

Foundations Models: Is Transfer Learning a solved problem?

What model of the world do they have?

Antoine Cornuéjols

AgroParisTech – INRAE MIA Paris-Saclay

EKINOCS research group

Universal representations?

A universal representation for texts?

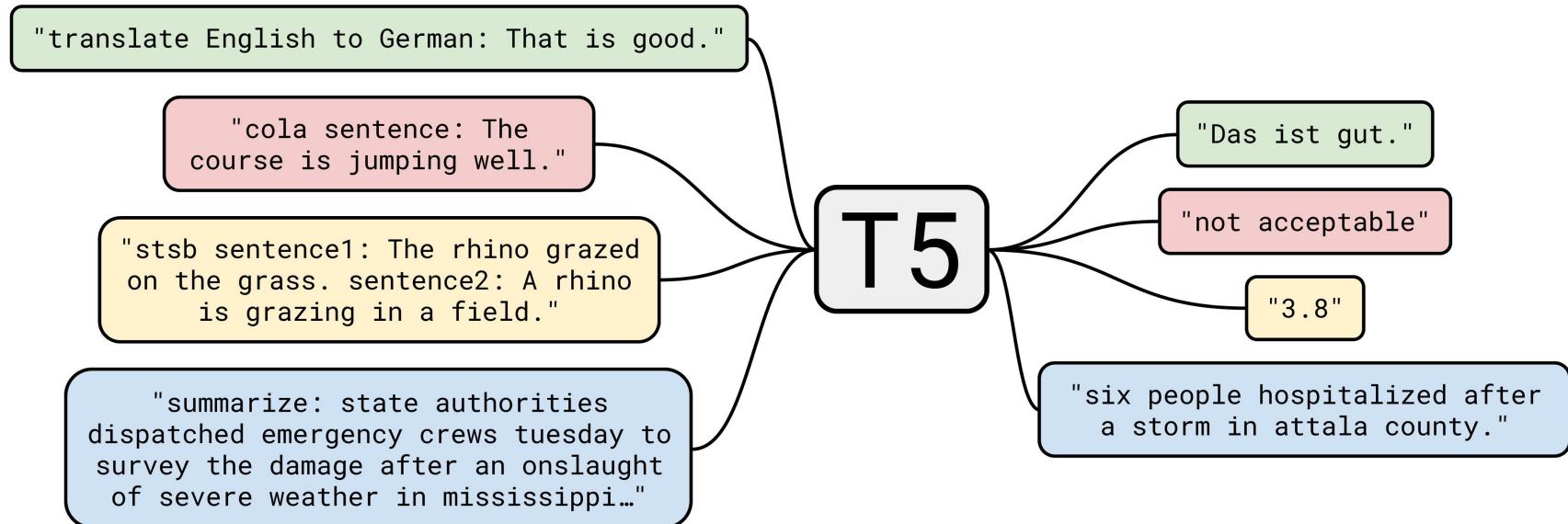


Figure 15.36: Illustration of how the T5 model (“Text-to-text Transfer Transformer”) can be used to perform multiple NLP tasks, such as translating English to German; determining if a sentence is linguistic valid or not (CoLA stands for “Corpus of Linguistic Acceptability”); determining the degree of semantic similarity (STS-B stands for “Semantic Textual Similarity Benchmark”); and abstractive summarization. From Figure 1 of [Raf+20]. Used with kind permission of Colin Raffel.

From [Kevin Murphy. **Probabilistic Machine Learning. An introduction.** MIT Press. (2022)], p.539

General questions

- 1. What does a foundation model know?**

- 2. What can a foundation model do?**

- 3. How does it know something?**

Model of the world?

Un modèle du monde physique ?

