On the Factual Consistency of Text-based Explainable Recommendation Models

Ben Kabongo
Sorbonne University, CNRS, ISIR
Paris, France
ben.kabongo@sorbonne-universite.fr

Vincent Guigue
AgroParisTech, UMR MIA Paris-Saclay
Palaiseau, France
vincent.guigue@agroparistech.fr

***Abstract*—Text-based explainable recommendation aims to generate natural-language explanations that justify item recommendations, to improve user trust and system transparency. Although recent advances leverage Large Language Models (LLMs) to produce fluent outputs, a critical question remains underexplored: are these explanations factually consistent with the available evidence? We introduce a comprehensive framework for evaluating the factual consistency of text-based explainable recommenders. We design a prompting-based pipeline that uses LLMs to extract atomic explanatory statements from reviews, together with domain-specific topics and sentiment labels, which are then used to construct ground-truth explanations. Applying this pipeline to five categories from the Amazon Reviews dataset, we create augmented benchmarks for fine-grained evaluation of explanation quality. We further propose statement-level alignment metrics that combine LLM-based and Natural Language Inference (NLI) approaches to assess both factual consistency and relevance of generated explanations. Across extensive experiments on six state-of-the-art explainable recommendation models, including recurrent, transformer-based, and LLM-enhanced architectures, we observe substantial gaps between surface-level text quality and factual accuracy. Our analysis shows frequent hallucination of explanatory content, with statement-level precision ranging from 4.38% to 32.88% across datasets. These findings underscore the need for factuality-aware evaluation in explainable recommendation and provide a foundation for developing more trustworthy explanation systems.** [1]

***Index Terms*—Explainable recommendation; Recommender systems; Factual consistency; Benchmarking and evaluation metrics; Large Language Models (LLMs); Natural Language Inference (NLI); Data quality and trust.**

I. Introduction

Recommender systems have become integral to modern digital platforms, guiding users through vast catalogs of products, content, and services. However, traditional recommendation approaches [1]–[4] often operate as black boxes, providing little insight into *why* a particular item is suggested. This opacity can undermine user trust and hinder the adoption of recommendation systems in domains where transparency is critical [5]. Explainable recommendation addresses this limitation by generating human-interpretable justifications for recommendations, thereby enhancing transparency, user satisfaction, and trust.

[1]We publicly release our implementation at https://github.com/BenKabongo25/factual_explainable_recommendation.

TABLE I
Non-Factuality of Text-based Explainable Recommender Models
(BLUE = correct, RED = hallucination)

Item title: skechers women's go walk 2-spark walking shoe

Rating: 4.0

Review: I really like the Skechers Go Walk. This updated version is a bit larger than my previous pair. They are very comfortable to walk in, but my feet got hotter in these as compared to the previous version.

Ground-truth Explanation: The user would appreciate this product because it is very comfortable to walk in. However, they may dislike that it causes feet to feel hotter compared to previous version. They seem indifferent to it is a bit larger than previous version.

Att2Seq: The user would appreciate this product because it are very light, it are comfortable, it are suitable for running [...]. However, they may dislike that it are not suitable for running due to lack of cushioning. They seem indifferent to it are not suitable for running due to foot shape.

XRec: The user would appreciate this product because it is comfortable, and it is suitable for casual wear. However, they may dislike that it does not provide adequate support, and it does not fit as expected. [...].

Among various explanation paradigms, text-based explainable recommendation has emerged as a particularly promising approach, leveraging the flexibility and expressiveness of natural language to convey personalized rationales [6]–[11]. Recent advances have increasingly turned to Large Language Models (LLMs) [12], [13], which can generate fluent, contextually appropriate explanations by drawing on vast pretraining corpora and sophisticated language understanding capabilities. These models have demonstrated impressive performance on standard text generation metrics [14]–[17], producing explanations that appear coherent and plausible at first glance.

Yet a critical question remains largely unexplored: *Are the explanations generated by state-of-the-art models factually consistent with the available evidence?* While surface-level fluency is important, the true value of an explanation lies in its *factuality*: whether the explanatory content accurately reflects the user's actual preferences as expressed in their reviews. As illustrated in Table I, even sophisticated models, such

as XRec [11], can hallucinate explanatory content, misrepresenting user sentiments expressed in reviews. Such factual inconsistencies can erode user trust and diminish the utility of explainable recommendation systems.

The challenge of evaluating factual consistency in explainable recommendation is compounded by several factors. First, reviews (the primary source of explanatory evidence) mix genuinely explanatory passages with irrelevant or noisy content, making it difficult to isolate the true explanatory signal. Second, widely used evaluation metrics such as BLEU [18], ROUGE [19], and even semantics-aware measures like BERTScore [17], BARTScore [16], BLEURT [15] focus primarily on surface-level similarity rather than factual accuracy [20], [21]. These metrics can assign high scores to explanations that are fluent but factually inconsistent. Third, existing factuality-oriented metrics from summarization [21], [22] and question-answering [23] were not designed for the unique characteristics of explainable recommendation, where explanations must capture fine-grained, aspect-level user preferences and sentiments.

To address these limitations, we introduce a comprehensive framework for evaluating the factual consistency of text-based explainable recommenders, comprising three key contributions:

(1) Statement-Level Ground-truth Construction. We design a prompting-based pipeline that leverages LLMs to extract atomic explanatory statements from user reviews with their associated domain-specific topics and sentiment labels. This fine-grained decomposition isolates explanatory content from noise while preserving all relevant information. A rule-based aggregation procedure then constructs ground-truth explanations that faithfully represent the complete statement set, avoiding information loss from word limits or prompt dependencies [10], [11].

(2) Augmented Benchmark Datasets. We apply our pipeline to five Amazon Reviews categories [24] (*Toys and Games*, *Clothing*, *Beauty*, *Sports*, and *Cellphones*) creating augmented benchmarks pairing each user–item interaction with extracted statements and derived ground-truth explanations. These datasets enable fine-grained evaluation across domains and provide a foundation for future factual consistency research.

(3) Statement-Level Factuality Metrics. Building on recent advances [21], [22], [25], we propose metrics tailored to explainable recommendation. Combining LLM-based and NLI approaches, our metrics assess both factual consistency (precision) and relevance (recall) at the statement level, capturing fine-grained alignment between explanatory passages.

We conduct extensive experiments on six state-of-the-art explainable recommendation models spanning three architectural families: recurrent models (NRT [7], Att2Seq [6]), transformer-based models (PETER [8], CER [26], PEPLER [9]), and LLM-enhanced models (XRec [11]). Our evaluation reveals substantial gaps between surface-level text quality and factual accuracy. While models achieve high scores on standard similarity metrics (e.g., BERTScore F1 ranging from 0.81 to 0.90), statement-level precision scores tell a different story, ranging from as low as 4.38% (NRT on *Toys*) to 32.88% (XRec on *Sports*). This discrepancy highlights the limitations of existing evaluation practices and underscores the prevalence of hallucinated explanatory content in current systems.

Our findings have important implications for the design and evaluation of explainable recommenders. They demonstrate that fluency and factuality are distinct dimensions of explanation quality, and that progress on one does not guarantee progress on the other. The low factual consistency scores we observe suggest that current models struggle to ground their explanations in verifiable evidence, relying instead on generic or hallucinated content. Addressing this challenge will require not only better evaluation methodologies, such as the statement-level metrics we propose, but also fundamental innovations in model architectures and training objectives that explicitly prioritize factual grounding.

In summary, our contributions are:

- A prompting-based pipeline for extracting atomic explanatory statements with topics and sentiments from reviews, enabling fine-grained ground-truth construction.
- Five augmented benchmark datasets spanning diverse product categories, providing a foundation for factuality-aware evaluation in explainable recommendation.
- A comprehensive suite of statement-level metrics combining LLM-based and NLI-based approaches to assess factual consistency and relevance.
- Extensive experiments on six state-of-the-art models revealing substantial gaps between surface-level quality and factual accuracy, with precision scores ranging from 4.38% to 32.88%.

II. Related Work

A. Explainable Recommendations

Explainable recommendation aims to provide insights into why an item is recommended, thereby improving transparency, effectiveness, and trust in recommender systems [5].

Explanation Types Early, traditional approaches justify recommendations using user–user or item–item similarity [27], but such explanations are often opaque or lack detail. A second line of work explains recommendations through item aspects, attributes, or features [28], [29], expressing the user's appraisal of the recommended item on specific aspects. Visual explanation methods have also been explored [30], leveraging visual characteristics to justify recommendations. Finally, text-based explainable recommendation [6]–[11], [31], [32] has gained momentum, especially in combination with Large Language Models (LLMs). In this paper, we concentrate on evaluating models in this latter category.

Text-based Explanation Early text-based explanation methods relied on predefined templates, mainly driven by item features and opinion words [33]–[36]. More recent work turns to free-form generation, producing tips [7], reviews [6], [8], [9], [32], or an explanatory paragraph derived from the

review [10], [11]. Within this line, the earliest approaches are based on recurrent neural networks [37]. Att2Seq [6] uses an attention-based Long Short-Term Memory (LSTM) [38] to generate reviews from attributes, including user and item representations. NRT [7] adopts a multi-task setup that predicts the overall rating from user and item representations and generates a tip-style explanation with a Gated Recurrent Unit (GRU) [39].

PETER [8] follows the multi-task paradigm but replaces recurrence with an untrained Transformer [40]. Several architectures extend PETER [26], [41], [42]. In particular, CER [26] addresses alignment between the generated explanation and the predicted rating by encouraging consistency between the model's rating prediction and the rating entailed by the explanation. To leverage language model pretraining, PEPLER [9] fine-tunes GPT-2 [43] to generate explanations from user and item embeddings. To improve factuality and informativeness, PRAG [32], inspired by Retrieval-Augmented Generation (RAG) [44], employs a personalized retriever to select past reviews that condition the language model during generation.

Another line of work incorporates aspect information (attributes or item features) to improve the factuality of explanations [8], [31], [45]–[48]. In particular, PETER+ [8] extends PETER by additionally conditioning the generation on item feature words. ELIXIR [31] learns global and aspect-specific representations of users and items, then constructs a prompt to guide explanation generation, in the same spirit as MAPLE [48]. The main limitation of these approaches is their reliance on aspect annotations, which are often unavailable in practice.

More recently, XRec [11] and G-Refer [10] incorporate collaborative-filtering signals learned with Graph Neural Networks (GNNs) [2] into LLMs to better capture complex user–item interaction patterns and user preferences. These models learn, in addition to latent representations, user and item profiles from reviews, which are inserted into the LLM prompt for generation. XRec [11] introduces a lightweight collaborative adapter that injects user and item latent representations into the LLM. G-Refer [10] adopts a hybrid graph-retrieval mechanism that gathers collaborative signals from both structural and semantic perspectives and integrates them into the LLM prompt to guide explanation generation.

B. Evaluation of Text-based Explainable Recommenders

Text-based explainable recommendation methods employ a variety of metrics to assess generation quality. Early work largely relied on n-gram overlap metrics such as BLEU [18] and ROUGE [19]. However, these metrics depend on exact word matching and n-gram overlap, which makes them struggle with synonymy and paraphrasing.

