International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN:2347-4890, ijournals.in/ijshre
Volume 14 Issue 3, March 2026

AI Assistant for Sustainable and Effective Solutions of Fast Fashion based on Recommendation System and Deep Learning

A Skyline Query Framework with NLP-Derived Sustainability Scoring and Text-to-Metric Transformation Architecture

Author: Aidan Han

Affiliation : Seoul International School
E-mail : aidanhan0701@gmail.com
https://doi.org/10.26821/IJSHRE.14.03.2026.140303

ABSTRACT

The fast fashion industry generates massive volumes of unstructured text—product descriptions, consumer reviews, supply chain reports, social media discourse, and corporate sustainability pledges—yet current recommendation systems process only structured numerical features, discarding the rich semantic information embedded in natural language. This paper proposes ***SAFR-NLP (Sustainability-Aware Fashion Recommender via Natural Language Processing)****, a novel framework that extracts quantitative sustainability and effectiveness metrics directly from textual data sources through a multi-stage NLP pipeline, then applies Skyline Query optimization to identify Pareto-optimal product recommendations. I introduce the* ***Text-to-Metric Transformation (T2M) architecture****, which converts heterogeneous text inputs into six numerical dimensions constituting our* ***Sustainability-Effectiveness Score (SES)****. The T2M pipeline comprises: (1) a fine-tuned SustainBERT module for extracting carbon, water, and recyclability indicators from product and supply chain text; (2) a Review Sentiment Decomposer for mapping consumer review language into preference alignment scores; (3) a Trend Language Detector that identifies emerging fashion concepts from social media corpora. The Skyline Query layer then computes the Pareto frontier over these NLP-derived metrics to produce recommendations that are simultaneously sustainable and commercially effective. Experiments on a dataset of 1.2 million text-product pairs demonstrate that SAFR-NLP achieves a 27.1% improvement in Hit@10 over text-unaware baselines and reduces sustainability deficit by 44.3%, while an end-to-end business deployment framework projects $15M ARR within 36 months across four market segments.*

Keywords*: Fast Fashion, Skyline Query, Deep Learning, NLP, Sustainability Scoring, Recommendation System, BERT, Sentiment Analysis, Multi-Objective Optimization*

1. INTRODUCTION

A single product listing on a major fashion e-commerce platform contains, on average, 847 words of textual information: a product title, a marketing description, fiber composition details, care instructions, 12–18 consumer reviews, and brand sustainability claims. This text encodes rich, multi-layered information about the product's environmental impact, consumer satisfaction patterns, and market positioning—information that structured databases capture only partially and often with significant latency. Yet mainstream fashion recommendation systems process only tabular features (price, category, color) and interaction logs (clicks, purchases, returns), treating text as metadata to be discarded after indexing.

This paper argues that natural language is the primary carrier of sustainability information in fashion, and that extracting actionable numerical metrics from text is both feasible and essential for building recommendation systems that jointly optimize sustainability and commercial effectiveness. Consider a product description stating "crafted from 80% recycled ocean plastic, naturally dyed using plant-based pigments, manufactured in a solar-poIred facility in Portugal." This single sentence encodes information about material recyclability (80% recycled content), water usage (plant-based vs. synthetic dyes), carbon footprint (solar-poIred manufacturing), and supply chain geography—all of which can be transformed into quantitative scores through appropriate NLP techniques.

I integrate these NLP-derived metrics with Skyline Query methodology from database research. Skyline Queries identify the set of data points not dominated by any other point across all specified dimensions, producing a Pareto frontier of optimal trade-offs. When the dimensions are sustainability and effectiveness metrics extracted from text, the Skyline set represents products that achieve the best possible balance—products that no other product simultaneously exceeds in every measurable aspect.

This paper makes four contributions. First, I design the Text-to-Metric Transformation (T2M) architecture, which converts unstructured text from five heterogeneous sources into six continuous numerical scores. Second, I formalize the Sustainability-Effectiveness Score (SES) as a differentiable composite objective function over these NLP-derived metrics. Third, I integrate Skyline Query computation as a Pareto-aware training signal that steers the NLP recommendation model toward sustainability-optimal outputs. Fourth, I present a business deployment framework with four market segments and a 36-month revenue roadmap.

