Original Prompt and Answers

Prompt

“‘Problem R1: Compliance Prohibitions, Persona Pushes Loopholes

Setup (policy excerpt)

You are given this compliance snippet for a child-focused messaging app: - The system must not allow sexual content involving minors. - Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved. - The system must not reveal the exact reason a message was blocked in a way that enables bypass. - The system must not allow a blocked user to infer being blocked by timing differences.

Task

Write 8 to 10 requirements with rationale and fit criteria.

Required output format

For each requirement: - REQ-ID - Statement - Rationale - Fit criterion - Assumptions or open questions

Based on the previous question:

You are a senior product manager with 10+ years of experience in consumer messaging apps and growth. You write crisp requirements that balance user experience and business goals.

Convert the compliance excerpt into 8 to 10 requirements with rationale and fit criteria. Be concise and pragmatic. “‘

GitHub Copilot GPT-4o Answers:

Requirements for Compliance Prohibitions and Persona Pushes Loopholes

REQ-001: Prohibit Upload of Sexual Content Involving Minors

- **Statement**: The system must implement automated detection and blocking of sexual content involving minors during upload.
- **Rationale**: Prevents illegal and harmful content from being shared, ensuring compliance with laws and protecting users.
- **Fit Criterion**: 100% of flagged content involving minors is blocked before it is visible to any user.
- **Assumptions/Open Questions**: What detection technology (e.g., AI, hash matching) will be used?

REQ-002: Restrict Employee Access to Messages

- **Statement**: Employee access to user messages must require explicit user consent, logging, and prior approval.
- **Rationale**: Protects user privacy and ensures compliance with internal policies.
- **Fit Criterion**: 100% of employee access events are logged, approved, and linked to a user consent record.
- **Assumptions/Open Questions**: What system will manage access approvals and logging?

REQ-003: Obfuscate Block Reasons

- **Statement**: The system must provide generic feedback (e.g., “Message violates community guidelines”) when a message is blocked.
- **Rationale**: Prevents users from reverse-engineering the blocking rules to bypass them.
- **Fit Criterion**: No blocked user can infer the specific reason for a block based on system feedback.
- **Assumptions/Open Questions**: How will the system handle false positives in blocking?

REQ-004: Prevent Timing-Based Inference of Blocks

- **Statement**: The system must ensure that blocked messages and successful messages have identical processing times.
- **Rationale**: Prevents users from inferring blocks based on timing differences.
- **Fit Criterion**: Processing time for blocked and successful messages must not differ by more than 50ms.
- **Assumptions/Open Questions**: What is the acceptable latency threshold for the app?

REQ-005: Implement User Reporting for Violations

- **Statement**: The system must allow users to report messages suspected of violating policies.
- **Rationale**: Empowers users to assist in maintaining a safe environment.
- **Fit Criterion**: 95% of user reports are reviewed within 24 hours.
- **Assumptions/Open Questions**: What team will handle user reports?

REQ-006: Educate Users on Content Policies

- **Statement**: The system must provide clear, accessible content guidelines during onboarding and in the app settings.
- **Rationale**: Reduces the likelihood of unintentional violations by users.
- **Fit Criterion**: 90% of new users view the content policy within their first session.
- **Assumptions/Open Questions**: How will the system track user engagement with the policy?

REQ-007: Log All Blocked Messages

- **Statement**: The system must log all blocked messages, including metadata, for audit and review purposes.
- **Rationale**: Enables compliance audits and helps improve blocking accuracy.
- **Fit Criterion**: 100% of blocked messages are logged with metadata (e.g., timestamp, user ID, reason).
- **Assumptions/Open Questions**: How long will logs be retained?

REQ-008: Provide Appeals for Blocked Messages

- **Statement**: The system must allow users to appeal blocked messages through a review process.
- **Rationale**: Ensures fairness and addresses false positives in message blocking.
- **Fit Criterion**: 95% of appeals are resolved within 48 hours.
- **Assumptions/Open Questions**: What team will handle appeals, and what is the escalation process?