More recent studies therefore incorporate semantics-aware metrics, including GPTScore [14], BERTScore [17], BARTScore [16], and BLEURT [15]. These approaches rely on pre-trained models to estimate similarity between texts, either through contextual embeddings [17] or by framing scoring as a conditional generation task [14]–[16]. While useful for comparing systems on surface-level generation quality, these metrics do not adequately capture the factuality of generated explanations [20]–[22], i.e., their consistency with the ground-truth explanation.

To assess model informativeness, the *Unique Sentence Ratio* (USR) has been proposed [33]. USR counts the number of unique sentences in a generated explanation using exact word matching. This metric has limitations: a single sentence may contain redundant information, and two sentences that are paraphrases conveying the same idea may still be counted as different.

C. Factual Consistency and Relevance Evaluation

Generated explanations in explainable recommendation must be factual to enhance usability and user trust. However, widely used text-similarity metrics [14]–[17], common in prior evaluations, do not adequately capture the factuality of existing text-based explainable recommendation models.

Some metrics have also been proposed to assess the consistency of generated explanations with respect to item features [33]. In particular, *Feature Matching Ratio* (FMR) measures whether a given input feature appears in the generated explanation, *Feature Coverage Ratio* (FCR) measures at the corpus level how well the set of generated explanations covers the input features, and *Feature Diversity* (DIV) measures the diversity of features mentioned across generated explanations. One limitation of these metrics is their reliance on exact word matching of feature names, which prevents handling synonyms and paraphrases. A second limitation is that they quantify the presence of features but do not indicate whether the sentiment associated with a feature matches between the ground-truth and the generated explanation.

Overall, relatively few works on explainable recommendation have examined the factuality of generated explanations, despite its importance. PRAG [32] introduces the *Entail* measure based on textual entailment models [49], reporting the fraction of generated explanations that entail at least one item-level reference explanation. While this captures factuality with respect to the item, it does not assess factual consistency between the generated explanation and the ground-truth.

Factual consistency is a central concern in summarization [20], [22] and, more broadly, in the evaluation of large language models [50], [51]. Initially, n-gram overlap metrics such as ROUGE [19], BLEU [18], and METEOR [52] were widely used, but they have well-known limitations and correlate weakly with human judgments. Subsequent work proposed a variety of metrics for evaluating factual consistency between texts [21], [22], [25], including NLI-based methods that use entailment models [21], [22], [53], LLM-based metrics [25], and QA-based approaches [23], [54].

NLI-based metrics include SummaC [21] and AlignScore [22], which are precision-oriented. They measure factual consistency between a summary and its source by decomposing both texts into chunks, paragraphs, or sentences, scoring pairs, and aggregating the scores into a single value. SEval [25]

uses an LLM to extract atomic statements from the summary and the source document, then evaluates each statement for factual support, yielding precision and recall. QuestEval [23] provides a unified Question Generation (QG) and Question Answering (QA) framework to measure factual consistency of the summary with respect to the document (precision) and the relevance of the document with respect to the summary (recall). Two texts are considered aligned when their answers agree on a set of generated questions.

To more accurately assess the factual consistency and relevance of explanations produced by explainable recommendation models, we adopt the suite of metrics comprising SummaC [21], AlignScore [22], SEval [25], and QuestEval [23]. We also derive statement-centric metrics, grounded in both NLI and LLM scoring, to evaluate factuality at a fine-grained level.

III. Framework for Factual Explainable Recommendation

To evaluate the factuality of explainable recommendation models, we propose a framework comprising three components: (1) a prompting-based pipeline that extracts atomic statement–topic–sentiment triplets from reviews to construct ground-truth explanations, (2) five augmented benchmark datasets from Amazon Reviews, and (3) a comprehensive suite of factuality metrics combining NLI and LLM-based approaches (detailed in Section IV).

A. Preliminaries

Let $\mathcal{U}$ be the set of users and $\mathcal{I}$ the set of items, with cardinalities $|\mathcal{U}|$ and $|\mathcal{I}|$, respectively. Given an explicit interaction between user u and item i, let r_{ui} denote the overall rating and $\mathbf{t}_{ui}$ the associated review. We denote the observed interaction data by:

$$\mathcal{R} = \{(u, i, r_{ui}, \mathbf{t}_{ui})\}. \tag{1}$$

The goal of text-based explainable recommendation is to jointly predict the user's rating for the item and generate a textual rationale $\mathbf{e}_{ui}$ that explains the underlying user–item interaction.

Several state-of-the-art methods [6]–[9], [31], [32], directly use portions of the review $\mathbf{t}_{ui}$ as a proxy to explain the interaction between user u and item i. More recent approaches [10], [11] design prompts to transform the review into a concise explanatory paragraph using an LLM, typically with a word limit. However, such limits and, more broadly, prompt structure can lead to the loss of explanatory information present in the full review, potentially omitting aspects that are crucial to explaining the interaction.

Reviews typically combine explanatory content that justifies the rating with noise, including irrelevant, non-explanatory text. Depending on the domain (e.g., clothing, toys, sports), explanatory content can be grouped into high-level topics (e.g., *fit, material, comfort* in clothing). Let $\mathcal{T}$ denote the set of topics of interest. Generated explanations should be faithful to the source review. A largely overlooked evaluation dimension is whether a model recovers and highlights the factual explanatory passages that appear in the review.

B. Statement-Topic-Sentiment Triplets

An *atomic statement* is a polarized fact expressing the user's opinion about a single attribute or topic of the item. Given the review $\mathbf{t}_{ui}$ written by user u for item i, our first goal is to extract the atomic explanatory statements together with their corresponding topics and polarities. Large Language Models (LLMs) have demonstrated strong capabilities in natural-language understanding and generation [12], [13], [55], [56]. We therefore employ a pre-trained LLM to perform this extraction. In particular, we used `Llama-3-8B-Instruct`[2] [13] in our experiments. To preserve the full explanatory content extracted from each review, we construct an explanatory paragraph from the triplets using a rule-based procedure, which we adopt as the ground-truth for all models considered in our experiments.

Triplets Extraction For each domain, we specify a set of topics of interest $\mathcal{T}$. These topics can be defined with input from domain experts. In our experiments, we first design a prompt to elicit a shortlist of domain-specific topics. Given the topic set $\mathcal{T}$ and the sentiment label set $\mathcal{P} = \{\textit{positive}, \textit{negative}, \textit{neutral}\}$, we design a domain-specific prompt to extract from $\mathbf{t}_{ui}$ the set of atomic statements and assign to each statement its topic and polarity, as follows:

$$\mathcal{S}_{ui} = \pi(\mathbf{t}_{ui} \mid \cdot, \mathcal{T}, \mathcal{P}) = \{(\mathbf{s}_1, t_1, p_1), \cdots, (\mathbf{s}_n, t_n, p_n)\}, \tag{2}$$

where π denotes the language model, $\mathbf{s}_k$ is the k-th atomic explanatory statement in $\mathbf{t}_{ui}$, $t_k \in \mathcal{T}$ is its topic, and $p_k \in \mathcal{P}$ is its sentiment label. Figure 1 illustrates our pipeline for extracting statement–topic–sentiment triplets from reviews.

Ground-truth Explanation After obtaining the triplet set $\mathcal{S}_{ui}$ for the interaction between user u and item i, we apply a rule-based procedure to compose a single explanatory paragraph that combines all statements. We first group statements by polarity; for each polarity present, we form a sentence that aggregates its associated statements. For the *positive* polarity, for example, we write: `The user would appreciate this product because [POSITIVE_STATEMENTS]`. We follow the same pattern with `The user may dislike ...` for *negative* and `The user seems indifferent to ...` for *neutral*. We then concatenate the resulting sentences with simple logical connectors to form a well-structured paragraph. This approach avoids the cost of using an LLM for this step and ensures that all extracted statements are preserved in the constructed explanation.

C. Datasets

We apply our triplet-extraction pipeline to five categories from the Amazon Reviews 2014 dataset[3] [24], recovering atomic explanatory statements together with their topics and

[2]meta-llama/Meta-Llama-3-8B-Instruct

[3]https://jmcauley.ucsd.edu/data/amazon/links.html

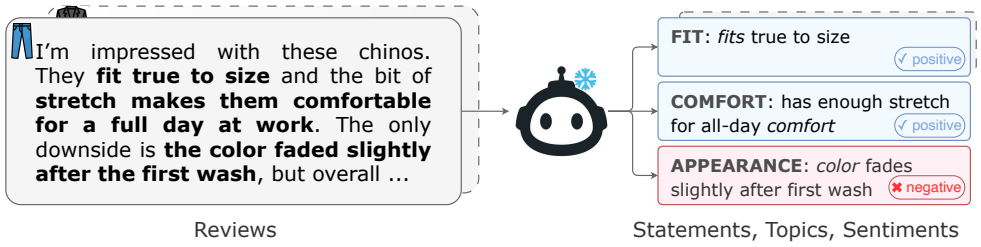

Fig. 1. Our review-to-statements pipeline. We prompt an LLM to extract, from each review, the explanatory passages as atomic statements, along with their domain-specific topic (here: clothing) and sentiment label.

TABLE II
DATASETS STATISTICS

	Toys	Clothes	Beauty	Sports	Cellphones
Users	19 398	39 385	22 362	35 596	27 873
Items	11 924	23 033	12 101	18 357	10 429
Interactions	163 711	274 774	197 621	293 244	190 194
Train	121 751	203 574	149 569	219 913	139 889
Validation	14 805	24 396	18 506	27 394	16 099
Test	22 441	41 995	27 862	42 675	28 901
Statements					
Avg/interaction	5.03	4.42	5.45	4.93	4.54
Avg/user	41.76	30.12	46.99	40.24	30.65
Avg/item	67.49	50.70	84.79	76.90	81.42
Unique	587 114	619 917	622 276	1 055 145	662 466
Total	823 932	1 215 270	1 076 769	1 447 240	863 036

sentiments. The categories are **Toys and Games** (*Toys*), **Clothing, Shoes and Jewelry** (*Clothes*), **Beauty** (*Beauty*), **Sports and Outdoors** (*Sports*), and **Cell Phones and Accessories** (*Cellphones*).

For each dataset, we define a list of ten topics of interest and design a dataset-specific prompt with a few illustrative examples to extract all explanatory triplets from every interaction. From each review and its extracted statements, we then construct the corresponding ground-truth explanation.

For evaluation, we perform a temporal split per user. We retain users with at least five interactions in total. For each user, we use 80% of interactions for training, 10% for validation, and 10% for testing in chronological order. Because most considered models are not robust to cold start, we remove from the validation and test sets any items not observed during training. Dataset statistics are reported in Table II.

IV. EVALUATION METRICS

We evaluate models with a suite of metrics, including LLM-as-a-judge, Natural Language Inference (NLI), Question Generation and Answering (QG-QA), and textual similarity measures. Building on these, we introduce LLM-based and NLI-based metrics tailored to our setting, focused on assessing factual consistency and relevance at the statement level. We first briefly review existing metrics and then present our proposed metrics.

A. Existing Metrics

We evaluate models using state-of-the-art metrics for factual consistency, relevance, and text similarity.

