailed results of these experiments are provided in Appendix C.

## 4 Experiments

We conduct a series of experiments to assess the performance of our proposed X-ICL framework.

### 4.1 Experimental Setup

**Tasks and datasets** We consider the Natural Language Inference (NLI) and paraphrasing identification tasks as our testbed. To ascertain the robustness of LLMs when employing the proposed approach, we evaluate it across eight adversarial datasets. For the NLI task, we include HANS, ISCS, ST, PICD, PISP, NaN, and ANLI. The first five datasets (HANS, ISCS, ST, PICD, PISP) are from Liu et al. (2020b), while NaN and ANLI are sourced from Truong et al. (2022) and Nie et al. (2020), respectively. Regarding the paraphrasing identification task, we use the PAWS-QQP (or PAWS) dataset (Zhang et al., 2019).

Additionally, the SNLI dataset (Bowman et al., 2015) and QQP (Wang et al., 2018), which are non-adversarial, are employed for a comparative purpose. The details of these datasets are provided in Appendix A.

**Language models and prompts** The evaluation of our approach is undertaken across five prominent LLMs: (1) Mistral, (2) Zephyr, (3) Vicuna, (4) Llama2, and (5) GPT3.5-turbo (version 0613).

| Models | Methods | Natural Language Inference | | | | | | | | Paraphrasing | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SNLI | HANS | ISCS | NaN | ST | PICD | PISP | ANLI | QQP | PAWS | |
| **Mistral 7B** | ICL | 59.8 ±3.4 | 54.0 ±2.2 | 51.9 ±1.4 | 55.0 ±1.3 | 44.4 ±1.7 | 58.2 ±2.6 | 23.0 ±2.6 | 39.8 ±4.6 | 69.9 ±1.7 | 68.3 ±2.7 | 50.3 |
| | X-ICL (Human) | 60.0 ±2.0 | 56.0 ±2.9 | 54.7$^{\triangledown}$ ±2.5 | 58.6$^{\triangledown}$ ±2.9 | 51.7$^{\blacktriangledown}$ ±4.0 | 56.9 ±3.3 | 35.8$^{\blacktriangledown}$ ±6.7 | 43.9$^{\blacktriangledown}$ ±1.7 | 69.9 ±0.8 | 66.4 ±1.5 | 53.5 |
| | zs-X-ICL (ChatGPT) | 56.7 ±6.3 | 51.8 ±5.1 | 47.7 ±3.5 | 55.9 ±5.0 | 44.9 ±4.8 | 56.7 ±6.6 | 25.1 ±8.9 | 28.8 ±4.4 | 67.3 ±2.3 | 64.7 ±3.1 | 46.4 |
| | fs-X-ICL (ChatGPT) | **61.8** ±3.1 | **58.2**$^{\blacktriangledown}$ ±2.5 | **57.2**$^{\blacktriangledown}$ ±2.2 | **62.4**$^{\blacktriangledown}$ ±2.6 | **55.2**$^{\blacktriangledown}$ ±1.5 | **59.2** ±2.7 | **47.6**$^{\blacktriangledown}$ ±1.8 | **46.9**$^{\blacktriangledown}$ ±2.3 | **70.3** ±1.1 | **72.5**$^{\triangledown}$ ±1.3 | **57.1** |
| **Zephyr 7B** | ICL | 67.1 ±3.4 | 71.0 ±1.8 | 63.4 ±1.2 | 65.7 ±1.8 | 60.5 ±1.0 | 64.8 ±1.5 | 48.4 ±1.4 | 47.1 ±1.6 | 76.9 ±0.4 | 57.7 ±1.1 | 59.8 |
| | X-ICL (Human) | 72.4$^{\blacktriangledown}$ ±4.3 | 64.3 ±6.7 | 58.3 ±5.5 | 62.0 ±5.3 | 57.0 ±6.3 | 60.6 ±9.7 | 52.0 ±6.7 | 49.4 ±3.0 | 75.8 ±1.7 | 61.4$^{\triangledown}$ ±2.3 | 59.3 |
| | zs-X-ICL (ChatGPT) | 67.2 ±3.9 | 72.7 ±2.6 | 60.4 ±5.3 | 64.0 ±5.2 | 61.4 ±5.7 | 64.1 ±5.4 | 50.8 ±5.2 | 40.9 ±3.8 | 74.7 ±1.8 | 59.1 ±2.4 | 58.1 |
| | fs-X-ICL (ChatGPT) | **74.2**$^{\blacktriangledown}$ ±3.6 | **77.4**$^{\blacktriangledown}$ ±2.2 | **67.0** ±1.6 | **67.7** ±2.3 | **69.3**$^{\blacktriangledown}$ ±1.5 | **70.0**$^{\blacktriangledown}$ ±2.1 | **65.6**$^{\blacktriangledown}$ ±2.5 | **52.1**$^{\triangledown}$ ±2.8 | **77.3** ±0.9 | **61.5**$^{\triangledown}$ ±1.0 | **65.5** |
| **Vicuna 30B** | ICL | 65.2 ±2.7 | 69.4 ±1.2 | 62.7 ±0.9 | 61.4 ±3.5 | 58.7 ±0.8 | **67.1** ±1.6 | 50.9 ±1.3 | 50.0 ±2.6 | **81.8** ±0.5 | 69.7 ±2.6 | 61.4 |
| | X-ICL (Human) | **67.8** ±3.2 | 62.9 ±3.7 | 60.9 ±2.2 | 64.2 ±1.2 | 57.3 ±2.0 | 63.7 ±7.2 | 55.0 ±5.8 | 48.2 ±4.7 | 77.4 ±2.8 | 63.4 ±3.5 | 59.8 |
| | zs-X-ICL (ChatGPT) | 64.2 ±5.9 | 61.4 ±7.7 | 64.9 ±2.3 | 60.2 ±4.0 | 61.7 ±3.1 | 57.9 ±8.7 | 51.8 ±8.7 | 49.7 ±3.6 | 72.1 ±3.2 | 61.8 ±4.9 | 58.8 |
| | fs-X-ICL (ChatGPT) | 65.0 ±3.1 | **74.5**$^{\triangledown}$ ±4.4 | **65.5**$^{\triangledown}$ ±1.6 | **66.3**$^{\triangledown}$ ±1.1 | **64.8**$^{\blacktriangledown}$ ±1.8 | 61.6 ±8.9 | **65.9**$^{\blacktriangledown}$ ±4.7 | **57.5**$^{\blacktriangledown}$ ±1.3 | 78.6 ±1.7 | **70.0** ±3.3 | **65.4** |
| **Llama2 70B** | ICL | 69.3 ±1.2 | 65.7 ±3.4 | **63.1** ±1.6 | 61.5 ±2.3 | 58.8 ±4.4 | 67.6 ±3.0 | 48.5 ±7.3 | 54.2 ±2.9 | **80.8** ±0.6 | 44.5 ±2.9 | 60.3 |
| | X-ICL (Human) | 73.0$^{\blacktriangledown}$ ±3.1 | 65.2 ±4.6 | 59.6 ±4.4 | 62.4 ±3.3 | 55.7 ±3.9 | 64.3 ±2.3 | 50.4 ±5.1 | 49.0 ±2.6 | 74.5 ±3.0 | 42.6 ±3.3 | 57.7 |
| | zs-X-ICL (ChatGPT) | 55.4 ±5.5 | 64.0 ±6.3 | 37.4 ±6.0 | 58.1 ±5.4 | 47.7 ±5.4 | 53.5 ±8.5 | 44.2 ±8.7 | 35.8 ±0.8 | 69.1 ±4.1 | 37.8 ±4.8 | 48.1 |
| | fs-X-ICL (ChatGPT) | **74.2**$^{\blacktriangledown}$ ±2.5 | **73.3**$^{\blacktriangledown}$ ±8.5 | 57.7 ±1.2 | **65.9**$^{\triangledown}$ ±3.2 | **63.1**$^{\triangledown}$ ±3.7 | **70.6**$^{\triangledown}$ ±6.5 | **55.8**$^{\blacktriangledown}$ ±5.9 | **59.2**$^{\blacktriangledown}$ ±1.6 | 77.6 ±0.6 | **46.5**$^{\triangledown}$ ±1.9 | **63.6** |
| **GPT3.5-turbo** | ICL | 71.9 ±1.4 | 72.4 ±0.6 | 64.4 ±0.9 | 70.0 ±0.8 | 62.1 ±1.6 | 64.0 ±3.1 | 51.2 ±0.4 | 56.1 ±2.0 | **81.5** ±0.3 | 42.9 ±2.8 | 62.4 |
| | X-ICL (Human) | **78.0**$^{\blacktriangledown}$ ±1.7 | 71.0 ±1.7 | 69.0$^{\triangledown}$ ±1.2 | 70.5 ±2.2 | 65.7$^{\triangledown}$ ±1.0 | 72.7$^{\blacktriangledown}$ ±1.3 | 59.3$^{\triangledown}$ ±1.9 | 59.8$^{\triangledown}$ ±2.3 | 76.0 ±3.9 | 53.4$^{\blacktriangledown}$ ±5.3 | 66.2 |
| | zs-X-ICL (ChatGPT) | 71.9 ±2.7 | 71.6 ±0.8 | 68.4$^{\triangledown}$ ±0.3 | 70.2 ±0.0 | 67.6$^{\triangledown}$ ±1.3 | 67.7$^{\triangledown}$ ±4.1 | 61.7$^{\blacktriangledown}$ ±1.9 | **60.4**$^{\blacktriangledown}$ ±2.0 | 80.4 ±0.8 | 51.2$^{\blacktriangledown}$ ±3.1 | 66.0 |
| | fs-X-ICL (ChatGPT) | 75.5$^{\triangledown}$ ±2.8 | **76.0**$^{\blacktriangledown}$ ±2.0 | **74.9**$^{\blacktriangledown}$ ±0.1 | **73.1**$^{\blacktriangledown}$ ±1.4 | **73.3**$^{\blacktriangledown}$ ±0.4 | **76.9**$^{\blacktriangledown}$ ±0.4 | **75.5**$^{\blacktriangledown}$ ±3.0 | 59.6$^{\triangledown}$ ±1.8 | 79.0 ±1.7 | **54.0**$^{\blacktriangledown}$ ±2.6 | **69.7** |

