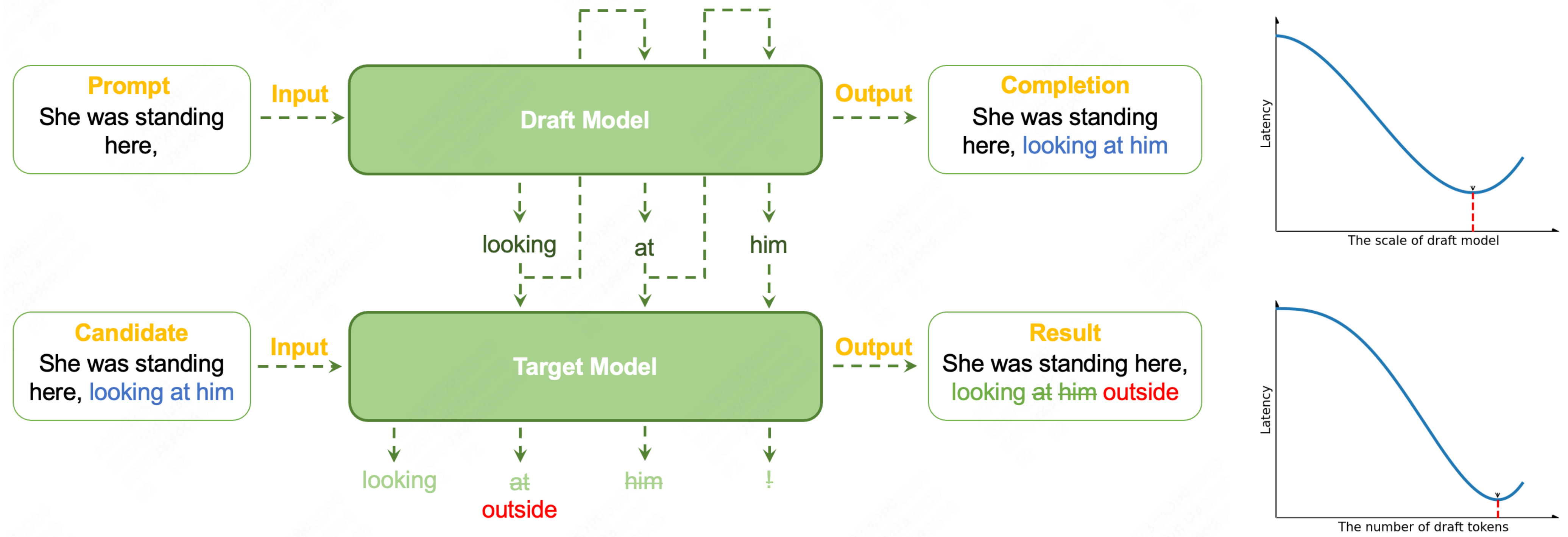
How Speculative Can Speculative Decoding Be?

