Sur la Cohérence Factuelle des Modèles de Recommendation Explicables fondés sur le Texte

Ben Kabongo [1] Vincent Guigue [2]
(1) Sorbonne University, CNRS, ISIR, Paris, France
(2) AgroParisTech, UMR MIA Paris-Saclay, Palaiseau, France
ben.kabongo@sorbonne-universite.fr, vincent.guigue@agroparistech.fr

RÉSUMÉ

Les systèmes de recommandation explicables fondés sur le texte génèrent des justifications en langage naturel pour les recommandations d'articles afin d'améliorer la confiance des utilisateurs et la transparence. Cependant, une question cruciale reste peu étudiée : ces explications sont-elles factuellement cohérentes avec les preuves disponibles ? Nous introduisons un cadre d'évaluation complet pour combler cette lacune. Nous proposons un pipeline basé sur le prompting qui exploite des LLM pour extraire, à partir des avis utilisateurs, des énoncés explicatifs atomiques et construire un ensemble de vérité factuellement fondé. En l'appliquant à cinq catégories d'Amazon Reviews, nous créons des benchmarks enrichis permettant une évaluation fine de la qualité des explications. Nous proposons aussi des métriques d'alignement au niveau des énoncés, combinant approches fondées sur les LLM et le NLI. Malgré des scores élevés de similarité sémantique (BERTScore F1 : 0,81–0,90), la factualité reste faible (précision LLM : 4,38 %–32,88 %).

ABSTRACT

On the Factual Consistency of Text-based Explainable Recommendation Models

Text-based explainable recommendation systems generate natural-language justifications for item recommendations to improve user trust and transparency. However, a critical question remains underexplored : are these explanations factually consistent with available evidence ? We introduce a comprehensive evaluation framework addressing this gap. We propose a prompting-based pipeline that leverages LLMs to extract atomic explanatory statements from user reviews, constructing a factually grounded truth set. Applying this pipeline to five Amazon Reviews categories, we create augmented benchmarks enabling fine-grained evaluation of explanation quality. We further propose statement-level alignment metrics combining LLM- and NLI-based approaches to jointly assess factual consistency and relevance. Extensive experiments across six state-of-the-art explainable recommendation models reveal a critical gap : despite high semantic similarity scores (BERTScore F1 : 0.81–0.90), factuality metrics expose alarmingly poor performance (LLM-based precision : 4.38%–32.88%).

MOTS-CLÉS : Recommandation Explicable, Systèmes de Recommandation, Consistence Factuelle, Benchmarking et Métriques d'Évaluation, Grand Modèles de Langue (LLMs), Inférence en Langage Naturel (NLI).

KEYWORDS: Explainable Recommendation, Recommender Systems, Factual Consistency, Benchmarking and Evaluation Metrics, Large Language Models (LLMs), Natural Language Inference (NLI).

1 Introduction

Les systèmes de recommandation sont incontournables, guidant les utilisateurs à travers de vastes catalogues de produits, de contenus et de services. Cependant, les approches traditionnelles (He *et al.*, 2020; Koren *et al.*, 2009) offrent peu d'informations sur les *raisons* pour lesquelles un article particulier est suggéré. La recommandation explicable répond à cette limite en générant des justifications interprétables, améliorant ainsi la transparence, la satisfaction des utilisateurs et la confiance (Zhang *et al.*, 2020).

La recommandation explicable fondée sur le texte s'est imposée comme une approche particulièrement prometteuse, en exploitant la flexibilité du langage naturel pour formuler des justifications personnalisées (Dong *et al.*, 2017; Li *et al.*, 2017, 2021, 2023, 2025; Ma *et al.*, 2024). Les avancées récentes se sont de plus en plus tournées vers les grands modèles de langage (LLM) (Achiam *et al.*, 2023; Dubey *et al.*, 2024). Ces modèles ont démontré des performances impressionnantes sur les métriques standards de génération de texte (Fu *et al.*, 2023; Sellam *et al.*, 2020; Yuan *et al.*, 2021; Zhang *et al.*, 2019), en produisant des explications qui paraissent cohérentes et plausibles au premier abord.

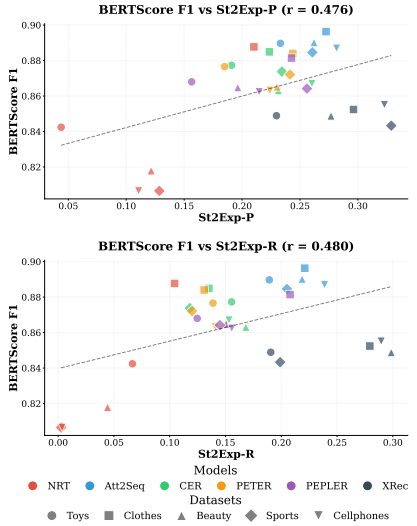

FIGURE 1 – BERTScore F1 comparé aux métriques au niveau des énoncés fondées sur les LLM, précision (en haut) et rappel (en bas). La corrélation de Pearson (r) est indiquée entre parenthèses.