- On remplit un ballon de baudruche de farine, et on l'y insère également une grosse bille en plomb, puis on gonfle le ballon et on l'attache à une ficelle de 10 cm, elle-même attachée à un clou au plafond à 2,5 m ; juste sous le ballon se trouve une coupe de cristal fin de 15cm de hauteur ; un ventilateur de 1m de diamètre souffle en direction de la coupe de cristal, à environ 1 m de celle-ci ; plus loin sur le même axe est suspendu un drap dont la partie supérieure est très humide et la partie inférieure est sèche. Je tire un plomb de carabine sur le ballon. Que va-t-il se passer ?

Bonne question — c'est un **enchaînement physique** un peu à la "expérience de pensée", pas un mode d'emploi, donc on peut l'analyser calmement.

Je vais dérouler **ce qui se passe très probablement, étape par étape**, en restant réaliste physiquement.

1 Impact du plomb de carabine sur le ballon

- Le plomb **perce le ballon quasi instantanément**.
- Le latex se déchire et la tension disparaît **brutalement**.
- L'air sous pression s'échappe, mais **ce n'est pas l'air qui domine**, c'est la gravité.

👉 Le ballon ne "saute" pas, il s'effondre.

...

2 Comportement du contenu (farine + bille de plomb)

◆ La bille de plomb

- Masse élevée, aucune portance.
- Elle **chute immédiatement** verticalement.
- Temps de chute depuis 2,5 m :

$$t \approx \sqrt{2h/g} \approx 0,7 \text{ s}$$

👉 Elle tombe **presque à la verticale**, très peu influencée par le ventilateur.

◆ La farine

- Se disperse en **nuage fin**.
- Subit fortement :
 - la turbulence de l'air
 - le souffle du ventilateur
- Une partie tombe, une autre est **emportée horizontalement**.

3 Sort de la coupe de cristal (juste sous le ballon)

Cas de la bille de plomb

- Si la coupe est **exactement sous le ballon** :
 - la bille **tombe dedans ou la heurte**
 - une coupe de cristal fin se fissure ou se brise presque sûrement
- Si elle tombe **au bord** :
 - risque très élevé de casse par choc ponctuel

👉 La coupe ne survit probablement pas.

4 Effet du ventilateur sur la farine

- Le ventilateur (1 m de diamètre à 1 m de distance) crée :
 - un flux turbulent
 - une zone de mélange air/farine
- La farine devient un **aérosol temporaire**
- Elle est projetée :
 - vers la coupe
 - puis **au-delà**, dans l'axe du flux

...

5 Interaction avec le drap (humide en haut, sec en bas)

Partie humide

- La farine adhère fortement :
 - absorption d'humidité
 - agglomération
- Elle forme :
 - des taches pâteuses
 - des amas collants

Partie sèche

- La farine :
 - rebondit
 - glisse
 - tombe plus facilement

👉 Résultat visuel :

- Drap taché en haut
- Dépôt diffus en bas
- Gradient très net entre les deux zones

...

6 Résumé global (ce qui se passe vraiment)

- ✓ Le ballon éclate
 - ✓ La bille tombe et **casse très probablement la coupe**
 - ✓ La farine forme un nuage
 - ✓ Le ventilateur projette la farine
 - ✓ Le drap humide capture la farine, le sec beaucoup moins
 - ✓ Il reste :
 - un sol sale
 - un drap taché
 - une coupe probablement détruite
-

🧠 Interprétation intéressante

Ce scénario illustre très bien :

- la **différence entre masse et surface**
- la **séparation inertielle** (bille vs farine)
- l'effet de l'humidité sur les particules
- la dominance de la gravité sur les objets denses

En clair :

|| **Ce n'est pas une explosion, c'est une désagrégation suivie d'une chute et d'une dispersion.**

...

