

Problem Statement and Objective

Dick’s Sporting Goods (DSG) has an online catalog of 2.4M+ SKUs, but key product attributes are **only ~40% complete**. This limits search and filter performance, product discoverability, and business analytics.

Our goal is to leverage a **Retrieval-Augmented Generation (RAG)–based LLM** pipeline to:



Impute missing product attributes



Improve recommendation quality and user engagement

Data

ID	CATEGORY
#1023	T-shirt
#1056	Shoe
#1091	Sneaker

Product Catalog

Structured data including product ID, category, and known attribute values



Product Description

Short text containing cues like fit, material, brand, and use-case

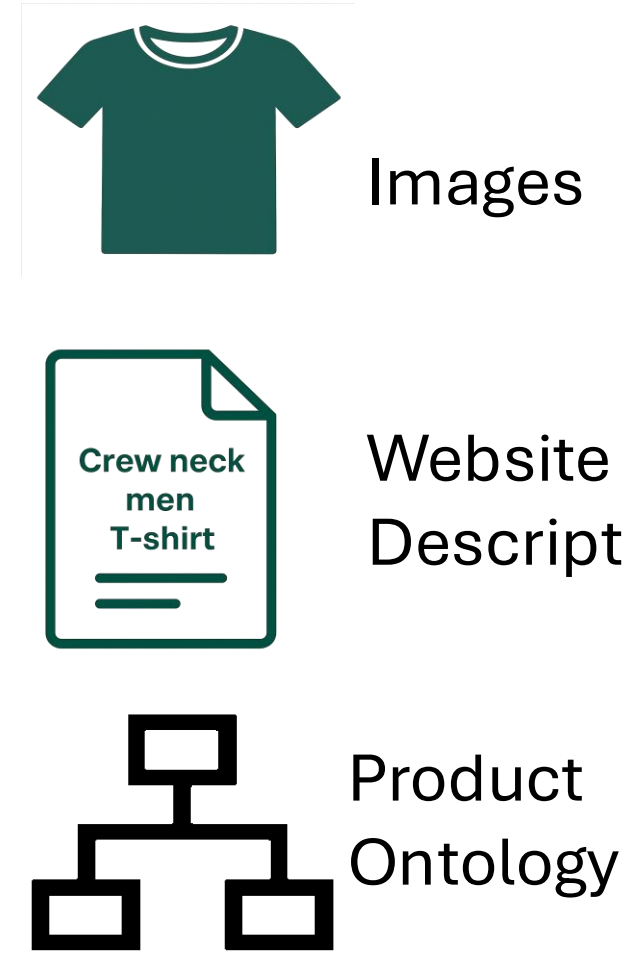


Product Image

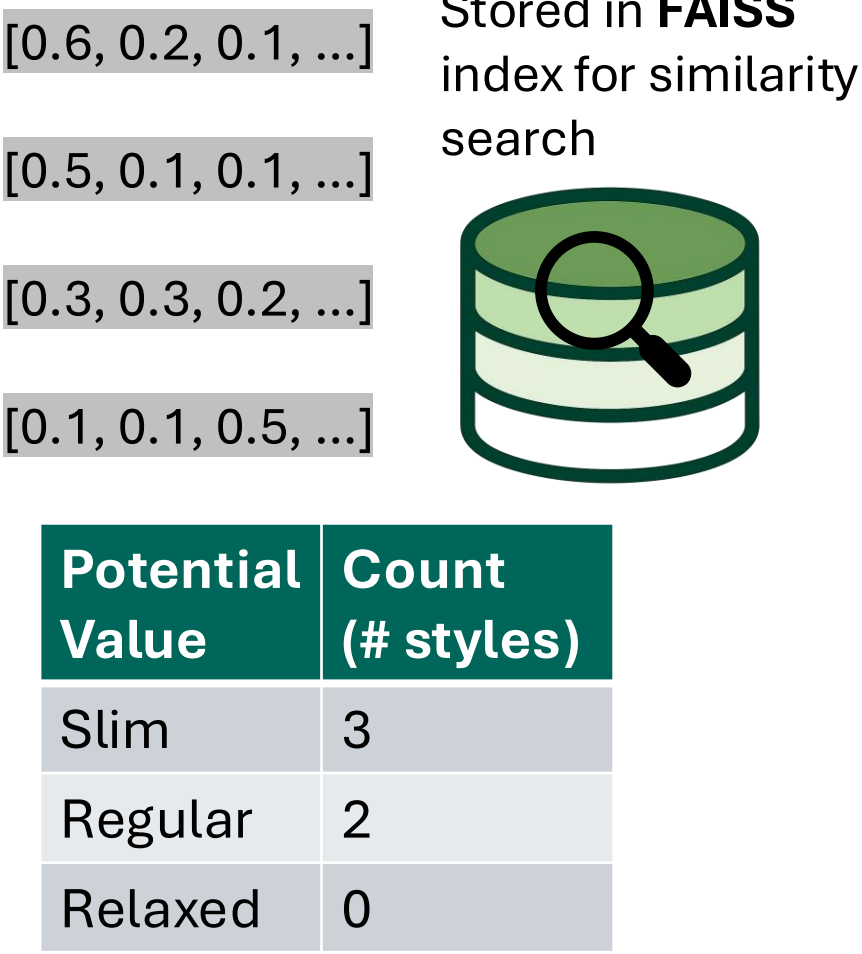
Visual input used to infer style-related details such as sleeve length, neckline, and cut

