RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled Translation

Gabriele Sarti[*†], **Phu Mon Htut**[‡], **Xing Niu**[‡], **Benjamin Hsu**[‡],
Anna Currey[‡], **Georgiana Dinu**[‡], **Maria Nadejde**[‡]
[†]University of Groningen [‡]AWS AI Labs
g.sarti@rug.nl, {hphu, xingniu, benhsu, ancurrey, gddinu, mnnadejd}@amazon.com

Abstract

Attribute-controlled translation (ACT) is a subtask of machine translation that involves controlling stylistic or linguistic attributes (like formality and gender) of translation outputs. While ACT has garnered attention in recent years due to its usefulness in real-world applications, progress in the task is currently limited by dataset availability, since most prior approaches rely on supervised methods. To address this limitation, we propose *Retrieval and Attribute-Marking enhanced Prompting* (RAMP), which leverages large multilingual language models to perform ACT in few-shot and zero-shot settings. RAMP improves generation accuracy over the standard prompting approach by (1) incorporating a semantic similarity retrieval component for selecting similar in-context examples, and (2) marking in-context examples with attribute annotations. Our comprehensive experiments show that RAMP is a viable approach in both zero-shot and few-shot settings.

1 Introduction

Text style transfer (TST) is a task that aims to control stylistic attributes of an input text without affecting its semantic content (Jin et al., 2022). Research in TST has largely focused on English, thanks to the availability of large monolingual English datasets covering stylistic attributes like formality and simplicity (Rao and Tetreault 2018, Zhu et al. 2010, *inter alia*). In recent years, however, multilingual and cross-lingual applications of TST have seen a steady gain in popularity (Briakou et al., 2021; Garcia et al., 2021; Krishna et al., 2022). A notable instance of cross-lingual TST is *attribute-controlled translation* (ACT), in which attribute[1] conditioning is performed alongside machine translation (MT) to ensure that translations are not only correct but match user-specified preferences, such as formality/honorifics (Sennrich et al., 2016; Niu et al., 2017; Michel and Neubig, 2018; Niu and Carpuat, 2020; Nadejde et al., 2022; Wang et al., 2022), gender (Rabinovich et al., 2017; Vanmassenhove et al., 2018; Saunders and Byrne, 2020), and length (Lakew et al., 2019; Schioppa et al., 2021). ACT is especially important for sectors like customer service and business communication, where stylistic differences can have an impact on user perception (e.g., misgendering customers or speaking to them in an appropriately informal tone can be offensive or disconcerting). Table 1 gives examples of ACT for formality and gender.

Formality-Controlled Translation (CoCoA-MT)	
Neutral Src (EN)	OK, then please follow me to your table.
Formal Ref (JA)	ではテーブルまで私について来てください。
Informal Ref (JA)	ではテーブルまで私について来て。
Gender-Controlled Translation (MT-GenEval)	
Neutral Src (EN)	After retiring from teaching, Cook became a novelist.
Feminine Ref (NL)	Nadat ze stopte met lesgeven, werd Cook schrijfster.
Masculine Ref (NL)	Nadat hij stopte met lesgeven, werd Cook schrijver.

Table 1: Examples of attribute triplets from CoCoA-MT and MT-GenEval. Attribute markers in the attribute-controlled translations are underlined.

Most prior work on ACT relies on a supervised adaptation component that conditions the generative model on the selective attribute. However, few annotated ACT datasets are available, and they generally cover only a limited set of languages and attributes. Thus, enabling few-shot or zero-shot ACT would facilitate applying attribute control to less-resourced attributes and langauges.

In this paper, we introduce a new approach for ACT: **R**etrieval and **A**ttribute-**M**arking enhanced **P**rompting (RAMP). Recent studies have shown that large language models (LLMs) can perform MT out of the box using the prompting paradigm (Brown et al., 2020; Lin et al., 2022; Chowdhery et al., 2022). We build on this, prompting LLMs to perform *attribute-controlled* MT through two innovations: (1) *retrieval of similar examples* and (2)

[*]Work conducted during an internship at Amazon.

[1]In this paper, we prefer the term *attribute* rather than *style*, since not all the attributes addressed here (e.g., gender) can be considered styles.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1476–1490

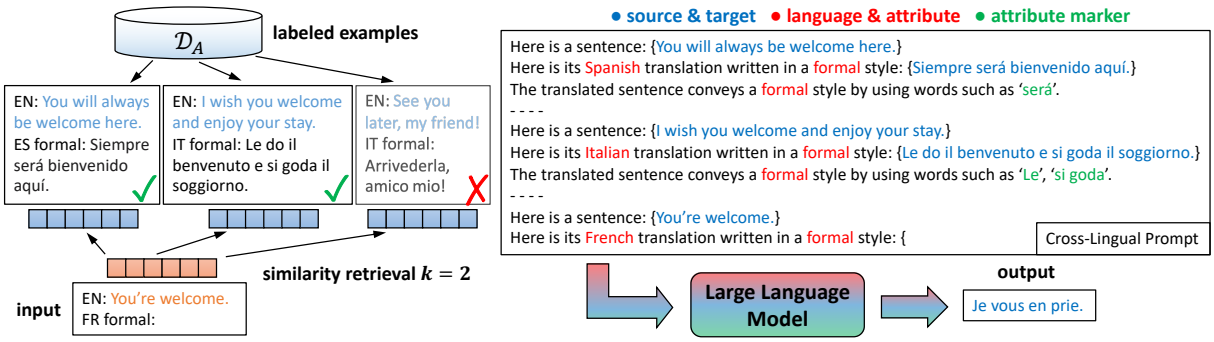

Figure 1: An example of RAMP using 2 in-context examples. (Left) The input sentence is embedded by a sentence similarity model, and the top-k most similar labeled examples are retrieved from a pool of training data to build the prompt context. (Right) Labeled cross-lingual examples are used to fill in the English prompt template, which is then provided to the LLM to generate the output.

explicit attribute marking.

