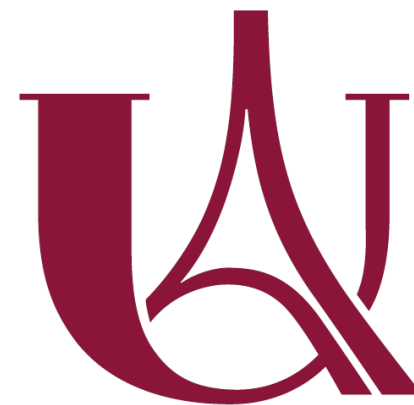


Explorer les capacités émergentes dans ChessGPT

Luc POMMERET

Université Paris Cité — IRIF

Encadré par Michel de Rougemont



**INSTITUT
DE RECHERCHE
EN INFORMATIQUE
FONDAMENTALE**

Perroquets stochastiques vs. Capacités émergentes

- “Perroquet stochastique” est une manière de désigner les LLMs comme incapables de dépasser ses données d’entraînement (Bender et al. 2021) 🦜
- Depuis, cette analogie a été battue en brèche par de nombreux articles qui vont dans le sens d’une émergence de capacités et d’un **dépassement** du niveau des données d’entraînement (Arora et al. 2023, Edelman et al. 2024)



LLM : Large Language Model

- Prédit le prochain mot en fonction des mots précédents, c'est la fonction `next_token`.
- Prédiction probabiliste, cela veut dire que le prochain token est tiré dans la distribution du prochain token.
- Permet d'adopter des stratégies : espace des stratégies plus vaste que l'espace des stratégies de Stockfish, ou AlphaZero. Ici le `next_move` est une itération du `next_token`.
- Permet d'adopter des stratégies très diverses, on peut par exemple vouloir perdre aux échecs !

« LLM », « transformers », késako ?

- Une liste de *tokens*, ou unité syntaxique minimale, ce sont les « mots » que notre modèle va voir.

Tokens

6

Characters

17

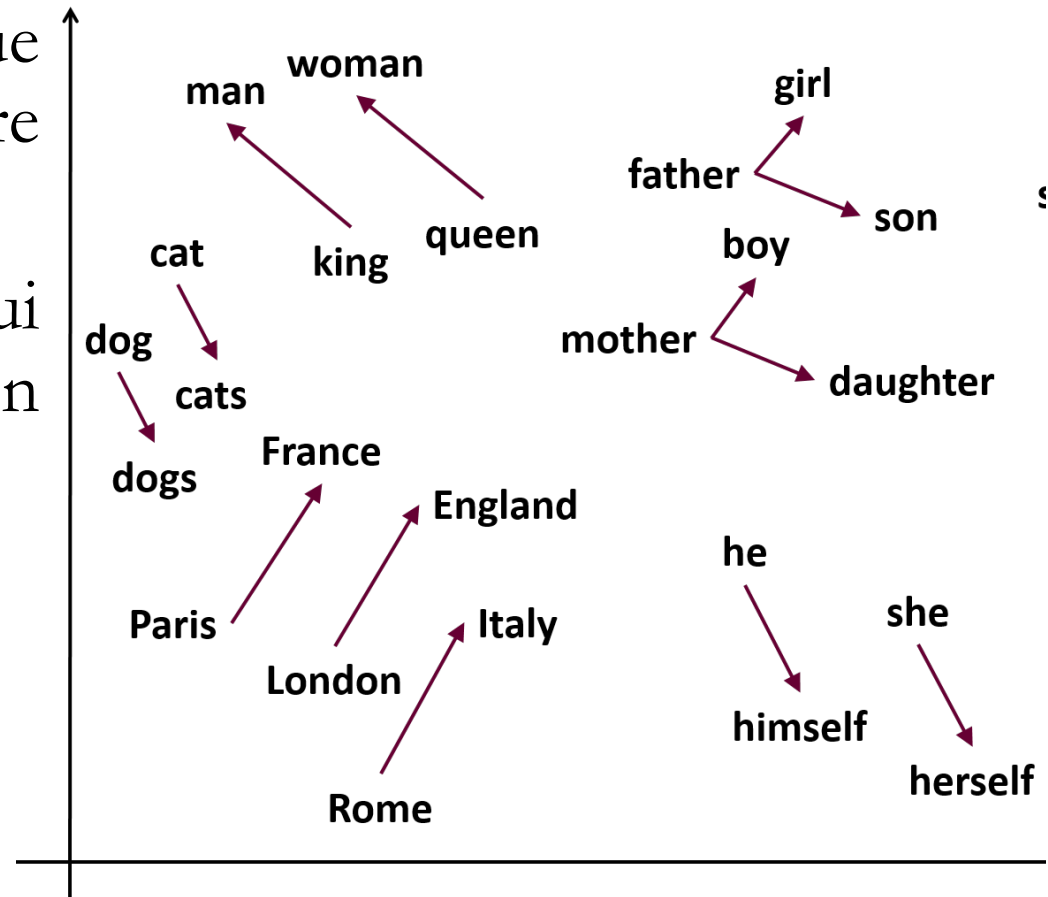
ChatGPT is great!

