Evaluating Out-of-Distribution Generalization in German-to-English Neural Machine Translation

Varun Rayamajhi
School of Informatics, University of Edinburgh

2025

Abstract

We train an encoder–decoder Transformer for German-to-English neural machine translation and study how its behavior changes on sentences that are longer than those seen during training. We compare our model with a pretrained translation model on in-distribution (i.i.d.) and out-of-distribution (o.o.d.) splits using loss curves, manual error annotation, BLEU, ChrF, output-length ratio, bigram repetition rate, and cross-attention coverage. Our model learns the training distribution without severe overfitting, but its o.o.d. validation loss flattens early and its translations become substantially shorter and less accurate. Manual analysis shows that omission and unsupported addition are the dominant o.o.d. errors. We also find that low attention coverage is a useful warning sign for under-translation, although it is noisy and model-dependent. Finally, tuning beam size and length penalty improves all four o.o.d. metrics, but only modestly. The results suggest that decoding changes can reduce some visible failures, while stronger out-of-distribution generalization requires changes to the training data, objective, or model architecture.

1 Introduction

Neural machine translation systems are usually evaluated on data that closely resembles their training distribution. This is useful for measuring average translation quality, but it does not fully explain what happens when sentence length or structure changes. In this report, we focus on that gap. We train a German-to-English encoder–decoder Transformer on a small dataset of short sentences and evaluate it on both an i.i.d. validation split and an o.o.d. split containing longer sentences of approximately 10–20 words.

Our goal is not only to report a single translation score. We ask four related questions. First, do the training curves already reveal the generalization problem? Second, what kinds of errors appear when the model leaves its training distribution? Third, can cross-attention coverage help identify under-translation? Finally, how much of the problem can be addressed by changing the decoding strategy without retraining the model?

To put our model in context, we compare it with a pretrained German-to-English translation model. The comparison is not intended as a matched-capacity benchmark: the pretrained model has been exposed to a much larger and more diverse corpus. Instead, it helps us distinguish failures caused by limited training data from failures introduced by evaluation and decoding choices.

2 Experimental Setup

2.1 Models and data splits

Our model is an encoder–decoder Transformer following the attention-based architecture introduced by Vaswani et al. [4]. It is trained on short German-to-English sentence pairs. The i.i.d. validation split follows the same sentence-length regime, while the o.o.d. split contains longer sentences that do not match the training conditions. We compare this model with a pretrained translation model on the same splits.

We evaluate four model–split combinations: our model on i.i.d. and o.o.d. data, and the pretrained model on i.i.d. and o.o.d. data. For the quantitative analysis, we use BLEU [1], ChrF [2], output-length ratio, and bigram repetition rate (REPR). BLEU and ChrF measure overlap with reference translations at the word and character levels. Length ratio helps reveal outputs that are too short or too long, while REPR tests for degenerative repetition.

2.2 Manual error taxonomy

Automatic metrics do not explain how a translation fails, so we also annotate six error types: omission, addition or hallucination, grammaticality or fluency, named entities/numbers/punctuation, incompleteness or truncation, and idioms/formulae/register. These labels are not always independent. In particular, truncation can induce both omission and grammatical errors. We therefore interpret the counts as a description of observed failure modes rather than mutually exclusive causes.

3 Training Behavior

Figure 1 shows the training, i.i.d. validation, and o.o.d. validation losses. Under correct training, we expect the training loss to decrease steadily, the i.i.d. validation loss to follow the same trend while remaining slightly higher, and the o.o.d. loss to decrease more slowly and remain above both.

The curves align with the first two expectations. Training and i.i.d. validation losses decrease together, suggesting that the model is learning and generalizing to i.i.d. sentences without severe overfitting despite the small dataset. The o.o.d. loss initially decreases but then flattens. By roughly 8,000 steps, training and i.i.d. losses are still improving slowly, while the o.o.d. loss remains nearly unchanged.