1) LLM-based Metrics: **SEval** [25] leverages the natural language understanding capabilities of LLMs to decompose inputs into complete atomic statements, score each statement, and then aggregate the scores into a single value.

Formally, given texts $\mathbf{a}$ and $\mathbf{b}$, the first step uses an LLM to extract their statements. This yields n statements $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ for $\mathbf{a}$ and m statements $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ for $\mathbf{b}$. For each statement $\mathbf{a}_k$, using an LLM, we compute a binary factual-consistency score $a_k \in \{0, 1\}$ with respect to $\mathbf{b}$, setting $a_k = 1$ if $\mathbf{a}_k$ is supported by the document $\mathbf{b}$. Symmetrically, for each $\mathbf{b}_\ell$ we compute b_ℓ with respect to $\mathbf{a}$. We then compute precision, recall, and F1 as:

$$\texttt{SEval-P}(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{k=1}^{n} a_k \quad \texttt{SEval-R}(\mathbf{a}, \mathbf{b}) = \frac{1}{m}\sum_{l=1}^{m} b_l,$$
$$\texttt{SEval-F1} = 2\,\frac{\texttt{SEval-P} \cdot \texttt{SEval-R}}{\texttt{SEval-P} + \texttt{SEval-R}}. \tag{3}$$

We use the public implementation[4] in our experiments.

2) Entailment-based Metrics (NLI-based): We present two entailment-based metrics: SummaC [21] and AlignScore [22].

SummaC [21] decomposes both input texts into paragraphs or sentences, applies an NLI model to score each pair, and aggregates the scores using two alternative schemes.

Formally, given texts $\mathbf{a}$ and $\mathbf{b}$, decompose them into n sentences $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and m sentences $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$, respectively. For each pair $(\mathbf{a}_k, \mathbf{b}_l)$, obtain from an NLI model the scores (E_{kl}, C_{kl}, N_{kl}) for entailment, contradiction, and neutral, respectively. This yields a score matrix X with entries X_{kl}, where X_{kl} denotes the final score between $\mathbf{a}_k$ and $\mathbf{b}_l$, typically taken as the entailment score E_{kl}. The aggregated score is computed from the matrix X using two schemes, corresponding to the two metric variants, `SummaC-ZS` and `SummaC-Conv`. In `SummaC-ZS`, for each sentence $\mathbf{a}_k$ we take the maximum score over all $\mathbf{b}_l$ and then average across sentences:

$$\texttt{SummaC-ZS}(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{k=1}^{n} \max_{l} X_{kl}. \tag{4}$$

[4]https://github.com/TanguyHsrt/seval-ex/tree/main

Given the sensitivity of SummaC-ZS to outliers, SummaC-Conv learns parameters for more robust score aggregation.

We report results for both variants in our evaluation. We use the publicly available implementation[5]. For SummaC-ZS, we adopt the NLI model DeBERTa-Large-mnli[6] [57] and the default scoring $X_{kl} = E_{kl} - C_{kl}$. For SummaC-Conv, we use the pretrained model provided in the repository.

AlignScore [22] introduces models pre-trained on a collection of unified tasks. The associated labels span ternary outputs (e.g., for NLI), binary outputs (e.g., for information retrieval), and continuous scores in $[0, 1]$ (e.g., for semantic textual similarity). Built on RoBERTa [58], the models are trained jointly over all datasets with three classification heads, each corresponding to one output type (continuous, binary, ternary).

In the same spirit as SummaC, AlignScore decomposes the inputs $\mathbf{a}$ and $\mathbf{b}$ into n chunks $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and m chunks $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$, and uses a SummaC-ZS-style aggregation over pairs $\big(\mathbf{a}_k, \mathbf{b}_l\big)$. The final score is given by:

$$\texttt{AlignScore}(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{k=1}^{n} \max_{l} \text{alignment}\big(\mathbf{a}_k, \mathbf{b}_l\big), \quad (5)$$

where $\text{alignment}(\cdot)$ denotes the probability of the ALIGNED class from the ternary classifier.

We use the publicly available implementation[7] in our experiments, along with the RoBERTa-Large[8] variant of the model.

3) QG–QA-based Metrics: **QuestEval** [23] couples Question Generation (QG) and Question Answering (QA) to assess both factual consistency and relevance of generated text. A text is considered consistent with a source if the answers to the generated questions match for both the source and the text.

Given a generated text and a reference, factual consistency (precision) is computed by evaluating the reference on questions derived from the generated text. Conversely, relevance is computed by evaluating the generated text on questions derived from the reference. The final score is the harmonic mean of these two scores. In our experiments, we use the publicly available implementation[9].

4) Reference-based Similarity Metrics: We present BERTScore [17] and the Semantic Textual Similarity (STS) measure.

BERTScore [17] measures similarity by comparing each token in a candidate sentence to every token in a reference sentence using contextual embeddings, and then applying greedy matching so that each token is paired with its most similar counterpart. We report precision, recall, and F1 variants as BERT-P, BERT-R, and BERT-F1, respectively. We use RoBERTa-Large[10] [58] and a publicly available implementation[11].

Semantic Textual Similarity (STS) models map text to sentence embeddings that capture semantic content and then measure their similarity [59], [60]. We use all-MiniLM-L6-v2[12] as the embedding model and report cosine similarity between ℓ_2-normalized sentence embeddings.

5) Learned and Likelihood-based Metrics: We consider two metrics here: BARTScore [16] and BLEURT [15].

BARTScore [16] frames text evaluation as a conditional generation problem: a pre-trained sequence-to-sequence model assigns higher likelihoods when the generated text better predicts the reference. We report this metric as BART. We use BART-Large[13] [61] and the publicly available code[14].

BLEURT [15] is a learned metric based on BERT [62] for evaluating text generation, and includes additional pretraining on synthetic data followed by fine-tuning on human judgements to train a model that scores system outputs. We use the publicly available implementation[15].

B. Our Metrics

A good explainable recommender should generate explanations whose passages are all supported by the reference (precision) while also covering as many of the reference's explanatory passages as possible (recall). Accordingly, we derive a set of metrics to measure the factual consistency and relevance of a generated explanation $\mathbf{e}'$ with respect to the ground-truth explanation $\mathbf{e}$ at the statement level. Our suite includes LLM-based metrics derived from SEval [25] and NLI-based metrics derived from SummaC [21] and AlignScore [22].

Statement Sets Given a user–item interaction, let $\mathbf{e}$ denote the ground-truth explanation and $\mathbf{e}'$ the explanation generated by a model. Using our statement–topic–sentiment extraction pipeline, we obtain the triplets associated with $\mathbf{e}$, written $\{(\mathbf{s}_1, t_1, p_1), \ldots, (\mathbf{s}_m, t_m, p_m)\}$. For each generated explanation $\mathbf{e}'$, we apply the same pipeline to extract $\{(\mathbf{s}'_1, t'_1, p'_1), \ldots, (\mathbf{s}'_n, t'_n, p'_n)\}$. To incorporate sentiment information, each triplet is formatted using the same pattern as in the ground-truth construction. For notational simplicity, the resulting statement sets are denoted $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$ for the ground-truth $\mathbf{e}$ and $\{\mathbf{s}'_1, \ldots, \mathbf{s}'_n\}$ for the prediction $\mathbf{e}'$. In what follows, we derive statement-based metrics to assess factual consistency and relevance of generated explanations.

Proposed Metrics Our approach centers on scoring each statement against a reference document (e.g., the generated or ground-truth explanation), followed by score aggregation to compute precision, recall, and their harmonic mean. We define a scoring function f that measures the factual consistency a statement with any text unit (explanation or another

[5]https://github.com/tingofurro/summac/tree/master
[6]https://huggingface.co/microsoft/deberta-base-mnli
[7]https://github.com/yuh-zha/AlignScore/tree/main
[8]https://huggingface.co/FacebookAI/roberta-large
[9]https://github.com/ThomasScialom/QuestEval
[10]https://huggingface.co/FacebookAI/roberta-large
[11]https://huggingface.co/spaces/evaluate-metric/bertscore
[12]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[13]https://huggingface.co/facebook/bart-large-cnn
[14]https://github.com/neulab/BARTScore
[15]https://huggingface.co/spaces/evaluate-metric/bleurt

statement). We consider two families of scoring functions. The first, f_{LLM}, is LLM-based (as in SEval [25]) and returns a binary score $\in \{0, 1\}$, where 1 indicates factual consistency. The second, f_{NLI}, is NLI-based (as in SummaC [21] and AlignScore [22]) and returns three probabilities (E, C, N) for entailment, contradiction, and neutrality, respectively. In this case, the statement score can be taken as either the entailment score E or the entailment-minus-contradiction score $E - C$.

1) LLM-based Metrics: The metrics introduced here use an LLM-based scoring function f_{LLM} to assess the factual consistency of a statement given a target text unit. We define: *Statement-to-Explanation Precision* (`St2Exp-P`), *Statement-to-Explanation Recall* (`St2Exp-R`), and *Statement-to-Explanation F1* (`St2Exp-F1`). All follow the SEval [25] framework.

Statement-to-Explanation Precision (`St2Exp-P`) measures the factual consistency of the generated explanation $\mathbf{e}'$ with respect to the ground-truth $\mathbf{e}$ at statement level. Formally, given n statements $\{\mathbf{s}'_1, \ldots, \mathbf{s}'_n\}$ extracted from $\mathbf{e}'$, we score each statement against the ground-truth: $s'_k = f_{\text{LLM}}(\mathbf{s}'_k, \mathbf{e})$. The metric is the fraction of positively supported statements, given by:

$$\texttt{St2Exp-P}(\mathbf{e}', \mathbf{e}) = \frac{1}{n}\sum_{k=1}^{n} s'_k \;=\; \frac{1}{n}\sum_{k=1}^{n} f_{\text{LLM}}(\mathbf{s}'_k, \mathbf{e}). \quad (6)$$

Statement-to-Explanation Recall (`St2Exp-R`) swaps the roles of $\mathbf{e}'$ and $\mathbf{e}$, thus measuring how well the generated explanation covers the reference's explanatory content. Given m statements $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$ extracted from $\mathbf{e}$, each statement is scored against the prediction: $s_l = f_{\text{LLM}}(\mathbf{s}_l, \mathbf{e}')$. The metric is given by:

$$\texttt{St2Exp-R}(\mathbf{e}', \mathbf{e}) = \frac{1}{m}\sum_{l=1}^{m} s_l \;=\; \frac{1}{m}\sum_{l=1}^{m} f_{\text{LLM}}(\mathbf{s}_l, \mathbf{e}'). \quad (7)$$

Statement-to-Explanation F1 (`St2Exp-F1`) is the harmonic mean of precision and recall, computed as:

$$\texttt{St2Exp-F1} \;=\; 2\,\frac{\texttt{St2Exp-P}\cdot\texttt{St2Exp-R}}{\texttt{St2Exp-P}+\texttt{St2Exp-R}}. \quad (8)$$

2) NLI-based Metrics: The metrics in this section use an entailment-based scoring function f_{NLI} to evaluate factual consistency between pairs of statements. We define two subgroups of metrics, differing in the choice of $f_{\text{NLI}}(\mathbf{s}_k, \mathbf{s}_l)$ for statements $\mathbf{s}_k$ and $\mathbf{s}_l$. In the first subgroup, the score is the entailment probability E_{kl}; we denote this by $f_{\text{NLI-ent}}$. In the second subgroup, the score is the difference between entailment and contradiction, $E_{kl} - C_{kl}$; we refer to this as the *coherence score* and denote it by $f_{\text{NLI-coh}}$. For the first subgroup, we also consider a binary variant $f_{\text{NLI-ent-bin}} \in \{0, 1\}$ that returns 1 when $E_{kl} = \max(E_{kl}, C_{kl}, N_{kl})$.

