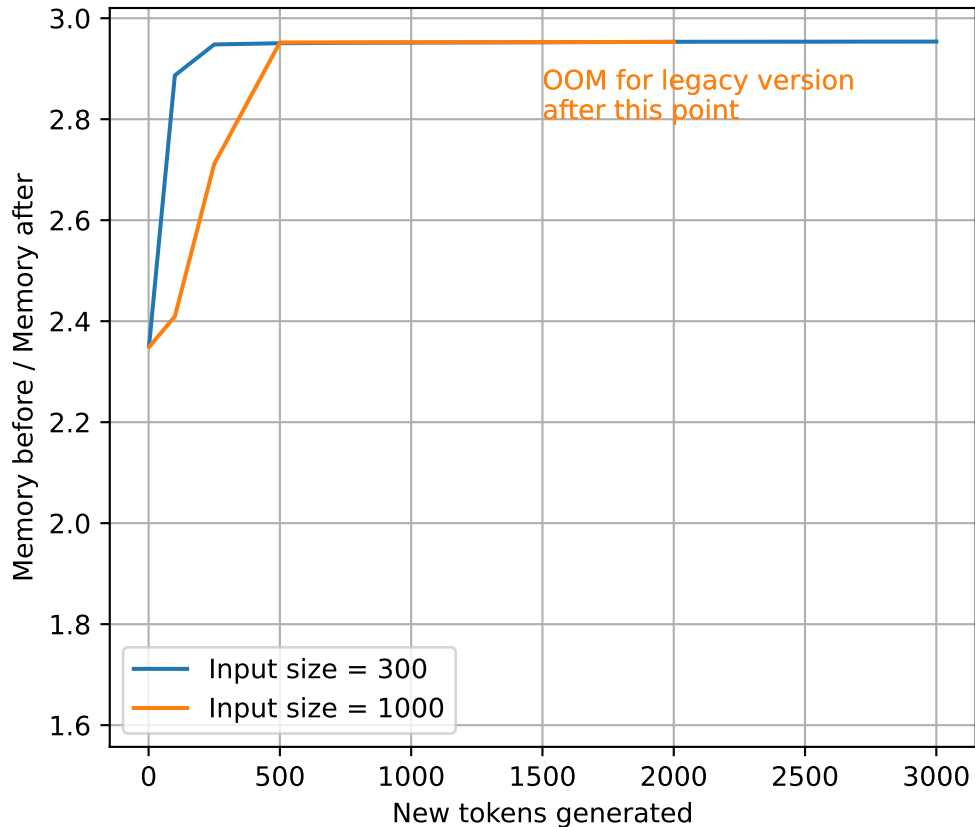


Beam-based decoding

Usual cache size (most models, e.g. Llama2)

Here: Llama2



Small efficient cache size (e.g. Mistral)

Here: Mistral

