Mitigating Hallucinations and Off-target Machine Translation with Source-Contrastive and Language-Contrastive Decoding

Rico Sennrich[1,2] **Jannis Vamvas**[1] **Alireza Mohammadshahi**[1,3]
[1]University of Zurich [2]University of Edinburgh [3]EPFL
{sennrich,vamvas}@cl.uzh.ch
alireza.mohammadshahi@epfl.ch

Abstract

Hallucinations and off-target translation remain unsolved problems in MT, especially for low-resource languages and massively multilingual models. In this paper, we introduce two related methods to mitigate these failure cases with a modified decoding objective, without either requiring retraining or external models. In source-contrastive decoding, we search for a translation that is probable given the correct input, but improbable given a random input segment. In language-contrastive decoding, we search for a translation that is probable, but improbable given the wrong language indicator token. Experiments on the massively multilingual models M2M-100 (418M) and SMaLL-100 show that these methods suppress hallucinations and off-target translations, reducing the number of translations with segment-level chrF2 below 10 by 67-83% on average, and the number of translations with oscillatory hallucinations by 75-92% on average, across 57 tested translation directions. In a proof of concept on out-of-English translation, we also show that we can suppress off-target translations with large language models. We release our source code.[1]

1 Introduction

Hallucinations are a long-standing well-known problem in machine translation (MT) (Koehn and Knowles, 2017) and natural language generation (Ji et al., 2023). While there has been extensive research on their identification and mitigation (Lee et al., 2019; Raunak et al., 2021; Mohammadshahi et al., 2022b; Guerreiro et al., 2023a; Dale et al., 2023, among others), they still persist as an issue, especially in low-resource settings.

Contrastive conditioning has previously been used for analysing specific translation errors such as disambiguation errors and undertranslation (Vamvas and Sennrich, 2021, 2022). The main idea is that translations that are equally or more probable given some corrupted source than the true source are likely to be erroneous with respect to the corrupted span. We can apply the same intuition to hallucinations and translations into the wrong language, so called off-target translations: if hallucinations are detached from the source, they should have a similar probability given the true source and given a random other source. A translation in the wrong language should have a similar or higher probability if that language is marked as desired.

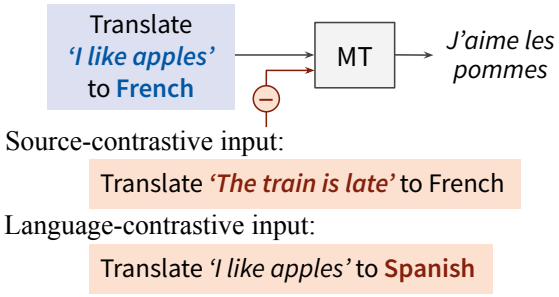

Figure 1: Our decoding objective yields a translation that is probable given the actual input, but improbable given a source-contrastive or language-contrastive input.

Inspired by this, we design decoding objectives that do not just search for the most probable translation, but search for a translation that maximizes the probability given the true input, but minimizes the probability given one or several contrastive inputs.

This paper makes the following contributions:

- We introduce contrastive decoding objectives to address two problems often observed in MT: hallucinations and off-target translations.
- By evaluating two massively multilingual MT models, M2M-100 (418M) and SMaLL-100, across 57 mostly low-resource translation directions, we show improvements in chrF2 by 1.3–1.7 points, and reduce the number of translations with chrF2 below 10 by 67-83%.

[1]https://github.com/ZurichNLP/ContraDecode

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 21–33

- Finally, we provide a proof of concept for applying our approach to LLM-based translation, where off-target issues are common.

2 Method

To suppress hallucinations, we pair each input X with a randomly selected input segment X'.[2] Rather than finding a translation that maximizes $p(Y|X)$, we search for one that both maximizes $p(Y|X)$ and minimizes $p(Y|X')$. We add a hyperparameter λ to control the strength of this contrastive penalty, yielding Eq. 1.

$$s(Y,X) = \sum_{i=1}^{|Y|} -\log\Big(p(y_i|y_{<i},X) - \lambda p(y_i|y_{<i},X')\Big) \quad (1)$$

We denote this **source-contrastive decoding**.

Off-target translations are a common failure mode in multilingual MT systems (Arivazhagan et al., 2019). They have been linked to the predominance of English in the training of multilingual systems (Rios et al., 2020). Production of text in the source language, often a copy of the input, is connected to the occurrence of copying in the training data, and the high probability of continuing to copy once a copy has been started (Ott et al., 2018).