Table 1: Accuracy of multiple LLMs using (1) standard ICL without NLEs, (2) X-ICL with human-written NLEs: X-ICL (Human), (3) X-ICL with ChatGPT-generated NLEs in a zero-shot scenario: zs-X-ICL (ChatGPT), (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario: fs-X-ICL (ChatGPT). The best performance for each task within a model is shown in **bold**. Significance testing was assessed via an unequal variances $t$-test in comparison with ICL: $\blacktriangledown$ (resp. $\triangledown$) represents a $p$-value lower than $10^{-3}$ (resp. $10^{-1}$). The results of ANLI are the average of ANLI R1, R2, and R3.

Specifically, the Mistral and Zephyr models have `7B` parameters each. For Vicuna and Llama2, we use the `30B` and `70B` versions, respectively.

We perform all X-ICL experiments in an 8-shot setting, wherein each experiment is conducted four times independently, thereby drawing 32 unique instances from the training-associated datasets as follows. Specifically, for NLI datasets (except ANLI, which includes its own training set and NLEs), we adhere to the established methodology of using the e-SNLI dataset as the demonstration set, as suggested by Liu et al. (2020b). The e-SNLI dataset is a modified version of SNLI, where each instance is annotated with NLEs written by humans. In the case of the QQP and PAWS datasets, the QQP dataset is utilized as the demonstration set. As no NLEs are available, we contribute the corresponding NLEs (refer to Appendix E).

Regarding the generation of NLEs via few-shot learning described in section 3.2, the methodology involves selecting a random instance from each label category within the training dataset to form the demonstration set. Consequently, the demonstration set comprises three instances for the e-SNLI dataset and two for the QQP dataset.

**Baselines** In addition to the proposed method, our study investigates two baselines for comparative analysis. The first baseline uses standard ICL without NLEs. The second employs human-written NLEs within the X-ICL process, referred to as X-

ICL (Human).

## 4.2 Main Results

This section examines ICL and X-ICL across the studied datasets using Mistral, Zephyr, Vicuna, Llama2, and GPT3.5-turbo. The results are summarized in Table 1.

The results demonstrate a consistent outcome across both scenarios: with and without the application of X-ICL. As the capabilities of the models increase, there is a noticeable improvement in average accuracy. This progression is evident when comparing the least potent model, exemplified by Mistral, to the most advanced one, represented by GPT3.5-turbo.

Table 1 demonstrates that X-ICL (Human) yields better predictive accuracy than ICL across all five LLMs assessed using the SNLI dataset, with enhancements of up to 6.1%. This performance elevation is, however, limited to the Mistral and GPT-3.5-turbo models when subjected to all adversarial NLI test sets. The advantage of X-ICL (Human) relative to ICL diminishes when applied to the QQP and PAWS datasets.

For fs-X-ICL (ChatGPT), both Mistral and Zephyr demonstrate a significant performance advantage in all evaluated tasks, outperforming ICL and X-ICL (Human) by at least 5.7% and 3.6%, respectively. Despite the notable improvement on ICL when employing GPT3.5-turbo in comparison to other LLMs, fs-X-ICL (ChatGPT) offers substantially additional gains, with an increase in absolute accuracy between 11%-24% on tasks such as ISCS, ST, PICD, PISP, and PAWS. This suggests that X-ICL enhances LLM effectiveness on in-distribution test sets and increases their robustness against adversarial test sets.

Remarkably, despite the predominant preference of human evaluators for NLEs generated by GPT3.5 over those written by humans, zs-X-ICL (ChatGPT) consistently produces less accurate results than X-ICL (Human) across all models under study. The exception to this trend is GPT3.5-turbo, where a tie is observed. Furthermore, it appears counter-intuitive that zs-X-ICL (ChatGPT) is outperformed by ICL for 4 out of the 5 LLMs analyzed, especially on Llama2. We conduct a systematic analysis in section 4.4 to understand this apparent discrepancy between human preferences and LLM performance.

In light of the encompassment of diverse robustness scenarios by the seven adversarial NLI datasets, our primary focus henceforth will be the examination of these NLI datasets.

| Models | Methods | SNLI | AdvNLI | Δ |
|---|---|---|---|---|
| Zephyr | ICL | 67.1 | 57.2 | **9.9** |
| | fs-X-ICL (ChatGPT) | 74.2 | **63.7** | 10.5 |
| | COSINE | 77.0 | 55.6 | 21.4 |
| | BM25 | 70.1 | 53.7 | 16.4 |
| | SET-BSR | **79.9** | 59.7 | 20.2 |
| GPT3.5-turbo | ICL | 71.9 | 61.4 | 10.5 |
| | fs-X-ICL (ChatGPT) | 75.5 | **69.8** | **5.6** |
| | COSINE | 75.0 | 58.1 | 16.9 |
| | BM25 | 71.4 | 56.0 | 15.4 |
| | SET-BSR | **77.4** | 59.5 | 17.9 |

Table 2: Performance of ICL, fs-X-ICL (ChatGPT) and three data selection approaches on SNLI and AdvNLI (*i.e.,* seven adversarial test sets). Δ indicates the difference between SNLI and adversarial NLI test sets. We report the average performance over all adversarial test sets.

