Generating Difficult-to-Translate Texts

Vilém Zouhar★ **Wenda Xu**■ **Parker Riley**■ **Juraj Juraska**■
Mara Finkelstein■ **Markus Freitag**■ **Daniel Deutsch**■
■Google ★ETH Zurich

Abstract

Machine translation benchmarks sourced from the real world are quickly obsoleted, due to most examples being easy for state-of-the-art translation models. This limits the benchmark's ability to distinguish which model is better or to reveal models' weaknesses. Current methods for creating difficult test cases, such as subsampling or from-scratch synthesis, either fall short of identifying difficult examples or suffer from a lack of diversity and naturalness. Inspired by the iterative process of human experts probing for model failures, we propose MT-breaker, a method where a large language model iteratively refines a source text to increase its translation difficulty. The LLM iteratively queries a target machine translation model to guide its generation of difficult examples. Our approach generates examples that are more challenging for the target MT model while preserving the diversity of natural texts. While the examples are tailored to a particular machine translation model during the generation, the difficulty also transfers to other models and languages.

1 Introduction

Test sets need to be difficult in order to be informative. If an NLP model performs flawlessly on a test set, then we learn very little about the model's shortcomings. Or, if we compare two NLP models that perform near-perfectly on a test set, we do not know which one is better. For machine translation, current models are able to translate even recent benchmarks with only a few mistakes across several language directions (Kocmi et al., 2024a). Attempts have been made to automatically select a subset of existing datasets that still pose challenges to the models (Proietti et al., 2025). However, these methods fall short due to the absence of a sufficient number of naturally occurring challenging examples. Another line of work makes use of LLMs to create difficult examples from scratch in a zero-shot manner (Pombal et al., 2025), which we show is inadequate due to its lack of difficulty and diversity.

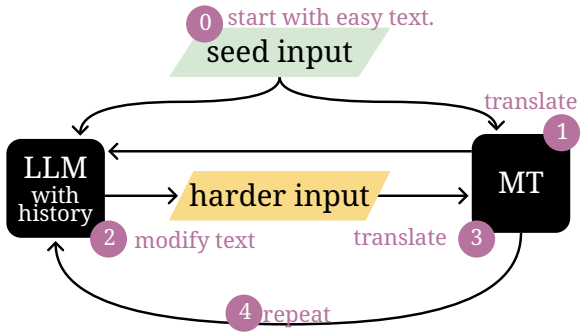

Figure 1: High-level overview of our interactive difficulty-to-translate text generation pipeline. The process is repeated for a fixed number of iterations.

A human practitioner who is trying to assess the quality of a model approaches this task differently: not by selecting difficult examples or synthesizing them from scratch (finding based on interviews in Section 3.7). Instead, the practitioner iteratively tries to find weaknesses in the model. They might start with some random inputs and then, based on the model's response, modify these inputs to cause the model to provide a faulty output. This way, the practitioner arrives at difficult examples tailored to a particular model that reveal its shortcomings. These insights can be then used to diagnose and improve the model.

In this work we propose MT-breaker, which uses an LLM to generate difficult-to-translate examples that mimic the process of the human practitioner (see Figure 1). We start with a seed text 0 that the machine translation model translates 1. The text and its translation is then passed to an LLM which modifies it 2 and passes it to the MT model again 3. The translation is again passed back to the LLM model with chat-like history and the whole process repeats 4.

★Work done during Google Internship.
Correspondence: vzouhar@ethz.ch

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 204–219
March 28, 2026 ©2026 Association for Computational Linguistics

Zeroshot (min)	Seeds	MTbreaker (seeded)
It is what it is.	Another one has been found!	Another rabbit hole of lies was found!
The quick brown fox jumps over the lazy dog.	I quit FB, removed all Meta and Pinterest tracker from my website and killed my private WhatsApp account.	I've quit FB and killed the private WhatsApp account to get rid of Meta's and Pinterest's gross ad-tracking tools that I hated with a fiery passion.
He bade them farewell and then he bade them all to hell.	Going back up tomorrow and we're doing stalls and slow flight.	Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.
It is what it is.	Heheh not one but three!	Heheh not one, two, but three!!!11!oneeleven

Table 1: Example of generated sources in English by Zeroshot, Seeds, and MT-breaker (seeded). Examples were selected to be short. See Appendix Table 7 for more examples.

MT-breaker(seed text s_0, steps N):
1: **for** $i = 1 \ldots N$ **do**
2: $\quad t_{i-1} \leftarrow \mathrm{MT}(s_{i-1})$ ▷ Translate
3: $\quad s_i \leftarrow \mathrm{LLM}_{\mathrm{step}}(\{s_k\}_{k=0}^{i-1}, \{t\}_{k=0}^{i-1})$ ▷ Generate next
4: **return** $\arg\min \mathrm{QE}(\{s_k\}_{k=0}^{N}, \{t\}_{k=0}^{N})$ ▷ Pick difficult

Algorithm 1: Given a source text, the LLM receives the MT model's translation which informs how to next change the text. The $\mathrm{LLM}_{\mathrm{step}}$ can be instructed to not stray too far from the original meaning or to preserve naturalness. At the end, we pick the source text from s that led to the worst translation.

We show that this approach maintains the diversity and naturalness of the source examples, while vastly increasing the difficulty of the generated examples. Despite the approach relying on a quality estimation metric, the same findings are confirmed by an expert MQM human-evaluation study. The examples are tailored to be difficult for a particular machine translation model, though we also show that the difficulty also transfers to other models and languages. The generated dataset can be used to find weaknesses of a particular model, which can, in turn, be used for hillclimbing during model development.

2 Methods

We now describe the various approaches of generating difficult-to-translate texts. The baseline, which we refer to as **seeds**, consists of simply taking existing texts from a dataset; i.e. not creating new ones.

Zero-shot benchmarking. This method simply prompts a large language model to generate an example to be part of a benchmark. This generation can be steered towards particular topics, domains, and also difficulty. This method is a simplified version of Pombal et al. (2025). We introduce a variant "(history)", where we include the history of previously generated texts in the prompt such that they are not repeated.

MT-breaker. Our method is shown in Figure 1 and Algorithm 1. It is based on mimicking a human expert that tries to find an input that triggers an error in the machine translation model, which is a strategy motivated by qualitative interviews with human experts (Section 3.7). We start with a piece of text, called a seed, which is translated by a machine translation model. This, together with the seed is the input to $\mathrm{LLM}_{\mathrm{step}}$, which modifies the seed. This is repeated for a certain number of steps, e.g. 10. At the end, we compute the quality estimation scores for all steps and select the most difficult source. A quality estimation metric is a model that takes in the source and its translation and outputs an scalar assessment of the translation's quality (Freitag et al., 2024). We also consider MT-breaker without the seed text "(seedless)", and a version which has the $\mathrm{LLM}_{\mathrm{step}}$ seeing the intermediate quality estimation scores "(seeded+qe)". The prompts for $\mathrm{LLM}_{\mathrm{step}}$ in all variants are shown in Appendix B.

3 Experiments and Results

We now describe experiments in which we compare and analyze the above listed methods for generating difficult-to-translate texts.