The majority of multilingual MT systems use special tokens to indicate the target language, following Johnson et al. (2017).[3] To penalize output in the wrong language, we can add contrastive inputs that only vary the language indicator token.

Let l_y be the target language. We replace its indicator token with contrastive variants $l_{y'} \in L_c$ for languages we wish to suppress. Based on the predominant off-target languages in multilingual MT (Arivazhagan et al., 2019), our set of contrastive languages L_c consists of English[4] and the respective source language. This results in Eq. 2.

$$s(Y,X) = \sum_{i=1}^{|Y|} -\log\Big(p(y_i|y_{<i},X,l_y) - \sum_{l_{y'}\in L_c} \lambda p(y_i|y_{<i},X,l_{y'})\Big) \quad (2)$$

We refer to decoding with contrastive translation directions as **language-contrastive decoding**. We can combine source-contrastive and language-contrastive decoding by summing all contrastive variants, and refer to the weights as λ_{src} and λ_{lang}.

3 Evaluation

3.1 Data and Models

We perform experiments with two massively multilingual MT models: M2M-100 (418M) (Fan et al., 2021), and SMaLL-100 (Mohammadshahi et al., 2022a), a distilled version of M2M-100 (12B).

We use beam size 5. We perform minimal hyperparameter tuning on the ps-ast translation direction with M2M-100 and set λ_{src} to 0.7.[5] Since only a small number of directions suffer from off-target outputs, we do not tune λ_{lang}, setting it to 0.1.

We test on three sets of translation directions:

- the 25 non-English-centric directions used by Guerreiro et al. (2023a) (**HLMT**). These are af-zu, ar-fr, be-ru, cs-sk, de-hr, de-hu, el-tr, fr-sw, hi-bn, hi-mr, hr-cs, hr-hu, hr-sk, hr-sr, it-de, it-fr, nl-de, nl-fr, ro-de, ro-hu, ro-hy, ro-ru, ro-tr, ro-uk, uk-ru.[6]
- 29 directions between 5 low-resource languages from different branches of Indo-European, plus Zulu from the Atlantic-Congo family (**X-branch**): af, ast, hr, ps, ur, zu.
- 4 high-resource translation directions: en-de, de-en, en-fr, fr-en (**high-res**).

We also report results for the union of the sets (**all**).

We evaluate with spBLEU (Goyal et al., 2022) and chrF2 (Popović, 2015) using sacreBLEU (Post, 2018)[7] on the Flores-101 devtest set (Goyal et al., 2022). We use OpenLID (Burchell et al., 2023) for language identification to measure off-target rates. To quantify the number of hallucinations, we employ a rough approximation following Lee et al. (2019); Müller and Sennrich (2021), counting the proportion of segments with chrF2 < 10.[8] Another automatic metric specific for oscillatory hallucinations is top n-gram (TNG) (Guerreiro et al., 2023b;

[2] In practice, by shuffling segments of the input document.

[3] The indicator token can be in the source (SMaLL-100), or output-initial and force-decoded (M2M-100).

[4] Unless English is the target language.

[5] We exclude ps-ast from average results reported.

[6] See Appendix B for full language names.

[7] BLEU|#:1|c:mixed|e:no|tok:flores101|s:exp|v:2.3.1
chrF2|#:1|c:mixed|e:yes|nc:6|nw:0|s:no|v:2.3.1

[8] Müller and Sennrich (2021) report a threshold of 1, but this is a typo (personal communication). This method does not distinguish between hallucinations and off-target translations.

	chrF2				spBLEU			
	HLMT	X-branch	high-res	all	HLMT	X-branch	high-res	all
				M2M-100				
baseline	46.4	28.8	61.3	39.0	22.0	8.3	37.2	16.4
C_{src}	46.7	31.4	60.8	40.3	21.6	9.1	36.4	16.6
$C_{src+lang}$	46.8	32.1	60.7	40.7	21.5	9.3	36.1	16.6
				SMaLL-100				
baseline	48.3	32.0	62.5	41.4	23.5	10.2	38.7	18.1
C_{src}	48.5	34.2	62.1	42.5	23.2	11.1	37.9	18.4
$C_{src+lang}$	48.7	34.6	62.0	42.7	23.3	11.2	37.6	18.4

Table 1: Automatic evaluation results. Averages over different sets of translation directions.

