Récap méthodes statapp

1 Méthodes *white-box*

— **Probabilité moyenne de la séquence** [1], [2] Mesure la vraisemblance moyenne d'une réponse, exprimée soit en probabilité, soit en log-probabilité.

$$P_{\text{mean}} = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log p_i\right) \qquad \Longleftrightarrow \qquad \text{Avg}(-\log p) = -\frac{1}{L}\sum_{i=1}^{L}\log p_i.$$

— **Probabilité totale de la séquence** [1] Forte sensibilité à la longueur ; reflète la vraisemblance brute.

$$P_{\text{seq}} = \prod_{i=1}^{L} p_i.$$

— **Probabilité des tokens clés** [1] Se concentre sur les tokens essentiels de la réponse (ex. le résultat final).

$$P_{\text{key}} = \prod_{t_i \in \mathcal{K}} p_i.$$

— **Incertitude maximale (log-probabilité)** [2] Identifie le token le moins probable d'une phrase.

$$\text{Max}(-\log p) = \max_i \big(-\log p_i\big).$$

— **Incertitude moyenne (entropie)** [2] Mesure l'incertitude globale du modèle pour chaque token.

$$\text{Avg}(H) = \frac{1}{L}\sum_{i=1}^{L}\left(-\sum_{\tilde{w}} p_i(\tilde{w})\log p_i(\tilde{w})\right).$$

— **Incertitude maximale (entropie)** [2] Repère la position où l'incertitude du modèle est la plus forte.

$$\mathrm{Max}(H) = \max_i \left(-\sum_{\tilde{w}} p_i(\tilde{w}) \log p_i(\tilde{w}) \right).$$

2 Méthode *white-box* de Prédiction Conforme (Conformal Prediction)

[3]

La **Prédiction Conforme** est une méthode statistique rigoureuse pour quantifier l'incertitude des modèles d'apprentissage automatique, y compris les grands modèles de langage (LLMs). Elle est **distribution-free** (non paramétrique) et **model-agnostic**, ce qui la rend particulièrement adaptée à l'évaluation de systèmes complexes comme les LLMs.

2.1 Principe fondamental

Étant donné un taux d'erreur $\alpha \in (0, 1)$ choisi par l'utilisateur, la prédiction conforme garantit que l'ensemble de prédiction $\mathcal{C}(X_t)$ contient l'étiquette vraie Y_t avec une probabilité d'au moins $1 - \alpha$:

$$\mathbb{P}\big(Y_t \in \mathcal{C}(X_t)\big) \geq 1 - \alpha.$$

Cette garantie probabiliste est obtenue via l'utilisation d'un **jeu de calibration** $\mathcal{D}_{\text{cal}} = \{(X_c^{(i)}, Y_c^{(i)})\}_{i=1}^n$, qui sert à calibrer le seuil de décision sans nécessiter d'hypothèses sur la distribution des données.

2.2 Processus de construction des ensembles de prédiction

1. **Calcul des scores de conformité** : Pour chaque exemple de calibration $(X_c^{(i)}, Y_c^{(i)})$, on calcule un score $s_i = s(X_c^{(i)}, Y_c^{(i)})$ qui mesure le désaccord entre l'entrée et sa vraie étiquette.
2. **Détermination du seuil** : On calcule le quantile d'ordre $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$

des scores de calibration :

$$\hat{q} = \text{quantile}\left(\{s_1, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right).$$

3. **Construction des ensembles** : Pour une nouvelle entrée X_t, on construit l'ensemble de prédiction :

$$\mathcal{C}(X_t) = \{Y' \in \mathcal{Y} : s(X_t, Y') \leq \hat{q}\}.$$

2.3 Fonctions de score couramment utilisées

Deux fonctions de score sont particulièrement adaptées aux tâches de classification avec LLMs :

LAC (Least Ambiguous set-valued Classifiers)

$$s_{\text{LAC}}(X, Y) = 1 - f(X)_Y$$

où $f(X)_Y$ est la probabilité softmax attribuée par le modèle à la vraie étiquette Y. Cette approche tend à produire des ensembles de petite taille en moyenne mais peut sous-couvrir les instances difficiles.

APS (Adaptive Prediction Sets)

$$s_{\text{APS}}(X, Y) = \sum_{\substack{Y' \in \mathcal{Y} \\ f(X)_{Y'} \geq f(X)_Y}} f(X)_{Y'}$$

Cette méthode somme les probabilités de toutes les étiquettes au moins aussi probables que la vraie étiquette, produisant généralement des ensembles plus grands mais avec une couverture plus adaptative.

