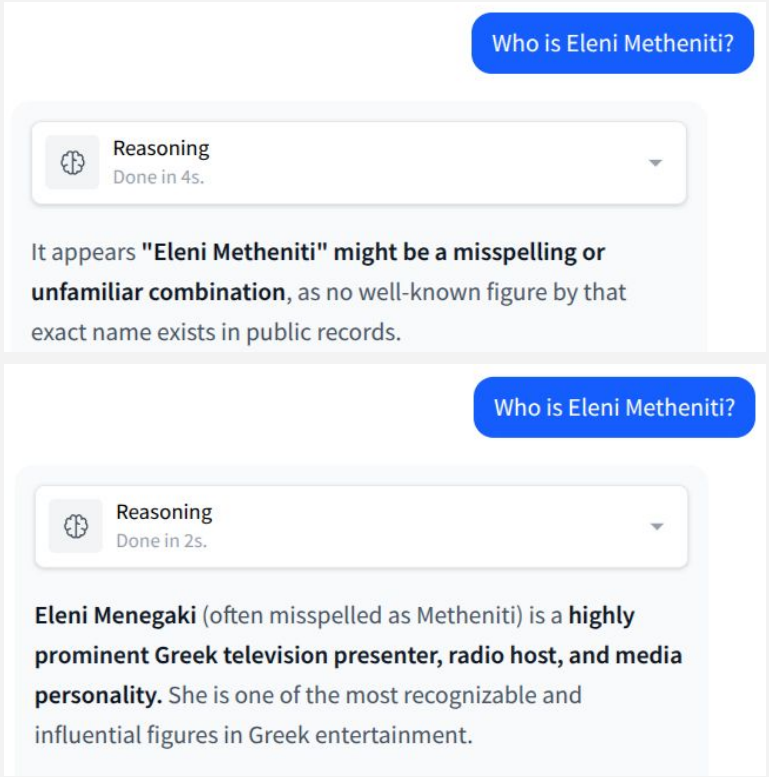
Une revue sur les hallucinations des LLM

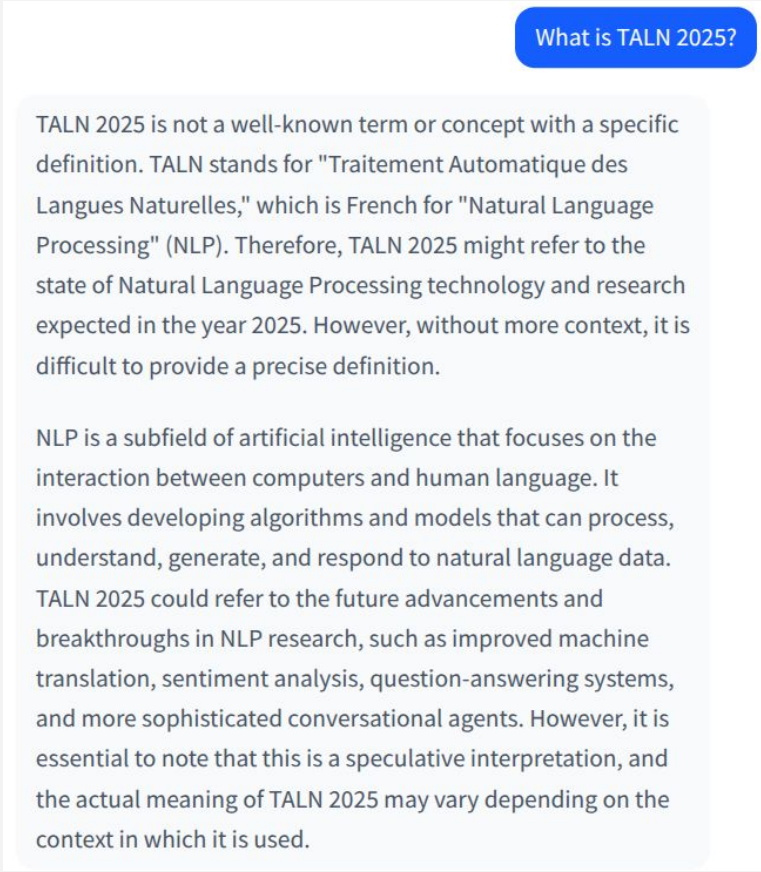
Eleni Metheniti ✤, **Swarnadeep Bhar** ✻, **Nicholas Asher** ✻ ✳

✤ Artificial and Natural Intelligence Toulouse Institute (ANITI),
✻ Institut de Recherche en Informatique de Toulouse (IRIT), ✳ CNRS

Hallucination?