We therefore expect additional training steps to improve i.i.d. performance, but not to solve the o.o.d. problem on their own. The mismatch is structural: longer sentences create alignment and generation demands that are not represented adequately in the training data. Improving o.o.d. performance likely requires a larger or more varied dataset, different training hyperparameters, or architectural changes.

4 Quantitative Evaluation

Table 1 reports the four evaluation metrics. The pretrained model outperforms our model on both splits. Our model also experiences a much larger drop between i.i.d. and o.o.d. evaluation.

For our model, BLEU drops from 0.294 to 0.096 and ChrF from 47.46 to 30.46. This indicates a substantial reduction in lexical and morphological overlap with the references, which matches the

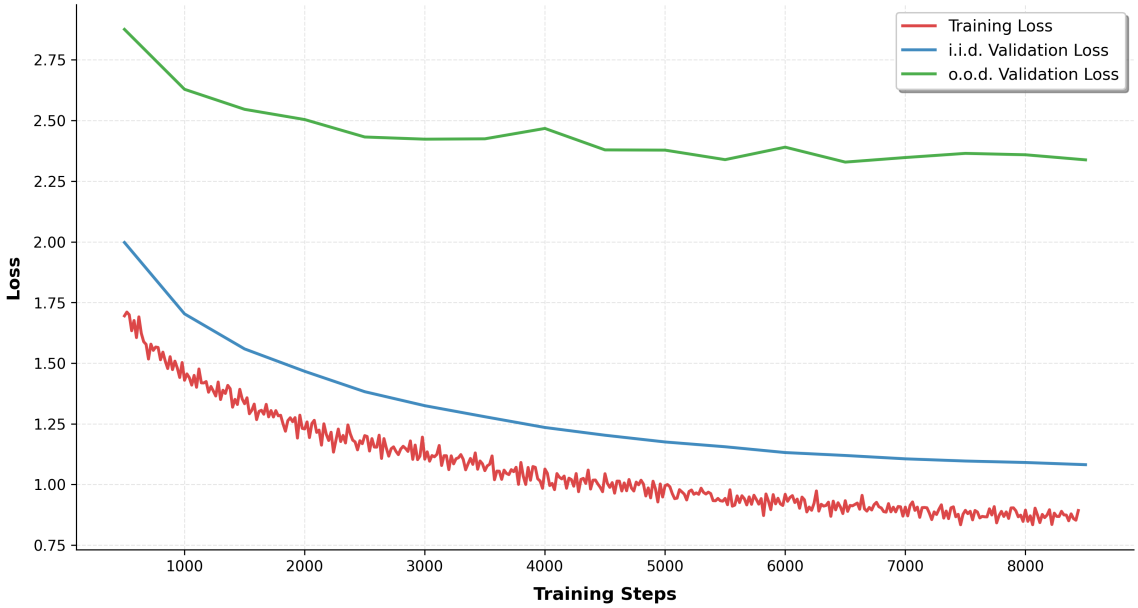

Figure 1: Training and validation losses for our model. The o.o.d. validation loss flattens early while the training and i.i.d. losses continue to decrease.

Table 1: Performance on the i.i.d. and o.o.d. splits.

Model	**BLEU**	**ChrF**	**Length ratio**	**REPR**
		I.I.D. split		
Our model	0.294	47.46	1.027	0.0036
Pretrained model	0.411	62.34	1.021	0.0004
		O.O.D. split		
Our model	0.096	30.46	0.588	0.0038
Pretrained model	0.348	58.54	0.935	0.0031

increase in omissions, additions, and grammatical errors found manually. The pretrained model degrades more moderately: BLEU decreases from 0.411 to 0.348, and ChrF from 62.34 to 58.54.

Length ratio is especially revealing. Both models produce outputs close to the source length on the i.i.d. split. On o.o.d. data, the pretrained model falls moderately to 0.935, whereas our model falls to 0.588. This is consistent with frequent under-translation and skipped clauses. REPR remains close to zero for all combinations, so neither model shows strong degenerative bigram loops.

