MIRAGE : une bibliothèque de métriques pour évaluer les hallucinations dans les textes générés

Benjamin Vendeville[1, 2, 3] Liana Ermakova[1, 3] Pierre De Loor[2, 4] Jaap Kamps[5]
(1) Université de Bretagne Occidentale, Brest, 29200, France
(2) Lab-STICC, Brest, 29200, France (3) HCTI, Brest, 29200, France
(4) ENIB, Brest, 29200, France (5) University of Amsterdam, Amsterdam, The Netherlands
benjamin.vendeville@univ-brest.fr, liana.ermakova@univ-brest.fr,
deloor@enib.fr, kamps@uva.nl

RÉSUMÉ

Les erreurs dans la génération de langage naturel, appelées hallucinations, restent un défi majeur dans des domaines tels que la santé ou la communication scientifique. Si plusieurs métriques ont été proposées pour les détecter, comme FactCC, QAGS, FEQA et FactAcc, elles sont souvent indisponibles, difficiles à reproduire ou incompatibles avec les workflows modernes. Nous présentons **MIRAGE**, une bibliothèque Python open-source qui réimplémente ces métriques au sein d'un cadre unifié construit sur Hugging Face, offrant modularité, reproductibilité et entrées/sorties standardisées. En adhérant aux principes FAIR, MIRAGE accélère l'expérimentation et soutient le développement de futures métriques. Nous le validons en réévaluant les métriques existantes sur des jeux de données de référence, démontrant des performances comparables avec une meilleure transparence.

ABSTRACT

MIRAGE : A Metrics lIbrary for Rating hAllucinations in Generated tExt

Hallucinations in natural language generation remain a critical challenge, particularly in high-stakes domains such as healthcare or science communication. While several automatic metrics have been proposed to detect and quantify them, such as FactCC, QAGS, FEQA, and FactAcc, these are often unavailable, difficult to reproduce, or incompatible with modern workflows. We introduce **MIRAGE**, an open-source Python library that re-implements key hallucination evaluation metrics in a unified framework built on Hugging Face, offering modularity, reproducibility, and standardized inputs and outputs. Adhering to FAIR principles, MIRAGE accelerates experimentation and supports the development of future metrics. We validate it by re-evaluating existing metrics on benchmark datasets, demonstrating comparable performance while significantly improving usability and transparency.

MOTS-CLÉS : Hallucination ; TALN ; Métriques automatiques.

KEYWORDS: Hallucination ; Natural Language Generation ; Automatic Metrics.

ARTICLE ACCEPTÉ À : Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea..
URL : https://dl.acm.org/doi/10.1145/3746252.3761644

Benjamin Vendeville et Liana Ermakova sont financés par l'ANR (ANR-22-CE23-0019-01) et MaDICS CNRS (https://www.madics.fr/ateliers/simpletext/). Jaap Kamps est soutenu par l'NWO (NWA #1518.22.105), NWO CI (#CISC.CC.016), l'Université d'Amsterdam (AI4FinTech) et l'ICAI (AI for Open Government Lab). Les opinions exprimées n'engagent pas les financeurs.