Anti-LM Decoding for Zero-shot In-context Machine Translation

Suzanna Sia **Alexandra DeLucia** **Kevin Duh**
Department of Computer Science
Johns Hopkins University
Baltimore, MD, USA
{ssia1, aadelucia}@jhu.edu, kevinduh@cs.jhu.edu

Abstract

Zero-shot In-context learning is the phenomenon where models can perform a task given only the instructions. However, pretrained large language models are known to be poorly calibrated for zero-shot tasks. One of the most effective approaches to handling this bias is to adopt a contrastive decoding objective, which accounts for the prior probability of generating the next token by conditioning on a context. This work introduces an Anti-Language Model objective with a decay factor designed to address the weaknesses of In-context Machine Translation. We conduct our experiments across 3 model types and sizes, 3 language directions, and for both greedy decoding and beam search. The proposed method outperforms other state-of-the-art decoding objectives, with up to 20 BLEU point improvement from the default objective in some settings.

https://github.com/suzyahyah/icl_Anti-LM_decoding

1 Introduction

Decoding strategies in supervised neural machine translation (NMT) models are typically designed to select the maximum likelihood response under the given model. However, in the era of in-context learning with large language models (LLMs), selecting for the maximum likelihood response should be re-examined, as LLMs are known to be poorly calibrated for their tasks and exhibit a strong prior bias (Zhao et al., 2021; Wang et al., 2023). There are two large classes of decoding strategies: sampling methods and decoding objectives. Sampling methods do not change the probability ranking of the next token but influence how it is sampled, such as beam search (Koehn, 2004; Freitag and Al-Onaizan, 2017), nucleus sampling (Holtzman et al., 2020), and top-k sampling (Fan et al., 2018). Decoding objectives modify the probability of the next token before sampling takes place, typically by adding or subtracting scores, i.e., a *contrastive objective*.

Decoding objectives are one of the most effective approaches for improving the output of a generative model. These often require no training and are typically "frustratingly simple", because they only involve manipulation of output probability distributions at inference time. However, as outlined by Zarrieß et al. (2021) in their recent large-scale survey, even for task-specific generation models, it is still surprisingly unclear what a good objective is for natural language generation.

In this work, we propose *an anti-language model (Anti-LM) decoding objective with exponential decay*, which is motivated by the observation that LLMs are performing Bayesian inference under the hood (Xie et al., 2022; Mirchandani et al., 2023), and are thus inclined to continue generating in the source language. We investigate this under a zero-shot setting where only the instructions are provided to the model without any examples.[1]

We hypothesise that poor translations may be due to a strong prior bias of the dominant source language, but that this bias should diminish over future decoding steps. Anti-LM modifies the original logits by taking the difference of the next token logits, conditioned on the source sentence to be translated. Penalising the conditional source sentence logits discourages the model from continuing the non-translated generation from the source sentence or regurgitating it. We consider this negative scenario as a "failure to translate".

Our work falls under the category of contrastive objectives, which were popularised by Li et al. (2016). We compare our approach against other contrastive objectives: "conditional domain PMI" (Holtzman et al., 2021) and "context-aware" decoding (Shi et al., 2023). Our method consistently

[1] An example of the complete input to the model is "*Translate English to French: English: He built a WiFi door bell, he said. French:*".

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3403–3420

outperforms competitive baselines across language directions and model sizes, and the default objective by up to 20 BLEU points. Analysis shows that most of the gains come from the "failure to translate" case which the method was designed for.

In addition, compared to the other contrastive decoding objectives, we only need to compute the contrastive logits (to be subtracted) once for the source sentence and not at every time step.

2 Related Work and Background

Zero-shot MT. Prior work on zero-shot MT has focused on error analysis for the "off-target" problem (Tan and Monz, 2023; Chen et al., 2023),[2] and techniques to improve translation performance (Gu et al., 2019; Chen et al., 2023; Wen et al., 2024).

Regarding error analysis, Tan and Monz (2023) found that performance varies with respect to language direction, vocabulary overlap, and linguistic properties, and Chen et al. (2023) identified lexical similarity between source and target language as an issue. To combat the off-target problem, Chen et al. (2023) introduced Language Aware Vocabulary Sharing (LAVS) to add language-specific tokens and decrease lexical similarity. Wen et al. (2024), similar to our work, use a decoding method to improve zero-shot performance. They introduced EBBS (Ensemble with Bi-level Beam Search), where multiple "components" influence the final generation. Other work uses a third language as a "pivot" (Gu et al., 2019). Unlike these prior works, our method does not involve training new tokenizers, adding linear complexity with ensembling, or translating through other languages.

Contrastive Decoding methods have been extensively explored for text generation, with different motivations behind each method. They have been used to reduce toxic language (Liu et al., 2021), improve general quality without further training (Li et al., 2023), improve factuality (Chuang et al., 2024), and reduce repetition (Yang et al., 2023). For summarization, van der Poel et al. (2022) used conditional PMI decoding to avoid model hallucinations and promote "faithfulness". Shi et al. (2023), Holtzman et al. (2021) and Kumar (2022) adopt a weighted PMI based objective, conditioned on the context to "penalise surface forms". The key difference between our Anti-LM formulation and other prior work, is that we compute the contrastive logits directly on x, the test source sentences to be translated, and not other "non-x context".

Our approach is motivated by improving the decoding of the target language by penalising source language continuations in Zero-shot MT. Within MT, concurrent work by Sennrich et al. (2024) also introduces a similar concept of "language contrastive decoding", but under a different formulation where they recompute the contrastive logits at each time step. In contrast, our method does not require recomputing the logits at every time step, which greatly speeds up inference.

Similarly, sampling methods are also designed for different purposes. Nucleus sampling is good for creative generation (DeLucia et al., 2021) while beam search is a popular choice for MT (Roberts et al., 2020). Decoding objectives can work in tandem with sampling methods but may have unexpected effects due to modification of the output probability space. We thus evaluate our objective with both Greedy Decoding and Beam Search.

3 Method

3.1 Problem Formulation

Let x refer to the source test sentence, y to the target test sentence to be generated, and u to the instructions provided as context to the model. Autoregressive LMs generate text by a series of next-token predictions conditioned on the partial sequence generated so far. Greedy decoding proceeds by sampling the argmax token y_t at every step t, given the previously sampled tokens $y_{<t}$, the test source sentence x and the instructions u based on the decoding objective $\log p(y_t|y_{<t}, x, u)$. Table 1 summarises the evaluated contrastive objectives.

3.2 PMI Decoding (Previous Work)

An intuitive formulation of contrastive decoding is Pointwise Mutual Information (PMI), where $\mathrm{PMI}(\mathbf{y};\mathbf{x})$ measures the association between the target sequence y and source sequence x. $\mathrm{PMI}(y;x)$ can be written as [3]

$$\begin{aligned}\mathrm{PMI}(y;x) &= \log \frac{p(x,y)}{p(x)p(y)} \\ &= \log p(y|x) - \log p(y)\end{aligned}$$

[2] Prior work defines "off-target" as the phenomena where the model returns a "translation" in the wrong language. In this work, we refer to both off-target and "empty" model generations as "failure to translate".

[3] Both $\log[\frac{p(y|x)}{p(y)}]$ and $\log[\frac{p(x|y)}{p(x)}]$ are equivalent forms. However $p(y|x)$ is more natural for autoregressive generation.

Name	RHS Expression	Example Conditional Text Input at $t = 5$	Note
ALMu	$\gamma^t \log p(y_1\|u)$	Translate from English to French:	Ablation with u
ALMx	$\gamma^t \log p(y_1\|x)$	In summer, you'll need to watch out for mosquitoes.	Our Method
PMIu	$\alpha \log p(y_t\|y_{<t}, u)$	Translate from English to French: En ete, il faudra	Holtzman et al. (2021)
PMIx	$\alpha \log p(y_t\|y_{<t}, x)$	In summer, you'll need to watch out for mosquitoes. En ete, il faudra	Shi et al. (2023)

Table 1: The four contrastive objectives evaluated in this work. The example shows the conditional input values for the following instruction (u), source sentence (x), and model generation ($y_{<t}$) at timestep $t = 5$: *Translate from English to French: English: In summer, you'll need to watch out for mosquitoes. French:* En ete, il faudra. PMI(u) and PMI(x) are shorthand for PMI($y; x|u$) and PMI($y; u|x$) respectively.

