An Empirical Study of In-context Learning in LLMs for Machine Translation

Pranjal A. Chitale[1,2*] **Jay Gala**[3*†] **Raj Dabre**[4]

[1]Nilekani Centre at AI4Bharat [2]IIT Madras
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]National Institute of Information and Communications Technology, Kyoto, Japan

cs21s022@cse.iitm.ac.in, jay.gala@mbzuai.ac.ae, prajdabre@nict.go.jp

Abstract

Recent interest has surged in employing Large Language Models (LLMs) for machine translation (MT) via in-context learning (ICL) (Vilar et al., 2023). Most prior studies primarily focus on optimizing translation quality, with limited attention to understanding the specific aspects of ICL that influence the said quality. To this end, we perform the first of its kind, an exhaustive study of in-context learning for machine translation. We first establish that ICL is primarily example-driven and not instruction-driven. Following this, we conduct an extensive exploration of various aspects of the examples to understand their influence on downstream performance. Our analysis includes factors such as quality and quantity of demonstrations, spatial proximity, and source versus target originality. Further, we also investigate challenging scenarios involving indirectness and misalignment of examples to understand the limits of ICL. While we establish the significance of the quality of the target distribution over the source distribution of demonstrations, we further observe that perturbations sometimes act as regularizers, resulting in performance improvements. Surprisingly, ICL does not necessitate examples from the same task, and a related task with the same target distribution proves sufficient. We hope that our study acts as a guiding resource for considerations in utilizing ICL for MT. Our code is available on https://github.com/PranjalChitale/in-context-mt-analysis.

1 Introduction

Large Language Models leverage In-Context Learning to effectively solve diverse downstream tasks (Brown et al., 2020; Dong et al., 2023; Liu et al., 2023; OpenAI et al., 2023; Chowdhery et al., 2022). This inference-only method involves conditioning the model with task-specific demonstrations consisting of input-output pairs within the prompt before the actual test example. Most recently, MT via ICL has become popular (Vilar et al., 2023), however, there is still a limited understanding of how ICL works and its aspects in this context. Raunak et al. (2023) make some initial explorations on partial test sets, but there are still open questions pertaining to whether ICL is example-driven or instruction-driven, whether all examples contribute equally or not, and its potential to enforce control over the generation to regulate the pre-training biases observed in models. To answer these questions, in this paper, we conduct experiments by perturbing either the instructions or the demonstrations while keeping the other clean to assess the model's ability to perform tasks appropriately. This allows us to understand if clear instructions can guide the model effectively even when examples are perturbed and, conversely, if clean examples can steer the model to perform the appropriate task despite the instruction being misleading. We assess 6 models across 2 language families, spanning 12 language pairs on complete test sets across various noise thresholds by considering realistic perturbation attacks.

When perturbing examples, we study 1) whether ICL treats all examples equally or favors those in spatial proximity to the test example, 2) whether utilizing in-context examples from related tasks can effectively guide models in performing another task to simulate the unavailability of demonstrations for some language pairs, 3) whether various choices of auxiliary source language for the demonstrations influence downstream MT, 4) whether models can implicitly make associations based on contextual information via a transitive MT setup and finally, 5) whether smaller models are susceptible to contextual misinformation or if their pre-training biases act as safeguards against it. Our work aims to holistically understand ICL and its driving factors for

[*] Equal contribution
[†] Work done at Nilekani Centre at AI4Bharat

Findings of the Association for Computational Linguistics: ACL 2024, pages 7384–7406

MT tasks with experiments spanning over *25K* runs across various models and language directions. We observe that examples significantly influence ICL, while instructions have a limited impact. Notably, target distribution plays a more significant role than source distribution in ICL demonstrations. While perturbations might seem potentially harmful, in some cases, they can act as a form of regularization, particularly evident in the Llama 2 7B model. Spatial proximity emerges as a critical factor, implying that clean examples should be placed closer and noisy ones farther during in-context example selection for optimal downstream performance. Furthermore, our findings suggest that examples from related tasks can suffice for ICL in MT, with the constraint that the target language of the demonstrations matches the language of the test example while the choice of the source language is inconsequential. The directionality of the demonstrations has minimal impact, with both target-original and source-original demonstrations being equally effective. ICL has the potential to override semantic priors and may be exploited to misguide models or generate misinformation. We believe our study offers valuable insights for practitioners and would be widely generalizable.

2 Related Works

In-context Learning: Large Language Models have demonstrated strong performance in various downstream tasks through supervised fine-tuning on task-specific data (Devlin et al., 2019; Raffel et al., 2020; Lewis et al., 2020; Radford et al., 2019). ICL (Brown et al., 2020; Dong et al., 2023; Liu et al., 2023; OpenAI et al., 2023; Chowdhery et al., 2022) is a training-free method that has emerged as a cost-effective approach given the increasing size of LLMs ($\geq$ 7B parameters). ICL aims to implicitly learn the task-solving abilities through a few hand-crafted demonstrations. These emergent capabilities (Wei et al., 2022; Lu et al., 2023) are driven by the distribution of the pretraining data (Chan et al., 2022; Briakou et al., 2023). Intuitively, ICL can be considered analogous to performing a gradient descent implicitly during the inference (Akyürek et al., 2023; Li et al., 2023; Zhang et al., 2023b). Vilar et al. (2023) explored few-shot prompting strategies for MT task on PaLM (Chowdhery et al., 2022) suggesting that the downstream MT performance is largely correlated with the quality of demonstrations. Subsequently, several recent works have also explored ICL capabilities of LLMs for the MT task (Robinson et al., 2023; Zhang et al., 2023a; Zhu et al., 2023; Kumar et al., 2023). However, none of these works delve into a fine-grained examination of the underlying factors influencing the model performance; a gap that we fill with our study.

Factors impacting in-context learning: Several aspects of the demonstrations, such as input distribution, output distribution, or input-output alignment, can play a significant role in the ICL performance of LLMs. While Min et al. (2022) observed limited significance of label correctness in demonstrations for classification tasks, however, Yoo et al. (2022); Kossen et al. (2023) argued that ICL is sensitive to input-label mapping and can affect downstream performance. Furthermore, Wei et al. (2023) showed that ICL can override semantic priors, being influenced even by semantically unrelated input-label mappings, particularly in larger models. While prior works (Min et al., 2022; Wei et al., 2023; Kossen et al., 2023) have explored various aspects of ICL for NLU tasks, it is crucial to examine its applicability to NLG tasks. Raunak et al. (2023), which is most related to our work, found that syntactic perturbations to demonstrations affect MT performance, with target text distribution having more impact than source text distribution. However, their experiments were limited to GPT-3 (Brown et al., 2020) with only 100 test examples per direction. Our study substantially expands on this by extensively experimenting across the multiple axes mentioned earlier.

Perturbation and Robustness: Text perturbation attacks (Xu et al., 2021; Moradi and Samwald, 2021; Zhang et al., 2022; Niu et al., 2020; Salinas and Morstatter, 2024) are frequently used to assess the resilience of models to input alterations, distinguishing them from adversarial attacks (Michel et al., 2019; Garg and Ramakrishnan, 2020; Zhang et al., 2021) with their intent to simulate real-world noise rather than to explicitly deceive models. These attacks, applied at either word level (Zhao et al., 2018; Cheng et al., 2020) or character level (Belinkov and Bisk, 2018; Ebrahimi et al., 2018; Formento et al., 2023), assess NMT model robustness through strategies like introducing random or linguistically aware perturbations, such as frequently confused characters. Additionally, perturbations may also involve random removal or addition of punctuations (Formento et al., 2023). Closest to our work is the work by Moradi and Samwald

(2021), however, our approach differs from theirs as we only perturb demonstrations while leaving test examples as is.

3 Methodology

Our experimentation first focuses on understanding whether in-context learning for MT is instruction-driven or example-driven, and then expands on the latter, exploring various aspects of examples. We categorize our experimentation along two principal axes: *instruction perturbations* and *demonstration perturbations*.

3.1 Instruction Variations

LLMs require extensive task descriptions for controlled generations and aligned outputs (Sondos Mahmoud Bsharat, 2023). We intend to understand if instructions are really necessary or not and whether in-context exemplars can guide the model to perform the MT task in cases of suboptimal or confounding instructions. Keeping the demonstrations fixed, we experiment with the task instructions described below:

Standard instruction: The baseline condition, a default instruction, indicates the MT task is used.

No instruction: Only in-context exemplars are used without any explicit instructions to see if the models can infer the task implicitly.

Generic instruction: A non-task-specific generic instruction is used. This is similar to the previous setting, however, we add a generic instruction to infer the task from the exemplars.

Constrastive instruction: An instruction with the opposite translation direction than the in-context exemplars is used to assess the model's susceptibility to off-target translations due to instruction changes.