In parallel with the LLM-based family, we define precision scores (`StEnt-P`, `StEnt-bin-P`, `StCoh-P`), recall scores (`StEnt-R`, `StEnt-bin-R`, `StCoh-R`), and their harmonic means (`StEnt-F1`, `StEnt-bin-F1`), each using the corresponding scoring function.

Precision-oriented metrics These metrics measure the factual consistency of the generated explanation $\mathbf{e}'$ with respect to the ground-truth $\mathbf{e}$:

$$\texttt{St}\star\texttt{-P}(\mathbf{e}', \mathbf{e}) \;=\; \frac{1}{n}\sum_{k=1}^{n} \max_{l} f_{\text{NLI-}\star}(\mathbf{s}'_k, \mathbf{s}_l), \quad (9)$$

where $\star$ indexes the chosen scoring function. This yields `StEnt-P` (entailment score), `StEnt-bin-P` (binary entailment), and `StCoh-P` (entailment minus contradiction). In particular, `StEnt-P` is equivalent to `SummaC-ZS` when applied at the statement granularity (rather than the sentence granularity).

Recall-oriented metrics As with the previous family, recall metrics swap the roles of $\mathbf{e}'$ and $\mathbf{e}$ to assess coverage:

$$\texttt{St}\star\texttt{-R}(\mathbf{e}', \mathbf{e}) \;=\; \frac{1}{m}\sum_{l=1}^{m} \max_{k} f_{\text{NLI-}\star}(\mathbf{s}_l, \mathbf{s}'_k). \quad (10)$$

This gives `StEnt-R`, `StEnt-bin-R`, and `StCoh-R`.

Aggregated metrics From precision and recall we compute F1:

$$\texttt{St}\star\texttt{-F1} \;=\; 2\,\frac{\texttt{St}\star\texttt{-P}\cdot\texttt{St}\star\texttt{-R}}{\texttt{St}\star\texttt{-P}+\texttt{St}\star\texttt{-R}}, \quad (11)$$

yielding `StEnt-F1` and `StEnt-bin-F1`.

V. Experiments

We now present a comprehensive empirical evaluation of our factuality-aware evaluation framework for text-based explainable recommendation. Our experiments are designed to address several key research questions:

RQ1: How do state-of-the-art explainable recommendation models perform when evaluated through the lens of factual consistency at the statement level?

RQ2: Do existing surface-level similarity metrics correlate with factual accuracy, or do they provide misleading signals about explanation quality?

RQ3: How do different evaluation approaches (LLM-based, NLI-based, and traditional similarity metrics) compare in their ability to capture factual consistency?

RQ4: What are the specific failure modes of current models in terms of precision (hallucination) and recall (coverage) of explanatory content?

To answer these questions, we evaluate six representative models spanning three architectural families on our augmented benchmarks across five diverse product categories. We employ our comprehensive suite of statement-level metrics alongside standard evaluation approaches to provide a multi-faceted view of model performance. Our analysis reveals striking disparities between surface-level quality and factual grounding, highlighting fundamental limitations in current evaluation practices and model designs.

A. Experimental Protocol

We describe the experimental protocol used to evaluate text-based explainable recommendation models.

TABLE III
LLM-BASED STATEMENT-LEVEL EVALUATION RESULTS
(GREEN DENOTES THE BEST PERFORMANCE; RED DENOTES THE WORST PERFORMANCE)

Dataset	Metric	NRT	Att2Seq	CER	PETER	PEPLER	XRec
Toys	St2Exp-P	0.0438±0.1772	**0.2331±0.2844**	0.1909±0.2793	0.1849±0.2827	0.1565±0.2731	0.2297±0.3039
	St2Exp-R	0.0666±0.1441	0.1893±0.2498	0.1555±0.2371	0.1388±0.2245	0.1247±0.2053	**0.1906±0.2617**
	St2Exp-F1	0.0091±0.0597	**0.1317±0.1954**	0.0988±0.1772	0.0917±0.1736	0.0728±0.1550	0.1124±0.1887
Clothes	St2Exp-P	0.2102±0.2482	0.2725±0.2998	0.2236±0.2935	0.2438±0.3096	0.2427±0.2639	**0.2962±0.2897**
	St2Exp-R	0.1044±0.1706	0.2211±0.2646	0.1351±0.2128	0.1309±0.2117	0.2079±0.2466	**0.2794±0.3003**
	St2Exp-F1	0.0830±0.1465	0.1613±0.2149	0.0982±0.1746	0.0987±0.1775	0.1521±0.2016	**0.1913±0.2257**
Beauty	St2Exp-P	0.1215±0.3267	0.2621±0.3020	0.2313±0.3281	0.2302±0.3261	0.1963±0.2860	**0.2768±0.3068**
	St2Exp-R	0.0443±0.1135	0.2187±0.2557	0.1683±0.2291	0.1501±0.2207	0.1508±0.2077	**0.2986±0.3188**
	St2Exp-F1	0.0391±0.1277	0.1552±0.2128	0.1203±0.1978	0.1102±0.1960	0.0999±0.1751	**0.1735±0.2266**
Sports	St2Exp-P	0.1286±0.3347	0.2607±0.2887	0.2344±0.3423	0.2414±0.3491	0.2560±0.3156	**0.3288±0.3663**
	St2Exp-R	0.0027±0.0269	**0.2051±0.2512**	0.1181±0.1979	0.1201±0.1997	0.1450±0.2145	0.1990±0.2626
	St2Exp-F1	0.0037±0.0377	**0.1511±0.2053**	0.0929±0.1810	0.0958±0.1848	0.1123±0.1853	0.1473±0.2216
Cellphones	St2Exp-P	0.1107±0.2981	0.2816±0.3033	0.2603±0.3435	0.2241±0.3297	0.2147±0.2993	**0.3229±0.3369**
	St2Exp-R	0.0036±0.0351	0.2388±0.2777	0.1531±0.2327	0.1409±0.2262	0.1556±0.2246	**0.2896±0.3301**
	St2Exp-F1	0.0005±0.0129	0.1679±0.2206	0.1169±0.1990	0.1027±0.1891	0.1071±0.1849	**0.1825±0.2430**

1) Baselines: We consider three families of state-of-the-art models in our evaluation: RNN–based models (Att2Seq [6] and NRT [7]), Transformer-based models (PETER [8] and PEPLER [9]), and LLM-based models (XRec [11]). PEPLER [9] is built on GPT-2 [43], and for XRec [11] we use `Llama-2-7b`[16] [63], following the original paper. For each model, we adopt the best hyperparameters reported in the corresponding paper. We train all models for 100 epochs. For XRec, we follow the authors' experimental protocol: we first train the collaborative filtering backbone (LightGCN [2]) to convergence and then fine-tune the LLM for one epoch. For all baselines, we rely on publicly available implementations.

2) Datasets: We evaluate on five categories from the Amazon Reviews dataset [24]: *Toys*, *Clothes*, *Beauty*, *Sports*, and *Cellphones*. The full statement–topic–sentiment triplets extraction process and dataset statistics are provided in Section III-C.

3) Evaluation Metrics: We evaluate all models with the suite introduced in Section IV .We employ `Llama-3.1-8B-Instruct`[17] [13] for our LLM-based methods and `DeBERTa-large-mnli`[18] [57] for our NLI-based methods. For all models, we train on the training split, use the validation split for model selection, and report, for each metric, the mean and standard deviation computed over all test examples.

B. LLM-based Statement Results

Table III reports results on our LLM-based statement-level factual-consistency metrics: `St2Exp-P` (precision), `St2Exp-R` (recall), and `St2Exp-F1` (F1). Across models and datasets, scores are uniformly low for both precision and recall. The highest average precision is achieved by XRec on *Sports* with 32.88% (std. 33.89%); despite being the best, this remains modest and can be problematic in settings where precision is critical. Overall, `St2Exp-P` indicates that state-of-the-art systems exhibit low precision in explanation generation. NRT yields the lowest precision, e.g., 4.38% on *Toys*, and performs poorly across datasets and metrics.

Recall (`St2Exp-R`) is generally even lower than precision, confirming that current models fail to recover most of the ground-truth explanatory passages. This shortfall is especially concerning in scenarios that require comprehensive coverage or when certain passages are particularly salient to users. XRec attains the best recall at 29.86% on *Beauty*, implying that in most cases more than 70% of explanatory passages are missed. The lowest recall is observed for NRT with 0.27% on *Sports*.

Finally, `St2Exp-F1` corroborates these findings, with values ranging from 0.05% (NRT on *Cellphones*) to 19.13% (XRec on *Clothes*), indicating, under our LLM-based metrics, that current state-of-the-art models exhibit limited factual consistency.

C. NLI-based Statement Results

Table IV and Figure 2 present results for our NLI-based metrics, which compute entailment and contradiction scores over statement pairs. Overall scores are low, corroborating LLM-based findings on limited factual consistency while inducing different system rankings.

Entailment Precision scores (`StEnt-P`) range from 4.66% (NRT on *Toys*) to 24.80% (NRT on *Cellphones*), showing substantial cross-dataset variability. Precision can be inflated for models generating few statements when those statements are common in the data. Recall scores (`StEnt-R`) support this interpretation: despite higher precision, NRT achieves only 3.67% recall on *Cellphones*. Recalls remain generally low, with a maximum of 14.02% for PEPLER on *Clothes*. The resulting `StEnt-F1` ranges from 0.61% (NRT on *Toys*) to 13.81% (PEPLER on *Clothes*), indicating weak entailment and factual consistency.