In PMI based objectives, the second term of Equation (1) functions as an anti-language model, and is typically weighted by $\alpha \in [0, 1]$.

$$\hat{y} = \text{argmax}_y \log p(y|x) - \alpha \log p(y) \quad (1)$$

PMI-based decoding (also known as Maximum Mutual Information Decoding Li et al. (2016)) and its variants (Holtzman et al., 2021; Kumar, 2022; Nandwani et al., 2023) have been widely adopted in neural text generation. It penalizes high-frequency generic responses, but may also penalise fluent ones and thus can lead to ungrammatical outputs.

Conditional PMI Decoding PMI can also be interpreted as penalising the "surface form" (Holtzman et al., 2021) of the target sequence, without having seen the source sequence in the context.

$$\log p(y_t|y_{<t}, x, u) - \alpha \log p(y_t|y_{<t}, u) \quad (2)$$

The objective contains a penalty term for the log probability over the next token, conditioned on the target sequence decoded $y_{<t}$, and the context u. In our case the natural choice of u would be the instructions *"Translate <L1> to <L2>."*.

3.3 Anti-LM Contrastive Decoding

We introduce our Anti-LM approach (ALM), which penalises the logits of the next token continuation of x, simply $\log p(y_1|x)$. The key difference between our Anti-LM objective and previous work is that we subtract logits conditioned directly on the test sentences x to be translated, and not other contexts u or any subsequent generations $y_{<t}$. Additionally, we use a discount factor γ^t to reduce the influence of the Anti-LM on future timesteps.

$$\begin{aligned} \text{ALM}(x) = &\log p(y_t|y_{<t}, x, u) \\ &- \gamma^t \log p(y_1|x) \end{aligned} \quad (3)$$

Unlike PMI decoding, the Anti-LM logits only need to be computed once for each source sentence. Note that $\log p(y_1|x)$ ensures that we never subtract the logits of the target language y if there is a "successful" translation. As a control condition, we experiment with the Anti-LM conditioned on u, which has the same context as conditional PMI.

Latency. Previous decoding methods require computation of the contrastive logits at every generation timestep, resulting in an additional time complexity of $O(n)$ where n is the length of the string generated. In contrast the proposed method (regardless of choice of discount factor) is only $O(1)$ as it only needs to compute the contrastive objective once, and makes use of the decay factor.

4 Experiments

Decoding Objectives We evaluate 4 decoding objectives in addition to the default maximum likelihood objective (summarised in Table 1).

Models. We use three models: XGLM (2.9B, 7.5B) (Lin et al., 2022), Bloom (3B, 7B) (Scao et al., 2022) and Llama 2 (7B, 7B-chat) (Touvron et al., 2023b).[4] All models are available on HuggingFace (Wolf et al., 2020), with the latter three having been advertised as "Multilingual Language Models." To our knowledge, there have not been any reports of sentence-level parallel corpora in their training datasets (Appendix A). In other words, these models were not trained with data that explicitly supports the translation task.

Data and Evaluation. We evaluate on the Wikipedia-based FLORES-101 (Goyal et al., 2022) in three bi-directions with English: French (en↔fr), German (en↔de), and Portuguese (en↔pt). As

[4]We also experimented with OPT2.7B (Zhang et al., 2022) but found that its in-context MT abilities were very poor. The RLHF version of Bloom, Bloomz7B reached a suspiciously high BLEU score of 60 and we suspect data leakage during its training.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM 2.9B	18.8	21.1	19.8	21.4	**21.5**	17.9	13.5	13.7	24.3	**26.6**
	XGLM 7.5B	21.7	25.9	25.3	25.5	**26.0**	13.2	20.9	17.4	28.1	**28.2**
	Bloom 3B	28.1	29.2	29.7	28.4	**30.0**	30.4	28.3	30.5	30.2	**34.0**
	Bloom 7B	32.5	32.7	33.3	32.7	**33.9**	34.6	32.8	32.9	34.8	**37.3**
	Llama 7B	37.2	36.6	**37.3**	37.0	36.6	38.1	35.4	34.4	38.6	**38.7**
	Llama-chat 7B	33.9	33.7	34.0	34.2	**34.3**	34.6	32.6	32.6	34.9	**35.2**
en → pt	XGLM 2.9B	9.0	14.7	12.9	18.2	**19.6**	5.9	7.4	14.8	19.8	**23.2**
	XGLM 7.5B	14.4	24.1	21.4	**25.4**	24.7	3.7	14.2	10.2	26.8	**27.0**
	Bloom 3b	29.9	30.3	**30.7**	30.0	30.6	32.1	30.3	31.3	31.7	**33.6**
	Bloom 7B	32.1	**33.0**	32.8	**33.0**	32.8	35.6	34.0	33.7	35.7	**35.8**
	Llama 7B	35.7	35.5	**35.9**	35.4	35.6	36.9	35.2	34.7	36.7	**37.4**
	Llama-chat 7B	32.9	33.0	33.0	33.2	**33.4**	34.0	31.9	31.7	**34.4**	**34.4**
en → de	XGLM 2.9B	12.0	**13.6**	12.7	13.2	13.3	11.9	8.9	8.4	16.0	**17.6**
	XGLM 7.5B	11.7	16.3	15.0	17.5	**17.8**	4.1	10.8	7.9	18.2	**18.5**
	Bloom 3b	3.3	3.9	3.6	3.8	**4.6**	3.5	3.8	3.7	3.7	**5.0**
	Bloom 7B	3.1	**8.2**	8.0	7.9	8.0	7.8	8.8.	7.4	8.1	**9.0**
	Llama 7B	25.5	25.1	**25.6**	25.3	25.5	25.5	24.7	23.8	26.0	**27.1**
	Llama-chat 7B	22.5	22.3	22.5	22.7	**23.2**	23.5	21.6	21.2	**23.7**	23.4

Table 2: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). Scores are reported with SacreBLEU (Post, 2018), where higher is better. "base" refers to default maximum likelihood decoding. The best scores are bolded and scores within 0.2 of the best are underlined. {} → en results are in Appendix Table 4.

these are zero-shot experiments, no separate dataset is required to be used as the prompt bank, and no randomness is associated with prompt selection.

For evaluation, we report BLEU (Papineni et al., 2002) and COMET-22 (Rei et al., 2022), two reference-based automatic metrics. COMET-22 is a neural-based method reported to correlate highly with human judgment (Freitag et al., 2022). We use the SacreBLEU (Post, 2018) implementation of BLEU with default arguments and Unibabel's COMET-22 implementation.[5]

Generation Settings. For the decoding objectives (Table 1), we chose $\alpha = 0.1$ for PMI Decoding and $\gamma = 0.3$ for Anti-LM decoding (see Appendix B) after hyperparameter search on only a single language direction (en → pt) using the dev set, thereafter applying the same α to all experiments. We evaluate on both greedy decoding and beam search (B=5).

Instructions. We provide instructions in the source (L1) language using the instructions "Translate from <L1> to <L2>" and the "masterful translator" prompt by (Reynolds and McDonell, 2021). See Appendix Table 5 for details.

[5] https://github.com/Unbabel/COMET

5 Results

We observe that **Anti-LM objective is best across most objectives, language directions, and sampling strategies** (see Table 2), although this is less pronounced in Llama7B. We find that PMI outperforms the default objective, which is consistent with previously reported work. For beam search, the Anti-LM objective is particularly effective for XGLM with an improvement of BLEU by up to 20 points. Example translations and COMET scores are in Appendix D.

6 Analysis

Failure to Translate. Models may fail to translate the provided sentence due to no generation or generation in the source (L1) language. Even for the "large" multilingual models (XGLM7.5B and Bloom7B), the models still make a sizeable number of such errors (10%-45%). Figure 1 shows the number of translation failures across models for $\text{PMI}(x)$ and Anti-LM(x) for en↔fr against the default (greedy) objective.

