



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	Shorts
#1091	Sneakers

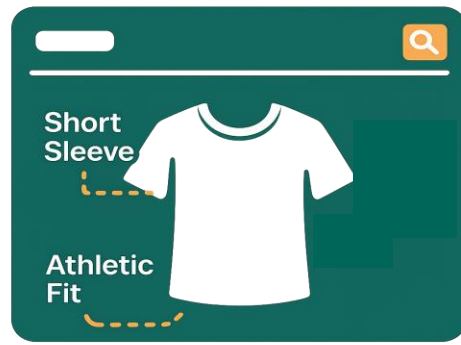
Product Catalog

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



Product Description

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

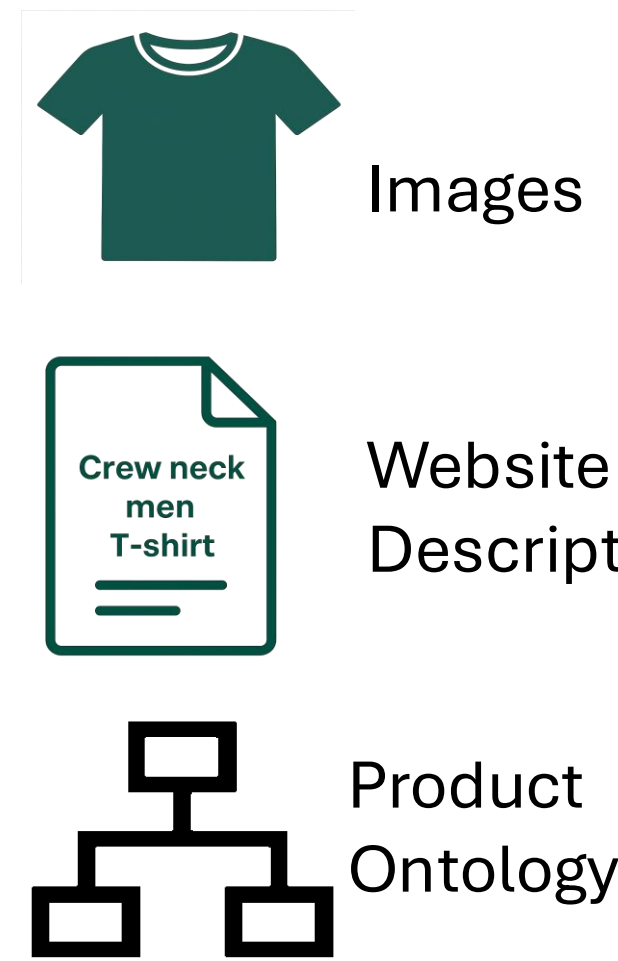


Product Image

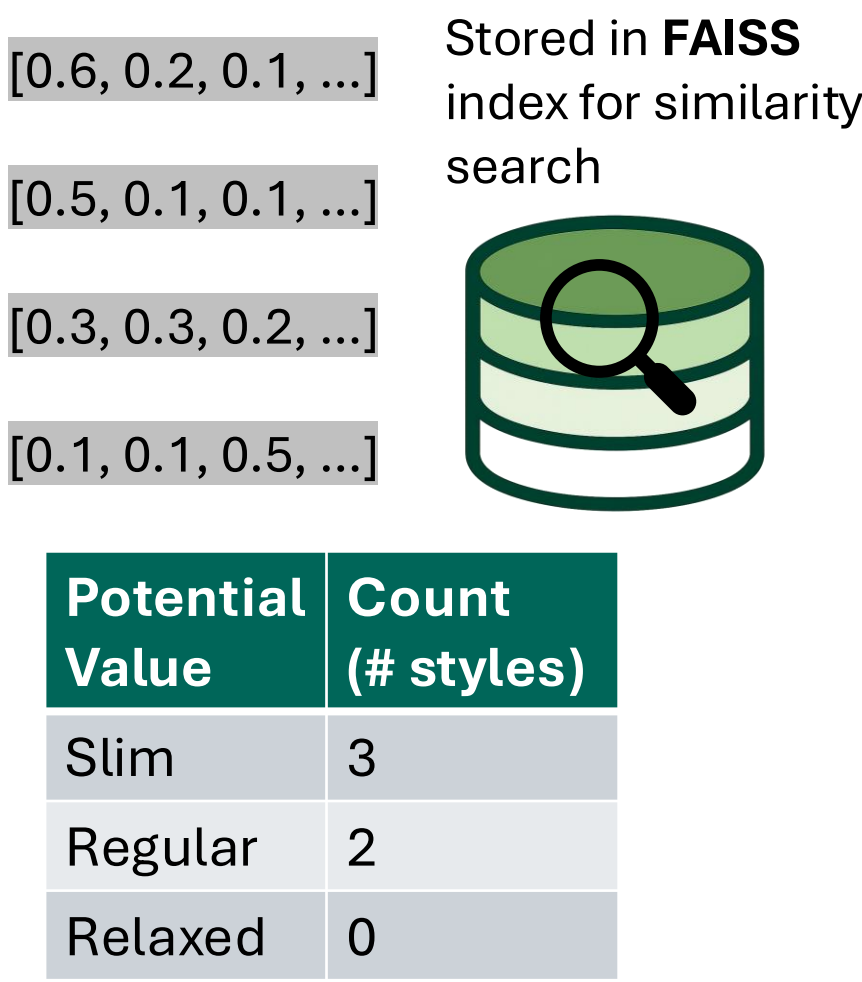
Visual input used to infer style-related details such as sleeve length, neckline, and cut