Cependant, une question cruciale demeure largement inexplorée : *les explications générées par les modèles les plus avancés sont-elles factuellement cohérentes avec les preuves disponibles ?* Si la fluidité de surface et le recouvrement lexical sont importants, la véritable valeur d'une explication réside dans sa *factualité* : c'est-à-dire dans la mesure où son contenu reflète fidèlement les préférences réelles de l'utilisateur telles qu'exprimées dans ses avis. Nous introduisons un cadre complet pour évaluer la cohérence factuelle des systèmes de recommandation explicable fondés sur le texte, qui s'articule autour de quatre contributions principales :

(1) Vérité de terrain au niveau des énoncés. Nous concevons un pipeline fondé sur des prompts qui utilise des LLM pour extraire, à partir des avis utilisateurs, des énoncés explicatifs atomiques, accompagnés de leurs thèmes spécifiques au domaine et de leurs étiquettes de sentiment associées. Une procédure d'agrégation à base de règles construit ensuite des explications de référence factuelles.

(2) Jeux de données de référence enrichis. Nous appliquons notre pipeline à cinq catégories d'Amazon Reviews (Ni *et al.*, 2019) (*Toys and Games*, *Clothing*, *Beauty*, *Sports* et *Cellphones*), créant ainsi des jeux de données enrichis qui associent chaque interaction utilisateur–item aux énoncés extraits et aux explications de référence dérivées.

(3) Métriques de factualité au niveau des énoncés. En nous appuyant sur des avancées récentes (Herserant & Guigue, 2025; Laban *et al.*, 2022; Zha *et al.*, 2023), nous proposons des métriques adaptées à la recommandation explicable. En combinant des approches fondées sur les LLM et sur l'inférence en langage naturel (NLI), nos métriques évaluent la cohérence factuelle (précision et rappel) au niveau des énoncés.

(4) Expériences. Nous menons des expériences approfondies sur six modèles de recommandation explicable à l'état de l'art, couvrant trois familles architecturales : les modèles récurrents (NRT (Li *et al.*, 2017), Att2Seq (Dong *et al.*, 2017)), les modèles fondés sur les transformeurs (PETER (Li *et al.*, 2021), CER (Raczyński *et al.*, 2023), PEPLER (Li *et al.*, 2023)) et les modèles enrichis par des LLM (XRec (Ma *et al.*, 2024)).

Notre évaluation met en évidence des écarts importants entre la qualité textuelle de surface et l'exactitude factuelle. Alors que les modèles obtiennent des scores élevés sur les métriques standards de similarité (BERTScore F1 de 0.81 à 0.90), nos métriques au niveau des énoncés racontent une tout autre histoire (Figure 1). En particulier, la précision fondée sur les LLM (`St2Exp-P` de 4.38% à 32.88%) et le rappel (`St2Exp-R` de 0.27% à 29.86%) sont très faibles. Cet écart met en lumière les limites des pratiques d'évaluation actuelles et souligne la prévalence de contenus explicatifs hallucinés dans les systèmes actuels. Nos résultats ont des implications importantes pour la conception et l'évaluation des systèmes de recommandation explicable. Relever ce défi nécessitera de meilleures méthodologies d'évaluation ainsi que des innovations fondamentales dans les architectures de modèles et les objectifs d'apprentissage, afin de privilégier explicitement un ancrage factuel. Nous fournissons notre code et les jeux de données pour supporter la reproductibilité et les travaux futurs. [1]

2 Travaux Connexes

Recommandation explicable fondée sur le texte. La recommandation explicable fondée sur le texte a connu un essor important, en particulier avec l'intégration de modèles de langage, capables de générer des des avis (Dong *et al.*, 2017; Kabongo *et al.*, 2025a,b; Li *et al.*, 2021, 2023; Xie *et al.*, 2023) ou des paragraphes explicatifs dérivés des avis (Li *et al.*, 2025; Ma *et al.*, 2024).

