Inference Parameters

Understanding how to control model behavior through parameter tuning for optimal AI responses.

Inference Parameters

Control Mechanism

Parameters that adjust model response during inference, influencing output generation patterns

Output Modification

Change the pool of possible outputs or limit final response characteristics

Two Key Categories

Randomness & Diversity parameters and Length control parameters

Randomness and Diversity

Control variation in model responses by adjusting probability distributions

Temperature

Scales log probabilities to control randomness level

Top K

Limits selection to K most probable tokens

Top P

Filters tokens by cumulative probability threshold

Greedy Sampling

> Always ordering the single most popular dish - "I'll have the Caesar salad"

Corresponds to Top-K=1 and Temperature=1.0, where model always chooses the most likely next word

Random Sampling

> Pulling menu items randomly from a hat - "I'll have the chicken fried steak soup"

Uses moderate Top-K values (50-100) with high Temperature (1.5-2.0) for creative but potentially incoherent outputs

Top-K Sampling

Restricts model vocabulary to the K most probable tokens at each generation step.

> Example: *"I'll have the..."*
>
> Vocabulary: {mat: 0.6, bucket: 0.3, couch: 0.2, bed: 0.1, chair: 0.05, bike: 0.01, car: 0.003}

With **Top-K=3**, only considers: {mat, bucket, bed}

Reduces incoherent output by limiting vocabulary choices to highest probability tokens.

Top-p Implementation

Filters tokens by cumulative probability threshold for adaptive selection

40%

Salad

First choice probability

30%

Burger

Second choice probability

10%

Pasta

Third choice probability

With Top-P=0.8, includes words until cumulative probability reaches 80%

Top-K vs Top-P Comparison

Top-K (Fixed)

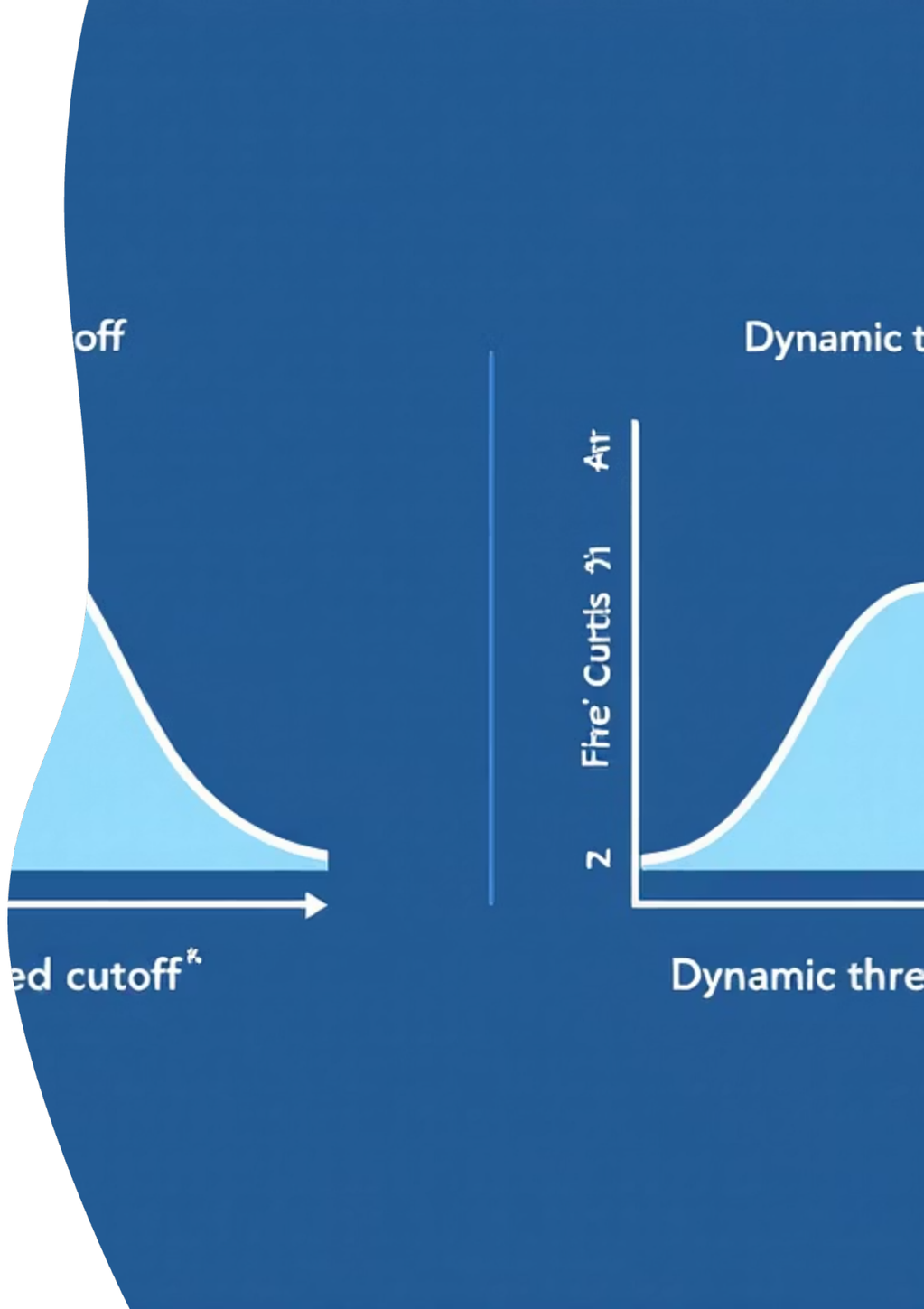
Fixed number of tokens (K=3)

Considers exactly 3 highest probability words regardless of them actual probabilities

Top-P (Dynamic)

Dynamic token selection (P=0.8)

Includes as many words as needed to reach 80% cumulative probability



Temperature Control

Low Temperature (<1)

Values like 0.2-0.5 make model more confident, peaking probability distribution

High Temperature (>1)

Values like 1.5-2.0 spread predictions, making model more "uncertain" and creative

Parameter Application Example

Prompt: *"I hear the hoof beats of..."*

0.7

horses

Highest probability candidate

0.2

zebras

Second most likely option

0.1

unicorns

Least probable choice

Combined Parameter Effects

Settings: **Temperature=0.8, Top-K=2, Top-P=0.7**

Temperature Scaling

Flattens distribution, increases "unicorns" probability, decreases "horses" probability

Top-K Filtering

Selects top 2 candidates: "horses" and "zebras"

Top-P Filtering

From Top-K set, keeps tokens until 70% cumulative probability reached

Length Control Parameters

Parameters that constrain response length and content boundaries.

1

Response Length

Exact minimum/maximum token count specifications

2

Stop Sequences

Character sequences that halt generation when encountered

3

Penalties

Penalize response length, repetition, frequency, or token types