TEXT

TOKEN IDS

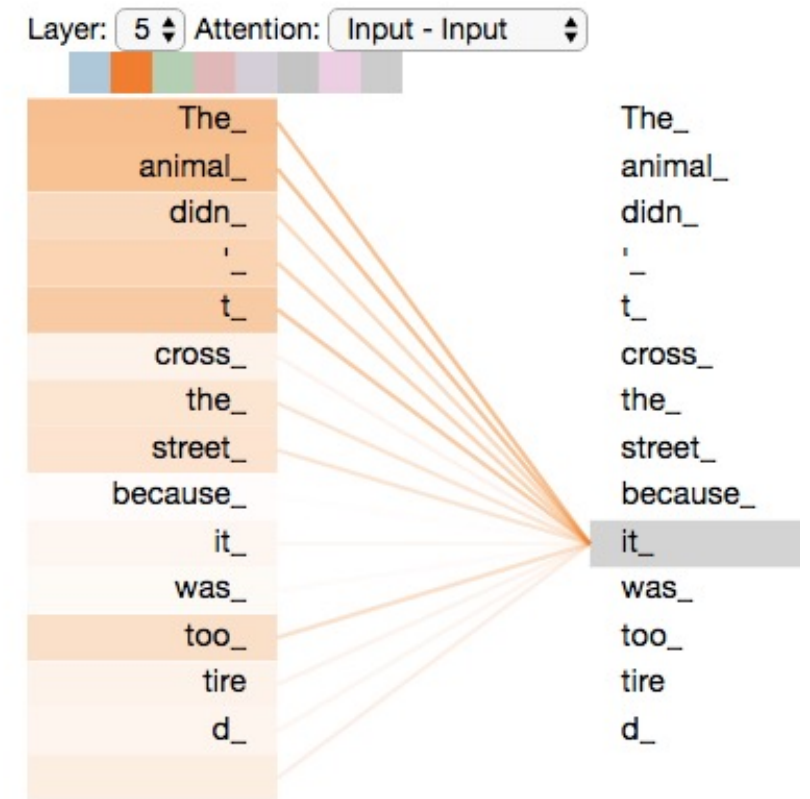
« LLM », « transformers », késako ?

- Une liste de *tokens*, ou unité syntaxique minimale, ce sont les « mots » que notre modèle va voir.
- Un mécanisme d'*embedding* des mots, qui encode leur sémantique en grande dimension (technique qui existe depuis word2vec).



« LLM », « transformers », késako ?

- Une liste de *tokens*, ou unité syntaxique minimale, ce sont les « mots » que notre modèle va voir.
- Un mécanisme d'*embedding* des mots, qui encode leur sémantique en grande dimension (technique qui existe depuis word2vec).
- Un mécanisme d'attention, qui permet d'avoir la grammaire, les dépendances entre les mots.



Un peu de vocabulaire échiquéen

- Stockfish : moteur d'échecs, actuellement le meilleur sur le marché, à plus de 3600 Elo
- Niveau Elo : score reflétant la probabilité de l'issue d'une partie d'échecs entre deux joueurs



Plan de la présentation

- Notation PGN
- ChessGPT
- Introduction de trois propriétés dont on étudie l'émergence :

1. Légalité des coups

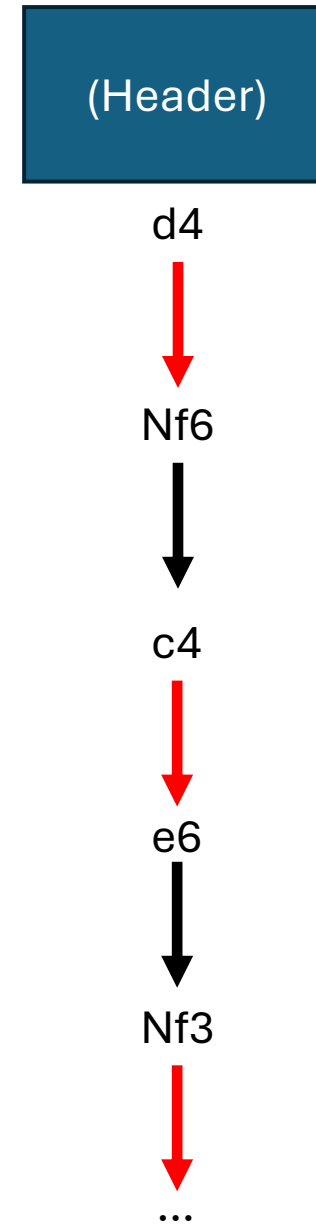
2. Résolution de puzzles

3. Prédiction de la position de l'échiquier (ici par probing)

La notation PGN aux échecs

```
[Event "Simulated Game"]  
[Site "Reykjavik"]  
[Date "1972.06.11"]  
[Round "1"]  
[White "Boris Spassky (2660)"]  
[Black "Bobby Fischer (2785)"]  
[Result "1-0"]
```

```
1. d4 Nf6  
2. c4 e6  
3. Nf3 d5  
4. Nc3 Bb4  
5. e3 0-0
```