Random instruction: This includes a random instruction from a pool of non-translation tasks, while the in-context exemplars still reflect the translation task, to see if the model is misled into performing a different task.

3.2 Demonstration Perturbations

In this set of experiments, we intend to investigate the impact of perturbing in-context demonstrations on the downstream MT performance, keeping instructions fixed. Specifically, our objective is to ascertain whether the model can effectively follow clear instructions and perform the tasks or whether the inclusion of suboptimal in-context exemplars adversely influences the subsequent downstream MT performance. To achieve this, we homogeneously either perturb the source or target distribution of in-context demonstrations by introducing different types of errors. Our perturbation methods are limited by a noise budget and influence the lexical (alterations to individual words), syntactic (alterations to structure or ordering of words), and semantic (alterations to meaning) properties of the in-context demonstrations and are listed below:

Span Noise: Random contiguous segments of the original text are modified by string operations such as deletion or replacement of characters within the selected spans similar to Maurya et al. (2023) and Joshi et al. (2020). We consider the number of characters to be perturbed by uniformly selecting 1-3 gram spans and uniformly choosing to delete or replace with a single random character until the chosen budget is exhausted.

OCR: Words from the original text are uniformly selected, and operations, such as fusion with adjacent words or splitting into two words to simulate noise, are performed, which might commonly be introduced in the pipelines involving OCR systems. The noise percentage determines the degree of perturbation, affecting how many words are manipulated with a uniform probability of fusing consecutive words together or splitting a word into two parts.

Word Ordering: The original word order of some portions of the original text is disrupted. It is important to note that this perturbation can have different implications across different languages due to differing levels of word order flexibility. We uniformly sample a set of words from the original text based on noise budget for reordering by generating a new order for these words.

Word Duplication: Redundancy is added to the original text without altering its semantics by duplicating certain words. We uniformly choose a set of words for duplication based on noise budget.

Punctuation: The syntactic cues and rhythm of the original text are altered by either inserting or deleting existing punctuations present in the original text. We uniformly add punctuation to words lacking it in addition and uniformly delete from

the original text in case of deletion, given a noise budget.

Table 3 in Appendix E provides a categorization of various perturbation methods considered as a part of this study with regard to the properties it impacts.

3.3 Role of Demonstration Attributes

While perturbations influence exemplar quality, there are other important attributes such as source distribution, target distribution, spatial proximity of the demonstration, and alignment between in-context demonstrations and test examples. We focus on these to further delve deeper into which attributes of an in-context demonstration play a critical role in downstream performance.

Heterogeneous Perturbations: We perform heterogeneous perturbations to investigate if spatial proximity to the test example affects the downstream performance. Specifically, we consider cases involving a mix of clean and noisy demonstrations: $\{(k-1\text{ c}, 1\text{ n}), (1\text{ c}, k-1\text{ n}), (1\text{ n}, k-1\text{ c}), (k-1\text{ n}, 1\text{ c})\}$ where c and n indicate clean and noisy and k denotes number of shots.

Directionality: While the influence of test set directionality and translationese effects on traditional MT is well-explored (Zhang and Toral, 2019; Federmann et al., 2022), its impact on the performance of LLMs using in-context examples remains largely unexplored (Raunak et al., 2023). We explore this by choosing in-context exemplars from source-original and target-original sets to analyze whether leveraging target-original in-context examples is a superior choice than source-original in-context examples.

Demonstrations from Allied Tasks as Proxy: When demonstrations for the specific task are not available during inference, a key question arises: *Can demonstrations from a related task serve as a proxy for the model?* This is particularly relevant in MT, where obtaining demonstrations for every low-resource language pair might not be feasible. In this scenario, if the translation direction at test time is from language X to Y, we provide exemplars from language A to Y to see if these are sufficient to guide the model in generating language Y effectively. Furthermore, we also experiment with various auxiliary languages as the source for these in-context exemplars to determine if the choice of auxiliary language affects downstream performance. Specifically, we consider whether the auxiliary language a) matches the test time source language (baseline), b) matches the script of the test time source language and is not the test time source language, c) is English, d) is a randomly selected non-English language that differs in script from the test time source language.

Misalignment: Effective in-context learners should extract pertinent information from the context and provide precise responses. However, this trait can be exploited by intentionally injecting misinformation into the examples. Therefore, we aim to understand if models are susceptible to such attacks. This differs from prior perturbation experiments (see Section 5.2.1), where attacks targeted either source or target distribution without introducing misinformation through alignment alterations. Our experiment emulates a pivot translation scenario. We present two in-context examples: the first in language X with its English translation and the second demonstrating the translation of the English sentence into language Y. During testing, the model is tasked with translating from language X to Y, resembling a multi-hop reasoning process. Here, we consider 4 alignment types: a) all aligned, b) pivot misaligned, c) pivot and target misaligned, d) target misaligned. Types a) and b) are unperturbed cases, while c) and d) present perturbations or misinformation.

4 Experimental setup

This section describes the different pre-trained models, evaluation benchmarks, in-context prompts, and decoding hyperparameters used for our current study.

Models: We conduct our experimentation on off-the-shelf open-source pre-trained LLMs like Llama 2 (Touvron et al., 2023), BLOOM (Fan et al., 2022) and their corresponding instruction-tuned variants, ALMA (Xu et al., 2023) and BLOOMZ (Muennighoff et al., 2023) respectively. Firstly, models like Llama 2 (Touvron et al., 2023) are predominantly trained with English or European languages and the tokenizer has poor fertility on Indic languages. Further, Ahuja et al. (2023) show that Llama 2 70B (Touvron et al., 2023) exhibits poor in-context MT performance on Indian languages. Hence, we evaluate Llama 2 (Touvron et al., 2023) and ALMA (Xu et al., 2023) only on European

languages and evaluate BLOOM (Fan et al., 2022) and BLOOMZ (Muennighoff et al., 2023) on the remaining three Indian languages. For consistency in our experiments, we add a task-specific fine-tuned baseline (BLOOM 7B FT) by fine-tuning the BLOOM 7B model on a subset of BPCC-seed data (Gala et al., 2023) using the same prompt structure as ALMA (Xu et al., 2023). The fine-tuning hyperparameters are outlined in Appendix A. Due to computational constraints, we restrict our experimentation to 7B models for a fair comparison across the base and instruction-tuned variants. We adopt a greedy decoding strategy without sampling to generate a maximum of 256 tokens with early stopping to ensure reproducibility in our results.

Languages: Broadly, our experimentation consists of 3 Indian languages like Bengali, Hindi, and Tamil, and 3 European languages like Czech, German, and Russian. However, we consider different sets of languages for certain experiments due to the nature of the experiment or the availability of the data. Table 4 in Appendix F provides the languages considered, and the test set for each experiment.

Benchmarks: FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) is a N-way parallel benchmark that covers 200 languages, including various low-resource languages. However, it is important to note that data from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) has been consumed in the multitask fine-tuning xP3 mixture of BLOOMZ (Muennighoff et al., 2023), suggesting data contamination. Therefore, we do not consider FLORES-200 for the evaluation of Indian languages and instead, we use the newly released IN22-Gen (Gala et al., 2023). IN22-Gen is a 22-way parallel benchmark for Indian languages that focuses on demography-specific content and serves as an unbiased evaluation set for both BLOOM and BLOOMZ models.

Prompt Details: In order to maintain consistency in terms of evaluation, we follow the same prompt that was used for fine-tuning ALMA (Xu et al., 2023) across all different models considered in this work (see Appendix C). However, we find Llama 2 Chat models do not perform well with the above standard prompt in the monolithic format, which is similar to Sclar et al. (2023). Therefore, we resort to passing each demonstration as a user-assistant turn and task description in the system prompt in the chat format. We use uniformly sampled high-quality translation pairs from the FLORES-200 dev set to compose the in-context demonstrations with $k = \{1, 4, 8\}$ across various noise scales with $\delta = \{0, 0.1, 0.25, 0.5, 0.75\}$ for evaluating LLMs.

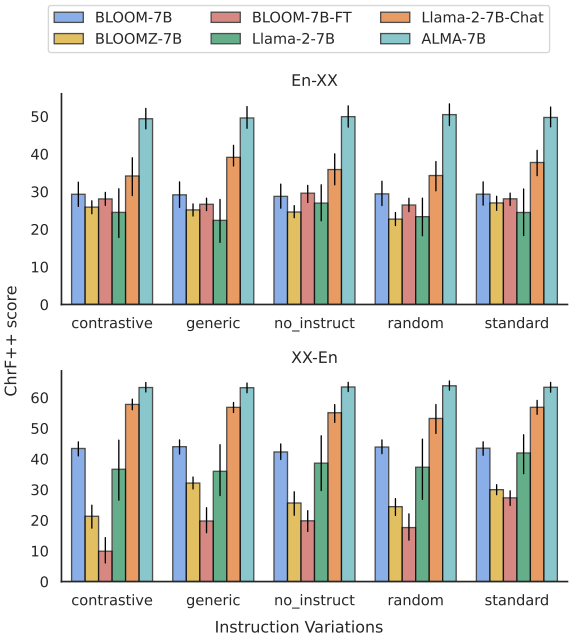

Figure 1: Comparison of ChrF++ scores for En-XX (top) and XX-En (bottom) across different instruction types for BLOOM and Llama 2 model families (averaged across different languages and shots).