Recent works adopting the prompting paradigm for text style transfer have mainly focused on the generalization capabilities of large English-centric LMs for zero-shot style transfer using previously unseen style descriptions (Suzgun et al., 2022; Reif et al., 2022). However, prior work on other NLP tasks has shown that cross-lingual prompting of multilingual LLMs can be effective (Zhao and Schütze, 2021; Zhou et al., 2022; Huang et al., 2022). As such, we leverage multilingual LLMs and extend their ACT capabilities cross-lingually to languages not covered by the in-context examples, thus enabling zero-shot ACT.

2 Method

2.1 Preliminaries

Attribute-Controlled Translation ACT takes two inputs, a sentence $\boldsymbol{x}$ and a desired target attribute $a \in A$ (with A being the space of attributes), and outputs a translation $\boldsymbol{y}$ that complies with the specified attribute. It can be formulated as a function $f : (\boldsymbol{x}, a) \rightarrow \boldsymbol{y}$. In our experiments, we use attribute values provided by the COCOA-MT formality translation dataset and the MT-GENEVAL gender translation dataset, i.e., $A = \{\text{formal, informal}\}$ or $\{\text{female, male}\}$.[2]

Prompting In the prompting paradigm for decoder-only LLMs, inputs are given as decoding prefixes to the model, usually combined with natural language instructions for output generation. In style-controlled translation, we formulate the prompt for target language l and attribute a using the text *"Here is a sentence: {$\underline{x}$} Here is its $\underline{l}$ translation written in a $\underline{a}$ style:"* to produce the output $\boldsymbol{y}$.[3] In the few-shot setting, we provide a sequence of k labeled *in-context examples* before the unlabeled input, which can be formulated as a function $f : \{(\boldsymbol{x}_1, l_1, a, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_{k+1}, l_{k+1}, a)\} \rightarrow \boldsymbol{y}_{k+1}$.

2.2 Our Approach: RAMP

RAMP builds on the success of the prompting paradigm on few-shot generation tasks such as monolingual text style transfer (Reif et al., 2022) and MT (Garcia and Firat, 2022; Agrawal et al., 2022) by creating more informative prompts through *similarity retrieval* and *attribute marking*. See Figure 1 for an illustration of RAMP.

Similarity Retrieval In standard prompting, in-context examples are sampled randomly from the pool of labeled examples $\mathcal{D}_A$. In RAMP, we select examples based on their similarity with the input text. We first embed both the input text and the source texts of $\mathcal{D}_A$ using all-MiniLM-L6-v2 (Wang et al., 2020). Then, the top-k most similar examples are retrieved for the input text based on cosine similarity. These are then used in a descending order w.r.t. similarity as the in-context examples in the inference prompt. As demonstrated in Figure 1, the in-context example "You will always be welcome here." has the highest similarity to the test example "You're welcome." so it is prompted first.

Attribute Marking In standard prompting, in-context examples are provided without explicit information on why they satisfy the prompting objective. Inspired by recent studies that have shown that decomposition of complex tasks can improve prompting quality (Nye et al., 2021; Wei et al.,

[2] See Section 5 for ethical considerations.

[3] We adopt prompt templates similar to the one used by Reif et al. (2022), and we write the prompt template in English. Complete templates are provided in Appendix A.

	AR	ES	FR	HI	PT	DE	IT	JA	RU	NL
CoCoA-MT		✓	✓	✓	✓	✓	✓	✓		✓
MT-GenEval	✓	✓	✓	✓	✓	✓	✓		✓	✓
XGLM	✓	✓	✓	✓	✓	✓	✓	✓	✓	
BLOOM	✓	✓	✓	✓	✓					

Table 2: Target languages in the test sets and languages **seen** by LLMs in pre-training. We report results on languages seen by both LLMs. Language codes are defined in Appendix B.

2022), we include for every in-context example an additional sentence directly after the target sentence that specifies which text spans convey the desired attribute (e.g., *"The translated sentence conveys a formal style by using words such as 'Vous'."*). In our experiments, we use the gold attribute spans included in the CoCoA-MT and MT-GenEval datasets. In section 4 we suggest possibilities for automatically deriving attribute spans when gold training labels are not available.

2.3 Cross-Lingual Prompting

The similarity retrieval component of RAMP requires a large pool D_A from which to find appropriate in-context examples for prompting. Low-resource attributes or language pairs may have insufficient or no annotated data from which to retrieve such examples. To mitigate this issue, we introduce *cross-lingual prompting*, in which the target side of the in-context examples differs from the desired target language of the translation task. As demonstrated in Figure 1, we study whether the system can leverage examples in one language (e.g., attribute indicators in Spanish) to produce the same attribute in another (e.g., French). Two main features of our RAMP model allow us to perform cross-lingual prompting: (1) the use of multilingual LLMs, and (2) the example retrieval step, which is done on the source language only.