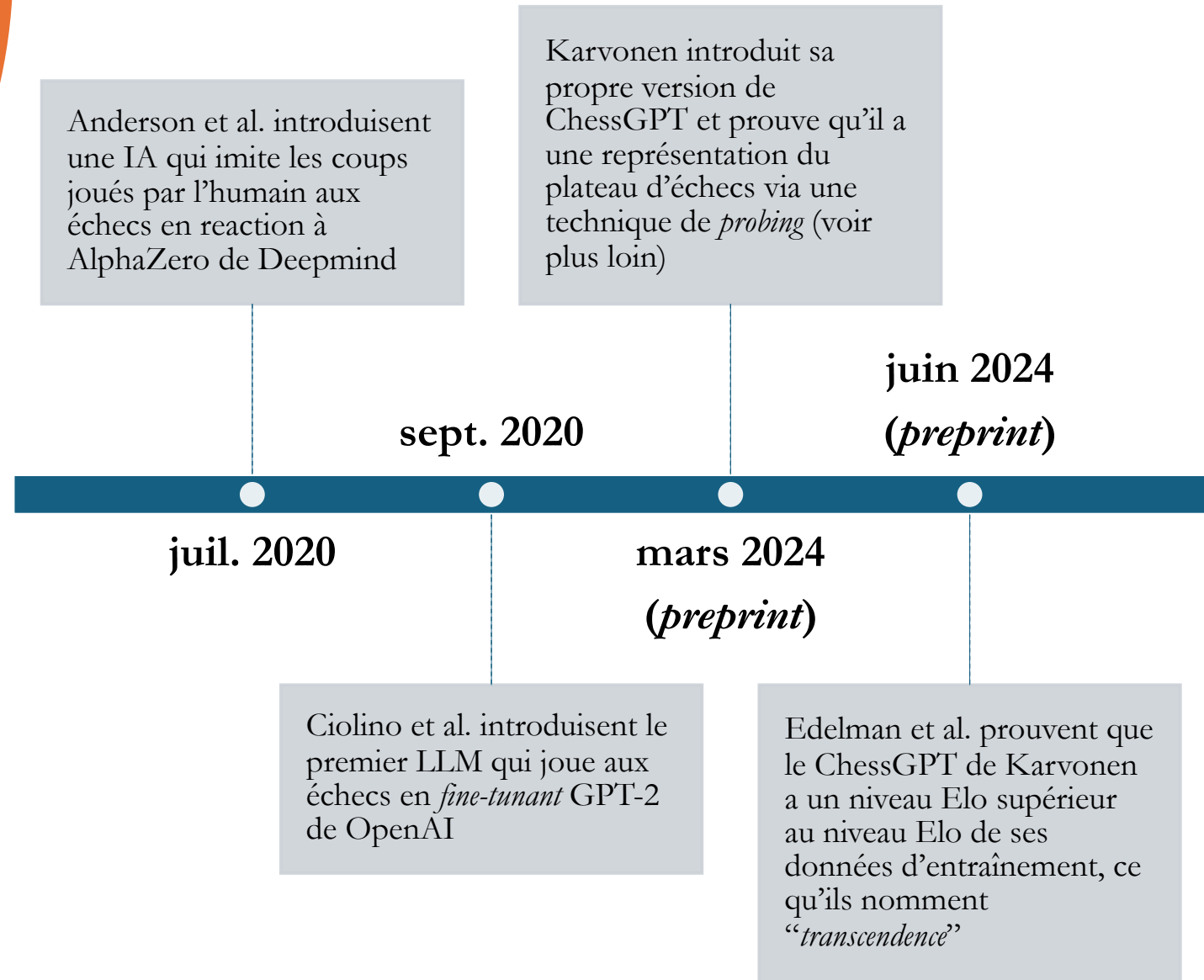


ChessGPT, le LLM qui joue aux échecs

- Grand modèle de langage qui a appris à jouer sur des milliers de parties d'échecs en notation PGN (1.e4 e5 2.Cf3 Cc6 3...)
- Il s'agit d'un modèle de langage monotâche, il ne fait que jouer aux échecs, en complétant le début de partie qu'on lui donne.
- Utile car cela permet de fixer de nombreux paramètres qui sont difficiles à fixer lorsqu'on étudie directement le langage naturel.

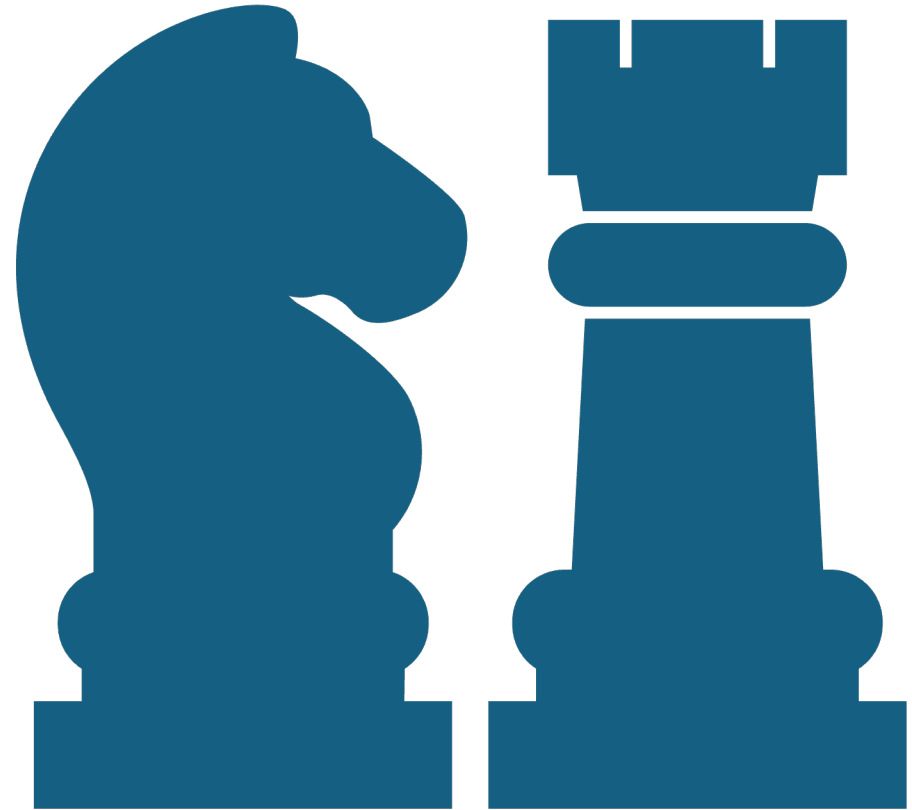
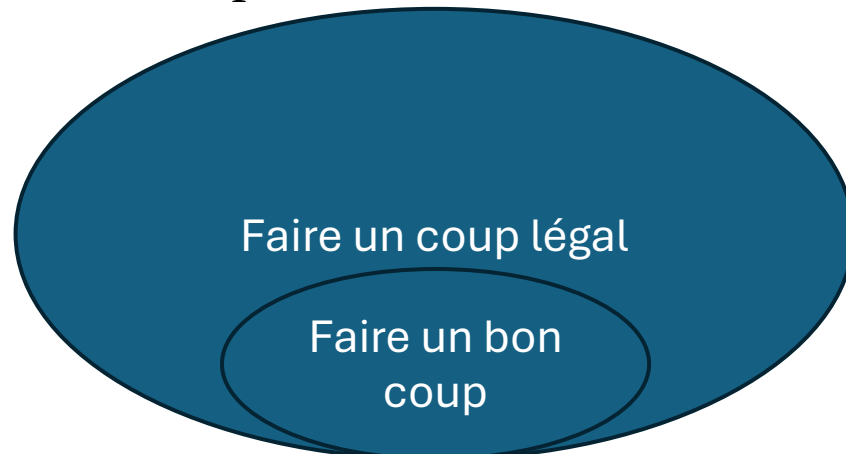