Raunak et al., 2022, 2021), which measures the number of sentences whose top repeating n-gram is more frequent than the top repeated source n-gram by at least t.[9]

3.2 Results

We report results using source-contrastive decoding (C_{src}), and combining source-contrastive and language-contrastive decoding ($C_{src+lang}$) in Table 1.[10] Across 57 translation directions, chrF2 improves by 1.3 (M2M-100) and 1.1 (SMaLL-100) points with source-contrastive decoding. Language-contrastive decoding brings additional gains of 0.4 (M2M-100) and 0.2 (SMaLL-100) points.

Improvements are more modest when measured with spBLEU (0.2 on M2M-100; 0.3 on SMaLL-100). We notice that hallucinations tend to be overlong, and can perversely improve BLEU by reducing the brevity penalty. We thus consider chrF2, which pairs precision with recall instead of a simplistic brevity penalty, to be our primary metric.

Off-target translations are relatively rare for the translation directions tested, especially for SMaLL-100 (see Table 2). With M2M-100, the highest proportion of English outputs in the baseline was detected for af-zu (9.1%), the highest percentage of outputs in the source language for hr-sr (4.2%)[11]. These are also among the translation directions that benefit the most from language-contrastive decoding: chrF2 increases by 2.3 for hr-sr[12], and by 2 for af-zu. However, we observe the largest increase in chrF2 (3.2) for ast-zu, a direction where source-contrastive decoding increases off-target outputs, and where the English output rate goes from 5.5% (baseline) to 9.9% (C_{src}) to 2.7% ($C_{src+lang}$).

	M2M-100		SMaLL-100	
	EN	SRC	EN	SRC
baseline	260	55	54	63
C_{src}	375	47	78	70
$C_{src+lang}$	88	28	16	21

Table 2: Number of off-target outputs (out of 57684), in English (EN) or the source language (SRC).

	HLMT	X-branch	high-res	all
		M2M-100		
baseline	2.1%	13.0%	0.0%	7.3%
C_{src}	1.0%	4.1%	0.0%	2.4%
$C_{src+lang}$	0.5%	2.0%	0.0%	1.2%
		SMaLL-100		
baseline	1.3%	10.6%	0.0%	5.6%
C_{src}	0.8%	4.3%	0.0%	2.5%
$C_{src+lang}$	0.4%	3.4%	0.0%	1.8%

Table 3: Proportion of translations with chrF2 < 10.

The proportion of translations with chrF2 below 10 is shown in Table 3. We observe large reductions in the number of defect translations, with a reduction from 7.3% to 1.2% (-83%) for M2M-100, and from 5.6% to 1.8% (-67%) for SMaLL-

	HLMT	X-branch	high-res	all
		M2M-100		
baseline	2.4%	16.9%	0.0%	9.3%
C_{src}	0.3%	3.7%	0.0%	2.0%
$C_{src+lang}$	0.1%	1.3%	0.0%	0.7%
		SMaLL-100		
baseline	0.7%	11.2%	0.0%	5.9%
C_{src}	0.1%	3.9%	0.0%	2.0%
$C_{src+lang}$	0.1%	2.9%	0.0%	1.5%

Table 4: Proportion of translations with oscillatory hallucinations according to TNG.

[9] We follow Guerreiro et al. (2023b) and use $n = 4$ and $t = 2$.

[10] See Appendix A for full results.

[11] This number may be an overestimate due to the close relationship between Serbian and Croatian, and the consequent difficulty of doing reliable language identification.

[12] This improvement is somewhat coincidental because both Latin and Cyrillic are accepted for Serbian, but Flores-101 has Cyrillic references. Penalizing output in Croatian, which uses the Latin alphabet, indirectly rewards output in Cyrillic.

100. When focusing on oscillatory hallucinations according to TNG in Table 4, the improvement is even more pronounced, with a reduction from 9.3% to 0.7% (-92%) for M2M-100, and from 5.9% to 1.5% (-75%) for SMaLL-100.

4 Ablation Studies

The fact that we pick contrastive inputs from the test sets at random raises a few questions about this approximation. We repeated the translation with M2M-100 across all 57 translation directions 3 times and find that the standard deviation is minimal (0.0107 for chrF2). Using a single random input as a contrastive variant is a heavy approximation, but our ablation study in Table 5 shows that this yields the majority of the performance gains, and using up to 3 inputs as contrastive examples[13] only yields an additional 0.1 point improvement in chrF2.

	chrF2	spBLEU
baseline	38.97	16.40
C_{src} (1)	40.31	16.60
C_{src} (2)	40.39	16.68
C_{src} (3)	40.41	16.67

Table 5: Ablation results for M2M-100 with different numbers of source-contrastive inputs. Average over all languages reported.