Évaluation des systèmes de recommandation explicable textuelle. Les premiers travaux reposaient principalement sur des métriques à base de n-grammes, tel que BLEU (Papineni *et al.*, 2002) et ROUGE (Lin, 2004). Toutefois, leur dépendance aux correspondances exactes les rend peu adaptées à la synonymie et à la paraphrase. Des métriques plus récentes, sensibles à la sémantique, telles que GPTScore (Fu *et al.*, 2023), BERTScore (Zhang *et al.*, 2019), BARTScore (Yuan *et al.*, 2021) et BLEURT (Sellam *et al.*, 2020), permettent une meilleure évaluation de la qualité textuelle de surface. Néanmoins, elles capturent encore mal la factualité des explications générées (Honovich *et al.*, 2022; Laban *et al.*, 2022; Zha *et al.*, 2023).

Évaluation de la cohérence factuelle. La cohérence factuelle occupe une place centrale dans l'évaluation du résumé automatique (Zha *et al.*, 2023; Honovich *et al.*, 2022) et des LLM (Huang *et al.*, 2024; Min *et al.*, 2023). Après les premières métriques à base de n-grammes (Banerjee & Lavie, 2005; Lin, 2004; Papineni *et al.*, 2002), des approches fondées sur l'inférence en langage naturel (NLI) (Laban *et al.*, 2022; Zha *et al.*, 2023) ou les LLM (Herserant & Guigue, 2025) ont été proposées. Cependant, ces méthodes s'appuient souvent sur des segments textuels trop grossiers pour permettre une analyse fine des explications en recommandation. Dans ce contexte, certaines métriques comme FMR, FCR et DIV (Li *et al.*, 2020) évaluent uniquement la cohérence des caractéristiques d'items par appariement exact, sans prendre en compte le sentiment, tandis que (Xie *et al.*, 2023) mobilise l'entailment sans comparaison explicite à une vérité de terrain. Nous proposons au contraire des métriques au niveau des énoncés, combinant NLI et évaluation par LLM, afin de mesurer finement la factualité en recommandation explicable.

1. https://github.com/BenKabongo25/factual_explainable_recommendation

3 Cadre pour la Recommandation Explicable Factuelle

Les avis utilisateurs mêlent généralement du contenu explicatif, qui justifie la note attribuée, et du bruit, comprenant des éléments non pertinents ou non explicatifs. L'objectif de la recommandation explicable fondée sur le texte est de prédire conjointement la note attribuée par l'utilisateur à un item et de générer une justification textuelle expliquant l'interaction utilisateur–item sous-jacente. Dans cette section, nous présentons notre pipeline de construction des jeux de données ainsi que les métriques utilisées pour évaluer la cohérence factuelle des systèmes de recommandation explicable textuels.

3.1 Extraction des énoncés et construction de la vérité de terrain

Un *énoncé explicatif atomique* est un fait polarisé exprimant l'opinion de l'utilisateur sur un unique attribut ou thème de l'item. Étant donné l'avis $\mathbf{t}_{ui}$ rédigé par l'utilisateur u pour l'item i, notre premier objectif consiste à extraire les énoncés explicatifs atomiques ainsi que leurs thèmes et polarités. Les thèmes permettent notamment de filtrer en amont le bruit non explicatif présent dans les avis. Afin de préserver l'ensemble du contenu explicatif extrait, nous construisons ensuite un paragraphe explicatif à partir de ces triplets au moyen d'une procédure à base de règles, comme illustré dans la Table 1.

Review : Got this sweater as a gift for my sister. *The material feels cheap*. I've been shopping with this brand for years now. *It runs true to size*. *The design is really cute* though ! Would recommend checking their other products.

Ground-truth Explanation : The user would appreciate this product because *it has a really cute design*. However, they may dislike that *the material feels cheap*. They seem indifferent to *it runs true to size*.

TABLE 1 – Illustration de la construction de la vérité de terrain.

Extraction des énoncés. Pour chaque domaine, nous définissons un ensemble de thèmes d'intérêt $\mathcal{T}$. Dans nos expériences, nous concevons d'abord un prompt visant à faire émerger une liste restreinte de thèmes spécifiques au domaine. Étant donné l'ensemble de thèmes $\mathcal{T}$ et l'ensemble d'étiquettes de sentiment $\mathcal{P} = \{\text{POS}, \text{NEG}, \text{NEU}\}$, nous élaborons un prompt spécifique au domaine afin d'extraire, à partir de $\mathbf{t}_{ui}$, un ensemble d'énoncés atomiques et d'associer à chacun son thème et sa polarité : $S_{ui} = \text{LLM}(\mathbf{t}_{ui} \mid \cdot, \mathcal{T}, \mathcal{P}) = \{(\mathbf{s}_1, t_1, p_1), \cdots, (\mathbf{s}_{n_{ui}}, t_{n_{ui}}, p_{n_{ui}})\}$, où $\mathbf{s}_k$ désigne le k-ième énoncé explicatif atomique extrait de $\mathbf{t}_{ui}$, $t_k \in \mathcal{T}$ son thème, et $p_k \in \mathcal{P}$ son étiquette de sentiment.

