# German Machine Translation Performance Report

This report provides an in-depth analysis of the performance of various machine translation models evaluated on **English-to-German** translations. The models were assessed using key metrics: BLEU, chrF, and TER. These metrics provide insights into the quality, accuracy, and readability of translations, revealing areas of strength and potential for improvement across models.

## Metrics Overview

The following metrics were used to assess translation quality in different aspects:

1. BLEU Score : Measures the overlap of n-grams between the model output and reference translation. Higher BLEU scores indicate a closer match to the reference. Ideal for evaluating translation fluency and word choice.

2. chrF Score : Focuses on character-level F-score, which is particularly useful for languages with rich morphology or when assessing models on complex sentences. Higher chrF scores suggest better character-level precision and recall.

3. TER Score : Translation Edit Rate (TER) evaluates the number of edits required to transform the model output to match the reference. Lower TER scores are better, indicating fewer changes are needed for an accurate translation.

## Summary of Model Performance

The following table summarizes the BLEU, chrF, and TER scores for each model. These scores indicate overall model performance, with TER focusing on error rate, BLEU on fluency, and chrF on character-level precision.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BLEU Score** | **chrF Score** | **TER Score** |
| gpt\_4o\_eng\_deu.txt | 54.76 | 71.83 | 34.51 |
| deepl\_eng\_deu.txt | 50.58 | 69.44 | 38.99 |
| gpt\_4o\_mini\_eng\_deu.txt | 52.22 | 70.43 | 36.95 |
| azure\_eng\_deu.txt | 50.98 | 69.9 | 37.7 |
| gpt\_4\_eng\_deu.txt | 52.99 | 71.08 | 36.2 |

**Observations:**

The GPT-4 model variants generally scored the highest in BLEU and chrF, with GPT-4O achieving the best BLEU score of 54.76. TER scores indicate the error rate, with GPT-4O again performing best, suggesting fewer modifications are needed to make it comparable to the reference translation. DeepL, while achieving a decent BLEU score, had a relatively higher TER, indicating that while it produces readable outputs, they may deviate from the exact reference.

## Detailed Analysis of Low BLEU Score Sentences

An analysis of sentences with low BLEU scores was conducted to identify specific challenges faced by each model. Common translation issues included the following:

1. \*\*Idiomatic Expressions\*\*: Difficulty in accurately translating idioms or culturally specific phrases.

2. \*\*Complex Sentence Structures\*\*: Problems with maintaining correct structure in long or nested sentences.

3. \*\*Word Order and Syntax\*\*: Especially in German, proper syntax can be challenging for models trained primarily on English.

4. \*\*Contextual Understanding\*\*: Models occasionally failed to capture nuanced meaning or context, affecting accuracy.