Methodology

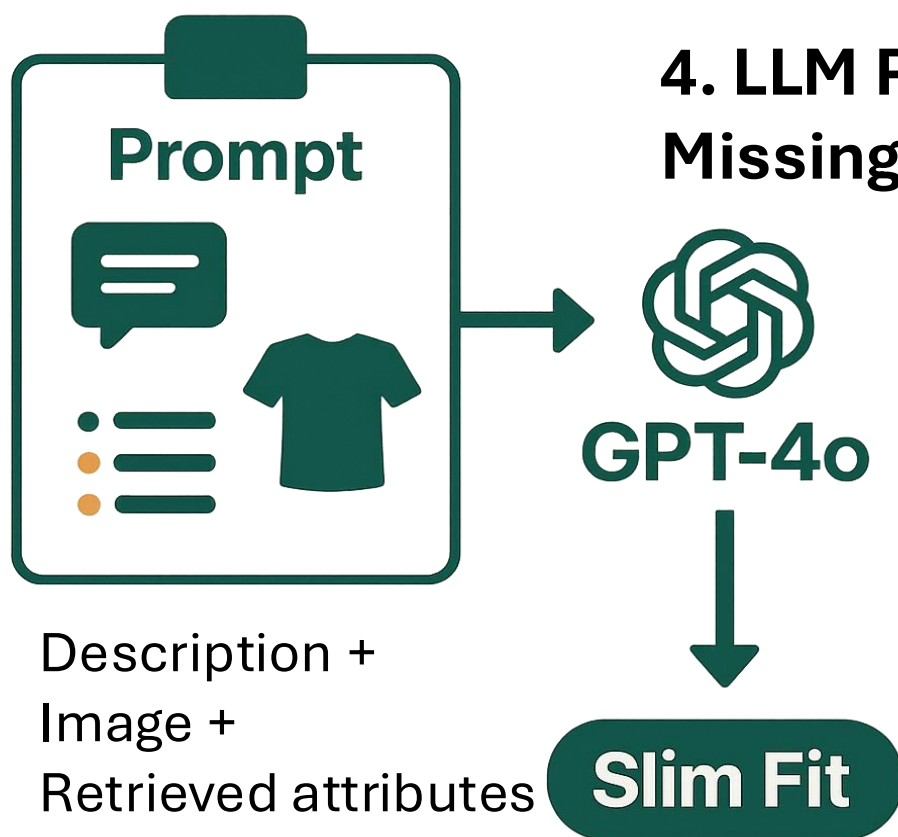
1.Gather Inputs



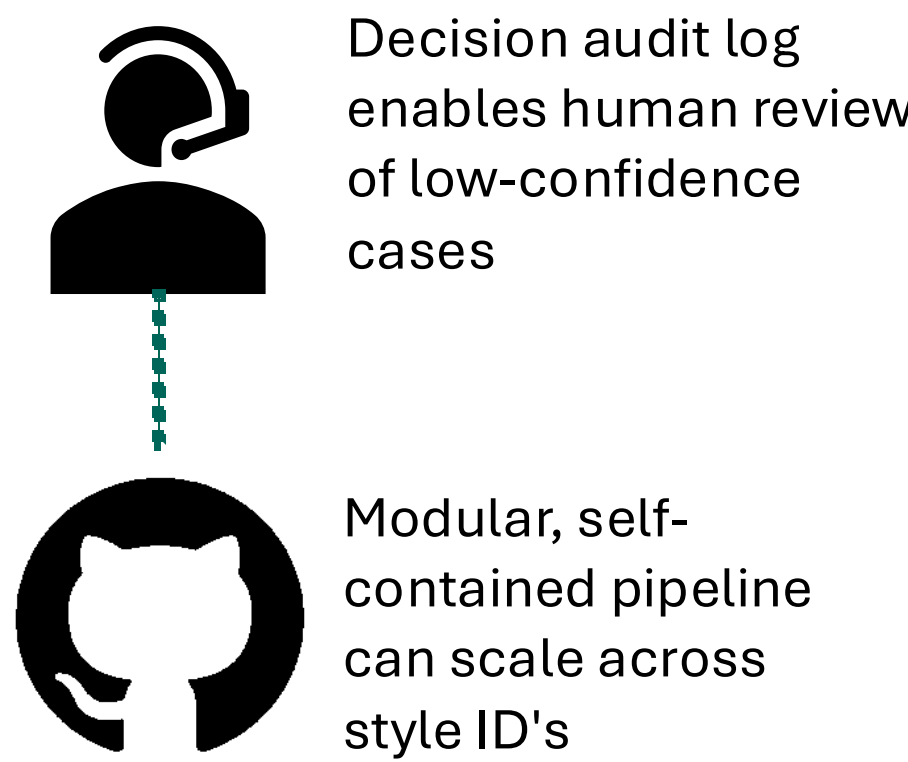
2. Retrieve neighbors



3. Build prompt

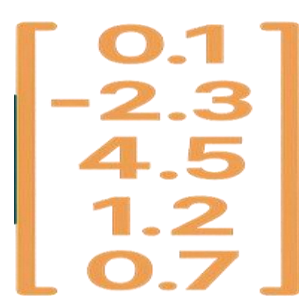