3 Experiments

3.1 Datasets

We experiment on two multilingual ACT datasets:

- **CoCoA-MT** (Nadejde et al., 2022) covers formality-controlled translation in the conversation domain. Source sentences are underspecified for formality, and references require formality markings (formal or informal).
- **MT-GenEval** (Currey et al., 2022) covers gendered translation in the Wikipedia domain. We use the *contextual* subset, in which sentences are gender ambiguous in the source while the reference requires gender marking. We do not use the disambiguating sentences, instead explicitly controlling target gender.

Both datasets have gold annotations for attribute-marked target spans, and both cover translation from English into multiple diverse target languages. We list their target languages in Table 2.

Dataset	Attribute	# Train	# Test	Acc.
CoCoA-MT	Formality	7,600	1,596	0.990
MT-GenEval	Gender	4,900	9,854	0.970

Table 3: Dataset statistics. We report **#** of triplets in the **train/test** split aggregated across all languages and the classification **acc**uracy on the test split of the classifiers.

3.2 Large Language Models (LLMs)

We select three massively multilingual decoder-only LLMs for the prompting experiments: XGLM (Lin et al., 2022), BLOOM (BigScience, 2022) and GPT-NeoX (Black et al., 2022). The selected models span three orders of magnitude in terms of number of parameters and differ in the languages that they cover (see Table 2). Appendix D motivates our choice of models in more detail. GPT-3 is not included because it is not freely accessible and it is not intended for multilingual use-cases.

3.3 Baseline

Attribute tagging is a standard method for ACT, so we include a baseline following the approach and configuration used by Nadejde et al. (2022): a transformer MT model (Vaswani et al., 2017) pre-trained on public parallel data and further fine-tuned on contrastive training pairs with attribute tags (from either CoCoA-MT or MT-GenEval). We refer to this as **adapted MT**.

3.4 Evaluation Metrics

We measure translation quality with BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). For attribute accuracy, we use both (1) the lexical matching metrics provided with CoCoA-MT and MT-GenEval (**Lexical-Accuracy**) and (2) sentence encoders trained on contrastive examples (**Sentential-Accuracy**). For (2), we train multilingual classifiers on top of the mDeBERTa-v3 encoder (He et al., 2021). High-performance pre-trained classifiers have been shown to produce attribute accuracy estimates closer to human judgments for style transfer (Lai et al., 2022). Table 3 presents the accuracy of the classification models on the test sets of their respective datasets, averaged over all languages.[4]

[4]More details of datasets and classifiers are in Appendix C.

			CoCoA-MT				MT-GenEval			
			BLEU	COMET	L-Acc	S-Acc	BLEU	COMET	L-Acc	S-Acc
Same-Language	XGLM 7.5B	base	28.6	**0.463**	0.835	0.846	23.7	0.445	0.790	0.727
		+mark	28.7	0.423	0.920	0.902	23.7	0.444	0.789	0.732
		RAMP	**30.0**	0.451	**0.938**	**0.923**	**24.8**	**0.473**	**0.836**	**0.820**
	BLOOM 175B	base	39.9	0.691	0.930	0.940	33.3	0.679	0.748	0.704
		+mark	40.3	0.688	0.970	**0.970**	33.1	0.674	0.759	0.725
		RAMP	**41.9**	**0.711**	**0.973**	**0.970**	**34.3**	**0.699**	**0.817**	**0.818**
	Adapted MT		38.5	0.454	0.691	0.693	39.6	0.750	0.842	0.864
Cross-Lingual	BLOOM 175B	base	32.1	0.644	0.567	0.596	28.5	0.469	0.777	0.633
		RAMP	31.8	0.646	**0.625**	**0.622**	**29.4**	**0.502**	**0.788**	**0.673**

Table 4: BLEU, COMET, **L**exical- and **S**ential-**Acc**uracy of selected LLMs using 16 same-language in-context examples on two tasks, alongside adapted MT models. Scores are aggregated across **seen** languages (w.r.t. BLOOM pre-training) and both attributes for each task. (Decomposed results are included in Table 6–9.)

Unlike lexical accuracy, the multilingual attribute classifier does not penalize text generated in incorrect languages. Thus, in cross-lingual prompting experiments, we include a step of language detection[5] so that generated sentences not in the requested target language are considered incorrect.

3.5 Results: Same-Language Prompting

We first evaluate the effectiveness of RAMP for formality- and gender-controlled translation where the language pair used for in-context examples is the same as the one used in the prompt candidate (e.g., EN→ES formality-controlled translation using EN→ES in-context examples). We test XGLM 7.5B and BLOOM 175B with 16 in-context examples on both tasks.[6] Table 4 presents our results alongside the adapted MT baseline. The base model uses in-context examples that are sampled randomly from the pool of labeled examples. We also include an ablation that adds attribute marking only on top of base, without similarity retrieval (**+mark**).

Using just attribute marking consistently improves attribute accuracy of the generated text, but it leads to degradation of COMET on CoCoA-MT. The complete RAMP with similarity retrieval not only compensates for the COMET degradation but also improves quality and attribute metrics across the board, especially for the high-capacity BLOOM 175B model.

