Detailed Turkish Translation Model Performance Report

This report evaluates the performance of multiple machine translation models translating from **English to Turkish**. Key metrics used in this analysis include BLEU, chrF, and TER scores, which assess the accuracy, fluency, and edit distance of the translations, respectively. The models tested are Azure, GPT-4, GPT-4O, GPT-4O-Mini, and DeepL.

# 1. Model Performance Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BLEU Score** | **chrF Score** | **TER Score** |
| Azure | 42.92 | 71.06 | 39.41 |
| GPT-4O | 42.11 | 70.35 | 40.44 |
| DeepL | 40.67 | 69.15 | 41.62 |
| GPT-4O-Mini | 38.29 | 67.15 | 45.45 |
| GPT-4 | 37.97 | 67.61 | 45.43 |

# 2. Detailed Analysis by Metric

## BLEU Score

Azure achieved the highest BLEU score at 42.92, indicating strong accuracy in capturing reference text n-grams. GPT-4 scored the lowest at 37.97, showing a moderate difference in accuracy compared to the other models. Both Azure and GPT-4O achieved high BLEU scores, suggesting these models closely align with the reference text in terms of structure and content.

## chrF Score

Azure also led in chrF score with 71.06, followed closely by GPT-4O with 70.35. This indicates that these models produced translations with higher fluency and appropriate lexical choices. GPT-4O-Mini scored the lowest with 67.15, though all models were close in range.

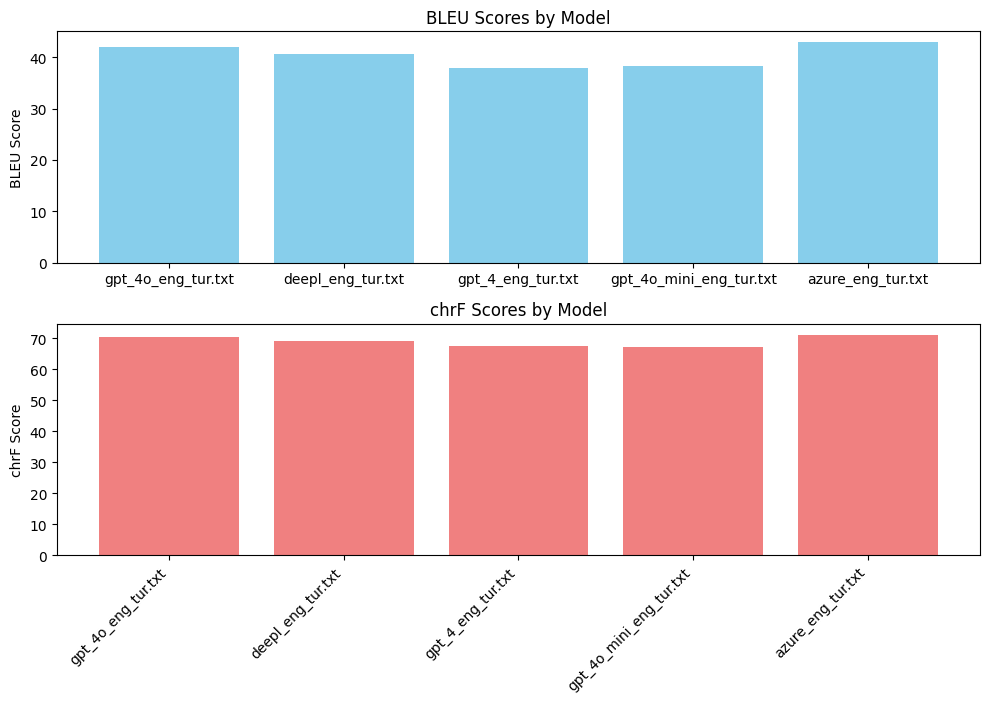
## TER Score

Azure achieved the lowest TER score (39.41), suggesting it requires the fewest edits to match the reference. GPT-4O-Mini had the highest TER score (45.45), indicating more discrepancies requiring adjustments to align with the reference. Azure's consistent performance across metrics highlights its effectiveness in maintaining meaning and grammatical structure.

