AI REIMAGINE PROJECT · FILE 03

Responsible AI & Risk Plan

FitSense - keeping the AI fit-advisor safe, fair & trustworthy

Prepared by Abhishek Katkar · June 2026

The spine of this plan is the confidence gate from File 2. FitSense advises on size; it does not decide for the shopper. Its single most important safety property is honesty about its own certainty - it must never present a confident size it has not earned. Everything below builds on that. Because the feature uses personal data (body measurements, purchase history) and reasons in natural language, privacy, fairness, and hallucination get specific attention.

1. Possible AI Failure Cases

Realistic failures specific to FitSense, not generic AI risks.

1. **FC1 - Confident-wrong recommendation.** Shows “High confidence” but the garment runs small or large, so the user gets a misfit. The most damaging failure - it weaponises the trust the confidence label creates.
2. **FC2 - Construction mis-extraction.** The LLM misreads the product description and labels a slim-fit garment “relaxed,” skewing both the size and the reason.
3. **FC3 - Cross-brand normalisation error.** Maps one brand’s “M” to the wrong reference, recommending the wrong size when the user switches brands.
4. **FC4 - Cold-start over-confidence.** A first-time buyer supplies only self-reported (often inaccurate) measurements, yet the system shows Medium or High confidence it has not earned.
5. **FC5 - Hallucinated reason.** The size may be fine, but the stated reason cites an attribute the garment does not have (“stretch fabric” when it has none), misleading the user.
6. **FC6 - Stale or contradictory profile.** Past returns were for non-fit reasons (wrong colour, changed mind), or the user’s body changed, so the history-derived fit signal is wrong.
7. **FC7 - Latency / timeout at peak.** During sale events the recommendation misses its budget; the PDP must degrade gracefully rather than stall or show nothing.

2. Responsible AI Risks

Risk	How it shows up in FitSense
Privacy	Body measurements and purchase history are sensitive personal data. The risks are over-collection, leakage or sharing with sellers, and use beyond fit (e.g., profiling).
Bias & fairness	Fit accuracy may be lower for under-represented body shapes, plus sizes, some genders, or regional body proportions (Indian proportions vary widely - File 1). That means unequal recommendation quality across users.

Risk	How it shows up in FitSense
Hallucination	The LLM may invent garment attributes in the reason, or assert a confidence level not grounded in the retrieved data - the size and the explanation can be fabricated.
Safety	Lower-stakes than medical or financial advice, but a confidently-wrong size drives returns and misleads. Free-form input or product text could attempt prompt injection, so inputs must be constrained.
User trust	Opaque or wrong recommendations erode trust quickly. A dishonest confidence display - looking sure when it is not - is actively anti-trust and worse than no feature.
Overdependence	Users may stop checking the size chart and over-trust the AI, amplifying harm exactly when it is wrong - most acute at cold-start where inputs are weakest.

3. Guardrails & Mitigations

Risk	Possible impact	Mitigation / guardrail
Confident-wrong recommendation	Misfit, return, broken trust	Calibrate confidence to historical accuracy; show “High” only when retrieval quality + profile completeness + consistency clear set thresholds; cap at “Medium” for cold-start with self-reported-only inputs.
Construction mis-extraction	Wrong size + misleading reason	Ground the reason strictly in retrieved attributes (RAG); if a construction attribute cannot be retrieved with confidence, omit it from the reason and lower the confidence.
Cross-brand error	Wrong size on a new brand	Validate normalisation against known anchor SKUs; when normalisation confidence is low, fall back to the brand’s own chart and lower confidence.
Hallucinated reason	Misinformation, lost trust	Constrain the reason to a template populated only from retrieved fields; post-generation check that every claim maps to a retrieved attribute; drop any unverifiable claim.
Bias / fairness	Unequal quality across bodies	Evaluate accuracy by body-size band, gender, and category; set a maximum accuracy gap as a launch and kill-switch gate; actively collect under-represented fit data.
Privacy	Leakage, misuse	Opt-in measurements; encryption in transit and at rest; never shared with sellers; user view/delete; purpose limitation; minimal retention.