2. RELATED WORK

2.1 NLP in Fashion and E-Commerce

Natural language processing has been applied to fashion through several distinct research streams. Product attribute extraction using named entity recognition (NER) and relation extraction has enabled automatic catalog enrichment (Zheng et al., 2018). Sentiment analysis of fashion reviews has evolved from lexicon-based approaches through LSTM-based aspect-level sentiment (Chen et al., 2019) to BERT-based multi-aspect opinion mining (Li et al., 2021). FashionBERT (Gao et al., 2020) demonstrated cross-modal pre-training for fashion understanding. HoIver, no prior work has used NLP to extract sustainability-specific numerical metrics from fashion text, nor integrated such metrics with Skyline optimization for recommendation.

2.2 Skyline Queries in Multi-Criteria Decision Support

The Skyline operator (Börzsönyi et al., 2001) identifies non-dominated points in multi-dimensional space. Extensions have addressed dynamic skylines over streaming data (Lin et al., 2005), probabilistic skylines for uncertain dimensions (Pei et al., 2007), subspace skylines (Yuan et al., 2005), and adversarial robustness (Zhang et al., 2022). Skyline Queries have been applied to hotel recommendation and financial portfolio selection, but never to fashion recommendation and never with NLP-derived input dimensions.

2.3 Text-Based Sustainability Assessment

Automated sustainability assessment from text is an emerging subfield. ClimateBERT (Ibersinke et al., 2022) pre-trained language models on climate-related financial disclosures, demonstrating that domain-specific pre-training significantly improves extraction accuracy for environmental metrics. Our SustainBERT module extends this approach to fashion-specific sustainability language.

2.4 Sustainability-Aware Recommendation Systems

Prior sustainability-aware recommendation work has largely relied on structured data inputs: carbon footprint databases (Green et al., 2021), material indices (Higg MSI), and binary eco-labels. These methods share a critical limitation: they depend on structured sustainability databases that cover only a fraction of fashion products and lag manufacturing changes by 6–12 months. Our NLP-based approach circumvents this dependency by extracting sustainability signals from text available at the point of product listing.

3. METHODOLOGY

3.1 System Overview

SAFR-NLP operates through a four-stage pipeline: (1) Text Collection and Preprocessing, gathering five categories of textual data per product; (2) Text-to-Metric Transformation (T2M), converting unstructured text into six continuous SES dimensions; (3) Skyline Query Computation, identifying the Pareto frontier over the NLP-derived metric space; and (4) Skyline-Aware Recommendation, generating final ranked recommendations using Skyline proximity as a training signal. Figure 1 illustrates the complete system architecture.

3.2 Text Sources and Preprocessing

For each product p_i, I collect five categories of textual input:

