From the sentence-level BLEU performed on each sample in each file, the mean BLEU scores across all the samples were calculated in order to determine the overall performance of the respective translation method. Based on the aggregated results, one main trend was observed. The translations into Spanish (en-es) across the models tested or translation methods had relatively low mean BLEU scores across all samples in the respective files, ranging from approximately 2.3 to 5.5. However, the translations into English (es-en or fr-en) had much higher mean BLEU scores, with the highest being approximately 40.2.

Another main trend that was observed was that generally, the "teacher-CoT-translation" method had higher mean BLEU scores across any of the models or language pairs tested. More specifically, for the “en-es” and “es-en” language pairs and “DeepSeek-R1-Distill-Llama-8B” model, the "teacher-CoT-translation" method performs the best; for the “fr-en” language pair and “DeepSeek-R1-Distill-Llama-8B” model, the "self-CoT-translation" method performs the best; for the “en-es”, “es-en”, and “fr-en” language pairs and “DeepSeek-R1-Distill-Qwen-1.5B” model, the "teacher-CoT-translation" method performs the best; for the “en-es” language pair and “DeepSeek-R1-Distill-Qwen-7B” model, the "self-CoT-translation" method performs the best; for the “es-en” and “fr-en” language pairs and “DeepSeek-R1-Distill-Qwen-7B” model, the "teacher-CoT-translation" method performs the best; for the “en-es” and “es-en” language pairs and “Qwen3-8B” model, the "teacher-CoT-translation" method performs the best; and lastly, for the “fr-en” language pair and “Qwen3-8B” model, the "teacher-synthesized-CoT-translation" method performs the best.

Therefore, according to BLEU, the “teacher-CoT-translation” method seems to perform the best overall across the language pairs and models. One limitation that was faced during this analysis was that there was no “self-CoT-translation” for the “Qwen3-8B” model for the “en-es” and “es-en” language pairs, signaling that the model “Qwen3-8B” can not parse Spanish text. Though, none of the other models had this much of an issue, where all 50 samples had empty translations.