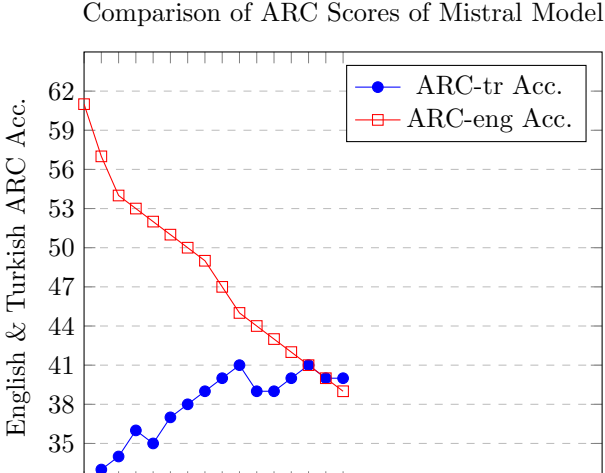

Figure 1: Accuracy comparison of Mistral models on English (left) and Turkish (right) version of the ARC dataset, showing an improvement in Turkish accuracy and a decrease in English accuracy as the number of Turkish tokens increases.

Accuracy comparison of Continued Pretrained models on English (Left, Right) and Turkish (Right) question answering tasks and demonstrating the original language catastrophic forgetting while learning the new language. In the table on the left, the performance of our Hamza$_{Mistral}$ and Hamza$_{GPT2-xl}$ models that are adapted on Turkish together with the original Mistral 7B and GPT2-xl. We present the result of our ablation study, where the performance of the adapted models is given by progressively enlarging the pretraining corpus size from 0.1 GB to 5 GB. Here, the zero and few-show accuracies were evaluated on the original ARC and TruthfulQA. The figure on the right illustrates the Mistral model's results on both Turkish and English versions of the ARC dataset, highlighting its improved performance in Turkish and decreasing performance in English with continued pretraining.