Quelques jalons de la littérature



Coup légal (legal move skill)

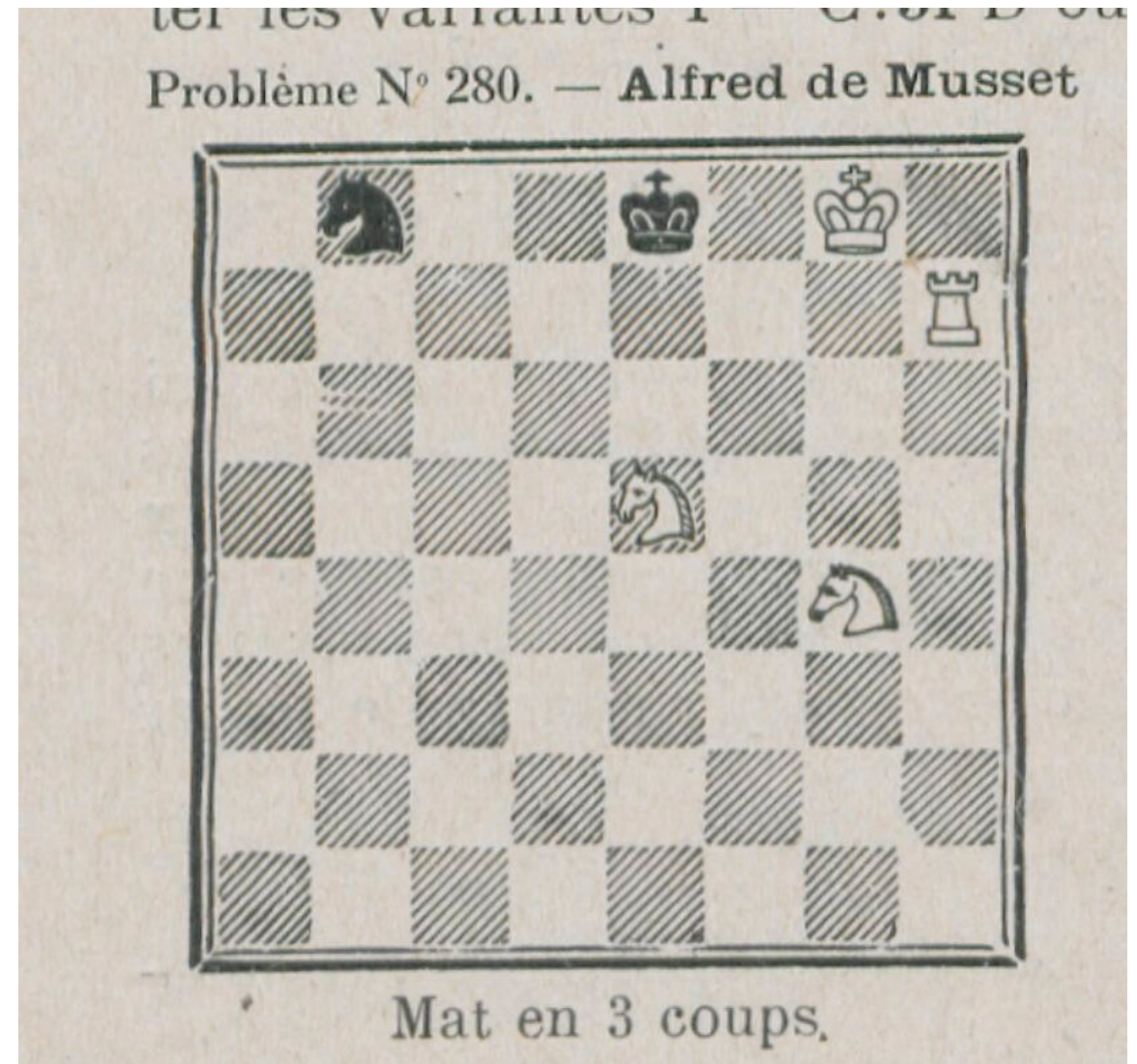
- Aux échecs, un coup légal est un coup qui étant donné une position peut être joué sans enfreindre les (nombreuses) règles du jeu d'échecs.
- Nous voulons d'abord tester ce skill, qui est une condition pour tester les suivants :



Problèmes d'échecs

(Puzzle solving skill)