[16]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[17]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[18]https://huggingface.co/microsoft/deberta-large-mnli

TABLE IV
NLI-BASED STATEMENT-LEVEL EVALUATION RESULTS
(GREEN DENOTES THE BEST PERFORMANCE; RED DENOTES THE WORST PERFORMANCE)

Dataset	Metric	NRT	Att2Seq	CER	PETER	PEPLER	XRec
Toys	StEnt-P	0.0466±0.1706	**0.0916±0.1542**	0.0805±0.1663	0.0831±0.1715	0.0729±0.1655	0.0538±0.1178
	StEnt-R	0.0127±0.0567	**0.0532±0.1131**	0.0522±0.1180	0.0530±0.1181	0.0441±0.1104	0.0435±0.1033
	StEnt-F1	0.0061±0.0328	**0.0410±0.0907**	0.0349±0.0880	0.0360±0.0902	0.0299±0.0831	0.0259±0.0674
	StCoh-P	-0.0261±0.2527	0.0048±0.2388	0.0160±0.2343	**0.0204±0.2381**	0.0098±0.2334	-0.0803±0.2594
	StCoh-R	-0.1639±0.2025	-0.0662±0.2217	-0.0590±0.2077	-0.0649±0.2125	-0.0895±0.2131	**-0.0588±0.2014**
Clothes	StEnt-P	0.2217±0.2074	0.1777±0.2106	0.2110±0.2329	0.2202±0.2444	**0.2422±0.2254**	0.1407±0.1679
	StEnt-R	0.1284±0.1803	0.1141±0.1690	0.1102±0.1694	0.1099±0.1682	**0.1402±0.1892**	0.1160±0.1686
	StEnt-F1	0.1259±0.1630	0.1016±0.1507	0.1077±0.1563	0.1088±0.1590	**0.1381±0.1708**	0.0895±0.1280
	StCoh-P	0.1222±0.3323	0.0931±0.2965	0.1188±0.3254	0.1341±0.3319	**0.1539±0.3222**	0.0250±0.2749
	StCoh-R	0.0372±0.2375	-0.0108±0.2761	-0.0112±0.2433	-0.0156±0.2448	**0.0390±0.2522**	0.0169±0.2640
Beauty	StEnt-P	0.1367±0.3265	0.1555±0.2022	0.1980±0.2693	**0.2076±0.2751**	0.2001±0.2447	0.1162±0.1641
	StEnt-R	0.0218±0.0713	**0.0915±0.1404**	0.0718±0.1308	0.0755±0.1314	0.0764±0.1330	0.0755±0.1282
	StEnt-F1	0.0323±0.1039	**0.0813±0.1297**	0.0753±0.1351	0.0806±0.1406	0.0810±0.1345	0.0597±0.1004
	StCoh-P	0.0963±0.3575	0.0823±0.2744	0.1326±0.3352	**0.1474±0.3401**	0.1373±0.3146	-0.0140±0.2975
	StCoh-R	-0.2454±0.1970	-0.0298±0.2420	-0.0746±0.2295	-0.0722±0.2297	-0.0570±0.2227	**-0.0286±0.2322**
Sports	StEnt-P	0.1588±0.3428	0.1521±0.1912	0.2063±0.2794	**0.2117±0.2868**	0.1574±0.2211	0.0973±0.1679
	StEnt-R	0.0215±0.0660	**0.0763±0.1303**	0.0706±0.1300	0.0704±0.1305	0.0744±0.1336	0.0488±0.1048
	StEnt-F1	0.0326±0.0991	0.0675±0.1161	**0.0709±0.1349**	0.0706±0.1353	0.0643±0.1204	0.0386±0.0874
	StCoh-P	0.0794±0.3867	0.0760±0.2571	0.1541±0.3338	**0.1590±0.3403**	0.0989±0.2761	-0.0673±0.3415
	StCoh-R	-0.3174±0.1908	**-0.0256±0.2180**	-0.0783±0.2248	-0.0814±0.2268	-0.0366±0.2100	-0.1245±0.2909
Cellphones	StEnt-P	**0.2480±0.3170**	0.1096±0.1697	0.1878±0.2654	0.1617±0.2582	0.2238±0.2545	0.1036±0.1664
	StEnt-R	0.0367±0.1009	0.0629±0.1235	0.0577±0.1220	0.0509±0.1135	**0.0661±0.1326**	0.0514±0.1128
	StEnt-F1	0.0399±0.1061	0.0463±0.0977	0.0580±0.1234	0.0494±0.1143	**0.0703±0.1328**	0.0402±0.0906
	StCoh-P	**0.1570±0.4285**	0.0157±0.2578	0.1078±0.3437	0.0782±0.3346	0.1365±0.3595	-0.0641±0.3295
	StCoh-R	-0.1805±0.2257	**-0.0575±0.2377**	-0.1101±0.2336	-0.1259±0.2222	-0.0692±0.2234	-0.0656±0.2322

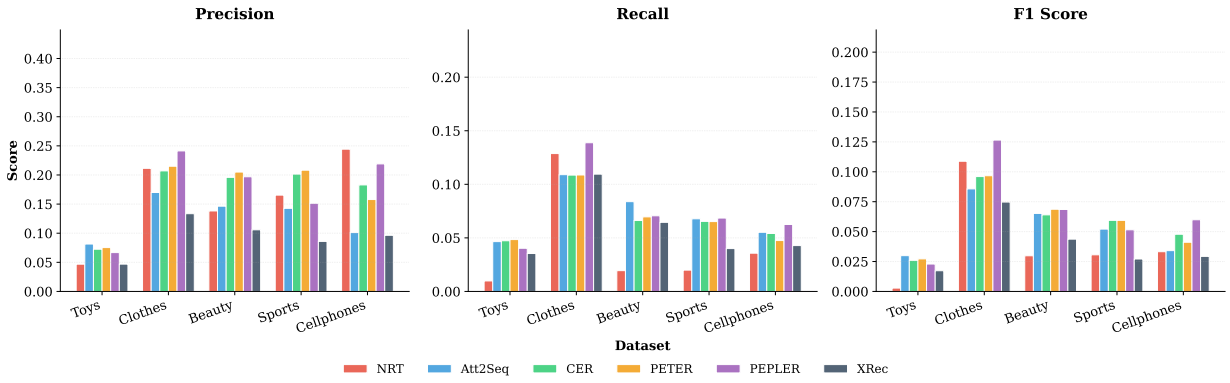

Fig. 2. StEnt-bin Results.

Binary variants confirm these trends (Figure 2): StEnt-bin-P peaks around 25% (NRT on *Cellphones*, PEPLER on *Clothes*), StEnt-bin-R remains below 15%, and StEnt-bin-F1 stays under 12.5%. Cross-dataset variability persists, with weakest scores on *Toys* and strongest on *Clothes*.

Coherence (Entailment minus Contradiction) The StCoh-⋆ metrics assess coherence via entailment-contradiction differences. Negative values indicate more contradiction than entailment. Even recent models like XRec yield negative precision and recall on several datasets, with only a few models (e.g., PEPLER on *Clothes*) achieving positive but modest scores. These results highlight that beyond low factual consistency, models can produce statements that directly contradict the reference.

LLM-based vs NLI-based Results Overall, we observe a disparity between our LLM-based and NLI-based metrics, attributable to their differing granularities. NLI metrics compare statement pairs, while LLM-based metrics compare each statement against the full explanation. Although NLI models enable efficient pairwise comparison at scale, prohibitively expensive for LLMs, the LLM-based approach better pre-

TABLE V
STANDARD NLI AND QG-QA METRICS RESULTS
(GREEN DENOTES THE BEST PERFORMANCE; RED DENOTES THE WORST PERFORMANCE)

Dataset	Metric	NRT	Att2Seq	CER	PETER	PEPLER	XRec
Toys	AlignScore	**0.3343±0.2647**	0.0720±0.1625	0.1236±0.2143	0.1336±0.2277	0.1652±0.2437	0.0756±0.1229
	SummaC-ZS	**-0.1903±0.2758**	-0.4111±0.3686	-0.3178±0.3531	-0.3120±0.3506	-0.2815±0.3296	-0.4006±0.3590
	SummaC-Conv	0.2021±0.0251	0.2053±0.0346	0.2046±0.0348	0.2044±0.0342	0.2042±0.0326	**0.2106±0.0468**
	QuestEval	0.3472±0.0721	0.3565±0.0817	0.3747±0.0817	**0.3763±0.0824**	0.3658±0.0847	0.3171±0.0744
Clothes	AlignScore	0.1644±0.2914	0.1277±0.2266	0.2167±0.3020	**0.2495±0.3134**	0.1677±0.2619	0.0892±0.1264
	SummaC-ZS	**-0.1696±0.3507**	-0.3831±0.4138	-0.1982±0.3570	-0.2154±0.3604	-0.1827±0.3636	-0.5305±0.3851
	SummaC-Conv	0.2081±0.0460	0.2098±0.0512	0.2079±0.0485	0.2077±0.0488	0.2124±0.0631	**0.2216±0.0705**
	QuestEval	**0.4114±0.0938**	0.3919±0.0937	0.4019±0.0936	0.4064±0.0941	0.3930±0.0930	0.3510±0.0762
Beauty	AlignScore	**0.4011±0.2904**	0.1317±0.2282	0.3064±0.3381	0.3052±0.3479	0.2536±0.3167	0.1007±0.1499
	SummaC-ZS	**-0.1409±0.2775**	-0.3660±0.3911	-0.2532±0.3435	-0.2313±0.3344	-0.2198±0.3305	-0.4440±0.3877
	SummaC-Conv	0.2034±0.0270	0.2073±0.0413	0.2056±0.0401	0.2052±0.0384	0.2068±0.0411	**0.2130±0.0534**
	QuestEval	**0.4142±0.0806**	0.3966±0.0909	0.4088±0.0892	0.4090±0.0916	0.3962±0.0884	0.3477±0.0844
Sports	AlignScore	0.2368±0.3148	0.1135±0.2005	0.2638±0.3149	**0.2743±0.3201**	0.2046±0.2608	0.1095±0.1636
	SummaC-ZS	**-0.1400±0.2566**	-0.3996±0.3754	-0.2252±0.3271	-0.2286±0.3299	-0.3262±0.3656	-0.3740±0.3639
	SummaC-Conv	0.2022±0.0178	0.2067±0.0373	0.2048±0.0368	0.2049±0.0366	0.2074±0.0403	**0.2096±0.0379**
	QuestEval	0.3712±0.0669	0.3802±0.0819	**0.4087±0.0871**	0.4056±0.0867	0.3811±0.0803	0.3076±0.0725
Cellphones	AlignScore	**0.6046±0.2822**	0.0852±0.1760	0.2569±0.3078	0.2367±0.2939	0.2212±0.2896	0.0920±0.1351
	SummaC-ZS	**-0.1709±0.2613**	-0.4743±0.3795	-0.2974±0.3617	-0.3055±0.3628	-0.2186±0.3288	-0.4420±0.3699
	SummaC-Conv	0.2010±0.0073	0.2068±0.0385	0.2047±0.0340	0.2040±0.0306	0.2059±0.0402	**0.2101±0.0433**
	QuestEval	0.3800±0.0753	0.3583±0.0811	0.3868±0.0924	**0.3874±0.0891**	0.3803±0.0912	0.3316±0.0761

serves the complete reference context when scoring individual statements. Nevertheless, both metric families converge on the same conclusion: models exhibit poor factual consistency.

D. Standard NLI and QG-QA Results

Table V reports model performance on state-of-the-art metrics, including NLI-based metrics (SummaC, AlignScore) and the QG-QA-based metric QuestEval. Unlike our statement-level metrics, these operate at different granularities: fixed-size chunks for AlignScore, sentences for SummaC, and full documents for QuestEval.

`AlignScore` ranks NRT as the most precise model on average, contradicting our statement-based findings. NRT achieves the highest score (60.48% on *Cellphones*), while Att2Seq and XRec, top performers on our LLM-based metrics, score poorly (7.20% for Att2Seq on *Toys*). Overall, alignment scores remain low. `SummaC-ZS` confirms this trend with consistent rankings. The negative scores across all models indicate that generated explanations contradict the ground-truth more often than they entail it, aligning with our `StCoh-*` findings. However, both `AlignScore` and `SummaC-ZS` are sensitive to outliers.