5 Application to Large Language Models

In this section, we demonstrate that our method can be applied to large language models (LLM). Previous work has achieved competitive translation quality for some directions by prompting models such as PaLM (Vilar et al., 2023; Garcia et al., 2023), GPT (Hendy et al., 2023) or BLOOM (Bawden and Yvon, 2023). However, LLM-based translation is still prone to hallucination and off-target translation (Zhang et al., 2023; Guerreiro et al., 2023a).

Our demonstration is based on the Llama 2 model family (Touvron et al., 2023) and specifically the instruction-tuned version (*Llama Chat*), exploiting the fact that MT examples were among the data used for instruction tuning (Wei et al., 2022; Chung et al., 2022). We generate translations by instructing the model to translate a segment into a given language, force-decoding the line *"Sure, here's the translation:"*, and then decoding until the next line break. The prompt we used is detailed in Appendix E.

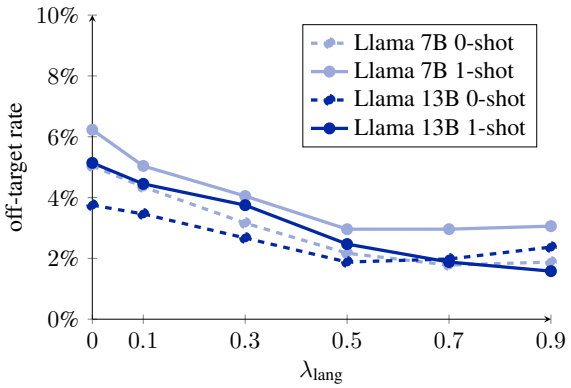

Figure 2: Off-target translation rate for Llama 2 Chat models when translating the English Flores-101 devtest set into German. Language-contrastive decoding tends to reduce off-target translation as λ_{lang} is increased.

When using this simple prompting approach in the en–de direction, we find that off-target output in English is very common. Moreover, providing a 1-shot example in the prompt, while improving translation quality, does not prevent the off-target issue. We thus apply language-contrastive decoding and add a contrastive prompt that instructs the model to "translate" into English instead of German. The decoding objective is analogous to Eq. 2. We use 4-bit precision (Dettmers and Zettlemoyer, 2023) and greedy decoding.

Figure 2 shows the percentage of off-target output for different λ_{lang}. Generally, we observe that the off-target rate falls with increasing λ_{lang}, demonstrating the effectiveness of our method with LLM prompting. English–French (Appendix C) has similar results. In terms of overall translation quality, we find that language-contrastive decoding improves chrF2 and spBLEU and only becomes detrimental for $\lambda_{\text{lang}} > 0.7$ (Appendix D).

6 Related Work

Hallucination Detection and Reduction

Various methods have been proposed to detect hallucinations, including identifying typical patterns in the output (Raunak et al., 2021), using internal information like attention patterns (Lee et al., 2019) or the contribution of the source to predictions (Dale et al., 2023), or measures of decoder confidence, including the output probability (Guerreiro et al., 2023b) or stability of samples under perturbation (Lee et al., 2019; Guerreiro et al., 2023b).

[13] we divide λ_{src} by the number of contrastive inputs.

Hallucination mitigation is more difficult, especially if we assume that models are already trained with best practices, and focus on training-free methods. Several studies use external models for mitigation, e.g. using other translation models as a fallback (Guerreiro et al., 2023a), or sample reranking based on quality estimation (QE) models (Guerreiro et al., 2023b). Our method has the advantage of not requiring external models, and we note that modern QE metrics are themselves prone to score certain hallucinations highly (Freitag et al., 2022; Yan et al., 2023).

Mitigation methods that do not rely on external models are typically sampling-based. Guerreiro et al. (2023b) report that even the translation model's own sequence probability can be used for sample reranking. A consensus translation can be identified via sampling-based Minimum Bayes Risk decoding (Eikema and Aziz, 2020), which benefits from the fact that hallucinations are dissimilar from each other (Müller and Sennrich, 2021).

Contrastive Decoding

Contrastive decoding is similar to contrastive learning (e.g. Hadsell et al., 2006; Socher et al., 2014; Gao et al., 2021) in that positive and negative examples are contrasted, but involves no training.

Li et al. (2023) introduce a form of contrastive decoding that contrasts the probability between different models, whereas our methods work with a single model, contrasting inputs. Su and Collier (2023) introduce a contrastive search where potential output tokens are compared to previous tokens, penalizing outputs that are similar to the context and thus suppressing repetition patterns.