## 4.3 Impacts of NLEs

Our research has demonstrated that using NLEs generated by GPT3.5 can substantially enhance the performance of X-ICL. To provide a more comprehensive understanding of the NLEs' influence, we conducted two investigations, presented below.

**Data selection vs. X-ICL.** The effectiveness of ICL in LLMs is closely linked to the quality of demonstrations provided, as these demonstrations are critical for the model's ability to understand and address the test instances (Zhao et al., 2021; Liu et al., 2022; Lu et al., 2022). Consequently, considerable research has focused on developing data selection techniques to optimize the curation of ICL demonstrations from relevant candidate data pools, aiming to enhance their alignment with the test instances (Gupta et al., 2023; Levy et al., 2023; Ye et al., 2023). While these approaches have proven to be highly effective on in-distribution test sets, their performance on adversarial test sets remains uncertain, as these sets have the potential to misguide the selection algorithms.

In this context, we compare the performance of fs-X-ICL (ChatGPT) to three prevalent data selection techniques: COSINE, BM25, and SET-BSR. COSINE incorporates sentence embeddings (Reimers and Gurevych, 2019) to identify the most relevant demonstrations for each test instance, while BM25 employs the BM25 algorithm (Sparck Jones et al., 2000) for retrieving candidate demonstrations. SET-BSR utilizes BERTScore (Zhang

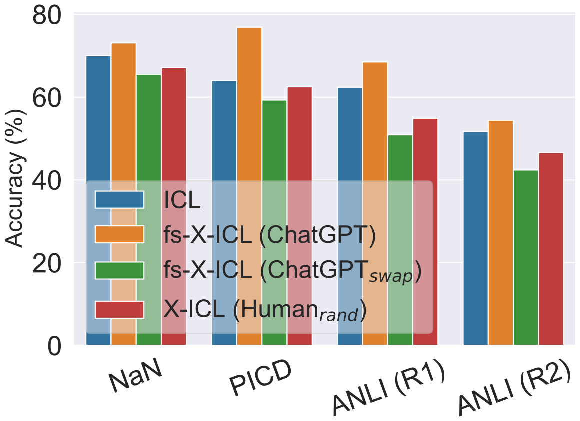


Figure 3: ICL performance of GPT3.5-turbo using (1) standard ICL without NLEs, (2) X-ICL with GPT3.5-generated NLEs in a few-shot scenario: fs-X-ICL (ChatGPT), (3) X-ICL with GPT3.5-generated NLEs, where the NLEs of the prompt are swapped and do not match the instances: fs-X-ICL (ChatGPT$_{swap}$), and (4) X-ICL with random human NLEs: X-ICL (Human$_{rand}$).

| |
|---|
| **Premise**: None of them supported her.<br>**Hypothesis**: One of them supported her.<br>**NLE [X-ICL (Human) ]**: If none of them supported her, then one of them did not support her.<br>**NLE [fs-X-ICL (ChatGPT) ]**: The hypothesis contradicts the given premise, which states that none of them supported her. |
| **Premise**: Not all people have had the opportunities you have had.<br>**Hypothesis**: Some people have not had the opportunities you have had.<br>**NLE [X-ICL (Human) ]**: If not all people have had the opportunities you have had, then some people have not had the opportunities you have had.<br>**NLE [fs-X-ICL (ChatGPT) ]**: The hypothesis is a direct result of the premise, and the label assigned is entailment. |

Table 3: Two test examples from the NAN dataset and the corresponding NLEs generated by X-ICL (Human) and fs-X-ICL (ChatGPT) using Zephyr.

et al., 2020), integrated with set theory, to ensure comprehensive information coverage and diversity within the selected instances (Gupta et al., 2023). Note that these data selection techniques are designed to sift through the entirety of the training data to choose demonstrations, a computationally demanding and computationally expensive process for generating NLEs for the full dataset. Therefore, our analysis is confined to applying ICL to these methods. To facilitate a generic comparison with the in-distribution set, we consider the average performance across all adversarial NLI test sets.

According to Table 2, as expected, the data selection approaches markedly enhance ICL performance on the SNLI dataset for all studied LLMs, with notable improvements observed in SET-BSR, achieving gains of up to 17.8% over standard ICL. However, this pronounced advantage diminishes considerably on adversarial test sets, particularly for COSINE and BM25 models, which are outperformed by ICL across all tested LLMs. This discrepancy results in a marked disparity between the in-distribution and adversarial test sets, contrary to what is observed in fs-X-ICL (ChatGPT). These results imply that current data selection approaches may be prone to overfitting on in-distribution tests, potentially leading to significant challenges in processing OOD and adversarial datasets due to their limited generalizability.

**Do proper NLEs really help?** The prevailing assumption argues that the benefits of the X-ICL primarily originate from the NLEs provided. To conclusively attribute these gains to the NLEs rather than any potential influence of additional sentences, we investigate two experimental setups. In the first setup, we randomly swap the NLEs within the prompt, leading to a mismatched NLE for each instance. This variant is henceforth referred to as fs-X-ICL (ChatGPT$_{swap}$). Regarding the second variant, for each instance in the demonstration set, we randomly select an unrelated human NLE from the corresponding training set, referred to as X-ICL (Human$_{rand}$).

As depicted in Figure 3, despite identical content being provided to GPT3.5-turbo, a misalignment between the NLE and the instance results in a marked reduction in the performance of fs-X-ICL (ChatGPT$_{swap}$) when compared to fs-X-ICL (ChatGPT). This decline is discernible across various datasets, including NaN, PICD, and ANLI (R1/R2).[4] It is also shown that an irrelevant and arbitrary NLE triggers a performance reduction within the X-ICL framework. Furthermore, the efficiency of both fs-X-ICL (ChatGPT$_{swap}$) and X-ICL (Human$_{rand}$) substantially lags behind that of ICL. Therefore, it can be inferred that the efficacy of the fs-X-ICL (ChatGPT) hinges on providing an accurate and relevant NLE.

[4]Similar patterns have been detected in other datasets

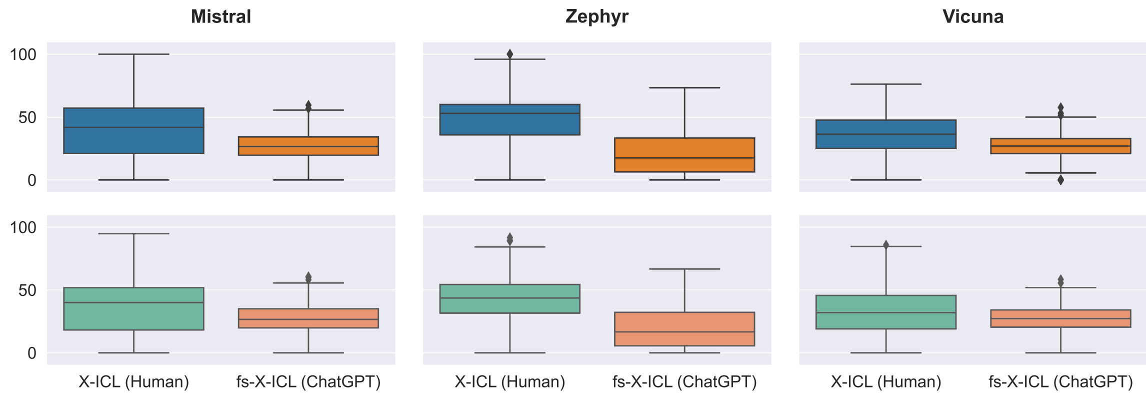


Figure 4: ROUGE-L between the NAN test set and the corresponding generated NLEs. **Top**: ROUGE-L between test premise and NLE. **Bottom**: ROUGE-L between test hypothesis and NLE.