# 3. Visual Comparison of Model Performance

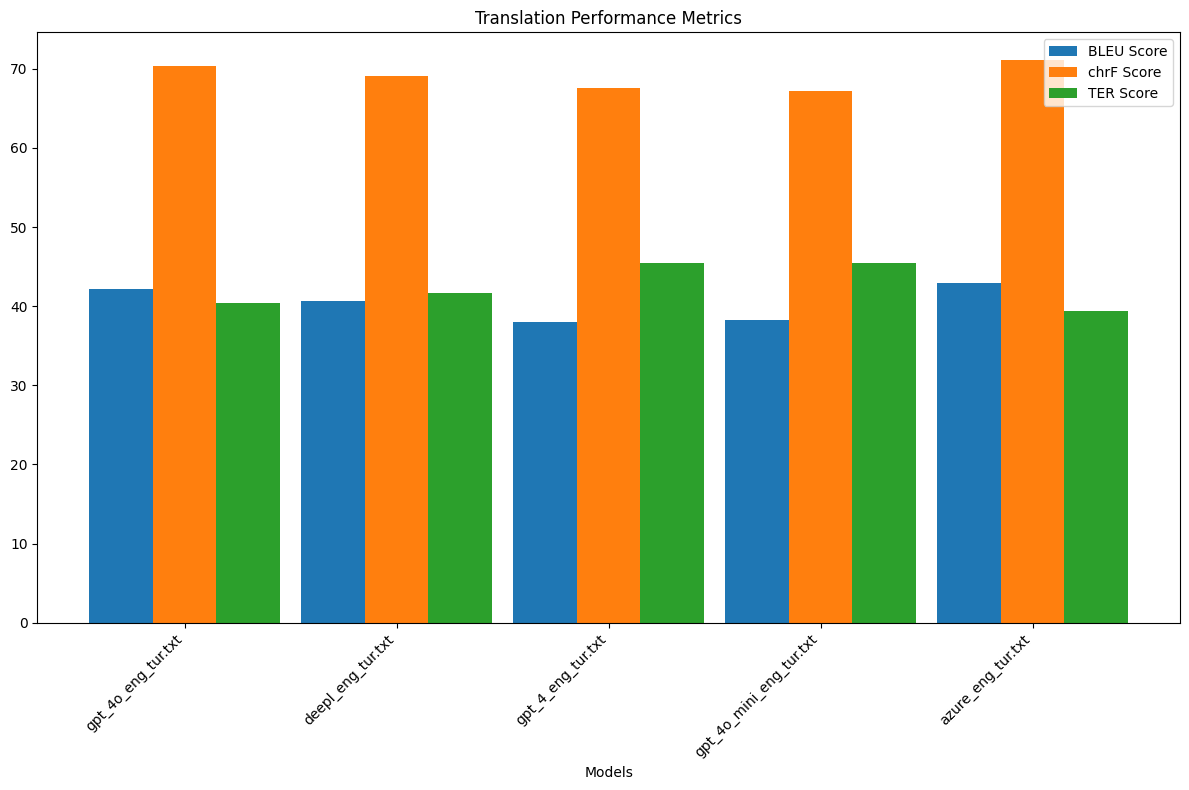
## BLEU Score Comparison

This plot shows that Azure and GPT-4O lead in BLEU scores, indicating they have strong n-gram accuracy with the reference translations. GPT-4 and GPT-4O-Mini scored slightly lower, suggesting they may miss certain n-gram matches.