Evaluation: Following Sai B et al. (2023), we use the ChrF++ metric (Popović, 2017), computed using SacreBLEU[1] (Post, 2018), as the primary metric for our analysis due to its strongest correlation with human judgments among automatic metrics.

5 Results and Analysis

We now describe results that attempt to reveal the various factors that influence ICL for MT.

5.1 Sensitivity to Instruction Variations

Figure 1 illustrates that base LLMs such as BLOOM-7B and Llama 2 7B exhibit less variability to instruction perturbations averaged across different shots. Furthermore, Figure 12 provides detailed results depicting performance variation across different shots subjected to instruction perturbation. We observe that chat models like BLOOMZ-7B and Llama 2 7B Chat show some sensitivity to instructions, particularly in a

[1] `sacreBLEU ChrF++ signature: nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.4.0`

1-shot setting, but this sensitivity diminishes as the number of shots increases. On the other hand, task-specific fine-tuned models like ALMA and BLOOM-7B-FT might be expected to be highly sensitive to instructions due to their training on data with specific instruction formats. Surprisingly, these models remain largely unaffected by perturbed instructions, even in a 1-shot setting. Although contrastive or random instructions aim to induce off-target or irrelevant generations, we observe that all the models consistently perform the translation task in the appropriate language, albeit with slightly reduced quality, highlighting that in-context examples primarily influence task recognition and learning.

5.2 Example Perturbations

Having established that models are mostly immune to instruction perturbations and are influenced more by examples, we focus on the latter.

5.2.1 Homogeneous Perturbations

Figure 2 indicates that perturbing in-context examples appear to adversely impact performance despite clear instructions. It is important to note that these instructions do not suffice as substitutes for noisy in-context examples. Generally, perturbations to the target distribution have a more severe impact than those to the source distribution, however, we observe instances where the impact of the source distribution cannot be disregarded. For instance, the BLOOMZ-7B model and ALMA-7B are susceptible to source-side perturbations, necessitating dedicated experimentation to ascertain the significance of the source distribution. We explore this aspect further in Section 5.2.3. Span noise emerges as the most detrimental perturbation across different model families, with the exception of BLOOM-7B-FT and Llama 2 7B models.

Across various perturbation attacks, we observe that the task-specific fine-tuned BLOOM-7B-FT model is most robust, followed by the BLOOM-7B model, while the multitask fine-tuned BLOOMZ-7B model is highly vulnerable. Llama 2 model generally demonstrates strong robustness to different attacks, except for the task-specific fine-tuned ALMA model, which is susceptible to span noise attack on the target side. Despite the relatively mild nature of word duplication attacks, the BLOOM-7B models surprisingly show high susceptibility. BLOOMZ-7B is robust to all attacks in XX-En translation, except for the span noise attack. Surprisingly, the Llama 2 7B model shows significant performance improvements when subjected to various attacks compared to the clean baseline across both translation and perturbation directions.

The current experiments primarily focus on homogeneous perturbations, indicating a detrimental impact on performance when in-context examples are perturbed uniformly. To investigate the potential influence of spatial proximity between perturbed in-context examples and the test examples, we delve into heterogeneous perturbations in Section 5.2.2. To limit the number of experiments due to computational constraints, we choose to exclude punctuation add and drop attacks, as the number of punctuations in a sentence is limited.

5.2.2 Heterogeneous Perturbations

We consider four cases categorized into two pairs based on the degree of noise perturbation: $\{(k-1\ \mathrm{c}, 1\ \mathrm{n}), (1\ \mathrm{n}, k-1\ \mathrm{c})$ and $\{(1\ \mathrm{n}, k-1\ \mathrm{c}), (k-1\ \mathrm{n}, 1\ \mathrm{c})\}$ where c and n indicates clean and noisy. We find that placing noisy examples closer to the test example has a more detrimental impact than placing them farther away (see Figure 3). Among these cases, $\{(1\ \mathrm{c}, k-1\ \mathrm{n})\}$ is identified as the most detrimental. Consistent with prior findings, we note an improvement in Llama 2 7B performance when exposed to noise, with the most significant improvement occurring when the noisy examples are furthest from the test example (optimal conditions at $(k-1\ \mathrm{c}, 1\ \mathrm{n})$). Llama 2 7B Chat is fairly robust and demonstrates improvement upon perturbation, except for the most detrimental case. ALMA shows slight susceptibility across all variants, with the most severe impact observed in the most detrimental case. BLOOM-7B demonstrates robustness against source-side perturbations but is heavily affected by target-side perturbations, particularly when $k-1$ noisy examples are closer to the test example. BLOOMZ-7B is generally affected across all settings, with the most severe impact in the most detrimental case. Conversely, BLOOM-7B-FT also demonstrates a fair amount of robustness. Our experimental results provide empirical evidence to practitioners that in settings involving a mixed pool of noisy and clean in-context examples, the optimal approach is to position noisy examples at the beginning and cleaner examples closer to the test example to maximize performance.

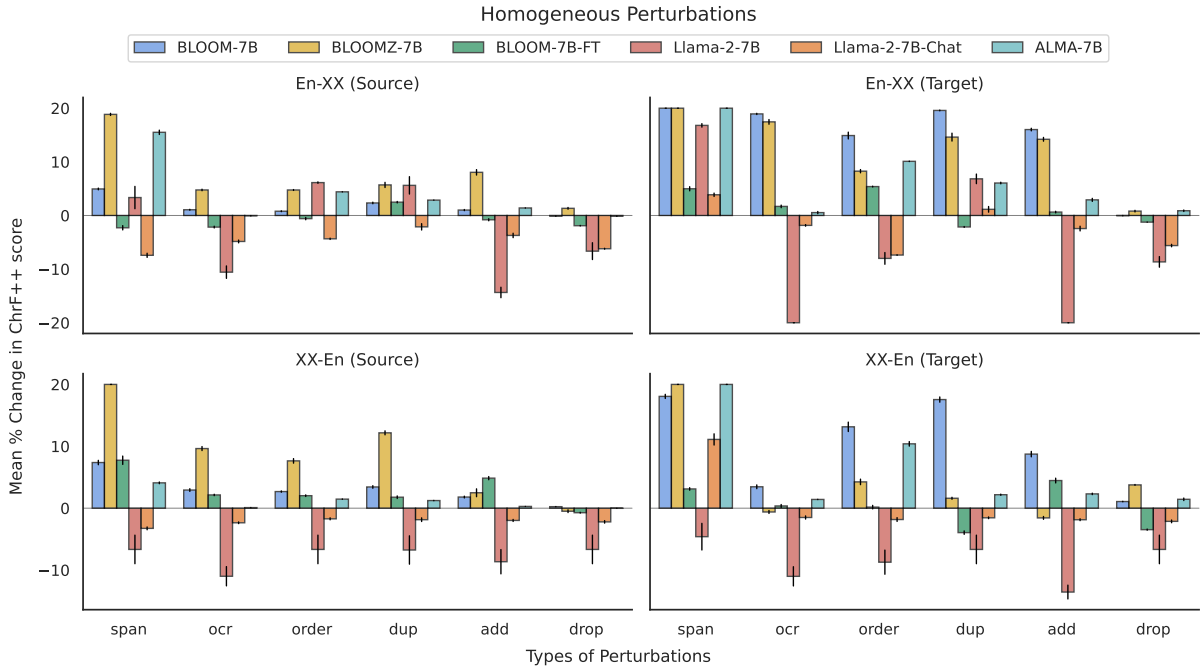

Figure 2: Mean percent change in ChrF++ score for each attack relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across shots and noise percentages (δ). "Span" denotes span noise, "order" represents word order attack, "dup" signifies word duplication attack, "add" indicates a punctuation addition attack, and "drop" signifies punctuation removal attack. Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for trend clarification.