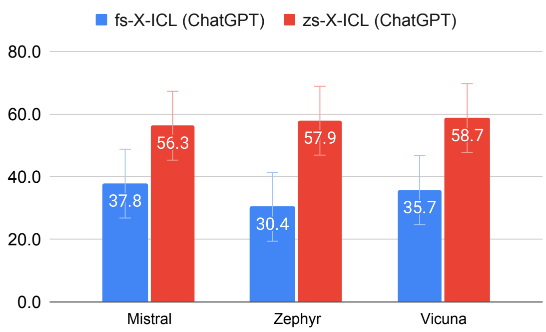


Figure 5: Average length (#words) of NLEs generated by fs-X-ICL (ChatGPT) and zs-X-ICL (ChatGPT).

| Methods | Mistral | Zephyr | Vicuna |
|---|---|---|---|
| X-ICL (Human) | 53.5 | 59.3 | 59.8 |
| zs-X-ICL (ChatGPT) | 46.4 | 58.1 | 58.8 |
| zs-X-ICL (ChatGPT$_S$) | 56.2 | 62.3 | 63.4 |
| fs-X-ICL (ChatGPT) | 57.1 | 65.5 | 62.1 |

Table 4: Average accuracy of X-ICL (Human), zs-X-ICL (ChatGPT), zs-X-ICL (ChatGPT$_S$) and fs-X-ICL (ChatGPT) among all test sets.

### 4.4 Further Analysis

**Why is fs-X-ICL (ChatGPT) producing the most accurate results?** Our study demonstrates that fs-X-ICL (ChatGPT) surpasses both X-ICL (Human) and zs-X-ICL (ChatGPT) in accuracy. However, the reasons behind this superior performance are not yet understood. Therefore, this section focuses on systematically analyzing the efficacy of fs-X-ICL (ChatGPT).

We first dissect the effectiveness of fs-X-ICL (ChatGPT) over X-ICL (Human). As shown in Table 3, NLEs from X-ICL (Human) are mere verbatim copies of inputs rather than insightful explanations. To substantiate this, we calculate the ROUGE-L scores between the NAN test set and the corresponding NLEs from X-ICL (Human) and fs-X-ICL (ChatGPT) as a means of similarity measurement. As depicted in Figure 4, NLEs from X-ICL (Human) often replicate the given premise and hypothesis, resulting in high ROUGE-L scores. Instead, fs-X-ICL (ChatGPT) can produce meaningful NLEs, demonstrating lower similarity to the test instances.

After analyzing the NLEs from zs-X-ICL (ChatGPT), we attribute the inefficiency to verbose NLEs. Specifically, Figure 5 shows that zs-X-ICL (ChatGPT) produces longer NLEs than fs-X-ICL (ChatGPT). As a result, we observe inconsistency within the NLEs, leading to incorrect predictions. As a remedy, we prompt ChatGPT to generate shorter NLEs in the zero-shot setting, denoted as zs-X-ICL (ChatGPT$_S$). Compared to zs-X-ICL (ChatGPT), the NLEs generated by zs-X-ICL (ChatGPT$_S$) are reduced to an average of 27 tokens. Consequently, with the help of the concise NLEs, we can improve the accuracy significantly and even surpass the X-ICL (Human) as shown in Table 4.

## 5 Summary and Outlook

We introduced a simple yet effective method called fs-X-ICL (ChatGPT), leveraging human-written NLEs to generate synthetic NLEs by prompting ChatGPT. fs-X-ICL (ChatGPT) significantly boosts accuracy across various adversarial datasets and five LLMs, compared to standard in-context learning and X-ICL using human-written NLEs. Additionally, our analysis revealed that data selection methodologies may exhibit overfitting within the in-distribution dataset, thus potentially failing to extend to unseen or adversarial datasets. In contrast, our approach employing NLEs has shown consistent performance in both in-distribution and adversarial contexts. Our work paves the way for more robust performance and enhanced explainability capabilities of LLMs.

## Limitations

One limitation of X-ICL might be the observed lack of fidelity in the NLEs generated by LLMs, despite their capability to provide accurate answers. These NLEs may sometimes include unfaithful or hallucinated information, which if relied upon by users for model trust, can lead to severe implications. Testing and enhancing the faithfulness of NLEs is a challenging open question (Atanasova et al., 2023). In this work, we show that X-ICL improves robustness, but we do not advocate using the generated NLEs as faithful explanations without further testing. Second, our approach exhibited promising results when tested against adversarial datasets in two notable NLP tasks: natural language inference and paraphrasing identification. However, further research is required to examine the performance of LLMs and their generalizability across diverse NLP tasks in the context of adversarial examples.

## Acknowledgements

Xuanli He was supported by an industry grant from Cisco. Oana-Maria Camburu was supported by a Leverhulme Early Career Fellowship. Pasquale Minervini was partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 875160, ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence) EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP; and is grateful to NVIDIA for the GPU donations. This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

## References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness Tests for Natural Language Explanations. In *ACL*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642. The 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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng

Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2023. Coverage-based example selection for in-context learning. *arXiv preprint arXiv:2305.14907*.

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.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. 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.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1244–1254.

Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. 2022. Explaining chest x-ray pathologies in natural language. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 701–713, Cham. Springer Nature Switzerland.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. *CoRR*, abs/1807.11546.

Michalis Korakakis and Andreas Vlachos. 2023. Improving the robustness of NLI models with minimax training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14339, Toronto, Canada. Association for Computational Linguistics.

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.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023. Explanation-based fine-tuning 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.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian Mcauley. 2022. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14786–14801. PMLR.

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.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium. 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.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *ACL*, pages 4885–4901. Association for Computational Linguistics.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Did chatgpt cheat on your test?

William A. Scott. 1962. Cognitive complexity and cognitive flexibility. *Sociometry*, 25(4):405–414.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing and Management*, 36(6):779–808.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023a. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

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*, ICML'23. JMLR.org.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

## A Details of Datasets

The details of all studied datasets are delineated as follows

- **SNLI Dataset**: The SNLI dataset, a benchmark in natural language inference, encompasses approximately 570,000 human-annotated sentence pairs, each pair formed by a premise and a hypothesis. These sentences originate from an existing corpus of image captions, thus offering a broad spectrum of common subjects and linguistic structures (Bowman et al., 2015).
- **HANS Dataset**: McCoy et al. (2019) developed a dataset with the express purpose of scrutinizing the performance of models when confronted with sentences characterized by several types of distracting signals. These signals encompass the presence of lexical overlap, sub-sequences, and constituent heuristics between the corresponding hypotheses and premises.
- **Datasets Sensitive to Compositionality (ISCS)**: As proposed by Nie et al. (2019), a softmax regression model was employed to utilize lexical features present in the premise and hypothesis sentences, thereby generating instances of misclassification. Here, the *Lexically Misleading Score* (LMS) denotes the predicted probability of the misclassified label. Adapting the approach of Liu et al. (2020b), we concentrated on the subsets possessing LMS values exceeding 0.7.
- **Not another Negation (NaN) NLI Dataset**: NaN dataset is developed to probe the capabilities of NLP models in comprehending sub-clausal negation (Truong et al., 2022).
- **Stress Test Datasets (ST)**: Our analysis also incorporates various stress tests described by Naik et al. (2018) such as "word overlap" (ST-WO), "negation" (ST-NE), "length mismatch" (ST-LM), and "spelling errors" (ST-SE). Specifically, ST-WO aims to identify lexical overlap heuristics between the premise and hypothesis, ST-NE seeks to detect intense negative lexical cues in partial-input sentences, ST-LM aspires to create misleading predictions by artificially lengthening the premise using nonsensical phrases, and ST-SE employs spelling errors as a means to deceive the model.
- **Datasets Detected by Classifier (PICD)**: In the approach proposed by Gururangan et al. (2018), fastText was applied to hypothesis-only inputs. Subsequent instances from the SNLI test sets (Bowman et al., 2015) that could not be accurately classified were designated as 'hard' instances.
- **Surface Pattern Datasets (PISP)**: Liu et al. (2020a) identified surface patterns that exhibit strong correlation with specific labels, thereby proposing adversarial test sets counteracting the implications of surface patterns. As suggested by Liu et al. (2020b), we employed their 'hard' instances extracted from the MultiNLI mismatched development set (Williams et al., 2018) as adversarial datasets.
- **Adversarial NLI (ANLI)**: ANLI dataset (Nie et al., 2020) is a challenging resource created for training and testing models on NLI, featuring adversarial examples intentionally curated to obfuscate or mislead benchmark models, thereby increasing its challenge factor. This dataset is constructed in multiple rounds, with each subsequent round featuring human-created examples specifically designed to outsmart models trained on the previous rounds. In total, the dataset comprises three distinct rounds, specifically ANLI R1, ANLI R2, and ANLI R3, highlighting the layered complexity of this resource.
- **Quora Question Pairs (QQP)**: QQP dataset (Wang et al., 2018) comprises pairs of questions sourced from the Quora community question-answering platform. The primary objective is to ascertain whether each question pair exhibits semantic equivalence.
- **Paraphrase Adversaries from Word Scrambling (PAWS)**: The PAWS-QQP dataset (Zhang et al., 2019), derived from the QQP datasets, targets the intricate task of paraphrasing identification, emphasizing the differentiation of sentences that, despite high lexical similarity, convey distinct meanings. It incorporates adversarial examples generated via word scrambling, presenting a stringent assessment for NLP models.