Setup. For brevity and computational feasibility, we consider English→{Czech, German, Chinese, Vietnamese, Polish} language directions. For seeds, we use 100 randomly sampled English sources from WMT 2024 (Kocmi et al., 2024a). We use the following publicly available machine translation models:

- Google Translate[1]
- Gemini 2.5-flash (Comanici et al., 2025)
- Gemini 2.0-flash (Comanici et al., 2025)
- Gemma 3-27b-it (Gemma Team et al., 2025)

[1] translate.google.com, accessed August 2025.

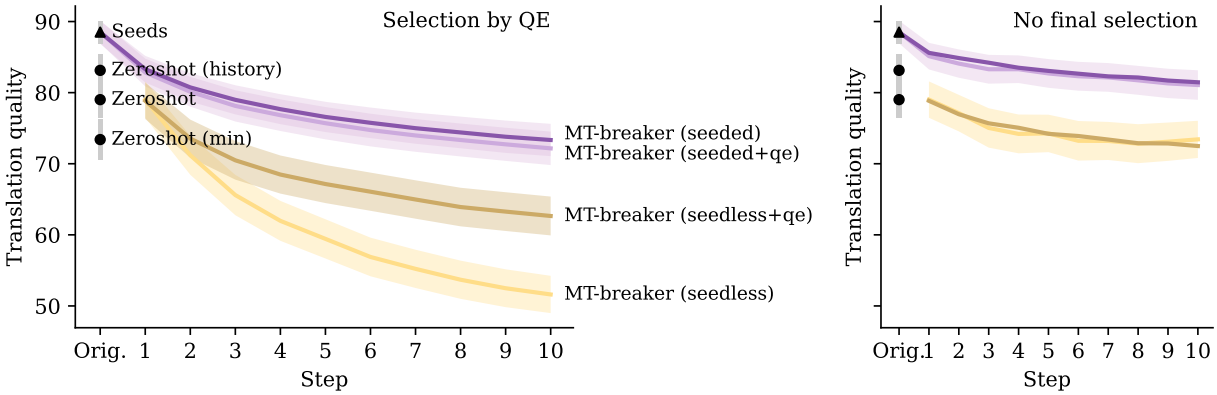

Figure 2: Development of difficulty with each iteration. Seeds require no iterations, Zeroshot and Zeroshot (history) require one LLM step, and the rest requires 10 steps. Right plot shows running hypotheses at a given step without final selection by quality estimation. This does not include Zeroshot (min) which is simply Zeroshot with selection. Values averaged across all models and languages. Shaded areas show 90% t-test confidence intervals averaged across all models and languages.

For the LLM_{step} we use Gemini 2.5-pro (Comanici et al., 2025) in the main paper, though explore others in Section 3.4. For measuring the quality of the translations, we use the average of MetricX-24-QE (Juraska et al., 2024) and prompted LLM-as-QE (see Appendix B for details). We also include a source-only version of MetricX, that only sees the source texts and no translation, which serves as a proxy for general difficulty (see Appendix A). All three are scaled from 0 (poor translation, high difficulty) to 100 (perfect translation, low difficulty). Same as Proietti et al. (2025), the quantity of "difficulty" has an opposite direction to quality estimation (100-QE). While we use the term "difficulty", we include the original quality estimation scores in tables and figures for simplicity.

We compare the following methods for obtaining a difficult test set:

- "Seed" segments from WMT
- "Zeroshot" benchmarking
- "Zeroshot (history)" benchmarking with history
- "Zeroshot (min)" benchmarking with selection[2]
- "MT-breaker (seeded)" with seeds
- "MT-breaker (seedless)" without seeds
- "MT-breaker (seeded+qe)" with seeds and with quality estimation
- "MT-breaker (seedless+qe)" without seeds and with quality estimation

For a fair comparison, we instruct the generation of the Zeroshot benchmarking and MT-breaker without seeds methods to have approximately the same number of words as the seeds used by the other methods.

See generated examples of Zeroshot (min) and MT-breaker (seeded) in Table 1 and Appendix Tables 7 to 9.

3.1 Increasing Difficulty

As shown in Figure 2, the translation quality of texts generated by MT-breaker methods decreases with each step but begins to plateau around step 10. Interestingly, having access to the quality estimation during the breaking process does not help the MT-breaker (seeded) and worsens MT-breaker (seedless). We explain this through different variances. The per-step averages (Figure 2 right) are the same for the seedless version. The reason for the lower minimum (shown cumulatively in the lines in Figure 2 left) is that the version with quality estimation has lower variance (σ^2=297) than the version without it (σ^2=325). This is likely caused by the model using most of the 10 steps to continue the breakage with the text with currently lowest quality estimation score, which can be a dead-end. The minimum of a variable with higher variance is lower than the minimum of a variable with lower variance, even when the means are the same.

Methods that use seeds for the initial source (Seeds, MT-breaker seeded) lead to lower difficulty than those without such restrictions (Zeroshot, MT-breaker seedless). In the next section, we show that this is part of a diversity-difficulty tradeoff.

[2]For a fair comparison, we also include a version of Zeroshot that is run 10x (same number of steps as MT-breaker) out of which (per batch of 10) we select the most difficult sources with quality estimation in the same manner.

	Seed	Zeroshot	Zeroshot (min)	Zeroshot (history)	MT-breaker (seedless)	MT-breaker (seeded)	MT-breaker (seedless+QE)	MT-breaker (seeded+QE)
Diversity (higher is better)								
Diversity (embd)	0.33	0.09	0.10	0.31	0.16	0.34	0.12	0.34
Diversity (chrF)	0.76	0.24	0.29	0.73	0.46	0.77	0.38	0.77
Diversity (topics)	247	126	131	225	148	262	147	262
Diversity (errors)	23.50	20.80	22.00	20.65	22.95	23.80	22.65	23.90
Diversity (words)	1566	1121	1177	1508	1532	2082	1527	2091
Complexity & Style								
Grammaticality	92.28	96.16	95.67	97.91	94.15	87.65	92.27	87.93
Naturalness	92.91	57.49	54.15	73.89	29.80	82.71	41.99	83.31
Word Rarity	17.67	46.24	46.14	28.45	54.80	35.69	53.23	34.77
Syntax Complexity	39.95	55.44	56.28	57.12	70.38	51.21	65.78	50.23
Avg. Word Count	33.25	31.21	31.25	33.34	31.32	39.62	37.67	40.18
Avg. Word Length	4.69	5.44	5.43	4.93	5.90	5.07	5.66	5.01
Quality Estimation (lower is better)								
QE (Gemini)	89.78	81.94	73.94	83.53	45.35	67.52	59.97	69.23
QE (MetricX)	87.25	76.09	72.87	82.73	57.88	76.81	65.33	77.45
SRCQE (MetricX)	94.47	91.69	91.55	93.29	90.54	91.96	90.12	92.01

Table 2: Quantitative automatic evaluation of difficult data generation approaches.