The metrics also complement the error taxonomy. BLEU is sensitive to omitted, added, or truncated word sequences. ChrF responds to similar failures at the character level and is also informative for malformed words and named entities. Length ratio directly exposes outputs shortened by omission or truncation, while REPR detects one specific failure mode in which repeated phrases fill the output.

5 Error Analysis

Our model produces substantially more annotated errors than the pretrained model. Figure 2 shows that its total error count rises from 47 on i.i.d. examples to 72 on o.o.d. examples, driven mainly by omission and addition. The pretrained model remains almost error-free on i.i.d. data, but reaches 20 errors on the o.o.d. split, primarily through omission, fluency problems, and truncation.

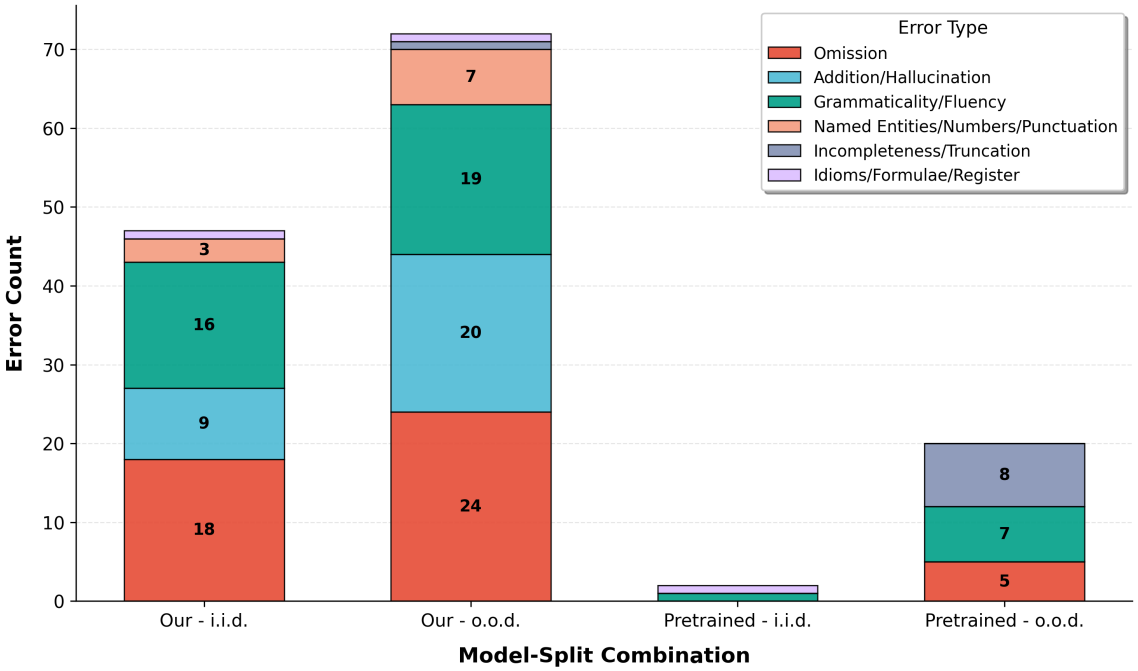

Figure 2: Distribution of manually annotated errors for each model–split combination.

Omission is the most frequent error for our model in both splits. With a small training corpus of short sentences, beam search often favors short fragments and drops lower-probability content instead of attempting a complete translation. This becomes more pronounced on longer o.o.d. sentences, where encoder–decoder alignments are weaker and later clauses are skipped.

For example, given *"Eine Lösung wäre ein großer Durchbruch, um 2011 zu einem vielversprechenden Jahr für eine Erweiterung zu machen."*, our model produces *"One solution would be a great deal for Iraq."* The output omits the central meaning and introduces "Iraq," which is unsupported by the source. Other failures include corrupted names ("Roth-Behrendt" becoming "Roth-in-Office"), malformed output such as *"What is the people of these people's people?"*, and failure to preserve idiomatic meaning.

