

Contrastive Search: How It Works

Step 1: Get top-8 tokens by probability ($\alpha=0.6$)

- city: 0.45
- town: 0.18
- area: 0.15
- place: 0.12

$$\text{score} = (1-\alpha) \times P(\text{token}) - \alpha \times \text{similarity}$$

$$\text{city: } 0.4 \times 0.45 - 0.6 \times 0.92 = 0.372$$

$$\text{town: } 0.4 \times 0.18 - 0.6 \times 0.75 = 0.378$$

$$\text{area: } 0.4 \times 0.15 - 0.6 \times 0.65 = 0.330$$

$$\text{place: } 0.4 \times 0.12 - 0.6 \times 0.60 = 0.312$$

Step 2: Compute similarity to recent context

- Context: "the"
- city: 0.92 (cosine similarity)
 - town: 0.75 (cosine similarity)
 - area: 0.65 (cosine similarity)
 - place: 0.60 (cosine similarity)

Winners: "town" (high prob, lower similarity)

Key Insight: Balance probability (coherence) with diversity (novelty)

 $\alpha=0$: Pure greedy | $\alpha=0.6$: Balanced | $\alpha=1.0$: Maximum diversity