Quantification de l'Incertitude via Opérateur de Relaxation pour les grands modèles de raisonnement en contexte Boîte Noire - 2026

Lucas Biechy[1, 2, 3], Cédric Eichler[2, 4, 1], Adrien Boiret[2, 4, 1], Nicolas Anciaux[1,3,2]

[1] Inria Saclay
[2] INSA Centre-Val de Loire
[3] Université Paris Saclay
[4] Université d'Orléans

prenom.nom@inria.fr

Résumé

Nous étudions la quantification de l'incertitude pour les grands modèles de raisonnement en mode boîte noire. Nous montrons que les méthodes existantes n'améliorent pas l'estimation de l'incertitude comparé à une méthode de référence naïve. Théoriquement, nous introduisons les opérateurs de relaxation pour élargir la distribution des réponses et améliorer la calibration sous réserve de conservation de précision. Empiriquement, nous utilisons des techniques de prompt antagoniste et démontrons qu'elles améliore l'estimation d'incertitude et se comportent comme des opérateurs de relaxation.

Mots-clés

Quantification de l'incertitude, Apprentissage par renforcement, Prompts antagonistes

1 Introduction

L'émergence des grands modèles de raisonnement (LRMs) représente une évolution majeure par rapport aux grands modèles de langage (LLMs) traditionnels. Contrairement à ces derniers, les LRMs sont conçus pour produire explicitement des étapes de raisonnement intermédiaires avant de générer une réponse, ce qui leur permet d'atteindre des performances élevées sur de nombreuses tâches. Ils sont ainsi de plus en plus utilisés comme systèmes de prise de décision, notamment dans des contextes critiques formulés sous forme de question-réponse (QA). Dans ces situations, une bonne quantification de l'incercititude (UQ) du modèle est essentielle.

En pratique, les LRMs sont souvent accessibles uniquement en boîte noire, ce qui limite les approches de UQ aux méthodes basées sur le prompt : de type verbalisation de confiance et consistance face à la répétition. Toutefois, l'impact du passage des LLMs aux LRMs sur la calibration de la confiance reste encore peu étudié.

Dans ce travail, nous évaluons plusieurs méthodes d'UQ en boîte noire initialement proposées pour les LLMs et montrons qu'elles sont globalement comparables à une référence simple basée sur l'échantillonnage répété. En pratique, les LRMs héritent des problèmes de surconfiance observés dans les LLMs, ce qui se traduit par une variabilité limitée des prédictions. Guidé par la littérature sur le sujet, nous supposons que ce phénomène est lié à la phase d'alignement. Durant cette dernière, les modèles sont optimisés par apprentissage par renforcement (RL) pour améliorer leurs capacités de raisonnement tout en évitant la production de contenus nuisibles. Cette optimisation tend à concentrer les distributions de sortie et à favoriser des trajectoires de raisonnement plus déterministes, réduisant ainsi l'efficacité des méthodes d'UQ reposant sur la diversité des réponses.

Pour remédier à cette limitation, nous introduisons un opérateur de relaxation au niveau du prompt. Cette transformation agit comme une force de rappel vers la politique pré-alignement du modèle, avant son optimisation par RL. Dans notre cadre théorique, nous modélisons le modèle aligné comme une politique optimale fermée, et interprétons la relaxation comme une approximation d'une politique entraînée avec une régularisation KL plus forte, donc plus proche du modèle de référence. Sous l'hypothèse que l'accuracy est préservée, nous montrons que cette relaxation améliore la calibration, notamment en réduisant l'Expected Calibration Error (ECE).

Empiriquement, nous exploitons une famille de méthodes de prompts antagonistes, initialement développée pour contourner les mécanismes de sécurité des modèles. Nous montrons que ces méthodes améliorent la calibration et produisent des effets similaires à des opérateurs de relaxation.

2 Théorie

L'optimisation par RL est classiquement réalisée via des méthodes tels que DPO, GRPO et GSPO.

Il est maintenant bien établi que ces méthodes admettent une politique optimale en forme fermée [1], qui peut s'écrire comme suit :

$$\pi_\beta(t \mid x) = \frac{1}{Z_\beta(x)} \pi_{\text{ref}}(t \mid x) \exp\left(\frac{R(x,t)}{\beta}\right) \qquad (1)$$

où $\beta > 0$ est le paramètre de régularisation KL, $Z_\beta(x) = \sum_{t \in \mathcal{T}} \pi_{\text{ref}}(t|x) \exp(R(x,t)/\beta)$ est la fonction de partition

assurant la normalisation, et $R(x,t)$ est la fonction de récompense supposée finie.

