Prédiction de préférences et génération de revue personnalisée basées sur les aspects et l'attention

AURA: Aspect-based Unified Ratings Prediction and Personalized Review Generation with Attention

Auteurs : Ben Kabongo, Vincent Guigue & Pirmin Lemberger

MLIA SORBONNE UNIVERSITÉ

02 juillet 2025

Systèmes de recommandation

Systèmes de recommandation
Moyen personnalisé d'accès à l'information.
Objectif : suggérer à chaque utilisateur les contenus les plus pertinents selon ses préférences.

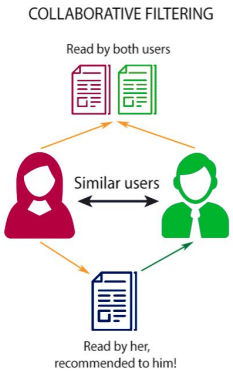
Systèmes de recommandation

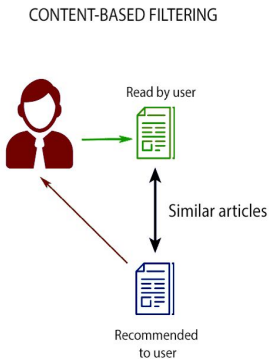
Données manipulées par les systèmes de recommandation
Items/Articles/Produits : éléments qui sont recommandés
Utilisateurs : personnes/entités à qui les recommandations sont faites

Transactions : données d'interaction de l'utilisateur avec les items
- ***Explicite :*** *note (entre 1 et 5), revue (commentaire textuel)*
- ***Implicite :*** *clicks, temps de visionnage*

Exemple de note et de revue utilisateur (tiré de tripadvisor.com)

Approches classiques de recommandation





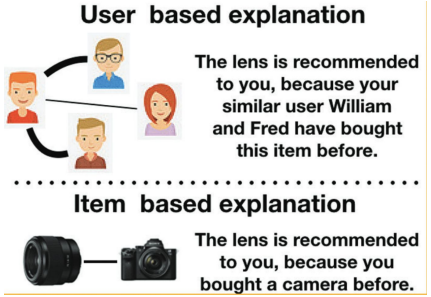
Filtrage collaboratif (CF)
Factorisation matricielle : MF
[Koren et al. 2009, He et al. 2017]
❗ ***Cold start (démarrage à froid)***

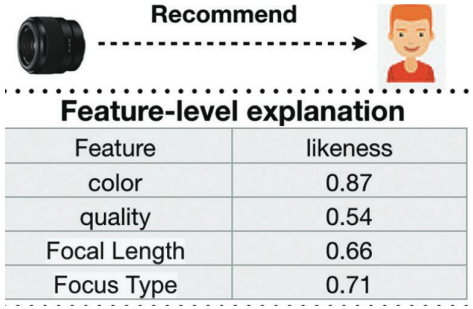
Approches basées sur le contenu (CB)
[Aggrawal et al. 2016]
❗ ***Diversité***

Approches hybrides
Prise en compte de la revue
DeepCoNN [Zheng et al. 2017]
✅ ***Gain de performance***

❗ **Profils utilisateurs peu interprétables**
❗ **Recommandations difficiles à expliquer**

Explicabilité des recommandations





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.

Explication traditionnelle

Basée sur la similarité entre utilisateurs ou items [Zhang et al. 2020]

Analyse des sessions et séquentialité

❗ ***Peu claires ou peu détaillées***

Explication basée sur les aspects

Revue : opinions sur les aspects d''intérêt
Exemple : service, propreté, localisation d'un hôtel

Recommandation aspect-based : décomposition du profil utilisateur en profils d'aspects [Chin 2018, Cheng 2018, Sun 2021]

Explication textuelle

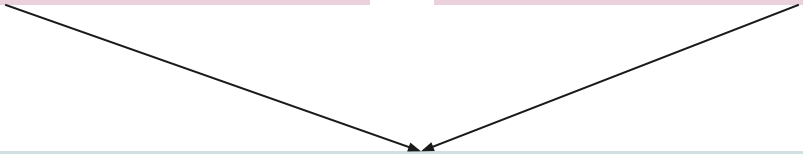
État de l'art :
Génération de revues/textes explicatifs avec des **LLMs**. [Li 2023, Ma 2024]

❗ **Textes générés parfois non factuels ou non alignés avec les préférences réelles**