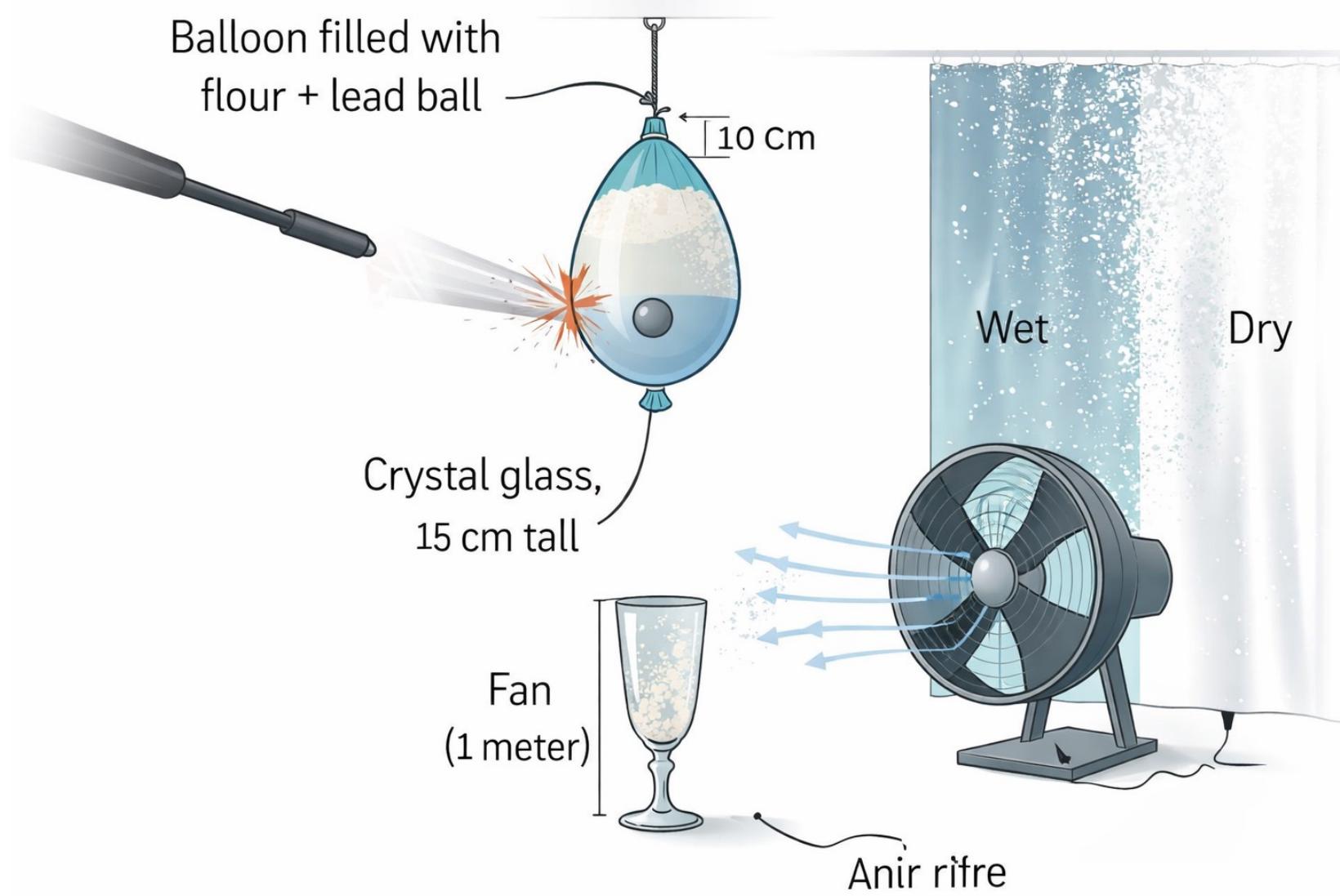
Ce qui m'interpelle

Ce qui m'interpelle

- C'est que chatGPT ne s'étonne pas que la partie humide du drap soit en haut.

Mais quand même:

Est-ce que chatGPT a un **modèle du monde physique**?



