CLAIMS

1. A processing system, comprising: at least one memory having executable instructions stored thereon; and one or more processors configured to execute the executable instructions to cause the processing system to: receive a plurality of sets of tokens generated based on an input prompt and a first generative artificial intelligence model, each set of tokens in the plurality of sets of tokens comprising a sequence of tokens corresponding to a candidate response to the input prompt, wherein the plurality of sets of tokens is organized into a tree data structure; select, using a second generative artificial intelligence model and recursive adjustment of a target distribution associated with the received plurality of sets of tokens, a set of tokens from the plurality of sets of tokens; and output the selected set of tokens as a response to the input prompt; wherein a root node of the tree data structure corresponds to the input prompt, and wherein each path through the tree data structure corresponds to a different sequence of tokens corresponding to the candidate response to the input prompt.

2. The processing system of claim 1, wherein: the tree data structure includes a plurality of levels, each level corresponding to a token in the sequence of tokens, and a number of tokens at a particular level in the tree data structure is based on a branching factor associated with an immediately prior level to the particular level in the tree data structure.

3. The processing system of claim 1, wherein a depth of the tree data structure corresponds to a parameter defining a maximum number of tokens generated by a single pass through the first generative artificial intelligence model.

4. The processing system of claim 1, wherein a size of each set of tokens is based on a computational complexity metric associated with generating a target set of tokens by the second generative artificial intelligence model.

5. The processing system of claim 1, wherein the recursive adjustment of the target distribution comprises: determining whether to accept or reject a first token in a set of tokens from the plurality of sets of tokens; and adjusting a probability distribution used to verify a second token in the set of tokens subsequent to the first token based on the determination of whether to accept or reject the first token.

6. The processing system of claim 5, wherein to adjust the probability distribution, the one or more processors are configured to cause the processing system to subtract a probability value associated with the first token from the probability distribution based on determining to reject the first token.

7. The processing system of claim 1, wherein to select the set of tokens from the plurality of sets of tokens, the one or more processors are configured to cause the processing system to: reject a first token at a first level of a tree data structure representing the plurality of sets of tokens; generate an adjusted probability distribution based on the rejection of the first token; discard or ignoring, from the tree data structure, children tokens of the first token at levels deeper than the first level of the tree data structure; and determine whether to accept or reject a second token at the first level of the tree data structure based on the adjusted probability distribution.

8. The processing system of claim 1, wherein to select the set of tokens from the plurality of sets of tokens, the one or more processors are configured to cause the processing system to: reject each set of tokens generated by the first generative artificial intelligence model; and sample, using the second generative artificial intelligence model, a token based on a target distribution that excludes probabilities associated with each set of tokens generated by the first generative artificial intelligence model, wherein the selected set of tokens comprises the sampled token.

9. The processing system of claim 1, wherein: the first generative artificial intelligence model corresponds to a draft model in a speculative decoding pipeline, and the second generative artificial intelligence model corresponds to a target model in the speculative decoding pipeline.

10. A processor-implemented method, comprising: receiving a plurality of sets of tokens generated based on an input prompt and a first generative artificial intelligence model, each set of tokens in the plurality of sets of tokens comprising a sequence of tokens corresponding to a candidate response to the input prompt, wherein the plurality of sets of tokens is organized into a tree data structure; selecting, using a second generative artificial intelligence model and recursive adjustment of a target distribution associated with the received plurality of sets of tokens, a set of tokens from the plurality of sets of tokens; and outputting the selected set of tokens as a response to the input prompt; wherein a root node of the tree data

structure corresponds to the input prompt, and wherein each path through the tree data structure corresponds to a different sequence of tokens corresponding to the candidate response to the input prompt.

11. The method of claim 10, wherein: the tree data structure includes a plurality of levels, each level corresponding to a token in the sequence of tokens, and a number of tokens at a particular level in the tree data structure is based on a branching factor associated with an immediately prior level to the particular level in the tree data structure.

12. The method of claim 10, wherein a depth of the tree data structure corresponds to a parameter defining a maximum number of tokens generated by a single pass through the first generative artificial intelligence model.

13. The method of claim 10, wherein a size of each set of tokens is based on a computational complexity metric associated with generating a target set of tokens by the second generative artificial intelligence model.

14. The method of claim 10, wherein the recursive adjustment of the target distribution comprises: determining whether to accept or reject a first token in a set of tokens from the plurality of sets of tokens; and adjusting a probability distribution used to verify a second token in the set of tokens subsequent to the first token based on the determination of whether to accept or reject the first token.

15. The method of claim 14, wherein adjusting the probability distribution comprises subtracting a probability value associated with the first token from the probability distribution based on determining to reject the first token.

16. The method of claim 10, wherein selecting the set of tokens from the plurality of sets of tokens comprises: rejecting a first token at a first level of a tree data structure representing the plurality of sets of tokens; generating an adjusted probability distribution based on the rejection of the first token; discarding or ignoring, from the tree data structure, children tokens of the first token at levels deeper than the first level of the tree data structure; and determining whether to accept or reject a second token at the first level of the tree data structure based on the adjusted probability distribution.

17. The method of claim 10, wherein selecting the set of tokens from the plurality of sets of tokens comprises: rejecting each set of tokens generated by the first generative artificial intelligence model; and sampling, using the second generative artificial intelligence model, a token based on a target distribution that excludes probabilities associated with each set of tokens generated by the first generative artificial intelligence model, wherein the selected set of tokens comprises the sampled token.

18. The method of claim 10, wherein: the first generative artificial intelligence model corresponds to a draft model in a speculative decoding pipeline, and the second generative artificial intelligence model corresponds to a target model in the speculative decoding pipeline.

19. A non-transitory computer-readable medium having executable instructions stored thereon which, when executed by a processor, perform an operation comprising: receiving a plurality of sets of tokens generated based on an input prompt and a first generative artificial intelligence model, each set of tokens in the plurality of sets of tokens comprising a sequence of tokens corresponding to a candidate response to the input prompt, wherein the plurality of sets of tokens is organized into a tree data structure; selecting, using a second generative artificial intelligence model and recursive adjustment of a target distribution associated with the received plurality of sets of tokens, a set of tokens from the plurality of sets of tokens; and outputting the selected set of tokens as a response to the input prompt; wherein a root node of the tree data structure corresponds to the input prompt, and wherein each path through the tree data structure corresponds to a different sequence of tokens corresponding to the candidate response to the input prompt.