5. Ensure write-back and QA pipeline

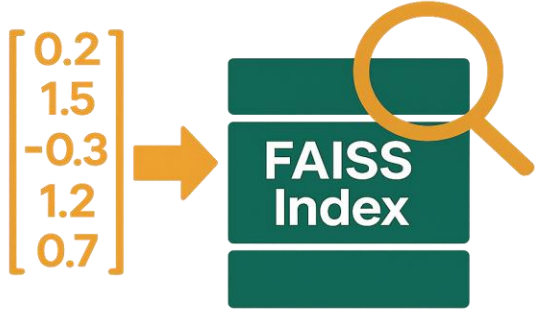


Retrieval Pipeline

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



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



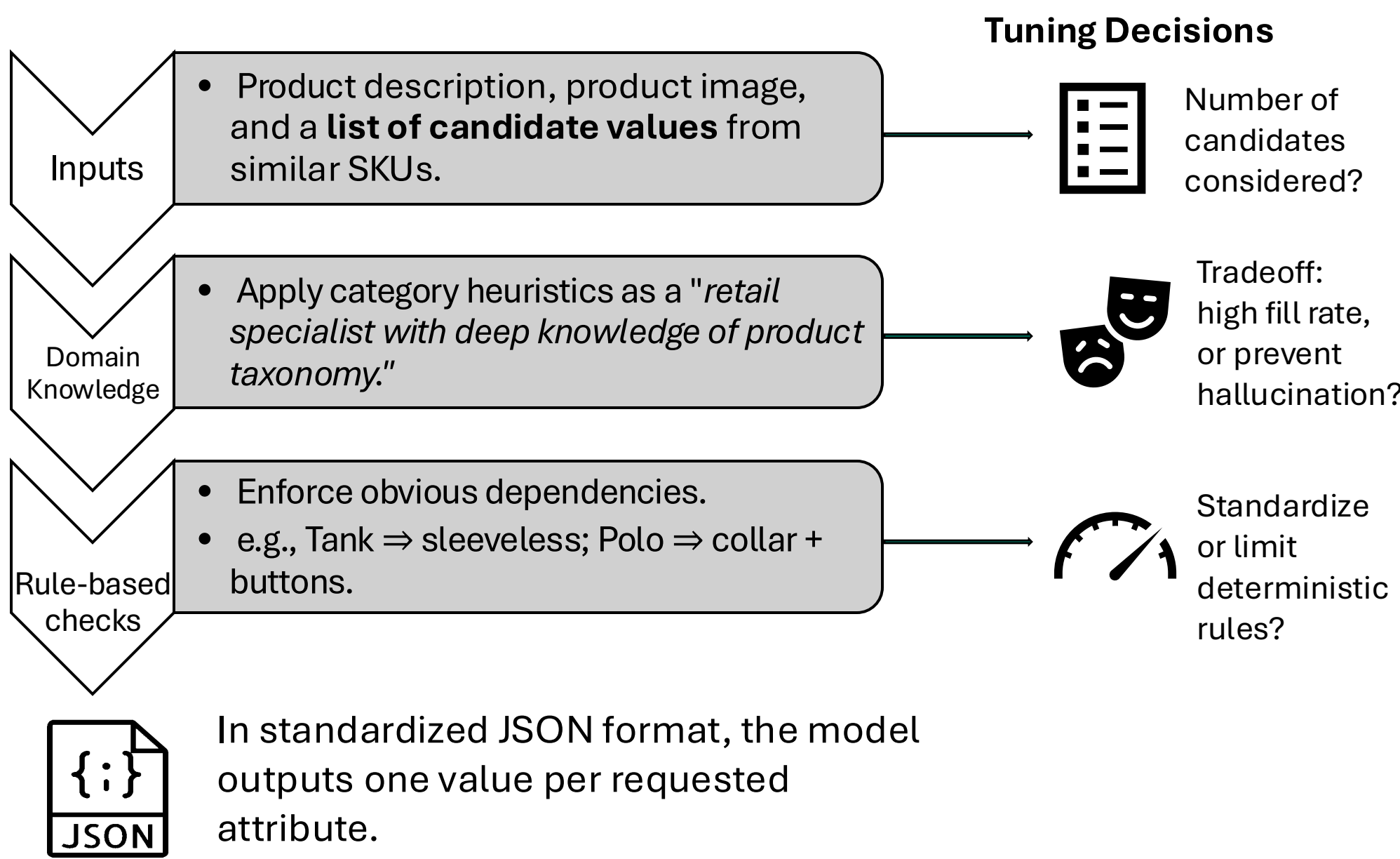
Use FAISS (Facebook AI Similarity Search) to index combined embeddings for fast, scalable similarity retrieval.