Risk	Possible impact	Mitigation / guardrail
Overdependence	Amplified harm when wrong	Keep the size chart one tap away; always show confidence + reason so the user co-decides; honest low-confidence states nudge verification.
Latency / availability	Blocked PDP, lost sale	Strict P95 budget + cached garment-level retrieval; on timeout or outage degrade to the static chart; never block PDP render.
Prompt injection / input abuse	Manipulated output	Constrain fit preference to chips (relaxed / true / snug), not free text; sanitise any free text; fix the model's role via system prompt; never follow instructions embedded in product descriptions or user input.

4. Fallback Behaviour

When the AI is…	What FitSense does
Unsure (low confidence)	Shows the honest "Low confidence - here is what would help" state with "Refine my fit." Never presents a single size as certain.
Unavailable (downtime, rate limit, timeout)	Degrades to Myntra's existing static size chart; the PDP never blocks; a silent retry runs in the background.
Returning an unsafe / off-policy reason	Suppresses the reason; shows the size from the profile without the fabricated detail, or falls back to the chart.
Asked something out of scope	Politely declines and redirects; non-fit issues (returns, delivery) go to support. FitSense advises on fit only.

Graceful degradation path. Full AI reason → confidence-only recommendation → static size chart → human support. The user always lands somewhere useful.

5. Human-in-the-Loop Plan

Which cases route to a human

- Monitored accuracy drops for a brand, category, or cohort trigger a catalogue / data-team review.
- Clusters of user-reported "bad recommendation" flags go into a review queue.
- Any flagged or abusive free-text input is routed to moderation.

How humans feed back into the AI

- Merchandising / data ops correct construction attributes and normalisation anchors; these corrections retrain extraction and update the cross-brand reference.

- Labelled kept/returned outcomes refine the confidence calibration over time.

Escalation paths

- A fairness-gap breach pauses further rollout pending human review.
- A confirmed privacy or data incident escalates to privacy and legal.
- Repeated confident-wrong results on one SKU pull that SKU's AI recommendation until fixed.

Core safeguard: FitSense is advisory, not autonomous - the shopper always makes the final size choice, which is itself a human-in-the-loop control.

6. AI Transparency

- **AI label.** The recommendation is clearly marked as an AI suggestion ("FitSense · AI suggestion") with a short "how this works" link.
- **Confidence indicator.** The High / Medium / Low pill is always visible - certainty is the product, not a footnote.
- **Reason.** Every recommendation carries a plain-language reason grounded in retrieved attributes.
- **Flag a bad response.** A "this didn't fit / wrong suggestion" control sits on the recommendation and post-delivery, feeding the review queue and the user's profile.
- **Escalation to a human.** A clear path to support for non-fit issues; the size chart stays one tap away so the user can always override.
- **Data transparency.** A visible note states what data is used (history, measurements), with a link to view or delete it.

7. AI Evaluation Metrics

Metrics that evaluate the AI itself, not just the product. (Product and business metrics - conversion, return rate, adoption - are specified in File 5.)

Metric	What it measures - and why it matters for FitSense
Fit accuracy	Recommended size = size kept. The core quality metric - the whole point of the feature.
Precision	Of sizes shown as high-confidence, how many actually fit. The error to control hardest, because confident-wrong is the costly failure.
Recall	Of sizes that would fit, how many we recommended. Guards against being uselessly cautious.
F1 score	A single balanced read of precision and recall.
Confidence calibration (ECE)	Does "High confidence" actually correspond to high accuracy? This proves the confidence display is honest - arguably the most important metric for this feature.

Metric	What it measures - and why it matters for FitSense
Hallucination rate	Share of reasons containing a claim not grounded in a retrieved attribute. Target near-zero.
Low-confidence / refusal rate	On thin-input or out-of-scope cases, are we correctly declining to over-claim? Healthy = honest; too low = overconfident.
Fairness gap	Maximum accuracy delta across body-size bands, gender, and category. A launch and kill-switch gate.
P95 latency	Slow recommendations get ignored - a UX requirement, not a nicety.
Cost per successful recommendation	Token economics per accepted-and-kept recommendation; unit economics are detailed in File 5.

Carries into File 4 and File 5. The low-confidence fallback and the flag/feedback controls must be visible in the File 4 prototype screenshots; calibration, hallucination rate, and the fairness gap anchor the evaluation and kill-switch criteria in File 5.