Objectifs

Approches classiques de recommandation
❗ **Profils utilisateurs peu interprétables**
❗ **Recommandations difficiles à expliquer**

Explication textuelle
❗ **Textes générés parfois non factuels ou non alignés avec les préférences réelles**



Construire des profils utilisateurs plus interprétables pour guider la génération d'explications textuelles avec des LLMs

Notations et Formulation du problème

Exemple de note globale, de notes d'aspects et de revue utilisateur (tiré de tripadvisor.com)

Notations

$R = \{(u, i, r_{ui}, t_{ui}, \{r^a_{ui}\}a \in A)\}$

u: utilisateur

i: item

a: aspect

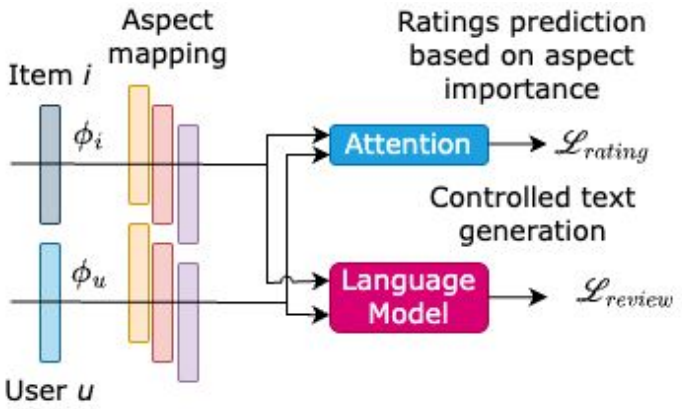
A: ensemble d'aspects

r_{ui}: note globale

r^a_{ui}: note de l'aspect a

t_{ui}: revue

Notations et Formulation du problème



Vue d'ensemble schématique de notre approche

Hypothèse
Les **aspects** permettent de mieux guider la génération de la revue [Sun 2021]

Objectifs :

- **Recommandation** : prédiction de la note globale
- **Explicabilité** : prédiction des notes d'aspects et génération de la revue

Proposition : le modèle AURA

- **Décomposition des profils en profils d'aspects**
- **Attention personnalisée** pour estimer l'importance des aspects
- Un **module pour la prédiction de notes** et un **module pour la génération de revue**

Module de prédiction de notes

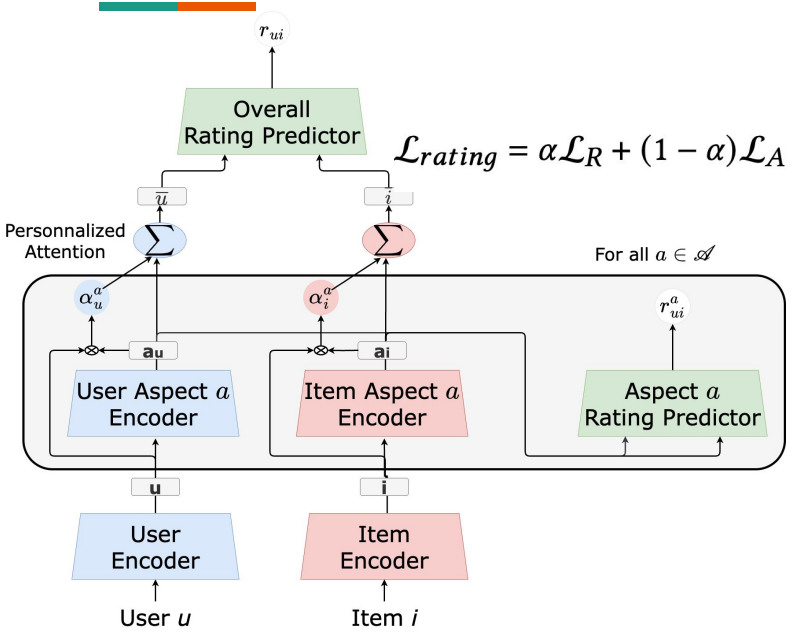


Schéma du module de prédiction de notes du modèle AURA

Décomposition des profils en profils d'aspects

$$\mathbf{a}_u = \phi_{\mathcal{U}}^a(\mathbf{u}), \quad \mathbf{a}_i = \phi_{\mathcal{I}}^a(\mathbf{i})$$