Methodology

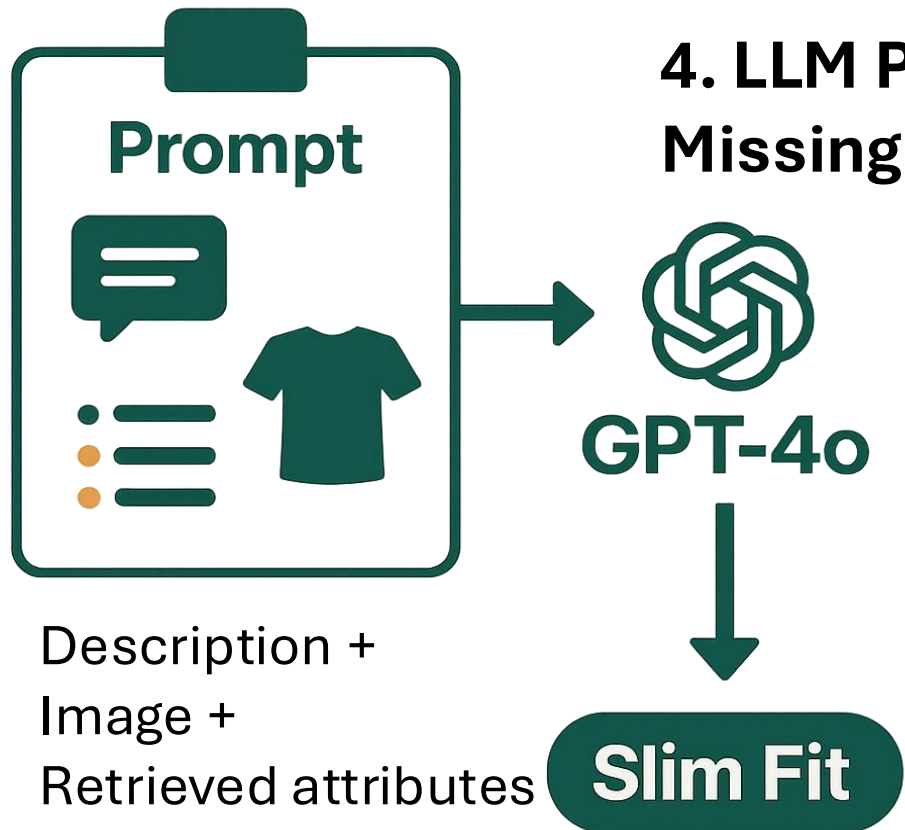
1.Gather Inputs



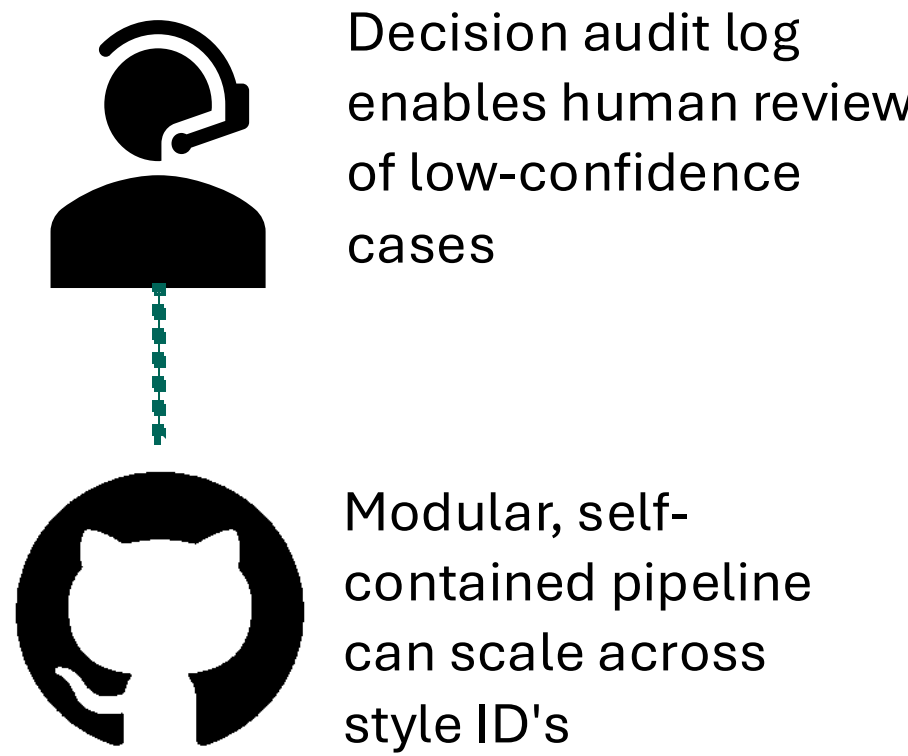
2. Retrieve neighbors



3. Build prompt



5. Ensure write-back and QA pipeline



Retrieval Pipeline

We retrieve similar styles using text and image embeddings to provide relevant context for LLM prediction. This improves prediction accuracy and reduces hallucination.

$\begin{bmatrix} 0.1 \\ -2.3 \\ 4.5 \\ 1.2 \\ 0.7 \end{bmatrix}$

Embed product text and image using CLIP (Contrastive Language–Image Pretraining) into a shared vector space.

