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

Anonymous ACL submission

Abstract

Recent studies have demonstrated that large language models (LLMs) excel in diverse tasks through in-context learning (ICL) facilitated by task-specific prompts and examples. However, the existing literature shows that ICL encounters performance deterioration when exposed to adversarial inputs. Enhanced performance has been observed when ICL is augmented with natural language explanations (NLEs) (we refer to it as X-ICL). Thus, this work investigates whether X-ICL can improve the robustness of LLMs on a suite of adversarial and challenging datasets covering natural language inference and paraphrasing identification. Moreover, we introduce a new approach to X-ICL by prompting an LLM (ChatGPT in our case) with few human-generated NLEs to produce further NLEs (we call it ChatGPT few-shot), which we show superior to both ChatGPT zero-shot and human-generated NLEs alone. We evaluate five popular LLMs (GPT3.5-turbo, LLaMa2, Vicuna, Zephyr, Mistral) and show that X-ICL with ChatGPT few-shot yields over 6% improvement over ICL. Furthermore, while prompt selection strategies were previously shown to improve ICL on in-distribution test sets significantly, we show that these strategies do not match the efficacy of the X-ICL paradigm in robustness-oriented evaluations.

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 test data after observing a small number of demonstrations. Remarkably, they achieve this without the necessity for further training or any modifications to their underlying parameters. This novel learning paradigm is referred to as *in-context learning* (ICL, Brown et al., 2020; Rae et al., 2021). However, it has been noted that ICL struggles with the execution of complex tasks, such as arithmetic, commonsense, and symbolic reasoning (Rae et al., 2021). To enhance the capability of ICL in solving tasks requiring complex reasoning, Wei et al. (2022b) draw inspiration from the extensive body of literature on natural language explanations (NLEs) to introduce a method denoted as the Chain-of-Thought (CoT) prompting. This method empowers LLMs to utilize human-written NLEs as a mechanism for deliberate thinking before delivering a prediction. Note that CoT and NLEs are interchangeable, describing the same concept. Since NLEs were introduced before CoT (Camburu et al., 2018; Hendricks et al., 2018), we used the former term in this paper. We denote the ICL equipped with NLEs as *X-ICL*. Albeit 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).

Similar to supervised learning, ICL demonstrates vulnerability to adversarial and misleading examples, causing a decline in performance (Wang et al., 2023a). Given that X-ICL promotes deliberate thinking in LLMs, we hypothesize that incorporating NLEs could enhance the resilience of LLMs against adversarial inputs, *aka* robustness. To this end, we leverage eight adversarial datasets to evaluate the added benefit of X-ICL to the robustness of LLMs.

Moreover, the effectiveness of X-ICL so far relies on human-written NLEs (Wei et al., 2022b), which usually require domain-specific expertise, thereby imposing constraints on its scalability. However, the advent of ChatGPT[1] uncovered a range of possibilities where LLMs can assist human annotators (Bang et al., 2023; Guo et al., 2023). Motivated by this development, we leverage ChatGPT (specifically, GPT3.5-turbo) to generate NLEs for examples from human-written NLEs. Following this generation step, four authors assess the quality of the human-written and ChatGPT-

[1] https://openai.com/blog/chatgpt

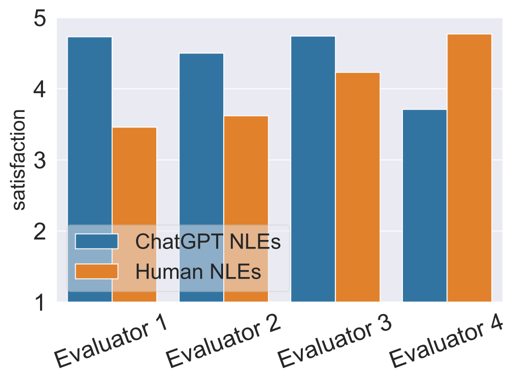

Figure 1: Human evaluation on ChatGPT-generated NLEs (ChatGPT NLEs) and human-written NLEs (Human NLEs). The satisfaction scores span from 1 (extremely dissatisfied) to 5 (extremely satisfied).

generated NLEs. As demonstrated in Figure 1, most of the evaluators (3 out of 4) exhibited a preference for the NLEs produced by ChatGPT over those crafted by humans. The details of this evaluation are presented in Appendix D.1.