	Toys	Clothes	Beauty	Sports	Cell
Utilisateurs	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
Énoncés					
Moy/interaction	5.03	4.42	5.45	4.93	4.54
Moy/utilisateur	41.76	30.12	46.99	40.24	30.65
Moy/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

TABLE 2 – Statistiques des jeux de données.

Construction de l'explication de référence. Une fois l'ensemble de triplets S_{ui} obtenu, nous appliquons une procédure à base de règles pour composer un unique paragraphe explicatif regroupant tous les énoncés extraits. Nous commençons par regrouper les énoncés selon leur polarité ; pour chaque polarité présente, nous formons une phrase agrégeant les énoncés correspondants. Les phrases obtenues sont ensuite concaténées à l'aide de connecteurs logiques simples afin de produire un paragraphe structuré. Cette approche évite le coût d'un recours supplémentaire à un LLM tout en garantissant la conservation de tous les énoncés extraits. La Table 1 en donne un exemple.

3.2 Jeux de données

Nous appliquons notre pipeline à cinq catégories du jeu de données Amazon Reviews 2014 [2] (Ni *et al.*, 2019) : **Toys and Games** (*Toys*), **Clothing, Shoes and Jewelry** (*Clothes*), **Beauty** (*Beauty*), **Sports and Outdoors** (*Sports*) et **Cell Phones and Accessories** (*Cellphones*). Dans nos expériences, nous utilisons `Llama-3-8B-Instruct` (Dubey *et al.*, 2024). Pour chaque jeu de données, nous définissons dix thèmes d'intérêt et élaborons un prompt spécifique, accompagné de quelques exemples illustratifs, afin d'extraire tous les triplets explicatifs de chaque interaction. À partir de chaque avis et des énoncés extraits, nous construisons ensuite l'explication de référence correspondante. Les statistiques des jeux de données sont présentées dans la Table 2.

3.3 Métriques d'évaluation

Un bon système de recommandation explicable doit générer des explications dont tous les passages sont appuyés par la référence (précision), tout en couvrant autant que possible les passages explicatifs de cette référence (rappel). Nous introduisons donc un ensemble de métriques mesurant la cohérence factuelle, en précision et en rappel, au niveau des énoncés. Étant donnés m énoncés $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$ extraits de l'explication de référence $\mathbf{e}$ et n énoncés $\{\mathbf{s}'_1, \ldots, \mathbf{s}'_n\}$ extraits de l'explication générée $\mathbf{e}'$, nous définissons deux familles de métriques : des métriques fondées sur les LLM et des métriques fondées sur la NLI.

Métriques fondées sur les LLMs. Ces métriques reposent sur une fonction de score f_{LLM} évaluant la cohérence factuelle d'un énoncé par rapport à une unité textuelle cible. Nous définissons ainsi la *Statement-to-Explanation Precision* (`St2Exp-P`), la *Statement-to-Explanation Recall* (`St2Exp-R`) et la *Statement-to-Explanation F1* (`St2Exp-F1`). Elles suivent toutes le cadre SEval (Herserant & Guigue, 2025) et sont définies par :

$$\texttt{St2Exp-P} = \frac{1}{n}\sum_{k=1}^{n} f_{\text{LLM}}(\mathbf{s}'_k, \mathbf{e}), \quad \texttt{St2Exp-R} = \frac{1}{m}\sum_{l=1}^{m} f_{\text{LLM}}(\mathbf{s}_l, \mathbf{e}'),$$
$$\texttt{St2Exp-F1} = 2\,\frac{\texttt{St2Exp-P}\cdot\texttt{St2Exp-R}}{\texttt{St2Exp-P}+\texttt{St2Exp-R}}. \tag{1}$$