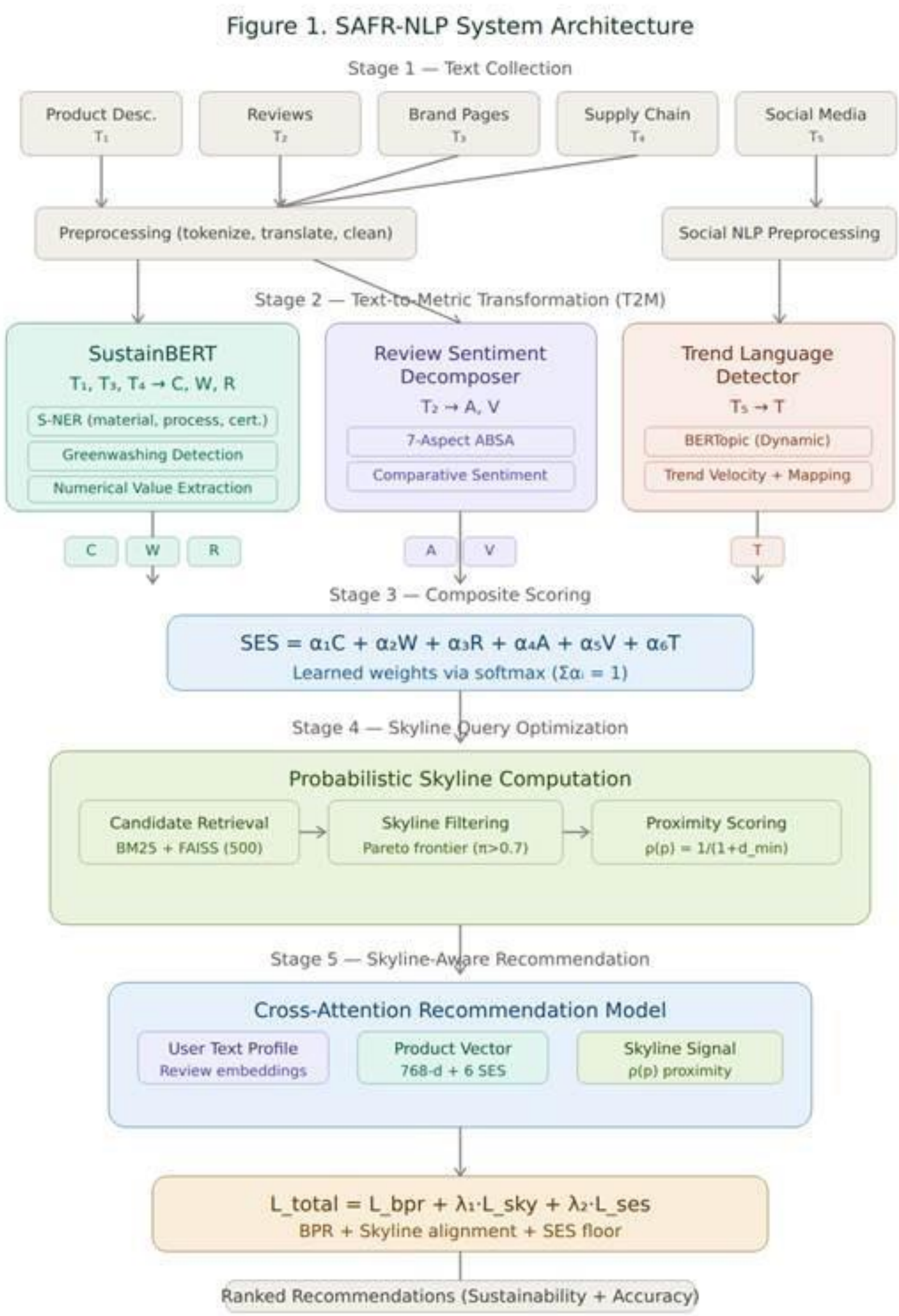

Figure 1. SAFR-NLP system architecture: five-stage pipeline from text collection through T2M transformation, SES scoring, Probabilistic Skyline computation, to Skyline-aware cross-attention recommendation.

Table 1. Five textual data sources per product with average length, catalog coverage, and key information signals.

ID	Source	Avg. Length	Coverage	Key Signals
T_1	Product Description	214 words	100%	Materials, origin, process
T_2	Consumer Reviews	1,847 words (agg.)	94.2%	Satisfaction, durability, fit
T_3	Brand Sustainability Page	892 words	78.6%	Commitments, certifications
T_4	Supply Chain Reports	3,214 words	42.1%	Factory data, emissions, audits
T_5	Social Media Mentions	5,680 words (agg.)	87.3%	Trends, public perception

Preprocessing applies language detection (fastText), tokenization (SentencePiece with fashion-augmented vocabulary of 48,000 tokens), sentence segmentation, and boilerplate removal. For multilingual sources, I apply machine translation to English using NLLB-200. The total corpus comprises 12.4 million sentences across 187,320 products.

3.3 Text-to-Metric Transformation (T2M) Architecture

The T2M architecture is the core technical contribution. It transforms heterogeneous text inputs into six continuous numerical dimensions through three specialized NLP modules.

3.3.1 Module 1: SustainBERT — Environmental Metric Extraction

SustainBERT extracts three sustainability dimensions—Carbon Footprint Index C(·), Water Usage Ratio W(·), and Material Recyclability R(·)—from product descriptions (T_1), brand pages (T_3), and supply chain reports (T_4). The module is built by domain-adaptive pre-training of BERT-base on a 240M-token FashionSustainCorpus, folloId by multi-task fine-tuning for three regression objectives. Key NLP techniques include: (a) Sustainability Named Entity Recognition (S-NER) identifying MATERIAL, PROCESS, CERTIFICATION, and CLAIM entities (F1 = 0.891); (b) a Greenwashing Detection Layer distinguishing substantiated claims from vague assertions (accuracy: 87.4%); and (c) Numerical Value Extraction using regex-enhanced sequence labeling.