We analyse the scores for the non-failure cases and find that there is largely equivalent proportion of sentences which are either better or worse than the baseline (Figure 2). This indicates that the

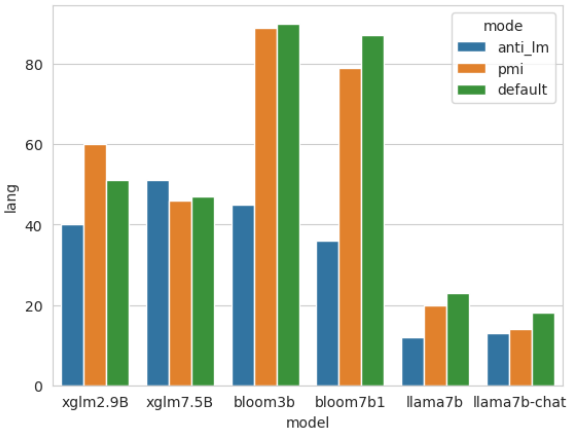

Figure 1: Number of non-target French sentences generated given the task *Translate English to French* which indicates a failure to translate.

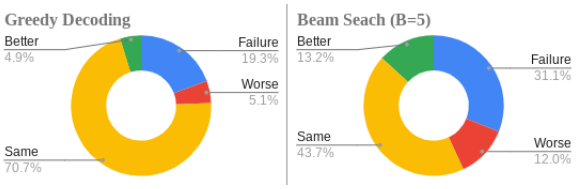

Figure 2: Proportion of failure to translate vs successful cases averaged across all models. For successful cases, we compute whether the Anti-LM objective improves, degrades, or does not affect the performance.

gains observed can be attributed to addressing failure to translate cases (see Appendix D).

Missing Entity Rate. One aspect of translation "faithfulness", considers the named entity retention from the source to the target (Alves et al., 2022). For example, the name "Ehud Ur" should be included as-is in the translation. This is one potential area for improvement of the proposed approach, as the contrastive objective would affect the logits of the "as-is" named entities (see Appendix D.2).

Choice of Discount Factor. The discount factor presented in Equation (3) is an exponential decay in timesteps t. We investigated the importance of the decay function by evaluating others: reverse-ReLU, logistic, and Gompertz decay, which is an asymmetric logistic function. Details of these functions are in Appendix E. We found that Gompertz decay and reverse-ReLU can sometimes outperform exponential, although their performances are quite similar (Table 3).

Instruction Language. Anti-LM has a positive effect in the L1→ L2 direction, *if* the instructions were given in the L1 (source) language. Our findings indicate that there is an unintended effect of source language dominance during zero-shot MT.[6] This suggests that without taking into account the Anti-LM calibration, the *true* zero-shot capabilities of GPT-style models may be under-reported.

		base	exp	r-relu	log	gomp
en → fr	XGLM 2.9B	19.1	**21.3**	21.0	20.6	21.0
	XGLM 7.5B	21.7	26.0	26.2	25.2	**26.4**
	Bloom 3B	28.2	**29.9**	29.8	29.1	**29.9**
	Bloom 7B	32.4	**33.8**	33.6	32.9	33.7
	Llama 7B	**37.2**	36.6	36.1	35.7	36.3
	Llama-chat 7B	33.9	34.3	34.5	**34.8**	34.4
en → pt	XGLM 2.9B	8.9	19.3	**20.0**	19.0	19.6
	XGLM 7.5B	14.3	25.1	**26.2**	24.7	25.6
	Bloom 3B	29.9	**30.7**	30.4	29.4	30.5
	Bloom 7B	32.1	**32.9**	32.7	31.9	32.8
	Llama 7B	**35.7**	35.6	**35.7**	35.1	35.4
	Llama-chat 7B	32.9	33.4	**33.6**	33.5	**33.6**
en → de	XGLM 2.9B	12.0	**13.6**	13.3	12.5	**13.6**
	XGLM 7.5B	11.8	17.7	17.5	16.8	**17.9**
	Bloom 3B	3.3	**4.5**	**4.5**	4.3	**4.5**
	Bloom 7B	7.3	**8.0**	7.6	7.5	7.8
	Llama 7B	**25.5**	**25.5**	25.3	23.7	25.3
	Llama-chat 7B	22.5	**23.2**	**23.2**	23.1	23.1

Table 3: Translation performance on FLORES with greedy decoding using different decay functions. These are the exponential (exp) as shown in eq 3, reverse-relu (relu), logistic (log), and gompertz decay (gomp). The best scores are bolded and scores within 0.2 of the best are underlined.

Elaborate Instructions. Anti-LM similarly outperforms the baseline and comparisons in an experiment with more elaborate instructions, specifically the "masterful translator" prompt by Reynolds and McDonell (2021). See Appendix F and Table 5 for details and results.

7 Conclusion

Decoding objectives are one of the most effective ways to improve a model's output, especially if it has strong prior bias from pre-training. We designed an Anti-LM objective with decay for zero-shot Machine Translation which has a much smaller computational overhead and is more effective than existing approaches. Our method outperforms strong baselines across language directions, model types and sizes, and decoding strategies, especially in failure to translate cases.

[6] We find that the approach is also effective for translating from other languages, e.g., French to Portuguese.

8 Limitations

Comparison to Few-shot. The approach described in this work while effective in the zero-shot setting was found to be less effective for K-shot examples setting. We did not tune the hyperparameter for this setting or investigate this thoroughly. The primary reason is that the K-shot examples has much less failure to translate cases, i.e., more consistent at giving an appropriate translation in the target language.

Low-resource Languages. We do not evaluate our method on 'low-resource' languages. However, what the MT community traditionally considers as 'low-resource' language is a misnomer when working with pre-trained language models, as a language might be 'low-resource' for the model if it is not explicitly collected in the training data. An example of this is German (de) for Bloom. While traditionally considered a 'high-resource' language, it is actually a 'low-resource' language for Bloom as it was not collected in the dataset.[7]

Human Evaluation. While we do evaluate with COMET-22, a metric well-correlated with human judgment, we did not include a human annotation study for the generations.

9 Ethics Statement

In the course of LLM generation, there may be unexpected outputs. The generations of our method may have hallucinated content and can be misleading. When deployed in real-world applications, special attention should be paid to avoid inappropriate generations. For example, one can use post-process steps such as fact-checking for named entities. With regard to toxic or unfair output, we believe that the method does not contribute to these biases that were not already previously present in the pre-trained models.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). 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

[7] https://huggingface.co/bigscience/bloom/discussions/221

benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Boyd Adriane. 2020. spacy: Industrial-strength natural language processing in python.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6*, pages 115–124. Springer.

Sawan Kumar. 2022. Answer-level calibration for free-form multiple choice question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–679.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Suvir Mirchandani, Fei Xia, Pete Florence, brian ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. In *7th Annual Conference on Robot Learning*.

Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi, and Luis Lastras. 2023. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10347, Singapore. Association for Computational Linguistics.

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.

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.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. 2020. Decoding and diversity in machine translation. *arXiv preprint arXiv:2011.13477*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Shaomu Tan and Christof Monz. 2023. Towards a better understanding of variations in zero-shot neural machine translation performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13553–13568, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaosu Wang, Yun Xiong, Beichen Kang, Yao Zhang, Philip S Yu, and Yangyong Zhu. 2023. Reducing negative effects of the biases of language models in zero-shot setting. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 904–912.

Yuqiao Wen, Behzad Shayegh, Chenyang Huang, Yanshuai Cao, and Lili Mou. 2024. Ebbs: An ensemble with bi-level beam search for zero-shot machine translation.

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

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Haoran Yang, Deng Cai, Huayang Li, Wei Bi, Wai Lam, and Shuming Shi. 2023. A frustratingly simple decoding method for neural text generation. *arXiv preprint arXiv:2305.12675*.

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. *Information*, 12(9).