- La résolution de problèmes est une tâche particulière, où étant donné une position sur l'échiquier, il faut trouver le coup gagnant.
- C'est un *skill* dont nous pouvons facilement contrôler l'utilisation, et est donc intéressant à étudier chez ChessGPT
- On donne en entrée un historique de partie en PGN



Sondes (probes)

- Une sonde au sens de Karvonen est un classifieur linéaire entraîné sur les activations des différentes couches du réseau de neurones de ChessGPT :

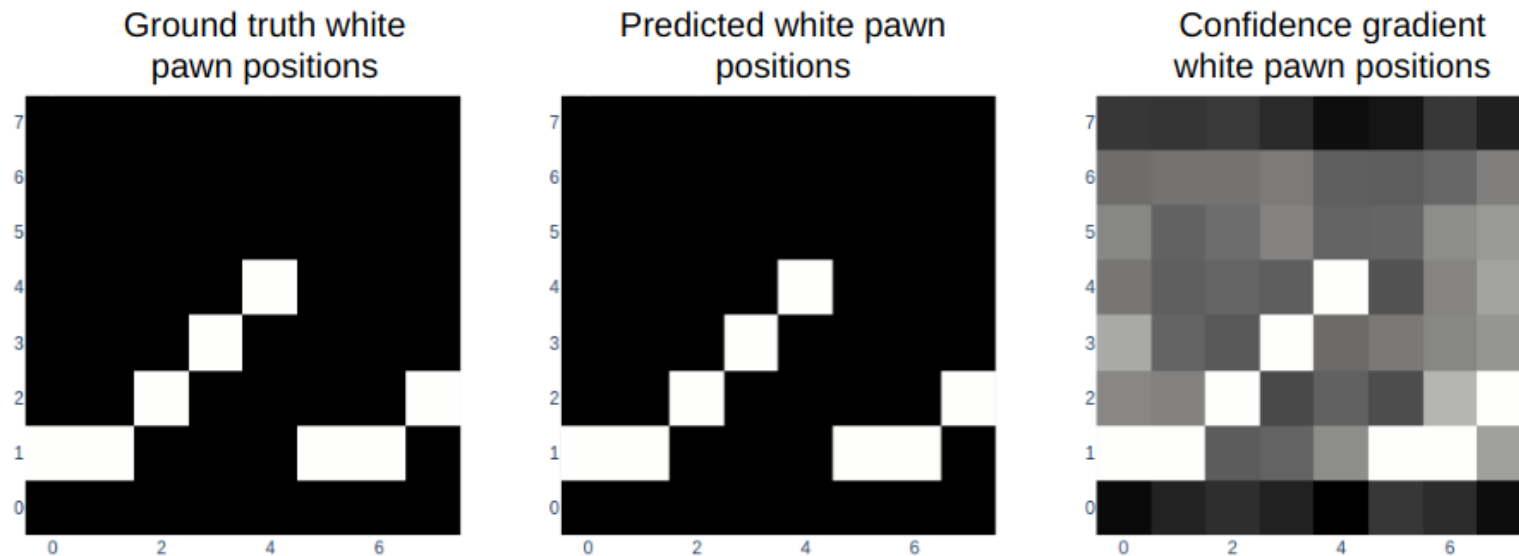
$$P_{i,l} = \text{Softmax}(W_{i,l} \cdot A_l)$$

$$g(x) = f(w^T x + w_0) = f \left(\sum_{j=1}^N w_j x_j + w_0 \right)$$

- $P_{i,l}$: Distribution de probabilité pour la case i prédite par la sonde à la couche l
- $W_{i,l}$: Matrice de poids de dimension 512×13
- A_l : Vecteur d'activation de dimension 512
- f : Fonction d'activation (Softmax)

Sondes (probes)

- Une sonde au sens de Karvonen est un classifieur linéaire entraîné sur les activations des différentes couches du réseau de neurones de ChessGPT
- On peut retrouver les positions de l'échiquier seulement en regardant les activations de ChessGPT :



Sondes (probes)

- Une sonde au sens de Karvonen est un classifieur linéaire entraîné sur les activations des différentes couches du réseau de neurones de ChessGPT
- On peut retrouver les positions de l'échiquier seulement en regardant les activations de ChessGPT
- ChessGPT a donc une représentation interne (*World Model*) de l'échiquier. Il joue ses coups en se représentant l'échiquier

Les vertus du bruit en machine learning

- Dimakis et al. (2020) montre que dans le domaine des images, entraîner un réseau de convolution avec du bruit permet d'avoir de meilleurs résultats avec un nombre d'images d'entraînement plus faible !

