1. Description dudataset

Le corpus a été conçu pour l'apprentissage et l'évaluation de systèmes de *question answering* (QA) sur des tables semi-structurées issues de Wikipédia. Il comporte 22 033 paires question–réponse réparties sur 2 108 tables HTML distinctes.

- Source : tables Wikipédia possédant au moins 8 lignes et 5 colonnes, extraites automatiquement.
- Format : chaque exemple contient un identifiant, une question en langage naturel (*utterance*), une table (*context*) et une ou plusieurs réponses exactes (*targetValue*). Les données sont fournies en TSV, avec métadonnées HTML/CSV associées et des annotations linguistiques (tokenisation, lemmatisation, NER) via CoreNLP.
- Splits officiels : *training* (14 152 ex.), *pristine-unseen-tables* (4 344 ex., test principal, tables jamais vues), *pristine-seen-tables* (3 537 ex.) et des sous-échantillons aléatoires pour la validation.

Les tables sont volontairement ouvertes : les colonnes et entités rencontrées en test sont différentes par rapport à l'apprentissage. Les questions requièrent souvent des opérations complexes : jointure de colonnes, agrégation (*count, sum*), superlatifs (*argmax/argmin*), comparaisons arithmétiques, voire composition de plusieurs règles logiques.

2. Biais:

1. **Biais de domaine** : Les tables proviennent uniquement de Wikipédia. Ça favorise des thématiques encyclopédiques (histoire, sport, géographie) et un style particulier de rédaction qui n'est pas forcément très humain.
2. **Biais linguistique** : Les questions ont été rédigées par des annotateurs anglophones d'Amazon Mechanical Turk avec des *prompts* incitant à la complexité (par ex. « the question should require calculation »). Cela peut induire un vocabulaire ou une grammaire artificiellement variés mais parfois peu naturels.
3. **Biais sur le prétraitement NLP:** Certaines tables contiennent des cellules ambiguës (listes, formats de date hétérogènes/non conventionnels) ou nécessitent une normalisation complexe (heures, notations sportives), ce qui pénalise différemment les systèmes selon leurs capacités de prétraitement.

3. Difficultés scientifiques:

- **Généralisation hors-vocabulaire** : colonnes et entités inconnues exigent des modèles capables de raisonnement symbolique ou de représentation sémantique flexible.
- **La complexité/difficulté combinatoire** : la recherche de formes logiques (*logical forms*) croît exponentiellement avec la profondeur de composition.
- **Ambiguïtés sémantiques** : plusieurs expressions peuvent produire le même résultat numérique, ce qui complique l'évaluation fondée uniquement sur la réponse finale.

4. Protocole expérimental proposé

Tâche: Pour une question et la table associée, prédire la ou les valeurs exactes.

Mesures d'évaluation:

- **Exact Match Accuracy** (*official evaluator*): proportion d'exemples dont la prédiction correspond exactement à la ou aux réponses cibles.
- **F1 macro** sur ensembles de réponses multiples (utile quand plusieurs cellules sont attendues).
- **Temps de calcul / taille du beam** pour évaluer la scalabilité face à la combinatoire.
- **Robustesse par type de question** : évaluer séparément sur les lookup simples, les superlatifs, l'arithmétique…
- **Temps / complexité** : comparer la vitesse d'inférence par exemple entre modèles neuronaux et parseurs symboliques.

Expériences envisagées

- *Encoder–decoder* (BERT ou T5) sur entrée concaténant question et table linéarisée.
- Parseur sémantique guidé par types pour tester la robustesse à la généralisation de schémas
- Comparaison entre apprentissage supervisé complet et apprentissage faible (uniquement paires question–réponse).
- Évaluer les performances par catégorie thématique (sport, politique, etc.) et par complexité logique (simple *lookup*, agrégation, superlatif, arithmétique) comme suggéré dans le papier.

Motivation

- Ce corpus reste une référence pour tester la capacité de raisonnement compositionnel au-delà du *text span QA*. Les données tabulaires sont omniprésentes (open data, rapports financiers) ; un système performant sur WikiTableQuestions pourrait s'adapter à des environnements hétérogènes.
- Confronter architectures purement neuronales et approches hybrides permet d'éclairer les limites actuelles du *semantic parsing*.