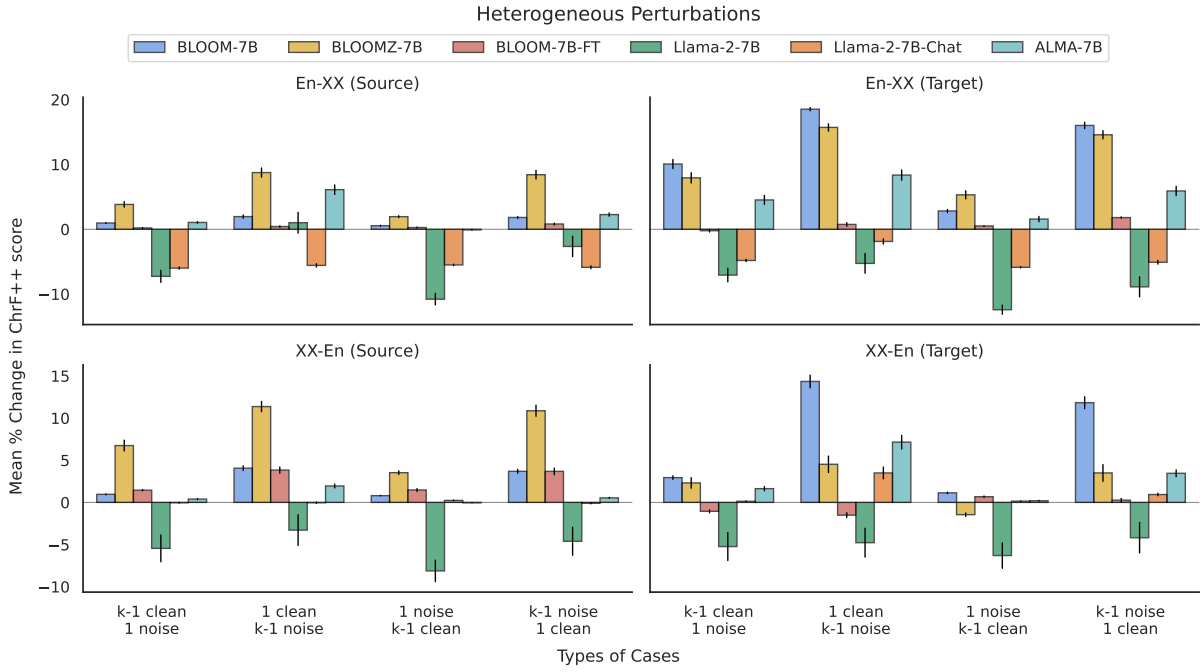

Figure 3: Mean percent change in ChrF++ score across different heterogeneous cases relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across attack types, shots and noise percentages (δ). Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for clarity in depicting overarching trends.

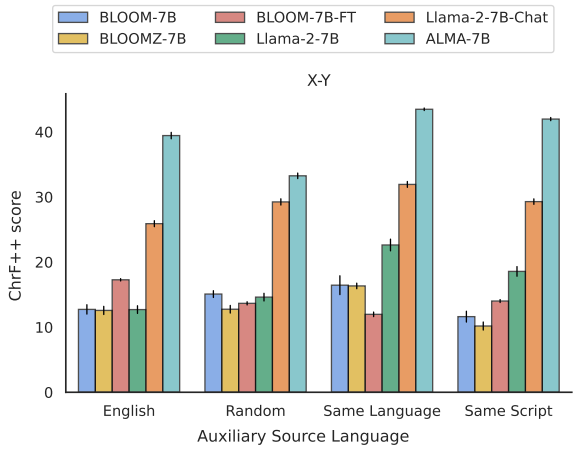

Figure 4: ChrF++ score of different models averaged across shots and translation directions comparing the choice of the auxiliary source language of the demonstration.

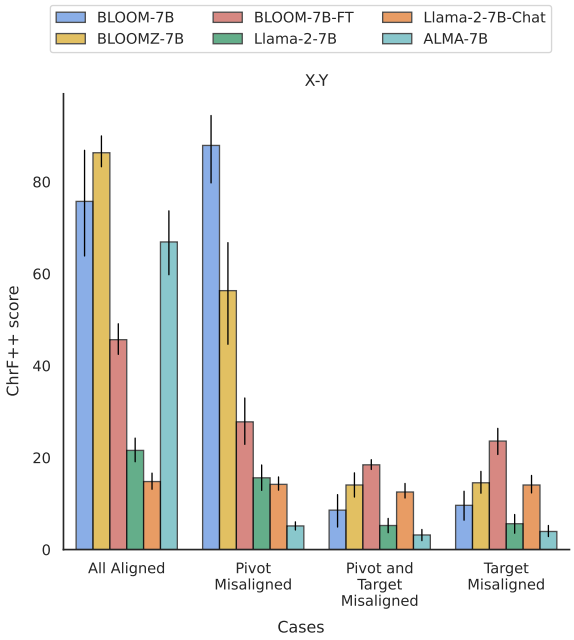

Figure 5: ChrF++ score of different models averaged translation directions comparing different forms of misalignment in a pivot-translation setting.

5.2.3 Demonstrations from Allied Tasks

Prior experiments (see Sections 5.1 and 5.2.1) indicate that examples are more influential than instructions, and the target distribution of the example has broadly more importance than the source distribution. However, our focus has been limited to in-context examples of the same task during testing. To explore the limits of example-based learning, we aim to investigate if in-context examples from a different source language but the same target language suffice as a proxy for guiding the model in translating into the desired target language, thus probing the limits of example-based learning.

Our initial experiments, selecting an auxiliary language, showed that using in-context examples from a related task can generally serve as a suitable proxy for the target task, barring a few exceptions. To delve deeper, we considered 2 additional settings: using the same script as the test time source language and using English, the predominant language models were trained on, as the auxiliary source. Figure 4 shows that the source distribution of in-context examples has a marginal effect on downstream MT performance. This suggests that any auxiliary source language can generally guide the model adequately, although there are a few exceptions, such as Llama 2 7B where using the same script as the auxiliary language outperforms English, and BLOOMZ-7B where a randomly chosen auxiliary language surpasses even using the same language. This suggests the limited impact of the choice of source language (source distribution). We conclude that in the absence of demonstrations from the exact pair as the test examples, demonstrations from any alternative source language can serve as a viable substitute, yielding comparable performance to that of the same source language baseline.

5.3 Misalignment

In this experimental setup, we aim to understand if semantic priors in models can protect against susceptibility to misinformation presented in context, rendering them effective and robust in-context learners. It is evident that most models, except Llama 2 7B Chat, can form transitive associations to varying degrees (see Figure 5). Models from the BLOOM family and ALMA are particularly susceptible to extensive copy-pasting, suggesting they can be easily misled by in-context demonstrations, especially in cases of the target being misaligned. Notably, the ALMA model heavily relies on the English sentence as a pivot for these associations, and perturbing the English example also severely impacts performance despite the target translation being present in the context. Llama 2 7B exhibits poor translation quality, being unable to deduce answers from the context using transitive relations in aligned cases, and the performance further degrades in misaligned cases, indicating vulnerability to perturbations. In contrast, Llama 2 7B Chat remains largely unaffected by misinformation-

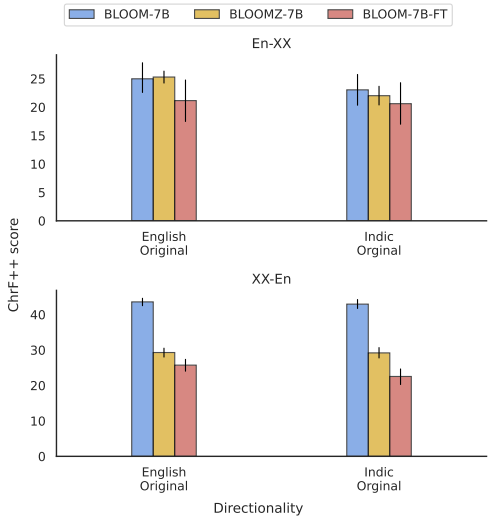

Figure 6: ChrF++ score of different models averaged across shots across translation directions comparing the choice of the source original and target original demonstrations.

based perturbations, showing robustness but poor downstream performance, indicating inadequacy in building transitive associations in the case of aligned examples. An ideal model should utilize its inherent biases to regulate its reliance on the information presented in context. Current 7B LLMs fall short in this regard, highlighting the necessity for future research to develop models that excel as in-context learners while minimizing susceptibility to misinformation within the context.

5.4 Directionality of ICL Demonstrations

Sections 5.2.1 and 5.2.3 shows the significance of target distribution; therefore, we further study if target-original demonstrations are superior to source-original demonstrations, depending on the translation direction. Focusing on Indic languages, we utilize an in-house multi-domain Indic Original set for sourcing Indic Original demonstrations and FLORES-200 dev set (Goyal et al., 2022; Costa-jussà et al., 2022) for English original ones. IN22-Gen (Gala et al., 2023), which is English-original, serves as the test set. The Indic Original set[2] serves as the target original set for En-XX direction and the source original set for XX-En, while FLORES-200 dev set (Goyal et al., 2022; Costa-jussà et al., 2022) acts as the target original set for XX-En direction and source original set for En-XX direction.