Source-contrastive decoding can also be seen as a variant of implicit language model (ILM) compensation, mirroring recent work by Herold et al. (2023). Our work is different in motivation in that ILM is typically used to allow the inclusion of an external LM, where we show the effectiveness of simply suppressing the ILM. Also, we show the effectiveness of a different, simple approximation.

Finally, language-contrastive decoding bears resemblance to negative prompting, a technique used to suppress concepts in image generation.

7 Conclusion

This paper shows that certain failure modes of MT can be addressed by contrastive decoding objectives that use pairs or sets of inputs for the prediction. Specific contrastive inputs address specific errors, and we introduce strategies to mitigate hallucinations and off-target translation.

Future work could expand on our work by exploring if other MT failure modes can be mitigated with appropriate contrastive inputs, or if other forms of control can be improved. For example, for models that use domain indicator tokens (Kobus et al., 2017), we could perform domain-contrastive decoding and achieve stronger domain control. Beyond MT, we expect that source-contrastive decoding can also be useful for other tasks, e.g. to penalize over-generic responses in dialogue systems.

8 Limitations

We only tested language-contrastive decoding in multilingual models that control the target language via language indicator tokens. It is possible to apply the same strategy to modular architectures that use language-specific components (Firat et al., 2016; Vázquez et al., 2019; Bapna and Firat, 2019), but its effectiveness remains to be tested. For bilingual translation models that suffer from off-target translations, e.g. because of noisy training data (Khayrallah and Koehn, 2018), we would need bilingual models for other translation directions to implement language-contrastive decoding, but this sacrifices the main strength of our approach: not relying on external models.

We perform minimal hyperparameter tuning for λ_{src}, and did not tune λ_{lang}. Using the same hyperparameters across translation directions and translation models results in performance degradations in some cases, most noticeably for high-resource translation directions. We consider it a positive result that we obtain improvements on average with minimal hyperparameter tuning, but future work may wish to use more complex strategies to weight (or disable) contrastive variants across translation directions.

9 Ethics Statement

This paper introduces new decoding objectives for machine translation, and we do not foresee any harms being caused by source-contrastive or language-contrastive decoding. More widely, we are interested in exploring novel contrastive inputs for risk mitigation, e.g. for model debiasing, but certain contrastive inputs could also have undesirable consequences, e.g. increasing model bias.

Acknowledgements

We thank the anonymous reviewers for their comments. This work was funded by the Swiss National Science Foundation (project MUTAMUR; no. 213976).

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

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

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference*

of the European Chapter of the Association for Computational Linguistics, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7114–7123, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

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.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022a. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022b. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4308–4329, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2020. Subword segmentation and a single bridge language affect zero-shot neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 528–537, Online. Association for Computational Linguistics.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation. *Transactions on Machine Learning Research*.

