Table 4: Memory slots *VS* Original contexts (∼512 tokens) on the PWC test set

System 1 (k **memory slots**)	**System 2** **(original context)**	**Judgement (%)** **win**	**lose**	**tie**	**on par (win+tie)**
Llama-7b (ICAE, k=128)	Alpaca	56.7	26.9	16.4	73.1
	StableLM-7b	74.1	18.8	7.2	81.3
	GPT-4 (gold)	3.4	69.4	27.2	30.6
Llama-2-7b-chat (ICAE, k=64)	Llama-2-7b-chat	13.6	51.6	34.8	48.4
	GPT-4 (gold)	1.9	44.7	53.4	55.3
Llama-2-7b-chat (ICAE, k=128)	Llama-2-7b-chat	19.6	45.4	35.0	54.6
	GPT-4 (gold)	2.8	25.8	71.4	74.2
Llama-2-7b-chat (ICAE, k=256)	Llama-2-7b-chat	22.0	22.2	55.8	77.8
	GPT-4 (gold)	3.8	20.5	75.7	79.5
Llama-2-13b-chat (ICAE, k=256)	Llama-2-13b-chat	21.9	20.8	57.3	79.2
	GPT-4 (gold)	4.0	19.2	76.8	80.8

Table 5: ICAE with different memory slot lengths and different pretraining setups. The last row is the comparison between 128-length ICAE's memory and 128-token summary produced by the GPT-4.

ICAE (Llama-2-7b-chat)	**Judgement** **win** (%)	**lose** (%)	**tie** (%)	**win/lose**
$k = 128$ (pretrained) **VS** $k = 64$ (pretrained)	57.6	19.5	22.9	3.0
$k = 64$ (pretrained) **VS** $k = 32$ (pretrained)	44.7	21.8	33.5	2.1
$k = 64$ (pretrained) **VS** $k = 128$ (no pretraining)	33.1	28.0	38.9	1.2
$k = 128$ (pretrained) **VS** $k = 128$ (no pretraining)	60.4	9.5	30.1	6.4
$k = 128$ (pretrained) **VS** $k = 128$ (pretrained only with AE)	36.4	28.5	35.1	1.3
$k = 128$ (pretrained) **VS** $k = 128$ (pretrained only with LM)	35.1	24.9	40.0	1.4
$k = 128$ (pretrained) **VS** 128-token summary (by GPT-4)	34.1	17.6	48.3	1.9

non-pretrained counterpart, emphasizing the importance of pretraining. By comparing the outputs generated via the pretrained and non-pretrained ICAE, we find the pretrained ICAE suffers less from hallucination than the non-pretrained counterpart (see the examples in Table 9 in Appendix D). We assume the pretraining of ICAE improves the LLM's working memory as it shares some analogies with humans enhancing their memory capacity via extensive memory training which improves the brain's memory encoding capabilities. We also examine pretraining objectives and find combining[3] AE and LM yields better results than using AE or LM individually (the 4th row in Table 5).

The last row of Table 5 compares ICAE's 128-length memory slots with a summary[4] within 128 tokens (∼100 words). Memory slots significantly outperform summaries under the same context length, with ∼2× win/lose ratio, proving to be more compact and informative than natural language.

3.3 Analysis

3.3.1 Scalability

As discussed above, ICAE should achieve better compression performance with a more powerful target LLM. To verify this assumption, we compare the ICAE's performance on three target LLMs: Llama-7b, Llama-2-7b and Llama-2-13b in Table 6, which align well with our expectations – more powerful target LLMs can achieve better context compression ratios.

3.3.2 Latency

We conducted an empirical test to evaluate the impact of ICAE's $4\times$ context compression on inference efficiency. For this efficiency test, we fix the context (i.e., input) length to either 512 or 2048 and the generation length to 128. Table 7 shows that context compression by ICAE is helpful to improve LLM (i.e., Llama-7b) inference efficiency, achieving over $2\times$ speedup. Its acceleration becomes

[3] $\mathcal{L}_{\text{pretrain}} = \lambda\mathcal{L}_{\text{AE}} + (1-\lambda)\mathcal{L}_{\text{LM}}$. We find $\lambda = 0.4 \sim 0.6$ leads to the best result.

[4] Produced by the GPT-4. The specific prompt text is presented in Appendix D.