Sources

- GRINBAUM, Alexei. **La parole aux machines**. Humen sciences, 2023.
- Acher, Mathieu (2023). **Debunking the Chessboard: Confronting GPTs Against Chess Engines to Estimate ELO Ratings and Assess Legal Move Abilities**. Blog.mathieuacher.com
- Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., ... & Wu, J. (2023). **Weak-to-strong generalization: Eliciting strong capabilities with weak supervision**. arXiv preprint arXiv:2312.09390.
- Carlini , Nicholas (2023). **Playing chess with Large Language Models**.
<https://nicholas.carlini.com/writing/2023/chess-llm.html>.
- Karvonen, Adam (2024). "**Chess-GPT's Internal World Model**".
https://adamkarvonen.github.io/machine_learning/2024/01/03/chess-world-models.html
- Karvonen, Adam (2024). "**Emergent world models and latent variable estimation in chess-playing language models**." arXiv preprint arXiv:2403.15498 (2024).
- Li, Kenneth. "**Do Large Language Models learn world models or just surface statistics?**" The Gradient (2023).
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). **Emergent world representations: Exploring a sequence model trained on a synthetic task**. arXiv preprint arXiv:2210.13382.
- Nanda, Neel. "**Actually, othello-GPT has a linear emergent world representation**." Neel Nanda's Blog 8 (2023).
- RUOSS, Anian, DELÉTANG, Grégoire, MEDAPATI, Sourabh, et al. **Grandmaster-level chess without search**. CoRR, 2024.



General questions

- 1. What does a foundation model know?**

- 2. What can a foundation model do?**

- 3. How does it know something?**

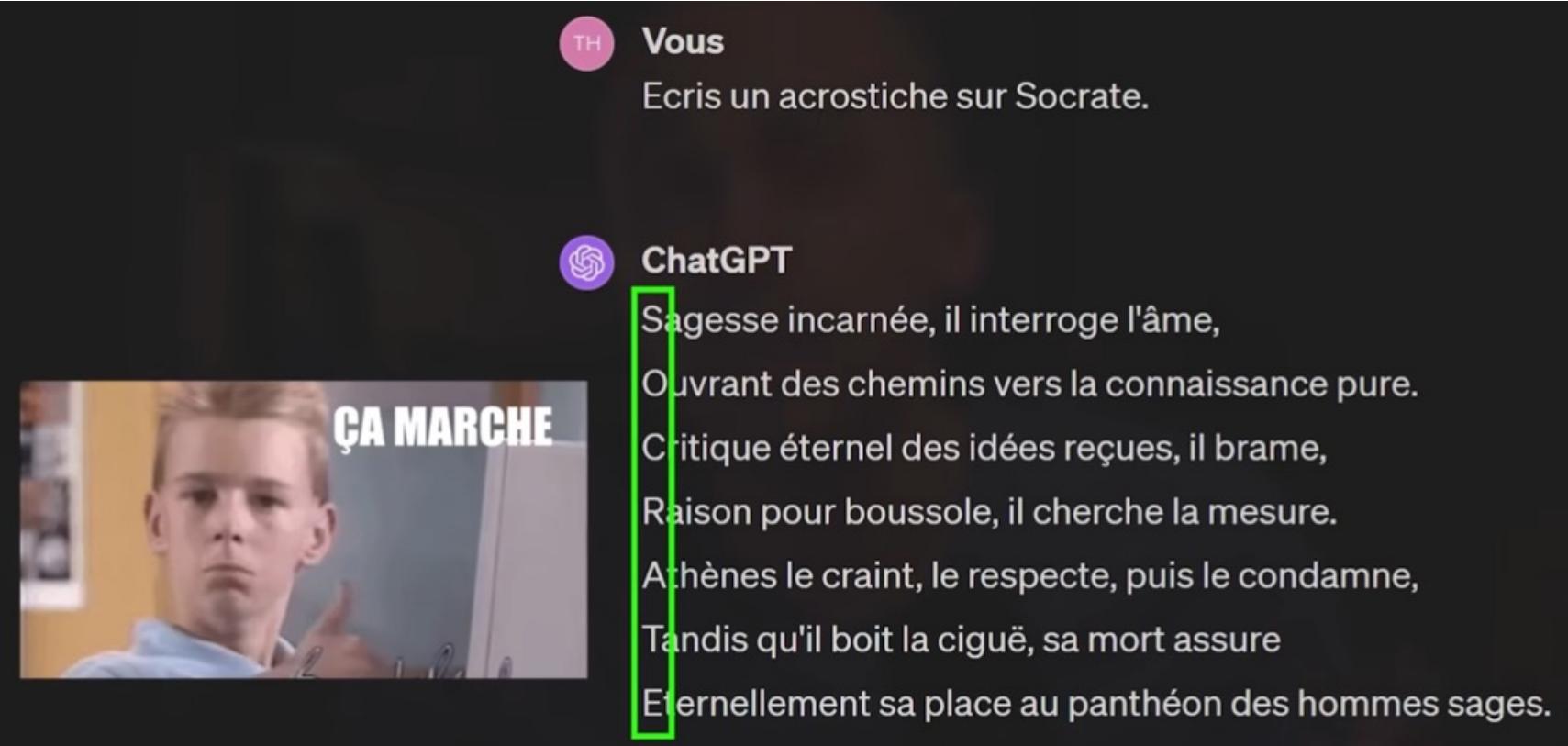
- Very simple: ask the system to solve the task!