`SummaC-Conv`, the pre-trained SummaC variant, offers a more robust ranking despite less interpretable scores [21]. XRec performs best across datasets, followed by PEPLER and Att2Seq, while NRT ranks poorly except on *Clothes*. These results better align with our LLM-based metrics.

`QuestEval` presents a different picture: XRec performs worst, while Transformer-based models (CER, PETER, PEPLER) rank best. However, QuestEval's heavy dependence on generated questions may bias results. We believe precision- and recall-oriented metrics better suited to explainable recommendation are needed.

E. Text Similarity Results

Explanation Generation Table VI presents model evaluation using similarity metrics common in recent explainable recommendation work, with models evaluated against statement-derived ground-truth explanations. Results show trends contrary to prior work evaluating on shorter explanations. XRec ranks among the lowest on similarity metrics, outperformed even by NRT on *Clothes*. Att2Seq performs best across datasets on most metrics, followed closely by CER, while PEPLER consistently achieves the highest BLEURT [15] scores.

Review Generation Table VII reports model performance (excluding XRec) on review generation using text similarity metrics. Results align with those reported in PEPLER and PETER (evaluated on review generation with n-gram metrics), confirming that Transformer-based models generate reviews more similar to actual reviews than RNN-based predecessors, though this does not guarantee factuality. The disparity between Tables VI and VII suggests models excel at either explanation or review generation, but not necessarily both. However, since reviews contain noise alongside explanatory content, neither high review nor explanation similarity guarantees factual generation.

Comparison with our LLM-based metrics Critically, comparing text similarity metrics with our LLM-based factuality metrics reveals a striking disconnect: despite low factual precision and recall scores, models achieve very high similarity scores, with BERTScore F1 ranges from 0.81 to 0.90. Overall,

TABLE VI
TEXT SIMILARITY METRICS RESULTS
(GREEN DENOTES THE BEST PERFORMANCE; RED DENOTES THE WORST PERFORMANCE)

Dataset	Metric	NRT	Att2Seq	CER	PETER	PEPLER	XRec
Toys	BERT-P	0.8152±0.0129	**0.8922±0.0285**	0.8711±0.0347	0.8702±0.0352	0.8572±0.0367	0.8393±0.0288
	BERT-R	0.8720±0.0254	**0.8878±0.0276**	0.8841±0.0296	0.8837±0.0299	0.8797±0.0306	0.8592±0.0249
	BERT-F1	0.8424±0.0142	**0.8897±0.0241**	0.8773±0.0283	0.8766±0.0288	0.8680±0.0301	0.8489±0.0229
	STS	0.3157±0.1117	**0.3859±0.1551**	0.3838±0.1564	0.3799±0.1559	0.3496±0.1532	0.3532±0.1523
	BART	-3.5259±0.8088	**-3.3336±0.7122**	-3.3533±0.7830	-3.3689±0.7865	-3.3890±0.7730	-3.6835±0.7166
	BLEURT	-1.1833±0.1045	-0.6324±0.3019	-0.5498±0.2577	**-0.5399±0.2582**	-0.6882±0.2794	-0.8049±0.2604
Clothes	BERT-P	0.8844±0.0180	**0.8985±0.0292**	0.8835±0.0405	0.8820±0.0379	0.8749±0.0309	0.8378±0.0218
	BERT-R	0.8913±0.0287	**0.8945±0.0277**	0.8871±0.0297	0.8867±0.0296	0.8884±0.0298	0.8678±0.0239
	BERT-F1	0.8877±0.0201	**0.8963±0.0248**	0.8849±0.0310	0.8840±0.0298	0.8814±0.0270	0.8524±0.0189
	STS	**0.4313±0.1368**	0.4132±0.1519	0.3958±0.1574	0.3932±0.1570	0.4123±0.1479	0.3665±0.1401
	BART	-3.1714±0.7369	**-3.1158±0.6694**	-3.1569±0.7283	-3.1536±0.7288	-3.1716±0.7284	-3.2874±0.6340
	BLEURT	**-0.3169±0.2352**	-0.3567±0.2891	-0.3235±0.2719	**-0.3169±0.2715**	-0.6337±0.2695	-0.6489±0.2601
Beauty	BERT-P	0.7928±0.0177	**0.8923±0.0293**	0.8512±0.0371	0.8539±0.0370	0.8542±0.0427	0.8400±0.0255
	BERT-R	0.8451±0.0220	**0.8885±0.0263**	0.8761±0.0271	0.8776±0.0267	0.8763±0.0289	0.8583±0.0246
	BERT-F1	0.8179±0.0131	**0.8902±0.0242**	0.8631±0.0276	0.8652±0.0272	0.8648±0.0323	0.8488±0.0214
	STS	0.2407±0.0901	**0.4523±0.1605**	0.3882±0.1476	0.4075±0.1507	0.4042±0.1513	0.3997±0.1536
	BART	-3.4498±0.6700	**-3.1902±0.6146**	-3.2496±0.6733	-3.2448±0.6779	-3.2570±0.6675	-3.4463±0.6735
	BLEURT	-0.4922±0.1824	-0.4882±0.2853	**-0.4668±0.2435**	-0.4708±0.2796	-0.4866±0.2722	-0.7397±0.2825
Sports	BERT-P	0.7571±0.0130	**0.8868±0.0313**	0.8729±0.0336	0.8701±0.0369	0.8539±0.0376	0.8357±0.0278
	BERT-R	0.8634±0.0245	**0.8829±0.0273**	0.8754±0.0286	0.8749±0.0288	0.8754±0.0288	0.8516±0.0235
	BERT-F1	0.8065±0.0127	**0.8846±0.0249**	0.8738±0.0263	0.8721±0.0279	0.8642±0.0290	0.8433±0.0221
	STS	0.1976±0.0911	**0.3539±0.1522**	0.3499±0.1532	0.3481±0.1525	0.3358±0.1516	0.3040±0.1475
	BART	-3.8072±0.8063	**-3.5409±0.7220**	-3.6498±0.7917	-3.6482±0.7935	-3.5736±0.7557	-4.0168±0.7066
	BLEURT	-1.0342±0.0996	-0.5480±0.2967	-0.5052±0.2891	**-0.4972±0.2892**	-0.5924±0.2760	-0.8095±0.2757
Cellphones	BERT-P	0.7658±0.0129	**0.8901±0.0290**	0.8613±0.0504	0.8563±0.0523	0.8541±0.0447	0.8501±0.0247
	BERT-R	0.8522±0.0237	**0.8842±0.0269**	0.8738±0.0304	0.8714±0.0306	0.8717±0.0307	0.8607±0.0253
	BERT-F1	0.8065±0.0114	**0.8869±0.0237**	0.8670±0.0362	0.8632±0.0371	0.8624±0.0334	0.8551±0.0212
	STS	0.2645±0.0892	**0.3864±0.1502**	0.3605±0.1435	0.3369±0.1404	0.3573±0.1427	0.3758±0.1471
	BART	-3.7354±0.7440	**-3.4603±0.6750**	-3.5410±0.7401	-3.5436±0.7427	-3.5803±0.7395	-3.7209±0.7122
	BLEURT	-1.1215±0.0980	-0.5769±0.2861	**-0.5232±0.2761**	-0.5334±0.2712	-0.7300±0.3129	-0.7622±0.2781

text similarity metrics measure semantic similarity without ensuring factual consistency in either precision or recall. This dramatic gap demonstrates that models can be semantically similar while being factually inconsistent: they use similar vocabulary and phrasing while making unsupported or contradictory claims. This underscores the need to revise experimental protocols for factuality-oriented models and develop robust, adapted metrics that properly evaluate factual consistency.

VI. DISCUSSION

Our experimental results reveal several critical insights about the state of factual consistency in text-based explainable recommendation systems.

The Factuality Gap Our experiments uncover a dramatic disconnect between surface-level text quality and factual accuracy. Models achieve impressively high scores on standard similarity metrics (BERTScore F1 ranges from 0.81 to 0.90 across all datasets) suggesting near-human quality text generation. However, when evaluated through our statement-level factual consistency metrics, these same models exhibit strikingly poor performance, with precision scores ranging from merely 4.38% to 32.88%. It suggests that models have learned to generate fluent, contextually appropriate text that *appears* explanatory but frequently fails to ground its claims in verifiable evidence.

Precision versus Recall Trade-offs A consistent pattern across our results is that models struggle with both precision and recall, though often in different ways. The low precision scores indicate frequent hallucination of explanatory content not supported by the evidence. The even lower recall scores reveal that models fail to recover most of the ground-truth explanatory passages, leaving critical aspects of the user's preferences unaddressed.

Cross-Dataset Variability Our evaluation across five product categories reveals substantial domain-specific challenges. The *Toys* dataset consistently yields the lowest factual consistency scores across models, while *Clothes* and *Sports* show relatively better performance. These domain effects underscore the importance of diverse evaluation benchmarks.

The Limitations of Standard Metrics Our results cast serious doubt on the adequacy of existing evaluation practices in explainable recommendation. Standard similarity metrics consistently fail to identify factual inconsistencies. Even more sophisticated metrics designed for factual consistency evaluation in summarization (SummaC, AlignScore, QuestEval) exhibit concerning limitations, suggesting sensitivity to the spe-

TABLE VII
TEXT SIMILARITY METRICS RESULTS ON REVIEW GENERATION
(GREEN DENOTES THE BEST PERFORMANCE; RED DENOTES THE WORST PERFORMANCE)

