SCOPE : un cadre d'entrainement auto-supervisé pour améliorer la fidélité dans la génération conditionnelle de texte

Song Duong, Florian Le Bronnec, Alexandre Allauzen, Vincent Guigue,
Alberto Lumbreras, Laure Soulier, Patrick Gallinari
Sorbonne Université, CNRS, ISIR, Paris, France
Équipe MILES, LAMSADE, Université Paris-Dauphine, Paris, France
AgroParisTech, UMR MIA-PS, Palaiseau, France
Criteo AI Lab, Paris, France
s.duong@criteo.com, florian.le-bronnec@dauphine.psl.eu

RÉSUMÉ

Les modèles de langage (LLM) produisent souvent des hallucinations lors de la génération conditionnelle de texte, introduisant des informations non fidèles ou non ancrées dans le contexte. Ce phénomène est particulièrement problématique en résumé automatique et en génération texte-à-partir-de-données, où les sorties doivent refléter précisément l'entrée. Nous proposons SCOPE, une méthode auto-supervisée innovante générant automatiquement des exemples non fidèles plausibles pour affiner les modèles par apprentissage par préférences. SCOPE pousse ainsi les modèles à préférer les sorties fidèles. Nous évaluons notre approche sur divers jeux de données de génération texte-à-partir-de-données et de résumé. Simple à implémenter, notre méthode nettement les alternatives existantes selon des métriques automatiques et des évaluations humaines ainsi qu'avec GPT-4.

ABSTRACT

SCOPE : A Self-supervised Framework for Improving Faithfulness in Conditional Text Generation

Large Language Models (LLMs) often produce hallucinations in conditional text generation, introducing information that is unfaithful or not grounded in the input context. This issue frequently arises in automatic summarization and data-to-text generation tasks, where outputs must accurately reflect input data. We introduce SCOPE, a novel self-supervised method that automatically generates plausible unfaithful training samples to refine models using preference-based learning. SCOPE encourages models to prefer faithful outputs over hallucinations. We evaluate our method across multiple summarization and data-to-text datasets, demonstrating significant improvements in faithfulness metrics. Easy to implement, our approach outperforms existing techniques according to automatic, human, and GPT-4-based evaluations.

MOTS-CLÉS : Génération conditionnelle de texte, Fidélité, Hallucinations, Auto-supervision.

KEYWORDS: Conditional text generation, Faithfulness, Hallucinations, Self-supervised learning.

ARTICLE : **Accepté à** ICLR 2025 (https://openreview.net/forum?id=dTkqaCKLPp).