From Figure 6, across both En-XX and XX-En translation directions, we observe a minimal difference in the performance when demonstrations are chosen from the respective source original and target-original sets for all BLOOM family models considered. This suggests that the directionality of demonstrations may not have much significance on the downstream performance. However, it's important to acknowledge that our findings are limited to experiments conducted for Indic languages and using automatic metric-based evaluations, which may not sufficiently capture whether using target-original in-context examples has any impact on the fluency of the generation or not. Further investigation through human evaluation is necessary to ascertain this, an aspect we leave for future work.

6 Conclusion

Although ICL has demonstrated good performance and is widely employed across different tasks, there is limited understanding of the aspects that influence the downstream performance of ICL. In this work, we aim to bridge this gap for MT by investigating the central question of whether ICL for MT is example-driven or instruction-driven and subsequently examining the role of different aspects of ICL. We find that: (1) ICL is primarily example-driven and choice of instruction has limited impact, with target distribution of the demonstrations being the most influential (2) Perturbation to demonstrations is generally detrimental but can have a regularization effect in some cases (3) Spatial proximity to the test example is an important factor in ICL (4) Demonstrations having the same target language can serve as a reliable proxy and choice of the source language of the demonstration is inconsequential (5) The directionality of the demonstrations has minimal impact. (6) ICL can potentially be exploited to mislead even smaller models, highlighting the need for further research to make models robust in-context learners.

7 Limitations

Our perturbation experiments focused on only 5 non-linguistic perturbations, which involved random textual attacks that were mostly uniform and language-agnostic in nature. However, it would also be interesting to explore linguistically aware perturbations such as causality alternation, entity replacement, and number replacement, similar to Chen et al. (2023). This would test if the model is susceptible to more fine-grained and subtle errors.

[2]Will be released later as a held-out set of https://www2.statmt.org/wmt24/multiindicmt-task.html

Additionally, our investigation was limited to the 7B model scale due to computational constraints, and exploring whether our current findings generalize across larger model scales would provide us a more comprehensive understanding of whether ICL capabilities models vary with scale. Furthermore, while the focus of this study was on the MT task, subsequent research should examine the transferability of these insights to diverse natural language generation tasks.

8 Ethical Considerations

Potential toxic and hateful outputs: Our work focuses on the robustness of models to various factors influencing MT via ICL. One aspect of our exploration is via perturbation which may lead to hallucinations as well as generate toxic and hateful content as a result of the model's distributions being perturbed. This may even be applicable to other generative tasks but that is not the intention of our work.

Safety Circumvention and Jailbreaking: Adversarial perturbations might be able to circumvent a model's safety parameters and enable the generation of biased and harmful outputs. We plan to release our perturbation scripts and resultant perturbed data for research, and we do not intend it to be used to perform adversarial attacks in practice intended to undo the security features, also known as jailbreaking, of a model.

9 Acknowledgements

We sincerely thank Varun Gumma (SCAI Fellow, Microsoft Research) for the insightful discussions and his generous review of this draft, which helped us organize our ideas more effectively and improve its overall readability.

References

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Megaverse: Benchmarking large language models across languages, modalities, models and tasks.

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

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. 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.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers.

Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. 2023. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–109, Toronto, Canada. Association for Computational Linguistics.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia,

Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé, editors. 2022. *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, virtual+Dublin.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, and See Kiong Ng. 2023. Using punctuation as an adversarial attack on deep learning-based NLP systems: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2023. In-context learning learns label relationships but is not conventional learning.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining multiple features for in-context example selection for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. 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 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov,

Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vikas Raunak, Arul Menezes, and Hany Awadalla. 2023. Dissecting in-context learning of translations in GPT-3. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 866–872, Singapore. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv: 2401.03729*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv: 2310.11324*.

Zhiqiang Shen Sondos Mahmoud Bsharat, Aidar Myrzakhan. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, E. Chi, Tatsunori Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*

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

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Weiwen Xu, Ai Ti Aw, Yang Ding, Kui Wu, and Shafiq Joty. 2021. Addressing the vulnerability of NMT in input perturbations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 80–88, Online. Association for Computational Linguistics.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2023b. Trained transformers learn linear models in-context.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. 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 1967–1977, Online. Association for Computational Linguistics.

Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. 2022. Interpreting the robustness of neural NLP models to textual perturbations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3993–4007, Dublin, Ireland. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv: 2304.04675*.

A Fine-tuning BLOOM 7B on MT

We fine-tune BLOOM (Fan et al., 2022) 7B model on a human-annotated subset of BPCC (Gala et al., 2023), namely BPCC-seed and this forms the task-specific fine-tuned baseline for the BLOOM family. Table 1 and Table 2 list the data scales and hyper-parameters used for fine-tuning respectively.

Language	Bitext pairs
Bengali	50K
Hindi	50K
Marathi	50K
Tamil	30.6K
Telugu	36.9K

Table 1: Statistics of the sample of the English-centric BPCC-Seed data (Gala et al., 2023) used for fine-tuning the BLOOM 7B model. Note that the data was formatted in both directions (En-XX and XX-En) directions and used for joint fine-tuning.

Hyperparameters	Value
Batch size	128
Learning rate	$5e-5$
Epochs	4
Scheduler	linear
Warmup ratio	0.03
Weight decay	0.

Table 2: Hyperparameters for fine-tuning the BLOOM 7B model on MT task.

B Use of model-based metrics

Model-based metrics such as COMET (Rei et al., 2022) have demonstrated a good correlation with human judgments across several languages (Kocmi et al., 2021). However, these metrics may not be well-calibrated for all the languages considered in this study (Gala et al., 2023). Furthermore, the scale of our experiments makes it impractical to compute model-based metrics like COMET due to the additional computational overhead needed. Consequently, we use the best string-based metric, ChrF++ (Popović, 2017) as outlined in (Sai B et al., 2023). For completeness, we also compute the COMET scores using the `wmt22-comet-da`[3] model for one of our primary experimental setups

[3] https://huggingface.co/Unbabel/wmt22-comet-da

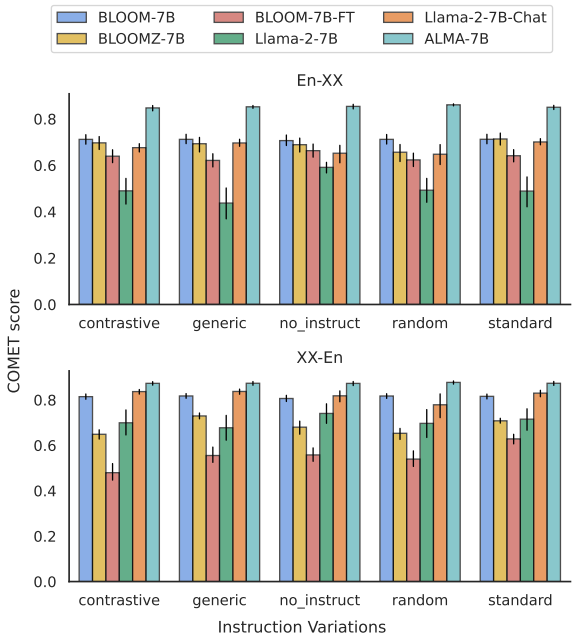

Figure 7: Comparison of COMET scores for En-XX (top) and XX-En (bottom) across different instruction types for BLOOM and Llama 2 model families (averaged across different languages and shots).

(instruction perturbation) and found the trends to be consistent with the ChrF++ metrics. This suggests that ChrF++ can serve as a reliable proxy, indicative of the overall trends. Figure 7 shows the trends in COMET scores for the instruction perturbation experiment, clearly mirroring the trends in ChrF++ scores presented in Figure 1.

C Prompt Templates

We use the standard prompt template (Figure 8) for most of our experiments except for experiments related to Sections 5.2.3 and 5.3. In the case of experiments in Section 5.2.3, we do not specify the source language in the instruction similar to the prompt illustrated in Figure 9. For the experiments in Section 5.3, we use the prompt template specified in Figure 10.