Attention personnalisée [Vaswani 2017] pour pondérer les aspects selon leur importance

$$\alpha_u^a = \frac{\exp(q_{\mathcal{U}}(\mathbf{u})^{\mathrm{T}} k_{\mathcal{U}}(\mathbf{a}_u))}{Z}, \quad \alpha_i^a = \frac{\exp(q_{\mathcal{I}}(\mathbf{i})^{\mathrm{T}} k_{\mathcal{I}}(\mathbf{a}_i))}{Z}$$

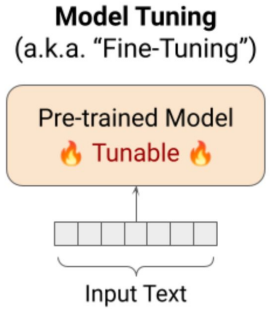
$$\tilde{\mathbf{u}} = \sum_{a \in \mathcal{A}} \alpha_u^a v_{\mathcal{U}}(\mathbf{a}_u), \quad \tilde{\mathbf{i}} = \sum_{a \in \mathcal{A}} \alpha_i^a v_{\mathcal{I}}(\mathbf{a}_i)$$

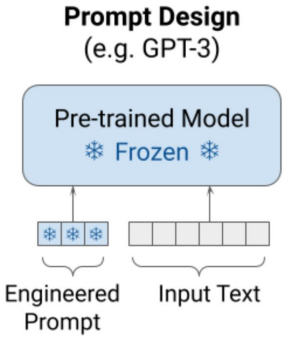
Prédiction de note globale et des notes des aspects

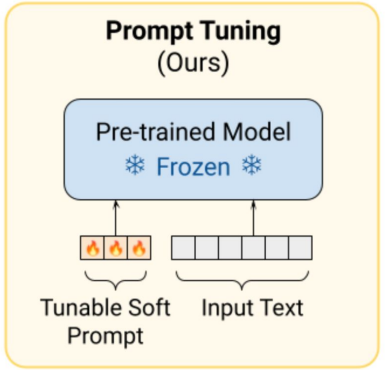
$$\hat{r}_{ui} = f(\tilde{\mathbf{u}}, \tilde{\mathbf{i}}) \quad \hat{r}_{ui}^a = g_a(\mathbf{a}_u, \mathbf{a}_i)$$

✅ **Explicabilité** : importance (attention) et notes d'aspects

Prompt tuning







Fine-tuning
Génération de texte : $P_{\theta_{LM}}(Y|X)$
X, Y: textes ; θ_{LM}: LLM pré-entraîné (GPT, Llama, T5)

Fine-tuning : spécialiser θ_{LM} pour chaque tâche

Prompting
$P_{\theta_{LM}}(Y|[P, X])$

P: prompt dépendant de θ_{LM}

Approche manuelle et non différentiable

Prompt tuning [Lester et al. 2021]
$P_{\theta_P, \theta_{LM}}(Y|[P, X])$
θ_P : 🔥 appris par tâche (params du prompt)
θ_{LM} : ❄ figé pour toutes les tâches

P: prompt dépendant uniquement de θ_P

Approche automatique et différentiable

Module de génération de revue

$$\mathcal{L}_{review} = -\frac{1}{|\mathcal{R}|}\sum_{(u,i)\in\mathcal{R}}\frac{1}{|t_{ui}|}\sum_{k=1}^{|t_{ui}|}\log P_{\theta_P,\theta_{LM}}(y_k \mid \mathbf{p}_{ui}, y_{<k})$$

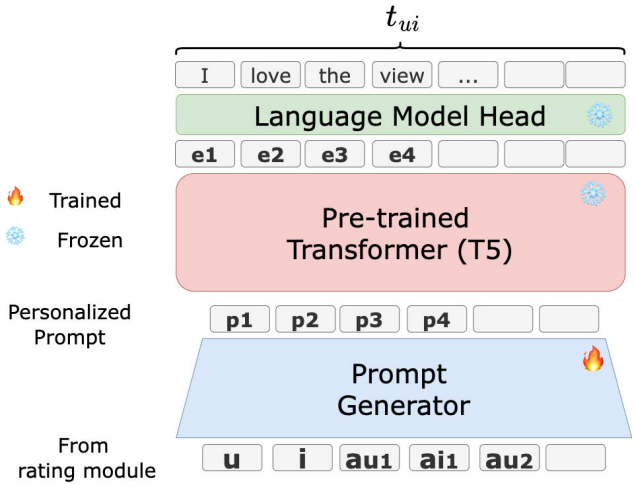


Schéma du module de génération de revue du modèle AURA

Génération de revues en recommandation

État de l'art : apprentissage des profils utilisateurs et items (2 tokens) pour guider la génération avec un LLM [Li 2021, Li 2023]

