Temperature Calculation: Step-by-Step

Given: Logits = [2.0, 1.0, 0.5, 0.2]
Tokens = ["cat", "dog", "bird", "fish"]

Step 1: Scale by T=0.5 (MORE PEAKED)
Step 2: Scale by T=1.0 (UNCHANGED)

Scaled = [2.0/0.5, ...] = [4.0, 2.0, 1.0, 0.4]
Scaled = [2.0, 1.0, 0.5, 0.2]
Softmax → [0.53, 0.19, 0.12, 0.10]
Softmax 53% on "cat" (BALANCED)

→ **73% on "cat" (VERY FOCUSED)**

Step 3: Scale by T=2.0 (FLATTER)
Scaled = [1.0, 0.5, 0.25, 0.1]
Softmax = [..., 0.17, 0.15]
28% (MUCH FLATTER)

$$p_i = \frac{\exp(\text{logit}_i/T)}{\sum \exp(\text{logit}_j/T)}$$

Lower T → More confident (peaky)
Higher T → More random (flat)