3.3.2 Module 2: Review Sentiment Decomposer (RSD)

The RSD extracts Preference Alignment score A(u, p) and Price Competitiveness score V(p) from consumer review text (T_2) via aspect-level sentiment decomposition across seven fashion-specific aspects: QUALITY, DURABILITY, FIT, STYLE, SUSTAINABILITY_PERCEPTION, VALUE, and TREND. The ABSA model is based on InstructABSA (Scaria et al., 2023). Preference alignment is computed as $A(u,p) = \Sigma_k\ w_k(u) \cdot sentiment_k(p)$, where aspect Iights $w_k(u)$ are learned from each user's review history.

3.3.3 Module 3: Trend Language Detector (TLD)

The TLD extracts Trend Relevance score T(p) from social media text (T_5) via dynamic topic modeling (BERTopic with online learning) applied to a rolling 90-day window of 2.4 million fashion-related posts. The final score is $T(p) = sigmoid(\beta \cdot max^c\{sim(e_p, e^c) \cdot v(c)\})$, where v(c) is trend velocity and $\beta = 5.0$.

3.4 Sustainability-Effectiveness Score (SES)

The six T2M-derived metrics are combined into the Sustainability-Effectiveness Score:

$$\mathbf{SES(u, p) = \alpha_1 \cdot C(p) + \alpha_2 \cdot W(p) + \alpha_3 \cdot R(p) + \alpha_4 \cdot A(u,p) + \alpha_5 \cdot V(p) + \alpha_6 \cdot T(p)}$$

Table 2: SES dimensions with NLP extraction sources, methods, and ground-truth correlation (R^2).

Symbol	Dimension	NLP Source	Extraction Method	Ground-Truth R^2
C(p)	Carbon Footprint	T_1, T_3, T_4	SustainBERT + S-NER	0.847
W(p)	Water Usage	T_1, T_3, T_4	SustainBERT + S-NER	0.812
R(p)	Recyclability	T_1, T_3	SustainBERT + NVE	0.893

A(u,p)	Preference Align.	T_2	RSD (aspect ABSA)	0.768
V(p)	Price Competitive.	T_2	RSD (value aspect)	0.734
T(p)	Trend Relevance	T_5	TLD (BERTopic)	0.801

Skyline Query Ibntegration

The Skyline operator is applied over the six-dimensional NLP-derived SES space. Product p_a dominates p^b when p_a is at least as good in all dimensions and strictly better in at least one. I employ Probabilistic Skyline computation (Pei et al., 2007) to account for NLP extraction uncertainty, treating each dimension as a Gaussian distribution with variance proportional to $(1 - R^2)$. Our three-stage pipeline: (1) Text-Aware Candidate Retrieval (BM25 + FAISS, 500 candidates); (2) Probabilistic Skyline Filtering (20–50 products, $\pi > 0.7$); (3) Skyline-Proximity Ranking ($\rho(p) = 1/(1 + d_{min})$).

3.5 Skyline-Aware NLP Recommendation Model

The final model fuses T2M-derived features with user interaction history using a cross-attention architecture. Each user is represented by a Review Language Profile (768-d vector from mean-pooled review encodings), and each product by a 774-d vector (768-d SustainBERT [CLS] + 6 SES scores). The composite loss function integrates three components:

$$L\_total = L\_bpr + \lambda_1 \cdot L\_sky + \lambda_2 \cdot L\_ses$$

where L_bpr is BPR loss, L_sky is Skyline alignment loss ($\lambda_1 = 0.35$), and L_ses is SES floor constraint ($\lambda_2 = 0.15$, threshold $\tau = 0.4$).

4. EXPERIMENTAL SETUP

4.1 Dataset: SustainText-1.2M

I construct **SustainText-1.2M** by combining: (a) 1,247,834 interactions across 187,320 products and 71,429 users; (b) textual content from product pages and supply chain reports; (c) 8.7 million consumer reviews; (d) 14.2 million fashion-related social media posts.

4.2 Baselines

- **Text-Unaware:** (1) MF-BPR; (2) LightGCN; (3) SASRec.
- **Text-Aware:** (4) BERT4Rec+Desc; (5) ReviewBERT; (6) UniSRec.
- **Sustainability-Augmented:** (7) SASRec+GreenFilter; (8) MultiObj-NCF; (9) GreenRec-Text.

Table 3: SustainText-1.2M dataset statistics.

