Prompt Controls Cheatsheet

1. Tokenization Context Window

- **Token**: The smallest unit the LLM reads (word, part of word, or character).
- **Tokenization**: "cats" ≠ "cat"+"s" – subword units like BPE or WordPiece.
- **Context Window**: Models can only “see” a limited number of tokens (e.g., 8k, 16k, 128k).
- **Truncation**: When the context length is exceeded, earliest tokens are discarded.
- **Controls**:
- `max_tokens` / `max_new_tokens`: Caps output length.
- **Tools**: [tiktoken](https://github.com/openai/tiktoken), [HuggingFace Tokenizers](https://huggingface.co/docs/tokenizers)

2. Role Structure Delimiters

- **System / User / Assistant**: Role separation helps guide behavior.
- `system`: Sets up tone/persona.
- `user`: Asks the question.
- `assistant`: Responds.
- **Delimiters**: Use markdown, quotes, or custom delimiters (e.g., `[INST]`, triple backticks) to separate instructions.

3. Length Output Structure

- Use prompts like:
- “Respond in exactly 15 words.”
- “List 3 bullet points.”
- “Format the output as a CSV table.”
- **Formatting Options**:
- Markdown, bullet points, numbered lists.
- Code formatting: `Wrap code in triple backticks`.

4. Style, Tone, and Register

- **Control voice & audience**:

- “Explain like I’m 5.”
- “Reply in a legal memo format.”
- “Use a formal business tone.”
- **Common Styles**:
- Poetic, casual, humorous, academic, persuasive.

5. Divergent vs. Convergent Prompts

- **Divergent**: Encourage ideation/creativity.
- e.g., “Give 10 ideas for a new product.”
- **Convergent**: Drive focus/summary.
- e.g., “Summarize this paragraph into one sentence.”

6. Decoding Parameters (API-Level Controls)

- **Temperature**:
- 0.0 = deterministic
- 1.0 = creative/random
- **Top-k Sampling**: Only consider the top-k most likely tokens.
- **Top-p (Nucleus) Sampling**: Only consider the top tokens whose cumulative probability reaches p.
- **Presence Penalty**: Discourages or encourages new ideas.
- **Frequency Penalty**: Discourages repetition of the same phrases.

7. Prompt Debugging Readability

- Use versioned edits and isolate variables.
- Be explicit with format and instruction.
- Use modular prompts: reuseable building blocks.
- Participate in prompt libraries and shared repositories.

8. Common Pitfalls Fixes

Problem	Fix

Response too short or cut off	Raise `max_tokens`
Unclear response	Add examples and rephrase
Too repetitive	Use frequency penalty
Off-topic or inconsistent	Clarify role and context

9. Quick Examples

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.

10. Quick Diagnostic

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

Reference Links

- [OpenAI API](https://platform.openai.com/docs)
- [HuggingFace Transformers](https://huggingface.co/docs/transformers)
- [Prompting Guide](https://www.promptingguide.ai)