Production System Settings (2025)

Default decoding parameters used by major LLM APIs and platforms

System/API	Default Method	Temperature	Top-p	Top-k	Other Parameters	Notes
GPT-4 API	Nucleus	0.7	1.0	—	frequency_penalty=0 presence_penalty=0	Can adjust all params
Claude API	Nucleus	1.0	0.999	—	max_tokens required	Temperature □ [0,1]
ChatGPT Web	Nucleus+Temp	0.7	0.95	—	Not adjustable	Optimized for chat
Gemini API	Top-k + Top-p	1.0	0.95	40	candidate_count=1	Both k and p used
Llama 2 (HF)	Configurable	1.0	0.9	50	repetition_penalty=1.0	Full control
Cohere API	Nucleus	0.75	0.999	0	frequency_penalty=0	k=0 means disabled
Mistral API	Nucleus	0.7	1.0	—	safe_mode=false	Similar to OpenAI
Together AI	Configurable	0.7	0.7	50	repetition_penalty=1.0	Multiple options

- Temperature: Lower = more deterministic, Higher = more creative
- Top-p (Nucleus): Cumulative probability threshold (typically 0.9-1.0)
- Top-k: Number of top tokens to consider (often 40-50 when used)
- Most production systems use Nucleus sampling as default
- ChatGPT and Claude optimize for conversational quality
- APIs generally allow full parameter customization

Provider Types:

Proprietary LLM
Open Source
Optimized for Chat