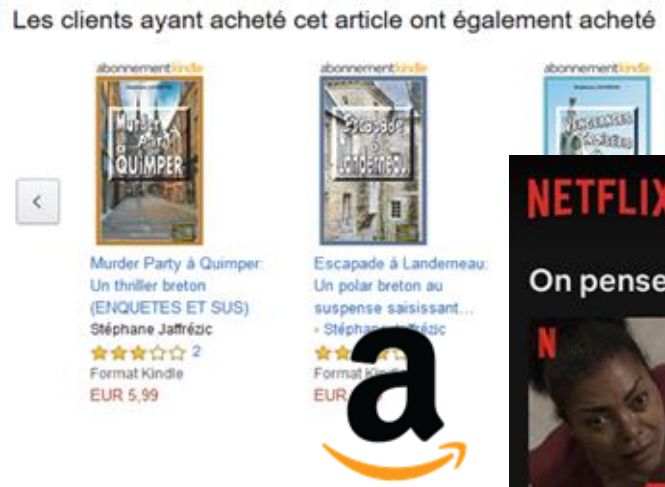
# On the Factual Consistency of Text-based Explainable Recommendation Models

Ben Kabongo, Vincent Guigue

Sorbonne University, CNRS, ISIR, Paris, France
AgroParisTech, UMR MIA Paris-Saclay, Palaiseau, France



# Recommender Systems





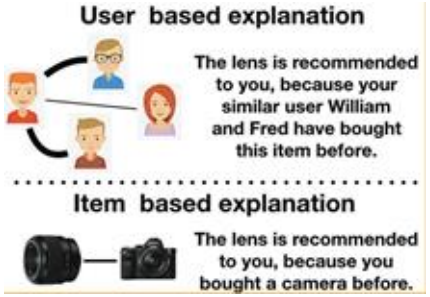

## Recommendation System

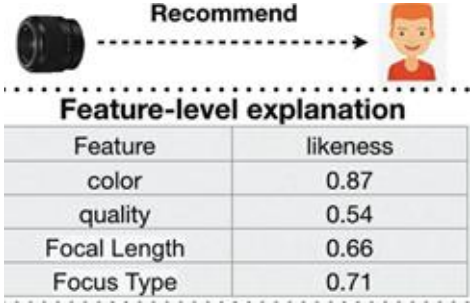
**A personalized way to access information.**

**Goal:** suggest the most relevant content to each user based on their preferences.

# Explainable Recommendation

=> provide insights to explain **why an item is recommended**




| Feature | likeness |
|---|---|
| color | 0.87 |
| quality | 0.54 |
| Focal Length | 0.66 |
| Focus Type | 0.71 |

**Sentence-level explanation**

**Structured: You might be interested in [feature] (can be quality, color, etc), on which this product performs well.**
**Unstructured: Great and deserve the price.**

### Traditional explanation

Based on **similarity** between users or items

***Such explanations are often opaque or lack detail***

### Aspect/Feature-based explanation

User's appraisal on item's specific aspects

***Aspects are not always available.***

### Text-based Explanation

Generating reviews and explanations with language models

# Text-based Explainable Recommendations

**(1) Template-based Explanation:** **Predefined templates**, (*item features* and *opinion words*)

**(2) Free-form Explanation:**

- **Review** (Att2Seq, NRT, PETER, CER, PEPLER)
- **Explanation** (XRec, G-refer)