## Performance Analysis Plots

**1. TER Scores by Model**

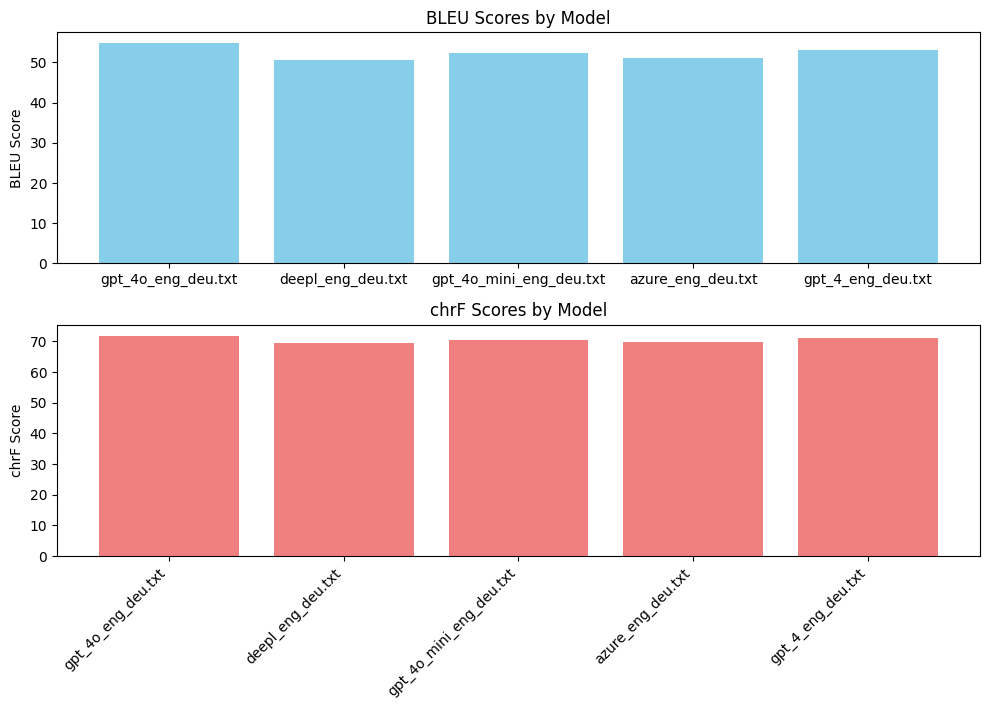


Figure 1: TER Scores for each model showing the error rate in translation. Lower TER scores indicate fewer required edits.

**2. chrF Scores by Model**

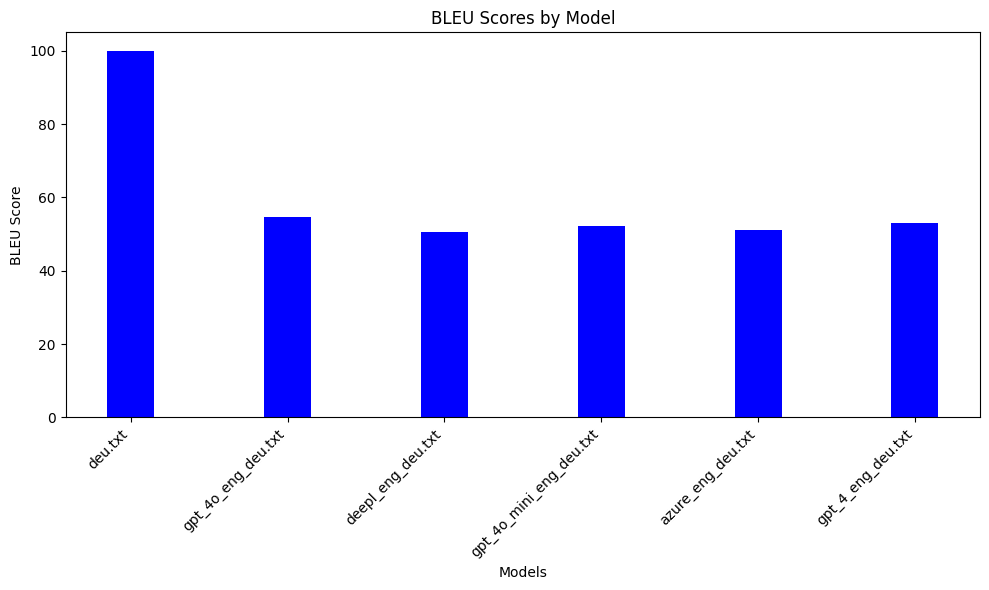


Figure 2: chrF Scores showing character-level precision and recall. Higher chrF scores suggest better word formation and readability.