## B Meta-prompts for Generating Synthetic NLEs

Table 5 and 6 present the meta-prompts and demonstration instances employed for producing NLEs utilizing ChatGPT in zero- and few-shot scenarios.

| Meta-prompt for zero-shot generation |
|---|
| 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: |
| **Meta-prompt and demonstration instances for few-shot generation** |
| 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:<br><br>**Premise**: A boy peers out of an open window.<br>**Hypothesis**: The boy looks out the window.<br>**Label**: entailment<br>**NLE**: The boy peers out of a window, so the boy looks out the window.<br>=====<br>**Premise**: A kid doing a trick on a skateboard.<br>**Hypothesis**: The kid eating lunch inside the cafeteria.<br>**Label**: contradiction<br>**NLE**: The kid cannot be doing a trick and eating lunch at the same time<br>=====<br>**Premise**: A man jumps off of his skateboard on the top of a cement ramp.<br>**Hypothesis**: a man jumps off a skateboard at the top of a ramp.<br>**Label**: neutral<br>**NLE**: A man can jump off a skateboard without being at the top of a ramp. |

Table 5: Meta-prompts used to generate NLEs via ChatGPT in zero- and few-shot scenarios for natural language inference tasks.

| Meta-prompt for zero-shot generation |
|---|
| Assume that you're an expert working on paraphrasing identification tasks. Given two sentences and the corresponding label. Please write a concise and precise reason to explain why the label is assigned to the example: |
| **Meta-prompt and demonstration instances for few-shot generation** |
| Assume that you're an expert working on paraphrasing identification tasks. Given two sentences 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:<br><br>**Q1**: Does life get harder as you get older?<br>**Q2**: Does life really get harder as you get older?<br>**Label**: duplicate<br>**NLE**: Both questions ask whether life does get harder as you get older.<br>=====<br>**Q1**: What is the National nanotechnology initiative?<br>**Q2**: What is the lead time for SSN4EGS411 board?<br>**Label**: not duplicate<br>**NLE**: completely different questions |

Table 6: Meta-prompts used to generate NLEs via ChatGPT in zero- and few-shot scenarios for paraphrasing identification tasks.

## C Supplementary Studies

**Using NLEs Generated by Vicuna and Llama2.** Our research demonstrates that the integration of NLEs generated by ChatGPT significantly enhances the performance of X-ICL for five advanced LLMs. To assess the efficacy of these ChatGPT-generated NLEs, we explore the generation of synthetic NLEs using Vicuna and Llama2, ranked as the third and second-best models respectively. Likewise, these NLEs are generated in a few-shot setting, referred to herein as $\text{Vicuna}_{\text{few}}$ and $\text{Llama2}_{\text{few}}$, respectively. To ensure a fair comparison, we employ Vicuna as the underlying model to evaluate fs-X-ICL(Vicuna), fs-X-ICL (Llama2), and fs-X-ICL (ChatGPT) on all studied datasets.

Our results, detailed in Table 7, highlight that X-ICL generally gains greater benefit from LLM-generated NLEs as opposed to those produced by humans. Meanwhile, fs-X-ICL (ChatGPT) consistently outperforms fs-X-ICL(Vicuna) and fs-X-ICL (Llama2) considerably, except for ANLI R1 and R2. These findings suggest that to harness the potential of AI-generated NLEs fully, the employment of a powerful LLM is integral.

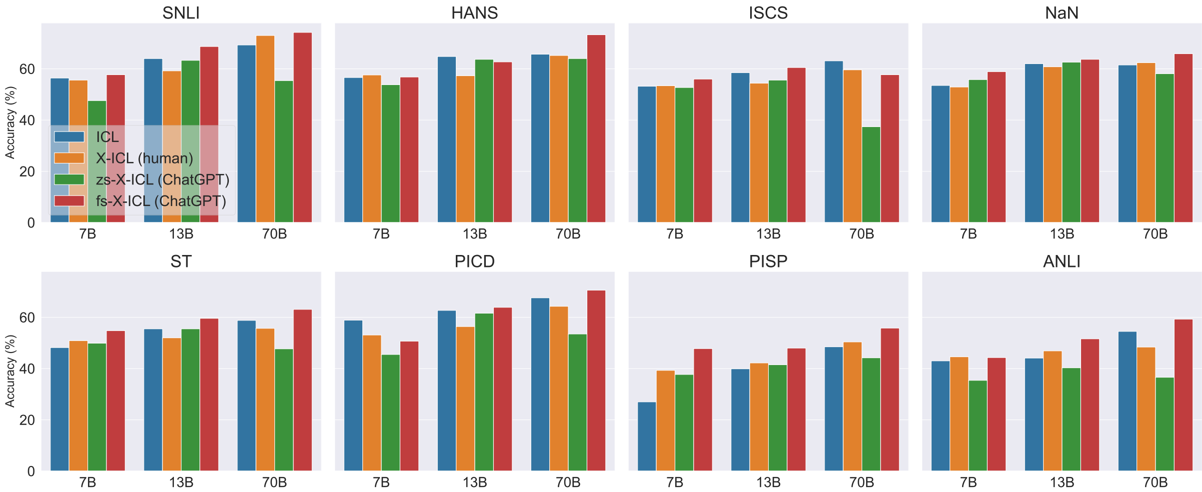


Figure 6: ICL performance of Llama2 (7B, 13B, 70B) using (1) standard ICL without NLEs, (2) X-ICL with human-written NLEs: X-ICL (Human), (3) X-ICL with ChatGPT-generated NLEs in a zero-shot scenario: zs-X-ICL (ChatGPT), (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario:fs-X-ICL (ChatGPT). ANLI is the average of R1, R2 and R3.

| Tasks | NLEs | | |
|---|---|---|---|
| | **fs-Vicuna** | **fs-Llama2** | **fs-ChatGPT** |
| SNLI | 62.9 ( −5.0) | 64.1 ( −3.7) | 65.0 ( −2.9) |
| HANS | 55.5 ( −7.4) | 67.4 ( +4.5) | 74.5 (+11.6) |
| ISCS | 65.1 ( +4.2) | 63.6 ( +2.7) | 65.5 ( +4.6) |
| NaN | 62.6 ( −1.6) | 65.1 ( +0.9) | 66.3 ( +2.1) |
| ST | 59.5 ( +2.2) | 61.9 ( +4.6) | 64.8 ( +7.5) |
| PICD | 60.2 ( −3.5) | 60.8 ( −2.9) | 61.6 ( −2.1) |
| PISP | 66.0 (+11.0) | 66.1 (+11.1) | 66.0 (+11.0) |
| ANLI (R1) | 66.1 ( +9.1) | 65.8 ( +8.8) | 64.9 ( +7.9) |
| ANLI (R2) | 55.4 ( +6.5) | 55.9 ( +7.0) | 55.5 ( +6.6) |
| ANLI (R3) | 49.6 (+10.8) | 50.7 (+11.9) | 52.0 (+13.2) |
| Average | 60.3 ( +3.8) | 62.1 ( +5.6) | **63.5** ( +6.9) |

Table 7: ICL performance of Vicuna using (1) standard ICL without NLEs, (2) X-ICL with Vicuna-generated NLEs in a few-shot scenario: fs-Vicuna, (3) X-ICL with Llama2-generated NLEs in a few-shot scenario: fs-Llama2, (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario: fs-ChatGPT. Numbers in the parentheses represent differences compared to X-ICL (Human).