## RNN-based

**Att2Seq** [Dong 2017]: attention + LSTM to generate reviews from attributes

**NRT** [Li 2017]: multi-task model (overall rating + explanation) tations), GRU-based

## Transformer-based

**PETER** [Li 2021]: multi-task + replaces recurrence with an untrained Transformer

**CER** [Raczynski 2023]: alignment between explanation and rating

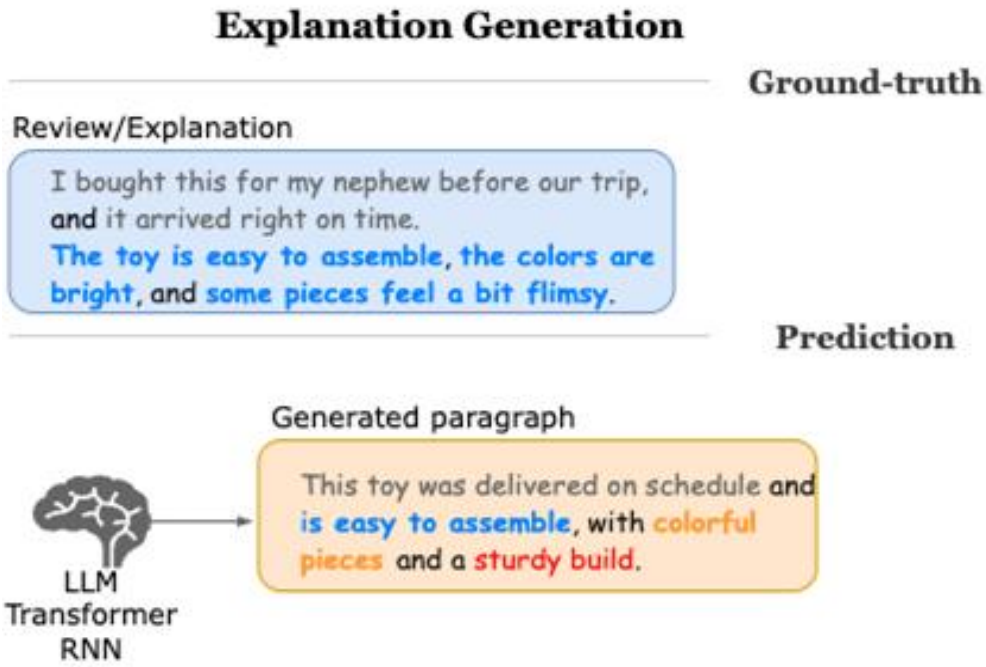
**PEPLER** [Li 2023]: fine-tunes GPT-2 to generate explanations from user and item embeddings

## LLM-enhanced

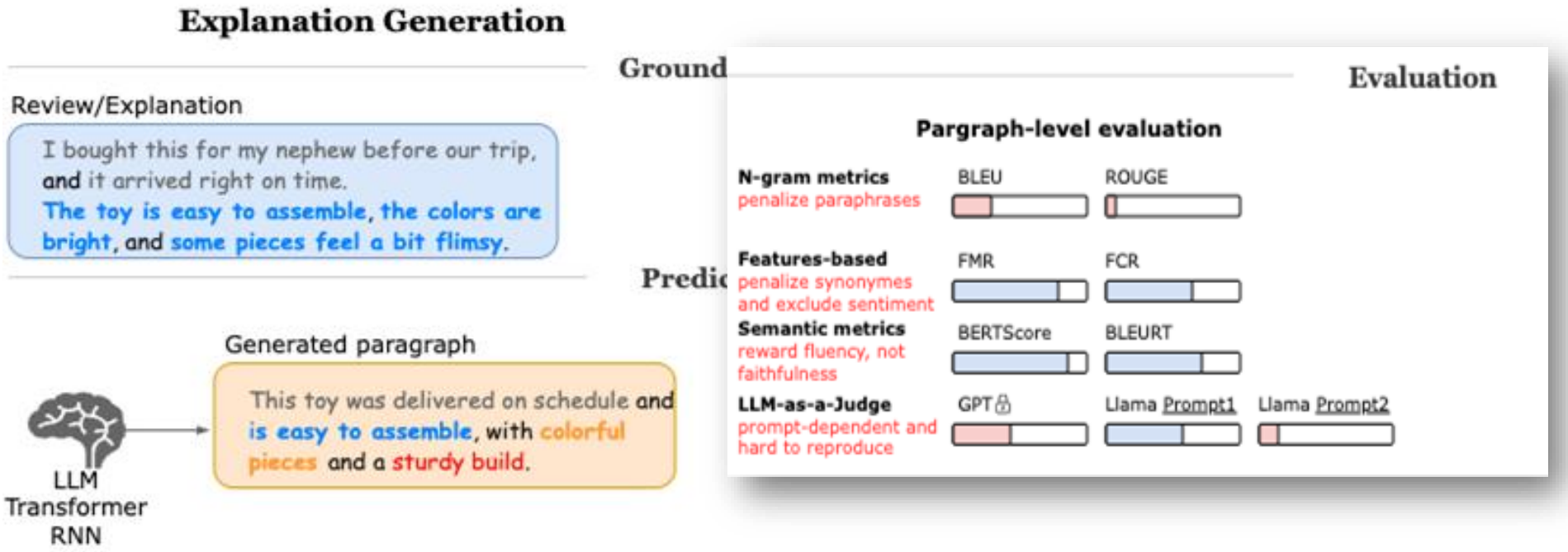
**XRec** [Ma 2024]: lightweight collaborative adapter + injects user and item latent representations into the LLM.