$\begin{bmatrix} 0.2 \\ 1.5 \\ -0.3 \\ 1.2 \\ 0.7 \end{bmatrix}$

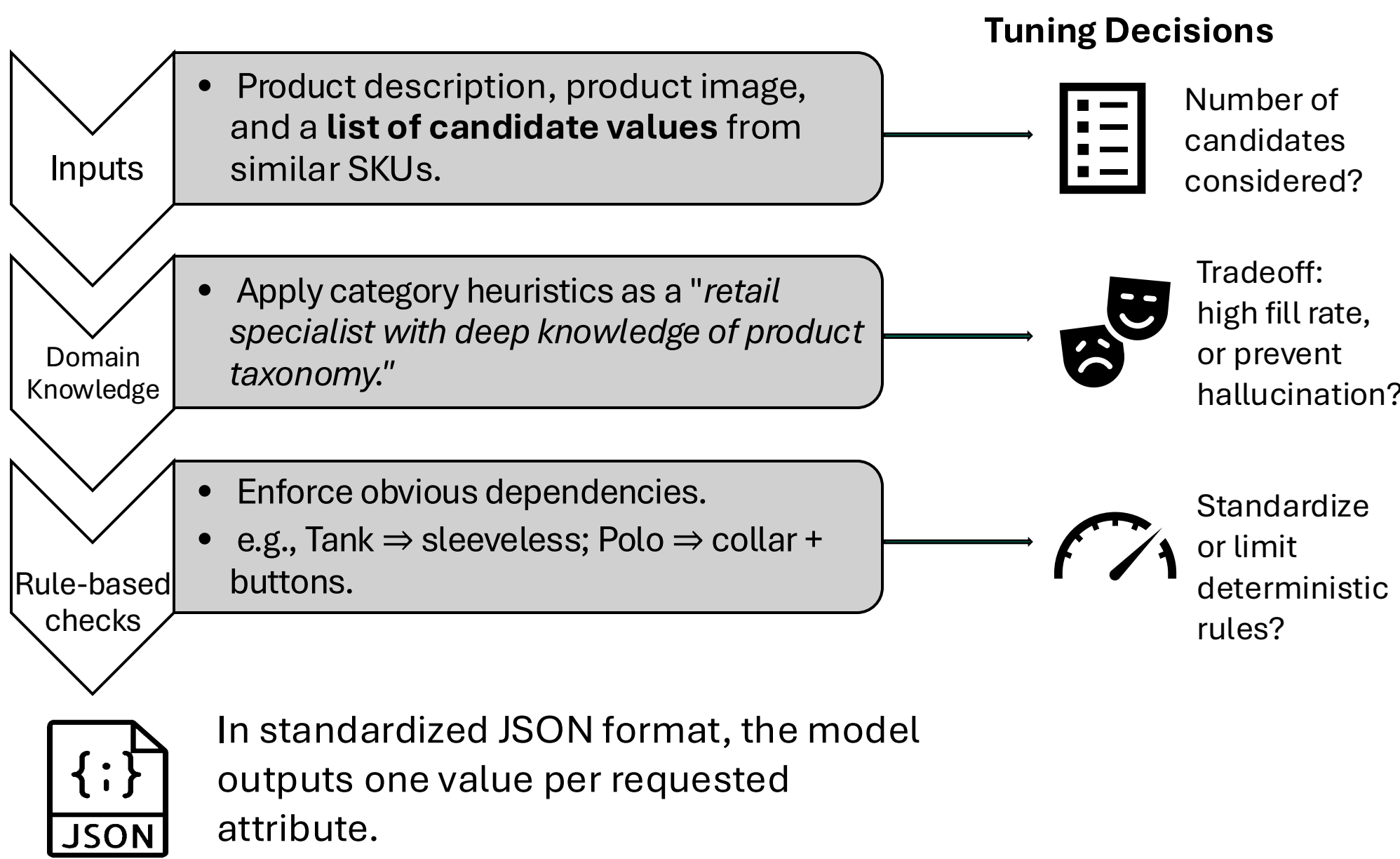
Use FAISS (Facebook AI Similarity Search) to index combined embeddings for