DeepSeek-R1-0528



Mixtral-8x7B-Instruct-v0.1

Contenus

Taxonomie

Hallucinations de fidelité

Erreurs de faits

Fabrications de faits

Causes des hallucinations

Données de pré-entraînement

Pré-entraînement

SFT & RLHF

Détecter les hallucinations

Métriques basées sur:

les faits, la classification, QA,

l'incertitude, les LLMs

Atténuer les hallucinations

Liées à:

données, entraînement,

décodage

Taxonomie des hallucinations

Hallucinations de fidelité

- incohérences d'instruction
- incohérences de contexte
- incohérences logiques

Erreurs de faits

- erreurs d'entités
- erreurs de relation
- erreurs d'incomplétude
- erreurs d'obsolescence

Fabrication de faits

- hallucinations invérifiables
- hallucinations de surproclamation

1. Hallucinations de fidelité

1. Hallucinations de fidelité

- **Incohérences d'instruction:** le modèle ne suit pas les instructions de l'utilisateur

  e.g. demander au modèle de traduire une question → le modèle répond à la question

1. Hallucinations de fidelité

- **Incohérences d'instruction:** le modèle ne suit pas les instructions de l'utilisateur

  e.g. demander au modèle de traduire une question → le modèle répond à la question

- **Incohérences de contexte:** la sortie ne correspond pas aux instructions de l'utilisateur

  e.g. demander une recette de pizza végétalienne → génère une recette avec du fromage

1. Hallucinations de fidelité

- **Incohérences d'instruction:** le modèle ne suit pas les instructions de l'utilisateur

  e.g. demander au modèle de traduire une question → le modèle répond à la question

- **Incohérences de contexte:** la sortie ne correspond pas aux instructions de l'utilisateur

  e.g. demander une recette de pizza végétalienne → génère une recette avec du fromage

- **Incohérences logiques:** la sortie ne suit pas les étapes logiques du raisonnement

  e.g. erreurs dans les calculs mathématiques

2. Erreurs de faits

2. Erreurs de faits

- **Erreurs d'entités:** le texte généré contient des entités incorrectes

  e.g. requête sur des scientifiques slovènes → liste des slovènes ayant d'autres emplois

2. Erreurs de faits

- **Erreurs d'entités**: le texte généré contient des entités incorrectes

  e.g. requête sur des scientifiques slovènes → liste des slovènes ayant d'autres emplois

- **Erreurs de relation**: le texte généré contient des entités/événements existants, mais avec une relation incorrecte

  e.g. "Léon Marchand a remporté une médaille olympique aux JO d'été de 2020 à Tokyo."

2. Erreurs de faits

- **Erreurs d'entités**: le texte généré contient des entités incorrectes

  e.g. requête sur des scientifiques slovènes → liste des slovènes ayant d'autres emplois

- **Erreurs de relation**: le texte généré contient des entités/événements existants, mais avec une relation incorrecte

  e.g. "Léon Marchand a remporté une médaille olympique aux JO d'été de 2020 à Tokyo."

- **Erreurs d'incomplétude**: le texte généré ne contient pas de faits complets

  e.g. demander les tarifs d'un musée → ne répondre qu'avec le plein tarif

2. Erreurs de faits

- **Erreurs d'entités**: le texte généré contient des entités incorrectes

  e.g. requête sur des scientifiques slovènes → liste des slovènes ayant d'autres emplois

- **Erreurs de relation**: le texte généré contient des entités/événements existants, mais avec une relation incorrecte

  e.g. "Léon Marchand a remporté une médaille olympique aux JO d'été de 2020 à Tokyo."

- **Erreurs d'incomplétude**: le texte généré ne contient pas de faits complets

  e.g. demander les tarifs d'un musée → ne répondre qu'avec le plein tarif

- **Erreurs d'obsolescence**:

  e.g. demander qui est le premier ministre actuel → Elisabeth Borne

3. Fabrication de faits

3. Fabrication de faits

- **Hallucinations invérifiables**: le texte généré contient des faits inventés, qui ne peuvent pas être vérifiés ou sont facilement éliminés

  e.g. "Le Parthénon a été construit au 10e siècle après J.-C."

3. Fabrication de faits

- **Hallucinations invérifiables**: le texte généré contient des faits inventés, qui ne peuvent pas être vérifiés ou sont facilement éliminés
  e.g. "Le Parthénon a été construit au 10e siècle après J.-C."

- **Hallucinations de surproclamation**: le texte généré contient des faits inventés, plausibles mais pas vérifiables
  e.g. l'utilisateur demande des informations médicales → réponse qui peut être correcte… ou dangereusement pas

Causes des hallucinations

Causes des hallucinations

- **Données de pré-entraînement:**
  - données contenants des erreurs, des biais, vérification impossible
  - impossible de MAJ les jeux de données tout le temps!

Causes des hallucinations

- **Données de pré-entraînement:**
  - données contenants des erreurs, des biais, vérification impossible
  - impossible de MAJ les jeux de données tout le temps!

- **Pré-entraînement:**
  - modélisation causale du langage → limitation des relations contextuelles
  - *seuil d'apprenabilité* → les faits fréquents sont favorisés
  - hallucinations inhérentes? observations mathématiques

Causes des hallucinations

- **Données de pré-entraînement:**
  - données contenants des erreurs, des biais, vérification impossible
  - impossible de MAJ les jeux de données tout le temps!

