

Temperature: Step-by-Step Calculation

Given: Logits = [2.0, 1.0, 0.5, 0.2]

Tokens = ["cat", "dog", "bird", "fish"]

T=0.5: [4.0, 2.0, 1.0, 0.4] → [0.73, 0.18, 0.07, 0.02]

→ **73% on "cat" (FOCUSED)**

T=1.0: [2.0, 1.0, 0.5, 0.2] → [0.42, 0.23, 0.16, 0.13]

→ **42% on "cat" (BALANCED)**

T=2.0: [1.0, 0.5, 0.25, 0.1] → [0.32, 0.26, 0.23, 0.19]

→ **32% on "cat" (FLAT)**

Lower T = more peaked | Higher T = more flat