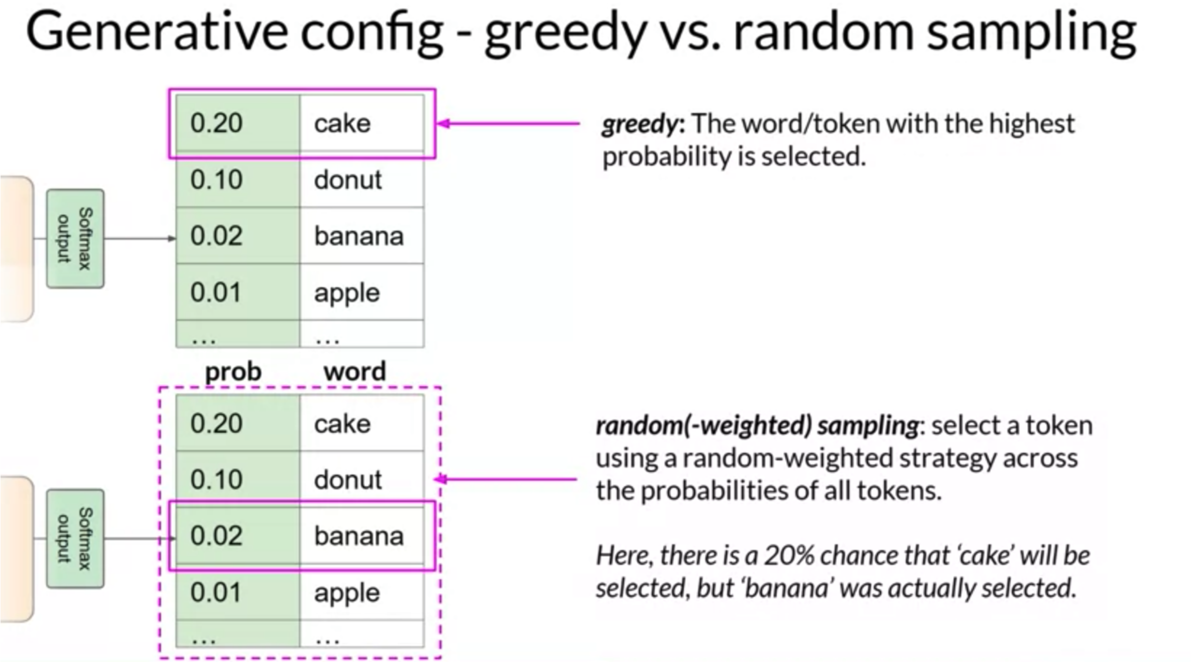
Generative Configuration

The output of Softmax Layer: a list of word/token with its probabilities.

- Greedy: the work/token with the highest probability is selected.
- Random sampling: select a token using a random-weighted strategy across the probabilities of all tokens.



Generative config - greedy vs. random sampling

prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
…	…

greedy**:** The word/token with the highest probability is selected.

prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
…	…

random(-weighted) sampling: select a token using a random-weighted strategy across the probabilities of all tokens.

Here, there is a 20% chance that 'cake' will be selected, but 'banana' was actually selected.

Generative Config - top p vs top k sampling

- select an output from the top-k results after applying random-weighted strategy using the probabilities
- select an output from the top-p: limit the random sampling to the top-ranked consecutive results by probability and with a cumulative probabilities ≤ p.

Generative Config - temperature

- Temperature setting affects affect the randomness of the output of the softmax layer.
- Cooler temperature (e.g < 1) strongly peaked probability distribution ⟶ text sounds more reliable
- Higher temperature (e.g > 1) broader flatter probability distribution; higher degree of randomness ⟶ text sounds more creative