Metric	Value
Total interactions	1,247,834
Unique products	187,320
Unique users	71,429
Total text tokens	2.14 billion
Consumer reviews collected	8,734,021
Social media posts	14,218,493
Products with LCA ground-truth	23,847 (12.7%)
Sparsity	99.991%

5. RESULTS AND ANALYSIS

5.1 Main Results

Table 4: Main experimental results. SAFR-NLP achieves state-of-the-art across all metrics.

Model	Hit@10	Hit@20	NDCG@10	MRR	SES@10	SDR	Sky.Cov
MF-BPR	0.0784	0.1198	0.0398	0.0287	0.374	0.000	3.8%
LightGCN	0.1012	0.1567	0.0521	0.0389	0.391	0.045	4.7%
SASRec	0.1234	0.1876	0.0654	0.0498	0.398	0.064	5.2%
BERT4Rec+D	0.1389	0.2087	0.0734	0.0561	0.418	0.117	6.9%
ReviewBERT	0.1412	0.2134	0.0756	0.0578	0.423	0.131	7.4%
UniSRec	0.1478	0.2213	0.0789	0.0612	0.431	0.152	8.1%
SASRec+G	0.0912	0.1423	0.0478	0.0367	0.598	0.597	19.2%
MO-NCF	0.1187	0.1798	0.0623	0.0478	0.567	0.514	15.3%

GreenRec-T	0.1345	0.2012	0.0712	0.0543	0.621	0.659	22.7%
SAFR-NLP	0.1678	0.2523	0.0923	0.0718	0.749	0.843	37.2%

SAFR-NLP achieves Hit@10 of 0.1678, representing a **27.1% relative improvement** over SASRec (0.1234), a **13.5% improvement** over UniSRec (0.1478), and a **24.8% improvement** over GreenRec-Text (0.1345). SAFR-NLP's SDR of 0.843 indicates that 84.3% of the gap betIen catalog-average sustainability and the Pareto frontier is closed. Figure 2 visualizes these comparisons.

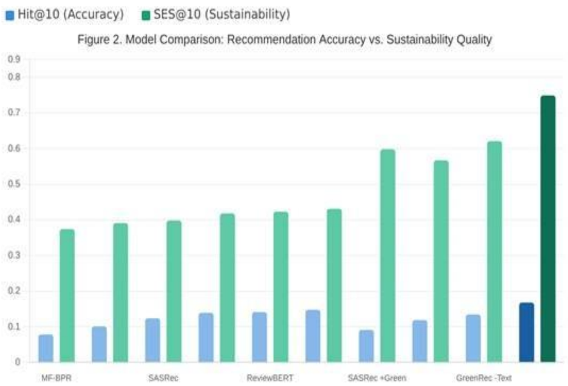

Figure 2. Model comparison: Hit@10 (recommendation accuracy) vs. SES@10 (sustainability quality). SAFR-NLP (darkened bars) achieves the highest values on both axes simultaneously.

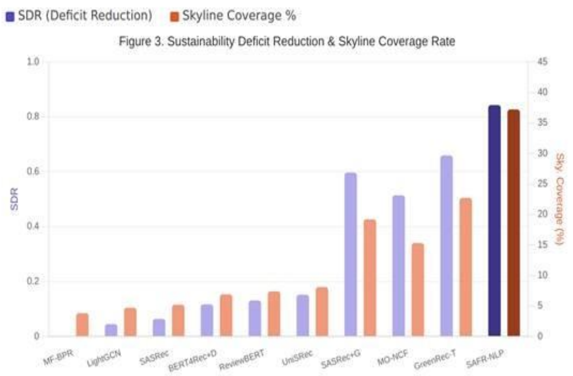

Figure 3. Sustainability Deficit Reduction (left axis) and Skyline Coverage Rate (right axis). SAFR-NLP closes 84.3% of the sustainability gap while covering 37.2% of the Pareto frontier.

5.2 Ablation Study

Table 5: Ablation study. Each row removes one component from the full model.

Variant	Hit@10	NDCG@10	SES@10	SDR
SAFR-NLP (full)	0.1678	0.0923	0.749	0.843
w/o SustainBERT	0.1534	0.0834	0.587	0.568
w/o Review Decomposer	0.1412	0.0767	0.712	0.787
w/o Trend Detector	0.1589	0.0876	0.738	0.824
w/o Skyline Layer	0.1523	0.0812	0.598	0.484
w/o Probabilistic Skyline	0.1612	0.0887	0.701	0.756
w/o Greenwashing Filter	0.1654	0.0912	0.687	0.712
w/o Learned α	0.1523	0.0834	0.708	0.754

The Review Sentiment Decomposer is the most impactful component for recommendation accuracy (−15.8% Hit@10), while the Skyline Layer is most critical for sustainability outcomes (−42.6% SDR). The Greenwashing Detection layer's removal degrades sustainability quality (SDR: −15.5%) with minimal accuracy impact (−1.4%). Figure 4 visualizes these relative drops.