**G-refer** [Li 2025]: graph retrieval (structural + semantic) + LM prompt to guide explanation generation.

# Evaluation of Text-based Explainable Recommenders



# Evaluation of Text-based Explainable Recommenders



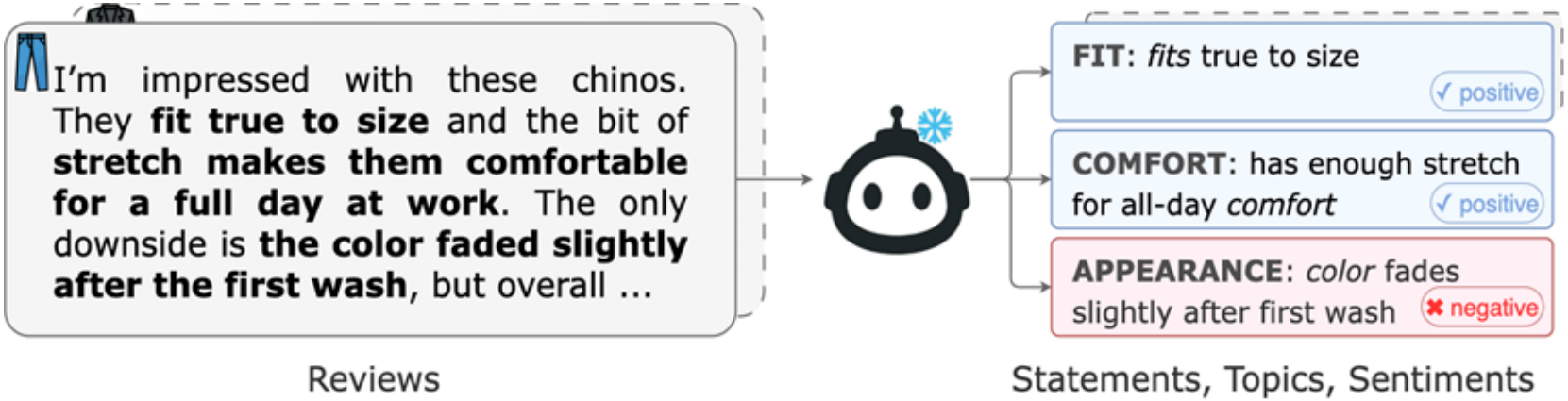
# Factual Consistency?

**Definition (& link to NLI - Natural Language Inference)**

Let two texts **a** and **b**. Text **a** is factually consistent with text **b** if all the information contained in **a** is also contained in **b** and does not contradict **b**.

**A critical question remains largely unexplored:**
**Are the explanations generated by state-of-the-art models factually consistent with the available evidence?**

# Contributions: Statement-Level Ground-truth




**(1) Statement-Level Ground-truth.**
A prompting-based pipeline to extract **atomic explanatory statements from user reviews**.

Hyp : Fine-grained decomposition isolates explanatory content from noise while preserving all relevant information.