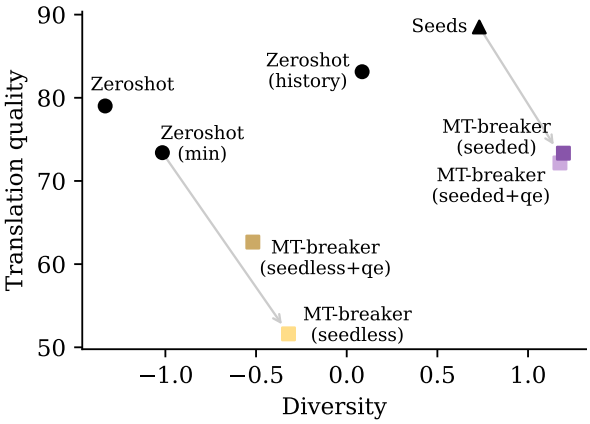

Figure 3: Comparison of difficulty and diversity of difficult-to-translate text generation methods. Diversity is z-normalized across each diversity measure and averaged. Gray arrows signify most fair comparison.

3.2 Data Quality

So far we have investigated only the method's capability of breaking the machine translation model. We now confirm the usability of the method for generating meaningful datasets. Due to the fact that there is no accepted way of intrinsically evaluating a dataset, we propose to examine several different statistics about the dataset. Prior works either do not investigate the data quality (Pombal et al., 2025), rely on guarantees of the data creation process, such as grammatical or human edits (Isabelle et al., 2017; Manakhimova et al., 2023), or resort to manual human inspection (Amrhein et al., 2022; Ármannsson et al., 2024).

For a generated set of examples, we measure the following variables:

- Diversity (in embedding space, pairwise chrF Popović, 2015,[3] unique topics, unique words), to make sure that the texts does not collapse into one subtopic.
- Diversity (errors), to make sure that the error mode is not simply being repeated.
- Source and word length, to make sure the difficulty is not being trivially induced by extremely long segments.
- Word rarity, to analyze if the difficulty stems from simply using rare words.
- Syntax complexity, to analyze if the difficulty is due to convoluted sentence structures.
- Grammaticality, to make sure the text is well-formed and not broken.
- Naturalness, to make sure the text sounds like it could have been written by a human.

See Appendix B for details on how some of these variables are measured using a prompted LLM.

The results in Table 2 show that methods without an explicit treatment for diversity (Zeroshot, Zeroshot min, MT-breaker seedless) indeed lack in the diversity, though this effect is most prevalent in Zeroshot. On the other hand, the methods without diversity guardrails lead to texts of higher difficulty. This tradeoff between diversity and difficulty is shown in Figure 3. If only the difficulty is the concern and not the diversity, MT-breaker seedless is

[3] `ref:1|case:mix|eff:y|nc:6|nw:2|sp:no|v:2.5.1`

		Target model (translation quality)				Target model (diversity)			
		Translate	Gemma3	Gemini2.0	Gemini2.5	Translate	Gemma3	Gemini2.0	Gemini2.5
MT-break. (seedless)	Gemma 3	48.00	57.31	59.77	58.99	0.05	0.19	0.16	0.06
	Gemini 2.0	35.92	42.36	42.97	39.79	-1.03	-0.72	-0.95	-1.09
	Gemini 2.5	52.56	60.51	63.12	62.18	0.48	0.64	0.57	0.65
	Gemini 2.5 pro	44.13	52.45	55.91	53.98	-0.72	-0.31	-0.30	-0.41
Zeroshot (min)	Gemma-3	53.00	68.28	66.31	66.75	-1.00	-0.89	-1.11	-1.23
	Gemini 2.0	45.15	54.34	51.63	55.67	-1.62	-1.60	-1.72	-1.55
	Gemini 2.5	56.68	63.62	65.81	64.31	0.02	0.18	-0.13	0.08
	Gemini 2.5 pro	71.78	72.63	74.59	74.61	-1.03	-0.85	-1.23	-1.05
MT-break. (seeded)	Gemma-3	53.69	63.22	65.05	63.68	0.89	0.85	0.75	0.74
	Gemini 2.0	41.67	52.59	53.56	52.46	1.13	0.71	0.66	0.53
	Gemini 2.5	59.86	68.46	73.11	69.77	1.37	1.34	1.09	1.11
	Gemini 2.5 pro	65.01	74.09	74.47	75.09	1.05	0.90	0.86	0.95
	Seeds	85.63	89.26	89.94	89.23	0.76	0.74	0.44	0.60

Table 3: Results (average quality estimation scores) for using different LLMs (rows) as part of MT-breaker. Left side shows translation quality (lower is better) and right side shows average of z-normalized diversity measures. Averaged across all languages.

an apt choice. If, however, one desires to stay close to some existing dataset in terms of diversity but needs more difficulty, MT-breaker seeded provides a balance between the two.

3.3 Transfer Between MT Models

When MT-breaker targets a specific model, that model's performance is most significantly impacted, indicated by the lowest scores on the diagonal in Table 4, especially in contrast to the low difficulty Seeds. While the difficulty does transfer to other models, the difficulty on other models is generally on par with or slightly lower than Zeroshot. For this specific experiment only due to compute constraints, we also include o4-mini as an additional machine translation model on which we evaluate the generated texts. The model's performance has similar patterns to the other models, with most difficult examples being generated by MT-breaker (seedless).

Targeting the weakest of the models, Google Translate, generally leads to lower difficulty for other models, though this effect is not consistent when targeting more performant models.

The Seeds baseline shows similar performance across Gemma 3, Gemini 2.0 and Gemini 2.5 (89.3, 89.9, 89.2). However, the Zeroshot approach reveals greater differences (74.9, 77.5, 77.2), and MT-breaker accentuates these distinctions even further (54.6, 58.9, 57.1). This shows that difficult testsets or evaluation setups are useful in better benchmarking of models because larger gaps between models generally lead to more statistically significant outcomes.

Finally, we also include a Multi version of MT-breaker, that is optimized against Google Translate, Gemma 3, Gemini 2.0 and Gemini 2.5 at the same time. This modification simply includes the outputs of all models instead of just one and for the final selection takes the average score across all models. This approach does not always lead to the highest difficulty for the individual models, but it is at worst second best and notably better than when transferring from other individual models. It also leads to higher difficulty for o4-mini, a machine translation model that was not part of the optimization, which suggests that the Multi version of MT-breaker taps into a more general notion of difficulty rather than model-specific one.

3.4 Transfer between LLMs

We now explore how the choice of the LLM in LLM_{step} in Algorithm 1 influences the generated data. This is important, as LLMs can have blind spots, such as not being able to generate difficult texts for themselves. For this, we generated the difficult texts using various LLMs that also ultimately serve as the machine translation models. The results in Table 3 suggest that there is no discernible bias. However, it shows that the outputs are systematically different. For example, when optimized with Gemini 2.0 (Zeroshot min, MT-

		Target model				
		Translate	Gemma3	Gemini2.0	Gemini2.5	o4-mini
MT-break. (seedless)	Translate	45.07	68.86	70.68	70.15	77.95
	Gemma3	60.35	54.55	68.76	66.92	76.70
	Gemini2.0	58.12	65.75	58.94	66.18	76.42
	Gemini2.5	56.42	64.89	67.29	57.09	75.18
	Multi	53.30	59.39	63.09	62.44	76.54
Zeroshot (min)	Translate	72.89	77.62	78.87	80.24	83.66
	Gemma3	73.17	74.91	77.45	78.53	82.22
	Gemini2.0	74.17	77.33	77.52	80.78	83.96
	Gemini2.5	73.08	75.71	76.90	77.21	82.33
	Multi	67.46	69.57	69.64	72.51	79.14
MT-break. (seeded)	Translate	66.97	81.66	83.20	82.16	86.05
	Gemma3	75.95	76.67	82.85	82.19	85.66
	Gemini2.0	74.90	80.13	77.02	81.14	84.77
	Gemini2.5	74.83	81.49	83.35	77.35	86.17
	Multi	70.38	77.49	78.29	78.04	85.19
	Seeds	85.63	89.26	89.94	89.23	92.00

Table 4: Results (average quality estimation scores) for difficulty transfer between target machine translation models. MT-breaker and Zeroshot are optimized towards one MT (rows) but the obtained examples are evaluated on another MT (columns). Averaged across all languages.

breaker seeded, MT-breaker seedless), the texts seem to be more difficult. However, this comes at the expense of much lower diversity, which supports the tradeoff findings of Section 3.1.