❗ **Les LLMs génèrent parfois des textes inventés et incohérents**

Proposition : Génération factuelle de revues personnalisées

Hypothèse : les **aspects** permettent de mieux guider la génération de la revue [Sun 2021]

Meta-prompt :

1- Génération d'un prompt personnalisé sur la base des profils d'aspects (apprentissage via prompt-tuning [Lester 2021])

$$\mathbf{p}_{ui} = \psi\left(\mathbf{u}, \mathbf{i}, \{\mathbf{a}_u, \mathbf{a}_i\}_{a\in\mathcal{A}}\right)$$

2- Génération des revues sur la base de ce prompt

$$P_{\theta_P,\theta_{LM}}\left(t_{ui}|\mathbf{p}_{ui}\right)$$

✅ **Approche frugale** : expérimentations avec un petit modèle de langue, T5-Small (60M Params.) [Raffel 2020]

Ablations

AURA -Attention : Ablation de l'attention personnalisée

Remplace l'attention pour l'agrégation des aspects par du max pooling

$$(\tilde{\mathbf{u}})_j = \max_{a \in \mathcal{A}} (\mathbf{a}_u)_j, \quad (\tilde{\mathbf{i}})_j = \max_{a \in \mathcal{A}} (\mathbf{a}_i)_j$$

AURA -Aspects : Ablation de la modélisation des aspects

Omet l'apprentissage des profils d'aspects => pas de note d'aspects, pas d'attention

$$\hat{r}_{ui} = f(\mathbf{u}, \mathbf{i}), \quad \mathbf{p}_{ui} = \psi(\mathbf{u}, \mathbf{i})$$

AURA -Global : Ablation des représentations aspectuelles

N'apprend que les représentations globales (u et i), pour l'ensemble des tâches

$$\hat{r}_{ui} = f(\mathbf{u}, \mathbf{i}), \quad \hat{r}^a_{ui} = g_a(\mathbf{u}, \mathbf{i}),$$

$$\mathbf{p}_{ui} = \psi(\mathbf{u}, \mathbf{i}).$$

Protocole Expérimental

Jeux de données

TripAdvisor (Hôtels) - Aspects: Cleanliness, Location, Service, Sleep, Rooms, Value
8K utilisateurs, 2K items, 62K revues/notes

RateBeer (Bières) - Aspects: Appearance, Aroma, Palate, Taste
8K utilisateurs, 5K items, 201K revues/notes

Modèles de référence

Prédiction de notes
Classiques : Average, MF Koren 2009, MLP, NeuMF He 2017
Aspect-based : ALFM Chin. 2018, ANR Cheng 2018
Multi-tâches : NRT, PETER, PEPLER

Génération de revues
RNN-based : Att2Seq Dong 2017, NRT Li 2017
Transformer-based : PETER Li 2021
Transformer pré-entraîné : PEPLER Li 2023 (GPT-2)
❗ **Négligent la modélisation des aspects**

Métriques

Métriques classiques
RMSE, MAE

Mesures de la qualité de la génération
BLEU, ROUGE, METEOR, BERTScore

Prédiction de la note globale

Table des résultats de la prédiction de note globale

	TripAdvisor		RateBeer	
Model	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Average	0.9325	0.6458	0.5711	0.4249
MF	0.8409	0.6463	**0.4114**	**0.3008**
MLP	0.8332	0.5656	0.4648	0.3244
NeuMF	0.8408	0.5702	0.4731	0.3295
ALFM	0.8967	0.6912	0.4335	0.3142
ANR	0.8473	0.6075	0.4231	0.3084
NRT	0.8592	0.5481	0.4208	0.3066
PETER	0.8078	0.5327	0.4156	**0.3008**
PEPLER	0.7792	0.4782	0.4305	0.3059
AURA	**0.7482**	**0.4477**	0.4166	0.3050
-Attention	0.7716	0.5132	0.4217	0.3111
-Global	0.8651	0.6320	0.4439	0.3326

AURA se classe parmi les meilleurs modèles pour la prédiction de note globale

Prédiction des notes des aspects

Table des résultats de la prédiction des notes des aspects
Nous reportons la moyenne et l'écart-type des métriques sur l'ensemble des aspects