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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

A Models

We ran all experiments on 24GB GeForce RTX 3090 and 32GB Tesla V100 GPUs for the models[8] under and over 7B parameters, respectively.

XGLM adopts a similar architecture to GPT-3 (Brown et al., 2020) and was trained on the large multilingual Common Crawl (CC100-XL, (Conneau et al., 2020)). Bloom has been trained on the ROOTS Corpus (Laurençon et al., 2022), a 1.6 TB collection of HuggingFace text datasets. Llama 2 was trained on an unspecified "new mix of publicly available online data", however, it is 90% English (Touvron et al., 2023b). This mix most likely includes the Llama training set from CommonCrawl, C4, GitHub, Wikipedia, Gutenburg, Books3, ArXiv, and StackExchange, some of which are multilingual (Touvron et al., 2023a).

B Hyperparameter Selection

B.1 Hyperparameter Sweep

The only hyperparameters associated with our experiments are α on PMI Decoding and γ for the discount factor on Anti-LM Decoding. We experiment with $\{-0.1, 0.1, 0.3, 0.5, 0.8, 1.0\}$ for both α and γ and observe that the best α is 0.1 and the best γ is 0.3 across models. Figure 3 is an example graph of the hyperparameter sweep for γ.

Note that we only search for the hyper-parameter once on the **dev set** of en → pt, and use the same hyper-parameter throughout all experiments. i.e., We did not tune the hyperparameter for every single language direction. Note also that the hyperparameters found generalises *across* models. We do not perform hyper-parameter search with Llama models and adopt the same hyperparameter that was found with other models.

C COMET Results

The COMET results are shown in Table 10. The COMET score ranges from 0 to 1, where a score of 1 is considered a good translation. While the scores appear high, they should be interpreted as a comparative score instead of an absolute score.

Improvements over the baselines are primarily seen with greedy decoding. And as in Table 2, XGLM benefits the most from the calibration offered from contrastive decoding. Also, the ALM objectives more consistently improved over the baselines than the PMIs.

[8] In an earlier version of this paper, we had included GPT-Neo results which show large positive effects in the greedy decoding case, and mixed results in the Beam Search case. We omit GPTNeo in this version due to space.

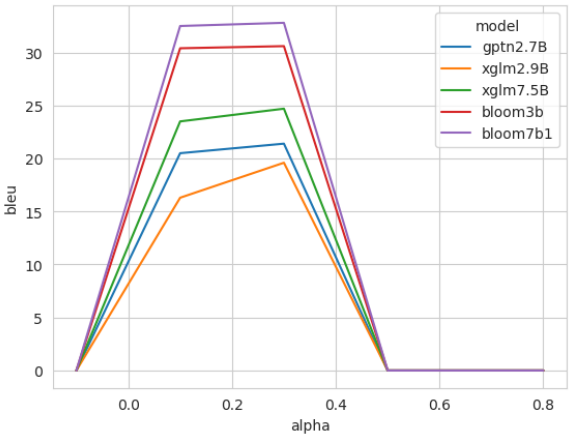

Figure 3: γ sweep for en → pt.

Comparing BLEU and COMET scores For a fair comparison to how we evaluated with BLEU, we did not remove the non-L2 generations. Unlike BLEU, COMET still awards higher than expected scores to "translations" that should be considered failure cases. For example, a model can get scores of 70 and 34 for simply repeating the source sentence (i.e., generating L1 instead of L2) and not generating anything, respectively. These failures are further discussed in Appendix D.

D Failure Analysis

We evaluate "failure to translate" in multiple ways. As discussed in Section 6, the most common failure case is when the model generates the L1 (source) language instead of L2 (target). Statistics on how often that occurs across all the methods and models are shown in Table 6.

D.1 Rate of Empty Generation (REG)

Separate from the L1 generation is when the model does not generate a response at all. We refer to this as "empty generation", and the Rate of Empty Generation (REG) is shown in Table 12. Since this measurement is the ratio of the number of empty generations to the number of generations, a score of 0 is best and a score of 100 is very poor. Though generating text does not mean it is correct or in L2, only that there was *some* output from the model.

An interesting note is that the REG of the baselines (i.e., without special decoding objectives) are never 0, which occurs more frequently with PMI and ALM objectives. Regardless of the decoding

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
fr → en	Xglm 2.9B	23.3	29.1	27.4	**30.1**	30.0	19.8	23.1	20.5	**31.4**	31.3
	Xglm 7.5B	24.6	33.1	31.9	34.4	**34.8**	12.0	31.2	26.8	34.8	**35.2**
	Bloom 3B	14.0	27.6	22.0	29.9	**32.7**	26.3	31.6	28.2	27.0	**34.8**
	Bloom 7B	24.7	31.9	29.7	32.0	**37.2**	25.7	36.6	34.1	27.3	**37.6**
	Llama7b	40.7	**41.4**	40.8	40.8	41.0	40.1	39.6	39.6	40.0	**40.4**
	Llama7bc	40.3	**40.5**	40.2	40.2	40.0	**40.3**	40.2	40.1	**40.3**	40.1
pt → en	Xglm 2.9B	29.7	32.3	32.1	33.2	**33.3**	27.1	30.3	29.2	**35.2**	34.6
	Xglm 7.5B	18.7	**38.3**	37.6	37.8	38.2	7.8	31.8	28.8	**39.4**	39.0
	Bloom 3B	19.1	34.9	33.3	28.9	**35.2**	25.1	34.0	31.4	34.9	**38.0**
	Bloom 7B	27.5	36.0	35.5	15.0	**37.4**	29.5	35.0	33.3	29.3	**39.6**
	Llama 7B	38.4	**44.0**	41.2	42.8	**44.0**	42.0	41.6	35.1	42.8	**43.7**
	Llama-chat 7B	43.2	**43.6**	42.8	43.2	43.2	43.2	**43.4**	43.1	43.2	43.3
de → en	Xglm 2.9B	5.3	4.7	5.1	8.8	7.2	2.0	3.8	3.3	2.2	2.2
	Xglm 7.5B	32.3	32.3	32.2	32.5	**33.0**	29.3	31.0	29.3	33.2	**33.8**
	Bloom 3B	7.1	7.2	7.4	7.9	**9.1**	6.4	6.2	6.3	6.7	**7.6**
	Bloom 7B	20.0	18.8	20.2	20.2	**21.1**	18.0	18.5	19.8	18.4	**19.5**
	Llama 7b	39.3	40.4	39.7	39.4	**39.5**	37.6	37.9	36.7	38.4	**38.5**
	Llama 7b-chat	**39.3**	39.2	39.0	39.1	39.1	38.9	**39.2**	39.0	39.0	38.9

Table 4: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). Scores are reported with SacreBLEU (Post, 2018), where higher is better. "base" refers to default maximum likelihood decoding. The best scores are bolded and scores within 0.2 of the best are underlined.

method, the ALM(u) objective has the best REG. From the ALM(u) and PMI(u) scores, it is apparent that conditioning on the instructions (u) reduces the number of empty generations. Overall, greedy decoding produces the highest rate of output from the models.

D.2 Missing Entity Rate (MER)

Another failure case, which can be used as a proxy for translation "faithfulness", considers the named entity retention from the source to the target. For example, a name such as "Ehud Ur" should be included as-is in the translation (see Table 8). We define the Missing Entity Rate (MER) as the ratio of the number of entities that *should* be in the translation (as determined by the reference) out of all entities in the source. We use the `en_core_web_trf` spaCy model to extract entities from the source sentences (Honnibal et al., 2020). The model is penalised for not generating any text, and only source sentences that have at least one detected entity are considered. This metric is similar in spirit to the "deviation in named entities" challenge from SMAUG (Alves et al., 2022). Similar to the trend with REG scores, we see that the contrastive objectives outperform the baselines (Table 11), and the improvements are greater when conditioning on the instructions (u).

E Decay Functions

The following decay functions were evaluated for the analysis in Table 3. The shape of these functions are shown in Figure 4.

