STRENGTHEN GUARDRAILS

Reduce hallucinations

Even the most advanced language models, like Claude, can sometimes generate text that is factually incorrect or inconsistent with the given context. This phenomenon, known as "hallucination," can undermine the reliability of your AI-driven solutions. This guide will explore techniques to minimize hallucinations and ensure Claude's outputs are accurate and trustworthy.

Basic hallucination minimization strategies

Allow Claude to say "I don't know": Explicitly give Claude permission to admit uncertainty. This simple technique can drastically reduce false information.

Example: Analyzing a merger & acquisition report

Role	Content
User	As our M&A advisor, analyze this report on the potential acquisition of AcmeCo by ExampleCorp. <report> {{REPORT}} </report> Focus on financial projections, integration risks, and regulatory hurdles. If you're unsure about any aspect or if the report lacks necessary information, say "I don't have enough information to confidently assess this."

Use direct quotes for factual grounding: For tasks involving long documents (>20K tokens), ask Claude to extract word-for-word quotes first before performing its task. This grounds its responses in the actual text, reducing hallucinations.

Example: Auditing a data privacy policy

Role	Content
User	As our Data Protection Officer, review this updated privacy policy for GDPR and CCPA compliance. <policy> {{POLICY}} </policy> 1. Extract exact quotes from the policy that are most relevant to GDPR and CCPA compliance. If you can't find relevant quotes, state "No relevant quotes found." 2. Use the quotes to analyze the compliance of these policy sections, referencing the quotes by number. Only base your analysis on the extracted quotes.

**Verify with citations: Make Claude's response auditable by having it cite quotes and sources for each of its claims. You can also have Claude verify each claim by finding a supporting quot after it generates a response. If it can't find a quote, it must retract the claim. </callout>

Example: Drafting a press release on a product launch

Role	Content
User	Draft a press release for our new cybersecurity product, AcmeSecurity Pro, using only information from these product briefs and market reports. <documents> {{DOCUMENTS}} </documents> After drafting, review each claim in your press release. For each claim, find a direct quote from the documents that supports it. If you can't find a supporting quote for a claim, remove that claim from the press release and mark where it was removed with empty [] brackets.

Chain-of-thought verification: Ask Claude to explain its reasoning step-by-step before giving a final answer. This can reveal faulty logic or assumptions.

Best-of-N verficiation: Run Claude through the same prompt multiple times and compare the outputs. Inconsistencies across outputs could indicate hallucinations.

Iterative refinement: Use Claude's outputs as inputs for follow-up prompts, asking it to verify or expand on previous statements. This can catch and correct inconsistencies.

External knowledge restriction: Explicitly instruct Claude to only use information from provided documents and not its general knowledge.

> Remember, while these techniques significantly reduce hallucinations, they don't eliminate them entirely. Always validate critical information, especially for high-stakes decisions.