Red Team Technique (Adversarial Self-Critique)

Was es ist

Eine zweistufige Methode:

1. Zuerst lässt du die KI **Inhalte erzeugen** (z. B. Lebenslauf, Business-Proposal, Cold-Outreach-Mail, Marketing-Entwurf).
2. Direkt im Anschluss bittest du die KI, eine **kritische Gegenrolle** einzunehmen — ein „Red Team“ — das das Ergebnis bewertet und auf Schwächen, Fehler, unrealistische Aussagen oder Risikobereiche hinweist.

Warum das funktioniert

- Hilft dir, **reale Einwände, Schwachstellen oder Dealbreaker** frühzeitig zu erkennen, bevor du Inhalte echten Menschen präsentierst.
- Ermöglicht gezielte Nachbesserung im Vorfeld — Inhalte werden belastbarer, glaubwürdiger und weniger naiv.
- Macht **Bias, blinde Flecken, schwache Argumentation oder unüberzeugende Passagen** sichtbar, die man selbst leicht übersieht.

Typische Workflows / Beispiele

- **Bewerbung**:
  - Schritt 1: Lebenslauf auf eine konkrete Stellenbeschreibung zuschneiden.
  - Schritt 2: KI als Hiring Manager einsetzen, um Red Flags zu markieren — hilft, Klarheit zu verbessern, Stärken zu schärfen und Fülltext zu entfernen.
- **Business-Proposal**:
  - Schritt 1: Vorschlag für CFO oder Management entwerfen.

    - Schritt 2: KI als CFO bewerten lassen, um finanzielle Risiken, schwache ROI-Argumente oder Begründungslücken aufzudecken.

- **Cold Outreach / Marketing-Mail**:

    - Schritt 1: Outreach-Mail schreiben.
    - Schritt 2: KI als gestressten Empfänger reagieren lassen — identifiziert spamartige, schwache oder unglaubwürdige Formulierungen zum Streichen oder Überarbeiten.

Best Practices für „Red Teaming“

- Nutze **sehr konkrete Personas** für die Kritik
  (z. B. „risikoscheuer CTO mit Fokus auf Datensicherheit“ statt nur „Kritiker“).
  Je klarer Motivation und Perspektive, desto relevanter das Feedback.
- **Schließe den Kreis**: Bitte die KI nach der Kritik gezielt, die schwächsten Stellen auf Basis der eigenen Einwände zu verbessern.
- Setze diese Methode als **Qualitätsfilter** für Inhalte mit hohem Einsatz ein (Bewerbungen, Angebote, externe Kommunikation) oder um überzeugende und formale Texte zu stärken.

Empfohlene Lektüre & Ressourcen

- Praxisnahe Übersicht, die Red Teaming als eine der zentralen fortgeschrittenen Prompting-Techniken aufführt.
  (Upaspro)
  (Five-in-One Amplifier (Content Amplification via AI))

- Akademische Diskussion zur Notwendigkeit strukturierter und adversarialer Prüfung im Prompt-Design, um Fragilität zu reduzieren und Skalierbarkeit von LLMs zu verbessern.
  (SSRN)

- Allgemeine Prompt-Engineering-Ressourcen zu rollenbasiertem Prompting, Thought-Sequenzen und dem größeren Kontext — hilfreich in Kombination mit Red Teaming.
  (Medium)

Revision #3

Created 2025-11-28 09:27:57 UTC by Carsten

Updated 2026-01-30 20:04:18 UTC by Carsten