Prompt Controls Cheatsheet

1. Tokenization & Context Window

Token = smallest unit (word/part/character). Models have max context length (e.g., 8k-128k tokens).

Controls: Use max_tokens, check count with tools like tiktoken. Truncation drops earliest tokens.

2. Role Structure & Delimiters

System/User/Assistant roles define behavior and responses. Use delimiters (`[INST]`, `<|user|>`, etc.) to structure prompts.

3. Length & Output Structure

Control output with instructions like 'respond in 20 words' or 'output as a CSV table'. Use markdown/code block formatting when needed.

4. Style, Tone, and Register

Examples: 'Write in poetic form', 'Explain as if to a child'. Tone matters: business, technical, humorous, casual.

5. Divergent vs. Convergent Prompts

Divergent: 'List 10 ideas'. Convergent: 'Summarize in one sentence'. Use based on goal.

6. Decoding Parameters

Temperature: randomness control. Top-k/top-p: diversity filters. Presence/Frequency penalties: reduce repeats or encourage novelty.

7. Prompt Debugging & Readability

Make prompts modular, versionable. Use clear formatting and iterate based on feedback from outputs.

8. Common Pitfalls & Fixes

Fix ambiguity with examples. Prevent truncation with proper token limits. Clarify roles when confusion arises.

Control	Example Prompt
Output Length	Write exactly 15 words about AI.
Style & Tone	Explain quantum mechanics to a 5th grader in a cheerful tone.

Role	You are a Unix shell. Output only shell commands.
Output Format	Respond with a JSON object summarizing the main point.
Creativity (Temperature)	(API param) temperature=0.8
Convergent/Divergent	Give three possible summaries vs. one single best summary.

Quick Diagnostic

Output too random? \rightarrow Lower temp/top-p. Repeating? \rightarrow Add frequency penalty. Format wrong? \rightarrow Clarify structure and give example.

Reference Links

- OpenAl API: https://platform.openai.com/docs
- HuggingFace Transformers: https://huggingface.co/docs/transformers
- Prompting Guide: https://www.promptingguide.ai