3.5 Transfer Between Languages

The true translation difficulty is certainly a function of both the source and target languages, because some concepts might be more easily expressed in some languages than others. However, we pose that it is mostly the property of the source language. When generating difficult texts for a specific target language with MT-breaker, the resulting examples are most challenging for that particular language. As shown in Table 5, the lowest translation quality scores consistently appear on the diagonal, where the optimization and evaluation languages match. While some difficulty transfers to other languages, the effect is most pronounced on the targeted one. In contrast, the Zeroshot (min) method shows no discernible or consistent pattern when transferring across different languages, despite the language direction being part of the prompt. Still, even in the non-target languages, the MT-breaker (seedless) leads to higher difficulty than Zeroshot (min).

Translations from other languages into English show similar ranking of methods, as shown in Appendix Figure 4.

		Language Pair				
		EnCs	EnDe	EnZh	EnVi	EnPl
MT-breaker (seedless)	EnCs	42.28	74.40	76.23	71.50	60.61
	EnDe	50.57	59.41	67.32	63.72	52.77
	EnZh	55.63	71.49	60.68	66.32	58.65
	EnVi	56.40	72.74	71.19	57.87	58.03
	EnPl	57.22	74.66	73.04	69.66	49.34
Zeroshot (min)	EnCs	63.42	83.54	83.02	78.56	71.17
	EnDe	59.76	79.78	78.54	74.40	67.29
	EnZh	66.04	83.90	79.24	79.08	71.93
	EnVi	64.04	84.61	81.43	78.58	70.14
	EnPl	72.91	88.62	86.28	84.60	77.13
MT-breaker (seeded)	EnCs	66.95	85.42	84.28	82.68	76.60
	EnDe	73.72	79.90	84.56	83.61	77.98
	EnZh	72.74	85.15	77.56	82.23	77.04
	EnVi	72.03	85.01	83.12	77.04	76.04
	EnPl	72.39	84.18	84.18	83.11	71.06
Seeds		84.19	92.50	90.01	90.31	85.56

Table 5: Results (average quality estimation scores) for difficulty transfer between language pair directions. MT-breaker is optimized on one language pair (rows) but the obtained examples are evaluated on another language pair (columns). Highlights are normalized within columns. Averaged across all models.

3.6 Human Translations and Annotations

To confirm the difficulty of generated texts and to determine if the texts difficult for machine translation models are also difficult for human translators, we take the 100 texts optimized against Gemini 2.5 in English→Czech and English→German and translate them with standard professional human translators. Then, both the Gemini 2.5 and human translations are assessed with professional annotators for translation quality using a mixture of MQM (Freitag et al., 2021) and ESA (Kocmi et al., 2024b) protocols.[4]

The results in Table 6 show similar difficulty patterns to Table 2: MT-breaker (seedless) is the most difficult, followed by Zeroshot and MT-breaker seeded. However, the annotators marked some segments as "untranslateable" for Zeroshot and MT-breaker (seedless) but less so for Seeds or MT-breaker (seeded). Therefore, while ungrounded generations might lead to the highest difficulty, the examples might be too contrived for some applications. Finally, there is very little difference between automatic and human translations, suggesting that the difficulty also transfers to human translators. We include the error category breakdown in Ap-

[4]Each translation gets a final score on the scale of 0 to 100, similar to ESA, but uses the full error taxonomy of MQM. Annotators are translators from a professional language resource company (anonymized) with MQM Athena annotation interface.

QE	Translator	Seeds	Zeroshot (min)	MT-breaker (seedless)	MT-breaker (seeded)
Gem.	Gemini2.5	91.76	79.35	46.93	73.94
	Human	87.22	79.28	59.63	77.66
MetricX	Gemini2.5	87.80	72.54	57.35	78.05
	Human	86.88	73.60	61.60	79.50
MQM	Gemini2.5	93.80	89.08	81.08	86.84
	Human	92.80	92.32	89.20	88.16
ESA	Gemini2.5	68.02	64.92	54.90	65.34
	Human	69.38	64.18	54.54	67.17

Table 6: Human translation of testsets and evaluation results. Shown is a subset where all humans provided the translation. See extended version in Appendix Table 10.

pendix Table 11 which shows that *Mistranslation* and *Creative reinpretation* are much more common with MT-breaker and Zeroshot than Seeds.

3.7 Interview with Experts

To inform the design of MT-breaker, which aims to mimic how human experts find weaknesses in machine translation models, we conducted a qualitative study with seven MT practitioners with at least 2 years of academic or industry experience in developing or evaluating machine translation models. The interview process involved two stages.

First, the participants were asked to interact with a machine translation model and to find its weaknesses. This was done with a concurrent think-aloud protocol (Lewis and Rieman, 1993), where the participants were commenting on their decisions and actions. The interviews were not recorded.

1. When you want to find weakpoints in a machine translation model, what is the first input you use?
2. What kind of language you consider for input?
3. Is your input choice influenced by the kinds of errors you're looking for apriori?
4. Are your next inputs a modification of/related to the first one?
5. How do you decide when to stop pursuing a specific direction before trying a new one?

Findings. Based on qualitative interviews with machine translation experts, a common strategy for identifying model weaknesses involves an iterative and targeted approach. Experts often begin with known challenging inputs, such as idioms and colloquialisms (I1, I5, I7), archaic expressions (I4), or specific domains like song lyrics (I3), drawing on their prior knowledge of typical MT failure points. All interviewees (I1 to I7) at some point searched and content from the Internet as the input, which they then manually edited. Upon finding a minor error, they tend to modify the input iteratively to provoke a larger failure, rather than switching to a new type of input immediately (I1 to I6). The process is guided by their intuition about which part of the current input triggered an error. Finally, some experts consider even non-standard input in order to trigger some unexpected behavior, like hallucinations. This included ungrammatical sentences (I5, I6), wrong formatting (I3 to I7), repeating words multiple times (I5 to I7), mixed or incorrect language input (I2, I5, I6, I7), or random sequence of characters (I5, I6). The patience in how long a certain input direction is being exploited when no error is found varies from 1-2 (I1, I5), 5 (I4), to 10 (I2, I3, I6).

