# MEthodology in Data Science

## Évaluation des LLMs et pertinence des benchmarks

Olivier Schwander
<olivier.schwander@sorbonne-universite.fr>

Master MIND
Sorbonne Université


2025-2026

# Au-delà des benchmarks et des scores



**Computer Science > Computation and Language**

*[Submitted on 3 Nov 2025]*

## Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H.S. Torr, Cozmin Ududec, Luc Rocher, Adam Mahdi

Evaluating large language models (LLMs) is crucial for both assessing their capabilities and identifying safety or robustness issues prior to deployment. Reliably measuring abstract and complex phenomena such as 'safety' and 'robustness' requires strong construct validity, that is, having measures that represent what matters to the phenomenon. With a team of 29 expert reviewers, we conduct a systematic review of 445 LLM benchmarks from leading conferences in natural language processing and machine learning. Across the reviewed articles, we find patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims. To address these shortcomings, we provide eight key recommendations and detailed actionable guidance to researchers and practitioners in developing LLM benchmarks.

Comments: 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks

# Pourquoi évaluer ?

## Rôle opérationnel

- Comparer plusieurs modèles
- Choisir un modèle pour une application
- Mesurer des progrès incrémentaux

## Rôle scientifique

- Tester des hypothèses sur les capacités des modèles
- Comprendre *pourquoi* un modèle réussit ou échoue
- Identifier les limites de généralisation

Réduire l'évaluation à un score revient à perdre l'objectif scientifique.

# Qu'est-ce qu'un benchmark ?

Un benchmark est une expérience standardisée, composée de :

- Une tâche : ce que le modèle doit produire
- Un jeu de données : quels exemples sont utilisés
- Une ou plusieurs métriques : comment les performances sont quantifiées
- Une interprétation : ce que l'on conclut à partir des scores

Même si elle n'est pas explicitée, l'interprétation fait partie du benchmark.

# Quelques benchmarks LLMs majeurs

## Compréhension du langage

- **GLUE / SuperGLUE** : agrégation de tâches NLP classiques pour mesurer la généralisation

## Raisonnement et connaissances

- **GSM8K** : résolution de problèmes mathématiques textuels
- **MMLU** : connaissances scolaires et professionnelles multi-domaines

## Capacités générales

- **BIG-bench** : large collection de tâches pour explorer des capacités émergentes

## Code

- **HumanEval / MBPP** : génération de code fonctionnel

# Validité de construit

## Définition (psychométrie)

- Un test possède une validité de construit s'il mesure effectivement le concept théorique qu'il prétend mesurer.
- Utile quand on ne peut pas mesure *directement* ce qui nous intéresse (intelligence, mémoire, raisonnement, anxiété), on a donc besoin de proxys.

## Transposition aux LLMs

- Un benchmark est valide si la performance obtenue est un indicateur fiable du phénomène abstrait ciblé.
- Sans validité de construit, les scores ont peu de sens.

# Remettre en question les benchmarks

*Que mesure-t-on réellement ?*

## Score élevé ?

- Véritable capacité cognitive
- Mémorisation massive
- Exploitation du format de la tâche
- Sur-optimisation de la métrique

Le benchmark seul ne permet pas toujours de trancher

# Biais dans les benchmarks

## Visual Question Answering

- (Historiquement) Souvent : pas besoin de regarder l'image…

## Exemples

- "Is there a clock in the room ?" → "Yes"
- "What color is the banana ?" → "yellow"

## Analyse : un bon score ne suffit pas

- Il faut comprendre pourquoi on répond juste
- Et comprendre qu'on répond juste pour de mauvaises raisons

## Solutions

- Observation des cartes d'activation en sortie des convolution
- Dataset moins biaisé

# Benchmark GSM8K

## Description officielle

*GSM8K (Grade School Math 8K) is a dataset of 8.5K high quality linguistically diverse grade school math word problems. The dataset was created to support the task of question answering on* ***basic mathematical problems*** *that require* ***multi-step reasoning****.*

- Format : problème textuel → réponse numérique
- Taille : ~8 500 questions
- Métrique standard : *exact match* sur la réponse finale

Objectif affiché : mesurer le raisonnement mathématique.