Adapted MT outperforms BLOOM 175B on MT-GenEval in all metrics, but underperforms it on CoCoA-MT. This suggests that it is challenging to do fine-grained comparison between LLMs and standard MT systems as they might have different domain coverage. BLOOM 175B consistently outperforms XGLM 7.5B in both generic translation quality and attribute control accuracy, so we proceed with using BLOOM 175B in the cross-lingual prompting setting.

3.6 Results: Cross-Lingual Prompting

We have demonstrated the effectiveness of selecting similar same-language examples to build the prompt, echoing contemporary work (Liu et al., 2022; Agrawal et al., 2022). In this section, we evaluate the cross-lingual prompting option, i.e., retrieving in-context examples from other target languages besides the desired language of translation. We test this zero-shot setting using the leave-one-out strategy, and results of tested language pairs are averaged.[7]

Table 4 presents our results using BLOOM 175B. On both test sets, compared to the baseline, we observe improved attribute accuracy and comparable or better generic translation quality when using RAMP with cross-lingual prompting.

We do observe translation quality degradation with RAMP on some target languages of CoCoA-MT, e.g., ES. Manual analysis shows that **repeated** inaccurate retrieval results could lead to hallucinations.[8] For example, RAMP retrieves multiple sentences containing *"million"* for the input *"If you got it why not? He is worth over 20 billion dollars after all."*. This results in mistranslation of *billion* to *million* (*millionario*): *"Si lo tienes, ¿por qué no? Es millonario después de todo."*. We give detailed examples in Appendix H.

[5] https://pypi.org/project/langdetect/

[6] We proceed with this setting based on a preliminary evaluation of 3 LLMs and 4 numbers of examples in Appendix E.

[7] Languages that are not seen during the LLM pre-training are included in the prompt but not tested.

[8] Vilar et al. (2022) also observe hallucinations when the retrieved examples have bad translations (i.e., non-parallel sentences).

4 Conclusions

We introduced the new RAMP in-context learning approach to leverage attribute annotations and similar same-language or cross-lingual examples for better prompting quality. We demonstrated its effectiveness with multilingual LLMs for both formality-controlled and gender-controlled translation. We use gold annotations for attribute marking, but we leave unsupervised automatic attribute span extraction as future work.

5 Limitations

- We currently rely on gold annotations for attribute marking, which are not always available depending on the dataset. However, RAMP could be easily extended to unsupervised settings through LLM feature attribution (Sarti et al., 2023), i.e., extracting salient tokens driving the attribute prediction. This approach builds upon recent techniques in unsupervised language generation metrics (Fomicheva et al., 2021, 2022; Leiter et al., 2022). We leave an empirical evaluation of its effectiveness to future work.

- Besides the choice of in-context examples, prompting is also sensitive to their ordering (Lu et al., 2022) and the design of the template (Jiang et al., 2020). We refrain from tuning example orders and templates to avoid introducing too many variables.

- Multilingual LLMs perform competitive MT out of the box for languages seen during their pre-training. However, we noticed that BLOOM 175B produces better EN-IT translations than XGLM 7.5B even though IT is not listed as a training language of BLOOM. This could possibly be due to typological similarity between Italian and the Romance languages included in BLOOM training. We leave experiments of unseen languages as future work.

- Multilingual LLMs like the ones used in this paper require larger GPU resources for inference than standard bilingual MT systems.

- One test set we use (MT-GENEVAL) provides only two gender values (female and male), but we do not intend to imply that other genders do not exist.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *CoRR*, abs/2212.02437.

BigScience. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, and et al. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of*

the 2nd Workshop on Evaluation and Comparison of NLP Systems, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Xavier Garcia, Noah Constant, Mandy Guo, and Orhan Firat. 2021. Towards universality in multilingual text rewriting. *CoRR*, abs/2107.14749.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *CoRR*, abs/2202.11822.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation. *CoRR*, abs/2203.11131.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8568–8575. AAAI Press.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. *CoRR*, abs/2302.13942.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *CoRR*, abs/2211.09102.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2022. Controlling styles in neural machine translation with activation prompt. *CoRR*, abs/2212.08909.

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

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2022. Enhancing cross-lingual prompting with mask token augmentation. *CoRR*, abs/2202.07255.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Prompt Templates

Formality-Controlled Translation `Here is a sentence: {`$\underline{x}$`} Here is its` $\underline{l}$ `translation written in a` $\underline{a}$ `style: {`$\underline{y}$`} The translated sentence conveys a` $\underline{a}$ `style by using words such as '`$\underline{w_1}$`', '`$\underline{w_2}$`'.`

Gender-Controlled Translation `Here is a sentence: {`$\underline{x}$`} Here is its` $\underline{l}$ `translation in which the person is` $\underline{a}$`: {`$\underline{y}$`} In the translation, the` $\underline{a}$ `gender of the person is made explicit by words such as '`$\underline{w_1}$`', '`$\underline{w_2}$`'.`

B Language Code

AR	Arabic	DE	German	EN	English
ES	Spanish	FR	French	HI	Hindi
IT	Italian	JA	Japanese	NL	Dutch
RU	Russian				

C Additional Details of Datasets Splits and Pre-Trained Attribute Classifiers

We use the original train/test split provided by the CoCoA-MT dataset. Each split contains *telephony* and *topical_chat* domains. We use the *topical_chat* domain in our experiments. MT-GenEval contains a dev and test split, and we use the dev split as training data for the classification model and prompting experiments.