Dataset	Metric	NRT	Att2Seq	CER	PETER	PEPLER
Toys	BERT-P	0.8012±0.0095	0.8587±0.0301	0.8601±0.0392	**0.8626±0.0374**	0.8549±0.0327
	BERT-R	0.8281±0.0175	0.8356±0.0316	0.8442±0.0251	**0.8443±0.0249**	0.8442±0.0283
	BERT-F1	0.8144±0.0105	0.8467±0.0267	0.8517±0.0278	**0.8530±0.0267**	0.8493±0.0262
	STS	0.3115±0.1104	0.4003±0.1769	0.4495±0.1582	0.4479±0.1582	**0.4810±0.1513**
	BART	-4.2007±0.6486	-3.9546±0.5838	-4.1143±0.6784	-4.1251±0.6772	**-3.8595±0.5553**
	BLEURT	-1.2260±0.0806	-1.0098±0.3625	-0.8877±0.2992	-0.9232±0.3031	**-0.8440±0.2419**
Clothes	BERT-P	0.7618±0.0081	0.8684±0.0295	0.8666±0.0352	0.8706±0.0353	**0.8823±0.0274**
	BERT-R	0.8171±0.0155	0.8430±0.0289	0.8510±0.0250	0.8503±0.0249	**0.8522±0.0257**
	BERT-F1	0.7884±0.0077	0.8552±0.0255	0.8584±0.0261	0.8600±0.0259	**0.8668±0.0229**
	STS	0.1226±0.0622	0.3957±0.1713	0.4492±0.1648	0.4389±0.1663	**0.4647±0.1614**
	BART	-4.4730±0.7239	-4.0133±0.6692	-4.0684±0.7352	-4.0944±0.7368	**-3.9681±0.6453**
	BLEURT	-1.2554±0.0521	-0.8794±0.3600	-0.7534±0.3073	**-0.7390±0.3053**	-0.7459±0.2827
Beauty	BERT-P	0.7649±0.0075	0.8495±0.0311	0.8521±0.0328	0.8545±0.0286	**0.8601±0.0327**
	BERT-R	0.8146±0.0148	0.8340±0.0260	0.8439±0.0231	**0.8444±0.0226**	0.8442±0.0261
	BERT-F1	0.7889±0.0075	0.8414±0.0244	0.8477±0.0242	0.8492±0.0214	**0.8518±0.0256**
	STS	0.0673±0.0628	0.4040±0.1600	0.4665±0.1587	0.4635±0.1544	**0.4899±0.1583**
	BART	-4.5875±0.7004	-3.9275±0.6297	-4.0094±0.6964	-4.0228±0.6942	**-3.8493±0.6139**
	BLEURT	-1.3628±0.0518	-0.7493±0.2862	-0.6844±0.2621	**-0.6368±0.2492**	-0.7033±0.2519
Sports	BERT-P	0.7583±0.0093	0.8539±0.0311	0.8599±0.0344	**0.8626±0.0325**	0.8581±0.0372
	BERT-R	0.8192±0.0146	0.8297±0.0305	0.8361±0.0246	0.8367±0.0240	**0.8385±0.0265**
	BERT-F1	0.7875±0.0075	0.8413±0.0262	0.8475±0.0238	**0.8491±0.0228**	0.8479±0.0281
	STS	0.0294±0.0668	0.2977±0.1879	0.3332±0.1938	0.3269±0.1957	**0.3806±0.1732**
	BART	-4.7838±0.7229	-4.2835±0.6853	-4.4563±0.7518	-4.4658±0.7437	**-4.1956±0.6385**
	BLEURT	-1.3193±0.0555	-1.0357±0.3797	-1.0646±0.3513	-1.0879±0.3605	**-0.8657±0.2614**
Cellphones	BERT-P	0.7690±0.0094	0.8496±0.0359	0.8575±0.0378	0.8541±0.0389	**0.8632±0.0328**
	BERT-R	0.8192±0.0156	0.8364±0.0307	**0.8464±0.0249**	0.8449±0.0258	0.8433±0.0285
	BERT-F1	0.7932±0.0083	0.8426±0.0279	0.8516±0.0276	0.8492±0.0290	**0.8529±0.0274**
	STS	0.1328±0.0867	0.3711±0.1719	0.4076±0.1561	0.3869±0.1601	**0.4234±0.1592**
	BART	-4.6615±0.7447	-4.0652±0.6761	-4.1434±0.7252	-4.1555±0.7281	**-4.0229±0.6607**
	BLEURT	-1.1876±0.0725	-0.9272±0.3365	-0.8290±0.2623	-0.8533±0.2579	**-0.8071±0.2619**

cific granularity and aggregation strategy used. `QuestEval` produces rankings inconsistent with both our LLM-based and NLI-based metrics, potentially due to biases in question generation and answering.

Limitations Our work has several limitations that suggest directions for future research: *(1) LLM-Based Extraction.* Our statement extraction pipeline relies on LLMs, which may introduce errors or biases. While we employ carefully designed prompts and domain-specific topics, the extraction process is not perfect. *(2) Granularity of Ground-truth.* Our rule-based aggregation of statements into explanations preserves all extracted content but may not reflect the natural structure or emphasis users would prefer. Alternative approaches, perhaps learning to select and organize statements based on user preferences or contextual relevance, could yield more realistic ground-truth.

VII. CONCLUSION

This paper presents a comprehensive investigation into the factual consistency of text-based explainable recommendation systems, revealing a critical gap between surface-level text quality and factual accuracy. Through the introduction of a statement-level evaluation framework, augmented benchmark datasets, and novel factuality metrics, we have demonstrated that current state-of-the-art models, despite achieving impressive fluency scores, frequently hallucinate explanatory content and fail to ground their outputs in verifiable evidence. This disconnect underscores a fundamental limitation in current evaluation practices, which prioritize semantic similarity over factual grounding. Second, our experiments reveal that models struggle with both precision (avoiding hallucinations) and recall (covering all relevant explanatory content), with recall scores often falling below precision. This dual deficiency indicates that generated explanations not only contain unsupported claims but also systematically omit critical information present in the source reviews. Our findings suggest that achieving factual consistency in explainable recommendation will require fundamental innovations in model architectures, training objectives, and evaluation.

REFERENCES

[1] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[2] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference*

on research and development in Information Retrieval, 2020, pp. 639–648.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 1, pp. 335–355, 2023.

[5] Y. Zhang, X. Chen *et al.*, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[6] L. Dong, S. Huang, F. Wei, M. Lapata, M. Zhou, and K. Xu, "Learning to generate product reviews from attributes," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 623–632.

[7] P. Li, Z. Wang, Z. Ren, L. Bing, and W. Lam, "Neural rating regression with abstractive tips generation for recommendation," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2017, pp. 345–354.

[8] L. Li, Y. Zhang, and L. Chen, "Personalized transformer for explainable recommendation," *arXiv preprint arXiv:2105.11601*, 2021.

[9] ——, "Personalized prompt learning for explainable recommendation," *ACM Transactions on Information Systems*, vol. 41, no. 4, pp. 1–26, 2023.

[10] Y. Li, X. Zhang, L. Luo, H. Chang, Y. Ren, I. King, and J. Li, "G-refer: Graph retrieval-augmented large language model for explainable recommendation," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 240–251.

[11] Q. Ma, X. Ren, and C. Huang, "Xrec: Large language models for explainable recommendation," *arXiv preprint arXiv:2406.02377*, 2024.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[14] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[15] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[16] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," *Advances in neural information processing systems*, vol. 34, pp. 27 263–27 277, 2021.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[19] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[20] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias, "True: Re-evaluating factual consistency evaluation," *arXiv preprint arXiv:2204.04991*, 2022.

[21] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "Summac: Re-visiting nli-based models for inconsistency detection in summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, 2022.

[22] Y. Zha, Y. Yang, R. Li, and Z. Hu, "Alignscore: Evaluating factual consistency with a unified alignment function," *arXiv preprint arXiv:2305.16739*, 2023.

[23] T. Scialom, P.-A. Dray, P. Gallinari, S. Lamprier, B. Piwowarski, J. Staiano, and A. Wang, "Questeval: Summarization asks for fact-based evaluation," *arXiv preprint arXiv:2103.12693*, 2021.

[24] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 188–197.

[25] T. Herserant and V. Guigue, "Seval-ex: A statement-level framework for explainable summarization evaluation," *arXiv preprint arXiv:2505.02235*, 2025.

[26] J. Raczyński, M. Lango, and J. Stefanowski, "The problem of coherence in natural language explanations of recommendations," in *ECAI 2023*. IOS Press, 2023, pp. 1922–1929.

[27] B. Abdollahi and O. Nasraoui, "Using explainability for constrained matrix factorization," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 79–83.

[28] X. He, T. Chen, M.-Y. Kan, and X. Chen, "Trirank: Review-aware explainable recommendation by modeling aspects," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1661–1670.

[29] Y. Hou, N. Yang, Y. Wu, and P. S. Yu, "Explainable recommendation with fusion of aspect information," *World Wide Web*, vol. 22, no. 1, pp. 221–240, 2019.

[30] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 765–774.

[31] B. Kabongo, V. Guigue, and P. Lemberger, "Elixir: Efficient and lightweight model for explaining recommendations," *arXiv preprint arXiv:2508.20312*, 2025.

[32] Z. Xie, S. Singh, J. McAuley, and B. P. Majumder, "Factual and informative review generation for explainable recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 816–13 824.

[33] L. Li, Y. Zhang, and L. Chen, "Generate neural template explanations for recommendation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 755–764.

[34] Y. Tao, Y. Jia, N. Wang, and H. Wang, "The fact: Taming latent factor models for explainability with factorization trees," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 295–304.

[35] N. Wang, H. Wang, Y. Jia, and Y. Yin, "Explainable recommendation via multi-task learning in opinionated text data," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 165–174.

[36] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 83–92.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[38] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[39] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] H. Cheng, S. Wang, W. Lu, W. Zhang, M. Zhou, K. Lu, and H. Liao, "Explainable recommendation with personalized review retrieval and aspect learning," *arXiv preprint arXiv:2306.12657*, 2023.

[42] R. Shimizu, T. Wada, Y. Wang, J. Kruse, S. O'Brien, S. HtaungKham, L. Song, Y. Yoshikawa, Y. Saito, F. Tsung *et al.*, "Disentangling likes and dislikes in personalized generative explainable recommendation," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 4793–4809.

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[44] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[45] J. Li, Z. He, J. Shang, and J. McAuley, "Ucepic: Unifying aspect planning and lexical constraints for generating explanations in recommendation," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 1248–1257.

[46] J. Ni and J. McAuley, "Personalized review generation by expanding phrases and attending on aspect-aware representations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 706–711.

[47] P. Sun, L. Wu, K. Zhang, Y. Su, and M. Wang, "An unsupervised aspect-aware recommendation model with explanation text generation," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 3, pp. 1–29, 2021.

[48] C.-W. Yang, Z.-Q. Feng, Y.-J. Lin, C.-W. Chen, K.-d. Wu, H. Xu, J.-F. Yao, and H.-Y. Kao, "Maple: Enhancing review generation with multi-aspect prompt learning in explainable recommendation," *arXiv preprint arXiv:2408.09865*, 2024.

[49] P. Bhargava, A. Drozd, and A. Rogers, "Generalization in nli: Ways (not) to go beyond simple heuristics," *arXiv preprint arXiv:2110.01518*, 2021.

[50] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang *et al.*, "Trustllm: Trustworthiness in large language models," *arXiv preprint arXiv:2401.05561*, 2024.

[51] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *arXiv preprint arXiv:2305.14251*, 2023.

[52] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[53] A. Mishra, D. Patel, A. Vijayakumar, X. L. Li, P. Kapanipathi, and K. Talamadupula, "Looking beyond sentence-level natural language inference for question answering and text summarization," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2021, pp. 1322–1336.

[54] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "Qafacteval: Improved qa-based factual consistency evaluation for summarization," *arXiv preprint arXiv:2112.08542*, 2021.

[55] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[56] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[57] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[59] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[60] J. Ni, G. H. Abrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[61] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[63] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

TABLE VIII
PROMPT FOR STATEMENT-TOPIC-SENTIMENT TRIPLET EXTRACTION (CLOTHES DATASET)

You are a review analysis expert tasked with extracting atomic explanatory statements from user reviews for the *Clothing* domain.

Your goal is to identify factual, objective statements that explain why a user liked or disliked specific product topics (from a fixed list), while filtering out noise and purely subjective opinions without explanatory value.

IMPORTANT — use the fixed list of accepted topics below. For extraction, use the short name (lowercase, singular) exactly as provided.

ACCEPTED TOPICS (short name (long name): description)

- fit (Fit and Sizing): Statements about how the item dimensions compare to expected sizes or how it fits the body (runs small/large, length, sleeve fit, recommended sizing).