**Does model size matter?** We have shown the efficacy of X-ICL across a range of LLMs of varying sizes. However, the variability in data and training processes among these models renders the applicability of our approach to smaller-scale models inconclusive, especially since the smaller models often exhibit less benefit from NLEs compared to larger models within the same family (Wei et al., 2022a). Therefore, we have evaluated our approach using three distinct sizes of Llama2 models: 7B, 13B, and 70B parameters.

Referring to Figure 6, one can find the performance of both ICL and X-ICL generally improves in correspondence with the escalation of model size, except for zs-X-ICL (ChatGPT). Moreover, the gap in performance between ICL and fs-X-ICL (ChatGPT) widens, indicating that models with greater capabilities derive increased benefits from NLEs. This observation aligns with the results reported by Wei et al. (2022a).

**Distribution Shift Prompting.** Previous works indicate that X-ICL can potentially encourage LLMs to engage in deliberate thinking, a predominant factor responsible for substantial performance improvements over the standard ICL in complex reasoning tasks (Wei et al., 2022b). In addition, our findings have demonstrated a dramatic enhancement in the robustness of LLMs due to X-ICL, which contributes to significant improvements in ICL when applied to various adversarial datasets.

Moreover, a previous study established that upon understanding the concept underlying particular tasks, humans can address similar tasks despite a distribution shift (Scott, 1962). To explore the robustness of ICL and X-ICL against distribution shifts, we employ the e-SNLI dataset as the demonstration set for ANLI (R1/R2), while utilizing the ANLI training set for testing NaN and PICD. Due to its outstanding performance, we use GPT3.5-turbo as the backbone model.

As suggested in Table 8, for NaN and PICD, using e-SNLI as the prompt proves to be more effective than ANLI for both ICL and fs-X-ICL (ChatGPT). This improvement can be attributed to the

| | NaN | | | PICD | | | ANLI (R1) | | | ANLI (R2) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | e-SNLI | ANLI | $\|\Delta\|$ | e-SNLI | ANLI | $\|\Delta\|$ | e-SNLI | ANLI | $\|\Delta\|$ | e-SNLI | ANLI | $\|\Delta\|$ |
| ICL | 70.0 | 69.4 | 0.6 | 64.0 | 64.1 | 0.1 | 52.6 | 62.4 | 9.7 | 43.9 | 51.7 | 7.8 |
| fs-X-ICL (ChatGPT) | 73.1 | 71.8 | 1.2 | 76.9 | 76.1 | 0.8 | 65.0 | 68.5 | 3.5 | 53.2 | 54.4 | 1.2 |

Table 8: Performance of ICL and fs-X-ICL (ChatGPT) employing e-SNLI and ANLI as prompts for testing NaN, PICD, and ANLI (R1/R2). $|\Delta|$ signifies the absolute difference in the performance outcomes when utilizing e-SNLI in contrast to ANLI. The backbone model is GPT3.5-turbo.

distribution shift. Likewise, the distribution shift results in a noticeable distinction between e-SNLI and ANLI for ICL on ANLI (R1/R2). Nonetheless, incorporating NLEs enables fs-X-ICL (ChatGPT) to substantially reduce this gap, from 9.7 to 3.5 for ANLI (R1), and from 7.8 to 1.2 for ANLI (R2). This finding indicates that X-ICL may improve the robustness of LLMs in the face of distribution shifts.

**Analysis on memorization** LLMs such as ChatGPT have occasionally replicated instances from renowned benchmark datasets, including MNLI and BoolQ (Sainz et al., 2023). This unintentional *'contamination'* might contribute to misconceptions regarding the superior performance of LLMs on these widespread benchmarks due to data memorization.

Following Carlini et al. (2023), we merge the premise and hypothesis of each test instance into a single sentence, using the first part as the prefix. If an LLM could perfectly replicate the second part, we labeled the instance as *'extractable'*. Evaluating all studied models, we observe that the proportion of extractable instances is under 0.001% across all datasets and backbone models, indicating that the superior performance of LLMs might not be ascribed to memorization.

# D Qualitative Analysis on NLEs

## D.1 Qualitative Analysis on NLEs for Demonstration Set

We first conducted a qualitative analysis of NLEs generated by ChatGPT under zero- and few-shot scenarios, using the demonstration set as a basis. Note that each instance in the demonstration set has three distinct NLEs: (1) the zero-shot NLE from ChatGPT, (2) the few-shot NLE from ChatGPT, and (3) the human-written NLE. From these three NLEs per instance, one was randomly selected, and both the instance and the chosen NLE were incorporated into the evaluation set.

Subsequently, this evaluation set was rated independently by four authors on a 5-point Likert scale to assess the quality of the NLEs. The scale ranges were 1 (extremely dissatisfied), 2 (dissatisfied), 3 (neutral), 4 (satisfied), and 5 (extremely satisfied). Finally, we calculated the average scores for both ChatGPT-generated and human-written NLEs for each evaluator.

## D.2 Qualitative Analysis on NLEs for Inference Set

We also conducted a qualitative analysis of NLEs generated by fs-X-ICL (ChatGPT), utilizing GPT3.5-turbo as the foundational model. A total of 280 randomly sampled, correctly predicted examples from fs-X-ICL (ChatGPT) were distributed evenly among seven evaluators. These evaluators were tasked to assess the quality of the NLE for each assigned instance, based on the premise-hypothesis pair and its corresponding correctly predicted label.

The evaluators were required to rate the quality of the NLE using the aforementioned 5-point Likert scale. In case of dissatisfaction, they were asked to identify the reason from a list of predefined factors, including:

- **template**: The NLE simply restates the input and employs it as a justification.
- **insufficient justification**: The NLE requires more support for the prediction.
- **too verbose**: The NLE is overly detailed and includes unnecessary information.
- **incorrect arguments**: Despite the prediction being accurate, the NLE fails to support it due to erroneous arguments.
- **contradict commonsense**: The NLE is incorrect and contradicts commonsense.
- **hallucinations**: The NLE includes fabricated information.

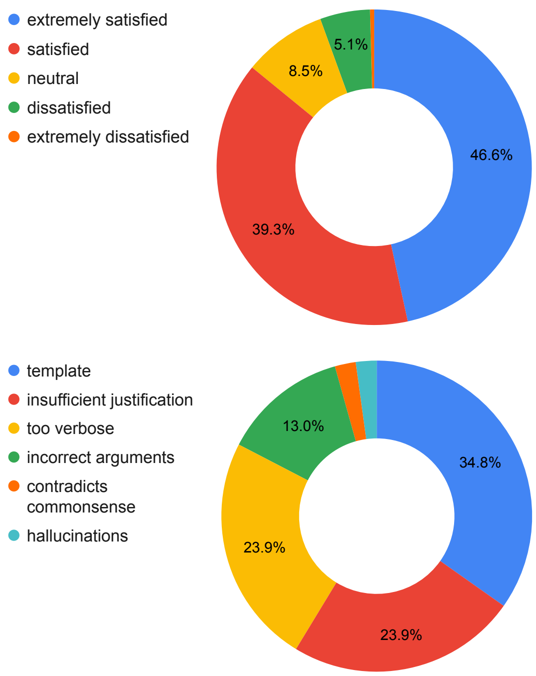


Figure 7: Human evaluation on ChatGPT-generated NLEs for the correct predictions from fs-X-ICL (ChatGPT). **Top**: distribution of satisfaction scores. **Bottom**: distribution of reasons for dissatisfaction.