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

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

A Full Results

direction	chrF2			spBLEU		
	baseline	C_{src}	$C_{src+lang}$	baseline	C_{src}	$C_{src+lang}$
af-zu	20.0	24.2	26.2	3.6	4.1	4.7
ar-fr	53.5	52.9	52.3	27.9	26.8	25.9
be-ru	42.6	43.8	43.7	15.8	16.5	16.5
cs-sk	55.5	55.1	55.0	33.7	33.0	32.8
de-hr	50.1	50.1	50.2	23.0	22.6	22.8
de-hu	49.1	48.7	48.8	23.2	22.3	22.3
el-tr	46.2	46.4	46.3	19.6	19.6	19.4
fr-sw	41.9	44.0	44.0	15.3	15.8	15.8
hi-bn	36.5	37.3	37.8	16.1	16.2	16.4
hi-mr	34.6	34.7	35.1	10.5	10.3	10.3
hr-cs	48.6	48.1	47.9	26.3	25.4	25.0
hr-hu	48.2	47.6	47.7	21.7	20.8	20.9
hr-sk	49.7	49.4	49.3	26.9	26.2	26.0
hr-sr	48.4	48.2	50.5	28.0	27.8	28.8
it-de	50.1	49.8	49.6	22.0	21.5	21.3
it-fr	56.8	56.2	55.9	32.7	31.7	30.9
nl-de	49.6	49.1	48.8	21.2	20.7	20.5
nl-fr	51.7	51.1	50.6	26.7	25.8	25.1
ro-de	52.5	52.3	52.1	25.0	24.7	24.3
ro-hu	49.5	49.1	48.8	23.5	22.8	22.6
ro-hy	24.1	28.7	29.3	4.7	6.3	6.4
ro-ru	48.7	48.4	48.3	23.6	23.1	22.8
ro-tr	50.3	50.4	50.3	24.2	24.0	23.7
ro-uk	48.2	47.9	47.9	23.8	23.4	23.4
uk-ru	53.8	53.4	53.3	29.9	29.5	29.3
avg (non-English-centric)	46.4	46.7	46.8	22.0	21.6	21.5
af-ast	45.1	46.3	46.2	19.3	19.2	18.9
af-hr	47.6	47.4	47.4	20.8	20.3	20.3
af-ps	22.8	24.4	24.5	5.4	5.7	5.8
af-ur	35.9	36.4	36.5	14.0	14.1	14.1
af-zu	20.0	24.2	26.2	3.6	4.1	4.7
ast-af	39.6	43.0	42.9	14.2	15.8	15.8
ast-hr	33.7	41.6	42.7	11.1	15.8	16.3
ast-ps	16.6	21.6	22.4	2.4	4.7	4.8
ast-ur	22.2	31.3	32.0	6.3	10.7	10.8
ast-zu	16.0	21.1	24.3	2.6	3.3	3.9
hr-af	46.3	46.4	46.3	17.6	17.5	17.5
hr-ast	45.3	46.5	46.4	18.8	18.6	18.6
hr-ps	21.8	23.4	23.7	4.4	5.0	5.1
hr-ur	35.1	35.8	36.1	13.6	13.6	13.8
hr-zu	18.6	23.0	24.9	3.0	3.6	4.1
ps-af	34.9	35.5	36.0	8.3	8.5	8.7
ps-ast	*32.2*	*34.3*	*34.2*	*7.8*	*9.4*	*9.1*
ps-hr	33.5	34.0	34.0	8.0	8.1	8.2
ps-ur	30.8	31.4	31.4	9.8	10.1	10.1
ps-zu	16.2	21.0	23.9	1.8	2.4	2.8
ur-af	35.3	36.1	36.6	9.0	9.1	9.3
ur-ast	29.7	33.6	34.1	7.1	9.1	9.1
ur-hr	34.2	35.1	35.4	8.9	9.1	9.2
ur-ps	21.2	22.8	23.5	4.2	4.8	4.9
ur-zu	16.0	19.5	22.2	1.4	1.7	2.1
zu-af	28.9	30.6	31.0	6.9	7.7	7.7
zu-ast	26.0	29.1	29.5	5.8	7.5	7.5
zu-hr	27.9	28.4	28.8	6.2	6.3	6.4
zu-ps	12.2	17.1	17.4	1.3	2.8	2.7
zu-ur	22.6	24.7	24.9	4.8	5.8	5.8
avg (X-branch)	28.8	31.4	32.1	8.3	9.1	9.3
de-en	61.4	61.2	61.0	36.6	36.0	35.9
en-de	57.2	56.6	56.5	31.1	30.1	29.8
en-fr	63.8	63.0	62.9	42.2	40.9	40.5
fr-en	62.8	62.5	62.4	38.9	38.6	38.4
avg (high-res)	61.3	60.8	60.7	37.2	36.4	36.1
avg (all)	39.0	40.3	40.7	16.4	16.6	16.6

Table 6: Full results for M2M-100. The direction ps-ast was used to tune λ_{src} and is excluded from the averages.