The pretrained model's most frequent o.o.d. failure is truncation. This is partly an evaluation artifact: the generation call used `max_length = 20`. Longer translations can reach this cap and terminate even when the model would otherwise continue. One output ends with *"the European Union and"*, simultaneously creating truncation, omission, and an incomplete clause. This example shows why a hierarchical annotation scheme would be useful: truncation is the primary error, while omission and grammaticality are secondary consequences.

6 Attention Coverage and Under-Translation

We next examine whether cross-attention coverage can expose under-translation. For source token i, we define coverage as the maximum attention weight it receives across decoder heads and target positions in the final decoder layer:

$$c_i = \max_{h,t} A_{h,t,i}. \tag{1}$$

A token is low-coverage when $c_i < \tau$, with $\tau = 0.2$. For a sentence with N_{src} non-padding source tokens, the low-coverage fraction is

$$f_{\text{low}} = \frac{1}{N_{\text{src}}} \sum_i \mathbf{1}[c_i < \tau]. \tag{2}$$

High f_{low} means that many source tokens never receive an attention weight above the threshold.

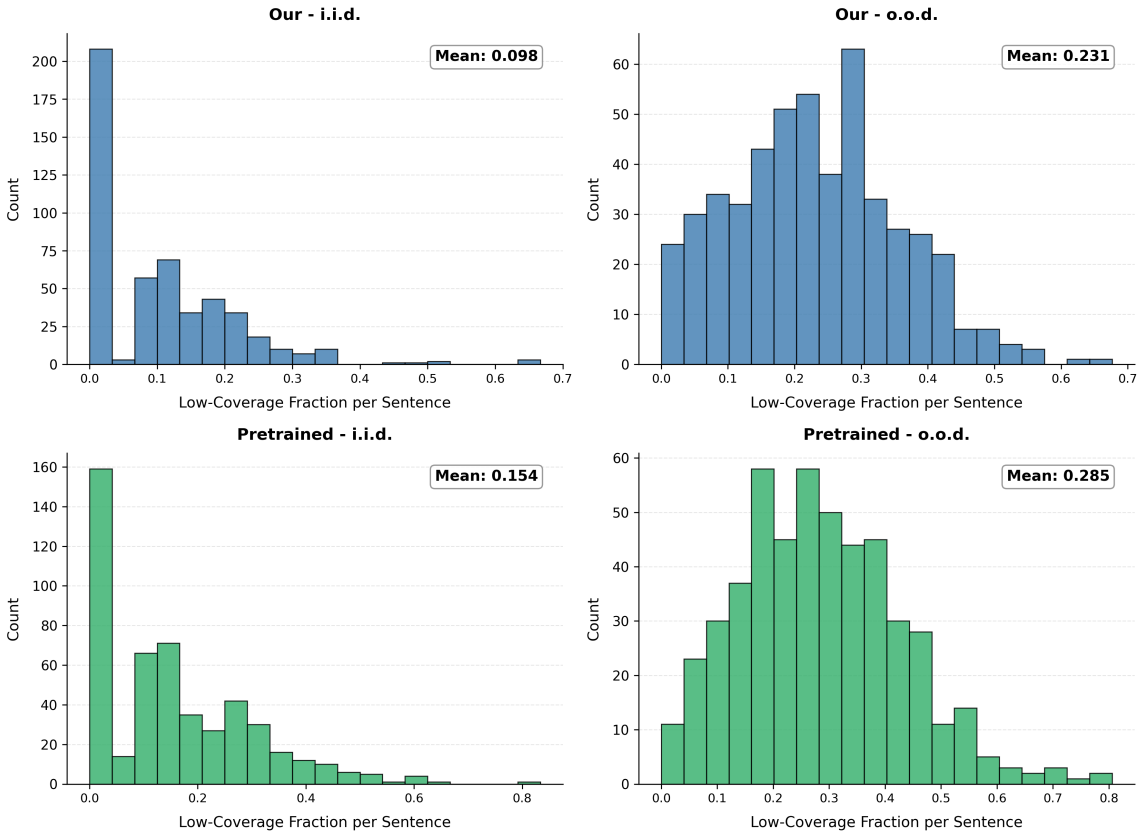

Figure 3: Distribution of low-coverage source-token fractions using $\tau = 0.2$.

