Résumé du papier : To Trust or Not to Trust? Enhancing Large Language Models’ Situated Faithfulness to External Contexts

# Résumé global

Ce papier traite du problème de confiance des grands modèles de langage (LLMs) lorsqu’ils sont confrontés à des contextes externes potentiellement incorrects. Les auteurs introduisent la notion de « situated faithfulness » : la capacité du modèle à déterminer quand faire confiance à son propre savoir interne ou au contexte externe pour répondre correctement.  
Pour cela, deux grandes méthodes sont proposées : Self-Guided Confidence Reasoning (SCR) et Rule-Based Confidence Reasoning (RCR), avec un nouveau dataset RedditQA permettant de tester les modèles sur des contextes incorrects réels.

# Détail des étapes des tests

## 1. Formulation du problème

Les auteurs modélisent la tâche comme une tâche de Question-Answering (QA) avec un contexte pouvant être correct ou incorrect. L’objectif est de maximiser la probabilité de produire une réponse correcte, indépendamment de la validité du contexte fourni.

## 2. Jeux de données utilisés

- RedditQA : nouveaux exemples avec des contextes incorrects extraits de Reddit.  
- NaturalQA : questions ouvertes réalistes.  
- TriviaQA : questions faciles souvent mémorisées par les LLMs.  
- PopQA : questions de différentes popularités.  
- FreshQA : questions sensibles au temps et complexes.  
- ClashEval : benchmark avec contextes synthétiques conflictuels.

## 3. Méthodes évaluées

### a. Direct Input Augmentation (DIA)

Méthode RAG classique : on concatène directement le contexte au prompt.

### b. Self-Guided Confidence Reasoning (SCR)

- ImplicitSCR : le modèle est informé que le contexte peut être faux et doit juger lui-même.  
- ExplicitSCR : le modèle produit séparément une réponse interne et une réponse basée sur le contexte, puis raisonne explicitement sur la confiance en chacune.  
- CR-DPO : méthode d’entraînement optimisée par préférence pour améliorer le raisonnement de confiance.

### c. Rule-Based Confidence Reasoning (RCR)

- InternalEval et ContextEval : le modèle évalue la fiabilité de chaque réponse (interne et contexte).  
- InternalConf et ContextConf : sélection via un score de confiance dépassant un seuil.  
- Token Probability Correction (TPC) : comparaison des probabilités de génération des deux réponses.

## 4. Résultats expérimentaux

- Sur GPT-4o et GPT-4o mini : SCR (notamment CR-DPO) surpasse toutes les autres méthodes, avec des gains allant jusqu'à +24.2% par rapport à DIA.  
- Sur LLaMA-3 8B : les méthodes RCR fonctionnent mieux que SCR car le modèle raisonne moins bien.  
- CR-DPO améliore la généralisation sur des tâches nouvelles en enseignant au modèle à préférer les raisonnements de confiance optimaux.

## 5. Limites des méthodes RCR

- Les règles peuvent être biaisées et sensibles aux seuils de calibration.  
- Les signaux de confiance sont parfois mal alignés avec les règles de décision.  
- Les règles dépendent fortement de la qualité des scores de confiance extraits.

## 6. Conclusion

Les LLMs doivent apprendre à gérer les conflits entre leurs connaissances internes et les contextes externes. Les approches basées sur le raisonnement explicite de confiance (SCR avec CR-DPO) donnent les meilleurs résultats sur les modèles avancés, tandis que les règles (RCR) sont encore utiles pour les modèles plus faibles.