- **Gompertz Decay** with parameters $a = 0.3, b = 20, c = 1$.

$$f(t) = a * \exp(-b * \exp(-c * t))$$

- **Exponential** with $a = 0.3$

$$f(t) = a^t$$

- **Logistic** with $a = 0.3, k = 1, t0 = 5$.

$$f(t) = -a/(1 + \exp(-k * (t - t0))) + a$$

- **Rev-Relu**[9] with $a = 0.3$.

$$f(t) = max(0, -a * t + a)$$

[9]This is not an official name.

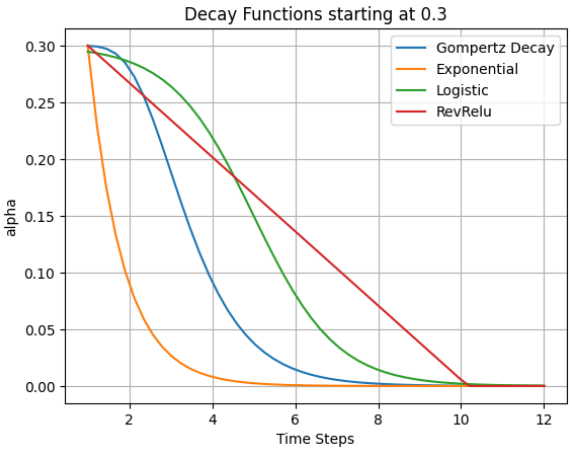

Figure 4: Weighting by different options for the decay functions, starting at 0.3 and ending at 0 in the Equation (1). Aside from exponential, all decay functions have the additional parameter of when the function reaches 0, we set this to 10 timesteps.

F Alternative Instructions

To ensure the results are not prompt-specific, we also ablate with another translation task prompt, the "masterful translator" from Reynolds and McDonell (2021): *"A <L1> phrase is provided. The masterful <L1> translator flawlessly translates the phrase into <L2>:"* (see Table 5). This prompt is more specific than the original, which is expected to cause the model to generate a response with the intended behavior, in this case, an accurate translation. Overall, we observed that while the model improved with a better prompt, the improvement of prompt-tuning *and* ALM decoding was greater. This indicates that zero-shot models can benefit from both prompt-tuning and decoding methods. See below for details for each metric.

BLEU From Table 15, we observe that while the individual scores for each model are higher for over half the models (64%, designated in the table with italics), the same trends as with the basic prompt exist; the ALM decoding methods improve over the baseline in almost all cases and decoding methods. The average ALM(x) improvement over the baseline is higher with the masterful prompt than with the basic prompt, with average gains of +2.52 with greedy search (compared to +0.55) and +5.53 with beam search (compared to +4.77).

COMET The COMET results follow the same trend as the BLEU results, with improving the model performance 58% of the time. The average ALM(x) improvement over the baseline is higher with the masterful prompt than with the basic prompt, with average gains of +4.10 with greedy search (compared to +3.61) and +7.75 with beam search (compared to +6.80).

Missing Entity Rate (MER) The improvements of the masterful prompt over MER are less pronounced as with BLEU and COMET. The model performance is only improved 41% of all cases. The average ALM(x) improvement over the baseline is higher with the masterful prompt than with the basic prompt, with average gains of -5.85 with greedy search (compared to -1.12) and -12.04 with beam search (compared to -11.42).

Rate of Empty Generation (REG) The REG improvements are more pronounced than with MER but still less than BLEU and COMET, with increased model performance 45% of the time. The average ALM(x) improvement over the baseline is higher with the masterful prompt than with the basic prompt, with average gains of -8.44 with greedy search (compared to -2.78) and -14.03 with beam search (compared to -13.67).

Prompt	(en)	Translate {L1} to {L2}: {L1}: {SOURCE} {L2}:
	(de)	Übersetzen Sie vom Deutschen ins Englische: Deutschen: {SOURCE} Englische:
	(fr)	Traduire du français vers l'anglais: français: {SOURCE} l'anglais:
	(pt)	Traduzir português para inglês: português: {SOURCE} inglês:
"Masterful" Prompt		A {L1} phrase is provided. The masterful {L1} translator flawlessly translates the phrase into {L2}: {L1}: {SOURCE} {L2}:

Table 5: The two prompt templates used in the experiments. {L1} refers to the source language and {L2} refers to the target language. {SOURCE} is replaced with the source-language sentence for translation. The prompt is translated into the source language for German (de), French (fr), and Portuguese (pt). The "masterful" prompt is from (Reynolds and McDonell, 2021).

		Non-failure			Failures
sampling	model	better	equal	worse	
Default	XGLM2.9B	0.12	0.78	0.12	0.30
	XGLM7.5B	0.08	0.83	0.08	0.33
	Bloom3B	0.07	0.88	0.05	0.30
	Bloom7B	0.05	0.90	0.05	0.13
	Llama7B	0.08	0.87	0.08	0.03
	Llama7B-chat	0.00	0.97	0.00	0.07
Beam Search	XGLM2.9B	0.22	0.52	0.23	0.43
	XGLM7.5B	0.20	0.57	0.22	0.62
	Bloom3B	0.17	0.62	0.20	0.40
	Bloom7B	0.17	0.62	0.20	0.22
	Llama7B	0.20	0.53	0.22	0.10
	Llama7B-chat	0.10	0.80	0.10	0.07

Table 6: Proportion of better, equal or worse scoring sentences where the difference is at least 5 BLEU points, when comparing the AntiLM approach against the baseline, when excluding 'failure to translate' cases. All values are aggregated across three language directions, en $\rightarrow$ fr, pt, de.

Source: He built a WiFi door bell, he said.
Target: Il dit avoir conçu une sonnette de porte Wi-Fi.

