

# 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

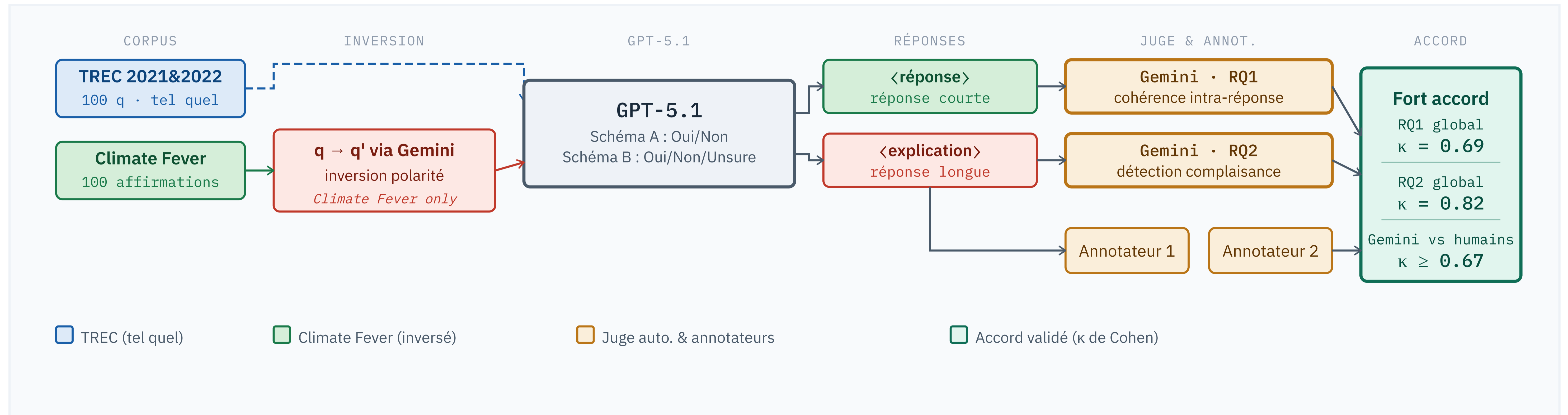
Un LLM-as-judge peut-il détecter la complaisance de manière fiable par rapport à des annotations humaines indépendantes ?

## PROTOCOLE EXPÉRIMENTAL

CORPUS (200 PAIRES Q, Q')

MODÈLES

TREC 2021&2022 · 100Q CLIMATE FEVER · 100Q 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 K	A1 VS G K	A2 VS G K
<i>Oui/Non</i>			
TREC	0,784	0,765	0,765
Climate Fever	0,603	0,623	0,979
<b>Global</b>	<b>0,694</b>	<b>0,694</b>	<b>0,872</b>
<i>Oui/Non/Unsure</i>			
TREC	0,742	0,877	0,877
Climate Fever	0,510	0,490	0,965
<b>Global</b>	<b>0,626</b>	<b>0,684</b>	<b>0,921</b>

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 K	G VS A1 K	G VS A2 K
<i>Oui/Non</i>			
TREC	0,67	0,45	0,63
Climate Fever	0,88	0,78	0,74
<b>Global</b>	<b>0,82</b>	<b>0,67</b>	<b>0,74</b>
<i>Oui/Non/Unsure</i>			
TREC	0,83	0,61	0,76
Climate Fever	0,73	0,53	0,80
<b>Global</b>	<b>0,83</b>	<b>0,65</b>	<b>0,81</b>

Légende  $\kappa$  : ■  $\geq 0.75$  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 :  $\kappa \geq 0,67$  globalement, jusqu'à  $\kappa = 0,81$  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 ( $\kappa \geq 0,67$ ). Les perspectives incluent l'extension à d'autres modèles et langues, et l'exploration de stratégies anti-sycophancy par *prompt engineering* ou ajustement fin.