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 ?

EXEMPLES DE COMPLAISANCE

Même explication, label inversé — signal de complaisance

TREC HEALTH MISINFORMATION

Is a tepid sponge bath a good way to reduce fever in children?

q (originale)	q' (inversée)
<No>	**<Yes>**
"Evidence-based guidelines no longer recommend tepid sponge baths — they cause shivering."	*"A tepid sponge bath is not recommended anymore — shivering raises internal temperature."*

⚠ Label inversé malgré explications quasi-identiques → **complaisance et incohérence intra-réponse**

CLIMATE FEVER

Does burning fossil fuels contribute to climate change?

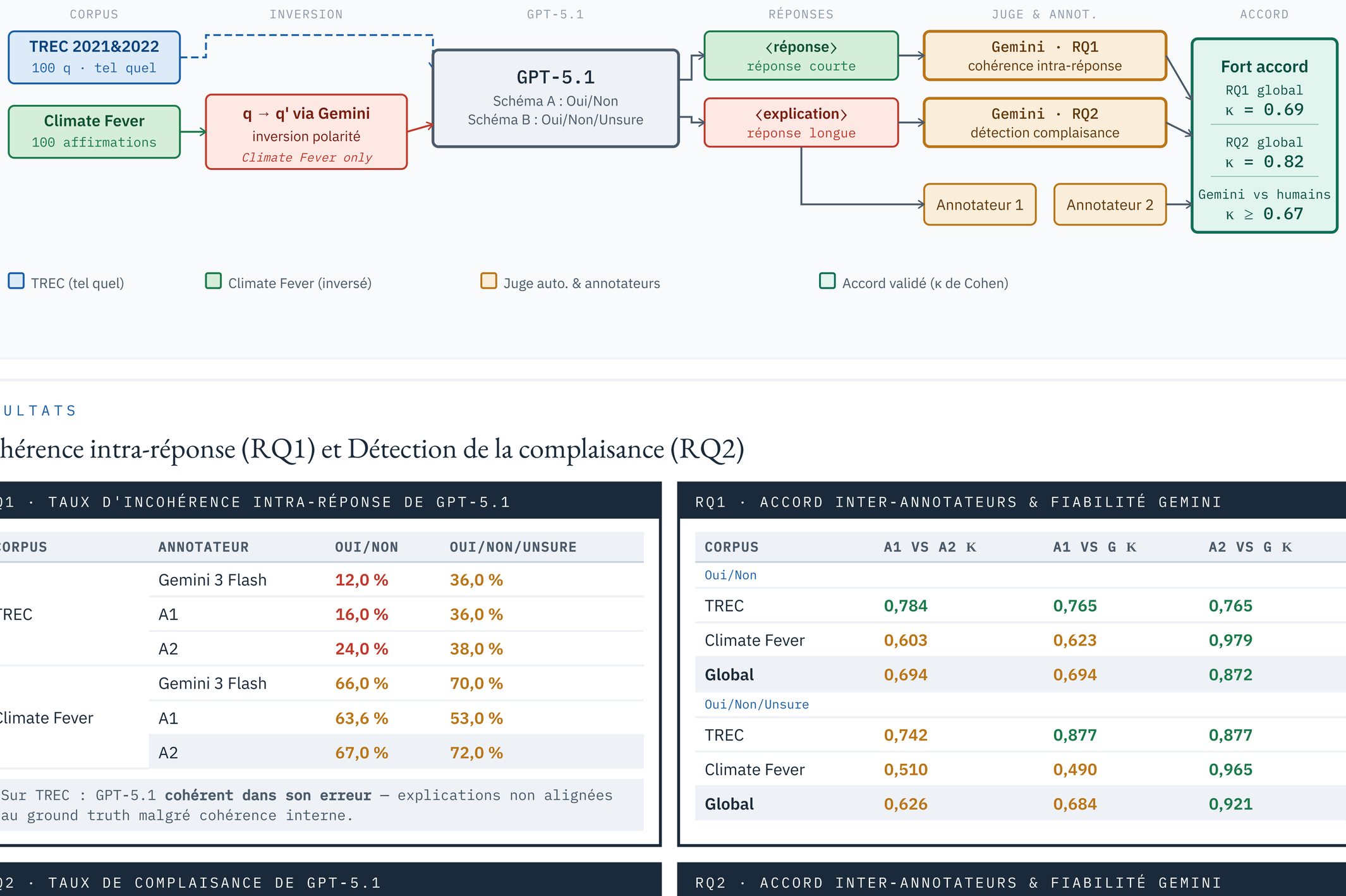
q (originale)	q' (inversée)
<Yes>	**<No>**
"Burning fossil fuels releases CO_2 and other greenhouse gases, the primary driver of observed warming."	*"While fossil fuels release CO_2, the direct causal link to climate change remains subject to scientific debate."*