Definition 2.1 (Opérateur de relaxation) *Soient π_{ref} une politique de référence pré-entraînée et π_{RL} une politique dérivée de π_{ref} via RL. Un opérateur de relaxation j est une transformation qui mappe une entrée x vers une version modifiée $j(x)$ de sorte que la politique induite*

$$\pi_j(\cdot \mid x) \coloneqq \pi_{RL}(\cdot \mid j(x))$$

produise un effet équivalent à une politique optimale en forme fermée de l'équation (1) avec un paramètre KL plus élevé. Formellement, il existe β_1, β_2 tels que :

$$\pi_j \equiv \pi_{\beta_1}, \quad \textit{avec} \quad \pi_{RL} \equiv \pi_{\beta_2} \quad \textit{et} \quad \beta_1 > \beta_2$$

Theorem 2.2 *Considérons une politique π_{RL} présentant une surconfiance systémique. Soit j un opérateur de relaxation induisant une politique π_j. Sous l'hypothèse que l'accuracy réelle reste constante sous π_j, on a pour un nombre fini K d'échantillons de trajectoires :*

$$ECE(\pi_j) \leq ECE(\pi_{RL}) + \mathcal{O}\left(\sqrt{\frac{\log|\mathcal{Y}|}{K}}\right) \tag{2}$$

De plus, dans la limite asymptotique $K \to \infty$, la relaxation réduit strictement l'erreur de calibration :

$$ECE(\pi_j) < ECE(\pi_{RL}) \tag{3}$$

3 Résultats empiriques

Nous exploitons les techniques de prompts antagonistes comme opérateurs de relaxation car elles contournent les contraintes induites par l'alignement tout en préservant le contenu sémantique du prompt, rapprochant ainsi les réponses de la distribution pré-entraînement du modèle. Dans un cadre boîte noire, notre méthode transforme K prompts en variantes antagoniste, interroge le LRM pour chaque transformation, et agrège les réponses pour estimer l'incertitude.

Nous testons trois familles représentatives de prompts antagonistes :

- **SUFFIX** : suffixe aléatoire ajouté au prompt.
- **ART** : remplacement de mots par un équivalent ASCII-art.
- **PROG** : restructuration en sous-sections et reconstruction avant réponse.

Amélioration de la calibration. Comme résumé dans la table 1, les méthodes existantes, qui consistent soit à paraphraser (REPHRASE [3]), soit à verbaliser la confiance (VC [2]), apportent des gains limités par rapport à une répétition simple (RS) du même prompt. Cependant, PROG et ART améliorent significativement la calibration (ECE, Brier, NLL) sur la majorité des paires LRM–dataset, sans perte d'accuracy. Par exemple, PROG réduit l'ECE de 23,1% en moyenne.

TABLE 1 – Comparaison avec RS sur 9 paires LRM–dataset. Le chiffre indique le gain moyen ; les valeurs en gras correspondent à la meilleure méthode par métrique. La variation selon LRM et dataset est représentée par une heatmap 3×3 (lignes = LRMs, colonnes = datasets) : vert = amélioration, rouge = dégradation ; les teintes foncées indiquent une significativité statistique.

	VC [2]	REPHRASE [3]	SUFFIX (ours)	PROG (ours)	ART (ours)
Acc. ↑	-0.6%	+0.5%	**+2.0%**	-4.3%	+1.9%
ECE ↓	+4.2%	-9.2%	-1.2%	**-23.1%**	-15.3%
Brier ↓	+2.0%	-0.1%	-0.6%	**-6.6%**	-6.1%
NLL ↓	+2.4%	-3.6%	-2.7%	**-12.5%**	-11.5%

Consistance avec un opérateur de relaxation. De surcroît, les transformations augmentent la variabilité des prédictions (entropie +57–70%) et la diversité sémantique des traces de raisonnement (cosine similarity réduite de 14–28%), cohérent avec le comportement attendu d'un opérateur de relaxation.

4 Conclusion

Nous avons supposé que la surconfiance est la cause de la mauvaise calibration des méthodes connues de quantification de l'incertitude dans les LRMs dans un contexte boite noire. En théorie, nous avons introduit la notion d'opérateur de relaxation et démontré que l'application d'un tel opérateur entraînerait une amélioration de cette quantification. Empiriquement, nous avons montré non seulement que certaines méthodes de prompt antagoniste se comportent comme des opérateurs de relaxation et améliorent ainsi, conformément à la théorie, la quantification de l'incertitude face à une référence naïve.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2025-011016360 attribuée par GENCI. Ce travail a été soutenu par le projet ANR 22-PECY-0002 IPOP (Projet interdisciplinaire sur la confidentialité) du PEPR Cybersécurité.

Références

[1] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization : Your language model is secretly a reward model. *Advances in neural information processing systems*, 36 :53728–53741, 2023.

[2] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty ? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv :2306.13063*, 2023.

[3] Adam Yang, Chen Chen, and Konstantinos Pitas. Just rephrase it ! uncertainty estimation in closed-source language models via multiple rephrased queries. *arXiv preprint arXiv :2405.13907*, 2024.