# Exemple GSM8K

**Question** :

```
Natalia sold clips to 48 of her friends in April, and then
she sold half as many clips in May. How many clips did
Natalia sell altogether in April and May?',
```

**Réponse** (utilisée uniquement en train) :

```
Natalia sold 48/2 = <<48/2=24>>24 clips in May.
Natalia sold 48+24 = <<48+24=72>>72 clips altogether
in April and May.
#### 72
```

**Sortie attendue** (utilisée pour l'évaluation) :

```
72
```

# Analyse GSM8K

## Évaluation exact-match

- Le raisonnement intermédiaire n'est pas évalué
- Seule la réponse finale compte

## Stratégies possibles pour le modèle

- Raisonnement étape par étape
- Reconnaissance de motifs fréquents
- Approximation numérique

## Manque de validité de construit

- Le benchmark ne distingue pas ces stratégies
- Le score ne garantit pas l'existence d'un raisonnement explicite

**Proxy partiel** du phénomène « raisonnement ».

# Phénomènes abstraits

## Objectifs de l'apprentissage

- Raisonnement
- Compréhension du langage
- Robustesse
- Alignement et sécurité

## Difficultés

- **Abstraits** : non directement observables
- **Multidimensionnels** : plusieurs sous-capacités
- **Contextuels** : dépendants des données et des tâches

# Manque de validité du construit

## Protocole d'évaluation machine learning

- Valide méthdologiquement (séparation train/test, etc)
- Bien décrit et formalisé
- Reproductible (split fournis, seeds, etc)

## Mais

- Phénomène mal défini ou trop vague
- Tâche couvrant seulement une partie du phénomène
- Métrique insensible à des comportements importants
- Agrégation de scores hétérogènes

**Un protocole valide ne fait pas un bon benchmark**

# Contamination des données

## Variante du sur-apprentissage

- Un bon score ne veut pas dire qu'on généralise bien
- Même avec un bon protococole d'évaluation
- On mesure la généralisation **dans** le benchmark mais pas **hors** de celui-ci

## Dangers

- Données de test accidentellement dans le train (ça peut arrive si on scrape son propre dataset...)
- Réponses déjà présentes dans les données (Wikipedia, StackOverflow)
- Exemples trop proches de ce qu'on trouve dans le train

**Ces risques augmentent avec le temps**

# Article Measuring what Matters

*Measuring what Matters : Construct Validity in Large Language Model Benchmarks*

## Question de recherche

Les benchmarks LLMs mesurent-ils réellement ce qu'ils prétendent mesurer ?

## Apport principal

- Introduction explicite de la validité de construit dans l'évaluation des LLMs
- Analyse systématique des benchmarks existants dans la littérature
- Propositions pour améliorer la qualité des benchmarks

# Méthodologie de l'étude

## Données

- Analyse de **445 benchmarks** publiés entre 2018 et 2024
- Conférences ML et NLP majeures
- Annonçant de nouveaux benchmarks pour les LLM
- Annotation qualitative via un *codebook*

## Dimensions analysées

- Définition du phénomène
- Tâches et données
- Métriques
- Interprétation des résultats

# Conclusions de l'article

## Constats récurrents

- Définitions absentes ou ambiguës
- Phénomènes contestés dans la littérature
- Tâches servant de proxies incomplets
- Peu de comparaisons avec des humains ou des baselines réalistes

## Implications scientifiques

- Risque de progrès illusoires
- Comparaisons trompeuses entre modèles
- Difficulté à comprendre le fonctionnement des modèles

**Une amélioration de score ne garantit pas une amélioration réelle de capacité**

# Recommandations des auteurs

## Règles de conception d'un benchmark

1. Définir clairement le phénomène ciblé
2. Évaluer uniquement ce phénomène
3. Construire un dataset représentatif
4. Faire attention à la réutilisation des données
5. Se préparer à la contamination
6. Utiliser des outils statistiques pour comparer les résultats
7. Étudier les erreurs du modèle
8. Analyser la validiter de construit

# Construire une projet de recherche

## Définir la question de recherche

## Choisir l'évaluation

- Quelle capacité votre méthode prétend-elle améliorer ?
- Comment cette capacité se distingue-t-elle d'autres ?
- Est-elle testable indirectement ?