fast, scalable similarity retrieval.



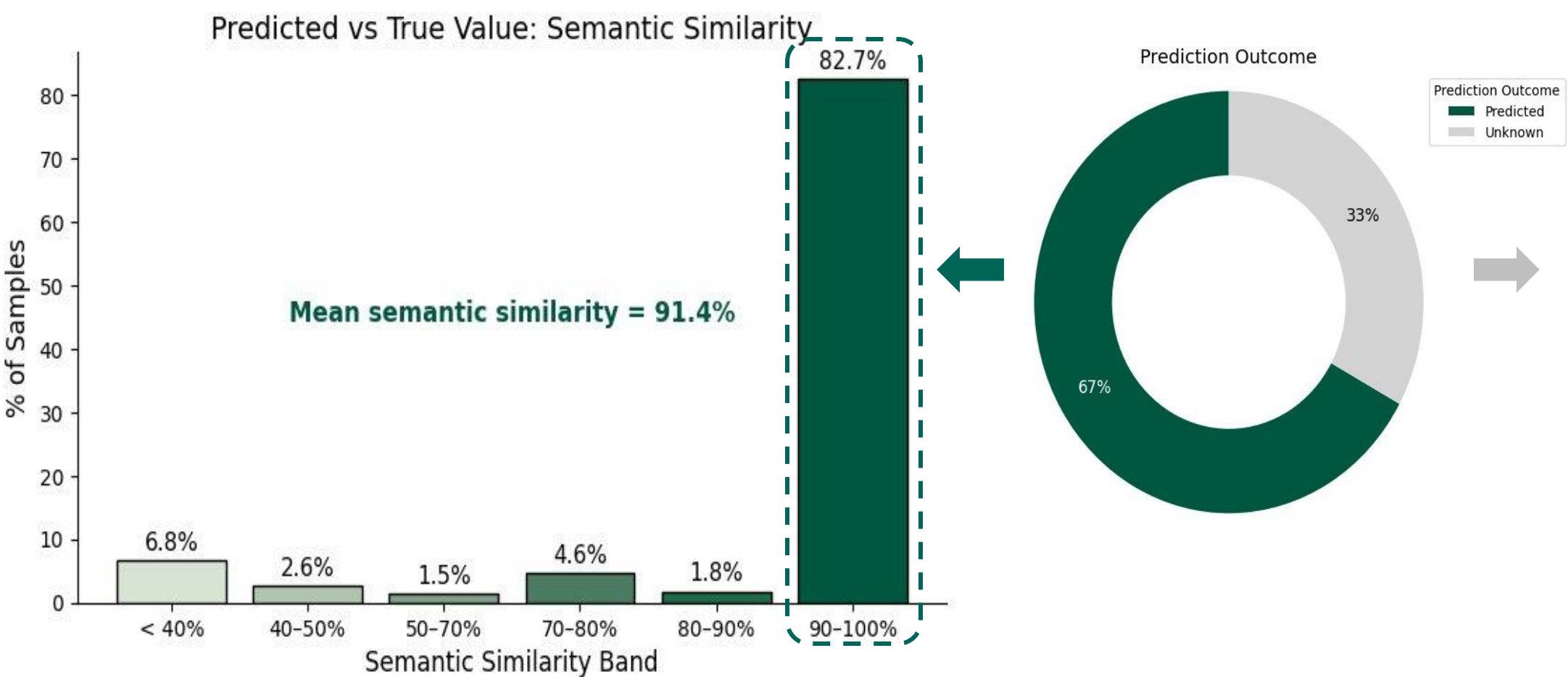
For each product with missing attributes, retrieve the top-k most similar styles with known attribute values.