In this paper, we evaluate the improvement in the robustness of LLMs provided by X-ICL in three regimes: utilizing NLEs generated by ChatGPT (generated in zero-shot and few-shot settings) and human-written 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 the challenging datasets.

Our experimental results suggest that X-ICL generally produces more accurate results than ICL on eight adversarial and challenging datasets. Furthermore, using few-shot ChatGPT-generated NLEs leads to more than 6% gains over ICL for the majority of the LLMs and datasets. The findings from our comprehensive study suggest that an integrated approach, combining human input with the capabilities of ChatGPT (*i.e.,* ChatGPT few-shot regime), provides a more effective solution than utilizing either human or ChatGPT (zero-shot) NLEs in isolation. Finally, while prompt-selection strategies (Gupta et al., 2023; Levy et al., 2023; Ye et al., 2023) considerably enhance ICL performance on in-distribution test sets, they are less effective on the adversarial datasets compared to the X-ICL approaches.

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 free-form natural language explanations (*i.e.,* NLEs) (Camburu et al., 2018; Narang et al., 2020; Wiegreffe and Marasovic, 2021). Our work concentrates on the latter category, namely, the generation of NLEs for justifying model predictions. Rajani et al. (2019) propose a two-stage training to improve the prediction performance for commonsense reasoning tasks. In their work, the first stage revolves around creating an NLE, which aids in label prediction during 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), medical (Kayser et al., 2022), and 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 continues to encounter difficulties with tasks that require reasoning abilities, including arithmetic, logical, and commonsense reasoning tasks (Rae et al., 2021; Srivastava et al., 2022). To augment the reasoning capabilities of LLMs, Wei et al. (2022b) introduced a method known as 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 the prediction. Subsequently, Zhou et al. (2023) demonstrate that dividing complex problems into simpler sub-problems and addressing them sequentially improves the performance of CoT prompting. Additionally, Wang et al. (2023b) propose an alternative method that enhances CoT prompting by combining multiple diverse reasoning paths generated by LLMs, surpassing the performance of a greedy CoT prompting approach. However, these aforementioned methods need human-written NLEs as CoT. Instead, our ChatGPT zero-shot regime harnesses the power of an LLM to synthesize NLEs without the need for human-written NLEs.

Learning Robust Models. Several works show that NLP models are prone to performance degradation when presented with adversarial datasets, 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). To mitigate biases within NLP models, various strategies have been proposed, *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). Additionally, data augmentation presents another viable option (Minervini and Riedel, 2018; Wu et al., 2021, 2022). Moreover, studies have shown that incorporation of rationalization methodologies into supervised models can significantly enhance the models' resilience against adversarial datasets (Chen et al., 2022; Stacey et al., 2022). Deviating from the precedent research, our study probes the robustness of X-ICL on eight existing adversarial datasets.

3 Methodology

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

3.1 ICL with NLEs (X-ICL)

LLMs greatly enhance their performance across various reasoning tasks when supplied with human-written NLEs (Wei et al., 2022b,a). We can define X-ICL as follows:

$$\underset{(\boldsymbol{r}',\boldsymbol{y}')\in\mathbb{R}\times\mathbb{Y}}{\arg\max}\; P_\theta\left((\boldsymbol{r}',\boldsymbol{y}')|(\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{y}_i)_{i=1}^{k},...,(\boldsymbol{x}')\right),$$

where $\boldsymbol{x}'$ denotes an unlabeled instance, $\boldsymbol{r} \in \mathbb{R}$ represents the corresponding NLE, and $\boldsymbol{y} \in \mathbb{Y}$ denotes the target label, where $\mathbb{R}$ is the space of all NLEs and $\mathbb{Y}$ is the set of possible labels for a given dataset. The objective of X-ICL is to maximize the likelihood of generating the optimal NLE, $\boldsymbol{r}' \in \mathbb{R}$, and its corresponding label, $\boldsymbol{y}' \in \mathbb{Y}$, given a demonstration set $(\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{y}_i)_{i=1}^{k}$ and an unlabeled instance $\boldsymbol{x}'$. Consequently, this prompts the LLM to produce the most plausible NLE and label combination.