	Objective	Model	Translation: Regular Prompt	Translation: Masterful Prompt
Beam Search (B=5)	ALM(u)	Bloom3B	Il construisit un interphone sans fil, il a dit.	Il construisit une sonnette WiFi, il a dit.
		Bloom7B	Il a construit un interphone sans fil, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un timbre à sonner par WiFi, il a dit.	Il a construit une sonnette WiFi, a-t-il dit.
		Llama2-7BChat	Il a construit un timbre Wi-Fi.	Il a construit un appareil de sonnette Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit une alarme WiFi, il a dit.
		XGLM7.5B	Il a construit un interphone WiFi, il a dit.	He a construit un WiFi porte-clefs, il a dit.
	ALM(x)	Bloom3B	Il a construit un sonnette WiFi, il a dit.	Il construisit un sonnette WiFi, il a dit.
		Bloom7B	Il a construit un interphone sans fil, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un appareil de sonnette Wi-Fi, il a dit.	Il a construit une sonnette Wi-Fi, a-t-il dit.
		Llama2-7BChat	Il a construit un timbre Wi-Fi.	Il a construit un appareil de sonnette Wi-Fi, il a dit.
		XGLM2.9B	Il a construit une WiFi porte-clef, il a dit.	Il a construit une WiFi porte-clés, il a dit.
		XGLM7.5B	Je l'ai construit un WiFi porte-clefs, il a dit.	Il a construit un interphone sans fil, il a dit.
	PMI(u)	Bloom3B	Il a construit un buzzer WiFi, dit-il.	Il a construit un interphone WiFi, il a dit.
		Bloom7B	Il a construit un bouton d'appel WiFi, il a dit.	Il a construit une sonnette WiFi, il a dit.
		Llama2-7B	Il a construit un interphone WiFi, il a dit.	Il a construit une sonnette WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi interphone.	Il a construit un WiFi interphone.
		XGLM7.5B	<EMPTY GENERATION>	Il a construit un interphone WiFi, il dit.
	PMI(x)	Bloom3B	Il a construit un porte-clés WiFi, dit-il.	Il a construit un petit interphone sans fil, il a dit.
		Bloom7B	Il a construit une sonnette WiFi.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un appareil de sonnette WiFi, il a dit.	Il a construit une sonnette WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un appareil de sonnette Wi-Fi, il a dit.
		XGLM2.9B	Il a construit une wifi porte-fenêtre.	Il a construit une alarme WiFi.
		XGLM7.5B	<EMPTY GENERATION>	<EMPTY GENERATION>
	base	Bloom3B	Il a construit un sonnette WiFi, il a dit.	Il a construit une sonnette WiFi, il a dit.
		Bloom7B	Il a construit un interphone sans fil, il a dit.	Il a construit une sonnette WiFi, il a dit.
		Llama2-7B	Il a construit un appareil de sonnette WiFi, il a dit.	Il a construit une sonnette WiFi, a-t-il dit.
		Llama2-7BChat	Il a construit un timbre Wi-Fi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit un WiFi.
		XGLM7.5B	Il a construit un interphone WiFi, il a dit.	Il a construit un WiFi porte-clefs, il a dit.
Greedy Search	ALM(u)	Bloom3B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Bloom7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit un WiFi, il a dit.
		XGLM7.5B	Il a construit un interphone WiFi, il a dit.	He a wifi porte-clefs, il a dit.
	ALM(x)	Bloom3B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Bloom7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un interphone WiFi, il a dit.	Il a fait un interphone WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	<EMPTY GENERATION>	Il a construit un WiFi, il a dit.
		XGLM7.5B	<EMPTY GENERATION>	Il a construit un WiFi porte-clefs, il a dit.
	PMI(u)	Bloom3B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Bloom7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Llama2-7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit un WiFi porte-fenêtre, il a dit.
		XGLM7.5B	Il a construit un interphone WiFi, il a dit.	Il a construit un WiFi porte-clefs, il a dit.
	PMI(x)	Bloom3B	Il a construit un appareil de sonnerie WiFi, il a dit.	Il a construit un interphone sans fil, il a dit.
		Bloom7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un appareil WiFi pour son interphone, il a dit.	Il a fait un interphone WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit un WiFi porte-fenêtre, il a dit.
		XGLM7.5B	Il a construit un WiFi porte-clefs, il a dit.	Il a construit un WiFi porte-clefs, il a dit.
	base	Bloom3B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone WiFi, il a dit.
		Bloom7B	Il a construit un interphone WiFi, il a dit.	Il a construit un interphone sans fil, il a dit.
		Llama2-7B	Il a construit un interphone WiFi, il a dit.	Il a fait un interphone WiFi, il a dit.
		Llama2-7BChat	Il a construit un timbre WiFi.	Il a construit un interphone Wi-Fi, il a dit.
		XGLM2.9B	Il a construit un WiFi, il a dit.	Il a construit un WiFi, il a dit.
		XGLM7.5B	<EMPTY GENERATION>	<EMPTY GENERATION>

Table 7: Example translation of a French sentence from various generation settings and models with the different prompts.

Objective	Model Translation	Missing Entities
Source	Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	
Target	Le Dr Ehud Ur, professeur de médecine à l'Université Dalhousie de Halifax (Nouvelle-Écosse) et président de la division clinique et scientifique de l'Association canadienne du diabète, a averti que la recherche en était encore à ses débuts.	
Base	<EMPTY GENERATION>	Ehud Ur, Halifax
PMI(x)	Dr. Ehud Ur, professeur de médecine à l'Université Dalhousie de Halifax, au Canada et président de la division clinique et scientifique de la Société canadienne du diabète, a souligné que les résultats de cette étude sont encore très précoces.	
ALM(x)	Dr. Ehud Ur, professeur de médecine à l'Université Dalhousie de Halifax, en Nouvelle-Écosse et président de la division de la recherche clinique et scientifique de la Société canadienne du diabète, a déclaré que le travail est encore en cours.	
PMI(u)	Dr. Ehud Ur, professeur de médecine à l'Université Dalhousie de Halifax, en Nouvelle-Écosse et président de la division clinique et scientifique de la Société canadienne du diabète, avertit que le travail est encore dans ses premières étapes.	
ALM(u)	Dr. Ehud Ur, professeur de médecine à l'Université Dalhousie de Halifax, en Nouvelle-Écosse et président de la division de la recherche clinique et scientifique de la Société canadienne du diabète, a déclaré que le travail est encore en cours.	

Table 8: Example translations from English to French across decoding methods for XGLM 7.5B model with greedy search. The empty generation from the default decoding objective is considered a failure to translate.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	66.59	**71.86**	69.98	1.73	71.32	64.75	59.99	62.08	75.07	**75.96**
	XGLM7.5B	68.26	74.82	73.54	74.77	**75.42**	56.49	66.31	63.27	**75.85**	75.47
	Bloom 3B	77.53	**79.67**	9.60	78.72	79.11	0.23	78.59	78.60	79.91	**80.39**
	Bloom 7B	80.98	1.70	**81.79**	1.68	81.03	81.76	81.70	81.19	**82.05**	1.95
	Llama 7B	3.86	3.88	**83.93**	3.77	83.49	84.68	82.99	82.87	4.89	**85.00**
	Llama-chat 7B	82.62	82.51	82.60	**82.88**	2.79	83.52	80.51	81.52	83.46	**83.72**
en → pt	XGLM2.9B	54.05	64.26	63.21	71.66	**73.26**	49.71	50.76	66.00	73.82	**77.22**
	XGLM7.5B	61.15	75.27	71.81	**78.78**	77.98	45.69	60.71	56.25	**78.16**	77.78
	Bloom 3B	82.76	**83.92**	83.52	83.34	83.54	4.25	83.33	83.86	84.02	**84.30**
	Bloom 7B	83.85	**84.89**	84.45	4.76	84.42	5.28	85.17	84.07	**85.41**	84.53
	Llama 7B	5.85	5.91	**86.04**	5.89	5.86	**86.62**	85.45	85.15	86.36	6.60
	Llama-chat 7B	83.49	83.90	83.85	4.21	**84.33**	84.35	82.21	82.40	4.96	**85.02**
en → de	XGLM2.9B	64.91	68.40	67.46	67.67	**68.71**	62.01	57.58	60.97	71.60	**73.17**
	XGLM7.5B	59.22	70.16	67.44	73.60	**74.00**	45.55	58.15	53.88	3.59	**73.63**
	Bloom 3B	53.38	**53.92**	53.48	53.45	46.99	56.39	55.62	51.55	**57.18**	49.09
	Bloom 7B	52.74	**54.49**	53.86	54.00	52.38	56.33	56.75	40.87	**57.63**	54.03
	Llama 7B	81.45	81.30	81.44	81.50	**81.98**	81.70	80.10	80.43	81.81	**83.17**
	Llama-chat 7B	77.94	77.99	78.07	78.64	**79.02**	79.59	75.76	76.91	0.19	**80.31**

Table 9: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). **Scores are reported with COMET-22** (Rei et al., 2022), where higher is better. "base" refers to default maximum likelihood decoding. The best scores are bolded.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	65.18	**70.03**	68.55	69.59	9.97	63.27	*61.07*	58.15	74.24	**75.49**
	XGLM7.5B	*68.62*	*76.30*	*75.43*	*76.46*	***77.25***	56.28	*68.47*	62.82	*77.41*	***79.03***
	Bloom 3B	*79.80*	***80.81***	*80.29*	*79.88*	80.65	*81.81*	*80.67*	*80.49*	*80.97*	***82.58***
	Bloom 7B	*82.77*	***83.66***	*83.45*	*82.97*	*82.76*	***84.62***	*83.76*	*83.33*	*84.08*	*84.23*
	Llama 7B	**83.75**	83.52	3.58	83.24	83.41	*84.98*	82.69	82.24	83.91	***85.37***
	Llama-chat 7B	2.41	82.24	82.37	2.44	**82.61**	83.79	*80.51*	*81.56*	*83.55*	***83.89***
en → pt	XGLM2.9B	52.49	*68.93*	62.44	70.79	**72.53**	*49.86*	*53.15*	48.72	*75.37*	***77.77***
	XGLM7.5B	59.65	*76.81*	71.68	***80.63***	80.54	45.34	*63.58*	52.88	*80.62*	***81.19***
	Bloom 3B	*83.48*	***83.98***	83.82	*83.69*	83.94	85.13	*84.30*	*84.20*	*84.88*	***85.21***
	Bloom 7B	*85.58*	85.79	***85.84***	*85.35*	*85.36*	***86.74***	*86.07*	*86.00*	86.71	*86.51*
	Llama 7B	5.83	85.51	5.87	85.64	***85.96***	*86.71*	85.16	84.81	86.31	***87.19***
	Llama-chat 7B	84.55	84.51	84.44	*84.41*	***84.62***	*85.30*	*83.27*	*82.93*	*85.46*	***85.66***
en → de	XGLM2.9B	*65.13*	68.00	*67.50*	66.78	***69.19***	*63.18*	*58.86*	58.70	71.39	***73.83***
	XGLM7.5B	59.08	*70.94*	*68.29*	*73.84*	***74.76***	45.17	*58.94*	*54.96*	73.16	***75.83***
	Bloom 3B	47.26	48.06	46.70	**48.52**	44.61	52.73	51.06	47.44	**53.33**	48.72
	Bloom 7B	51.27	**54.10**	52.46	53.77	51.94	55.96	54.38	*52.15*	***57.80***	*56.07*
	Llama 7B	81.02	80.68	81.31	80.50	**81.64**	*82.02*	*80.26*	80.31	80.83	***83.45***
	Llama-chat 7B	***78.40***	77.99	78.30	8.37	8.35	*80.35*	*76.50*	76.87	*80.28*	***80.70***