	TripAdvisor		RateBeer	
Model	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Average	1.014 (0.0879)	0.8014 (0.0572)	0.6054 (0.0117)	0.4893 (0.0231)
AURA	**0.7532 (0.0811)**	**0.4514 (0.0538)**	**0.4657 (0.0347)**	**0.3540 (0.0307)**
-Attention	0.7851 (0.0736)	0.5541 (0.0514)	0.4866 (0.0316)	0.3731 (0.0303)
-Global	0.8607 (0.0760)	0.6313 (0.0561)	0.4950 (0.0359)	0.3832 (0.0318)

AURA et ses ablations prédisent mieux les notes des aspects que le modèle basé sur la moyenne par aspect

AURA explique également les recommandations à travers les notes des aspects

Intégration des aspects : AURA -Attention > AURA -Global

- Apprendre des représentations aspectuelles en plus des représentations globales pour capturer finement les préférences

Attention personnalisée : AURA > AURA -Attention

- L'attention personnalisée permet de mieux agréger l'information des aspects

Génération de revue

Table des résultats de la génération de revue

TripAdvisor	METEOR ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERT-P ↑	BERT-R ↑	BERT-F1 ↑
Att2Seq	18.6113	04.6900	28.7839	06.4736	18.5239	85.3487	83.6769	84.4902
NRT	17.2198	03.4053	25.8336	05.1943	17.5390	82.8282	81.5335	82.1613
PETER	17.9550	03.9435	27.9742	05.9062	18.2520	85.0379	83.8235	84.4064
$\text{PEPLER}_{\text{GPT-2}}$	24.3400	11.4000	33.8312	11.6797	22.4529	82.6355	84.9450	83.7264
$\text{AURA}_{\text{T5-Small}}$	**42.7527**	**33.5446**	**53.2856**	**37.8780**	**44.0538**	**90.6867**	**88.4785**	**89.5549**
-Aspects	27.6423	10.0294	39.0768	21.9701	29.5941	88.0013	85.1939	86.5400
RateBeer	METEOR ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERT-P ↑	BERT-R ↑	BERT-F1 ↑
Att2Seq	18.6113	04.6900	28.7839	06.4736	18.5239	85.3487	83.6769	84.4902
NRT	24.9634	08.7375	32.5892	11.4721	26.6292	85.0467	82.9921	83.9859
PETER	28.8189	11.5183	35.5043	13.6200	29.6688	87.3401	85.6216	86.4486
$\text{PEPLER}_{\text{GPT-2}}$	28.2665	10.1432	32.4444	11.1827	26.2481	84.0207	86.0634	84.9906
$\text{AURA}_{\text{T5-Small}}$	**40.7637**	**24.1609**	**46.3715**	**25.8183**	**39.4616**	**90.4830**	**89.1356**	**89.7921**
-Aspects	32.6755	13.6520	39.0688	17.1069	32.4644	89.3652	87.3239	88.3102

AURA obtient des meilleures scores que l'ensemble des modèles de référence sur toutes les métriques considérées

Efficacité de notre architecture frugale : L'ablation basée sur les aspects surpasse également tous les autres modèles

Importance de l'intégration des aspects : AURA performe mieux que sa version ablatée sur les aspects

Impact du nombre de tokens du prompt

Table des résultats sur l'impact du nombre de tokens du prompt

η	METEOR ↑	BLEU ↑	ROUGE-2 ↑
PEPLER	24.3400	11.4000	11.6797
2	12.1594	01.2821	04.4362
5	16.7304	03.5462	06.0468
10	21.1600	07.6928	10.3301
20	<u>29.3798</u>	<u>17.0189</u>	<u>20.2004</u>
50	**42.7527**	**33.5446**	**37.8780**