3.2 Generating NLEs via ChatGPT

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 ChatGPT (or GPT3.5-turbo). This preference is driven by noting that NLEs produced by ChatGPT 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 ChatGPT. The primary aim is to encourage the LLMs to generate a correct NLE (*i.e.,* ground-truth arguments) for the correctly-predicted answer for a test instance. The NLE is generated as follows:

$$\underset{\boldsymbol{r}'\in\mathbb{R}}{\arg\max}\; P_\theta(\boldsymbol{r}'|\boldsymbol{s},(\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{r}_j)_{j=1}^{m},...,(\boldsymbol{x}',\boldsymbol{y}')), \quad (1)$$

where $\boldsymbol{s}$ is a meta-prompt representing the task. For the details of the meta-prompt, please refer to 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, ChatGPT is 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 into the existing X-ICL framework as delineated in Section 3.1. These are referred to as X-ICL ($\text{ChatGPT}_{\text{few}}$) and X-ICL ($\text{ChatGPT}_{\text{zero}}$), respectively.[2]

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

[2] In addition, we explore the application of two other widely-used, open-source LLMs for the generation of NLEs. Detailed results of these experiments are provided in Appendix C.

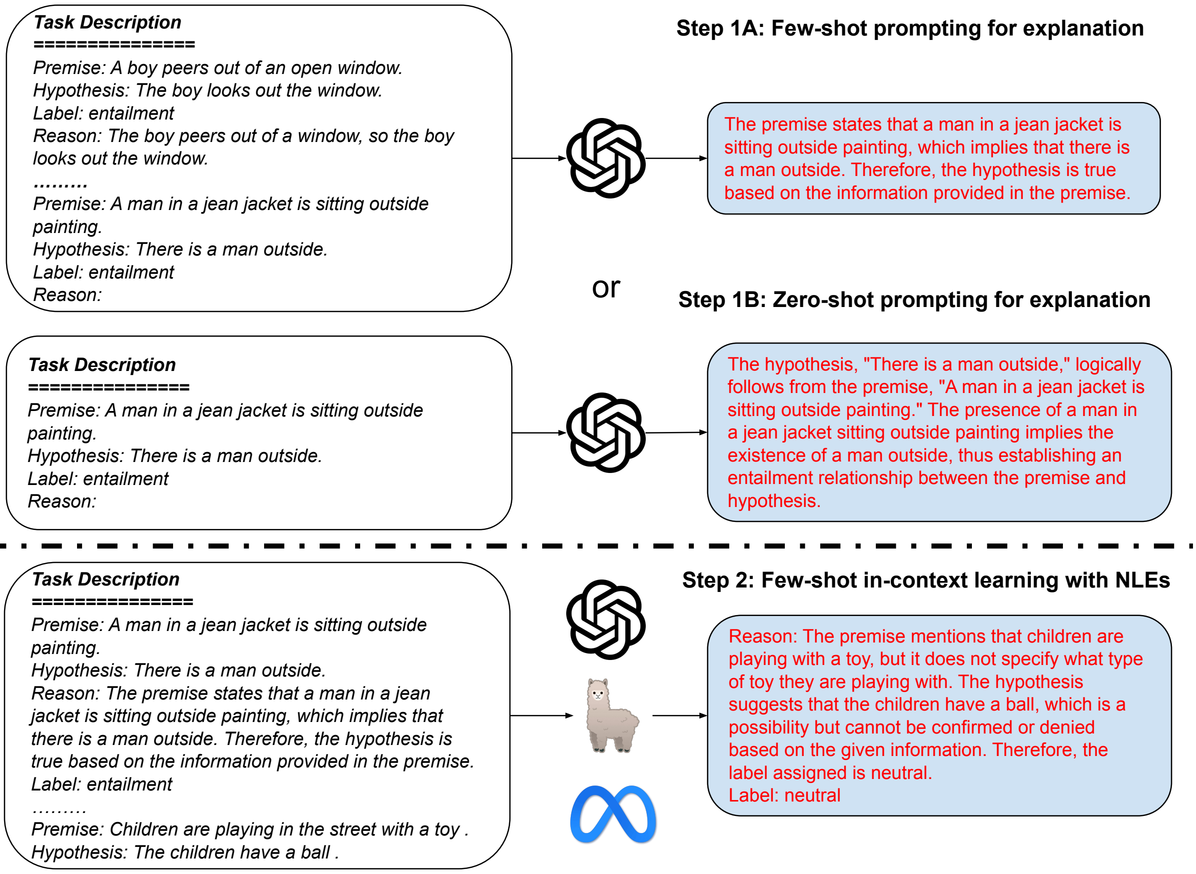

Figure 2: Our approach of using ChatGPT-generated NLEs for ICL consists of two steps: (1) prompt an LLM in a few-shot or zero-shot manner to generate NLEs for new instances; (2) prompt LLMs using ICL with the NLEs generated in step 1.