Figure 3 shows that mean f_{low} increases on the o.o.d. split to 0.231 for our model and 0.285 for the pretrained model. This broadly agrees with the error and metric analysis: o.o.d. translations contain more missing content and have worse BLEU, ChrF, and length ratio. Inspecting sentences with the highest and lowest coverage fractions suggests that low coverage is related to under-translation, but the relationship is noisy and model-dependent, consistent with earlier work on coverage in NMT [3].

For our model, high f_{low} often corresponds to genuine under-translation. In the bottom example of Figure 4, the model reduces a longer source sentence to *"Is it a matter of crisis or a crisis?"* At the same time, the metric can produce false negatives: some under-translations have $f_{\text{low}} = 0$, meaning that every token receives strong attention somewhere even though the content is not

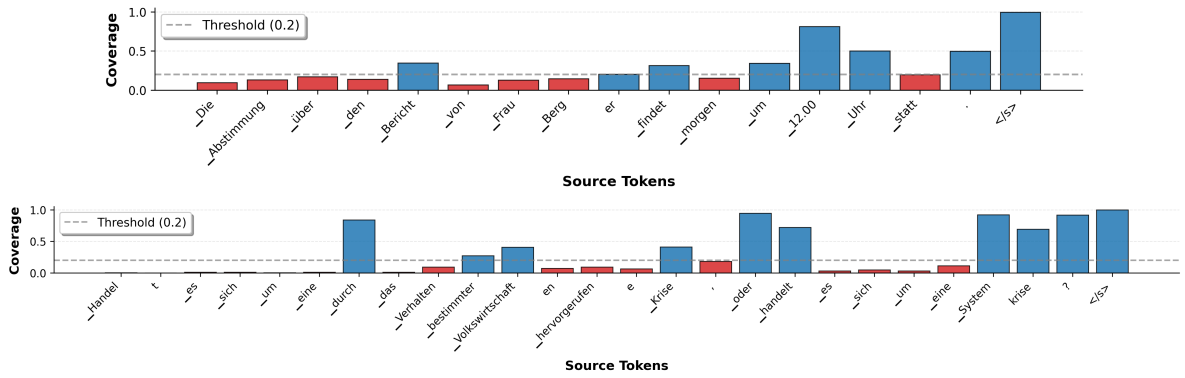

Figure 4: High-f_{low} examples: an adequate pretrained-model translation ($f_{\text{low}} = 0.471$, top) and an under-translation from our model ($f_{\text{low}} = 0.615$, bottom).

translated successfully. Coverage measures whether a token is attended, not how that information is used.

The pretrained model reveals a different limitation. It often concentrates attention on a small number of tokens, including the end-of-sentence token, which may act as summary tokens. The coverage metric penalizes this concentration even when the translation is correct. The top example of Figure 4 has $f_{\text{low}} = 0.471$ but produces a nearly perfect translation. We therefore treat high f_{low} as a heuristic for identifying translations worth inspecting, not as a stand-alone measure of quality.

7 Improving the Decoding Strategy

Because our o.o.d. errors are dominated by omission and short outputs, we first tune decoding rather than changing the architecture. We keep model parameters fixed and vary beam size and length penalty, evaluating both splits with the same four metrics.

Moving from greedy decoding to moderate beam sizes of 4 or 8 improves BLEU, ChrF, and repetition behavior on both splits. Larger beams produce negligible additional gains while increasing decoding time. We therefore choose a beam size of 8 as a reasonable quality–cost trade-off.

Increasing length penalty improves o.o.d. length ratio and also raises BLEU and ChrF. Beyond 1.5, however, BLEU and ChrF flatten while repetition increases. This suggests that aggressive penalties encourage longer output by repeating high-probability phrases rather than covering additional source content. We add a no-repeat trigram constraint and select

$$\text{beam} = 8, \qquad \text{length penalty} = 1.5, \qquad \text{no-repeat trigram} = 3. \tag{3}$$

Figure 5 shows that the tuned configuration improves all four o.o.d. metrics. The gains are modest: BLEU rises from 0.096 to 0.100, ChrF from 0.305 to 0.315 after rescaling, and length ratio from 0.587 to 0.628. On the i.i.d. split, BLEU decreases slightly and length ratio overshoots 1. The baseline already produces appropriately sized i.i.d. outputs, so the higher length penalty encourages longer, more generic translations and slightly harms exact overlap.