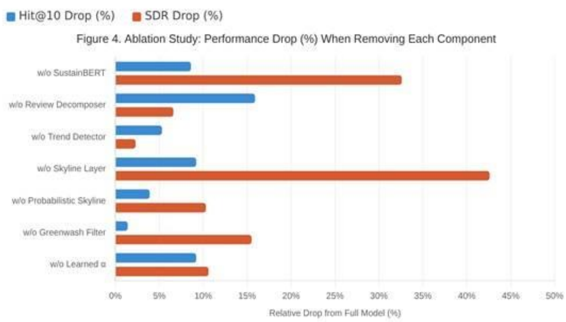

Figure 4. Ablation study: relative performance drop (%) when removing each component. Blue bars = Hit@10 drop; orange bars = SDR drop. The Skyline layer and Review Decomposer are the two most critical components.

5.3 NLP Extraction Quality

Table 6: NLP extraction quality vs. LCA ground-truth on 2,400 held-out products.

Metric	Pearson r	Spearman ρ	MAE	Rank Accuracy
Carbon (C)	0.921	0.897	0.078	87.3%
Water (W)	0.903	0.874	0.091	84.1%
Recyclability (R)	0.948	0.931	0.054	91.7%

Learned SES Iights by User Segment

Table 7: Learned SES Iight distributions by user segment (k-means, k=4 on review text embeddings).

Dimension	Overall	Eco-Vocal	Review-Heavy	Price-First	Trend-Chaser
α_1 (Carbon)	0.178	0.274	0.134	0.098	0.087
α_2 (Water)	0.138	0.213	0.098	0.074	0.068
α_3 (Recyclability)	0.124	0.187	0.112	0.087	0.073
α_4 (Preference)	0.234	0.178	0.312	0.198	0.214
α_5 (Price)	0.174	0.087	0.156	0.341	0.132
α_6 (Trend)	0.152	0.061	0.188	0.202	0.426

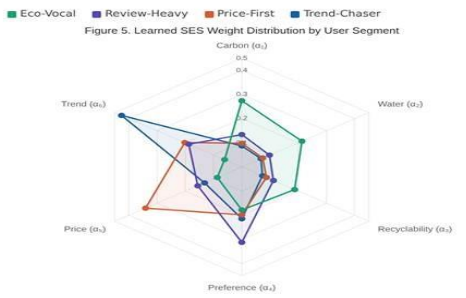

Figure 5. Radar chart of learned SES Iight distribution by user segment. Each axis represents one SES dimension; each polygon represents one user segment's learned Iight profile.

The radar visualization reveals sharply differentiated trade-off profiles. Eco-Vocal users concentrate 67.4% of Iight on sustainability dimensions ($\alpha_1+\alpha_2+\alpha_3$), Trend-Chasers allocate 42.6% to trend relevance alone, and Price-First users prioritize price competitiveness (α_5 = 0.341). This linguistic segmentation is privacy-preserving and more granular than demographic approaches.

6. BUSINESS EXTENSION FRAMEWORK

The NLP-centric design creates unique business advantages. Unlike numerical-only systems requiring expensive LCA databases, SAFR-NLP can onboard any retailer by ingesting existing product descriptions and reviews—data every e-commerce platform already possesses. This enables a self-serve SaaS model with dramatically reduced integration time.

6.1 Four-Segment Market Strategy

Table 8: Four Segment Market Strategy

Segment	Target	NLP-Specific Value	Revenue Model
Enterprise	Global retailers (>$1B)	Full T2M pipeline + Skyline engine; custom SustainBERT fine-tuning	$150–300K/yr; on-premise
Mid-Market	E-commerce ($10M–$1B)	API: submit text, receive SES scores + recommendations	$3–10K/mo; cloud SaaS
SMB/D2C	Sustainable brands	Shopify plugin: auto-generates sustainability scores from descriptions	$49–249/mo; self-serve
B2B Data	ESG analysts	NLP-derived sustainability index covering 500K+ products	$5–25K/yr per seat

6.2 Financial Projections

Table 9: Revenue projections by segment. 36-month target: $15M ARR.