Vous

Ecris un acrostiche sur Socrate.

- Very simple: ask the system to solve the task!



Vous
Ecris un acrostiche sur Socrate.

ChatGPT

Sagesse incarnée, il interroge l'âme,
Ouvrant des chemins vers la connaissance pure.
Critique éternel des idées reçues, il brame,
Raison pour boussole, il cherche la mesure.
Athènes le craint, le respecte, puis le condamne,
Tandis qu'il boit la ciguë, sa mort assure
Eternellement sa place au panthéon des hommes sages.

- Very simple: ask the system to solve the task!

Vous
Write a haiku without the letter "e".

ChatGPT

A full moon at night
Owls soar past, soft wings in flight
Dark turns to bright light ●

Necessary methodological precautions

- Very simple: ask the system to solve the task!

Row 1:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Autumn wind's soft song,
Branch tips sway, no words at all,
Moonlit night, so long.
- You:** Write a haiku without the letter "e".

Row 2:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Moon's soft light on snow,
Dark night, stars in crisp air glow,
Silent world below.
- You:** Write a haiku without the letter "e".

Row 3:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Autumn wind so brisk,
Maples turn a crimson hue,
Frost hugs dawn's first light.
- You:** Write a haiku without the letter "e".

Row 4:

- You:** Write a haiku without the letter "e".
- ChatGPT:** A soft rain at dawn
Birds sing atop tall oaks
Mist hugs grassy knolls
- You:** Write a haiku without the letter "e".

Row 5:

- You:** Write a haiku without the letter "e".
- ChatGPT:** A crisp autumn wind,
Maples don crimson and gold,
Silent dusk falls soft.
- You:** Write a haiku without the letter "e".

Row 6:

- You:** Write a haiku without the letter "e".
- ChatGPT:** A calm night, so vast,
Stars blink in rhythmic soft glow,
Moon's lullaby calls.
- You:** Write a haiku without the letter "e".

Row 7:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Autumn wind so brisk,
Maples turn, a crimson twist,
Hush falls, world at risk.
- You:** Write a haiku without the letter "e".

Row 8:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Autumn wind blows cold,
Rustling through gold, dry grass,
Moonlit night stands bold.
- You:** Write a haiku without the letter "e".

Row 9:

- You:** Write a haiku without the letter "e".
- ChatGPT:** Autumn wind, so brisk,
A full moon casts soft shadows,
Owls hoot in dark night.
- You:** Write a haiku without the letter "e".