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 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 (Jiang et al., 2023), (2) Zephyr (Tunstall et al., 2023), (3) Vicuna (Chiang et al., 2023), (4) LLaMA2 (Touvron et al., 2023), and (5) GPT3.5-turbo. Specifically, Mistral and Zephyr are models with `7B` parameters. For Vicuna and LLaMA2, we use `30B` and `70B-chat` versions, respectively.

We perform all 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, barring 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 (Camburu et al., 2018) is a modified version of the SNLI dataset, enhanced with NLEs written by humans. In the case of the QQP and PAWS datasets, the QQP dataset is utilized as the demonstration set, including NLEs contributed by four authors.

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.

Models	Methods	SNLI	HANS	ISCS	NaN	ST	PICD	PISP	ANLI	QQP	PAWS	Avg.
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▽ ±2.5	58.6▽ ±2.9	51.7▼ ±4.0	56.9 ±3.3	35.8▼ ±6.7	43.9▼ ±1.7	69.9 ±0.8	66.4 ±1.5	53.5
	X-ICL (ChatGPT$_{zero}$)	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
	X-ICL (ChatGPT$_{few}$)	**61.8** ±3.1	**58.2**▼ ±2.5	**57.2**▼ ±2.2	**62.4**▼ ±2.6	**55.2**▼ ±1.5	**59.2** ±2.7	**47.6**▼ ±1.8	**46.9**▼ ±2.3	**70.3** ±1.1	**72.5**▽ ±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▼ ±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▽ ±2.3	59.3
	X-ICL (ChatGPT$_{zero}$)	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
	X-ICL (ChatGPT$_{few}$)	**74.2**▼ ±3.6	**77.4**▼ ±2.2	**67.0** ±1.6	**67.7** ±2.3	**69.3**▼ ±1.5	**70.0**▼ ±2.1	**65.6**▼ ±2.5	**52.1**▽ ±2.8	**77.3** ±0.9	**61.5**▽ ±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
	X-ICL (ChatGPT$_{zero}$)	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
	X-ICL (ChatGPT$_{few}$)	65.0 ±3.1	**74.5**▽ ±4.4	**65.5**▽ ±1.6	**66.3**▽ ±1.1	**64.8**▼ ±1.8	61.6 ±8.9	**65.9**▼ ±4.7	**57.5**▼ ±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▼ ±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
	X-ICL (ChatGPT$_{zero}$)	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
	X-ICL (ChatGPT$_{few}$)	**74.2**▼ ±2.5	**73.3**▼ ±8.5	57.7 ±1.2	**65.9**▽ ±3.2	**63.1**▽ ±3.7	**70.6**▽ ±6.5	**55.8**▼ ±5.9	**59.2**▼ ±1.6	77.6 ±0.6	**46.5**▽ ±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**▼ ±1.7	71.0 ±1.7	69.0▽ ±1.2	70.5 ±2.2	65.7▽ ±1.0	72.7▼ ±1.3	59.3▽ ±1.9	59.8▽ ±2.3	76.0 ±3.9	53.4▼ ±5.3	66.2
	X-ICL (ChatGPT$_{zero}$)	71.9 ±2.7	71.6 ±0.8	68.4▽ ±0.3	70.2 ±0.0	67.6▽ ±1.3	67.7▽ ±4.1	61.7▼ ±1.9	**60.4**▼ ±2.0	80.4 ±0.8	51.2▼ ±3.1	66.0
	X-ICL (ChatGPT$_{few}$)	75.5▽ ±2.8	**76.0**▼ ±2.0	**74.9**▼ ±0.1	**73.1**▼ ±1.4	**73.3**▼ ±0.4	**76.9**▼ ±0.4	**75.5**▼ ±3.0	59.6▽ ±1.8	79.0 ±1.7	**54.0**▼ ±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: X-ICL (ChatGPT$_{zero}$), (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario: X-ICL (ChatGPT$_{few}$). 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: ▼ (resp. ▽) represents a *p*-value lower than 10^{-3} (resp. 10^{-1}). The results of ANLI are the average of ANLI R1, R2, and R3.