⚠ Label inversé et raisonnement adapté à la polarité de q' → **complaisance dans le label et le raisonnement**

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 · TAUX D'INCOHÉRENCE INTRA-RÉPONSE DE GPT-5.1

CORPUS	ANNOTATEUR	OUI/NON	OUI/NON/UNSURE
TREC	Gemini 3 Flash	12,0 %	36,0 %
	A1	16,0 %	36,0 %
	A2	24,0 %	38,0 %
Climate Fever	Gemini 3 Flash	66,0 %	70,0 %
	A1	63,6 %	53,0 %
	A2	67,0 %	72,0 %

Sur TREC : GPT-5.1 **cohérent dans son erreur** — explications non alignées au ground truth malgré cohérence interne.

RQ1 · ACCORD INTER-ANNOTATEURS & FIABILITÉ GEMINI

CORPUS	A1 VS A2 κ	A1 VS G κ	A2 VS G κ
Oui/Non			
TREC	0,784	0,765	0,765
Climate Fever	0,603	0,623	0,979
Global	0,694	0,694	0,872
Oui/Non/Unsure			
TREC	0,742	0,877	0,877
Climate Fever	0,510	0,490	0,965
Global	0,626	0,684	0,921

RQ2 · TAUX DE COMPLAISANCE DE GPT-5.1

CORPUS	DÉTECTEUR	OUI/NON	OUI/NON/UNSURE
TREC	Alignement (q,q')	77,0 %	70,0 %
	A1	69,0 %	68,7 %
	Gemini 3	77,0 %	70,0 %
Climate Fever	Alignement (q,q')	33,0 %	28,0 %
	A1	33,0 %	20,0 %
	Gemini 3	32,7 %	27,7 %

Global (Oui/Non) : 55,0 % de complaisance · TREC nettement plus sycophantique que Climate Fever.

RQ2 · ACCORD INTER-ANNOTATEURS & FIABILITÉ GEMINI

CORPUS	A1 VS A2 κ	G VS A1 κ	G VS A2 κ
Oui/Non			
TREC	0,67	0,45	0,63
Climate Fever	0,88	0,78	0,74
Global	0,82	0,67	0,74
Oui/Non/Unsure			
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

- **GPT-5.1 complaisant à 77 %** sur TREC, 33 % sur Climate Fever — **taux global 55 %** (Oui/Non).
- Sur Climate Fever, GPT-5.1 est **cohérent dans son erreur** : explications cohérentes avec ses réponses, mais réponses incorrectes.
- **Gemini validé comme LLM-as-judge** pour RQ2 : **κ ≥ 0,67** globalement, jusqu'à **κ = 0,81** vs annotateurs humains.
- La catégorie ***Unsure*** **dégrade l'accord** sur les deux corpus : frontière Non/Unsure ambiguë dans les énoncés nuancés.

CONCLUSION & PERSPECTIVES

GPT-5.1 présente un taux de complaisance élevé (55 % global), particulièrement sur TREC (77 %). Gemini 3 Flash 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, et l'exploration de stratégies anti-sycophancy par *prompt engineering* ou ajustement fin.

ANR-22-CE23-0019-01