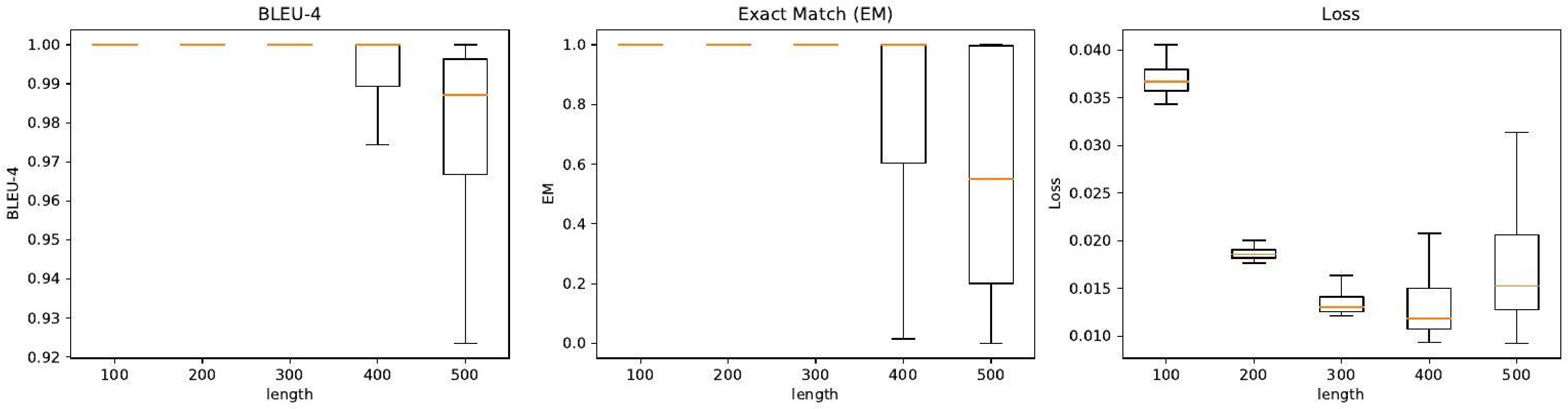

Figure 4: Autoencoding results of the ICAE based on the Llama-7b with memory length $k = 128$. The horizontal axis represents the original context length of test examples. For example, the horizontal axis value of 100 refers to the test examples with context lengths ranging from 95 to 105.

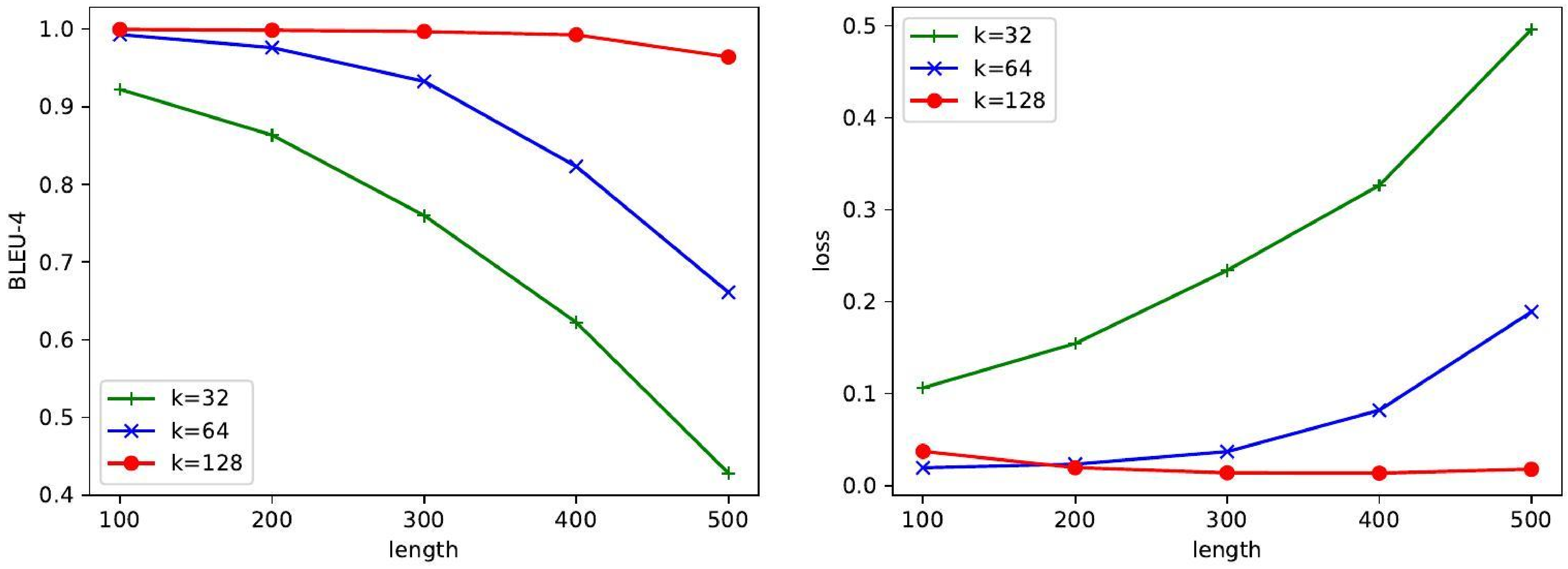

Figure 5: BLEU and loss at different memory slot lengths k.

Compared to $k = 128$ where the BLEU score can still reach over 95% at a context length of 500, the BLEU scores become much less satisfactory for k values of 64 and 32, indicating an inability to losslessly retain the original context. This observation is also evident from the loss curve, suggesting that achieving over $4\times$ compression is rather challenging.

Table 1: Text continuation evaluation for the pretrained ICAE. Similar to the autoencoding evaluation, a higher compression ratio tends to result in more pronounced losses in language modeling.

Context length	Text Continuation		
	PPL (w/ original context)	PPL (w/ 128 memory slots)	Δ
128→128 (1×)	9.99	10.15	+0.16
256→128 (2×)	9.45	9.77	+0.32
512→128 (4×)	9.01	9.50	+0.49

Similarly, the text continuation evaluation presented in Table 1 also illustrates that a higher compression ratio tends to result in more pronounced losses in language modeling.

Table 2 presents 1 specific example of the ICAE performing text restoration, demonstrating an interesting behavior: "*large pretrained language model*" is restored as "*large pretrained model*" and "*The results prove*" is restored as "*The experimental evidence proves*". These restoration errors resemble mistakes humans would make when memorizing the same text. This suggests that, like humans, the model selectively emphasizes or neglects certain parts of the information during the memorization based on its own understanding. It is also consistent with Peng et al. (2023): the stronger the LLM, the fewer it needs to memorize, and thus the smaller the memorization effort. This is similar to human learning: knowledgeable individuals tend to learn more effortlessly, while those with limited knowledge often rely on rote memorization to acquire new information.

To further look into the memorization insight, we test restoration performance for different types of 512-token texts with 128 memory slots produced by ICAE to investigate whether its memorization capability is consistent across different content types. According to Table 3, in contrast to compressing normal texts which can be well restored, compressing and restoring less common texts (i.e., random texts) becomes very challenging, reflected by much worse loss and BLEU scores. All these results strongly support our intuition that an LLM's memorization pattern is highly similar to humans.