Based on this, we designed MT-breaker to iteratively modify and exploit a specific input up to the depth of 10. Contrary to some of the practices of the experts, we instructed LLM_{step} to only make grammatical edits, so that we could fairly compare the methods. Additionally, the quality estimation metrics are trained on mostly grammatical inputs, so they might not be able to accurately assess outputs with out-of-distribution error modes.

4 Related Work

Our work builds upon three lines of research for creating challenging test sets for machine translation: selecting examples from existing corpora, generating examples from scratch, and creating adversarial examples.

Selecting examples. A common approach to building challenge sets is to select difficult examples from large, existing corpora. Chen et al. (2023) select sentences based on lexical rarity, structural complexity (sentence length), syntactic rarity (low-frequency parse trees), and model-based difficulty (high translation entropy). Proietti et al. (2025) train a source-based metric that predicts difficulty. While effective, these selection-based methods are limited by the diversity of the source corpora and can fail to uncover novel or complex failure modes not present in naturally occurring text. Lastly, Zouhar et al. (2025) show that samples with more errors are more useful for comparing models.

Generating examples. To overcome the constraints of selection, generative approaches create new examples. Early and highly-controlled methods rely on manual, linguistically-informed construction (Isabelle et al., 2017), or on semi-automated rule-based models to evaluate specific linguistic phenomena, which are then verified by human experts (Manakhimova et al., 2023).

Closest to our work is zero-shot benchmarking (Pombal et al., 2025), which uses LLMs at scale to generate examples from scratch in a zero-shot manner but with a control of the domain and, importantly, difficulty. However, such from-scratch generation often lacks sufficient difficulty and diversity for state-of-the-art models. Our work combines the scalability of LLM generation with preserving diversity through seed texts and the targeted difficulty through an iterative feedback loop.

Adversarial examples. The goal of adversarial examples is not to create inherently difficult text, but to induce model failure through minimal, often imperceptible, perturbations to an input that is otherwise easy to translate. Zou et al. (2020) use reinforcement learning to apply token-level edits that degrade translation quality while preserving the meaning of the source text. Similarly, Zhang et al. (2021) apply a sequence of heuristic edits to probe model brittleness to superficial input changes.

Interactive discovery. Several works have investigated mimicking the human expert who probes the model for weaknesses. Lu et al. (2025) propose Automated Capability Discovery, where one foundation model acts as a "scientist" to autonomously generate open-ended tasks (from a template) that reveal the capabilities and failures of another model. Similarly, Tjuatja and Neubig (2025) introduce BehaviorBox, a method to automatically discover fine-grained, interpretable features where one model systematically outperforms another. None of these works, however, made use off the specifics of machine translation, where we have access to accurate quality estimation metrics.

5 Conclusion

Discussion. The MT-breaker (intentionally) introduces a bias against the particular machine translation model which it is trying to break and find and amplify the weaknesses of. Consequently, the generated dataset is tailored to that model's particular vulnerabilities and cannot be used for a fair comparison across different models. Using this dataset to evaluate other models would be misleading, as it would disproportionately favor models that do not share the same specific weaknesses as the target model. Instead, we suggest using MT-breaker in two of the following ways: (1) Use the MT-breaker-generated dataset against a particular machine translation model m_1 to find its weaknesses, which can be used to hillclimb on. (2) Compare the scores of two MT-breaker-generated datasets against m_1 and m_2 (i.e. try to break m_1 and m_2 independently) to discover the models' worst-case performances.

Summary. We introduced MT-breaker, a novel simple method that mimics human experts to iteratively generate texts that are difficult to translate. This approach increases translation difficulty for specific machine translation models (confirmed by an expert human evaluation) while maintaining the naturalness and diversity found in the original seed texts. While the difficulty is most pronounced for the targeted model and language, it also transfers to other models and languages.

Future work. MT-breaker uses signal from the specific model to obtain breaking inputs. However, we hypothesize that some texts generated by both Zeroshot and MT-breaker methods may be difficult for humans to translate, not for machine translation models. This is because many sources in the training data (Internet) discuss difficulty with respect to a human translator, so LLM_{step} might mimic this.

Limitations

We have relied heavily on the quality estimation metric as part of both the optimization objective and the quantity of interest. To this end we used Zeroshot (min) variant, which uses a similar selection process to make the comparison more fair, and also confirm difficulty with human evaluation.

The Zeroshot method introduced by Pombal et al. (2025) is conditioned on specific domains. This could lead to more diverse outputs, though we did not include this in our implementation of Zeroshot so that it is more comparable to the other methods.

Lastly, the created texts fall outside of the expected user distribution (see Naturalness in Table 2). However, this is intentional as most texts are not challenging (Proietti et al., 2025) and we wish to focus on the more tail of the distribution with less likely but more difficult texts that reveal more the models' worst-case performances.

Ethics Statement

The data used within the human study are sourced from the WMT 2024 testset (Kocmi et al., 2024a) and have thus already been screened for disturbing content. The translators and annotators were fairly compensated, consented to participate in a research study, and no personal data was collected.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinthor Steingrimsson. 2024. Killing two flies with one stone: An attempt to break LLMs using English-Icelandic idioms and proper names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted challenge set for evaluating machine translation performance. In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Clayton Lewis and John Rieman. 1993. *Task-Centered User Interface Design: A Practical Introduction*. University of Colorado, Boulder, CO.

Cong Lu, Shengran Hu, and Jeff Clune. 2025. Automated capability discovery via foundation model self-exploration. *Preprint*, arXiv:2502.07577.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform

NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models. *Preprint*, arXiv:2504.01001.

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. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Lindia Tjuatja and Graham Neubig. 2025. BehaviorBox: Automated discovery of fine-grained performance differences between language models. *Preprint*, arXiv:2506.02204.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977. Association for Computational Linguistics.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2020. A reinforced generation of adversarial examples for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3486–3497. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Preprint*, arXiv:2501.18251.

	CsEn	DeEn	ZhEn	ViEn	PlEn
MT-breaker (seedless)	55.17	63.78	50.19	63.58	53.84
Zeroshot (min)	61.52	74.30	67.88	76.05	64.56
MT-breaker(seeded)	74.85	80.29	82.97	78.92	76.15
Seeds	90.02	92.81	94.47	92.75	88.91

Figure 4: Results for difficulty for translation into English, compare to diagonal in Table 5.

A Source-only MetricX

We also trained a source-only version of MetricX to estimate the difficulty without the translation model's output. Similar to Sentinel-src (Perrella et al., 2024; Proietti et al., 2025) or PreCOMET (Zouhar et al., 2025), the input to the source-only MetricX is only the source text. The model is trained to predict the average human assessment score aggregated across all translations for the given source in the data. For example, if two models translate the same source and their human assessments are 76 and 98 MQM scores, the source-only model will be trained to produce 87. The rest of the training follows the procedure of MetricX-23 (Juraska et al., 2023).

B Prompts

The first prompt is the same across all methods: **Zeroshot, MT-breaker.** We want to find a text in {LANG1} that's exceptionally difficult for a machine translation model to translate into {LANG2}. The goal is to expose a wide range of translation errors and severely challenge the MT model's capabilities. The text should be approximately {SEED_LENGTH} words. At the end of your response write SOURCE |||<SOURCE_TEXT>|||'.