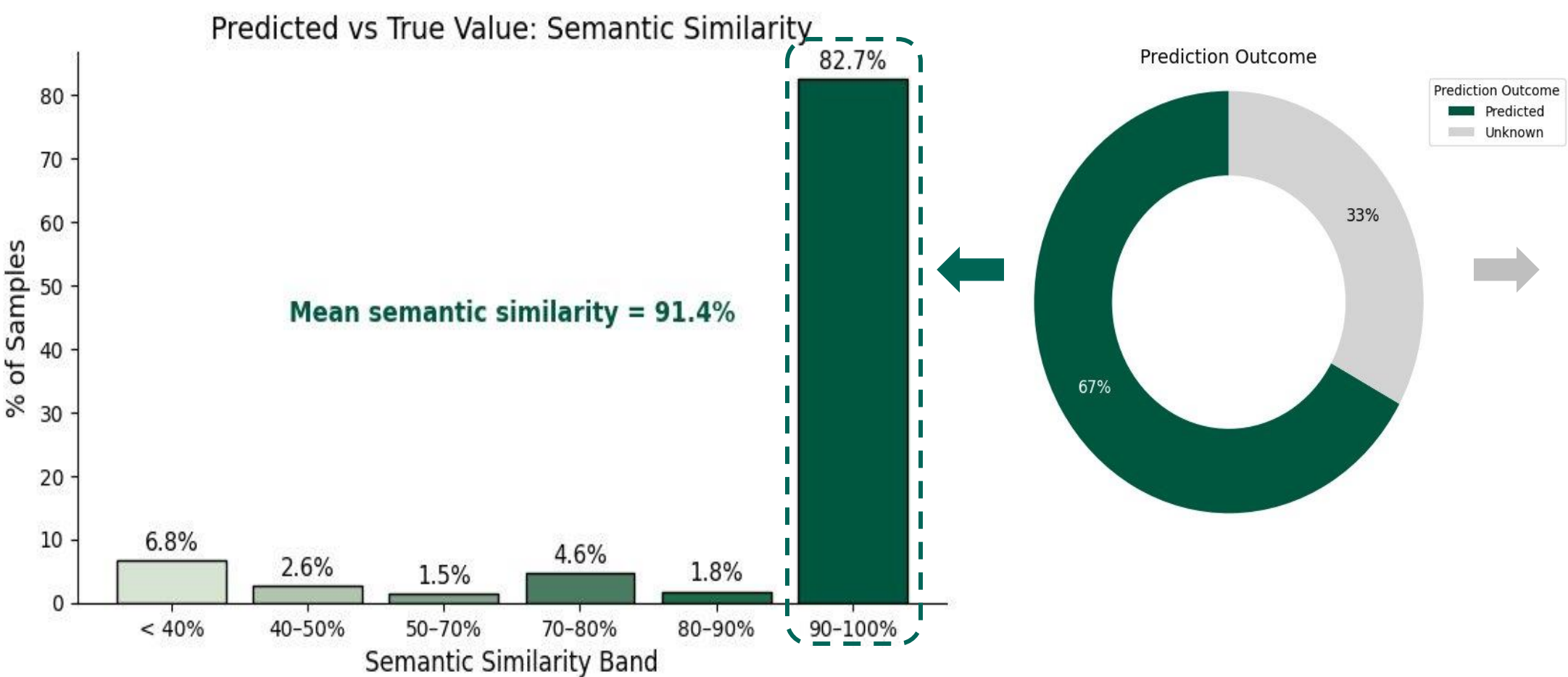
For each product with missing attributes, retrieve the top-k most similar styles with known attribute values.

LLM Inference and Guardrails

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



Result and Impact



Why the Model Predicted "Unknown"

- Ambiguous product info
- No clear visual/textual cues
- Missing context from similar items
- Attribute too subtle to detect

Estimated annual incremental revenue

\$3.0M

$S \times \Delta Rel@48 \times \gamma \times CVR \times \100 AOV

S: ~99M searches

$\Delta Rel@48$: +0.10

γ : 0.15 clicks/rel

CVR: 2%

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

Cost $\approx$ \$5k • ROI $\approx$ 600 $\times$

S derivation: 12-14M daily impressions $\div$ 48 results $\times$ 365 $\approx$ 91-106M. Assumption-based; to be validated via A/B. Figures illustrative.