D Instruction Templates

We outline the different types of instructions considered in the instruction perturbation experiments mentioned in section 5.1. For each type of instruction, an example is provided in Appendix C. Additionally, for the random instruction type, any one of the prefix instructions is selected at random.

```
Translate this from {src_lang} into
{tgt_lang}:

{src_lang}: {src_text}
{tgt_lang}: {tgt_text}

...
...

{src_lang}: {src_text}
{tgt_lang}:
```

Figure 8: Standard prompt for translation

```
Translate this into {tgt_lang}:

{aux_lang}: {aux_text}
{tgt_lang}: {tgt_text}


...
...

{src_lang}: {src_text}
{tgt_lang}:
```

Figure 9: Prompt for Allied Task Setup

```
Translate this from {src_lang} into
{pivot_lang}:

{src_lang}: {src_text}
{pivot_lang}: {pivot_text}


Translate this from {pivot_lang} into
{tgt_lang}:

{pivot_lang}: {pivot_text}
{tgt_lang}: {tgt_text}


Translate this from {src_lang} into
{tgt_lang}:

{src_lang}: {src_text}
{tgt_lang}:
```

Figure 10: Misalignment prompt for translation

E Examples of Perturbation Variants

Table 3 categorizes the different perturbations used in this study based on attributes they impact, along with an example.

F Languages and Directions Considered

Table 4 describes the languages considered and respective benchmarks used for each experiment conducted as a part of this study.

G Additional results

Figures 12 to 18 illustrate fine-grained details such as shot-wise and case-wise trends for the aspects outlined in Section 3. The overall trends have been described in Section 5.

```
# Standard

Translate this from {src_lang} into {tgt_lang}:


# Generic

Perform the task based on the examples provided:


# Random

Complete the description with an appropriate ending:

I am hesitating between two options. Help me choose the more likely cause or effect.

Generate a headline for the following article(s) as accurately as possible.

Predict the sentiment of the review. The possible choices for the sentiment are: 'positive' and
'negative'.

Answer whether the hypothesis is more likely to be true (entailment), false (contradiction), or
unknown (neutral) based on the given premise.

The following are multiple choice questions (with answers) about subjects.



# Contrastive

Translate this from {tgt_lang} into {src_lang}:
```

Figure 11: Different types of instructions

Perturbation Method	Lexical	Syntactic	Semantic	Example
Clean				Wow! That place is so wonderful, and I would love to go there again.
Span Noise	✓	✓	×	Wow! That place is po wonde9l, a I uld love to go thyre again.
OCR	✓	×	×	Wow!That place isso wonderful, and I would lo ve to go there again .
Word Ordering	×	✓	✓	Wow! and place is so to would I That wonderful, love go there again.
Word Duplication	✓	×	✓	Wow! That place is is so so wonderful, and I I would love to go there again. again.
Punctuation$_{add}$	✓	✓	✓	Wow! That% place is so wonderful, and I would" love. to go there again.
Punctuation$_{drop}$	✓	✓	✓	Wow That place is so wonderful, and I would love to go there again

Table 3: Categorization of different perturbation methods for the different attributes.

Experiment	Models	Translation direction	Test set	In-context set	Languages
Instruction variation	Llama 2	En-X	Flores200	Flores200	ces_Latn, deu_Latn, rus_Cyrl
	BLOOM		IN22-Gen	Flores200	ben_Beng, hin_Deva, tam_Taml
Demonstration perturbation	Llama 2	En-X	Flores200	Flores200	ces_Latn, deu_Latn, rus_Cyrl
	BLOOM		IN22-Gen	Flores200	ben_Beng, hin_Deva, tam_Taml
Directionality	BLOOM	En-X	IN22-Gen	Flores200 IndicOG set	ben_Beng, guj_Gujr, hin_Deva, tel_Telu
	BLOOM		IN22-Gen	Flores200 IndicOG set	ben_Beng, guj_Gujr, hin_Deva, tel_Telu
Demonstrations from allied task as proxy	Llama 2	En-X X-Y	Flores200	Flores200	ces_Latn - rus_Cyrl (deu_Latn, eng_Latn, hin_Deva) deu_Latn - rus_Cyrl (ces_Latn, eng_Latn, hin_Deva) srp_Cyrl - deu_Latn (rus_Cyrl, eng_Latn, hin_Deva)
	BLOOM		IN22-Gen	Flores200	srp_Cyrl - ces_Latn (rus_Cyrl, eng_Latn, hin_Deva) mar_Deva - tam_Taml (hin_Deva, eng_Latn, ben_Beng) asm_Beng - hin_Deva (ben_Beng, eng_Latn, tam_Taml)
Transitivity	Llama 2	X-Y	Flores200	Flores200	ces_Latn, deu_Latn, rus_Cyrl
	BLOOM		IN22-Gen	Flores200	ben_Beng, hin_Deva, tam_Taml

Table 4: Details about the models, test sets, in-context sets and languages considered for different experiments. LLama2 family indicates Llama2-7B, Llama2-Chat-7B, ALMA while BLOOM family indicates BLOOM-7B, BLOOMZ-7B and a task-specific fine-tuned BLOOM model on MT. En-X in the translation direction indicates English-centric evaluation, while X-Y indicates non-English-centric evaluation. FLORES200 in the test set column indicates the FLORES200 devtest set, while in the in-context set column indicates FLORES200 dev set.

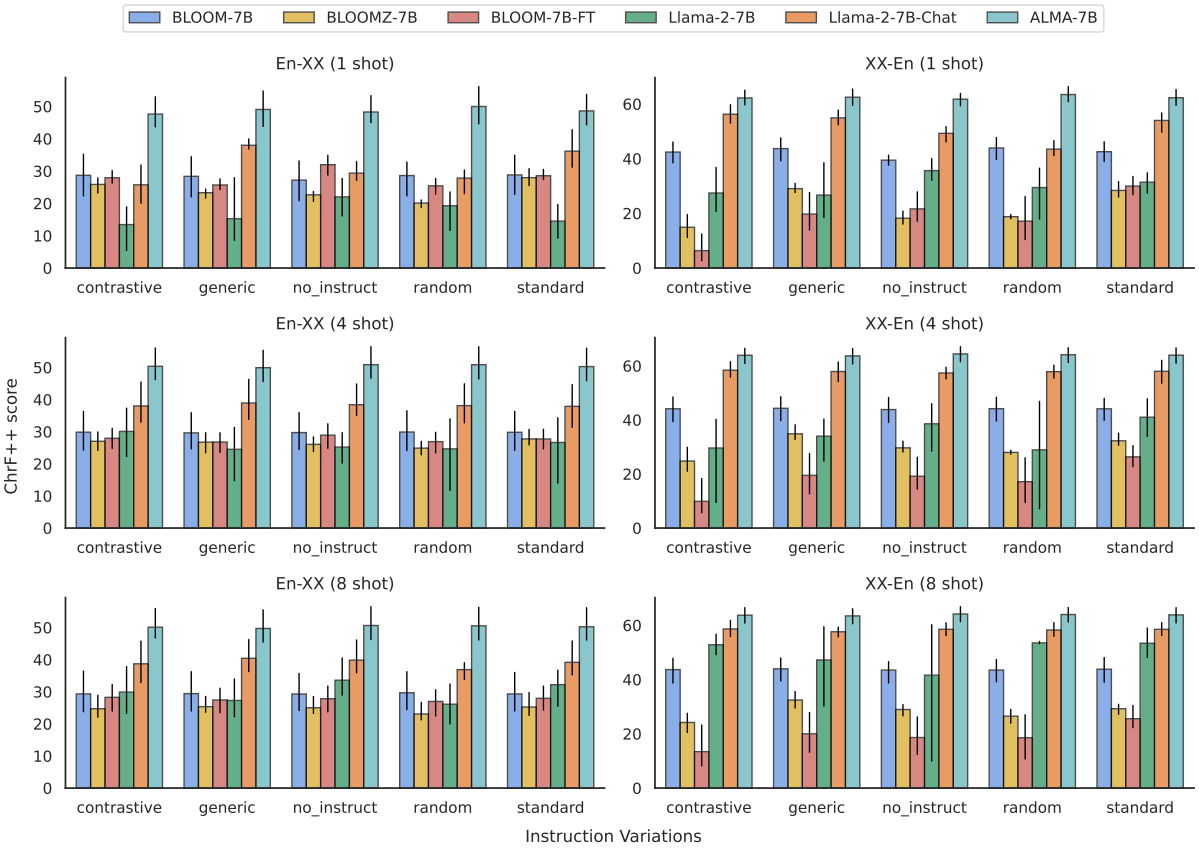

Figure 12: Shot-wise comparison of ChrF++ scores for En-XX (left) and XX-En (right) across different instruction types for BLOOM and Llama 2 model families (averaged across different languages).