MT-breaker (qe), follow-up. This is the machine translation of the source text.
TRANSLATION |||{TARGET_TEXT}|||
And this is the score of the translation: SCORE |||{QE_SCORE:.1f}%||| (0=bad translation, 100=perfect)
Try to make it even more difficult to translate. Modify at most 75% of the best SOURCE (one with lowest score) so far.

Source text analysis. Analyze the following text and return the answer in JSON. We want to determine the following attributes:
- grammaticality: on a scale from 0 (complete gibbersish) to 100 (perfect, no errors).
- naturalness: on a scale from 0 (wholy unnatural) to 100 (fully human-like and would occur in a corpus).
- word rarity: on average, how rare are the words from 0 (average modern human would use this word daily) to 100 (average modern human would not understand the word).
- syntax complexity: on a scale from 0 (simplest possible sentence) to 100 (most complex and hard to understand).
- topics: list of 1 to 5 topics that the text is about.

Provide only the output in JSON and nothing else. The output should look like this (no extra backticks or newlines):
{
"grammaticality": 90,
"naturalness": 80,
"word rarity": 50,
"syntax complexity": 70,
"topics": ["science", "technology"],
}
The sentence to analyze is:
> {SOURCE_TEXT}

Target text analysis. Analyze the following translation and return a list of reasons the translation might be incorrect (in JSON). The options are: "idioms", "metaphor", "gender", "style", "ambiguity", "named entities", "numerical expressions", "dates and times", "cultural nuances", "tone", "context", "syntax", "semantics", "lexical choice", "register", "collocations", "terminology", "cohesion and coherence", "omissions", "additions", "misinterpretations", "negation", "tense and aspect", "modality", "pronoun resolution", "punctuation", "formatting".
Feel free to add extra if you think it's appropriate. Output a list of reasons (even 0, depending on if the translation is erroneous). Provide only the output in JSON and nothing else. The output should look like this (no extra backticks or newlines):

"error_modes": ["idiom", "gender", "style"],

The sentence to analyze is:
> {SOURCE_TEXT}
The translation is:
> {TARGET_TEXT}

LLM-as-QE. Evaluate the quality of the translation on a scale from 0 100. Roughly:
100 - Perfect
95 - Excellent (closely aligned with the source)
80 - Very good (minor style choice)
60 - Fair (some inaccuracies or fluency errors)
40 - Poor (multiple inaccuracies or fluency errors)
0 - Inadequate (unrelated, completely wrong)
First, think about all the errors in the translation and their severity (very briefly, max few words per error). At the end, output a single line in the format like as follows: 'SCORE |||70.8|||' The last line is important because it will be matched with a regex, so make sure to use the |.

SOURCE: |||{SOURCE_TEXT}|||
TRANSLATION: |||{TARGET_TEXT}|||

C Data Examples

Zeroshot (min)	Seeds	MTbreaker (seeded)
It is what it is.	Another one has been found!	Another rabbit hole of lies was found!
The grizzled grackle, having grabbled a grub, gargled a guttural greeting. Its gambrel legs, like spindly stilts, straddled a gnarled branch, its gaze a glassy glint. The gloaming gathered, a gossamer shroud, as the grotesque gargoyle, its grimace a grotesque geometry, glowered from the gable. Below, a gaggle of geese honked a garrulous goodnight, their glossy plumage a ghostly gleam in the waning light.	We're just going to let that sit and rehydrate for a couple minutes. With ultralight cooking, we're not actually cooking on the stove. We're primarily boiling water so that we can rehydrate our food. It's also one of the reasons why we use instant grits. Less cooking time, less fuel, and less weight. Oh yeah. So this is about what we've got going on here. Perfect consistency.	We're gonna let this grub rehydrate a touch. Ultralight cooking's not really about cooking, per se. We mostly just boil water to rehydrate our grub. That's why we use instant grits, man. Less cook time, fuel, and weight. Gnarly, a perfect consistency. So this is what we've got going on right here. Far out. We're gonna grub on this for a minute. It's gonna be so bomb, bro. Fire, no cap. Let's get this bread. It's bussin'. This slaps.
The quick brown fox jumps over the lazy dog.	I quit FB, removed all Meta and Pinterest tracker from my website and killed my private WhatsApp account.	I've quit FB and killed the private WhatsApp account to get rid of Meta's and Pinterest's gross ad-tracking tools that I hated with a fiery passion.
The wily, woolgathering wordsmith, having just quaffed a flagon of kvass, began weaving a web of wonton words. He spun a yarn about a gaggle of geese, each given a gimcrack gewgaw, gossiping about a gimpy gnu. Their gabbles grew grandiose, morphing into a grandiloquent gallimaufry of gobbledygook. The gnu, a grizzled grammarian, grimaced, groaning at the grammatical grotesqueries. This ghastly gibberish, he groused, was a genuine gerrymander of language, a grotesque gargoyle of grammar. What a galling, garrulous goose-chase! I'm totally knackered.	The Department of Justice's overall approach to corporate crimes has come under intense scrutiny from advocates. While watchdogs have lauded some steps the department has taken during Merrick Garland's tenure as attorney general - like finally establishing a database on corporate crime - Biden's DOJ has also leaned heavily on leniency agreements that allow companies to defer or avoid prosecution, and encouraged companies to scapegoat individual employees so as to avoid a broader charge. Kenneth Polite Jr., former assistant attorney general for DOJ's Criminal Division, revised the division's corporate enforcement policy so as to limit prosecutions.	The Department of Justice's overall approach to corporate crimes has come under intense scrutiny from advocates. While watchdogs have lauded some steps the department has taken during Merrick Garland's tenure as attorney general - like finally establishing a database on corporate crime - Biden's DOJ has also leaned heavily on leniency agreements that allow companies to defer or avoid prosecution, and encouraged companies to scapegoat individual employees so as to avoid a broader charge. Kenneth Polite Jr., former assistant attorney general for DOJ's Criminal Division, revised the division's corporate enforcement policy so as to limit prosecutions.
The jabberwocky, with eyes of flame, came whiffling through the tulgey wood, and burbled as it came! 'Twas brillig, and the slithy toves did gyre and gimble in the wabe; all mimsy were the borogoves, and the mome raths outgrabe.	As they drew closer, the little heat signatures turned from blobs to more distinct shapes. As they drew closer, one of them ran for something and grabbed it before rusing back to the small ridge they were on. The figures dropped down into firing positions.	As they drew closer, the little heat signatures turned from blobs to more distinct shapes. As they drew closer, one of them ran for something and grabbed it before rusing back to the small ridge they were on. The figures dropped down into firing positions.
He bade them farewell and then he bade them all to hell.	Going back up tomorrow and we're doing stalls and slow flight.	Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.
It is what it is.	Heheh not one but three!	Heheh not one, two, but three!!!11!oneeleven
The cantankerous clockmaker, a man of mercurial moods and intricate mechanisms, grumbled as he tinkered with a recalcitrant timepiece. Its delicate gears, a labyrinth of brass and steel, refused to cooperate, their stubborn silence echoing his own frustration. He muttered a litany of obscure horological terms, his voice a gravelly counterpoint to the ticking of a dozen other clocks. Each tick was a tiny hammer blow against his patience, a reminder of time's relentless march. He yearned for a world without time, a place of blissful stasis where gears wouldn't jam and springs wouldn't break.	0430 Itania time. Cohren hated waking up this early. Mainly because it was his responsibility to wake everyone else up in his company. Silently stepping out of his bunk, he put on his combat fatigues, throwing on his assault vest, and packed up his belongings from his locker, being careful with a particular small box. He shoulded his rifle and put his helmet on a small hook connected to his belt. It had all its fixing, respirator, goggles, everything needed. After checking his ammo pouches, he quietly walked out of the barracks, mud and ice squelching under his boots.	0430 Itania time, not that Cohren cared for punctuality. The wee hours bit, stingier than a scrooge's handshake. His lot: rousing the grunts. He crept from his bunk, a ghost in combat gear, shrugging on a vest heavy with unspoken promises. From his locker, he snagged his kit, a particular small box handled with the reverence of a holy relic. Rifle kissed his shoulder, helmet dangled from his belt like a shrunken head, bristling with fixings – respirator, goggles, the whole shebang. Ammo? Check. He slunk out, the pre-dawn muck groaning under his boots, a symphony of squelch.
The grizzled grackle, having grabbled a grub, gargled a guttural garble, a gratingly grandiloquent guffaw that grated on the great, grey granite gravestones. It was a ghastly, ghoulish glee, a grim gambol against the gloomy gloaming, a grotesque gesture against the gathering gloom. Gravely, the groundskeeper grumbled, "Gadzooks! That garrulous grackle gives me grave grievances!"	Stealthily I made my way towards the back of the store, murmurs and talking heard beyond a metal grate which took a few yanks to tear off. Climbing into the vent was the safest and quietest option for me right now; climbing my way through, I'd pick up on conversations and talks between strangely dressed men. I noticed Al inside, chained. Figures he'd get caught eventually.	With gusto, I wriggled my way toward the back of the entrepot, susurrations and parleying heard beyond a wrought iron grille which took a few goes to heave off. Subterfuge by vent was the soundest and hushest option for me right now; scrabbling my way along, I'd cotton on to consultations and confabs between queerly-garbed men. I clocked Al inside, shackled. Figures he'd get his collar felt eventually. I had to spring him.