We finetune mDeBERTa-v3-base model[9] on the contrastive examples in the respective training sets to get the attribute classifiers. We finetune the classifier for 2 epochs with a batch size of 8, learning rate 2e-5, 500 warm up steps, max sequence length of 256, and save checkpoint every 500 steps. We do not do hyperparameter tuning, and thus, a validation set is not used.

D Selection of Large Language Models

XGLM (Lin et al., 2022) is a 7.5B-parameter model trained on a balanced corpus containing 30 languages (excluding NL). It was shown to outperform much larger models such as GPT-3 on tasks related to machine translation and cross-lingual language understanding. We select it due to its broad linguistic coverage and its manageable size.

BLOOM (BigScience, 2022) is a model available in multiple sizes, trained on a curated corpus spanning 46 natural languages (and 13 programming languages). However, many of the test set languages are not part of its pre-training corpus (see Table 2). We evaluate two variants of the model (7.1B and 175B parameters) to assess how it is affected by a massive scaling in model parameters. The larger variant has a parameter count comparable to the one of GPT-3, while it is presently the largest publicly available multilingual LLM.

GPT-NeoX (Black et al., 2022) is a 20B-parameter model trained on The Pile (Gao et al., 2021), a large English-centric corpus covering a broad range of domains. While the model saw mainly English data during pre-training and as such is not intended for multilingual usage, it exhibits interesting generalization performances for many of our target languages.

E Preliminary Evaluation of Same-Language Prompting

We conduct preliminary evaluations aimed at reducing the number of experimental settings. We perform formality-controlled translation using CoCoA-MT, and evaluate LLMs by varying the number of in-context examples (i.e., 4-8-16-32, selected based on the feasible context length[10]).

Figure 2 presents results averaged across all four languages **seen** by BLOOM during its pre-training.[11] Observations:

- `RAMP` generally outperforms base prompting (i.e., random in-context examples and no attribute marking) across most LLMs and example settings for both BLEU and formality accuracy.
- BLEU and formality accuracy improve with increased model size and with the number of examples, until this number reaches 16.

Based on these results we move forward with the XGLM 7.5B and BLOOM 175B models and 16 examples.

F Detailed Scores of Aggregated Results

- Table 5: Detailed scores of same-language prompting on CoCoA-MT (preliminary evaluation).[12]

[9] https://huggingface.co/microsoft/mdeberta-v3-base

[10] BLOOM 175B encountered out-of-memory errors with 32 in-context example using eight 40GB A100 GPUs.

[11] Detailed scores are included in Table 5.

[12] We set maximum output length as 50 tokens in the preliminary evaluation, while we use 100 tokens in the main

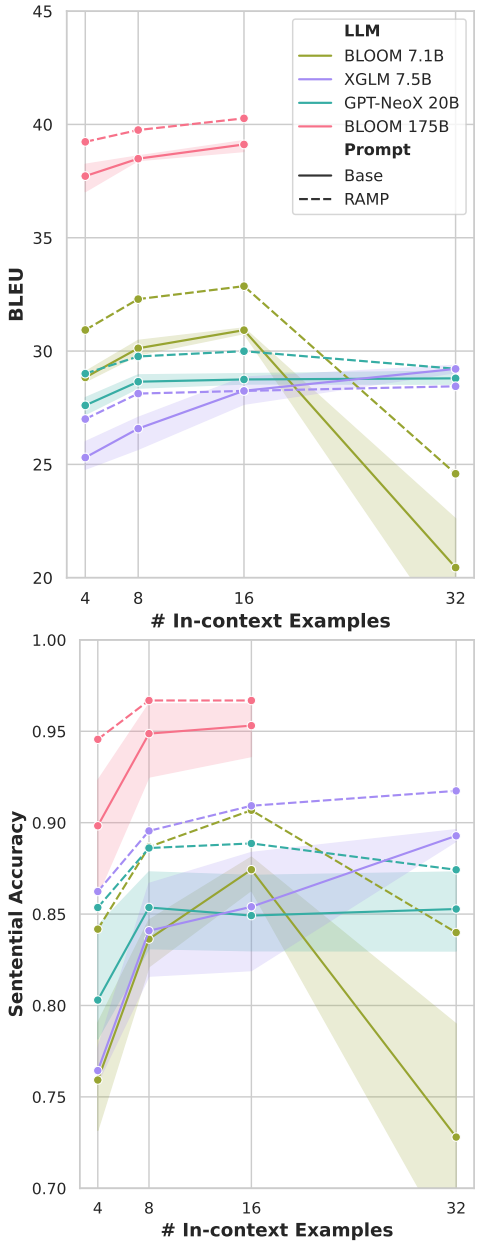

Figure 2: BLEU and sentential formality accuracy of prompt outputs on CoCoA-MT test set for different amounts of in-context examples. Confidence intervals are obtained base setting by sampling in-context examples using 3 seeds.

- Table 6: Decomposed results of same-language prompting on CoCoA-MT (full evaluation).
- Table 7: Decomposed results of same-language prompting on MT-GenEval (full evaluation).
- Table 8: Decomposed results of cross-lingual prompting on CoCoA-MT.
- Table 9: Decomposed results of cross-lingual prompting on MT-GenEval.

evaluation. Early truncating leads to slightly lower scores in Table 5 than in Table 4.