Table 10: Translation performance on FLORES with greedy decoding and beam search ($B = 5$) **with the "masterful" prompt**. **Scores are reported with COMET-22** (Rei et al., 2022), where higher is better. "base" refers to default maximum likelihood decoding. The best scores are bolded. Scores that are better than when using the basic prompt are italicized.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	31.20	**18.26**	25.52	23.35	59.11	38.41	40.88	48.12	**14.38**	16.92
	XGLM7.5B	29.72	**14.53**	20.28	18.45	55.43	53.22	27.41	37.70	18.49	**17.11**
	Bloom 3B	20.58	**13.65**	15.91	15.86	21.90	13.55	**11.27**	16.76	11.79	16.29
	Bloom 7B	13.05	**9.12**	10.69	11.58	17.71	9.42	**7.60**	11.38	9.35	13.04
	Llama 7B	9.71	**7.34**	9.57	9.28	9.59	6.93	**5.98**	9.45	7.48	7.15
	Llama-chat 7B	10.11	**9.35**	9.77	9.54	9.96	9.44	**7.72**	11.62	8.78	9.21
en → pt	XGLM2.9B	52.01	32.37	37.22	**19.95**	20.10	63.81	55.19	41.95	17.18	**15.56**
	XGLM7.5B	47.71	18.47	28.26	**12.39**	18.67	72.63	46.02	56.46	**13.62**	17.05
	Bloom 3B	12.41	9.62	11.75	**9.59**	14.51	**9.08**	9.94	13.53	9.31	12.86
	Bloom 7B	10.17	**8.94**	9.60	9.65	11.15	6.79	**5.62**	10.12	8.80	9.42
	Llama 7B	**3.60**	4.48	3.76	3.70	4.63	3.60	**3.03**	5.40	4.01	6.07
	Llama-chat 7B	6.22	**5.30**	5.68	**5.30**	**5.30**	6.17	6.10	7.41	**5.30**	**5.30**
en → de	XGLM2.9B	26.49	**16.53**	19.32	19.97	21.55	37.05	38.75	47.19	**14.20**	17.41
	XGLM7.5B	43.85	21.29	27.13	**13.59**	15.18	67.14	36.51	53.01	**16.21**	16.58
	Bloom 3B	12.01	**8.67**	10.55	11.41	22.23	7.66	7.94	19.19	**7.06**	19.17
	Bloom 7B	19.64	**12.92**	16.18	16.26	20.99	18.81	**11.39**	20.87	13.82	18.90
	Llama 7B	**5.05**	5.41	6.30	5.11	7.63	5.65	5.51	8.04	**4.64**	6.51
	Llama-chat 7B	9.12	8.47	8.98	7.08	**6.88**	7.83	7.97	9.52	**6.48**	7.01

Table 11: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). Scores are reported for the **missing entity rate (MER)**. A lower score is better. “base" refers to default maximum likelihood decoding. The best scores are bolded.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	11.07	0.89	3.95	**0.00**	46.08	19.86	6.13	9.78	**0.00**	**0.00**
	XGLM7.5B	14.33	0.79	3.66	**0.00**	45.63	36.86	13.83	21.34	**0.10**	0.20
	Bloom 3B	3.26	0.10	0.79	**0.00**	2.45	0.99	0.30	3.56	**0.00**	1.09
	Bloom 7B	1.48	0.10	0.40	**0.00**	2.14	0.40	0.10	1.09	**0.00**	0.89
	Llama 7B	0.10	**0.00**	**0.00**	**0.00**	**0.00**	0.10	**0.00**	0.30	**0.00**	0.10
	Llama-chat 7B	0.40	**0.00**	0.10	**0.00**	**0.00**	0.49	**0.00**	0.30	**0.00**	**0.00**
en → pt	XGLM2.9B	39.53	20.16	19.66	0.20	**0.00**	50.30	26.38	7.41	1.28	**0.69**
	XGLM7.5B	33.20	7.02	13.44	**0.00**	**0.00**	63.44	30.63	41.11	**0.69**	0.99
	Bloom 3B	1.58	0.20	0.89	**0.00**	1.09	0.69	0.59	0.89	**0.10**	0.99
	Bloom 7B	0.79	**0.00**	0.49	**0.00**	0.89	0.49	0.20	0.69	**0.00**	0.79
	Llama 7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.10	**0.00**	0.10
	Llama-chat 7B	1.38	0.20	0.69	**0.00**	**0.00**	1.28	0.20	0.69	**0.00**	**0.00**
en → de	XGLM2.9B	6.82	0.49	2.47	**0.00**	**0.00**	17.00	5.34	7.81	0.10	**0.00**
	XGLM7.5B	30.34	7.11	14.33	**0.00**	**0.00**	58.40	25.59	39.53	0.49	**0.20**
	Bloom 3B	1.98	0.20	0.30	**0.00**	1.28	2.77	1.28	5.14	**0.00**	2.57
	Bloom 7B	4.15	0.20	1.58	**0.00**	2.67	3.95	0.89	1.68	**0.00**	3.75
	Llama 7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.20	**0.00**	0.10	**0.00**	0.30
	Llama-chat 7B	1.88	0.59	1.28	0.10	**0.00**	1.48	0.30	0.89	0.10	**0.00**