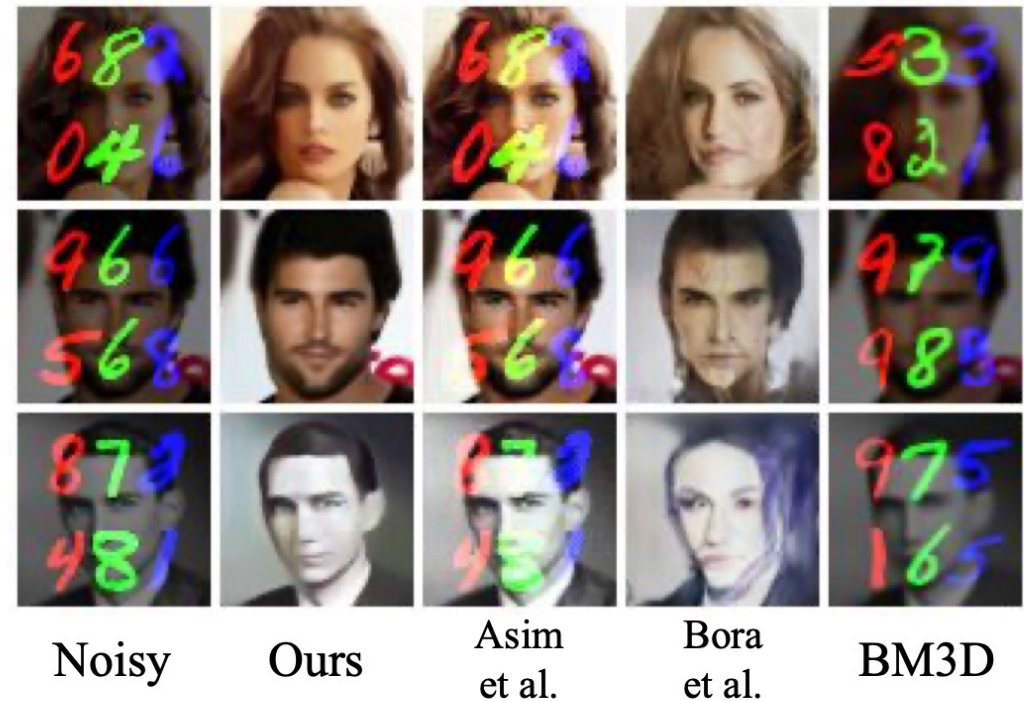
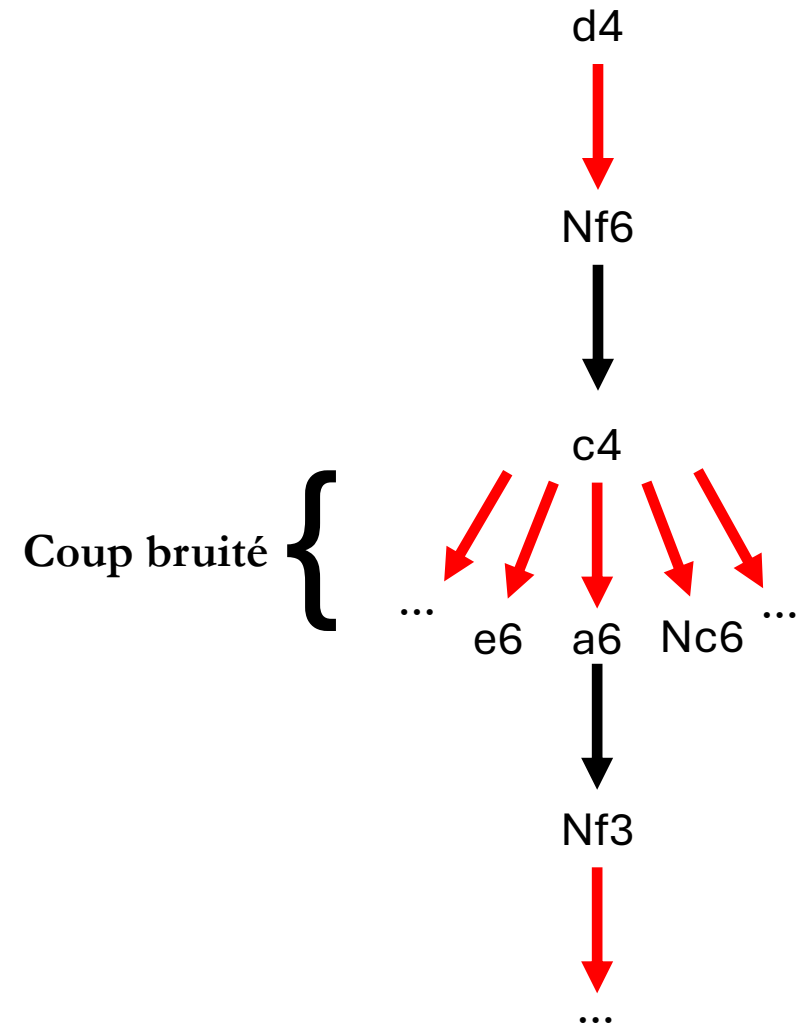


Figure 1. Result of denoising MNIST digits. The first column contains noisy observations, and subsequent columns contain reconstructions.

Génération de parties synthétiques avec un taux de bruit fixé

- Je définis le bruit aux échecs comme le fait de jouer un coup légal mais aléatoire (uniforme)
- Par exemple, pour le premier trait des blancs, il y a 20 possibilités.
- Le taux de bruit est la chance que le coup joué soit bruité
- Un taux de 5% de bruit signifie donc que chaque coup a 5% de chance d'être bruité (et donc 95% de chance d'être déterministe)
- Ici, le moteur d'échecs (non bruité) est Stockfish



