Discord In-DM Image Scam Detection — Product Requirements Document

Author: Benson Nguyen **Status:** Draft / Concept Exploration **Date:** August 2026

---

1. Problem Statement

Discord members in gaming and crypto/finance-adjacent communities are increasingly targeted by image-based phishing scams, including screenshots of fake gift offers, doctored "official Discord" notices, and QR codes, that evade the platform's existing text-based moderation systems. Discord removed over 3 million accounts for scam-related activity in the past year, yet current automated defenses were not built to catch threats embedded in images rather than text. This leaves members to rely on manual vigilance to avoid financial and account-security harm.

How this differs from Discord's existing unknown-sender warnings: Discord already warns users when a message comes from a stranger or new account, a sender-identity signal. This does not address cases where the threat comes from a compromised but familiar account (e.g., a hacked friend's or admin's account being used to mass-message real contacts, as documented in real-world incidents), since the message appears to come from someone the recipient already trusts. This proposal instead evaluates the content of the image itself, regardless of who sent it, closing a gap that sender-based warnings structurally cannot cover.

---

2. Target User

Primary: An active member, roughly ages 16–34, matching Discord's core demographic, of a mid-to-large gaming or crypto-adjacent server, who regularly receives DMs or sees posts from unfamiliar bots or accounts and has limited technical background to independently verify whether an image-based offer is legitimate.

Secondary/beneficiary: Server moderators, who would see a reduction in scam-related reports requiring manual review. Included as a supporting stakeholder, not the primary design target.

---

3. Why Now

The shift to image-based scam delivery is a 2026-specific attack pattern that emerged specifically to evade existing text-scanning moderation. This is a live, unaddressed gap, not a solved problem. Building this natively also lets it apply platform-wide by default, rather than depending on individual servers to install and configure a third-party bot.

---

4. Scope

In scope (v1): Direct messages (DMs) only. DMs are structurally private, meaning server moderators have no visibility into them, unlike public channels where existing moderation and human oversight already catch a meaningful share of scam activity.

Out of scope (v1): Server/public channel scanning, video content scanning, any mechanism requiring sender opt-in.

---

5. Goals

1. Reduce the number of Discord members who fall victim to image-based phishing scams delivered via direct message.
2. Surface a clear, low-friction credibility signal at the moment a user is about to engage with a risky image, without requiring the user to have any security expertise.
3. Address a gap that server-level moderation cannot reach, since mods have no visibility into private DMs between a scammer and a target.

---

6. Requirements (MoSCoW, v1 Scope)

Must have

- Automatic scanning of images sent in DMs for embedded text and QR codes (OCR-based), checked against known scam patterns and link databases
- An in-DM warning indicator shown directly on a flagged image, before the user taps or clicks anything in it
- A simple "why was this flagged" explanation panel (e.g., "this image contains a link commonly associated with phishing")
- A one-tap report and block action from the warning panel

Should have

- A short, dismissible first-time education prompt explaining what image-based scams are, shown the first time a user encounters a flag
- Confidence-level tiering on warnings (low / medium / high), rather than a single flat flag, so users are not overwhelmed by false positives

Could have

- User-level reporting trends visible to Discord's Trust & Safety team, feeding back into the detection model over time

Won't have (v1)

- Server/public channel scanning
- Scanning of video content
- Any feature requiring the sender to opt in, since scammers will not opt into being flagged

---

7. Evidence & Validation

Platform-level evidence: Discord's 2026 Transparency Report shows the platform removed over 3 million accounts for deceptive practices, including phishing and scams, over a 12-month period. Industry research indicates image-based phishing (screenshots, QR codes) became the dominant attack vector in 2026 specifically because it evades text-based moderation systems.

Peer survey (n=120): To validate this problem directly, I surveyed 120 Discord users across multiple gaming, crypto/finance, and university-affiliated servers.