[...]

TASK REQUIREMENTS

1. Extract atomic explanatory statements: break the review into the smallest possible factual statements that explain product characteristics or user experience, using only factual/explanatory content.

2. Assign topic: for each statement, assign exactly one topic from the ACCEPTED TOPICS list using the short name.

3. Determine sentiment polarity: label each statement as ”positive”, ”negative”, or ”neutral” with respect to the topic.

OUTPUT FORMAT

Return a JSON array (only the JSON array, no surrounding text). Each item must be an object with keys:

- ”statement”: string — the atomic, factual explanatory statement (present tense, minimal and decomposed).

- ”topic”: string — one of the short names from the ACCEPTED TOPICS list.

- ”sentiment”: string — one of ”positive”, ”negative”, or ”neutral”.

RULES FOR ATOMIC STATEMENTS

- Factual and explanatory only: include information that explains product characteristics or behavior (*e.g., ”shrinks after wash”*), exclude pure opinions without explanation (*e.g., ”I love it” — discard*).

- Smallest possible unit: each statement must be atomic and cannot be further decomposed while remaining factual and explanatory.

- Relevant to product evaluation: exclude personal context or unrelated commentary (*e.g., gift recipients, ages*) unless it directly explains a topic (*e.g., height/weight when explaining fit*).

- No extraneous temporal details: remove or omit timing references that are not explanatory of the product (*e.g., drop ”on Tuesday”*); allow minimal time if it explains product behavior (*e.g., ”shrinks after wash” is allowed; ”after two years” is allowed only if it explains durability*).

- Use plain, natural language and present tense.

RULES FOR TOPIC ASSIGNMENT

- Use exactly one topic short name per statement (from the ACCEPTED TOPICS list).

- Use the most general matching topic (e*.g., prefer ”construction” for pilling or seam issues; ”material” for fiber feel*).

- Topic must be lower-case and singular (as in the short names list).

RULES FOR SENTIMENT

- positive: the statement reports a desirable or good outcome for the topic.

- negative: the statement reports a problem, failure, or undesirable outcome for the topic.

- neutral: purely descriptive facts without clear positive/negative valence (e.g., ”is machine washable”).

SPLITTING RULE

- If a single sentence contains multiple explanatory facts about different topics, split them into separate atomic statements and label each with the appropriate topic.

- If a phrase is purely opinion without explanation (*e.g., ”I hate it”, ”it’s great”*), do not include it.

ADDITIONAL NOTES

- Always choose topics from the fixed list above; never invent or substitute new topic names.

- If uncertain between two topics for a factual statement, choose the topic that best describes the product characteristic (*material vs construction vs appearance*).

- When given a review, produce only the JSON array of extracted atomic explanatory statements following the format and rules above.

EXAMPLES
Example 1:
[...]

Please extract atomic explanatory statements following the format above. Return only the JSON array with no additional text or explanation.

TABLE IX
PROMPT FOR LLM-BASED FACTUAL CONSISTENCY EVALUATION

You are a factual verifier. SINGLE TASK: decide whether a STATEMENT is fully supported by at least one passage in the DOCUMENT.

DECISION RULES (apply strictly):

1) Output ”1” only if the DOCUMENT contains a passage that clearly ENTAILS the entire factual content of the STATEMENT (paraphrases/synonyms allowed; if the STATEMENT is negated, the negation must be explicit in the DOCUMENT).

2) If the STATEMENT has multiple sub-claims joined by “and”/commas, each sub-claim must be supported (sub-claims may be supported by different passages). If any sub-claim is unsupported, output ”0”.

3) Use NO knowledge outside the DOCUMENT. If evidence is missing, ambiguous, contradictory, or merely suggestive, output ”0”.

4) Numbers, quantities, dates, and named entities must match (obvious equivalences allowed, e.g., “dozen” = “12”).

5) Ignore off-topic content, opinions without factual content, and metadata without probative value.

6) If the DOCUMENT is empty or unreadable, output ”0”.

STRICT OUTPUT FORMAT:

- Respond with EXACTLY ONE character: ”1” (supported) or ”0” (not supported).

- No explanations, no extra text, no spaces, no quotes, no punctuation, and no additional newlines.

- Do not repeat the question or the statement.

Examples (NEVER reproduce in the output):

[...]

TABLE X
TOYS DATASET - TOPICS

Topic (Long name)	Description
safety (*Safety and Choking Risk*)	Mentions of hazards, small parts, sharp edges, toxic materials, battery-compartment security, or compliance with safety standards.
age (*Age Appropriateness and Skill Level*)	Suitability for the recommended age range and whether the toy’s complexity matches the child’s development or skill level.
durability (*Durability and Build Quality*)	Structural robustness under play, breakage, paint chipping, loose parts, and longevity.
educational (*Educational and Developmental Value*)	Learning or developmental benefits (motor skills, cognitive, STEM, problem solving, creativity).
engagement (*Play Value and Entertainment*)	How engaging the toy is, replayability, attention span it secures, imaginative or social play value.
materials (*Material Quality and Finish*)	Material type and tactile/finish quality (wood/plastic/fabric, texture, paint/finish condition, non-toxic claims).
functionality (*Functionality and Features*)	Whether moving parts, electronics, lights, sounds, batteries, or accessories work as intended.
assembly (*Assembly, Instructions, Packaging, Size and Completeness*)	Ease of setup, clarity of instructions, missing parts/tools, packaging condition on arrival, and physical dimensions vs expectations.
battery (*Power, Battery Life and Charging*)	Battery requirements, battery life/drain, charging behavior, and battery-compartment safety.
price (*Price and Value*)	Perceived cost-effectiveness relative to quality, durability, and comparable alternatives.

TABLE XI
CLOTHES DATASET - TOPICS

Topic (Long name)	Description
fit (*Fit and Sizing*)	How the item dimensions compare to expected sizes or how it fits the body (runs small/large, length, sleeve fit, recommended sizing).
material (*Material and Fabric Quality*)	Fabric type, feel, thickness, fiber content, pilling tendency, or perceived fabric quality.
comfort (*Comfort and Wearability*)	Wearing comfort, itchiness, breathability, stretch, ease of movement, or wearing-related comfort.
appearance (*Appearance and Color Accuracy*)	Look, color accuracy vs. photos, sheen, pattern accuracy, or visual defects.
construction (*Construction and Durability*)	Stitching, seams, zippers, buttons, durability, rips, or structural failures (including pilling and seam failure).
price (*Price and Value*)	Perceived value for money, price fairness, or cost-quality tradeoff.
care (*Care and Maintenance*)	Washing, shrinkage, colorfastness, drying, or care instructions and outcomes.
functionality (Functionality and Features)	Practical features (pockets, closures, hood, lining, insulation, pockets usability).
shipping (*Shipping, Delivery and Packaging*)	Delivery time, packaging condition, or shipping-related problems.
service (*Returns, Refunds and Customer Service*)	Return/refund experience, seller responsiveness, or customer service outcomes.

TABLE XII
BEAUTY DATASET - TOPICS

Topic (Long name)	Description
efficacy (*Efficacy and Performance*)	Whether the product delivers promised results (e.g., hydration, acne reduction, coverage, pigmentation) and measurable outcomes.
compatibility (*Skin/Hair Compatibility and Reactions*)	How the product performs for different skin or hair types and whether it causes irritation, breakouts, or other adverse reactions.
ingredients (*Ingredients and Formulation*)	Composition and formulation details (actives, presence/absence of parabens/sulfates, clean claims) and their role in safety or effectiveness.
texture (*Texture, Feel and Application*)	Sensory and application characteristics (creamy, greasy, sticky, lightweight, blendability, spreadability, pilling).
longevity (*Longevity and Wear Time*)	Staying power and persistence of effects (does it last all day, smudge/fade/crease, need for reapplication).
color (*Color Match, Coverage and Shade Range*)	Color accuracy, undertone, coverage/pigmentation, and availability of suitable shades for diverse tones.
price (*Price and Value*)	Perceived cost-effectiveness relative to size, performance, and alternatives.
packaging (*Packaging Quality and Design*)	Functionality and quality of packaging and applicators (dispensing, hygiene, travel-friendliness).
scent (*Scent and Fragrance*)	Fragrance profile and its acceptability or role in causing reactions (pleasant, overpowering, fragrance-free).
service (*Delivery, Returns and Brand/Customer Service*)	Shipping, packaging on arrival, returns/refunds, seller responsiveness, and overall brand trustworthiness.

TABLE XIII
SPORTS DATASET - TOPICS

Topic (Long name)	Description
performance (*Performance and Effectiveness*)	How well the product performs its intended sport function (power, responsiveness, control, accuracy, shock absorption, moisture-wicking, etc.).
durability (*Durability and Longevity*)	Structural robustness under use, resistance to wear, breakage, seam failure, and long-term longevity.
comfort (*Comfort and Support*)	Fit-related comfort, cushioning, chafing, breathability, ergonomic support during activity.
fit (*Fit and Sizing*)	How sizes, shapes and dimensions match expectations and bodies (true-to-size, need to size up/down) and the effect of fit on performance.
material (*Material Quality and Construction*)	Material types and workmanship (fabric/compound quality, composites vs metals, padding, stitching) that affect feel and longevity.
safety (*Safety and Injury Prevention*)	Protective performance, injury-prevention features, stability, and hazard risks for high-impact use.
portability (*Portability and Storage*)	Weight, packability, carrying convenience, and storage footprint.
usability (*Ease of Use, Setup and Maintenance*)	Ease of assembly, adjustment, operation, maintenance, and clarity of instructions.
price (*Price and Value*)	Perceived value for money — whether cost is justified by performance, durability, and comparable alternatives.
service (*Customer Service and After-sales*)	Brand responsiveness, warranty/returns experience, refunds, and support for issues.

TABLE XIV
CELLPHONES DATASET - TOPICS

Topic (Long name)	Description
performance (*Performance and Responsiveness*)	Real-world speed, app responsiveness, multitasking, thermal throttling and gaming/app performance.
battery (*Battery life and Charging*)	Runtime per charge, charging speed (wired/wireless), charging heat and long-term battery degradation.
camera (*Camera and Imaging*)	Photo and video quality, low-light performance, stabilization, zoom and camera features.
display (*Display, Screen and Touch Quality*)	Brightness, color accuracy, resolution, refresh rate and touch responsiveness.
specs (*Storage, Memory, Connectivity and Network*)	On-device storage and RAM, expandability, cellular/Wi-Fi/Bluetooth performance, and network behavior.
durability (*Durability and Build Quality*)	Physical robustness, scratch/drop resistance, water/dust ratings, and fit/finish.
software (*Software, User Interface and Updates*)	OS experience, UX/shell behavior, bloatware, update cadence and security patching.
usability (*Ease of Use, Compatibility and Accessory Quality*)	Setup and ergonomics, ease of daily use, and accessory interoperability/fit (chargers, cases, protectors).
price (*Price and Value*)	Cost versus features, perceived value-for-money and comparisons to competing models.
service (*Customer Service and Warranty*)	Warranty support, repairs, returns/refunds and vendor responsiveness.