direction	chrF2			spBLEU		
	baseline	C_{src}	$C_{src+lang}$	baseline	C_{src}	$C_{src+lang}$
af-zu	26.2	31.4	31.8	4.4	6.9	7.0
ar-fr	53.9	53.6	53.3	28.2	27.7	27.0
be-ru	45.1	45.2	45.1	17.3	17.5	17.3
cs-sk	55.3	55.1	55.2	33.0	32.6	32.8
de-hr	51.2	51.3	51.1	24.5	24.3	24.1
de-hu	49.7	49.4	49.5	23.7	23.1	23.1
el-tr	46.2	46.2	46.1	19.0	18.5	18.3
fr-sw	48.9	50.1	50.2	22.9	23.3	23.3
hi-bn	43.1	43.1	42.6	24.0	23.4	22.8
hi-mr	38.8	38.8	38.9	14.8	14.2	14.5
hr-cs	49.3	48.9	49.0	26.3	25.7	26.1
hr-hu	49.2	49.0	48.8	22.5	22.2	22.1
hr-sk	50.8	50.4	50.4	27.8	27.2	27.1
hr-sr	47.3	47.1	52.6	28.0	27.7	30.5
it-de	51.0	51.2	51.1	23.5	23.5	23.3
it-fr	57.2	56.8	56.8	33.1	32.0	31.9
nl-de	50.2	50.2	50.1	22.1	22.0	21.8
nl-fr	52.7	52.2	52.2	27.8	26.8	26.7
ro-de	54.2	53.6	53.7	27.4	26.4	26.4
ro-hu	50.0	50.1	49.9	23.8	23.7	23.5
ro-hy	34.5	35.3	35.9	11.0	11.3	11.6
ro-ru	49.4	49.3	49.3	24.1	23.7	23.8
ro-tr	50.4	50.2	50.0	23.5	23.0	22.9
ro-uk	49.2	49.0	49.2	24.5	24.1	24.1
uk-ru	54.1	53.8	53.9	30.1	29.7	29.7
avg (non-English-centric)	48.3	48.5	48.7	23.5	23.2	23.3
af-ast	48.3	49.7	49.3	22.0	21.6	21.5
af-hr	50.6	50.6	50.4	23.5	23.4	23.3
af-ps	24.8	24.9	25.1	6.4	6.2	6.1
af-ur	36.3	36.3	36.7	13.9	13.8	14.0
af-zu	26.2	31.4	31.8	4.4	6.9	7.0
ast-af	49.2	49.4	49.5	22.8	22.7	22.7
ast-hr	47.1	47.9	47.9	21.1	21.1	20.9
ast-ps	22.3	22.7	23.0	4.8	4.8	5.0
ast-ur	31.4	33.0	33.4	10.5	11.6	11.8
ast-zu	13.7	25.3	27.9	1.8	4.9	5.6
hr-af	50.8	50.7	50.9	23.4	23.3	23.2
hr-ast	47.3	48.5	48.3	20.6	20.1	20.0
hr-ps	24.0	24.1	24.4	5.6	5.4	5.4
hr-ur	35.2	35.4	35.7	13.3	13.4	13.3
hr-zu	21.7	28.9	30.4	3.2	6.0	6.3
ps-af	39.0	39.2	39.2	12.0	12.2	12.3
ps-ast	*29.9*	*34.8*	*35.0*	*6.0*	*9.3*	*10.0*
ps-hr	35.3	35.7	35.8	9.4	9.8	9.8
ps-ur	31.5	31.5	31.8	10.2	10.4	10.4
ps-zu	15.8	21.1	23.2	1.0	2.3	3.0
ur-af	42.6	42.9	43.1	15.1	15.1	15.1
ur-ast	33.7	38.5	38.3	8.3	12.1	12.1
ur-hr	40.4	40.4	40.6	13.4	13.3	13.2
ur-ps	23.5	23.8	23.9	5.1	5.1	5.2
ur-zu	11.6	19.5	20.6	0.6	2.1	2.6
zu-af	33.8	35.5	35.6	8.9	11.1	11.2
zu-ast	26.8	31.4	32.0	4.9	7.5	8.6
zu-hr	29.1	31.4	31.8	5.5	7.4	7.7
zu-ps	15.1	18.2	18.1	1.4	2.6	2.4
zu-ur	22.0	25.1	25.2	3.4	5.2	5.2
avg (X-branch)	32.0	34.2	34.6	10.2	11.1	11.2
de-en	62.7	62.3	62.2	38.3	37.4	37.3
en-de	59.3	58.9	58.8	33.7	33.2	32.9
en-fr	64.8	64.2	64.1	43.4	41.9	41.8
fr-en	63.2	63.0	62.7	39.4	39.0	38.6
avg (high-res)	62.5	62.1	62.0	38.7	37.9	37.6
avg (all)	41.4	42.5	42.7	18.1	18.4	18.4

Table 7: Full results for SMaLL-100. Averages exclude ps-ast translation direction for comparability to M2M-100.

B Languages

language code	language
af	Afrikaans
ar	Arabic
ast	Asturian
be	Belarusian
bn	Bengali
cs	Czech
de	German
el	Greek
en	English
fr	French
hi	Hindi
hr	Croatian
hu	Hungarian
hy	Armenian
it	Italian
mr	Marathi
nl	Dutch; Flemish
ps	Pushto; Pashto
ro	Romanian; Moldavian; Moldovan
ru	Russian
sk	Slovak
sr	Serbian
sw	Swahili
tr	Turkish
uk	Ukrainian
ur	Urdu
zu	Zulu

Table 8: List of languages in our experiments, sorted by ISO 639-1 language code.

