# Résumé

*Benchmarking LLMs via Uncertainty Quantification*

## Contexte

Les Large Language Models (LLMs) sont évalués classiquement par l'accuracy seule (ex. classements ouverts type HuggingFace), ce qui ignore un aspect : l'incertitude des prédictions. Deux modèles peuvent atteindre la même précision tout en ayant des degrés de certitude très différents. L'article résume une approche rigoureuse pour quantifier l'incertitude et l'intégrer au *benchmarking* des LLMs, en s'appuyant sur la prédiction conforme.

**Idée clé :** La prédiction conforme transforme une mesure heuristique (p. ex. scores *softmax*) en une garantie statistique *distribution-free* et *agnostique au modèle*, en produisant, pour chaque entrée, un ensemble de prédiction censé contenir la bonne réponse avec une probabilité au moins $1 - \alpha$. L'incertitude est alors capturée par la taille de cet ensemble.

## Méthode : prédiction conforme

Soit $f$ un classifieur à $K$ classes $Y = \{1, \ldots, K\}$, et $(X_t, Y_t)$ un point de test. La prédiction conforme construit un ensemble $C(X_t) \subseteq Y$ tel que

$$\mathbb{P}(Y_t \in C(X_t)) \;\geq\; 1 - \alpha, \tag{1}$$

où $\alpha \in (0, 1)$ est fixé par l'utilisateur (taux d'erreur). Cette garantie s'obtient via un petit ensemble de calibration $D_{\text{cal}}$, en :

1. choisissant une notion heuristique d'incertitude (ex : *softmax*) ;
2. définissant un score conforme $s(X, Y) \in \mathbb{R}$ ;
3. calculant les scores sur $D_{\text{cal}}$ et un seuil $\hat{q}$ au quantile $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ ;
4. formant $C(X_t) = \{Y' \in Y : \; s(X_t, Y') \leq \hat{q}\}$.

Deux fonctions de score sont étudiées :

**LAC** (*Least Ambiguous set-valued Classifier*) : $s(X, Y) = 1 - f(X)_Y$. Il minimise en moyenne la taille de $C(X)$, mais peut sous-couvrir des cas difficiles.

**APS** (*Adaptive Prediction Sets*) : $s(X, Y) = \sum_{\{Y' : f(X)_{Y'} \geq f(X)_Y\}} f(X)_{Y'}$, qui exploite les scores de toutes les classes ; il corrige une limite de LAC mais produit souvent des ensembles plus grands.

## Protocole d'évaluation

**Tâches et jeux de données.** Cinq tâches NLP (10 000 instances chacune) sont reformulées en QCM à 6 choix (A, B, C, D, E, F), en ajoutant *'I don't know'* et *'None of the above'* pour standardiser : **QA** (MMLU), **RC** (CosmosQA), **CI** (HellaSwag), **DRS** (HaluDial), **DS** (HaluSum, dérivé de CNN/DailyMail).

**Stratégies de prompting** Trois variantes sont comparées pour limiter la sensibilité aux prompts : *Base Prompt*, *Shared Instruction Prompt*, *Task-specific Instruction Prompt*. Les scores *softmax* sont extraits à partir des *logits* de la tête LM sur le dernier token.

**Métriques** On évalue la **précision** (Acc), l'**incertitude** via la **taille d'ensemble** (SS) et on vérifie la **couverture** (CR) :

$$\text{Acc} = \frac{1}{|D_{\text{test}}|} \sum_{(X_t,Y_t)\in D_{\text{test}}} \mathbf{1}\{Y_p = Y_t\}, \tag{2}$$

$$\text{SS} = \frac{1}{|D_{\text{test}}|} \sum_{(X_t,Y_t)\in D_{\text{test}}} |C(X_t)|, \tag{3}$$

$$\text{CR} = \frac{1}{|D_{\text{test}}|} \sum_{(X_t,Y_t)\in D_{\text{test}}} \mathbf{1}\{Y_t \in C(X_t)\}. \tag{4}$$

Dans les expériences, $\alpha = 0{,}1$ (cible CR $\approx 90\%$) et les résultats agrègent LAC/APS et les trois prompts.