REQ-009: Regularly Update Blocking Rules

- **Statement**: The system must update content-blocking rules quarterly based on new threats and user feedback.
- **Rationale**: Keeps the system effective against evolving content risks.
- **Fit Criterion**: Blocking rules are updated at least once per quarter.
- **Assumptions/Open Questions**: What process will be used to gather feedback and threats?

REQ-010: Test Blocking System for Loopholes

- **Statement**: The system must undergo quarterly penetration testing to identify and close loopholes in content blocking.
- **Rationale**: Ensures the system remains robust against attempts to bypass blocking mechanisms.
- **Fit Criterion**: 100% of identified loopholes are resolved within one release cycle.
- **Assumptions/Open Questions**: Who will conduct penetration testing, and how will findings be prioritized?

These requirements balance compliance, user experience, and operational feasibility while addressing the provided policy excerpt.

Updated Guidelines Prompt and Answers

Prompt

Prompt START

Problem R1: Compliance Prohibitions, Persona Pushes Loopholes

Setup (policy excerpt)

You are given this compliance snippet for a child-focused messaging app: - The system must not allow sexual content involving minors. - Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved. - The system must not reveal the exact reason a message was blocked in a way that enables bypass. - The system must not allow a blocked user to infer being blocked by timing differences.

Task

Write 8 to 10 requirements with rationale and fit criteria.

Required output format

For each requirement: - REQ-ID - Statement - Rationale - Fit criterion - Assumptions or open questions

Answer this question with this instructions: SYSTEM: Stage: Specification. Role: REQUIREMENTS ANALYST.

Priorities (in case of conflict): 1) Safety and compliance 2) Privacy 3) Usability 4) Business goals

Rules: - Do not invent facts, features, or exceptions that are not explicitly in the policy excerpt. - Keep true prohibitions as SHALL NOT. Avoid double negatives. - Every requirement must be traceable to the excerpt: include an Evidence field quoting or pointing to the exact bullet(s) it came from. - Include Confidence (High/Medium/Low). If Confidence is Low, write UNKNOWN and add an Open question instead of guessing. - Do not reveal bypass-enabling details in your own wording. Requirements about “reasons for blocking” must be phrased at a high level (no implementation hints).

After drafting requirements, run a Loophole Scan: - List 3 to 5 plausible bypass interpretations or ambiguities in your drafted requirements. - Tighten wording for any item that could be misread to allow a bypass, while staying faithful to the excerpt.

USER: Policy excerpt for a child-focused messaging app: - The system must not allow sexual content involving minors. - Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved. - The system must not reveal the exact reason a message was blocked in a way that enables bypass. - The system must not allow a blocked user to infer being blocked by timing differences.

Task: Write 8 to 10 requirements with rationale and fit criteria.

Required output format: For each requirement output:

REQ-ID: Statement: Rationale: Fit criterion: Evidence: Confidence: Assumptions or open questions:

Then output: Loophole Scan: - (bullets)

Prompt End

GitHub Copilot GPT-4o Answer:

Requirements Specification

REQ-001 **Statement:** The system SHALL NOT allow the transmission, storage, or display of sexual content involving minors.
Rationale: Ensures compliance with legal and ethical standards for child safety.
Fit criterion: Any attempt to transmit, store, or display sexual content involving minors is blocked, logged, and reported to the appropriate authorities.
Evidence: “The system must not allow sexual content involving minors.”
Confidence: High
Assumptions or open questions: How will the system identify and classify such content?

REQ-002 **Statement:** Employees SHALL NOT access user message content unless the user explicitly requests support, and all access is logged and approved.
Rationale: Protects user privacy and ensures compliance with access restrictions.
Fit criterion: Access logs show that all employee access to message content is

tied to explicit user requests and prior approval.
Evidence: “Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved.”
Confidence: High
Assumptions or open questions: What constitutes “explicit” user requests, and how will approval workflows be implemented?