# Contributions: Augmented Benchmarks Datasets

| | 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 |

Classical Dataset

Factual GT

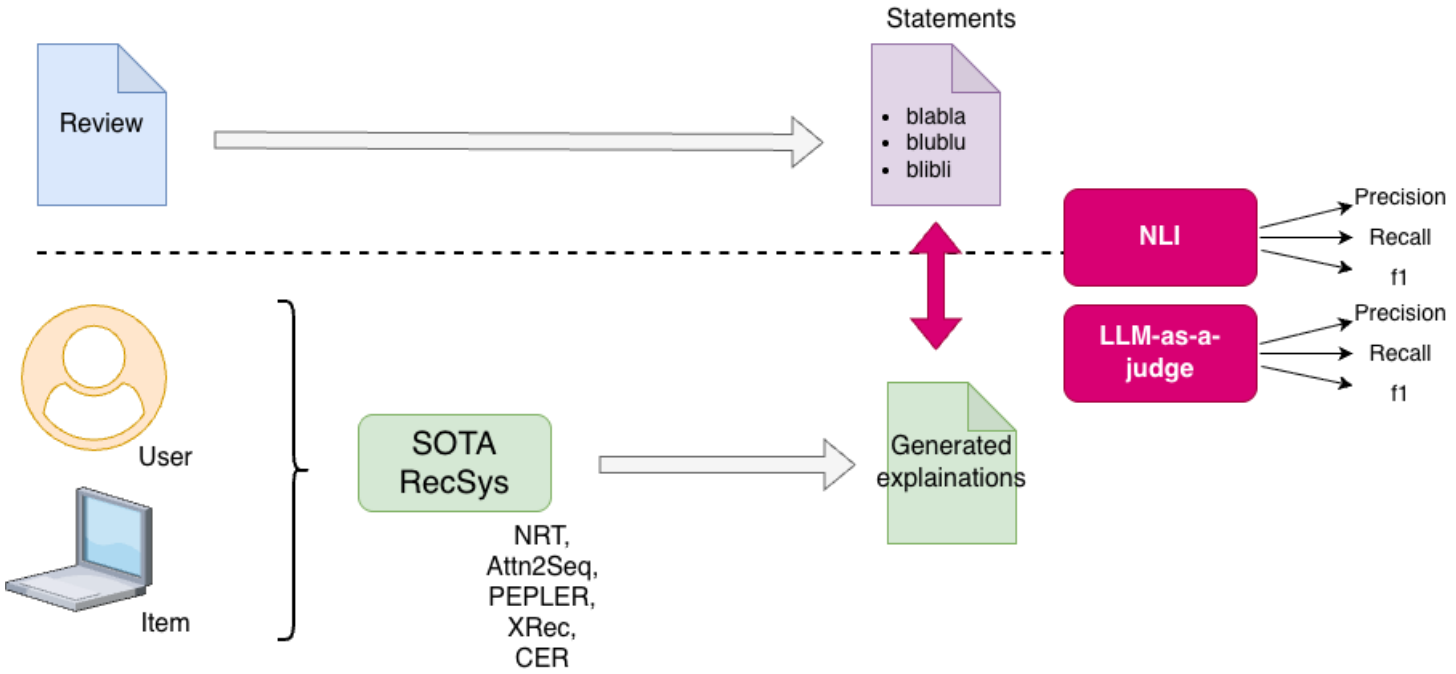
**(2) Augmented Benchmark Datasets.**

=> fine-grained evaluation across domains / foundation for factual consistency.
=> Publicly available