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.

Firstly, the results reveal a predictable outcome for both scenarios, namely with and without using X-ICL. As the models' abilities escalate, an upward trajectory can be discerned in the average accuracy. This progression is evident when comparing the least potent model, exemplified by Mistral, to the highest-performing one, represented by GPT3.5-turbo.

Table 1 demonstrates that X-ICL (Human) yields a 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.

In the context of X-ICL (ChatGPT$_{few}$), the evidence points to a commanding lead in all evaluated tasks on both the Mistral and Zephyr, surpassing the results of ICL and X-ICL (Human) by margins of at least 5.7% and 3.6%, respectively. Despite the notable improvement on ICL when employing GPT3.5-turbo in comparison to other LLMs, X-ICL (ChatGPT$_{\text{few}}$) offers substantially additional gains, with an increase in absolute accuracy between 11%-24% on tasks such as ISCS, ST, PICD, PISP and PAWS. In essence, X-ICL augments LLM performance on the in-distribution test and bolsters LLMs' robustness in the face of adversarial test sets.

Remarkably, despite the predominant preference of human evaluators for NLEs generated by ChatGPT over those written by humans, X-ICL (ChatGPT$_{\text{zero}}$) 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 X-ICL (ChatGPT$_{\text{zero}}$) is outperformed by ICL for 4 out of the 5 LLMs analyzed, especially on LLaMA2. This apparent discrepancy between human preferences and LLM performance strongly underlines the necessity for additional investigations to enhance our understanding of this intriguing phenomenon. Since this investigation deserves a thorough and systematic study, we leave it for future work.

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.

4.3 Impacts of NLEs

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

Data selection vs. X-ICL. The efficacy of ICL in LLMs is significantly influenced by the demonstrations provided, as the model depends on these demonstrations to comprehend and address the test instances (Zhao et al., 2021; Liu et al., 2022; Lu et al., 2022). Consequently, a spectrum of research has been directed towards optimizing data selection techniques that curate ICL demonstrations from a pertinent pool of candidate data in relation to 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.

Models	Methods	SNLI	AdvNLI	Δ
Mistral	ICL	59.8	45.1	14.7
	X-ICL (ChatGPT$_{\text{few}}$)	61.8	**53.4**	**8.4**
	COSINE	67.9	46.0	21.9
	BM25	65.2	44.2	21.0
	SET-BSR	**77.6**	52.2	25.4
Zephyr	ICL	67.1	57.2	**9.9**
	X-ICL (ChatGPT$_{\text{few}}$)	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
Vicuna	ICL	65.2	57.8	7.4
	X-ICL (ChatGPT$_{\text{few}}$)	65.0	**63.5**	**1.5**
	COSINE	72.5	53.6	18.9
	BM25	67.2	52.2	15.0
	SET-BSR	**79.5**	56.4	23.1
LLaMA2	ICL	69.3	58.7	**10.6**
	X-ICL (ChatGPT$_{\text{few}}$)	74.2	**62.7**	11.5
	COSINE	71.9	57.3	14.6
	BM25	70.8	55.6	15.2
	SET-BSR	**76.7**	59.2	17.5
GPT3.5-turbo	ICL	71.9	61.4	10.5
	X-ICL (ChatGPT$_{\text{few}}$)	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, X-ICL (ChatGPT$_{\text{few}}$) and three data selection approaches on SNLI and AdvNLI (*i.e.,* 7 adversarial test sets). Δ indicates the difference between SNLI and adversarial NLI test sets. We report the average performance over all adversarial test sets.

