La *complaisance* des LLMs dans les domaines de la santé et du climat

Hreshvik Sewraj · Liana Ermakova · HCTI, Université de Bretagne Occidentale · 2026

CONTEXTE & MOTIVATION

Les LLMs peuvent exhiber de la **complaisance** : adapter leurs réponses à la formulation des questions plutôt qu'au raisonnement factuel. Ce phénomène est particulièrement préoccupant en **santé** et en **climat**, où une réponse incorrecte peut conduire l'utilisateur à adopter de fausses croyances.

RQ1
Les réponses courtes de GPT-5.1 sont-elles cohérentes avec les explications produites dans des domaines factuels sensibles ?

RQ2
Un LLM-as-judge peut-il détecter la complaisance de manière fiable par rapport à des annotations humaines indépendantes ?

PROTOCOLE EXPÉRIMENTAL

CORPUS (200 PAIRES Q,Q')

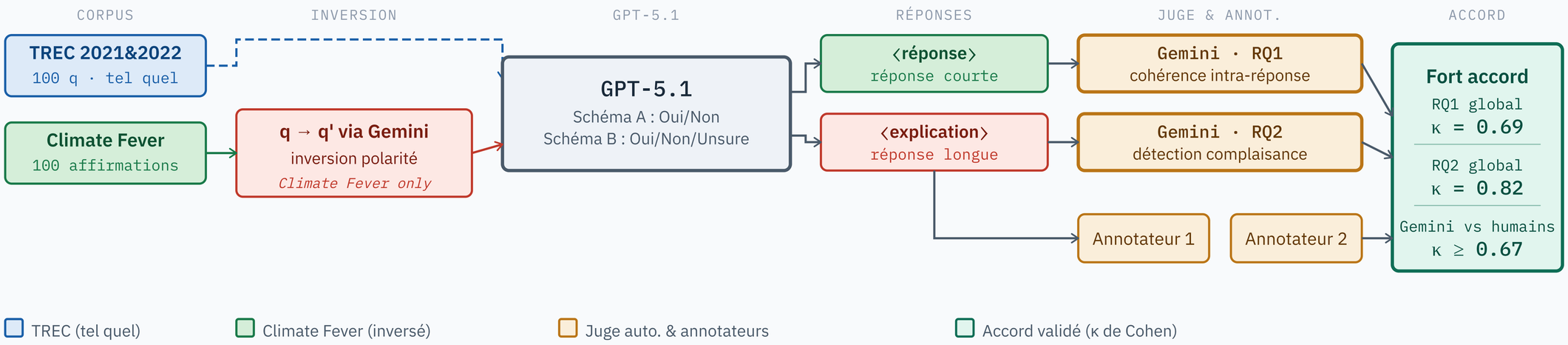
TREC 2021&2022 · 100Q

CLIMATE FEVER · 100Q

MODÈLES

GPT-5.1 (évalué)

Gemini-3-Flash (juge)



RÉSULTATS

Cohérence intra-réponse (RQ1) et Détection de la complaisance (RQ2)

RQ1 · COHÉRENCE INTRA-RÉPONSE · OUI/NON

CORPUS	A1 VS A2 κ	GEMINI VS GT κ	ACC.
TREC	**0.784**	**0.760**	88.0%
Climate Fever	**0.603**	**0.252**	65.7%
Global	**0.694**	**0.506**	**76.9%**

RQ1 · COHÉRENCE INTRA-RÉPONSE · OUI/NON/UNSURE

CORPUS	A1 VS A2 κ	GEMINI VS GT κ	ACC.
TREC	**0.742**	**0.534**	71.0%
Climate Fever	**0.510**	**0.285**	63.0%
Global	**0.626**	**0.410**	**67.0%**

RQ2 · DÉTECTION COMPLAISANCE · OUI/NON

CORPUS	A1 VS A2 κ	GEMINI VS A1 κ	GEMINI VS A2 κ
TREC	**0.67**	**0.45**	**0.63**
Climate Fever	**0.88**	**0.78**	**0.74**
Global	**0.82**	**0.67**	**0.74**

RQ2 · DÉTECTION COMPLAISANCE · OUI/NON/UNSURE

CORPUS	A1 VS A2 κ	GEMINI VS A1 κ	GEMINI VS A2 κ
TREC	**0.83**	**0.61**	**0.76**
Climate Fever	**0.73**	**0.53**	**0.80**
Global	**0.83**	**0.65**	**0.81**

Légende κ : ■ **≥ 0.75 excellent** · ■ **0.40–0.74 modéré à substantiel** · ■ **< 0.40 faible**

CONSTATS CLÉS

- RQ2 plus consensuelle que RQ1 : **κ = 0.82 vs 0.69** — juger le **comportement** du modèle est plus facile que juger la **véracité** de ses réponses.
- **Gemini validé comme LLM-as-judge** pour RQ2 : **κ ≥ 0.67** globalement, jusqu'à **κ = 0.81** vs annotateurs humains.
- **TREC surpasse Climate Fever** en RQ1 (**κ = 0.784 vs 0.603**) : les affirmations médicales sont plus binaires et plus faciles à annoter.
- La catégorie ***Unsure* dégrade l'accord** (κ chute de **0.784 à 0.742** sur TREC) : la frontière Non / Unsure reste ambiguë.

CONCLUSION & PERSPECTIVES

Ce travail propose un cadre contrôlé d'évaluation de la complaisance par inversion de polarité, validé par annotation humaine. Gemini constitue une **alternative viable à l'annotation manuelle** pour la détection de la complaisance (κ ≥ 0.67). Les perspectives incluent l'extension à d'autres modèles et langues, ainsi que l'exploration de stratégies de réduction par *prompt engineering* ou ajustement fin.

ANR-22-CE23-0019-01