2.4 Mesure d'incertitude : Set Size (SS)

L'incertitude est quantifiée par la taille moyenne des ensembles de prédiction :

$$SS = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(X_t, Y_t) \in \mathcal{D}_{\text{test}}} |\mathcal{C}(X_t)|.$$

Un SS proche de 1 indique une grande certitude (l'ensemble ne contient qu'une seule étiquette), tandis qu'un SS proche du nombre total de classes K indique une grande incertitude.

3 Méthodes *black-box*

Les méthodes *black-box* reposent soit sur la **confiance verbalisée** (auto-déclarée par le LLM), soit sur la **génération multiple** et l'analyse de cohérence entre sorties.

3.1 Confiance verbalisée (réponses courtes, vérifiables)

Ces méthodes supposent que la réponse du LLM est un format simple (QCM, numérique, vrai/faux) et que la confiance peut être *déclarée* par le modèle. Elles sont propres à [1].

— **Confiance verbalisée directe (Vanilla)** [1] Le LLM fournit une réponse et un score de confiance auto-déclaré.
— **Confiance après raisonnement (Chain-of-Thought)** [1] Le LLM explique son raisonnement puis déclare une confiance finale.
— **Auto-évaluation de la réponse (Self-Probing)** [1] Le LLM juge lui-même la probabilité que sa réponse soit correcte.
— **Confiance multi-étapes (Multi-Step)** [1] Le raisonnement est décomposé en étapes, chacune avec une confiance :

$$C_{\text{global}} = \prod_{i=1}^{n} C_i.$$

— **Top-K verbalized confidence** [1] Le LLM donne ses K meilleures hypothèses et leur probabilité déclarée.

3.2 Génération multiple (*sampling*) et mesures de cohérence (deux articles)

Ces méthodes reposent sur l'idée que si le LLM connaît la réponse, les échantillons générés seront similaires ; sinon, ils divergeront. Approche commune à [1] et [2].

— **Self-Random Sampling** [1], [2] Générer plusieurs réponses indépendantes avec température > 0.
— **Paraphrasing Sampling** [1] Reformuler la question pour tester la stabilité des réponses.
— **Misleading Sampling** [1] Introduire volontairement un indice erroné pour voir si le LLM change d'avis.
— **Cohérence majoritaire (Majority Consistency)** [1], [2]

$$C = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\{Y_i = Y^*\}.$$

Une forte cohérence indique une réponse fiable.
— **Agrégation Pair-Rank** [1] Exploite les classements Top-K pour inférer des préférences robustes.
— **Agrégation Avg-Conf** [1] Combine verbalisation de confiance et cohérence entre réponses.

Ces méthodes fonctionnent bien lorsque la réponse attendue est courte ou vérifiable.

3.3 Méthodes avancées pour les réponses longues, textuelles et non vérifiables

Ces méthodes sont spécifiques aux sorties composées de plusieurs phrases (résumés, biographies, descriptions), où la vérité n'est pas directement accessible. Elles appartiennent à [2].

— **Consistance sémantique (BERTScore Consistency)** [2] Mesure la similarité sémantique entre phrases sur plusieurs échantillons.
— **Modèle n-gram/unigram entraîné sur les échantillons** [2] Évalue si le texte d'origine est probable selon les n-grammes des générations. Variante : repérage du token le plus rare dans les échantillons.
— **Détection de contradiction via NLI** [2]

$$C_{\text{NLI}} = \frac{1}{M} \sum_{i=1}^{M} P(\text{contradiction} \mid Y, Y_i).$$

Identifie les incohérences factuelles.
— **QA-Consistency** [2] Generer des questions sur la réponse, puis tester

si les autres échantillons donnent la même réponse.

— **LLM Judge (Oui/Non)** [2] Le LLM répond à : *"Cette phrase est-elle supportée par ce contexte ?"* Score = proportion de réponses "Non".

Références

[1] *Can LLMs Express Their Uncertainty ? An Empirical Evaluation of Confidence Elicitation in LLMs*, Xiong et al., ICLR 2024.

[2] *SelfCheckGPT : Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*, Manakul et al., 2023.

[3] *Benchmarking LLMs via Uncertainty Quantification*, Ye et al., 2024.