C LLM Off-Target Analysis for English–French

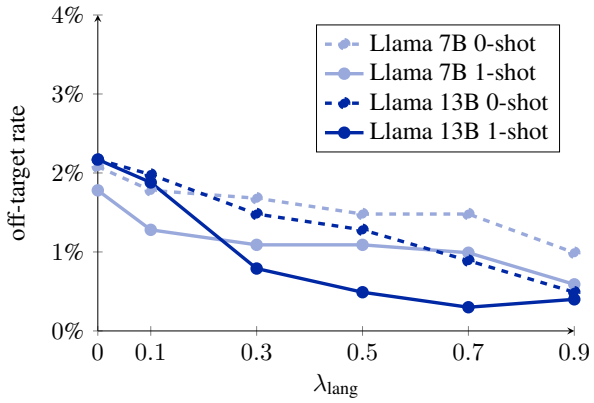

Figure 3: Off-target translation rate for Llama 2 Chat models when translating the English Flores-101 devtest set into French. As with German (Figure 2), language-contrastive decoding tends to reduce off-target translation as λ_{lang} is increased.

D LLM Automatic Evaluation Results

en-de	chrF2						spBLEU					
	baseline	$\lambda_{\text{lang}} = 0.1$	0.3	0.5	0.7	0.9	baseline	$\lambda_{\text{lang}} = 0.1$	0.3	0.5	0.7	0.9
Llama 7B 0-shot	50.0	49.9	50.2	50.3	49.9	49.4	23.8	23.7	23.8	23.7	23.3	22.3
Llama 7B 1-shot	50.5	50.9	51.1	51.4	50.9	49.7	24.4	24.7	24.8	25.1	24.3	22.6
Llama 13B 0-shot	54.2	54.5	54.5	54.7	54.3	53.3	29.1	29.4	29.3	29.3	29.0	27.8
Llama 13B 1-shot	54.4	54.5	54.7	55.1	54.9	53.7	29.4	29.5	29.7	29.9	29.5	27.4
Average	*52.3*	*52.5*	*52.6*	*52.9*	*52.5*	*51.5*	*26.7*	*26.8*	*26.9*	*27.0*	*26.5*	*25.0*

Table 9: English–German: Automatic evaluation of LLM-based translation on the Flores-101 devtest set. The scores tend to increase with smaller values of λ_{lang}, but decline with larger values.

en-fr	chrF2						spBLEU					
	baseline	$\lambda_{\text{lang}} = 0.1$	0.3	0.5	0.7	0.9	baseline	$\lambda_{\text{lang}} = 0.1$	0.3	0.5	0.7	0.9
Llama 7B 0-shot	58.3	58.7	58.8	58.6	58.1	57.2	35.2	35.6	35.7	35.5	34.9	33.5
Llama 7B 1-shot	58.4	58.7	58.7	58.4	58.0	56.7	35.8	36.2	36.1	35.7	35.1	33.2
Llama 13B 0-shot	62.4	62.5	62.6	62.6	62.6	62.0	40.6	40.6	40.8	40.8	40.6	39.7
Llama 13B 1-shot	62.1	62.2	62.6	62.6	62.6	61.7	40.6	40.7	41.0	41.2	41.1	39.8
Average	*60.3*	*60.5*	*60.7*	*60.6*	*60.3*	*59.4*	*38.0*	*38.3*	*38.4*	*38.3*	*37.9*	*36.6*

Table 10: English–French: Automatic evaluation of LLM-based translation on the Flores-101 devtest set, showing patterns similar to English–German.

E LLM Implementation Details

Our input to Llama consists of a system prompt and an instruction. We force-decode the prefix of the assistant response to make sure that the next generated line is the actual translation and not a prologue by the assistant.

Zero-shot

```
<s>[INST] <<SYS>>
You are a machine translation system that translates sentences from English to
German.  You just respond with the translation, without any additional comments.
<</SYS>>[INST] "We now have 4-month-old mice that are non-diabetic that used to be
diabetic," he added.

Translate to German [/INST]Sure, here's the translation:
```

One-shot

```
<s>[INST] <<SYS>>
You are a machine translation system that translates sentences from English to
German.  You just respond with the translation, without any additional comments.

Example instruction:

On Monday, scientists from the Stanford University School of Medicine announced the
invention of a new diagnostic tool that can sort cells by type:  a tiny printable
chip that can be manufactured using standard inkjet printers for possibly about one
U.S. cent each.
Translate to German

Example response:

Sure, here's the translation:
Am Montag haben die Wisenschaftler der Stanford University School of Medicine die
Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ
sortiert werden können:  ein winziger, ausdruckbarer Chip, der für jeweils etwa
einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.
<</SYS>>[INST] "We now have 4-month-old mice that are non-diabetic that used to be
diabetic," he added.

Translate to German [/INST]Sure, here's the translation:
```