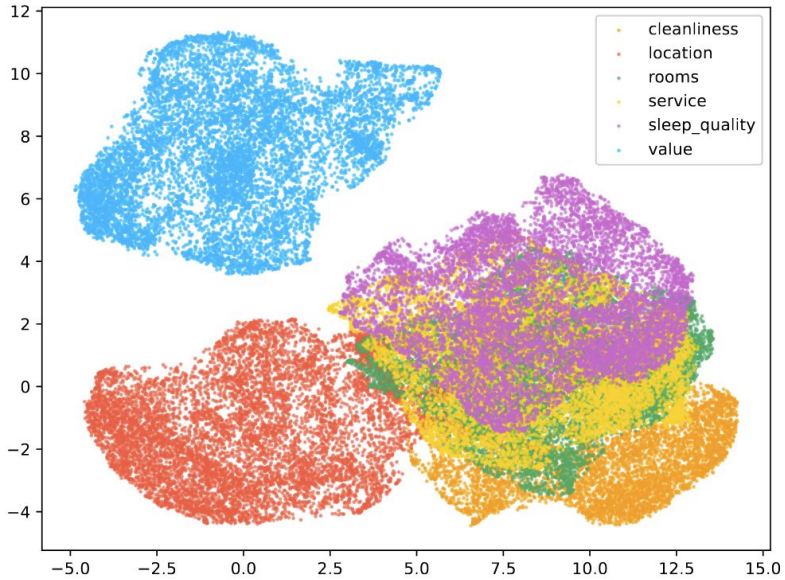
À partir de 20 tokens, AURA surpasse PEPLER sur la génération de revue

Les modèles de référence ne considèrent généralement que 2 tokens (utilisateur + item)

Efficacité de notre architecture : AURA apprend un prompt personnalisé encapsulant l'ensemble des informations sur les utilisateurs, les items et les aspects

Modélisation des aspects

Figure : Projection des représentations des aspects des utilisateurs (TripAdvisor)



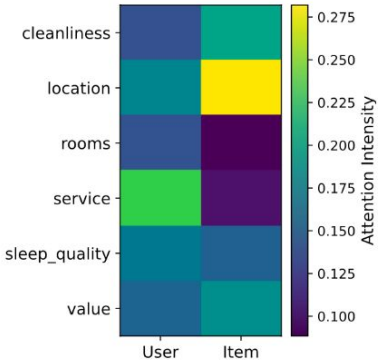
Observation d'une séparation sémantique cohérente des aspects

(Cleanliness, Rooms, Service, Sleep quality) & Location & Value

Intégration des aspects : amélioration des performances sur l'ensemble des tâches (prédiction des notes et génération de revue)

Attention personnalisée

Figure : Visualisation de l'attention sur les aspects et de l'alignement avec la revue (TripAdvisor)



Aspect	Rating	Ground truth review
Cleanliness	4.9 (5.0)	if we go back to paris, we are staying here again. the place is so charming and overlooks the beautiful luxembourg gardens. the staff were sooo hospitiable. always asking what they could do to help us. they arranged two tours for us, recommended places to eat and then made the reservations for us, arranged transportation from and to the airport, etc. royce and xavier, i can't thank you enough ! also, so many places are in walking distance, like notre dame and the louvre. you can't help but fall in love with this place !
Location	5.0 (5.0)	
Rooms	5.0 (5.0)	
Service	5.0 (5.0)	
Sleep	5.0 (5.0)	
Value	5.0 (5.0)	
Overall	4.9 (5.0)	

La comparaison des poids d'attention avec le contenu de la revue révèle un fort alignement entre les préférences de l'utilisateur et l'importance des aspects déduite par AURA

Attention personnalisée : déduit l'importance relative des aspects pour chaque utilisateur et chaque item, permettant également d'expliquer les recommandations

Conclusion

AURA
Modèle multi-tâche (prédiction de notes et génération de revue)

Intègre l'information des aspects via du prompt tuning pour mieux guider la génération de revue

Explique les recommandations par la revue, les aspects et l'attention

Expérimentations et analyses
Utilisation d'un LLM relativement petit : T5-Small

Le modèle se classe parmi les meilleurs sur l'ensemble des tâches => efficacité de notre approche frugale

En particulier, sur la génération de revue, AURA surpasse les modèles de référence

Limitations
Annotations en aspects : AURA repose sur des jeux de données annotés en aspects

Factualité : Un pas vers la lutte contre les hallucinations, encore présentes

Architecture : T5 est moins spécialisé que des LLMs plus récents => borne les performances

Travaux Futurs

Extraction d'aspects

Dans la réalité, les jeux de données ne sont pas souvent annotés en aspects

Proposer des techniques d'extraction d'annotations : *ABSA en few-shot avec un LLM*

Vers d'autres architectures et des explications plus factuelles

Passer de T5 (encoder-decoder) à des architectures decoder-only

Étudier des pistes et des architectures qui favorisent la factualité : *memory network pour la prédiction d'extraits de revues*

Dialogue utilisateur-système

L'explicabilité ouvre la voie la transparence et le contrôle utilisateur

Redonner la main à l'utilisateur pour qu'il comprenne et modifie son profil, en dialoguant avec le système