LLM Inference and Guardrails

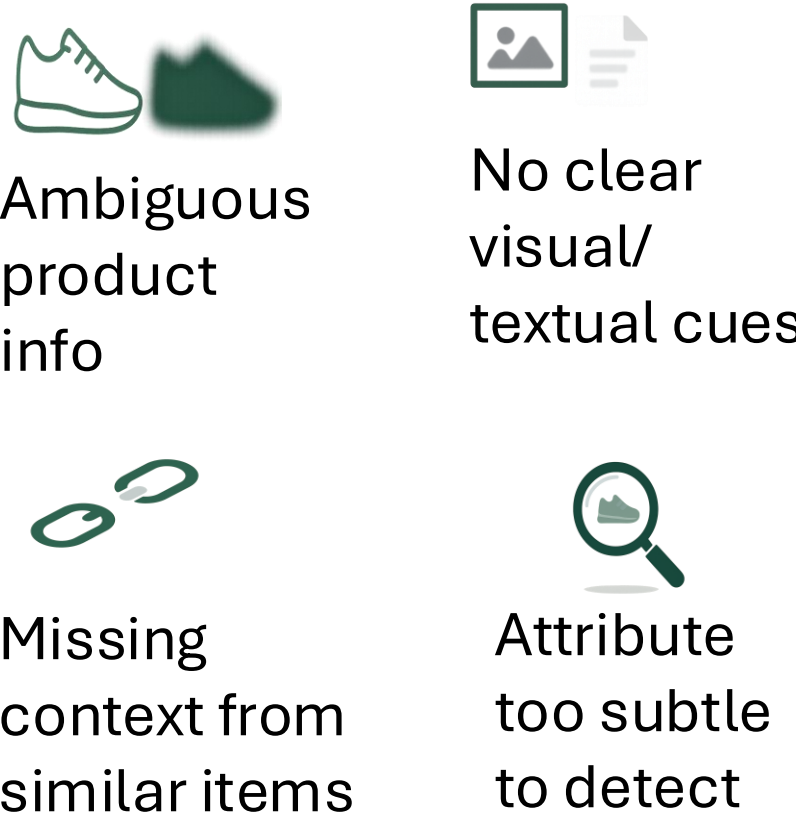
A large language model infers the most likely value per missing attribute.



Result and Impact



Why the Model Predicted “Unknown”



Estimated annual incremental revenue

\$3.0M

S × ΔRel@48 × γ × CVR × \$100 AOV

S: ~99M searches

ΔRel@48: +0.10

γ: 0.15 clicks/rel

CVR: 2%

Extra clicks: 1.5M $\rightarrow$ ~30k orders

Cost ≈ \$5k • ROI ≈ 600×

S derivation: 12-14M daily impressions ÷ 48 results x 365 ≈ 91-106M. Assumption-based; to be validated via A/B. Figures illustrative.