AI REIMAGINE PROJECT · FILE 05

Metrics, Experimentation & GTM

FitSense - measuring the bet and rolling it out

Prepared by Abhishek Katkar · June 2026

The bet, stated plainly. FitSense reduces size-driven returns and lifts confident conversion - concentrated on the cold-start and cross-brand moments the existing purchase-history recommendation cannot reach - while keeping its own confidence honest. So the measurement plan does two things at once: it proves the business outcome on the right cohort, and it proves the AI is calibrated (its confidence means what it says). We do not measure FitSense head-to-head against the history signal for repeat buyers; that is not where it wins.

1. Metrics Framework

North Star: size-related return rate for FitSense-exposed orders. Everything else is a supporting or diagnostic metric.

Tier	Metric	What it measures - and why	Goal
North Star	Size-related return rate (exposed orders)	Share of exposed-cohort orders returned for fit/size. The core problem from File 1 - the whole point of the feature.	Down
Product	Recommendation acceptance rate	Share of exposed users who add the suggested size to bag - direct read on trust and usefulness.	Up
Product	Refine engagement rate	Share who open/complete “Refine my fit” - shows the hybrid valve is used where confidence is low.	Healthy
Product	Add-to-bag conversion	PDP -> bag for exposed vs control - does an explained, confident recommendation move purchase intent.	Up / flat
Business	Reverse-logistics cost saved	Returns avoided x cost per return - the financial case for the feature.	Up
Business	Revenue per exposed session	Net revenue effect after returns - ties the feature to money.	Up
AI / model	Fit accuracy	Recommended size = size kept. The quality of the recommendation itself.	Up
AI / model	Confidence calibration (ECE)	Does “High” actually mean high accuracy - the proof the confidence display is honest.	Low ECE
AI / model	Hallucination rate	Reasons containing a claim not grounded in retrieved data (File 3).	Near 0
AI / model	Fairness gap	Max accuracy delta across body-size band, gender, and category - a launch gate.	Below cap

2. Guardrail Metrics

These must not degrade, whatever the headline numbers do. A breach blocks ship or triggers rollback.

Guardrail	Must stay…
Overall PDP -> purchase conversion	No worse than control - FitSense must not suppress purchases by over-warning.
Exchange rate	No spike - a recommendation that drives wrong-size exchanges is a hidden failure.
PDP latency (P95)	≤ 1.5s to render; never block PDP load (degrade to chart on timeout).
Bad-suggestion flag rate	Below threshold - a rising “flag” rate is an early trust-erosion signal.
Confidence calibration (ECE)	Below threshold - if High-confidence accuracy slips, the display is dishonest and the feature is pulled.
Fairness gap	Within the cap across cohorts - no group is silently served worse.

3. Experiment Design (A/B)

Element	Detail
Hypothesis	For cold-start / cross-brand shoppers, showing FitSense (explained, confidence-scored) reduces the size-related return rate and lifts add-to-bag, versus the current PDP - without hurting overall conversion.
Unit & split	User-level randomisation, 50/50. Control = current PDP (size chart + purchase-history recommendation). Treatment = FitSense.
Primary metric	Size-related return rate, measured on the pre-registered cold-start / cross-brand cohort, after the return window closes.
Secondary	Recommendation acceptance, add-to-bag conversion, refine engagement, and (treatment only) confidence calibration.
Cohort discipline	The cold-start / cross-brand segment is the primary analysis unit - that is where FitSense structurally wins. The full population is tracked for guardrails only.
Powering	Two-proportion test. Illustratively, a baseline fit-return rate of ~15% and a 10-15% relative reduction needs roughly 4,000-8,500 exposed orders per arm (80% power, 5% significance). Because returns lag, the test runs 6-8 weeks plus the return window.
Decision rule	Ship if the primary metric improves with significance AND no guardrail is breached. Otherwise iterate (prompt, data coverage, UX) or hold.

Why not a global head-to-head? Measuring FitSense against the history signal across all users would dilute the effect and likely lose - the history signal is strong for repeat buyers. Scoping the primary metric to the cold-start / cross-brand cohort measures the feature where its advantage is real and defensible.

4. Kill-Switch & Rollback

- **Calibration breach.** If High-confidence accuracy falls below its floor, pull the AI recommendation and degrade to the static size chart - an over-confident model is worse than none.
- **Fairness-gap breach.** If accuracy diverges across cohorts beyond the cap, pause rollout for review.
- **Latency breach.** P95 over budget -> degrade to chart; never block the PDP.
- **Outcome regression.** If the exposed cohort's return rate worsens, roll back the treatment.
- **SKU-level pull.** A cluster of confident-wrong results on one SKU pulls that SKU's recommendation until its data is fixed (File 3).

5. Go-to-Market & Phased Rollout

Phase	What happens	Gate to advance
P0 - Offline eval	Backtest accuracy and calibration on historical orders (predicted size vs size kept).	Accuracy and ECE thresholds met.
P1 - Shadow	Run FitSense in production silently (no user exposure); log predictions vs actual kept/returned.	Live calibration holds; fairness gap within cap.
P2 - Limited A/B	One high-return apparel category, cold-start / cross-brand cohort, ~5-10% of traffic.	Primary metric improves; no guardrail breach.
P3 - Expansion	Ramp by category and brand as construction-data coverage grows; widen traffic.	Sustained lift; data coverage sufficient.
P4 - GA	All apparel; continuous monitoring and the learning loop (kept/returned -> profile + calibration).	-

Positioning & channels

- **Shopper story:** "the right size, explained" - a trust feature at the point of decision, not a gimmick.
- **Business story:** a returns-reduction lever with a measured cost-saved number, concentrated where returns are worst.

- **Channels:** on-PDP as the primary surface; an opt-in measurement capture during onboarding to seed cold-start profiles; and a post-delivery feedback loop that feeds the learning loop.

6. Carries Into File 6

File 6 (Panel Defence) defends the choices behind this plan: why the cold-start cohort is the right unit of measurement, why calibration is the metric that proves the feature is trustworthy, and how the kill-switches keep an honest-but-fallible model safe in production.