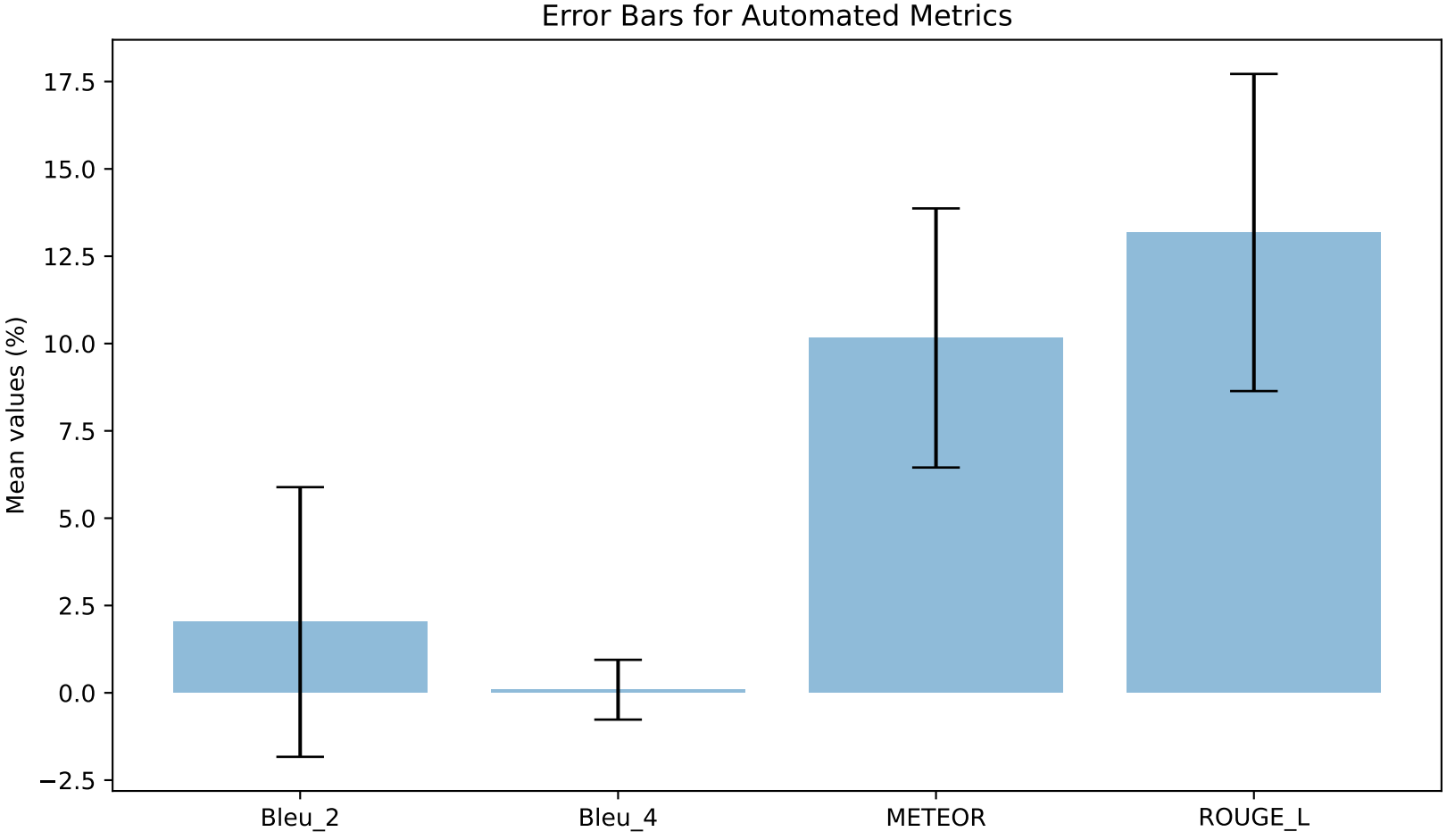

Note that in the main paper, we report a corpus-level BLEU score which is calculated by considering n-gram matches across the entire dataset, typically yielding a higher score compared to the average of instance-level scores. This is primarily due to the unique nature of BLEU's computation which benefits from larger text corpora. Consequently, the mean and standard deviation of instance-level scores, while providing insights into variability at the instance level, do not coincide with the corpus-level BLEU score reported in the main results.