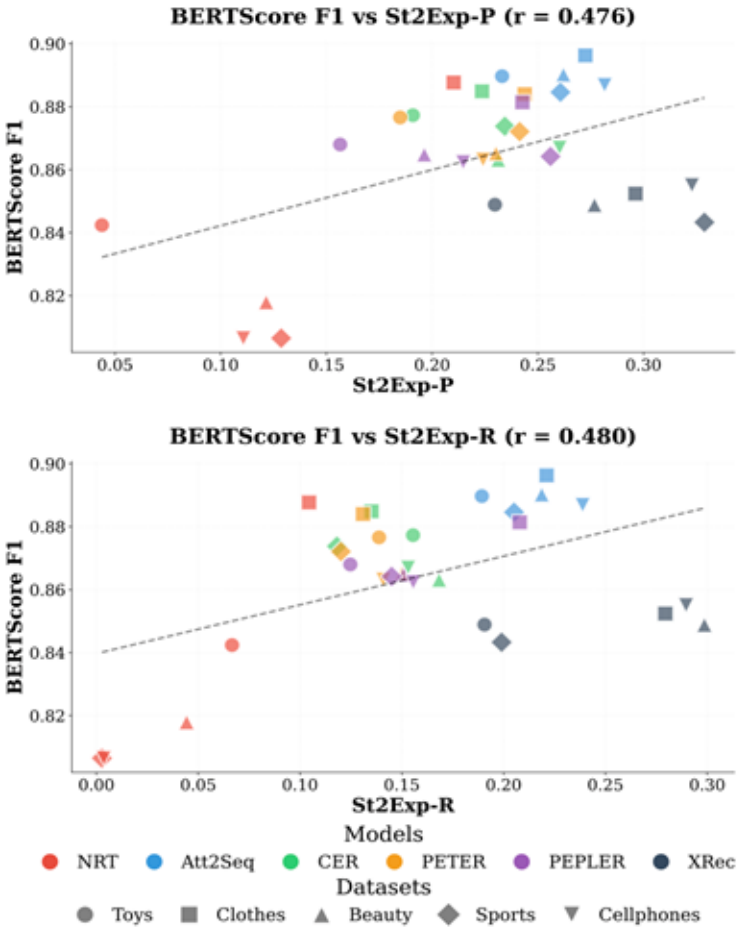
# Contributions: Statement-Level Factuality Metrics

**(3) Statement-Level Factuality Metrics.**

Factual consistency (**precision**) + relevance (**recall**) (& **f1**) at the statement level.



# Metrics Correlation




**Summary**

- Models achieve high scores on standard similarity metrics (BERTScore F1 from **0.81** to **0.90**).
- **Statement-level metrics tell a different story**.
- E.g. LLM-based precision (**4.38%** to **32.88%**) and recall (**0.27%** to **29.86%**) !!

# LLM-based Metrics Results

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** |

**Precision**

- The highest average precision is achieved by XRec on *Sports* with **32.88% (std 33.89%)**
- NRT yields the lowest precision, e.g., **4.38%** on *Toys*
- **Current state-of-the-art models exhibit limited factual consistency**

**Recall**

- Generally even lower than precision: **0.27%** (NRT on *Sports*) - **29.86%** (XRec on *Beauty*)
- **Current models fail to recover most of the ground-truth explanatory passages**

# NLI-based Metrics Results

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 |

**Entailment**

Precision: **4.66%** (NRT on *Toys*) to **24.80%** (NRT on *Cell*), showing substantial *cross-dataset variability*

Recall: NRT achieves only **3.67%** recall on *Cell*
Max: **14.02%** (PEPLER on *Clothes*)

**Coherence (entail. minus contrad.)**

Even XRec yield negative scores on several datasets

**Models can produce statements that directly contradict the reference**

# Discussion

### The Factuality Gap

**Dramatic disconnect** between **surface-level text quality** and **factual accuracy**.

BERTScore F1 ranges from 0.81 to 0.90 across all datasets **suggesting near-human quality text generation**

Statement-level factual consistency metrics, exhibit poor performance (**4.38% to 32.88%**)

### Precision vs Recall Trade-offs

Our results is that **models struggle with both precision and recall**

Low precision = **frequent hallucination of explanatory content**

Low recall scores = **models fail to recover most explanatory passages**

### Limitations

**LLM-Based Extraction.** Our statement extraction pipeline relies on LLMs, which may introduce errors or biases.

**Granularity of Ground-truth**. Our rule-based approach may not reflect the natural structure or emphasis users would prefer.

**(small) Inconsistency between our LLM-based and NLI-based metrics**

# Thank you for your attention!

Paper

Dataset

Code