Additional experiments on PALM-2 models with end-to-end latency comparisons

Anonymous Author(s)
Affiliation
Address
email

Below we report block efficiency and end-to-end (wall clock) latency comparisons on state-of-the-art PALM-2 models [1]. We use PALM-2-Gecko and PALM-2-Bison (where Bison is a larger model) as the small model and large model, respectively. We report average results over 1000 test prompts from the LM1B dataset. The wall clock speed-up is normalized by the wall clock latency of baseline autoregressive decoding.

Table 1: Experimental results on the LM1B dataset with PALM-2-Gecko as the small model and PALM-2-Bison as the large model. All results are over 1000 test prompts averaged over three different random seeds.

Algorithm	K	L	Number of decoded tokens per serial call	Relative wall clock speed-up (normalized by baseline)
Baseline	-	-	1.0	1.0
Speculative	1	4	2.4	1.67
SpecTr	8	4	**3.1**	**2.08**
Speculative	1	8	2.9	1.56
SpecTr	8	8	**4.0**	**2.13**

References

[1] Google AI. Introducing palm 2, 2023. https://blog.google/technology/ai/google-palm-2-ai-large-language-model/.