G Amended Details of Cross-Lingual Prompting

We test the zero-shot setting using the leave-one-out strategy, i.e. we retrieve in-context examples from every languages except the desired language of translation. We ensure that we retrieve an equal number of examples from all languages: the number of examples retrieved from each language is the total desired number of in-context examples divided by number of training languages. In CoCoA-MT, we retrieve 14 in-context examples from 7 languages. In MT-GenEval, we retrieve 8 in-context examples from 8 languages. We reduced the number of in-context examples in this setting to avoid out-of-memory errors with BLOOM 175B.

H Error Analysis of Cross-Lingual Prompting

Table 10 shows two examples where RAMP performs significantly worse than the base model in terms of COMET. In the first example, having multiple in-context examples containing *"million"* led the model to mis-translate *"billion"* to *"million"*. In the second example, we observe that the color related in-context examples led the model to produce hallucinated output about clothing colors.

Repeated misleading in-context examples are less observed on MT-GenEval and in the same-language setting because (1) CoCoA-MT translates the same set of English sentences to different languages while MT-GenEval collects English sentences independently; (2) There are no duplicated source (English) sentences for each language. (Therefore, if RAMP retrieves duplicated English sentences as in Table 10, their reference translations are guaranteed to be in different languages.)

		BLEU					COMET					Sentential Accuracy				
		0	**4**	**8**	**16**	**32**	**0**	**4**	**8**	**16**	**32**	**0**	**4**	**8**	**16**	**32**
BLOOM	base	21.8	28.8	30.1	30.9	20.5	0.162	0.578	0.594	0.603	-0.092	0.558	0.759	0.836	0.875	0.728
7.1B	RAMP		30.9	32.3	32.9	24.6		0.597	0.613	0.621	0.150		0.842	0.887	0.907	0.840
XGLM	base	11.8	25.3	26.6	28.3	29.2	-0.534	0.443	0.449	0.499	0.517	0.524	0.764	0.841	0.854	0.893
7.5B	RAMP		27.0	28.1	28.2	29.5		0.450	0.480	0.474	0.484		0.862	0.896	0.909	0.918
GPT-	base	22.7	27.6	28.7	28.8	28.8	0.108	0.268	0.272	0.272	0.275	0.559	0.803	0.854	0.849	0.953
NEOX 20B	RAMP		29.0	29.8	30.0	29.2		0.284	0.310	0.307	0.284		0.854	0.886	0.889	0.874
BLOOM	base	29.9	37.7	38.5	39.1	–	0.476	0.731	0.744	0.750	–	0.612	0.898	0.949	0.953	–
175B	RAMP		39.2	39.75	40.3	–		0.740	0.744	0.761	–		0.946	0.967	0.967	–

Table 5: Detailed scores of same-language prompting on COCOA-MT (preliminary evaluation). Numbers in the header represent the number of in-context examples used for prompting, including zero-shot prompting (0). Scores are averaged across two available formality values (formal, informal) and languages (ES,FR,HI,PT).

			ES		FR		HI		PT		
			F	I	F	I	F	I	F	I	AVG
XGLM 7.5B	base	BLEU	30.1	33.0	30.7	28.8	18.5	16.9	35.7	35.4	28.6
		COMET	0.500	0.527	0.348	0.350	0.454	0.425	0.547	0.554	0.463
		L-Acc	0.524	0.966	0.977	0.633	0.976	0.744	0.931	0.928	0.835
		S-Acc	0.507	0.958	0.953	0.840	0.963	0.748	0.888	0.912	0.846
	+mark	BLEU	31.0	33.2	29.4	27.4	19.2	18.6	35.7	35.5	28.7
		COMET	0.498	0.541	0.207	0.188	0.439	0.409	0.552	0.552	0.423
		L-Acc	0.728	0.972	0.985	0.923	0.986	0.860	0.960	0.947	0.920
		S-Acc	0.697	0.958	0.963	0.917	0.983	0.838	0.927	0.937	0.902
	RAMP	BLEU	32.8	33.5	32.7	31.0	21.0	20.3	34.2	34.4	30.0
		COMET	0.480	0.511	0.314	0.302	0.502	0.491	0.488	0.522	0.451
		L-Acc	0.842	0.963	0.989	0.926	0.993	0.885	0.961	0.943	0.938
		S-Acc	0.803	0.952	0.975	0.922	0.98	0.873	0.928	0.948	0.923
BLOOM 175B	base	BLEU	44.3	45.0	42.9	41.0	27.1	25.8	47.3	45.7	39.9
		COMET	0.728	0.759	0.611	0.600	0.673	0.645	0.762	0.750	0.691
		L-Acc	0.795	0.96032	0.987	0.890	0.978	0.885	0.987	0.954	0.930
		S-Acc	0.889	0.963	0.987	0.888	0.980	0.863	0.987	0.960	0.940
	+mark	BLEU	45.8	44.5	43.3	41.8	28.4	27.1	46.4	45.3	40.3
		COMET	0.726	0.745	0.610	0.594	0.677	0.659	0.751	0.745	0.688
		L-Acc	0.930	0.987	0.996	0.958	0.995	0.936	0.989	0.972	0.970
		S-Acc	0.942	0.985	0.992	0.957	0.992	0.925	0.990	0.977	0.970
	RAMP	BLEU	46.4	46.2	43.9	42.9	30.8	29.2	48.8	47.4	41.9
		COMET	0.718	0.759	0.611	0.610	0.721	0.713	0.782	0.771	0.711
		L-Acc	0.956	0.984	0.998	0.952	0.991	0.947	0.993	0.962	0.973
		S-Acc	0.957	0.982	0.995	0.945	0.993	0.935	0.990	0.967	0.970
Adapted MT		BLEU	44.4	43.7	43.4	37.8	19.1	17.0	53.0	49.9	38.5
		COMET	0.712	0.724	0.559	0.547	-0.191	-0.263	0.783	0.764	0.454
		L-Acc	0.697	0.598	0.822	0.377	0.869	0.449	0.972	0.744	0.691
		S-Acc	0.700	0.600	0.810	0.400	0.680	0.600	0.950	0.800	0.693