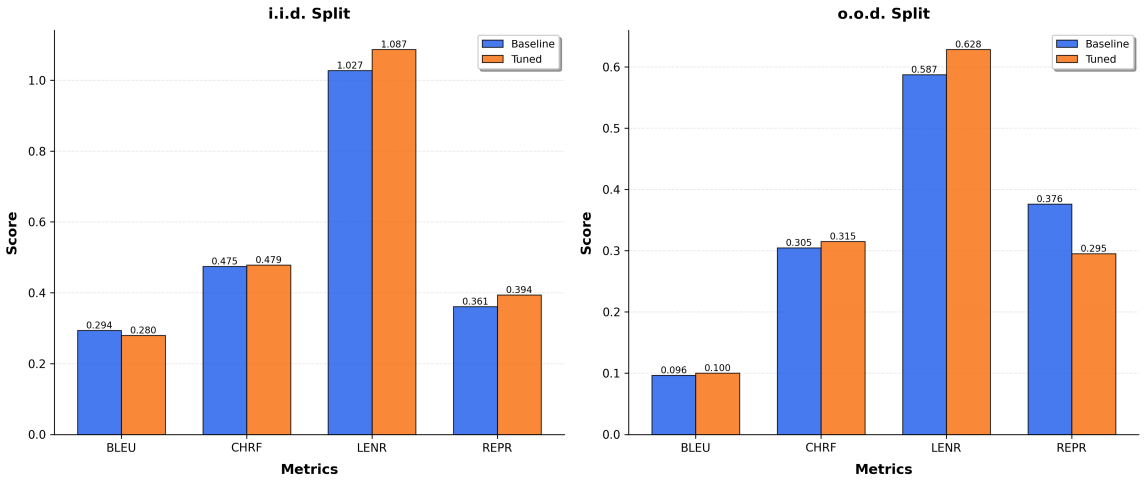

Figure 5: Baseline and tuned decoding configurations. ChrF is rescaled by 0.01 and REPR by 100.

8 Discussion and Limitations

The experiments show a consistent generalization gap, but several limitations affect how the results should be interpreted. First, our model and the pretrained model differ substantially in data scale and likely in capacity, so their absolute scores are not a controlled comparison. Second, the o.o.d. split changes sentence length, meaning that "out of distribution" here refers primarily to length rather than a broad change in domain. Third, the pretrained model's truncation count is inflated by the fixed maximum generation length. Future evaluation should use a length limit that scales with the source or allows the decoder to reach its end-of-sentence token naturally.

Our attention statistic is also deliberately simple. Taking the maximum over heads and target positions ignores how attention is distributed through layers and how information is compressed into contextual source representations. Better analysis could compare multiple layers, separate specialized heads, or combine coverage with output confidence and token-level alignment. Finally, manual error labels would benefit from primary and secondary categories so induced failures are not counted as independent causes.

9 Conclusion

Our Transformer learns the in-distribution translation task without severe overfitting, but it struggles to generalize to longer sentences. The o.o.d. loss flattens early, BLEU and ChrF fall sharply, output length drops to roughly 59% of source length, and manual errors shift toward omission and unsupported addition. A pretrained model is more robust, although its apparent truncation failures reveal the importance of generation settings.

Attention coverage provides a useful but imperfect signal for under-translation. High low-coverage fractions often identify missing content, but specialized attention patterns create both false positives and false negatives. Tuning beam size and length penalty improves o.o.d. translation modestly, but does not close the generalization gap. The main lesson is that search-time changes can reduce symptoms; improving the underlying behavior will require better training coverage, objectives, or model design.

References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[2] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

[3] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*, 2016.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.