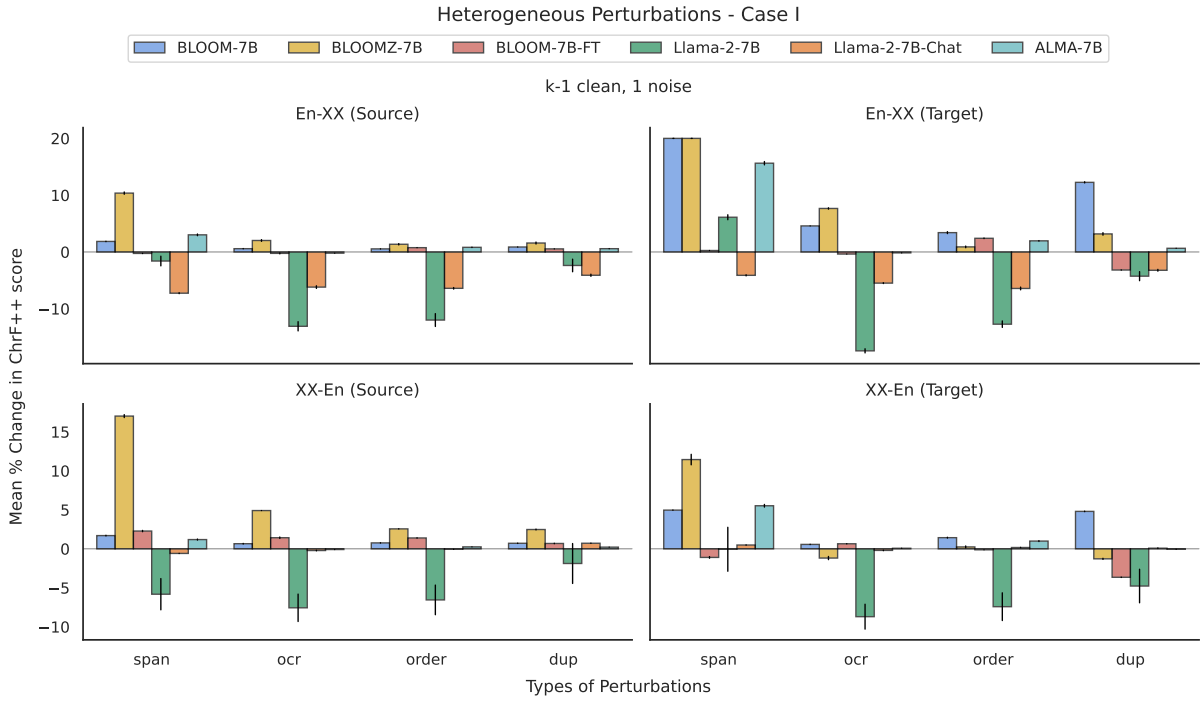

Figure 13: Mean percent change in ChrF++ score across heterogeneous case I relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across attack types, shots, and noise percentages. Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for clarity in depicting overarching trends.

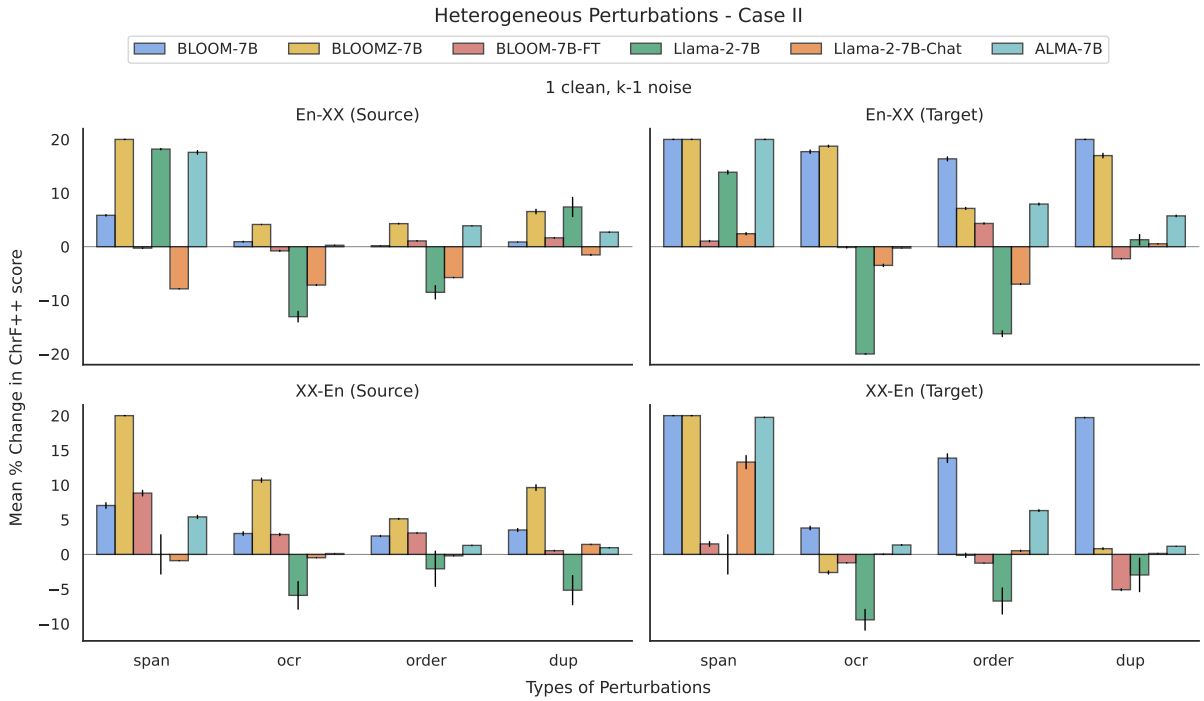

Figure 14: Mean percent change in ChrF++ score across heterogeneous case II relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across attack types, shots, and noise percentages. Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for clarity in depicting overarching trends.

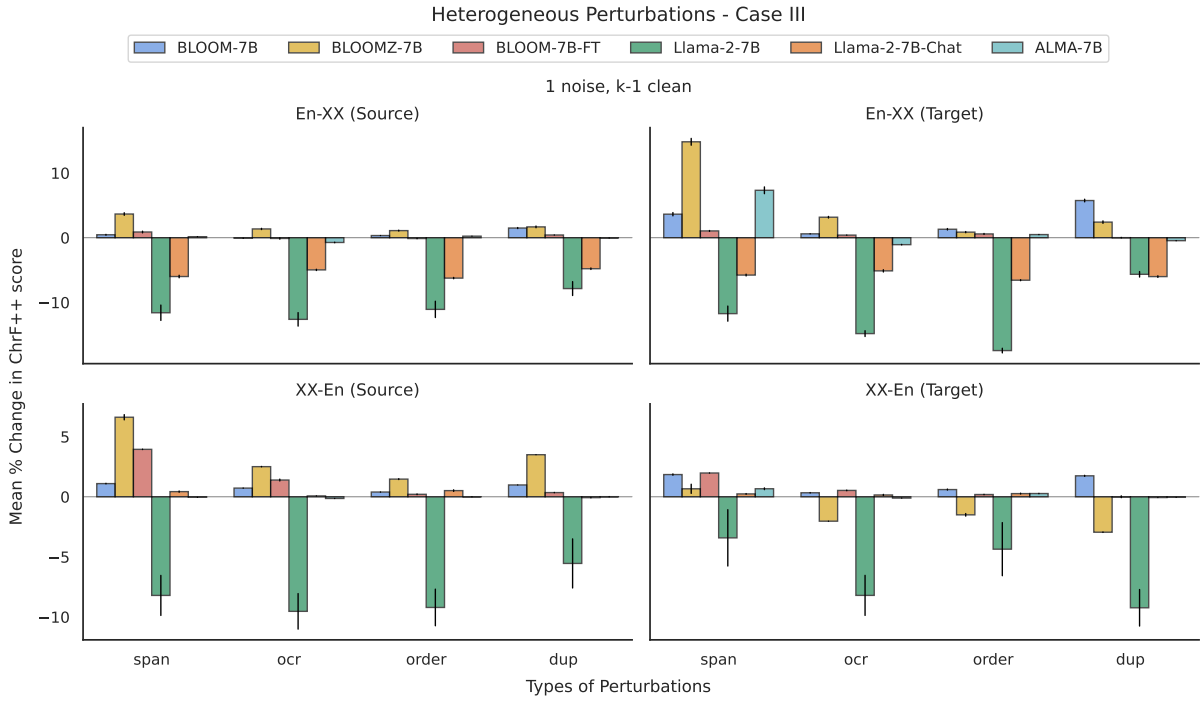

Figure 15: Mean percent change in ChrF++ score across heterogeneous case III relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across attack types, shots, and noise percentages. Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for clarity in depicting overarching trends.