- 75.8% (91 of 120) reported having received a suspicious image, screenshot, or QR code in a Discord DM from an unfamiliar sender or bot
- Nearly 1 in 5 respondents had a near-miss or worse: 14.3% almost interacted with the image before realizing it was a scam, and 5.5% did interact before catching on
- The most commonly cited red flags were unexpected DMs (22.5%) and suspicious links or QR codes (20.8%), directly reinforcing the image/QR-based attack pattern identified in platform research
- Only 36.7% had ever reported a scam, while 47.5% had not and 15.8% didn't know how, indicating a real gap in existing reporting awareness, not just enforcement
- 83.3% said they would be very or somewhat likely to read an automated warning label before dismissing it, supporting the core design decision to include a visible, explained warning rather than silent filtering

- Respondents were drawn from gaming (45.8%), crypto/finance (29.2%), and mixed (18.3%) communities, confirming the target user segmentation

Competitive landscape: Existing tools like Phantom's Anti-Scam module address image-based scams primarily at the server/public-channel level. No major solution, including Discord's native moderation, currently addresses this pattern within DMs, where server moderators have no visibility.

---

8. Success Metrics

Framed as hypotheses and target directions to validate post-launch, not claimed results.

- **Reduce near-miss/interaction rate:** baseline from validation survey shows 19.8% of users almost interacted or did interact with a scam image before recognizing it. Success means a measurable drop in this rate among users exposed to the in-DM warning.
- **Warning engagement rate:** 83.3% of surveyed users said they'd likely read an automated warning. Success means confirming this holds in practice, measured as the percentage of users who open the "why flagged" panel rather than immediately dismissing it.
- **Reporting gap closure:** only 36.7% of surveyed users had ever reported a scam, with 15.8% not knowing how. Success means an increase in report submissions directly from the warning panel's one-tap report action.
- **False positive rate:** not directly measurable from survey data, but critical to track post-launch, since over-flagging legitimate images would erode trust in the warning system. Flagged here as an open question rather than an invented number.

---

9. Risks & Open Questions

False positives eroding trust If the system flags legitimate images too often, users may start ignoring warnings altogether, defeating the purpose. This is the single biggest risk to the product's effectiveness and would need to be closely monitored post-launch, likely starting with a conservative, high-confidence-only threshold in early rollout.

Privacy concerns with scanning DM content Discord DMs are not end-to-end encrypted, and Discord's existing Terms of Service and Privacy Policy already permit automated scanning of DM content for high-harm categories such as malware, spam, and CSAM. This proposal does not introduce a new category of platform access, it extends an existing, already-disclosed scanning capability to one additional threat type (image-based scams). The genuine open question is not whether Discord can technically do this, but whether adding this category should

require more explicit user-facing transparency than current malware/spam scanning receives, and whether an opt-out should exist for users who would rather forgo the protection than have any additional content category scanned. This is a real product and legal judgment call, not a purely technical one, and would need input from Discord's policy and legal teams before launch.

Scammer adaptation Scammers already shifted from text to images specifically to evade moderation; a successful image-scanning system would likely push them toward new evasion methods (e.g., embedding text in video frames, using steganography, or altering images to defeat OCR). This is an open question requiring ongoing detection updates rather than a one-time fix.

Survey sample limitations The validation survey (n=120) draws from gaming, crypto/finance, and university-affiliated servers, which may not represent Discord's full user base (e.g., older users, non-English-speaking communities, casual/general-use servers were underrepresented). Broader validation would be needed before a full platform-wide launch decision.

Detection accuracy dependency The entire product depends on OCR and pattern-matching being reliably accurate across varied image formats, languages, and evolving scam templates. This is a genuine open technical question this PRD does not attempt to solve, and would require close collaboration with Discord's Trust & Safety and ML teams to validate feasibility.

---

10. Rollout Considerations

A phased rollout, starting with opt-in beta testing in a subset of gaming and crypto-adjacent servers similar to those surveyed, would allow the false-positive rate and warning engagement data to be validated in practice before expanding platform-wide.