On choisit les benchmarks en fonction de ce qu'on veut analyser (ou on en crée un).

# Conception des tâches

## Bonnes pratiques

- Utiliser une **famille de tâches**
- Introduire des variations contrôlées
- Tester des cas limites

Empêcher les stratégies de contournement (*shortcuts*).

## Exemple rapide

- Pour le VQA : tester avec des bananes vertes ou pas de banane

# Choisir les métriques

## En plus d'un score global

- Scores continus ou probabilistes : pas juste vrai/faux si possible
- Analyse par sous-catégorie : par classe, par langue, par tâche, par biais possible
- Étude qualitative des erreurs : regarder manuellement les cas d'erreur
- Tests statistiques

Les métriques et les benchmarks influencent directement ce que les modèles apprennent à optimiser.

## Loi de Goodhart

- Sciences sociales : *Quand une mesure devient un objectif, elle cesse d'être une bonne mesure*
- Ne pas regarder que le score
- Ne pas réutiliser tout le temps les mêmes données ou benchmarks

# Interpréter les résultats

## Questions indispensables

- Où le modèle réussit-il systématiquement ?
- Où échoue-t-il ?
- Ces résultats correspondent-ils au phénomène ciblé ?

## Interprétation

- Information sur le comportement du modèles
- Connaissances sur les contributions scientifiques proposées

# Benchmark GLUE

## Description officielle

*The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.*

- 9 tâches NLP classiques
- Classification, similarité, inférence
- Score agrégé

Objectif affiché : mesurer la **compréhension linguistique générale**.

# Exemple GLUE

## Tâche Natural Language Inference

**Prémisse** :

`A man is playing a guitar.`

**Hypothèse** :

`A person is making music`

**Sortie attendue** :

`entailment (implication)`

## Limites

- Tâches superficielles
- Fortes corrélations lexicales
- Saturé dès 2019 (c'est devenu trop facile)

# Benchmark SuperGLUE

## Description officielle

*In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. [...] We take into account the lessons learnt from original GLUE benchmark and present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.*

## Plus difficile

- Tâches plus complexes
- Moins de biais lexicaux
- Comparaison avec performance humaine

Objectif affiché : mesurer une **compréhension plus profonde**.

# Exemple SuperGLUE

## Tâche COPA (Choice of Plausible Alternatives)

**Prémisse** :

`The man broke his toe.`

**Question** :

`What was the CAUSE of this?`

**Réponse possibles** :

`1: He got a hole in his sock.`
`2: He dropped a hammer on his foot.`

# Analyse GLUE vs SuperGLUE

## Meilleure validité de construit ?

- Moins de biais lexicaux
- Tâches difficiles mêmes pour des humains
- Tâches plus délicates (COPA) nécessitant vraiment un raisonnement

## Limites

- Toujours des risques de bonne réponse lexicale
- “Understanding” toujours pas définit suffisament explicitement

# Benchmark HumanEval

## Description officielle

*The HumanEval dataset released by OpenAI includes 164 programming problems with a function sig- nature, docstring, body, and several unit tests. They were handwritten to ensure not to be included in the training set of code generation models.*

- Domaine : programmation Python
- Format : docstring → fonction
- Métrique : **pass@k** (tests unitaires)

Objectif affiché : évaluer le **raisonnement algorithmique**.