In this context, we examine the performance of X-ICL (ChatGPT$_{\text{few}}$) in relation 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 et al., 2020), coupled with strategies to ensure information coverage at the set level, to promote the choice of informative and diverse demonstration sets (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 process that is both computationally demanding and cost-inefficient for generating NLEs for the full dataset. Therefore, our analysis is confined

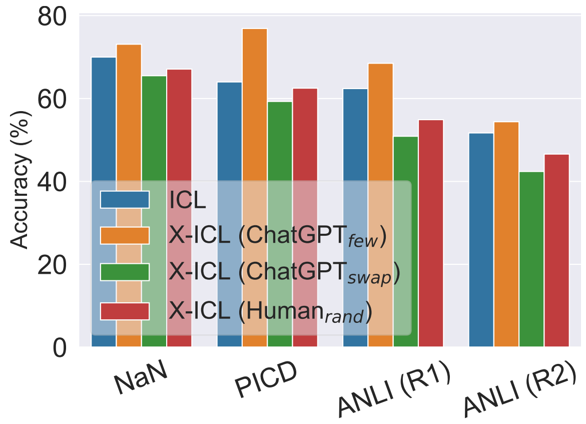

Figure 3: ICL performance of GPT3.5-turbo using (1) standard ICL without NLEs, (2) X-ICL with ChatGPT-generated NLEs in a few-shot scenario:X-ICL (ChatGPT$_{few}$), (3) X-ICL with ChatGPT-generated NLEs, where the NLEs of the prompt are swapped and do not match the instances: X-ICL (ChatGPT$_{swap}$),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

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 test set and adversarial test sets, contrary to what is observed in X-ICL (ChatGPT$_{few}$). 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 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 X-ICL (ChatGPT$_{swap}$) when compared to X-ICL (ChatGPT$_{few}$). This decline is discernible across various datasets, including NaN, PICD, and ANLI (R1/R2).[3] It is also shown that an irrelevant and arbitrary NLE triggers a performance reduction within the X-ICL framework. Furthermore, the efficiency of both 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 X-ICL (ChatGPT$_{few}$) hinges on providing an accurate and relevant NLE.

4.4 Supplementary Studies

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 4, one can find the performance of both ICL and X-ICL generally improves in correspondence with the escalation of model size, except for X-ICL (ChatGPT$_{zero}$). Moreover, the gap in performance between ICL and X-ICL (ChatGPT$_{few}$) 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,

[3]Similar patterns have been detected in other datasets

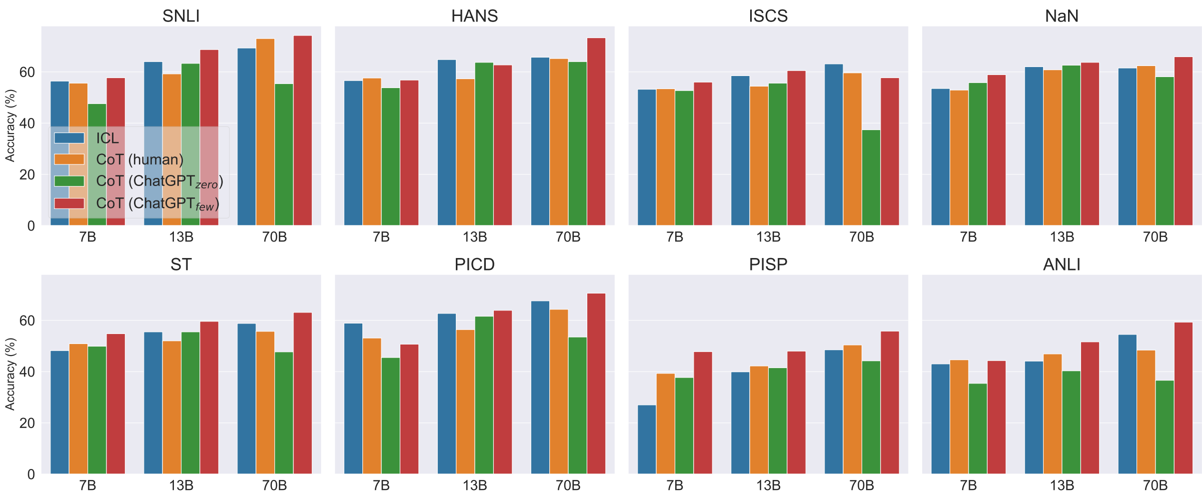

Figure 4: ICL performance of LLaMA2 (7B, 13B, 70B) using (1) standard ICL without NLEs, (2) X-ICL with human-written NLEs: X-ICL (Human), (3) X-ICL with ChatGPT-generated NLEs in a zero-shot scenario: X-ICL (ChatGPT$_{\text{zero}}$), (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario:X-ICL (ChatGPT$_{\text{few}}$). ANLI is the average of R1, R2 and R3.