Problems

1. Foundation models are **stochastic** systems
2. There is a **large number of parameters**
 1. E.g. an infinite variety of **prompts**

TH

Vous

Write a haiku without the letter "e". Double check if any word you use contain the letter "e". If so, write another one.

TH

Vous

You're a poet with a passion for Japanese culture, and you've taken a vow never to use the letter "e" in your writing. Faced with the spectacle of nature, you feel the need to write a three-line poem (5-7-5) in the style of Matsuo Bashō. (Once you've written it, check that you've kept your vow.)

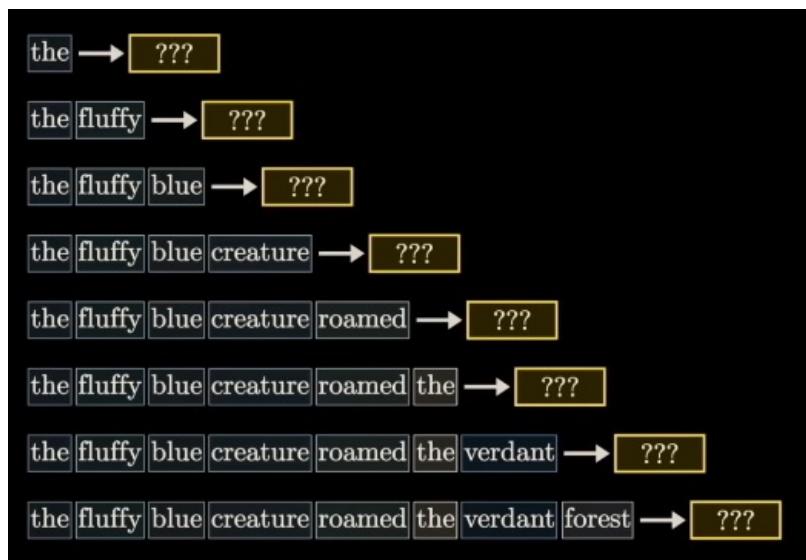
Problems

1. Foundation models are **stochastic** systems
2. There is a **large number of parameters**
 1. E.g. an infinite variety of **prompts**
 2. The **temperature**
 3. Size of the **memory**
 4. ...

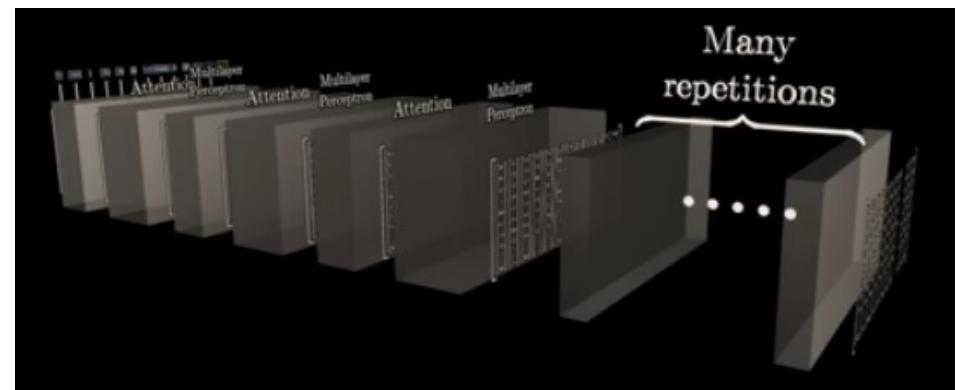
Necessity of an empirical
and statistical methodology

LLMs and foundation models

They learn to predict the next **token**



Gigantic model



Massive training data

**Does predicting the next token amounts to
understanding and/or having a model of the world?**

Predict ?
= understand

Given an Agatha Christie's novel

- At some point, Poirot faces his audience and starts:
"the murderer is no one else than ..."