## Jeux de modèles

Neuf familles *open-source* couvrant 6B–14B paramètres principalement : Llama-2, Mistral-7B, Falcon, MPT-7B, Gemma-7B, Qwen, Yi, DeepSeek, InternLM-7B (poids HuggingFace).

## Résultats principaux

— **Couverture.** La plupart des CR atteignent $\geq 90\%$ ; le plus bas cas rapporté est 89,56% (Qwen-7B/DS), et les moyennes par modèle dépassent 90%, confirmant la validité des ensembles produits.

— **Décorrélation Acc–SS.** En pratique, plus d'accuracy n'implique pas moins d'incertitude . Les rangs par Acc et par SS divergent souvent ; on observe même des inversions entre paires de modèles (p. ex. InternLM-7B vs MPT-7B en DRS). Conclusion : il faut évaluer précision *et* incertitude.

## Analyses complémentaires

**Changement d'échelle (scaling).** Sur la série Qwen (1,8B $\rightarrow$ 72B), la précision augmente globalement avec la taille ; l'incertitude (SS) baisse surtout jusqu'à $\sim$14B, puis les gains deviennent faibles et plus variables (ex. Qwen-72B plus incertain que 14B en RC/DRS malgré une meilleure Acc).

**Fine-tuning par instruction.** Sur Llama-2, la version *chat* (format *Chat-V1*) dégrade **Acc** et augmente **SS** à toutes tailles ; *Chat-V2* (même format que *Base*) améliore parfois l'Acc (7B/13B) mais **augmente systématiquement l'incertitude** vs *Base*. En bref : l'instruction-finetuning tend à accroître l'incertitude.

**Comparaison à l'entropie/perplexité.** L'entropie est transformée en **perplexité** $PPL = 2^H$ (dans $[1, |Y|]$) pour comparer à SS. Pour InternLM-7B, le CR basé sur $PPL$ varie fortement selon les tâches (jusqu'à 83,44% en QA), alors que la prédiction conforme reste $\geq 90\%$. En ECE, la prédiction conforme surpasse entropie et probabilité maximale moyenne (meilleur calibrage). **Conclusion : la prédiction conforme fournit une quantification d'incertitude plus fiable.**

**LLMs fermés.** Sans logits (GPT-3.5/4), les probabilités par option sont estimées en échantillonnant 50 sorties/question + *softmax* à température. GPT-4 affiche **meilleure Acc** et **plus faible SS**. Pour Qwen-72B, l'approximation par échantillonnage est proche des logits (JSD moyen 0,05).

**Génération libre (TriviaQA).** On génère 20 réponses/question et on utilise la **perplexité de chaque génération** comme score conforme ; SS $\in [1, 20]$. SS varie entre modèles et renseigne l'incertitude, mais la **couverture n'est plus garantie** (la bonne réponse peut ne pas figurer parmi les 20). Les modèles plus forts respectent plus souvent la couverture.

**Lien SS–accuracy (stratification).** En regroupant par SS (InternLM-7B/QA), l'accuracy décroît nettement quand SS augmente (SS=1 $\Rightarrow$ très forte Acc ; SS élevé $\Rightarrow$ Acc plus faible), confirmant SS comme **indicateur utile d'incertitude**—mais des cas justes subsistent même pour SS maximal.

## Implications pratiques (à retenir)

— **Évaluer** un LLM ne doit pas se limiter à l'accuracy : il faut **mesurer l'incertitude** (SS, CR) et **vérifier le calibrage**.

— **Comparer** des LLMs par profils (Acc, SS, CR) révèle des *trade-offs* utiles pour le déploiement (ex : sécurité, supervision humaine).

— **Conception** de systèmes : déclencher une relecture humaine ou une vérification externe quand SS dépasse un seuil.

## Conclusion

L'étude montre, sur 9 familles de LLMs et 5 tâches, qu'un benchmark centré uniquement sur l'accuracy est insuffisant. La prédiction conforme fournit un cadre rigoureux pour quantifier l'incertitude (via SS avec garantie de couverture), met en évidence la décorrélation entre précision et certitude, et s'étend aux LLMs fermés et à la génération libre.