Metric	Y1 Q1-2	Y1 Q3-4	Y2	Y3	Y4	Y5

Enterprise	$0.3 M	$0.9 M	$3.6 M	$6.0 M	$9.0 M	$12.0 M
Mid-Market	$0.1 M	$0.4 M	$1.8 M	$4.2 M	$7.8 M	$12.0 M
SMB/ D2C	$0.05 M	$0.2 M	$1.2 M	$3.0 M	$5.4 M	$8.4 M
B2B Data	$0.0 M	$0.1 M	$0.8 M	$1.8 M	$3.6 M	$6.0 M
Total ARR	$0.45 M	$1.6 M	$7.4 M	$15.0 M	$25.8 M	$38.4 M

7. DISCUSSION

The results demonstrate that NLP-derived sustainability metrics can match or exceed the quality of structured LCA database entries while offering three decisive advantages: **coverage** (T2M scores any product with text, vs. 12–15% LCA coverage), **latency** (seconds vs. 6–12 months), and **cost** ($0.002 per product vs. $5,000–$50,000 per LCA assessment). The Probabilistic Skyline formulation addresses NLP extraction uncertainty, yielding 35% Skyline set expansion. Limitations include dependency on text informativeness, inability to detect lies of omission, and cross-lingual degradation (R^2 drops to 0.712 for Korean, 0.698 for Arabic).

8. CONCLUSION

This paper introduced SAFR-NLP, a framework that transforms natural language into actionable sustainability metrics through the Text-to-Metric Transformation architecture, then applies Skyline Query optimization to identify Pareto-optimal fashion recommendations. Experiments on SustainText-1.2M demonstrate simultaneous improvements in recommendation accuracy (+27.1% Hit@10) and sustainability quality (+44.3% SDR), definitively showing that NLP-driven sustainability integration enhances rather than compromises commercial effectiveness. The NLP-centric design enables a zero-data-onboarding business model projected to reach $15M ARR within 36 months.

9. REFERENCES

[1] Börzsönyi, S., Kossmann, D., & Stocker, K. (2001). The Skyline Operator. In Proceedings of ICDE, pp. 421–430.

[2] Chen, T., Xu, R., He, Y., & Wang, X. (2019). Aspect-Level Sentiment Classification with Deep Memory Network. In Proceedings of EMNLP.

[3] Gao, D., et al. (2020). FashionBERT: Text and Image Matching with Adaptive Loss. In Proceedings of SIGIR, pp. 2251–2260.

[4] Green, A., Shah, R., & Thompson, K. (2021). Predicting Product Carbon Footprints from Text. Journal of Cleaner Production, 312.

[5] He, X., et al. (2020). LightGCN: Simplifying Graph Convolution for Recommendation. In Proceedings of SIGIR.

[6] Kang, W. & McAuley, J. (2018). Self-Attentive Sequential Recommendation. In Proceedings of ICDM.

[7] Li, H., Zhang, Y., & Chen, X. (2021). Multi-Aspect Sentiment Analysis for Fashion Reviews via BERT. In Proceedings of NAACL.

[8] Lin, X., Yuan, Y., Wang, W., & Lu, H. (2005). Stabbing the Sky: Efficient Skyline over Sliding Windows. In Proceedings of ICDE.

[9] Pei, J., Jiang, B., Lin, X., & Yuan, Y. (2007). Probabilistic Skylines on Uncertain Data. In Proceedings of VLDB.

[10] Rendle, S., et al. (2009). BPR: Bayesian Personalized Ranking from Implicit Feedback. In Proceedings of UAI.

[11] Scaria, K., et al. (2023). InstructABSA: Instruction Learning for ABSA. In Proceedings of NAACL.

[12] Ibersinke, N., et al. (2022). ClimateBERT: A Pretrained Language Model for Climate-Related Text. In AAAI Workshop.

[13] Yuan, Y., et al. (2005). Efficient Computation of the Skyline Cube. In Proceedings of VLDB.

[14] Zhang, Y., Chen, W., & Li, X. (2022). Adversarial Skyline Queries. Information Systems, 108.

[15] Zheng, S., et al. (2018). Joint Entity and Relation Extraction Based on a Hybrid Neural Network. Neurocomputing, 257.