Claude vs GPT vs Gemini vs Grok Comparison

Reference Card · Updated April 2026

Pricing Comparison

MODEL	INPUT / OUTPUT (PER 1M TOKENS)	CONTEXT WINDOW	BEST BENCHMARK
Claude Opus 4.6	$5 / $25 (premium >200k)	1M (beta)	~80.8% SWE-Bench Verified
Claude Sonnet 4.6	$3 / $15	1M (beta)	79.6% SWE-Bench Verified
Grok 4.20	$2 / $6	2M	Real-time data + parallel agents
GPT-5.4	$2.50 / $15	1M (API)	87.3% investment benchmarks
Gemini 3.1 Pro	$2 / $12 (≤200k)	1M	Multimodal leader

Task Recommendations

TASK	BEST MODEL	WHY	PROMPTING TIP
Coding	Claude Sonnet 4.6	Near-Opus performance at one-fifth the price. Handles real GitHub-style patches reliably. Opus only for the absolute hardest multi-file refactors.	Wrap everything in XML tags. Use `<instructions>`, `<context>`, `<code>`, and `<output_format>`. Be explicit about edge cases.
Reasoning	Claude Opus 4.6	Strongest on complex, long-horizon chains that require careful step-by-step thinking.	Use XML tags for each reasoning stage. Add `<scratchpad>` for visible chain-of-thought before final answer.
Creative	GPT-5.4 or Gemini 3.1 Pro	GPT-5.4 for fluent prose and financial/structured creative work. Gemini for anything visual or stylistic.	For Gemini, lead with the image or video when possible. Keep temperature at 1.0. Describe desired aesthetic precisely.
Agentic workflows	Grok 4.20	Built-in parallel agents that debate and fact-check in real time. Real-time web/X data is a bonus.	Explicitly instruct it to activate agent mode or multiple perspectives. Say "use your research, math, and creative agents in parallel."
Long context	Grok 4.20 or Claude Sonnet 4.6	Grok has 2M native. Claude's 1M works cleanly with XML structure. GPT-5.4's long context has pricing gotchas above 272k.	For Claude, chunk documents with numbered `<document id="1">` tags. Summarize sections first before full analysis.
Multimodal	Gemini 3.1 Pro	Native video, image, and audio understanding in one model. Best scores on mixed media benchmarks.	Describe what you want analyzed first ("analyze the UI in this screenshot for accessibility issues") then attach media. Reference timestamps on video.
Budget-sensitive	Claude Sonnet 4.6 or Gemini 3.1 Pro	Sonnet gives near-frontier quality for coding and reasoning at reasonable rates. Gemini wins on pure multimodal value.	Cache prompts where available. Use Sonnet for 80% of work and only escalate to Opus when you hit a wall.

Key API Prompting Differences

Claude (Opus 4.6 and Sonnet 4.6) Claude responds best to structured XML tags. It treats them like clear partitions. Use this skeleton every time:

```
<instructions>
Your exact task goes here. Number steps if complex.
</instructions>

<context>
All background information.
</context>

<examples>
Any few-shot examples.
</examples>

<input>
The actual data or code to process.
</input>

<output_format>
Return only a JSON object with keys: summary, issues, recommendation.
</output_format>
```

This reduces hallucinations and improves consistency. Claude almost never ignores well-tagged instructions.

GPT-5.4 OpenAI-style markdown and system prompts work fine. Function calling and structured outputs are reliable. It handles JSON mode cleanly but doesn't need XML. Best prompt pattern: clear system role first, then user message with explicit formatting requests like "Respond in valid markdown with code blocks labeled by language."

Grok 4.20 OpenAI-compatible API so basic prompting carries over. The real power is in parallel agents. Prompt it with: "Activate your research agent for current data, math agent for calculations, and debate any assumptions before final answer." It pulls real-time information from X and web naturally. Mention "use parallel agents" when you want the full multi-perspective treatment.

Gemini 3.1 Pro Native multimodal means you can interleave text and media without special syntax. Best practice: describe the media explicitly in text first ("In the attached image, focus on the button layout and color contrast"). Temperature recommendation is 1.0 for most creative or analysis tasks. Keep instructions clear and separated from the media content.

Use Sonnet 4.6 for most coding and reasoning work. Escalate to Opus 4.6 only when Sonnet fails twice on the same task. Pick Gemini for any image or video input. Grok when you need live data or built-in debate. GPT-5.4 when investment modeling or polished creative output matters most.

If your workflow crosses these boundaries, route each subtask to the model that wins that column above. Mixing them beats forcing one model to do everything.