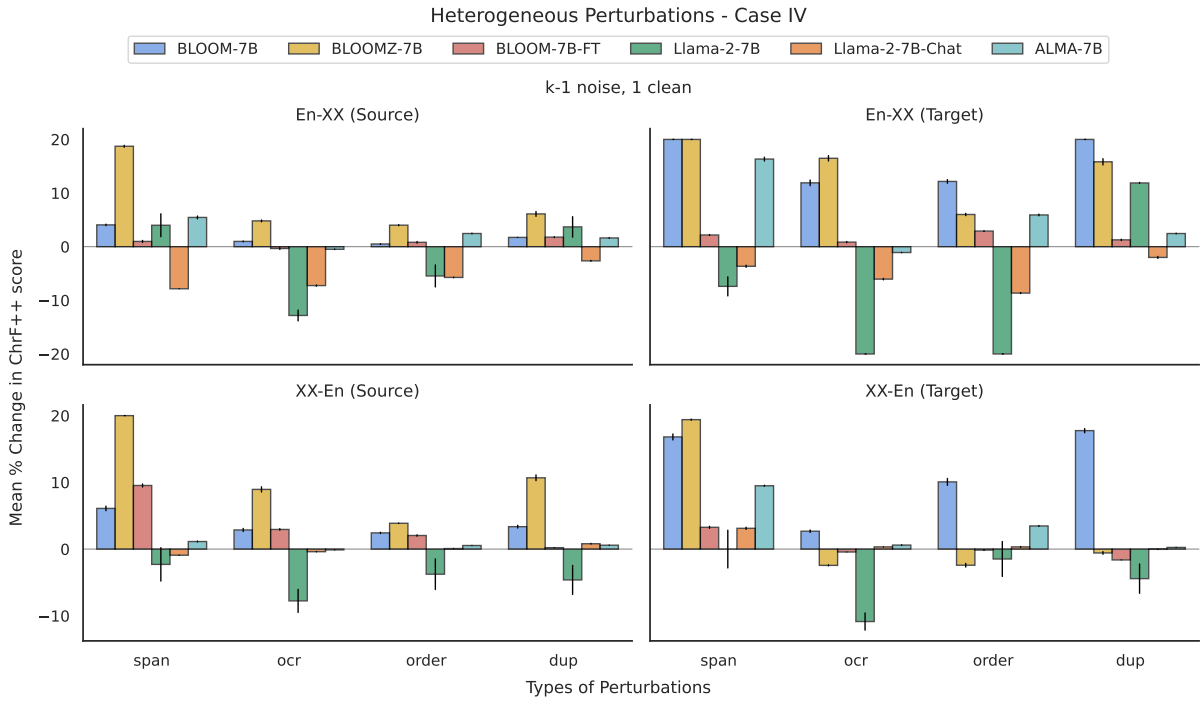

Figure 16: Mean percent change in ChrF++ score across heterogeneous case IV relative to the 0 noise baseline for each model across both translation directions (En-XX and XX-En) and both perturbation directions. Scores are averaged across attack types, shots, and noise percentages. Positive values indicate the performance decreased post perturbation while the negative values indicate that performance increases post perturbation. Note: In certain cases, scores are bounded within minimum and maximum values for clarity in depicting overarching trends.

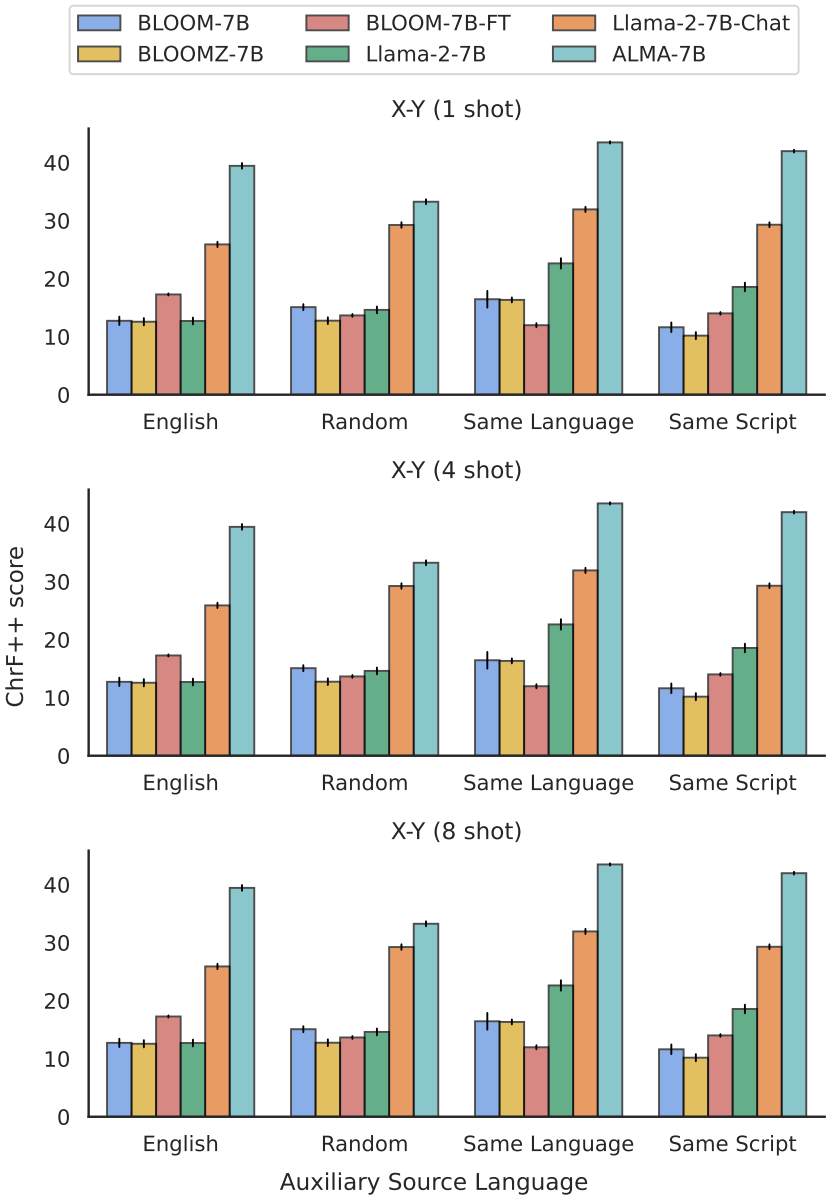

Figure 17: Shot-wise comparison of ChrF++ score of different models averaged across translation directions comparing the choice of the auxiliary source language of demonstrations.

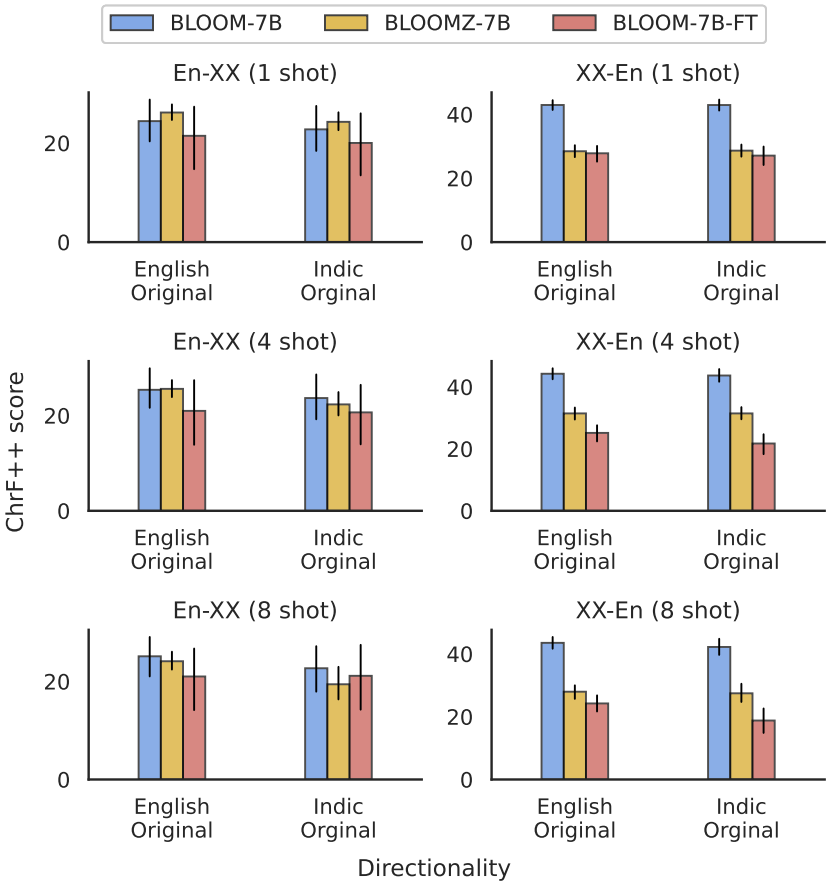

Figure 18: Shot-wise comparison of ChrF++ score of different models averaged across translation directions, comparing the choice of the source original and target original demonstrations