

THE FUNDAMENTAL TRUTH

Parameters don't operate in isolation. Temperature reshapes the probability distribution, then Top-P/Top-K filter from that reshaped distribution. Penalties apply after sampling pools are defined. Understanding this pipeline is the difference between random tweaking and intentional control.

⚡ PARAMETER PROCESSING ORDER



PARAMETER INTERACTIONS

TANDEM • PARALLEL • OVERRIDE

TEMPERATURE + TOP-P

PARALLEL

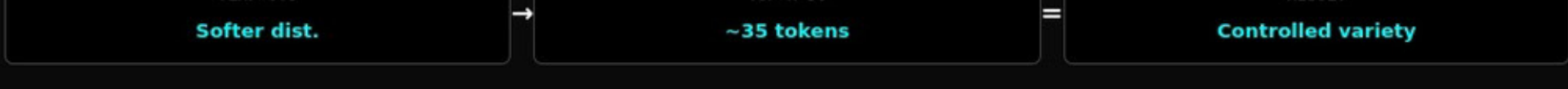
Both control randomness but through different mechanisms. Temperature reshapes the entire distribution; Top-P then filters that distribution. Using both aggressively creates unpredictable compounding effects.

⚠️ **OFFICIAL GUIDANCE**
OpenAI recommends altering one or the other, not both. Keep one at default when adjusting the other.

TEMPERATURE + TOP-K

TANDEM

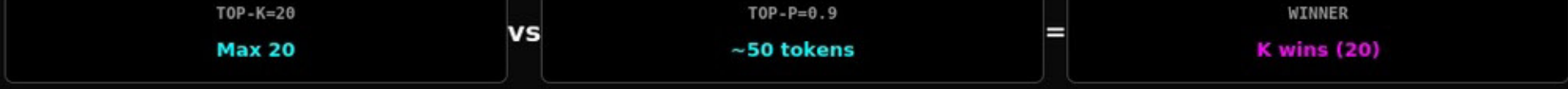
These work in sequence: Temperature flattens or sharpens probabilities, then Top-K provides a hard safety net. High temp + moderate K gives creativity with guardrails. Low temp makes K less relevant.



TOP-P + TOP-K

OVERRIDE

Whichever is more restrictive wins. If K=20 but P=0.9 would include 50 tokens, K wins. If P=0.5 includes only 10 tokens but K=50, P wins. They don't compound—they compete.



FREQ PENALTY + PRESENCE PENALTY

TANDEM

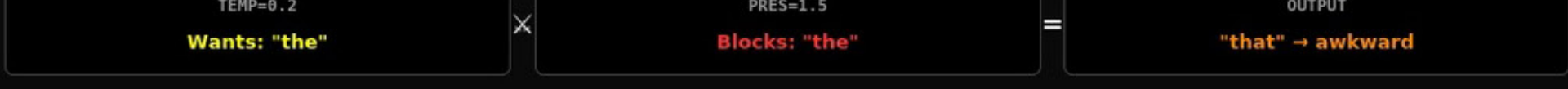
These stack additively. Frequency penalizes proportionally to repetitions. Presence penalizes on first occurrence. Combined, they create aggressive variety but can force unnatural word choices.



LOW TEMP + HIGH PENALTIES

CONFLICT

These fight each other. Low temperature wants predictable, high-probability tokens. High penalties push away from those safe choices. Result: forced into awkward vocabulary.



HIGH TEMP + LOW TOP-P

BALANCED

A sophisticated combo. Temperature flattens probabilities, making rare tokens more viable. Low Top-P then filters to only the most probable among those flattened options. Result: creative but not chaotic.



📊 INTERACTION EFFECTS MATRIX

COMBINATION	RELATIONSHIP	NET EFFECT	WATCH OUT FOR
Temp↑ + Top-P↑	Compound amplification	Maximum chaos	Incoherence risk
Temp↓ + Top-P↓	Double restriction	Deterministic	Repetitive, dull
Temp↑ + Top-P↓	Expand then filter	Controlled creativity	Sweet spot—experiment here
Temp↓ + Top-P↑	Minimal effect	Mostly deterministic	Top-P largely irrelevant
FreqPen + PresPen	Additive stacking	Strong anti-repetition	Forced vocabulary at high values
High Temp + High Penalties	Wild variety	Unpredictable novelty	May lose semantic coherence
Top-K + Top-P	Restrictive wins	Minimum of both	One may be doing nothing

PROVEN RECIPES

TESTED COMBINATIONS

📄 TECHNICAL WRITING

Temp:
0.3
Top-P:
1.0 (default)
FreqPen:
0.3
PresPen:
0.0

Low temp for accuracy, slight freq penalty prevents jargon loops

📖 CREATIVE FICTION

Temp:
0.9
Top-P:
0.85
FreqPen:
0.5
PresPen:
0.5

High temp for ideas, Top-P prevents nonsense, penalties ensure fresh prose

💬 CHATBOT

Temp:
0.7
Top-P:
0.9
FreqPen:
0.2
PresPen:
0.2

Balanced for natural conversation with light variety

📊 DATA ANALYSIS

Temp:
0.0-0.2
Top-P:
1.0
FreqPen:
0.0
PresPen:
0.0

Near-deterministic for consistent, reproducible outputs

💡 BRAINSTORMING

Temp:
1.2
Top-K:
50
FreqPen:
0.8
PresPen:
0.6

High temp for wild ideas, Top-K as safety net, penalties force novelty

📧 MARKETING COPY

Temp:
0.8
Top-P:
0.9
FreqPen:
0.7
PresPen:
0.4

Creative but coherent, penalties prevent buzzword overuse



THE GOLDEN RULE

Adjust one parameter at a time. Observe. Then layer another. Multi-parameter changes make debugging impossible.



THE OVERRIDE PRINCIPLE

When Top-K and Top-P conflict, the more restrictive wins. One of them may be doing nothing—test to find out which.



THE DIMINISHING TRAP

Low temperature minimizes the impact of Top-P and penalties. At Temp=0, most other parameters become nearly irrelevant.

QUICK START: Temp=0.7 | Top-P=0.9 | FreqPen=0.3 | PresPen=0.3 → Adjust ONE at a time