Zhuorui Liu, Chen Zhang, Dawei Song

Beijing Institute of Technology

Motivation



Prompt
She was standing here,

Input

Draft Model

Output

Completion
She was standing here, looking at him

looking
at
him

Candidate
She was standing here, looking at him

Input

Target Model

Output

Result
She was standing here, looking ~~at~~ ~~him~~ outside

looking
~~at~~
outside
~~him~~
~~!~~

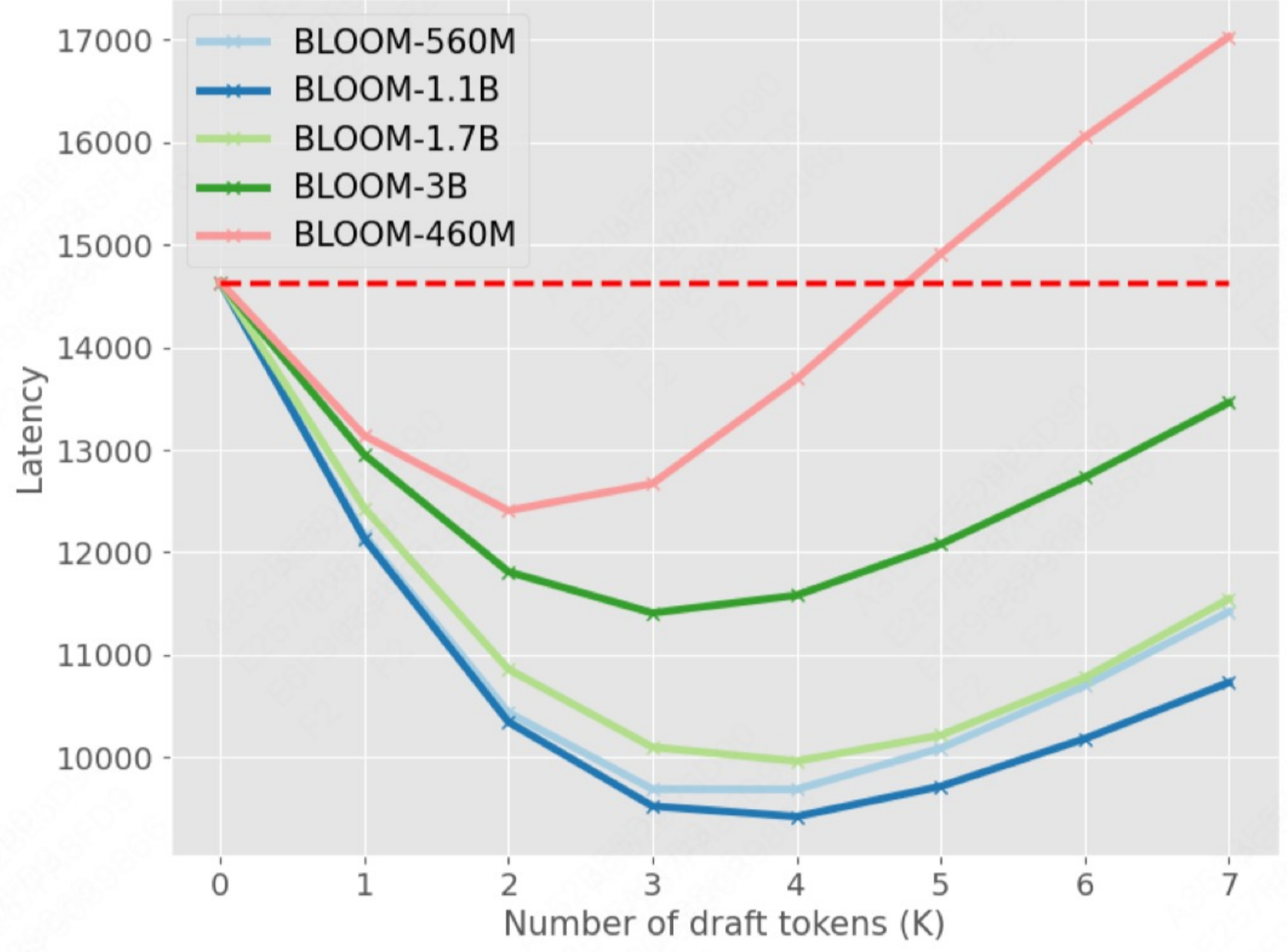
- Speculative Decoding may be influenced by two key factors, namely scale of draft model and number of draft tokens.
- How to confirm an optimal decisions will be important for making good use of exist resources.
- Is there a law for different scales of target model to select the optimal factors mentioned above? This is the key point we want to explore.

Method

- For models, we use several model families, each have one target model an some draft models with different size.
- In each group, we set different number of draft tokens from 1 to 7.
- Due to the characteristic of autoregressive models, we measure the average generation latency in every setting to judge the acceleration result, further more to detect the optimal factors.

Results

Target Model	Sampling Method	Draft Model	K	Speed Up	Scaling
Pythia-2.8B	Autoregressive	None	0	1×	1×
	SpS	Pythia-70M	5	1.984×	40×
	AsG	Pythia-70M	5	1.766×	40×
BLOOM-7.1B	Autoregressive	None	0	1×	1×
	SpS	BLOOM-1.1B	4	1.583×	6.45×
	AsG	BLOOM-560M	2	1.053×	12.68×
Cerebras-GPT-6.7B	Autoregressive	None	0	1×	1×
	SpS	Cerebras-GPT-111M	4	1.507×	60.36×
	AsG	Cerebras-GPT-111M	4	1.387×	60.36×
GPT2-XL	Autoregressive	None	0	1×	1×
	SpS	GPT2-Smallest	5	1.695×	12.097×
	AsG	DistilGPT2	5	1.827×	18.29×
LLaMA-13B	Autoregressive	None	0	1×	1×
	SpS	TinyLLaMA-1.1B	3	1.506×	11.82×
	AsG	TinyLLaMA-1.1B	2	1.096×	11.82×



- Left table is the optimal acceleration levels attained. We showcase the optimal scale of draft model and the optimal number of draft tokens in specific target model.
- Right figure is an example for the experiments results. In this figure we use BLOOM-7.1B as the target model and select several different scales of draft models, under the different number of draft tokens, measuring the average latency.

Conclusions

- For number of draft tokens, 3-5 seems to be the optimal range across different models.
- For scale of draft models, small draft model always have the better performance than larger models.
- But we still believe there is a border in draft model scale, we just don’t achieve the limitation. Therefore, how to construct a useful and smaller draft model to align with the target model for speculative decoding is still an open research question worth further exploration. Solving the problem would lay a solid basis for us to more accurately detect the scale bounds.