Table 7: Example of generated sources in English by Zeroshot, Seeds, and MT-breaker (seeded). Extends Table 1 from the main text.

Step 0 QE (Gemini): 65%, QE (MetricX): 72%
Source: Going back up tomorrow and we're doing stalls and slow flight.
Translation: Zítra jdeme zase nahoru a budeme dělat pádové rychlosti a pomalý let.

Step 1 QE (Gemini): 45%, QE (MetricX): 79%
Source: Going back up tomorrow; we're doing stalls and spins, and then slow flight.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak pomalý let.

Step 2 QE (Gemini): 45%, QE (MetricX): 78%
Source: Going back up tomorrow; we're doing stalls and spins, and then some slow flight.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak pomalý let.

Step 3 QE (Gemini): 40%, QE (MetricX): 63% •
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak neobvyklé polohy.

Step 4 QE (Gemini): 65%, QE (MetricX): 74%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recoveries.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik neobvyklých poloh.

Step 5 QE (Gemini): 45%, QE (MetricX): 66%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak nácvik neobvyklých poloh.

Step 6 QE (Gemini): 70%, QE (MetricX): 73%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.

Step 7 QE (Gemini): 55%, QE (MetricX): 72%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik neobvyklých poloh.

Step 8 QE (Gemini): 62%, QE (MetricX): 73%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.

Step 9 QE (Gemini): 60%, QE (MetricX): 73%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
Translation: Zítra jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.

Step 10 QE (Gemini): 48%, QE (MetricX): 62%
Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
Translation: Zítra jdeme zase nahoru; budeme dělat pády a výkruty a pak nácvik neobvyklých poloh.

Table 8: Example of one run of MT-breaker across 10 steps for English→Czech. Step 3 • is selected as the final candidate because of lowest average score. Underlined are major errors.

Step 0 QE (Gemini): 100%, QE (MetricX): 88%

Source: Going back up tomorrow and we're doing stalls and slow flight.

Translation: Morgen geht es wieder hoch und wir machen Strömungsabrisse und Langsamflug.

Step 1 QE (Gemini): 82%, QE (MetricX): 80%

Source: Going back up tomorrow; we're doing stalls and coffin corner.

Translation: Morgen geht es wieder hoch; wir machen Strömungsabrisse und den kritischen Geschwindigkeitsbereich.

Step 2 QE (Gemini): 100%, QE (MetricX): 82%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice.

Translation: Morgen geht es wieder hoch; wir machen Strömungsabriss- und Coffin Corner-Übungen.

Step 3 QE (Gemini): 100%, QE (MetricX): 83%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills.

Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und zum "Coffin Corner".

Step 4 QE (Gemini): 95%, QE (MetricX): 87%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it.

Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das hinzubekommen.

Step 5 QE (Gemini): 85%, QE (MetricX): 74% •

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride.

Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt perfekt hinzubekommen.

Step 6 QE (Gemini): 95%, QE (MetricX): 73%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam.

Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt perfekt hinzubekommen.

Step 7 QE (Gemini): 75%, QE (MetricX): 76%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...wish me luck.

Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt hinzubekommen...drückt mir die Daumen.

Step 8 QE (Gemini): 88%, QE (MetricX): 74%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...wish me luck...not.

Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt hinzukriegen ... wünscht mir Glück ... nicht.

Step 9 QE (Gemini): 70%, QE (MetricX): 71%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...not.

Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem Sargkurvenbereich, um das bei der Prüfungsfahrt perfekt hinzubekommen...nicht.

Step 10 QE (Gemini): 72%, QE (MetricX): 73%

Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...not!

Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um es bei der Prüfungsfahrt perfekt hinzubekommen...nicht!

Table 9: Example of one run of MT-breaker across 10 steps for English→German. Step 5 • is selected as the final candidate because of lowest average score. Underlined are major errors.

D Human Evaluation

EnCs+EnDe average					
QE	**Translator**	**Seeds**	**Zeroshot** (min)	**MT-breaker** (seedless)	**MT-breaker** (seeded)
Gemini	Gemini2.5	91.76	79.35	46.93	73.94
Gemini	Human	87.22	79.28	59.36	77.66
MetricX	Gemini2.5	87.80	72.54	57.35	78.05
MetricX	Human	86.88	73.60	61.60	79.50
MQM	Gemini2.5	93.80	89.08	81.08	86.84
	Human	92.80	92.32	89.20	88.16
ESA	Gemini2.5	68.02	64.92	54.90	65.34
ESA	Human	69.38	64.18	54.54	67.17
Gemini	Gemini2.5 (NS)	92.00	40.25	57.79	55.67
MetricX	Gemini2.5 (NS)	95.71	51.94	64.33	71.96
NS	Human	0.5%	4.0%	28.5%	1.5%