Table 6: Decomposed results of same-language prompting on COCOA-MT (full evaluation).

			AR		ES		FR		HI		PT		
			F	M	F	M	F	M	F	M	F	M	AVG
XGLM 7.5B	base	BLEU	7.6	7.5	35.5	38.2	27.1	28.6	13.8	16.4	29.2	33.1	23.7
		COMET	-0.040	-0.012	0.694	0.738	0.509	0.555	0.304	0.332	0.661	0.713	0.445
		L-Acc	0.848	0.947	0.688	0.808	0.715	0.880	0.585	0.956	0.621	0.855	0.790
		S-Acc	0.617	0.866	0.651	0.938	0.581	0.920	0.303	0.962	0.494	0.934	0.727
	+mark	BLEU	7.7	7.8	35.4	38.2	27.5	28.7	14.0	16.7	29.1	32.4	23.7
		COMET	-0.038	-0.020	0.704	0.735	0.508	0.556	0.300	0.317	0.663	0.714	0.444
		L-Acc	0.868	0.939	0.665	0.811	0.701	0.881	0.581	0.955	0.626	0.860	0.789
		S-Acc	0.664	0.856	0.612	0.937	0.562	0.919	0.355	0.966	0.519	0.927	0.732
	RAMP	BLEU	9.2	8.8	37.5	39.4	27.5	29.2	14.8	16.6	31.4	33.3	24.8
		COMET	0.037	0.043	0.723	0.759	0.528	0.571	0.325	0.337	0.681	0.723	0.473
		L-Acc	0.939	0.961	0.750	0.806	0.781	0.885	0.667	0.956	0.759	0.854	0.836
		S-Acc	0.836	0.901	0.722	0.936	0.716	0.937	0.509	0.974	0.729	0.940	0.820
BLOOM 175B	base	BLEU	14.8	16.9	45.6	50.3	38.1	41.7	20.8	24.6	37.6	42.2	33.3
		COMET	0.282	0.395	0.837	0.892	0.719	0.770	0.599	0.629	0.807	0.861	0.679
		L-Acc	0.665	0.966	0.578	0.814	0.660	0.902	0.480	0.951	0.594	0.872	0.748
		S-Acc	0.411	0.934	0.515	0.965	0.581	0.961	0.212	0.973	0.525	0.960	0.704
	+mark	BLEU	15.2	17.1	45.8	50.0	37.9	41.3	20.3	23.8	37.6	42.2	33.1
		COMET	0.294	0.387	0.843	0.887	0.712	0.767	0.576	0.606	0.807	0.861	0.674
		L-Acc	0.707	0.969	0.610	0.818	0.663	0.902	0.493	0.958	0.594	0.872	0.759
		S-Acc	0.482	0.936	0.568	0.973	0.588	0.962	0.284	0.974	0.525	0.960	0.725
	RAMP	BLEU	16.7	17.6	47.9	50.2	39.5	41.8	22.2	25.0	39.3	42.7	34.3
		COMET	0.358	0.407	0.860	0.895	0.734	0.787	0.632	0.646	0.810	0.858	0.699
		L-Acc	0.841	0.972	0.709	0.809	0.765	0.906	0.633	0.953	0.701	0.886	0.817
		S-Acc	0.721	0.940	0.707	0.964	0.732	0.971	0.518	0.973	0.683	0.972	0.818
Adapted MT		BLEU	23.3	24.4	53.2	54.2	44.2	46.4	29.3	32.3	43.4	45.7	35.9
		COMET	0.496	0.522	0.876	0.902	0.759	0.797	0.722	0.743	0.825	0.857	0.528
		L-Acc	0.910	0.981	0.932	0.921	0.919	0.956	0.762	0.837	0.922	0.961	0.853
		S-Acc	0.940	0.970	0.910	0.960	0.950	0.960	0.280	0.750	0.930	0.990	0.863

Table 7: Decomposed results of same-language prompting on MT-GENEVAL (full evaluation).

			ES		FR		HI		PT		
			F	I	F	I	F	I	F	I	AVG
BLOOM 175B	base	BLEU	40.9	46.3	33.7	32.0	21.8	18.9	33.9	29.0	32.1
		COMET	0.785	0.823	0.611	0.615	0.409	0.436	0.772	0.705	0.644
		L-Acc	0.211	0.990	0.899	0.656	0.944	0.123	0.704	0.010	0.567
		S-Acc	0.200	0.930	0.880	0.715	0.940	0.100	0.975	0.025	0.596
	RAMP	BLEU	39.4	44.6	35.3	34.7	22.4	18.4	32.2	27.5	31.8
		COMET	0.749	0.788	0.575	0.614	0.488	0.480	0.770	0.702	0.646
		L-Acc	0.169	0.978	0.949	0.770	0.973	0.143	1.000	0.015	0.625
		S-Acc	0.175	0.950	0.930	0.790	0.975	0.140	0.975	0.040	0.622

Table 8: Decomposed results of cross-lingual prompting on COCOA-MT.