3. BLEU Scores by Model

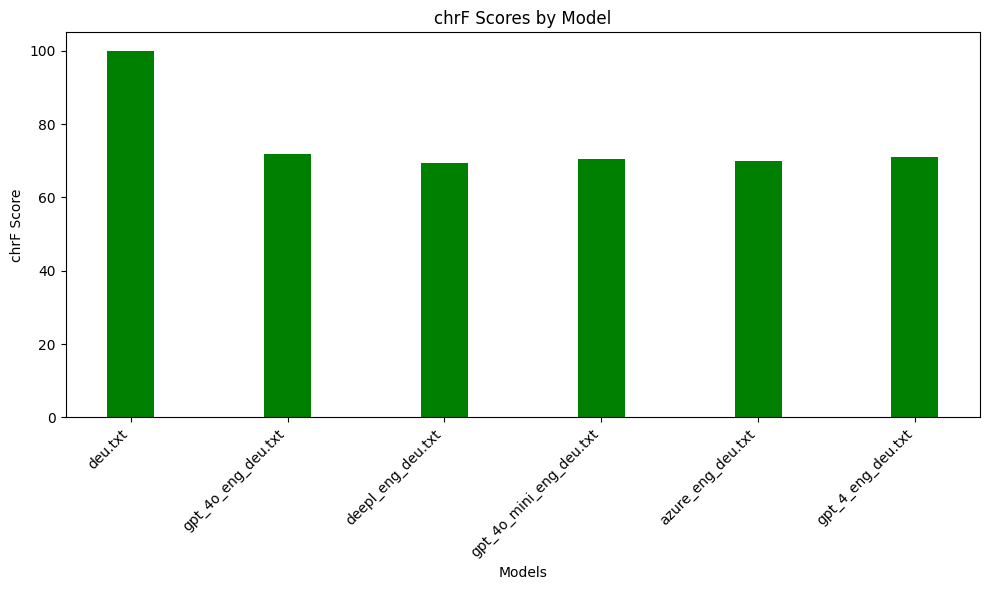


Figure 3: BLEU Scores highlighting n-gram overlap with the reference translation.

\*\*4. Combined Metrics for Comparison\*\*

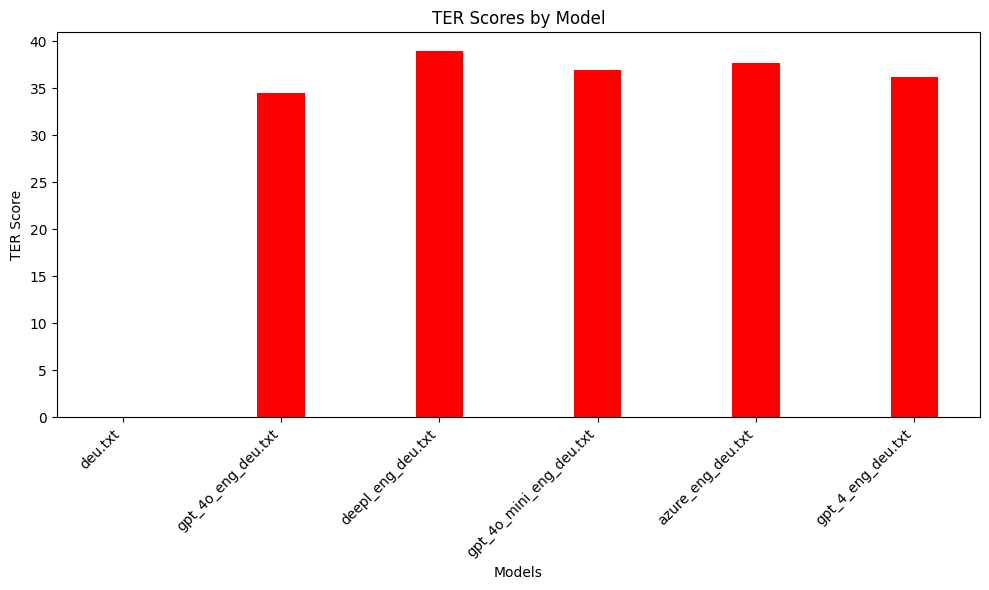


Figure 4: Comparative analysis of BLEU, chrF, and TER scores, providing a consolidated view of model performance.

## Insights and Conclusion

In summary, the GPT-4O model variant consistently performed the best across the metrics, demonstrating a balance between accuracy (BLEU), precision (chrF), and minimal required edits (TER). DeepL, while achieving a reasonable BLEU score, showed limitations in terms of TER, indicating room for improvement in achieving closer fidelity to the reference. Overall, the findings suggest that GPT-4 variants provide robust translation quality for English-to-German translation tasks, especially in terms of readability and structural integrity.