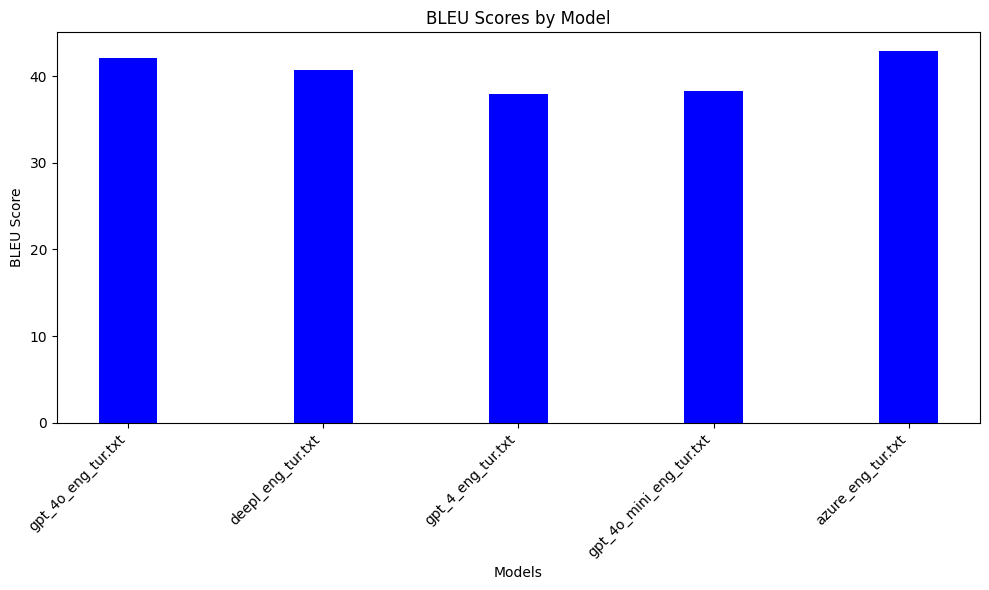
## chrF Score Comparison

This plot illustrates chrF scores, with Azure again leading, closely followed by GPT-4O. The chrF score reflects character-level fluency, indicating Azure produces the smoothest translations.



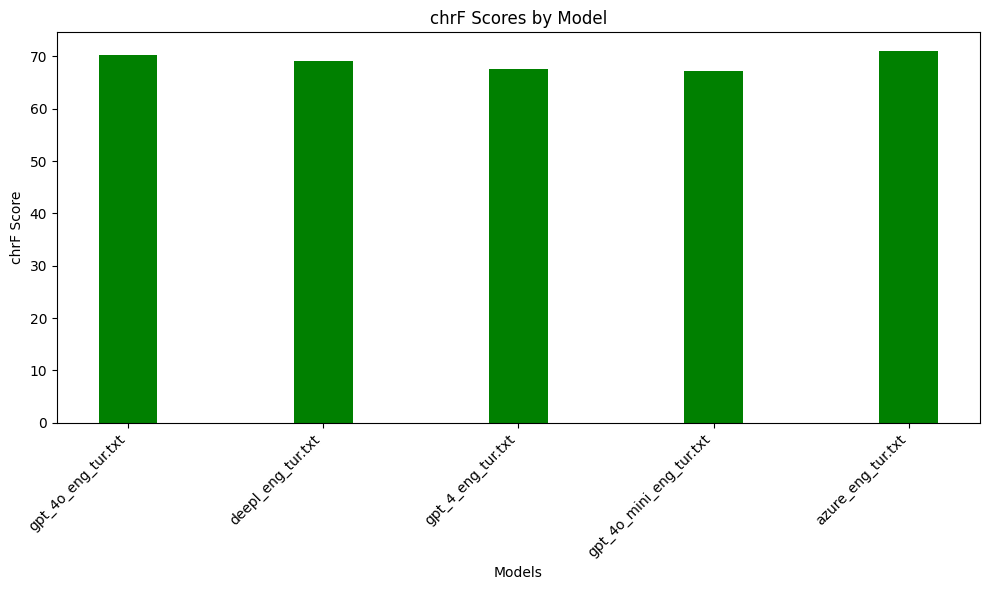
## TER Score Comparison

In this chart, Azure has the lowest TER score, meaning it requires the fewest edits to align with the reference, while GPT-4O-Mini requires the most edits. A lower TER score generally indicates better accuracy in maintaining meaning.



## Combined Translation Performance Metrics

This chart provides a holistic view of the models’ performances, showing that Azure consistently ranks at the top across BLEU, chrF, and TER scores, while GPT-4 and GPT-4O-Mini show room for improvement in accuracy and coherence.



# 4. Conclusions and Recommendations

**Best Performing Model:**  
Azure emerges as the top-performing model across BLEU, chrF, and TER scores, suggesting it provides the most accurate, fluent, and coherent translations with minimal required edits.  
  
**Areas for Improvement:**  
GPT-4 and GPT-4O-Mini lag behind slightly, particularly in BLEU and TER scores, indicating they may benefit from additional fine-tuning to improve accuracy and alignment with reference translations.  
  
**Recommended Usage:**  
For tasks where translation accuracy and fluency are critical, Azure and GPT-4O are recommended due to their strong performance across all evaluated metrics. DeepL also shows competitive scores, making it a reliable alternative.  
  
This analysis confirms Azure’s superior translation quality, making it the preferred choice for high-stakes translation needs.