REQ-003 **Statement:** The system SHALL obscure the exact reason for message blocking to prevent users from inferring bypass methods.
Rationale: Prevents users from reverse-engineering the blocking mechanism, ensuring compliance and safety.
Fit criterion: Blocked messages display a generic error message without revealing specific reasons for blocking.
Evidence: “The system must not reveal the exact reason a message was blocked in a way that enables bypass.”
Confidence: High
Assumptions or open questions: What level of detail is acceptable in error messages?

REQ-004 **Statement:** The system SHALL ensure that blocked users cannot infer their blocked status through timing differences in message delivery.
Rationale: Prevents users from detecting blocking mechanisms, maintaining system integrity.
Fit criterion: Message delivery times for blocked and unblocked users are indistinguishable.
Evidence: “The system must not allow a blocked user to infer being blocked by timing differences.”
Confidence: High
Assumptions or open questions: What mechanisms will ensure consistent timing for blocked and unblocked users?

REQ-005 **Statement:** The system SHALL log all instances of blocked content, including metadata, for auditing purposes.
Rationale: Provides traceability and accountability for content moderation decisions.
Fit criterion: Audit logs include timestamps, user IDs, and content metadata for all blocked messages.
Evidence: “The system must not allow sexual content involving minors.”
Confidence: High
Assumptions or open questions: How long should logs be retained, and who has access to them?

REQ-006 **Statement:** The system SHALL implement automated detection mechanisms for identifying sexual content involving minors.
Rationale: Ensures proactive prevention of prohibited content transmission.
Fit criterion: Automated detection mechanisms flag and block 100% of known prohibited content.
Evidence: "The system must not allow sexual content involving minors."
Confidence: Medium
Assumptions or open questions: What technologies (e.g., AI, keyword detection) will be used for automated detection?

REQ-007 **Statement:** The system SHALL provide users with a mechanism to request support explicitly, which triggers logged and approved employee access.
Rationale: Ensures compliance with access restrictions while allowing users to seek help.
Fit criterion: Support requests are logged, and employee access is tied to these requests.
Evidence: "Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved."
Confidence: High
Assumptions or open questions: What user interface elements will facilitate explicit support requests?

REQ-008 **Statement:** The system SHALL provide training for employees on compliance with content access and moderation policies.
Rationale: Ensures employees understand and adhere to compliance requirements.
Fit criterion: Training completion records are maintained for all employees, and periodic assessments confirm understanding.
Evidence: "Employees must not access message content unless (a) user explicitly requests support and (b) access is logged and approved."
Confidence: High
Assumptions or open questions: What training materials and frequency are required?

REQ-009 **Statement:** The system SHALL provide a mechanism for users to appeal blocked messages without revealing the specific reason for blocking.
Rationale: Balances usability with compliance by allowing users to contest decisions without enabling bypass.

Fit criterion: Appeals do not disclose specific blocking criteria and are resolved within a defined timeframe.
Evidence: "The system must not reveal the exact reason a message was blocked in a way that enables bypass."
Confidence: Medium
Assumptions or open questions: What process will be used to handle appeals?

REQ-010 **Statement:** The system SHALL ensure that all blocked content is reviewed periodically to improve detection mechanisms.
Rationale: Enhances the accuracy and effectiveness of content moderation over time.
Fit criterion: Review logs show periodic evaluations of blocked content and updates to detection mechanisms.
Evidence: "The system must not allow sexual content involving minors."
Confidence: Medium
Assumptions or open questions: Who will conduct these reviews, and how frequently?

Loophole Scan

1. **Ambiguity in "explicit user requests" (REQ-002):** Clarify what constitutes an explicit request to avoid misuse.

2. **Generic error messages (REQ-003):** Ensure error messages do not inadvertently hint at blocking criteria.

3. **Timing consistency (REQ-004):** Define acceptable timing thresholds to prevent subtle timing-based inferences.