Métriques fondées sur la NLI. Nos métriques fondées sur la NLI utilisent une fonction de score par entailment f_{NLI} pour évaluer la cohérence factuelle entre paires d'énoncés. Nous considérons deux variantes de cette fonction $f_{\text{NLI}}(\mathbf{s}_k, \mathbf{s}_l)$. Dans la première, le score correspond à la probabilité d'entailment E_{kl}, notée $f_{\text{NLI-ent}}$. Dans la seconde, le score correspond à la différence entre entailment et contradiction, $E_{kl} - C_{kl}$; nous l'appelons *coherence score* et la notons $f_{\text{NLI-coh}}$. Ces mesures ne sont pas symétriques. Nous définissons alors des métriques orientées précision (`StEnt-P` et `StCoh-P`), orientées rappel (`StEnt-R` et `StCoh-R`) ainsi qu'une moyenne harmonique (`StEnt-F1`) :

$$\texttt{St}\star\texttt{-P} = \frac{1}{n}\sum_{k=1}^{n} \max_{l} f_{\text{NLI-}\star}(\mathbf{s}'_k, \mathbf{s}_l), \quad \texttt{St}\star\texttt{-R} = \frac{1}{m}\sum_{l=1}^{m} \max_{k} f_{\text{NLI-}\star}(\mathbf{s}_l, \mathbf{s}'_k),$$
$$\texttt{St}\star\texttt{-F1} = 2\,\frac{\texttt{St}\star\texttt{-P}\cdot\texttt{St}\star\texttt{-R}}{\texttt{St}\star\texttt{-P}+\texttt{St}\star\texttt{-R}}, \tag{2}$$

où $\star \in \{\text{ent}, \text{coh}\}$ désigne la fonction de score retenue.

2. https://jmcauley.ucsd.edu/data/amazon/links.html

4 Expériences

4.1 Protocole expérimental

Baselines. Nous considérons trois familles de modèles à l'état de l'art dans notre évaluation : les modèles fondés sur les RNN (Att2Seq (Dong *et al.*, 2017) et NRT (Li *et al.*, 2017)), les modèles fondés sur les transformeurs (PETER (Li *et al.*, 2021), CER (Raczyński *et al.*, 2023) et PEPLER (Li *et al.*, 2023)) ainsi que les modèles fondés sur les LLM (XRec (Ma *et al.*, 2024)). Pour XRec (Ma *et al.*, 2024), nous utilisons `Llama-2-7b` [3] (Touvron *et al.*, 2023), conformément à l'article original. Pour chaque modèle, nous retenons les meilleurs hyperparamètres rapportés dans les travaux correspondants.

Jeux de données. Nous évaluons les modèles sur cinq catégories du jeu de données Amazon Reviews (Ni *et al.*, 2019) : *Toys*, *Clothes*, *Beauty*, *Sports* et *Cellphones*. Les détails de construction et les statistiques des jeux de données sont présentés dans la Section 3.2.

Métriques d'évaluation. Nous évaluons tous les modèles à l'aide de l'ensemble de métriques introduit dans la Section 3.3. Pour les métriques fondées sur les LLM, nous utilisons `Llama-3.1-8B-Instruct` [4] (Dubey *et al.*, 2024), et pour les métriques fondées sur la NLI, `DeBERTa-large-mnli` [5] (He *et al.*, 2021). Pour tous les modèles, l'entraînement est effectué sur l'ensemble d'apprentissage, la sélection de modèle sur l'ensemble de validation, et nous rapportons pour chaque métrique la moyenne et l'écart-type calculés sur l'ensemble de test.

4.2 Résultats en similarité textuelle

Génération d'explications. La Figure 2 (haut) présente l'évaluation des modèles à l'aide des métriques de similarité sémantique couramment utilisées dans les travaux récents sur la recommandation explicable, en comparant les sorties des modèles aux explications de référence dérivées des énoncés. Les résultats révèlent des tendances différentes de celles observées dans des travaux antérieurs évaluant des explications plus courtes. XRec se classe parmi les modèles les moins performants sur ces métriques, étant même dépassé par NRT sur *Clothes*. Att2Seq obtient les meilleurs résultats sur la plupart des métriques et des jeux de données, suivi de près par CER, tandis que PEPLER atteint systématiquement les meilleurs scores en BLEURT.

Génération d'avis. La Figure 2 (bas) présente les performances des modèles (hors XRec) en génération d'avis à l'aide de métriques de similarité textuelle. Les résultats sont cohérents avec ceux rapportés dans PEPLER et PETER, confirmant que les modèles fondés sur les transformeurs génèrent des avis plus proches des avis réels que leurs prédécesseurs fondés sur les RNN. Toutefois, cette proximité ne garantit pas la factualité. L'écart entre les résultats en génération d'explications et en génération d'avis suggère que les modèles excellent soit dans l'une, soit dans l'autre tâche, mais pas nécessairement dans les deux. De plus, comme les avis contiennent à la fois du bruit et du contenu explicatif, une forte similarité, qu'elle porte sur les avis ou sur les explications, ne garantit pas une génération factuelle.