EnCs						EnDe			
QE	**Translator**	**Seeds**	**Zeroshot** (min)	**MT-breaker** (seedless)	**MT-breaker** (seeded)	**Seeds**	**Zeroshot** (min)	**MT-breaker** (seedless)	**MT-breaker** (seeded)
Gemini	Gemini2.5	91.78	76.66	56.64	73.11	91.74	82.10	27.11	74.79
Gemini	Human	85.84	75.77	66.12	74.84	88.60	82.87	45.55	80.58
MetricX	Gemini2.5	83.11	64.22	47.38	69.42	92.53	81.04	77.71	86.95
MetricX	Human	80.93	62.35	52.70	70.22	92.89	85.09	79.79	89.07
MQM	Gemini2.5	90.04	85.40	72.44	78.48	97.60	92.76	89.82	95.20
MQM	Human	89.84	91.08	90.96	82.08	95.76	93.56	87.48	94.28
ESA	Gemini2.5	56.82	57.19	51.94	55.80	79.21	72.65	57.13	74.89
ESA	Human	58.94	57.46	56.26	57.15	79.82	70.90	53.53	77.19
Gemini	Gemini2.5 (NS)	-	00.00	19.50	-	92.00	64.40	60.68	55.67
MetricX	Gemini2.5 (NS)	-	10.87	20.42	-	95.71	76.59	67.64	71.96
MQM	-	80.00	40.00	-		100.00	96.80	85.16	94.68
ESA	Gemini2.5 (NS)	-	52.00	49.00	-	51.00	66.00	54.71	65.00
NS	Human	0.0%	3.0%	4.0%	0.0%	1.0%	5.0%	53.0%	3.0%

Table 10: Human translation and annotations. QE data is shown on a subset where all humans provided the translation and 'NS' shows the segments that were not translated by human translators. Tables are split across two language directions (EnCs and EnDe) and also averaged. The MQM is converted to 100+4·MQM to be on similar scale as MetricX. The tables xtend Table 6.

	Error type	All	EnCs	EnDe	Gemini2.5	Human
Seeds	Accuracy/Mistranslation	18.5%	21.8%	14.1%	18.6%	18.3%
	Style/Unnatural or awkward	16.1%	22.1%	8.1%	16.7%	15.5%
	Accuracy/Creative Reinterpretation	6.1%	6.5%	5.6%	3.8%	8.3%
	Fluency/Grammar	4.1%	6.8%	0.4%	4.2%	4.0%
	Fluency/Punctuation	3.3%	2.3%	4.7%	2.7%	4.0%
	Style/Bad sentence structure	4.4%	5.9%	2.6%	5.7%	3.2%
	Accuracy/Source language fragment	2.0%	0.7%	3.8%	2.7%	1.4%
	Accuracy/Omission	2.6%	1.6%	3.8%	2.7%	2.5%
	Fluency/Spelling	1.8%	2.0%	1.7%	0.0%	3.6%
	Accuracy/Gender Mismatch	0.9%	1.6%	0.0%	0.4%	1.4%
	Accuracy/Addition	0.4%	0.3%	0.4%	0.0%	0.7%
	Style/Archaic or obscure word choice	0.6%	1.0%	0.0%	1.1%	0.0%
	Terminology/Inappropriate for context	0.9%	0.7%	1.3%	0.8%	1.1%

	Error type	All	EnCs	EnDe	Gemini2.5	Human
Zeroshot (min)	Accuracy/Mistranslation	27.8%	27.2%	28.4%	31.1%	24.1%
	Style/Unnatural or awkward	14.6%	18.8%	9.7%	15.4%	13.7%
	Accuracy/Creative Reinterpretation	10.2%	9.2%	11.4%	8.2%	12.4%
	Fluency/Grammar	4.3%	7.0%	1.3%	4.9%	3.7%
	Fluency/Punctuation	3.9%	4.4%	3.4%	4.9%	2.9%
	Style/Bad sentence structure	2.2%	2.9%	1.3%	2.2%	2.1%
	Accuracy/Source language fragment	3.3%	1.5%	5.5%	5.6%	0.8%
	Accuracy/Omission	1.4%	2.2%	0.4%	0.0%	2.9%
	Fluency/Spelling	2.2%	2.2%	2.1%	0.4%	4.1%
	Accuracy/Gender Mismatch	0.8%	1.5%	0.0%	0.7%	0.8%
	Accuracy/Addition	1.4%	1.1%	1.7%	0.4%	2.5%
	Style/Archaic or obscure word choice	1.0%	0.4%	1.7%	1.5%	0.4%
	Terminology/Inappropriate for context	0.2%	0.0%	0.4%	0.0%	0.4%

	Error type	All	EnCs	EnDe	Gemini2.5	Human
MT-breaker (seeded)	Accuracy/Mistranslation	29.8%	35.4%	21.3%	30.2%	29.4%
	Style/Unnatural or awkward	16.9%	21.6%	9.7%	19.2%	14.4%
	Accuracy/Creative Reinterpretation	9.2%	10.1%	7.9%	8.8%	9.7%
	Fluency/Grammar	3.3%	4.9%	0.7%	2.5%	4.1%
	Fluency/Punctuation	5.5%	4.7%	6.7%	5.1%	5.9%
	Style/Bad sentence structure	3.0%	4.4%	0.7%	3.1%	2.8%
	Accuracy/Source language fragment	3.4%	1.2%	6.7%	4.0%	2.8%
	Accuracy/Omission	1.6%	0.7%	3.0%	2.0%	1.2%
	Fluency/Spelling	1.6%	1.7%	1.5%	0.0%	3.4%
	Accuracy/Gender Mismatch	1.2%	2.0%	0.0%	1.1%	1.2%
	Accuracy/Addition	0.6%	0.7%	0.4%	1.1%	0.0%
	Style/Archaic or obscure word choice	0.6%	0.5%	0.7%	0.3%	0.9%
	Terminology/Inappropriate for context	0.9%	0.5%	1.5%	1.4%	0.3%

	Error type	All	EnCs	EnDe	Gemini2.5	Human
MT-breaker (seedless)	Accuracy/Mistranslation	33.1%	35.7%	22.4%	39.0%	24.4%
	Style/Unnatural or awkward	17.9%	17.4%	20.0%	21.2%	13.1%
	Accuracy/Creative Reinterpretation	12.9%	14.0%	8.2%	15.1%	9.7%
	Fluency/Grammar	5.3%	6.6%	0.0%	5.0%	5.7%
	Fluency/Punctuation	3.2%	3.4%	2.4%	1.5%	5.7%
	Style/Bad sentence structure	1.1%	1.4%	0.0%	0.8%	1.7%
	Accuracy/Source language fragment	1.4%	1.4%	1.2%	1.5%	1.1%
	Accuracy/Omission	2.3%	1.4%	5.9%	1.2%	4.0%
	Fluency/Spelling	2.1%	1.4%	4.7%	0.4%	4.5%
	Accuracy/Gender Mismatch	1.4%	1.7%	0.0%	0.8%	2.3%
	Accuracy/Addition	1.4%	1.4%	1.2%	1.2%	1.7%
	Style/Archaic or obscure word choice	0.7%	0.6%	1.2%	0.4%	1.1%
	Terminology/Inappropriate for context	0.2%	0.3%	0.0%	0.0%	0.6%

Table 11: Distribution of MQM error categories. Only categories with at least 10 overall occurences are shown. The percentages are normalized for each column in each table independently. Only segments where human translation exists are considered.