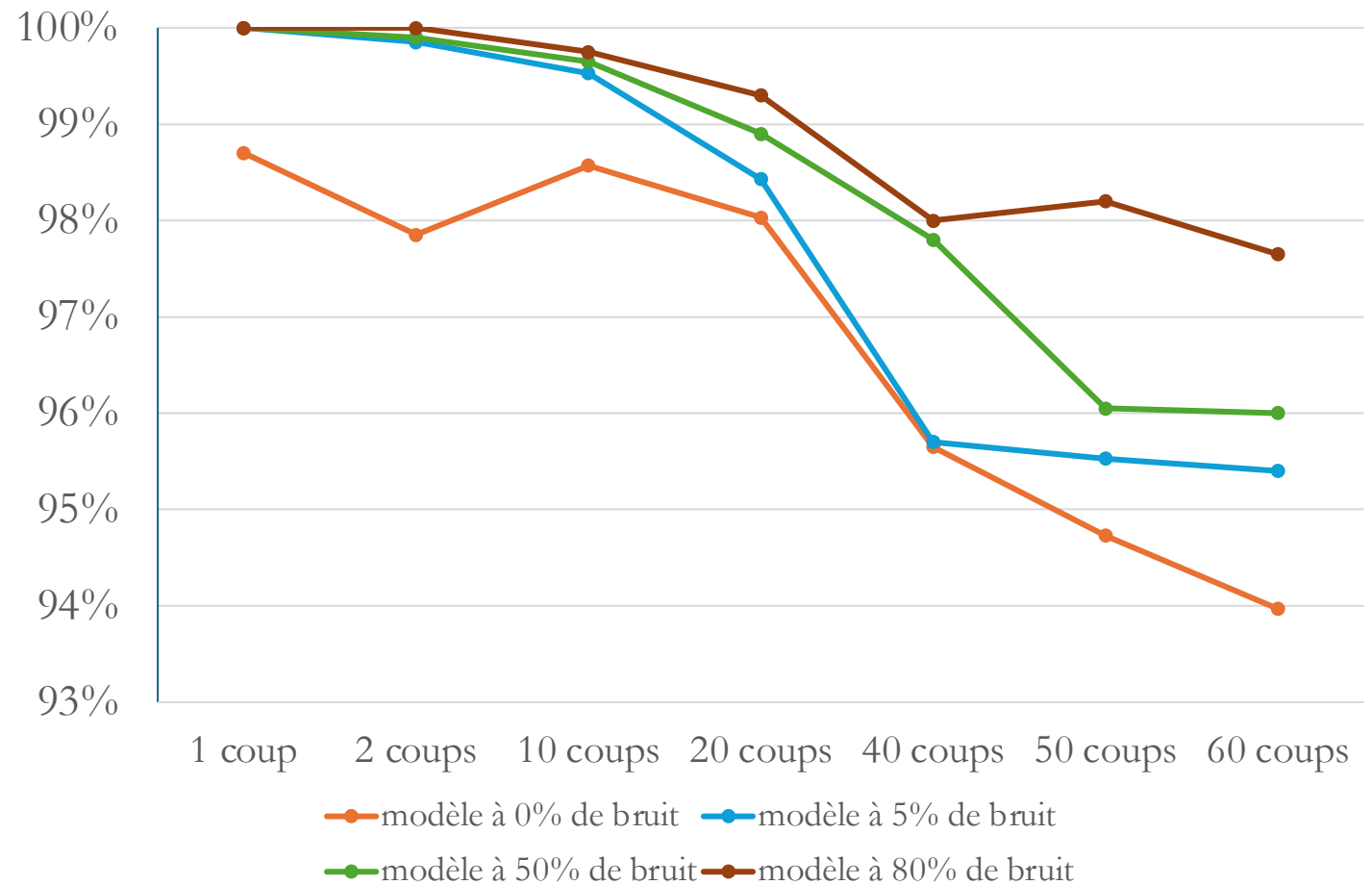


Tout le code des
expériences est disponible
à cette adresse :

https://github.com/zualmerul/chessGPT_noise.git

Impact du bruit sur les capacités de ChessGPT

Taux de coups légaux en fonction de la quantité de bruit dans les données d'entraînement du modèle et du nombre de coups considérés