# Exemple HumanEval

**Entrée** :

```
The Brazilian factorial is defined as:
brazilian_factorial(n) = n! * (n-1)! * (n-2)! * ... * 1
```

**Sortie attendue** :

```
def brazilian(n):
    fact_i = 1
    special_fact = 1
        for i in range(1, n+1):
            fact_i *= i
            special_fact *= fact_i
    return special_fact
```

**Tests unitaires** :

```
assert candidate(4) == 288, "Test 4"
assert candidate(5) == 34560, "Test 5"
```

# Analyse HumanEval

## Limites

- Le code peut passer les tests sans être robuste
- La compréhension du problème n'est pas évaluée
- Forte dépendance aux tests fournis

## Pas différent du génie logiciel

- C'est pas parce que ça passe les tests que c'est correct

## Bonnes pratiques

- Exemples rédigés par des humains pour éviter la contamination
- Tests variés, avec des commentaires "simple case", "edge case"

# Benchmark MMLU

## Description officielle

*This is a massive multitask test consisting of multiple-choice questions from various branches of knowledge. The test spans subjects in the humanities, social sciences, hard sciences, and other areas that are important for some people to learn. This covers 57 tasks including elementary mathematics, US history, computer science, law, and more. To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability.*

- 57 domaines (maths, droit, médecine, histoire, etc.)
- Questions à choix multiples
- Comparaison directe avec des performances humaines

Objectif affiché : mesurer une **compréhension experte multi-domaines**.

# Exemple MMLU

## Exemple professional_medicine

**Question** :

A 67-year-old woman comes to the physician for a followup examination. She had a pulmonary embolism [...]. Which of t following is the most likely cause of this patient's decrea sensation?

**Choix** :

- Cerebral infarction during the hospitalization
- Complication of the IVC filter placement
- Compression of the lateral femoral cutaneous nerve
- Hematoma of the left thigh

**Réponse attendue** :

# Analyse MMLU

## Manque de validité de construit

- Mesure principalement la **récupération d'informations**
- Forte sensibilité à la contamination des données
- Peu d'évidence de raisonnement profond

## Bonnes réponses

- Mais pas forcément de compréhension du sens des questions

# Benchmark BIG-bench

## Description officielle

*The Beyond the Imitation Game Benchmark (BIG-bench) is a collaborative benchmark intended to probe large language models and extrapolate their future capabilities.*

- +200 tâches
- Créées par une large communauté
- Inclut des tâches peu conventionnelles

Objectif affiché : explorer ce que les LLMs peuvent ou ne peuvent pas faire.

# Exemple BIG-bench

## Exemple `self_awareness/assess_own_capabilities`

**Entrée** :

```
Can you slightly modify the universal gravitational constan
```

**Sortie attendue** :

```
No
```

## Exemple `physical_intuition`

**Entrée** :

```
 The bonds in gold are of what type?
```

**Réponses possibles** :

```
Ionic
```

# Analyse BIG-bench

## Forces

- Grande diversité de phénomènes
- Exploration qualitative riche
- Utilise un canari pour détecter la contamnination
  https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/persian_idioms

## Limites

- Tâches hétérogènes
- Scores difficiles à agréger
- Interprétation souvent narrative

# Évaluation et généralisation

## Une forme sur-apprentissage

- Évaluation valide du point de vue machine learning
- Donc en théorie on mesure bien la généralisation
- Mais en fait on risque de sur-apprendre surles benchmarks

## Pour généraliser

- Le score ne suffit pas
- Le but n'est pas d'améliorer un score
- Éviter d'utiliser toujours les même benchmarks
- Et les mêmes données même avec des tâches ou des métriques différentes

**Un modèle avec une bonne validité de construit mesure vraiment la généralisation**

# Comment juger un travail

## Dès qu'un benchmark est utilisé

- dans un article que vous lisez
- dans un article à reviewer
- dans votre propre travail

## Questions critiques

- Le phénomène est-il défini clairement ?
- Les tâches sont-elles justifiées ?
- Les métriques sont-elles appropriées ?
- Les conclusions sont-elles raisonnablement étayées ?

# Conclusion

## Importance de l'évaluation

- Aussi important que le modèle et l'apprentissage
- Donne du sens à ce qu'on fait
- Pour comprendre ce qui marche et ce qui ne marche pas
- Pour améliorer les performances
- Pour en tirer des conclusions scientifiques

## Le score ne suffit pas

- Donner du sens au score
- Analyser les erreurs
- Varier les tâches, les données, les sorties