According to Figure 7, 46.6% and 39.3% of NLEs are marked as 'extremely satisfied' and 'satisfied' respectively, constituting 85.9% of the total 280 NLE samples. This suggests a high-quality output from GPT3.5-turbo in general. As for the lower-quality NLEs, the primary reasons for dissatisfaction include 'template', 'insufficient justification', and 'too verbose'. Interestingly, this suggests that, despite the expressed dissatisfaction, evaluators generally did not find incorrect justifications in most instances.

## E Human-written NLEs for QQP

Given the absence of NLEs for the QQP dataset, we have conducted a randomized sampling of 32 instances from the QQP training set. For each selected instance, we crafted a corresponding NLE. The details of these QQP instances and their respective NLEs are presented in Table 11, 12, 13, and 14.

| |
|---|
| **Premise**: The organizing committee said 65 countries have entered the Lillehammer Olympic Games, matching the number of nations at the 1992 Winter Games in Albertville, France.<br>**Hypothesis**: The Lillehammer Olympic games take place 12 years after the Albertville Olympic games.<br>**NLE**: The premise does not provide any information about the time difference between the Lillehammer Olympic Games and the Albertville Olympic Games. Therefore, there is no evidence to support or contradict the hypothesis, resulting in a neutral label.<br>**Predicted Label**: neutral<br>**Correct Label**: neutral |
| **Premise**: A woman was not injured when she was carjacked in west Charlotte Tuesday morning. According to Charlotte-Mecklenburg police, two men reportedly pulled a gun on a woman who was stopped at a stop sign on State Street at Whitehaven Avenue around 3:30 a.m. Police say the men then stole the woman's Honda Fit. Officers have not found the stolen vehicle. No other details were released. Copyright 2018 WBTV. All rights reserved.<br>**Hypothesis**: A woman stole a Honda Fit in west Charlotte at 3:30 AM but was not injured.<br>**NLE**: The premise states that a woman was carjacked in west Charlotte at 3:30 AM and that two men stole her Honda Fit. The hypothesis suggests that the woman stole a Honda Fit, which contradicts the information provided in the premise. Therefore, the hypothesis is false and leads to a contradiction.<br>**Predicted Label**: contradiction<br>**Correct Label**: contradiction |
| **Premise**: New question. Last question. For you, Governor. And this flows somewhat out of the Boston debate. You, your running mate, your campaign officials have charged that Vice President Gore exaggerates, embellishes and stretches the facts, etcetera. Are you – do you believe these are serious issues? This is a serious issue that the voters should use in deciding which one of you two men to vote for on November 7?<br>**Hypothesis**: the Vice President exaggerates the truth.<br>**NLE**: The premise states that the question is about whether the Governor believes that the Vice President exaggerates, embellishes, and stretches the facts. Therefore, the hypothesis is a restatement of the question and can be confirmed as true based on the information provided in the premise.<br>**Predicted Label**: entailment<br>**Correct Label**: entailment |
| **Premise**: Budapest consists of two parts, Buda and Pest, which are situated on opposite sides of the river and connected by a series of bridges.<br>**Hypothesis**: Buda and Pest function as separate cities.<br>**NLE**: The premise states that Budapest consists of two parts, Buda and Pest, which are connected by a series of bridges. However, it does not suggest that Buda and Pest function as separate cities. Therefore, the hypothesis cannot be confirmed or refuted based on the given information, resulting in a neutral label.<br>**Predicted Label**: neutral<br>**Correct Label**: neutral |

Table 9: Test instances and the corresponding generated NLEs from GPT-3.5-turbo for the correct predictions.

| |
|---|
| **Premise**: Mulberry () is a 1986 South Korean drama film directed by Lee Doo-yong. Based on the eponymous classic story by Na Do-hyang, the film became known for its erotic subject matter, made possible by the government's gradual relaxation of censorship and control over the film industry during the 1980s. It was part of the wave of "Folk erotic" films that had become popular in South Korea at this time.<br>**Hypothesis**: Mulberry () is a 1986 South Korean drama film directed by Lee Doo-dong.<br>**NLE**: The hypothesis is identical to the premise, except for a minor spelling error in the director's name. Therefore, the hypothesis is true based on the provided premise.<br>**Predicted Label**: entailment<br>**Correct Label**: contradiction |
| **Premise**: In the Company of Men is a 1997 Canadian/American black comedy written and directed by Neil LaBute and starring Aaron Eckhart, Matt Malloy, and Stacy Edwards. The film, which was adapted from a play written by LaBute, and served as his feature film debut, won him the Independent Spirit Award for Best First Screenplay.<br>**Hypothesis**: Neil LaBute worked with Aaron Eckhart for the comedy that was his first screenplay.<br>**NLE**: The premise states that Neil LaBute wrote and directed In the Company of Men, which was his feature film debut and won him the Independent Spirit Award for Best First Screenplay. It also mentions that the film starred Aaron Eckhart, but it does not provide any information about whether or not LaBute worked with Eckhart specifically for this film. Therefore, the hypothesis cannot be either true or false based on the provided premise.<br>**Predicted Label**: neutral<br>**Correct Label**: entailment |
| **Premise**: Barry Cunningham (born 19 December 1952) is a British publisher, who worked for various publishers including Penguin Books and Bloomsbury before setting up Chicken House publishing in 2000. He is most well known for signing J.K Rowling, and publishing "Harry Potter and the Philosopher's Stone" in 1997.<br>**Hypothesis**: Barry Cunningham is most proud of signing J.K Rowling, and publishing "Harry Potter and the Philosopher's Stone" in 1997.<br>**NLE**: The premise states that Neil LaBute wrote and directed In the Company of Men, which was his feature film debut and won him the Independent Spirit Award for Best First Screenplay. It also mentions that the film starred Aaron Eckhart, but it does not provide any information about whether or not LaBute worked with Eckhart specifically for this film. Therefore, the hypothesis cannot be either true or false based on the provided premise.<br>**Predicted Label**: entailment<br>**Correct Label**: neutral |
| **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".<br>**Hypothesis**: Gyula Trebitsch was nominated for the Academy Award for Best Foreign Language Film for his work on "The Captain of Kopenick" at the age of 43.<br>**NLE**: The premise states that Gyula Trebitsch was nominated for the Academy Award for Best Foreign Language Film for "The Captain of Kopenick" in 1956, but it does not provide information about his age at the time of the nomination. Therefore, the hypothesis cannot be either true or false based on the provided premise.<br>**Predicted Label**: neutral<br>**Correct Label**: contradiction |

Table 10: Test instances and the corresponding generated NLEs from GPT-3.5-turbo for the incorrect predictions.