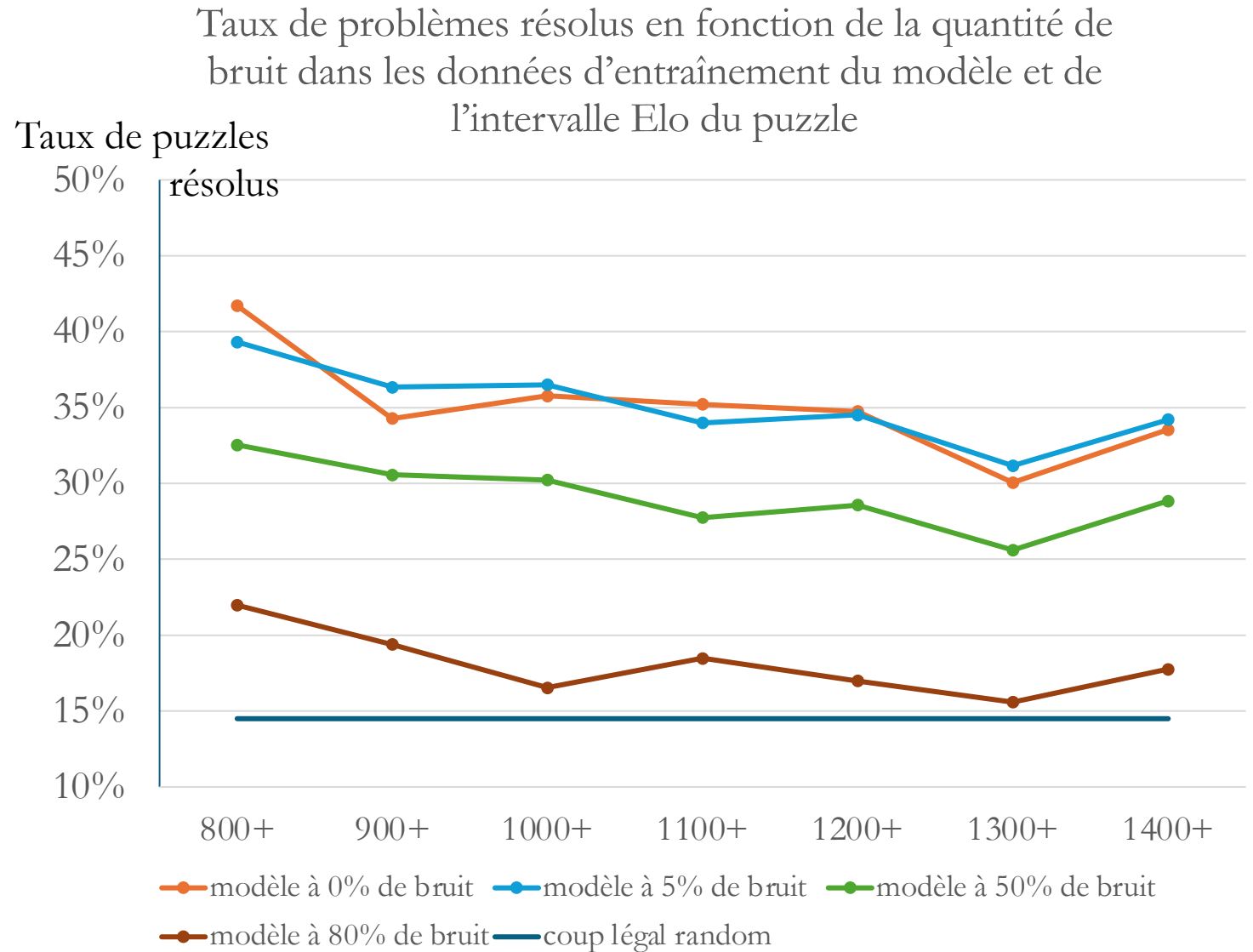


Échantillon par modèle : 10 000
20

Interprétation

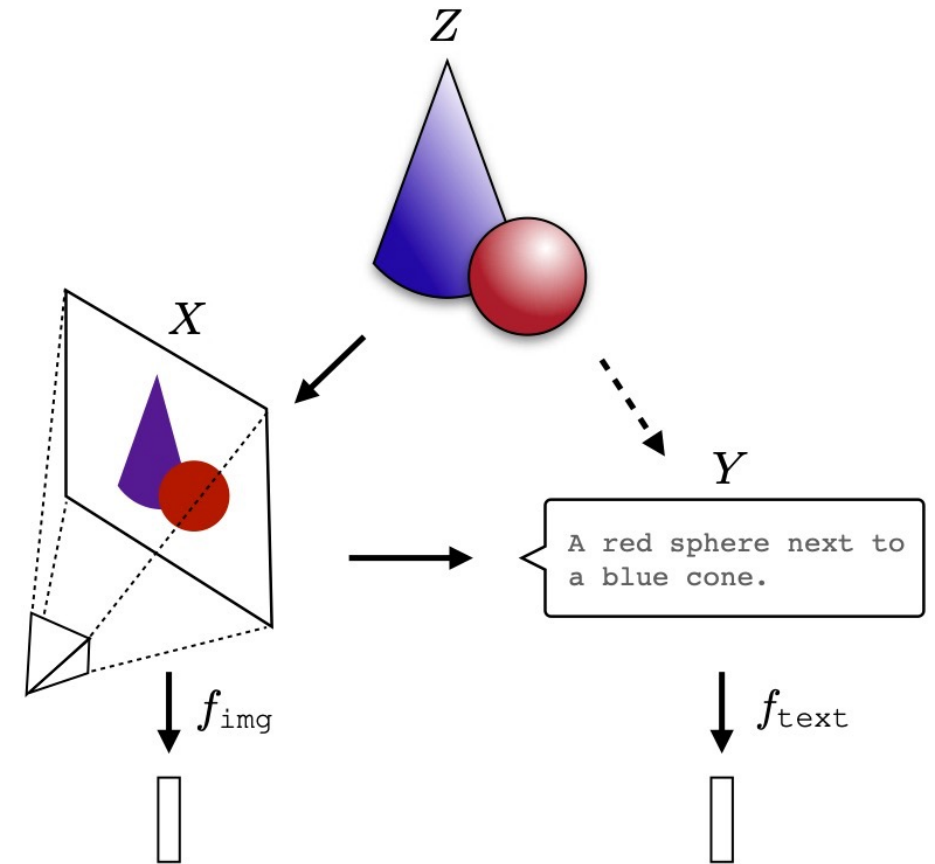
- Les LLMs entraînés sur du bruit ne sont entraînés que sur la tâche « faire un coup légal », tandis que les LLMs déterministes sont entraînés sur la tâche « faire un coup légal qui soit aussi un bon coup ».
- La seconde tâche est plus difficile à apprendre

Impact du bruit sur les capacités de ChessGPT



Travaux à venir

- Étudier la *Platonic Hypothesis Representation* (PRH, Isola et al. 2024), qui affirme que des réseaux de neurones entraînés sur des medias différents convergent vers la même représentation du monde.
- Utiliser ChessGPT pour tester cette hypothèse, étendre l'entraînement à un réseau de convolution (pour les images de l'échiquier)
- Enquêter pour voir si le bruit fait converger plus rapidement la représentation



Le sens du jeu

- Dans *Le joueur d'échecs* (1943), Stefan Zweig raconte l'apprentissage (forcé) d'un autrichien dans sa cellule de centaines de parties dans le format PGN. Cela le mène tout droit à la folie et finit par le libérer
- Dans *Le Joueur* (1866), Dostoïevsky donne les raisons pour lesquelles on adopte une stratégie, et montre qu'elles ne sont pas internes au jeu, mais plutôt contextuelles. Par exemple, la grand-mère joue tout son argent de manière impulsive... pour éviter que son neveu, le Général, touche son héritage.

Avez-vous des
questions ?

Merci pour votre
attention !