3. https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
4. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
5. https://huggingface.co/microsoft/deberta-large-mnli

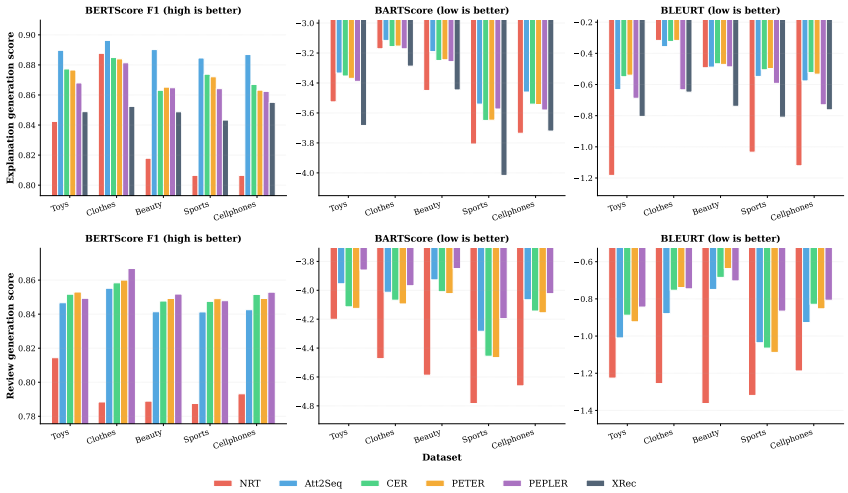

FIGURE 2 – Résultats de similarité textuelle sur la génération d'explications (en haut) et la génération d'avis (en bas).

4.3 Évaluation au niveau des énoncés fondée sur les LLM (St2Exp)

La Table 3 présente les résultats obtenus avec nos métriques de cohérence factuelle fondées sur les LLM. Globalement, les scores restent faibles, aussi bien en précision qu'en rappel.

Précision. La meilleure précision moyenne est obtenue par XRec sur *Sports*, avec 32.88% (écart-type : 33.89%) ; bien qu'il s'agisse du meilleur résultat, ce score reste modeste, voire problématique dans des contextes où la précision est essentielle. Dans l'ensemble, `St2Exp-P` montre que les systèmes à l'état de l'art présentent une faible précision en génération d'explications. À l'inverse, NRT obtient les plus faibles scores, par exemple 4.38% sur *Toys*, et se montre faible sur l'ensemble des jeux de données et des métriques.

Rappel. `St2Exp-R` est généralement encore plus faible que la précision, ce qui confirme que les modèles actuels ne parviennent pas à retrouver la majorité des passages explicatifs de référence. Cette limite est particulièrement préoccupante dans les scénarios où une couverture complète est requise, ou lorsque certains passages sont particulièrement importants pour les utilisateurs. XRec atteint le meilleur rappel avec 29.86% sur *Beauty*, tandis que le plus faible est observé pour NRT avec 0.27% sur *Sports*. Enfin, `St2Exp-F1` confirme cette tendance, avec des valeurs allant de 0.05% (NRT sur *Cellphones*) à 19.13% (XRec sur *Clothes*).

Comparaison avec les métriques de similarité textuelle. La comparaison entre les métriques de similarité textuelle et nos métriques de factualité fondées sur les LLM met en évidence un décalage frappant : malgré des scores très faibles en précision et en rappel factuels, les modèles obtiennent des scores très élevés en similarité, avec un BERTScore F1 compris entre 0.81 et 0.90 (Figure 1). Cet écart montre que des modèles peuvent produire des explications sémantiquement proches des évidences tout en restant factuellement incohérents : ils mobilisent un vocabulaire et des formulations similaires, tout en produisant des affirmations non fondées, voire contradictoires. Cela souligne la nécessité de revoir les protocoles expérimentaux orientés factualité et de développer des métriques adaptées à l'évaluation rigoureuse de la cohérence factuelle.

	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**

TABLE 3 – Résultats de l'évaluation au niveau des énoncés fondée sur les LLMs.

4.4 Évaluation au niveau des énoncés fondée sur la NLI (StEnt/StCoh)

La Table 4 présente les résultats de nos métriques fondées sur la NLI, qui calculent des scores d'entailment et de contradiction entre paires d'énoncés. Les scores globaux restent faibles, ce qui confirme les observations issues des métriques fondées sur les LLM, tout en produisant des classements de systèmes quelque peu différents.

Entailment. Les scores de précision (`StEnt-P`) varient de 4.66% (NRT sur *Toys*) à 24.80% (NRT sur *Cellphones*), ce qui révèle une forte variabilité selon les jeux de données. Cette précision peut être artificiellement élevée pour des modèles générant peu d'énoncés lorsque ceux-ci correspondent à des motifs fréquents dans les données. Les scores de rappel (`StEnt-R`) confirment cette interprétation : malgré une précision plus élevée, NRT n'atteint que 3.67% de rappel sur *Cellphones*. De manière générale, les rappels restent faibles, avec un maximum de 14.02% pour PEPLER sur *Clothes*.