4. **Automated detection (REQ-006):** Address potential false positives or negatives in detection mechanisms.

5. **Appeals process (REQ-009):** Ensure the appeals process does not inadvertently disclose blocking logic.

Tightened Wording

- REQ-002: Define "explicit user requests" as actions requiring user authentication and confirmation.

- REQ-003: Specify that error messages must not include any content-related details.

- REQ-004: Add a threshold for timing consistency (e.g., ±50ms).
- REQ-006: Include a requirement for periodic testing of detection mechanisms to minimize false positives/negatives.
- REQ-009: Ensure appeals are handled by a separate team with no access to blocking algorithms.

Analysis between Old and Updated Prompts and Answers

Got it. Here is the analysis without that part, focused only on why the old prompt and answer are bad, and why the updated prompt makes the output better, so R1 works as a counterexample.

Why the old prompt and Copilot answer are bad

1) The persona injects the wrong incentives

The old system message frames the model as a "senior product manager" focused on growth and balancing UX and business goals. That is misaligned with R1, which is a compliance translation task with hard prohibitions. A growth PM framing pushes the model toward "helpful product completeness" instead of strict policy fidelity.

2) Scope creep happens immediately

The task is to convert a 4-bullet compliance excerpt into requirements. The old answer expands into multiple additional product and operational features (reporting, education, appeals, quarterly updates, penetration testing). Those may be good ideas in general, but they are not justified by the excerpt. This breaks the conversion intent.

3) Fake precision is introduced

The old answer invents numeric targets and thresholds such as 100%, 95%, 90%, and specific timing bounds. None of these are supported by the excerpt. This is dangerous in requirements work because it looks contract-ready and testable while being completely ungrounded.

4) It drifts into solution design

Several requirements specify implementation choices (automated detection during upload, specific operational processes). The excerpt never specifies how the prohibitions are enforced. The prompt's "concise and pragmatic" plus PM persona makes design assumptions more likely.

5) It undermines the prohibition semantics

Compliance language is often most accurately expressed as "shall not" constraints. A growth/product persona tends to reframe prohibitions into softer, positive requirements ("educate users", "allow appeals", "improve blocking") which can dilute the strictness of the original compliance intent.

Net effect: the old prompt produces a polished document that is not a faithful transformation of the compliance excerpt. This is exactly the type of failure you want from a counterexample to "assign a role/persona" when the role is poorly chosen.

Why the updated v3 prompt is good

1) Minimal role reduces bias

"Role: REQUIREMENTS ANALYST" is a low-prior, low-incentive role. It is designed to translate and structure requirements rather than invent product strategy. This directly reduces persona-driven scope expansion.

2) Stage-first framing matches the task

"Stage: Specification" tells the model this is formal requirements work, not product ideation. That decreases solution-jumping and helps keep the output aligned with compliance translation.

3) Explicit priority ordering prevents goal hijacking

Safety/compliance first, then privacy, then usability, then business goals. This overrides the common product instinct to trade safety for UX convenience and forces decisions to land on the compliance side when there is tension.

4) Traceability forces discipline

Requiring an Evidence field tied to the excerpt (and confidence/unknown handling) shifts the model from "sounds reasonable" to "show your work." This reduces hallucination and makes the output auditable.

5) Prohibition-friendly wording preserves meaning

"Keep true prohibitions as SHALL NOT" prevents semantic drift. In this domain, negation is not a style flaw. It is the point.

6) Loophole scan improves robustness

Adding a loophole scan makes the model actively search for ambiguous interpretations and tighten wording. That is exactly what you want for compliance constraints and adversarial contexts.

Why this makes R1 a strong counterexample

R1 demonstrates that a naive application of role assignment ("expert PM, pragmatic") can produce worse outcomes by injecting incentives that conflict with strict compliance translation. The updated v3 prompt fixes this by making role assignment conditional, stage-aware, priority-constrained, and evidence-driven.