Table 12: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). Scores are reported for the **rate of empty generation (REG)**, or how often the model does not produce an output. A lower score is better. “base" refers to default maximum likelihood decoding. The best scores are bolded.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	33.49	**20.78**	28.62	*23.95*	*26.91*	*40.45*	*39.43*	*51.68*	***13.02***	17.67
	XGLM7.5B	***27.16***	***12.32***	*16.66*	*12.70*	*15.09*	*48.74*	*24.56*	*38.71*	*12.76*	***11.23***
	Bloom 3B	*17.13*	***12.99***	*14.64*	*15.47*	*19.93*	*12.78*	***9.19***	*14.26*	*9.70*	*15.49*
	Bloom 7B	13.94	**11.22**	*12.28*	*11.49*	*16.96*	12.00	**9.85**	13.53	10.97	15.69
	Llama 7B	10.16	**9.31**	10.16	9.95	12.23	7.17	6.24	*11.67*	***5.98***	9.35
	Llama-chat 7B	*10.11*	***9.05***	9.92	10.30	*10.49*	*9.40*	*8.92*	*10.99*	*9.28*	***8.85***
en → pt	XGLM2.9B	*60.08*	*25.07*	43.90	**20.35**	20.73	*64.63*	*53.92*	*71.57*	*15.41*	***13.11***
	XGLM7.5B	*48.95*	*16.13*	*27.09*	*10.12*	***9.81***	*73.10*	*41.05*	*59.08*	***11.82***	*13.76*
	Bloom 3B	***10.49***	***9.93***	*10.47*	*9.98*	*13.13*	*9.16*	***8.02***	*12.68*	*8.33*	*10.58*
	Bloom 7B	*10.81*	***8.86***	10.35	11.15	13.20	**7.21**	*7.49*	*9.45*	9.24	11.04
	Llama 7B	6.58	**4.99**	7.00	5.50	7.66	5.04	**3.86**	7.87	*4.73*	*5.30*
	Llama-chat 7B	8.10	6.92	7.59	**6.46**	*6.87*	***5.68***	6.38	10.37	5.94	5.94
en → de	XGLM2.9B	27.03	**17.80**	*21.04*	*18.78*	*22.91*	*31.87*	*37.26*	*45.13*	***8.20***	*15.93*
	XGLM7.5B	*43.32*	*19.15*	*27.27*	***11.29***	*11.54*	69.09	*39.85*	*52.79*	***11.05***	*12.79*
	Bloom 3B	15.47	**13.06**	17.03	13.64	25.62	10.57	*12.32*	***17.07***	***10.33***	*17.83*
	Bloom 7B	26.25	**15.51**	*23.45*	*15.69*	*24.11*	*18.57*	13.72	*23.18*	***10.60***	21.00
	Llama 7B	5.74	*5.80*	***5.63***	6.42	*9.87*	*5.12*	6.07	9.50	**4.96**	7.86
	Llama-chat 7B	***8.00***	8.51	9.29	8.48	10.41	*7.92*	***7.39***	11.99	7.83	8.36

Table 13: Translation performance on FLORES with greedy decoding and beam search ($B = 5$) **with the "masterful" prompt**. Scores are reported for the **missing entity rate (MER)**. A lower score is better. "base" refers to default maximum likelihood decoding. The best scores are bolded. Scores that are better than when using the basic prompt are italicized.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM2.9B	*10.77*	*0.79*	*3.66*	***0.00***	***0.00***	21.05	6.23	10.28	0.10	***0.00***
	XGLM7.5B	16.11	1.38	3.85	***0.00***	***0.00***	41.01	*13.83*	25.79	0.20	***0.00***
	Bloom 3B	*0.40*	***0.00***	*0.20*	***0.00***	*0.49*	*0.20*	***0.00***	*0.49*	***0.00***	*0.40*
	Bloom 7B	*0.79*	***0.00***	*0.20*	***0.00***	*0.89*	***0.00***	*0.10*	*0.40*	***0.00***	*0.69*
	Llama 7B	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***
	Llama-chat 7B	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	*0.20*	***0.00***	***0.00***	***0.00***	***0.00***
en → pt	XGLM2.9B	43.48	*7.21*	21.64	***0.10***	**0.10**	50.59	*22.43*	37.06	*1.19*	***0.20***
	XGLM7.5B	40.42	7.61	17.39	***0.00***	0.20	66.60	31.32	51.28	0.79	***0.49***
	Bloom 3B	*0.40*	***0.00***	*0.20*	***0.00***	*0.40*	**0.00**	**0.00**	*0.20*	***0.00***	*0.49*
	Bloom 7B	***0.00***	***0.00***	***0.00***	***0.00***	*0.79*	*0.10*	***0.00***	*0.10*	***0.00***	*0.20*
	Llama 7B	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***
	Llama-chat 7B	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	*0.20*	***0.00***	***0.00***	***0.00***	***0.00***
en → de	XGLM2.9B	*6.72*	*0.49*	2.67	***0.00***	***0.00***	*14.82*	5.83	8.40	***0.10***	**0.10**
	XGLM7.5B	31.72	8.10	14.72	***0.00***	***0.00***	59.49	28.66	*38.93*	0.69	***0.20***
	Bloom 3B	*0.59*	*0.10*	*0.20*	***0.00***	*0.30*	*1.19*	*0.30*	*1.48*	***0.00***	*0.59*
	Bloom 7B	9.68	1.19	4.84	**0.10**	6.32	5.53	1.28	5.53	***0.00***	5.34
	Llama 7B	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***
	Llama-chat 7B	*0.30*	***0.00***	*0.30*	***0.00***	***0.00***	*0.20*	***0.00***	*0.30*	***0.00***	***0.00***

Table 14: Translation performance on FLORES with greedy decoding and beam search ($B = 5$) **with the "masterful" prompt**. Scores are reported for the **rate of empty generation (REG)**, or how often the model does not produce an output. A lower score is better. "base" refers to default maximum likelihood decoding. The best scores are bolded and scores within 0.2 of the best are underlined. Scores that are better than when using the basic prompt are italicized.

		Greedy					Beam Search (B=5)				
		base	PMI_u	PMI_x	ALM_u	ALM_x	base	PMI_u	PMI_x	ALM_u	ALM_x
en → fr	XGLM 2.9B	17.4	19.9	18.8	19.0	**20.4**	15.5	*13.8*	10.9	22.7	**24.8**
	XGLM 7.5B	*22.5*	*27.2*	*26.6*	*27.4*	**28.1**	*13.9*	*22.3*	*17.9*	*28.7*	**30.8**
	Bloom 3B	*30.3*	**31.5**	*31.3*	*30.1*	*31.0*	*33.5*	*31.6*	*31.5*	*32.9*	**35.3**
	Bloom 7B	*35.3*	*35.7*	**36.4**	*35.6*	*35.3*	**39.1**	*36.5*	*36.7*	*38.6*	*38.6*
	Llama 7B	**36.2**	35.4	35.7	35.2	35.5	*38.7*	34.7	32.4	36.1	**38.7**
	Llama-chat 7B	33.7	33.5	33.7	33.7	**34.1**	*35.4*	32.5	*32.8*	*35.1*	**35.7**
en → pt	XGLM 2.9B	7.1	*15.7*	11.9	15.8	**19.0**	5.1	*7.5*	4.8	18.6	**23.9**
	XGLM 7.5B	13.6	*25.5*	*22.3*	**28.1**	*27.9*	3.5	*16.3*	8.1	*29.7*	**30.8**
	Bloom 3B	*30.3*	*30.6*	*30.7*	*30.6*	**30.8**	*33.9*	*32.0*	*31.5*	*32.9*	**34.5**
	Bloom 7B	**34.3**	*34.2*	*34.3*	*34.0*	*34.1*	*37.4*	*35.3*	*35.4*	**37.6**	*37.1*
	Llama 7B	35.3	34.5	**35.3**	34.8	35.1	37.3	34.1	33.0	36.3	**38.3**
	Llama-chat 7B	33.8	*33.5*	*33.7*	*33.7*	**34.0**	*34.9*	*32.6*	*32.3*	*34.9*	**35.2**
en → de	XGLM 2.9B	11.8	13.1	12.5	12.3	**13.2**	*12.1*	*9.1*	7.6	15.6	**17.3**
	XGLM 7.5B	11.2	*17.0*	*15.2*	*17.6*	**18.4**	*4.1*	*11.6*	*8.2*	17.4	**19.6**
	Bloom 3B	*5.3*	*5.8*	*5.6*	*5.3*	**6.2**	*5.2*	*5.4*	*5.9*	*5.1*	**6.4**
	Bloom 7B	*7.8*	**8.9**	*8.4*	*8.9*	*8.3*	*9.3*	*8.8*	*8.0*	**9.6**	*9.6*
	Llama 7B	24.7	24.3	**25.0**	23.9	24.6	*25.7*	23.8	22.9	24.4	**27.3**
	Llama-chat 7B	**22.2**	21.9	22.1	22.1	22.0	23.2	*21.6*	20.7	23.2	**23.6**

Table 15: Translation performance on FLORES with greedy decoding and beam search ($B = 5$). **Scores are reported with SacreBLEU** (Post, 2018), where higher is better. “base" refers to default maximum likelihood decoding. The best scores are bolded and scores within 0.2 of the best are underlined. The instructions used are "A <L1> phrase is provided. The masterful <L1> translator flawlessly translates the phrase into <L2>.”, a verbose instruction phrase recommended by Reynolds and McDonell (2021).