Cohérence. Les métriques `StCoh-*` évaluent la cohérence à partir de la différence entre entailment et contradiction. Des valeurs négatives indiquent que la contradiction domine l'entailment. Même des modèles récents comme XRec obtiennent des précisions et rappels négatifs sur plusieurs jeux de données, et seuls quelques modèles, comme PEPLER sur *Clothes*, atteignent des scores positifs, mais modestes. Ces résultats montrent qu'au-delà d'une faible cohérence factuelle, les modèles peuvent produire des énoncés qui contredisent directement la référence.

Résultats fondés sur les LLM vs fondés sur la NLI. Dans l'ensemble, nous observons un écart entre les métriques fondées sur les LLM et celles fondées sur la NLI, ce qui s'explique par leur granularité différente. Les métriques NLI comparent des paires d'énoncés, tandis que les métriques fondées sur les LLM évaluent chaque énoncé par rapport à l'explication complète. Si les modèles NLI permettent des comparaisons par paires efficaces à grande échelle, là où une approche LLM serait trop coûteuse, l'approche fondée sur les LLM préserve mieux le contexte complet de référence lors de l'évaluation

	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

TABLE 4 – Résultats de l'évaluation au niveau des énoncés fondée sur la NLI.

de chaque énoncé. Nous observons également que les métriques fondées sur les LLM produisent des résultats plus cohérents et plus stables que les métriques fondées sur la NLI. Néanmoins, les deux familles de métriques conduisent à la même conclusion : les modèles présentent une faible cohérence factuelle.

5 Discussion

Nos résultats expérimentaux mettent en évidence plusieurs enseignements importants sur l'état actuel de la cohérence factuelle dans les systèmes de recommandation explicable fondés sur le texte.

L'écart de factualité. Nos expériences révèlent un décalage marqué entre la qualité textuelle de surface et l'exactitude factuelle. Les modèles obtiennent des scores très élevés sur les métriques classiques de similarité, ce qui pourrait laisser penser à une génération presque humaine. Pourtant, lorsqu'ils sont évalués avec nos métriques de cohérence factuelle au niveau des énoncés, leurs performances chutent fortement. Cela suggère que les modèles ont appris à produire un texte fluide et contextuellement plausible, qui *semble* explicatif, sans pour autant appuyer systématiquement ses affirmations sur des éléments vérifiables.

Compromis entre précision et rappel. Un constat récurrent de nos résultats est que les modèles peinent à la fois en précision et en rappel, bien que de manière différente. Les faibles scores de précision indiquent une hallucination fréquente de contenu explicatif non soutenu par les données. Les scores de rappel, souvent encore plus faibles, montrent que les modèles ne retrouvent qu'une faible partie des passages explicatifs de référence, laissant de côté des aspects importants des préférences de l'utilisateur.

Limites. Notre travail présente plusieurs limites qui ouvrent des perspectives de recherche futures. *(1) Extraction fondée sur les LLM.* Notre pipeline d'extraction des énoncés repose sur des LLM, qui peuvent introduire des erreurs ou des biais. Malgré l'usage de prompts soigneusement conçus et de vérifications qualitatives manuelles régulières, ce processus reste imparfait. Des travaux futurs pourraient explorer des méthodes d'extraction plus robustes, éventuellement avec validation humaine dans les applications sensibles. *(2) Granularité de la vérité de terrain.* Notre agrégation à base de règles préserve l'ensemble du contenu extrait, mais ne reflète pas nécessairement la structure ni l'accentuation qu'un utilisateur jugerait naturelles. Des approches alternatives, capables d'apprendre à sélectionner et organiser les énoncés selon les préférences des utilisateurs ou la pertinence contextuelle, pourraient produire des références plus réalistes.

6 Conclusion

Cet article propose une étude approfondie de la cohérence factuelle des systèmes de recommandation explicable fondés sur le texte, et met en évidence un écart critique entre qualité textuelle de surface et exactitude factuelle. Grâce à l'introduction d'un cadre d'évaluation au niveau des énoncés, de jeux de données enrichis et de nouvelles métriques de factualité, nous montrons que les modèles actuels à l'état de l'art, malgré leur grande fluidité, hallucinent fréquemment du contenu explicatif et ancrent insuffisamment leurs sorties dans des éléments vérifiables. Ce décalage souligne une limite fondamentale des pratiques actuelles d'évaluation, qui privilégient la similarité sémantique au détriment de l'ancrage factuel. Nos résultats suggèrent que l'amélioration de la cohérence factuelle en recommandation explicable nécessitera des avancées de fond dans les architectures de modèles, les objectifs d'apprentissage et les protocoles d'évaluation.