- **Pré-entraînement:**
  - modélisation causale du langage → limitation des relations contextuelles
  - *seuil d'apprenabilité* → les faits fréquents sont favorisés
  - hallucinations inhérentes? observations mathématiques

- **SFT :** risque de surapprentissage, forcé de répondre

Causes des hallucinations

- **Données de pré-entraînement:**
  - données contenants des erreurs, des biais, vérification impossible
  - impossible de MAJ les jeux de données tout le temps!

- **Pré-entraînement:**
  - modélisation causale du langage → limitation des relations contextuelles
  - *seuil d'apprenabilité* → les faits fréquents sont favorisés
  - hallucinations inhérentes? observations mathématiques

- **SFT :** risque de surapprentissage, forcé de répondre
- **RLHF :** forcé de s'aligner sur le *feedback* humain que ses compétences

Risque de **sycophantie**!

Détecter les hallucinations

Détecter les hallucinations

- **Métriques basées sur les faits:**
  - extractions de faits + comparer à des sources de connaissances externes

Détecter les hallucinations

- **Métriques basées sur les faits:**
  - extractions de faits + comparer à des sources de connaissances externes

- **Métriques basées sur la classification:**
  - methodes traditionnelles de TALN possibles, mais pas toujours applicables

Détecter les hallucinations

- **Métriques basées sur les faits:**
  - extractions de faits + comparer à des sources de connaissances externes
- **Métriques basées sur la classification:**
  - methodes traditionnelles de TALN possibles, mais pas toujours applicables
- **Métriques basées sur QA:**
  - choisir des réponses → produire des questions →
    évaluer la fidélité des réponses en comparant source - cible

Détecter les hallucinations

- **Estimation de l'incertitude:**
  - probabilités, *weights*, entropie etc... disponible que pour modèles open-source
  - méthodes *adversarial*

Détecter les hallucinations

- **Estimation de l'incertitude:**
  - probabilités, *weights*, entropie etc... disponible que pour modèles open-source
  - méthodes *adversarial*

- **Métriques basées sur les LLM**
  - "LLMs-as-judge" ou "Auto-evaluation"
  - générer questions de vérification, évaluer le contenu d'une réponse générée
  - répondre + évaluer en même temps
  - générer une explication, son raisonnement

Atténuer les hallucinations

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données:**

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données:**
  - Filtrage des données : e.g. surechantillonnage, déduplication, *hashing*

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données:**
  - Filtrage des données : e.g. surechantillonnage, déduplication, *hashing*
  - *Knowledge Graphs*: bases de connaissances en graphe/topologie

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données:**
  - Filtrage des données : e.g. surechantillonnage, déduplication, *hashing*
  - *Knowledge Graphs*: bases de connaissances en graphe/topologie
  - *Retrieval-Augmented Generation*: extraire + ajouter des compétences extérieures

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**

- **Atténuer les hallucinations liées à l'entraînement:**
  - Hallucinations liées au pré-entraînement: fenêtres contextuelles, attention, objectifs

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**

- **Atténuer les hallucinations liées à l'entraînement:**
  - Hallucinations liées au pré-entraînement: fenêtres contextuelles, attention, objectifs
  - Révision du modèle: incorporer des connaissances, MAJ des couches, *meta-learning*

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**

- **Atténuer les hallucinations liées à l'entraînement:**
  - Hallucinations liées au pré-entraînement: fenêtres contextuelles, attention, objectifs
  - Révision du modèle: incorporer des connaissances, MAJ des couches, *meta-learning*
  - Hallucinations liées à l'alignement: annotations, données synthétiques, paires

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**
- **Atténuer les hallucinations liées à l'entraînement**
- **Atténuer les hallucinations liées au décodage:**
  - **Décodage de la factualité: ajuster la probabilité, encodage hiérarchique**

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**
- **Atténuer les hallucinations liées à l'entraînement**
- **Atténuer les hallucinations liées au décodage:**
  - Décodage de la factualité: ajuster la probabilité, encodage hiérarchique
  - Décodage post-édition: auto-correction

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**
- **Atténuer les hallucinations liées à l'entraînement**
- **Atténuer les hallucinations liées au décodage:**
  - Décodage de la factualité: ajuster la probabilité, encodage hiérarchique
  - Décodage post-édition: auto-correction
  - Décodage avec Consistance du contexte: distribution de sortie

Atténuer les hallucinations

- **Atténuer les hallucinations liées aux données**
- **Atténuer les hallucinations liées à l'entraînement**
- **Atténuer les hallucinations liées au décodage:**
  - Décodage de la factualité: ajuster la probabilité, encodage hiérarchique
  - Décodage post-édition: auto-correction
  - Décodage avec Consistance du contexte: distribution de sortie
  - Cohérence logique: *Chain-of-Thought (CoT)*, modèles enseignant/élève, expressions symboliques

Merci de votre attention!