	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
X-ICL (ChatGPT$_{\text{few}}$)	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 3: Performance of ICL and X-ICL (ChatGPT$_{\text{few}}$) 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.

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 3, for NaN and PICD, using e-SNLI as the prompt proves to be more effective than ANLI for both ICL and X-ICL (ChatGPT$_{\text{few}}$). This improvement can be attributed to the 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 X-ICL (ChatGPT$_{\text{few}}$) 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.

5 Summary and Outlook

We introduced a simple yet effective method called X-ICL (ChatGPT$_{\text{few}}$), leveraging human-written NLEs to generate synthetic NLEs by prompting ChatGPT. X-ICL (ChatGPT$_{\text{few}}$) 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 for the usage of the generated NLEs as faitfhul 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.

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.

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.

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.

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.

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.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

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 4 and 5 present the meta-prompts 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 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:

Table 4: 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 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:

Table 5: 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 Vicuna$_{\text{few}}$ and LLaMA2$_{\text{few}}$, respectively. To ensure a fair comparison, we employ Vicuna as the underlying model to evaluate X-ICL(Vicuna$_{\text{few}}$), X-ICL (LLaMA2$_{\text{few}}$), and X-ICL (ChatGPT$_{\text{few}}$) on all studied datasets.

Tasks	NLEs		
	Vicuna$_{\text{few}}$	**LLaMA2$_{\text{few}}$**	**ChatGPT$_{\text{few}}$**
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 6: ICL performance of Vicuna using (1) standard ICL without NLEs, (2) X-ICL with Vicuna-generated NLEs in a few-shot scenario: Vicuna$_{\text{few}}$, (3) X-ICL with LLaMA2-generated NLEs in a few-shot scenario: LLaMA2$_{\text{few}}$, (4) X-ICL with ChatGPT-generated NLEs in a few-shot scenario: ChatGPT$_{\text{few}}$. Numbers in the parentheses represent differences compared to X-ICL (Human).

Our results, detailed in Table 6, highlight that X-ICL generally gains greater benefit from LLM-generated NLEs as opposed to those produced by humans. Meanwhile, X-ICL (ChatGPT$_{\text{few}}$) consistently outperforms X-ICL(Vicuna$_{\text{few}}$) and X-ICL (LLaMA2$_{\text{few}}$) considerably, except for ANLI R1 and R2. These findings suggest that in order to fully harness the potential of AI-generated NLEs, the employment of a powerful LLM is integral.

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 X-ICL (ChatGPT$_{\text{few}}$), utilizing GPT3.5-turbo as the foundational model. A total of 280 randomly sampled, correctly predicted examples from X-ICL (ChatGPT$_{\text{few}}$) 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.

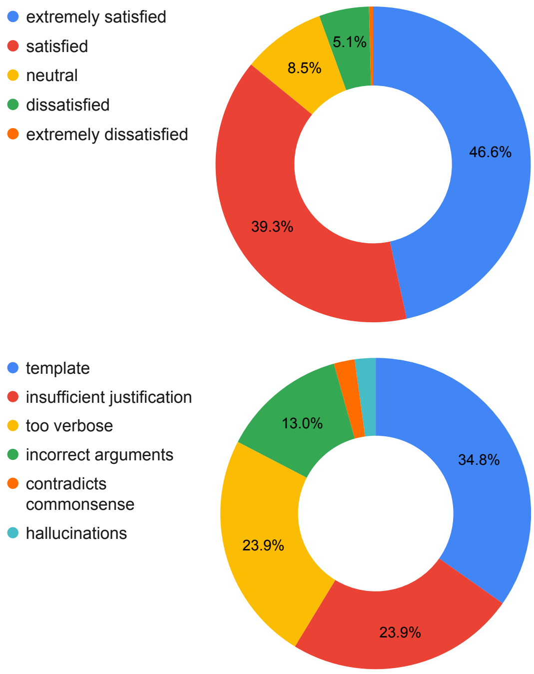

Figure 5: Human evaluation on ChatGPT-generated NLEs for the correct predictions from X-ICL (ChatGPT$_{\text{few}}$). **Top**: distribution of satisfaction scores. **Bottom**: distribution of reasons for dissatisfaction.

According to Figure 5, 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.