Références

ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

BANERJEE S. & LAVIE A. (2005). 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*, p. 65–72.

DONG L., HUANG S., WEI F., LAPATA M., ZHOU M. & XU K. (2017). 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*, p. 623–632.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., YANG A., FAN A. *et al.* (2024). The llama 3 herd of models. *arXiv e-prints*, p. arXiv–2407.

FU J., NG S.-K., JIANG Z. & LIU P. (2023). Gptscore : Evaluate as you desire. *arXiv preprint arXiv :2302.04166*.

HE P., LIU X., GAO J. & CHEN W. (2021). Deberta : Deberta : Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

HE X., DENG K., WANG X., LI Y., ZHANG Y. & WANG M. (2020). 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*, p. 639–648.

HERSERANT T. & GUIGUE V. (2025). Seval-ex : A statement-level framework for explainable summarization evaluation. *arXiv preprint arXiv :2505.02235*.

HONOVICH O., AHARONI R., HERZIG J., TAITELBAUM H., KUKLIANSY D., COHEN V., SCIALOM T., SZPEKTOR I., HASSIDIM A. & MATIAS Y. (2022). True : Re-evaluating factual consistency evaluation. *arXiv preprint arXiv :2204.04991*.

HUANG Y., SUN L., WANG H., WU S., ZHANG Q., LI Y., GAO C., HUANG Y., LYU W., ZHANG Y. *et al.* (2024). Trustllm : Trustworthiness in large language models. *arXiv preprint arXiv :2401.05561*.

KABONGO B., GUIGUE V. & LEMBERGER P. (2025a). Elixir : Efficient and lightweight model for explaining recommendations. *arXiv preprint arXiv :2508.20312*.

KABONGO B., GUIGUE V. & LEMBERGER P. (2025b). Prédiction des préférences et génération de revue personnalisée basées sur les aspects et attention. In *Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA)*, p. 151–170.

KOREN Y., BELL R. & VOLINSKY C. (2009). Matrix factorization techniques for recommender systems. *Computer*, **42**(8), 30–37.

LABAN P., SCHNABEL T., BENNETT P. N. & HEARST M. A. (2022). Summac : Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, **10**, 163–177.

LI L., ZHANG Y. & CHEN L. (2020). Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, p. 755–764.

LI L., ZHANG Y. & CHEN L. (2021). Personalized transformer for explainable recommendation. *arXiv preprint arXiv :2105.11601*.

LI L., ZHANG Y. & CHEN L. (2023). Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, **41**(4), 1–26.

LI P., WANG Z., REN Z., BING L. & LAM W. (2017). 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*, p. 345–354.

LI Y., ZHANG X., LUO L., CHANG H., REN Y., KING I. & LI J. (2025). G-refer : Graph retrieval-augmented large language model for explainable recommendation. In *Proceedings of the ACM on Web Conference 2025*, p. 240–251.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text summarization branches out*, p. 74–81.

MA Q., REN X. & HUANG C. (2024). Xrec : Large language models for explainable recommendation. *arXiv preprint arXiv :2406.02377*.

MIN S., KRISHNA K., LYU X., LEWIS M., YIH W.-T., KOH P. W., IYYER M., ZETTLEMOYER L. & HAJISHIRZI H. (2023). Factscore : Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv :2305.14251*.

NI J., LI J. & MCAULEY J. (2019). 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)*, p. 188–197.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

RACZYŃSKI J., LANGO M. & STEFANOWSKI J. (2023). The problem of coherence in natural language explanations of recommendations. In *ECAI 2023*, p. 1922–1929. IOS Press.

SELLAM T., DAS D. & PARIKH A. P. (2020). Bleurt : Learning robust metrics for text generation. *arXiv preprint arXiv :2004.04696*.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

XIE Z., SINGH S., MCAULEY J. & MAJUMDER B. P. (2023). Factual and informative review generation for explainable recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, p. 13816–13824.

YUAN W., NEUBIG G. & LIU P. (2021). Bartscore : Evaluating generated text as text generation. *Advances in neural information processing systems*, **34**, 27263–27277.

ZHA Y., YANG Y., LI R. & HU Z. (2023). Alignscore : Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv :2305.16739*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

ZHANG Y., CHEN X. *et al.* (2020). Explainable recommendation : A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, **14**(1), 1–101.