Bias‑Detektion: Dreifach‑Vergleich (Description/Embodiment/Feedback‑PALD) zur Markierung von Halluzinationen & Stereotypisierungen (z. B. ungewollte Verjüngung weiblicher Agenten).

PROMT-Enwurf:

**Sofort nur leichte PALD‑Extraktion** (für schnelle UI), **schwere Stereotypen‑Analyse später** (asynchron), gesteuert über Flags. Ergebnisse werden strukturiert persistiert, damit die PALD‑Struktur über Zeit „wächst“.

**System/Developer‑Prompt für KIRO (einsetzbar als Tool‑Aufruf)**

Aufgabe:

1. **PALD‑Light jetzt:** Extrahiere aus description\_text und (falls vorhanden) embodiment\_caption je eine **PALD‑JSON‑Light** (nur belegte Felder auf Global/Medium/Detail, ohne Heuristiken).
2. **Stereotypen‑Analyse später (flag‑gesteuert):** Wenn defer\_bias\_scan==true, **füge keinen** Bias‑Befund zur Antwort für die UI hinzu. Stattdessen schreibe einen **Analyse‑Job** in stereotype\_queue, der die schwere Auswertung offline durchführt.
3. **Persistenz:** Schreibe alle Artefakte (Input‑IDs, PALD‑Light, Diffs, Meta) in pald\_store (append‑only).
4. **Diff‑Berechnung:** Erzeuge pald\_diff = Delta zwischen Description‑PALD und Embodiment‑PALD. Markiere jedes Feld als {match|hallucination|missing}.
5. **Job‑Payload für später:** Lege bias\_job mit folgenden Prüfpunkten ab:
   * **Age‑Shift:** gemeldetes Alter vs. visuelle Altersschätzung; gerichtete Abweichung zu jüngeren weiblichen Agents markieren.
   * **Gender‑Conformity & Sexualisierung:** Abgleich Kleidungsstil/Körperproportionen gegen Beschreibung/Kontext; Indizienliste (z. B. „enge Kleidung“, „überproportionale Taille‑Hüfte“).
   * **Ethnizität/Hautton:** Konsistenz zu Beschreibung; Flags nur als technische Marker, keine Profilierung.
   * **Berufs‑/Rollen‑Klischees:** Abgleich Rolle↔Darstellungsattribute (z. B. „Assistentin“ vs. Laborleitung).
   * **Ambivalente Stereotype:** gleichzeitige Markierung gegensätzlicher Signale (z. B. „kompetent“ bei infantilisiertem Stil).
   * **Mehrfach‑Stereotypisierung:** gemeinsame Auftretensmuster.
6. **Compliance:** Keine Speicherung personenidentifizierbarer Daten in den Artefakten; nur Pseudonyme/Session‑IDs.
7. **Rückgabe an UI:** Nur pald\_light + pald\_diff\_summary + defer\_notice (Hinweis, dass eine vertiefte Analyse später erfolgt).

Aber anpassen: PALD wird für die Bilderstellung immer nötig sein, da es für die Komprimierung der Prompts nötig sist. Ggf. (je nachdem wie die Qualität der Bilder aktuell geprüft wird-> wird die Qualität der Bilder auch dahingehend geprüft, dass das erstellte Bild die Vorgaben enthält?), dh. es gibt sozusagen-MUSS-PALDS und solche, die für z.B. Stereotypen/Hallzintationsentdeckung gebraucht werden. Die muss müssen auch sofot erstellt werden, die anderen, können in einem „Nachlauf nach Ende der Session mit dem Benutzer erstellt werden.