			AR		ES		FR		HI		PT		
			F	M	F	M	F	M	F	M	F	M	AVG
BLOOM 175B	base	BLEU	10.6	11.6	43.3	47.4	34.2	38.2	11.4	15.0	34.4	38.6	28.5
		COMET	0.071	0.138	0.805	0.857	0.648	0.719	-0.135	-0.003	0.766	0.822	0.469
		L-Acc	0.843	0.956	0.627	0.810	0.561	0.899	0.653	0.962	0.588	0.874	0.777
		S-Acc	0.541	0.785	0.529	0.936	0.389	0.944	0.051	0.745	0.475	0.939	0.633
	RAMP	BLEU	10.0	10.5	44.6	47.8	35.7	39.1	13.9	16.6	36.0	39.4	29.4
		COMET	-0.044	0.020	0.818	0.860	0.686	0.739	0.139	0.212	0.779	0.816	0.502
		L-Acc	0.845	0.956	0.660	0.815	0.608	0.900	0.574	0.961	0.680	0.882	0.788
		S-Acc	0.479	0.703	0.605	0.953	0.497	0.956	0.105	0.870	0.613	0.951	0.673

Table 9: Decomposed results of cross-lingual prompting on MT-GENEVAL.

In-context examples (EN)	1.	**Maybe he should. What did you think about that guy findin 3 million dollars worth of old baseball cards in his grandpas attic.**
	2.	**Yeah that makes sense, did you heard about the $10 million bunker he has?**
	3.	I have. I heard that he started a library in 1895 with 32,000 books in it. All from his personal collection. Can you imagine?
	4.	**Yeah that makes sense, did you heard about the $10 million bunker he has?**
	5.	**Yeah that makes sense, did you heard about the $10 million bunker he has?**
	6.	**Maybe he should. What did you think about that guy findin 3 million dollars worth of old baseball cards in his grandpas attic.**
	7.	That is really expensive I agree, did you watch the Lego Batman movie?
	8.	**Yeah that makes sense, did you heard about the $10 million bunker he has?**
	9.	That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office
	10.	That is really expensive I agree, did you watch the Lego Batman movie?
	11.	That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office
	12.	That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office
	13.	He doesnt look like he has 56 years! I heard he made 75000000 from Mission Impossible 3
	14.	Really? I guess he made a valuable contribution to science and also to medicine, did you hear of that species of flying snakes
Input (EN)		If you got it why not? He is worth over 20 billion dollars after all.
Reference (ES)		Si lo tiene, ¿por qué no? Al fin y al cabo, vale más de 20 000 millones de dólares.
RAMP (ES)		Si lo tienes, ¿por qué no? Es millonario después de todo.
base (ES)		Si lo tienes, ¿por qué no? Él vale más de 20 mil millones de dólares después de todo.
In-context examples (EN)	1.	thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?
	2.	For sure lol, it was so nice talking with you, say hi to your cats for me!
	3.	thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?
	4.	**What can't dogs do! I know they aren't color blind like we were taught when young. It was so nice chatting with you, tell yuki hi!**
	5.	For sure lol, it was so nice talking with you, say hi to your cats for me!
	6.	**Oh yeah. That sucker was mean let me tell you. I think I would have a rabid dog chase me, which by the way dogs are not color blind!**
	7.	**Thanks you too! And watch out for dogs, they're not actually colorblind but they just don't see as many colors as a normal human can**
	8.	LOL That's funny! Google prefers dogs over cats! The wrote in their code of conduct that they are a dog company. It's been fun chatting with you!
	9.	Such gorgeous dogs! Do you & your dogs live in a northern climate?
	10.	LOL That's funny! Google prefers dogs over cats! The wrote in their code of conduct that they are a dog company. It's been fun chatting with you!
	11.	thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?
	12.	Such gorgeous dogs! Do you & your dogs live in a northern climate?
	13.	**haha me too! I heard that they wore clothes that were full of color. Surprised me because I always thought of them as wearing black and white.**
	14.	For sure lol, it was so nice talking with you, say hi to your cats for me!
Input (EN)		lol I had an African grey that could say everything I said around the house. Like clean your room! as well any loud sounds.
Reference (PT)		"lol Tinha um papagaio-cinzento que conseguia dizer tudo o que ouvia em casa. Tipo ""Limpe o seu quarto!"" e também quaisquer sons altos"
RAMP (PT)		haha eu também! Eu ouvi dizer que eles usam roupas de todas as cores. Surpreendeu-me porque eu sempre pensei neles usando preto e branco.
base (PT)		hahaha eu tinha um papagaio cinza africano que dizia tudo o que eu dizia em casa. Como limpar o quarto! Bem como qualquer som alto.

Table 10: Examples of CoCoA-MT (formal) where RAMP performs worse than the base model in cross-lingual zero-shot setting. Potentially problematic in-context examples leading to mistranslations or hallucinations are highlighted.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.