

Build an AI Content Moderation System with AI

Execution Path

A structured route from idea to working output - the steps, outcomes, workflows, resources, and tools, in order.

Category: AI Systems Steps: 8 Phases: 3

Exported: 2026-07-21 (UTC)

Source: <https://newprompt.net/blueprints/build-ai-content-moderation-system-with-ai/execution-map>

PATH SUMMARY

8 steps across 3 phases.

Use this document as a portable execution guide. Each step lists the expected outcome and the workflow, resource, and tool used in NewPrompt.

PHASE 1 · Define & Scope

Clarify what you're building and for whom.

Step 1 · Define the moderation policy and label taxonomy

Frame the policy before the model touches anything - the categories of violating content, the label set, and the edge cases that decide borderline calls - so classification applies an explicit standard instead of an implied one.

OUTCOME

A moderation policy and a closed label taxonomy the system enforces.

USED IN THIS STEP

Workflow AI Product Requirements Workflow

Resource Product Requirements Assistant

Tool System Prompt Generator

PHASE 2 · Build & Refine

Build, test, secure, and make it production-ready.

Step 2 · Extract the signals from user content

Pull the parts of the content that matter to the policy - the text, the claims, the context - out of the surrounding noise, so the classifier judges the right unit rather than a whole messy payload.

OUTCOME

The policy-relevant signals isolated from raw user content.

USED IN THIS STEP

Workflow AI Data Extraction Workflow
Resource Extract Data From Text with AI
Tool Extraction Prompt Generator

Step 3 · Classify content against policy

Assign a label from the closed taxonomy with tie-breaking rules, so the system returns one allowed verdict per item consistently instead of inventing a category or hedging on a call an enforcement path can't act on.

OUTCOME

One validated policy label per item, applied consistently.

USED IN THIS STEP

Workflow AI Classification Workflow
Resource Text Classification Prompt - the Anatomy
Tool Data Classification Prompt

Step 4 · Produce structured moderation decisions

Shape the verdict into a strict, machine-actionable decision - label, confidence, reason, recommended action - so an enforcement system or a review queue consumes a predictable contract every time, not free-form text.

OUTCOME

A strict moderation-decision contract an enforcement system can act on.

USED IN THIS STEP

Workflow AI Reliable JSON Output Workflow
Resource Turn a JSON Schema into a Prompt
Tool JSON Output Prompt Builder

Step 5 · Evaluate false positives, negatives, and edge cases

Build the test set the system must pass - content with known correct verdicts, including the borderline and adversarial cases - so false positives and false negatives are measured before the system acts on real user content.

OUTCOME

Moderation accuracy measured against known and adversarial cases.

USED IN THIS STEP

Workflow AI Agent Evaluation Workflow
Resource Agent Test Scenario Prompt
Tool Test Case Prompt Generator

Step 6 · Wire enforcement actions and review queues

Design the integrations that turn a decision into an outcome - auto-actions on high-confidence verdicts, a human review queue for the rest, webhook events into the product - with the retries and failure handling a consequential action demands.

OUTCOME

Enforcement and review-queue integration contracts with human-in-the-loop.

USED IN THIS STEP

Workflow AI Integration & Webhook Workflow
Resource Turn a JSON Schema into a Prompt
Tool JSON Output Prompt Builder

Step 7 · Review security and abuse risks

Review the system the way an adversary would - evasion attempts, prompt injection hidden in user content, reviewer-queue access - and back the findings with tests, because a moderation system is a target the moment it has teeth.

OUTCOME

Evasion, injection, and reviewer-access risks reviewed for vulnerabilities.

USED IN THIS STEP

Workflow AI Security Review Workflow
Resource Security Code Review Prompt
Tool Code Review Prompt Generator

PHASE 3 · Ship & Validate

Ship with confidence and validate results.

Step 8 · Prepare the release

Cross from built to live - readiness checks, a deploy and rollback plan, and monitoring on decision quality and queue volume - so launch is deliberate and a spike in false positives or a flood of missed violations is caught fast.

OUTCOME

The moderation system shipped with a rollback path and quality monitoring.

USED IN THIS STEP

Workflow AI Deployment & Release Workflow
Resource Production Readiness Review Prompt
Tool Markdown Output Builder

This PDF is a static export of the NewPrompt execution path. It does not include workspace progress, personal notes, saved outputs, or project state.