| |
|---|
| **Q1**: Is 6 foot 4 too tall as an ideal height for a man?<br>**Q2**: My height is 5'6 and I'm 14 year old boy, my mom is 5'4 and my dad is 5'7. How tall will I be?<br>**Label**: not duplicate<br>**NLE**: Predicting future height given parents' heights concerns genetic factors of height, whereas ideal height for man concerns more about its social aspect. |
| **Q1**: Approximately how many hours have you spent on the internet till date?<br>**Q2**: What amount of time do you spent on the Internet?<br>**Label**: not duplicate<br>**NLE**: Total number of hours spend on Internet till date not just depend on the average hours on internet per day, but also many other factors such as the age the user started using it. |
| **Q1**: What are the most ridiculous statements made by Donald Trump?<br>**Q2**: My black friend supports Donald Trump, is that ridiculous?<br>**Label**: not duplicate<br>**NLE**: Asking the most ridiculous statement made by Donald Trump is different than asking why a supporter support him. A supporter can support him for other reasons. |
| **Q1**: "What is the origin of the phrase ""pipe dream""?"<br>**Q2**: "How did the phrase ""toe head"" originate?"<br>**Label**: not duplicate<br>**NLE**: The two questions asked about the origin of two different words. |
| **Q1**: What is a good first programming language to learn?<br>**Q2**: What is the most valuable programming language for the future to learn?<br>**Label**: duplicate<br>**NLE**: When picking a good first programming language to learn, people may consider the most valuable one language if they learn it for making money. |
| **Q1**: What is best way for earning money?<br>**Q2**: How can I start making money? What are the best ways to make money?<br>**Label**: duplicate<br>**NLE**: Both questions ask about what are best ways to make money |
| **Q1**: Does the Indian education system need a reformation?<br>**Q2**: Should the education system be changed in India? If so why or why not?<br>**Label**: duplicate<br>**NLE**: Both questions essentially inquire about the necessity and justification for changing the Indian education system. |
| **Q1**: What is the application of quantum physics?<br>**Q2**: What are some applications of quantum physics?<br>**Label**: duplicate<br>**NLE**: The two questions both seek information about the practical use of quantum physics. |

Table 11: QQP instances and the corresponding NLEs.

| |
|---|
| **Q1**: How is the word 'calumny' used in a sentence?<br>**Q2**: How is the word 'mischievous' used in a sentence?<br>**Label**: not duplicate<br>**NLE**: The two questions ask about two different words with different meanings. |
| **Q1**: What are your views on the abolishment of 500 rupees note?<br>**Q2**: How will the ban of Rs 500 and Rs 1000 notes affect Indian economy?<br>**Label**: not duplicate<br>**NLE**: The former question asks specifically about the abolishment of the Rs 500 note, while the latter asks about the Rs 500 and the Rs 1000 notes. |
| **Q1**: What are the valence electrons of titanium?<br>**Q2**: What is the number of valence electrons in hydrogen? How is this determined?<br>**Label**: not duplicate<br>**NLE**: The former question asks about titanium, while the latter is about hydrogen. |
| **Q1**: Do movie actors get paid each time their movie is played on TV?<br>**Q2**: Why are film actors so highly paid whereas scientists are paid relatively quite little?<br>**Label**: not duplicate<br>**NLE**: The former question asks some details about how actors get paid, while the latter asks about the gap between actor and scientist salaries. |
| **Q1**: How do I build an electromagnetic propulsion engine?<br>**Q2**: How would I build a magnetic propulsion system?<br>**Label**: duplicate<br>**NLE**: Both question asks about building magnetic propulsion systems. |
| **Q1**: Why is salt water taffy candy imported in France?<br>**Q2**: Why is Saltwater taffy candy imported in The Bahamas?<br>**Label**: duplicate<br>**NLE**: Both questions ask about the reasons behind importing salt water taffy candy. |
| **Q1**: Why do we call Java platform independent language when it still requires platform dependent JVM to get executed?<br>**Q2**: How is the Java platform independent when we need to have JVM on every machine to run Java programs?<br>**Label**: duplicate<br>**NLE**: Both questions ask why do we call Java platform-independent, since it still depends on the availability of a JVM. |
| **Q1**: What are the various ways through which one can earn money online?<br>**Q2**: How do you make easy money online?<br>**Label**: duplicate<br>**NLE**: Both questions ask how to make money online. |

Table 12: QQP instances and the corresponding NLEs.

| |
|---|
| **Q1**: Why can’t some people think for themselves?<br>**Q2**: Why don’t people think for themselves?<br>**Label**: not duplicate<br>**NLE**: "some people" means not all people as the second question seems to imply |
| **Q1**: Why don’t we use Solar Furnace to produce electricity?<br>**Q2**: Why don’t we make Solar Cars?<br>**Label**: not duplicate<br>**NLE**: using Solar Furnace you can produce some amount of electricity but it may not enough to power a whole car |
| **Q1**: What is an intuitive explanation of the fractional quantum Hall effect?<br>**Q2**: What is an intuitive explanation of the Quantum Hall effect?<br>**Label**: not duplicate<br>**NLE**: fractional quantum Hall effect is different than the Quantum Hall effect, which refers to the integer quantum Hall effect |
| **Q1**: Can INTPs become successful entrepreneurs?<br>**Q2**: I am business associate in tcs?<br>**Label**: not duplicate<br>**NLE**: completely different questions |
| **Q1**: How can I be like Sheldon Cooper?<br>**Q2**: How do I become like Sheldon Cooper?<br>**Label**: duplicate<br>**NLE**: "be like" and "become like" someone is the same thing |
| **Q1**: What do people think about Anonymous?<br>**Q2**: What do you think about the ’Anonymous’ option on Quora?<br>**Label**: duplicate<br>**NLE**: "what do people think" and "what do you think" are usually used interchangeably |
| **Q1**: What’s the meaning of life?<br>**Q2**: "What is the meaning of ""Life""?"<br>**Label**: duplicate<br>**NLE**: same question with minor different spellings |
| **Q1**: What is it in for the Ibibo group employees with the Makemytrip merger / Buyout?<br>**Q2**: How do Ibibo employees feel about MakeMyTrip acquiring Ibibo?<br>**Label**: duplicate<br>**NLE**: "the Makemytrip merger / Buyout" refers to "MakeMyTrip acquiring Ibibo" and "what is it in for the employees" means "how do the employees feel about" |

Table 13: QQP instances and the corresponding NLEs.

| |
|---|
| **Q1**: Why is Lionel Messi so brilliant?<br>**Q2**: Is Lionel Messi a genius?<br>**Label**: not duplicate<br>**NLE**: the first question asks for the reason, while the second question inquires about yes or no |
| **Q1**: What are some of the best CyanogenMod 12.1 themes?<br>**Q2**: How do I make my own cyanogen 12.1 themes?<br>**Label**: one asks for the best, whereas the other asks for how |
| **Q1**: Study tips to pas ca ipcc?<br>**Q2**: If you are unhappy with your current job, would you quit right away & find another job or wait until you find a job. What are the pros & cons of each?<br>**Label**: not duplicate<br>**NLE**: completely different questions |
| **Q1**: How long does Klonopin (Clonazepam) stay in your system?<br>**Q2**: How long does 1 mg of Klonopin keep working in your system?<br>**Label**: not duplicate<br>**NLE**: the second question gives the exact amount, but the first question doesn't |
| **Q1**: Is a third World War imminent?<br>**Q2**: How close is a World War III?<br>**Label**: duplicate<br>**NLE**: "imminent" means will happen very soon, which is equivalent to "close" |
| **Q1**: What are some of the resources to learn about IoT?<br>**Q2**: What are the best resources to learn about the Internet of Things (IoT)?<br>**Label**: duplicate<br>**NLE**: both ask for the resources for IoT |
| **Q1**:Which are some of the best movies of 2016?<br>**Q2**: What has been the best movie of 2016?<br>**Label**: duplicate<br>**NLE**: both ask for the best movie of 2016 |
| **Q1**: Why is Saltwater taffy candy imported in Switzerland?<br>**Q2**: Why is Saltwater taffy candy imported in the Philippines?<br>**Label**: duplicate<br>**NLE**: both ask for the import of Saltwater taffy candy, albeit the different locations |

Table 14: QQP instances and the corresponding NLEs.