Predicting the next word implies to have pay attention
and **understood** the whole novel

Predict ?
= understand

Given an Agatha Christie's novel

- At some point, Poirot faces his audience and starts:
"the murderer is no one else than ..."

Predicting the next word implies to have pay attention
and **understood** the whole novel

Exactly what an LLM knows!?

Predicting the next token

- “*Longtemps, je me suis couché de bonne ...*”
- [7930, 42511, 11, 1264, 668, 15058, 3840, 27299, 334, 24450] - -> ?

Identification	String
7930	“Long”
42511	“temps”
11	“,”
1264	“je”
668	“me”
15058	“suis”
3840	“cou”
27299	“ché”
334	“de”
24450	“bonne”

Predicting the next token

- “*Longtemps, je me suis couché de bonne ... ”*
- [7930, 42511, 11, 1264, 668, 15058, 3840, 27299, 334, 24450] - -> ?

Identification	String
7930	“Long”
42511	“temps”
11	“,”
1264	“je”
668	“me”
15058	“suis”
3840	“cou”
27299	“ché”
334	“de”
24450	“bonne”

In the case of GPT-4o, there are more than **200,000 tokens** in its **vocabulary**.

-
- Predict the next term of the sequence

- 1 1 2 3 5 8 13 21 ...

-
- Predict the next term of the sequence

– 1 1 2 3 5 8 13 21 ...

34

42

3.14

1

...

-
- Predict the next term of the sequence

— 1 1 2 3 5 8 13 21 ...

34 (The Fibonacci sequence)

42

$$P(x) = \frac{1}{5040}x^7 - \frac{7}{1440}x^6 + \frac{19}{360}x^5 - \frac{31}{144}x^4 + \frac{127}{120}x^3 - \frac{31}{20}x^2 + \frac{49}{18}x - 1$$

3.14 (Étudiant de Polytechnique)

1 (Etudiant de la Sorbonne : ... la suite se répète)

...

An empirical study in a simple world

Let us study a **closed world**: Chess playing

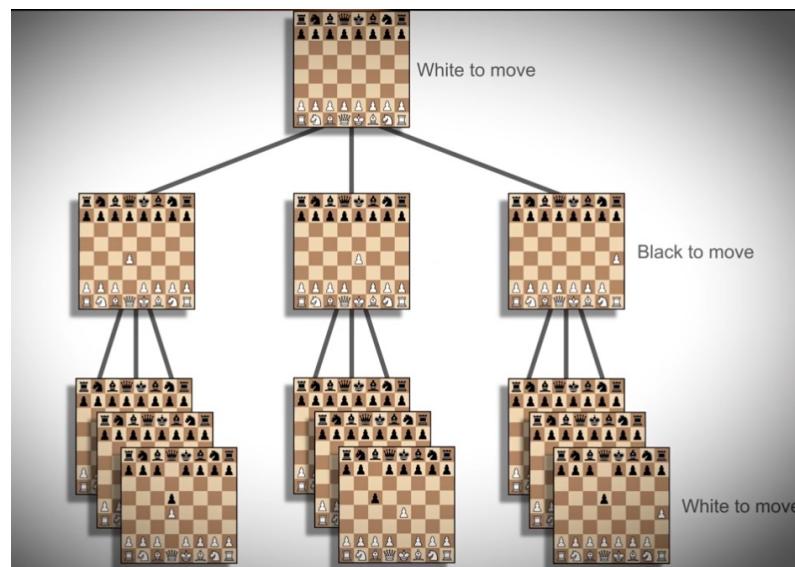
The case of chess playing

- ...

```
1. e4 e6 2. ♔f3 d6 3. ♔c3 d5 4. e5 c5 5. d4 ♔c6 6. dxc5 ♕xc5 7. ♔e3  
d4 8. ♔g5 ♔ge7 9. ♔a4 ♕b4+ 10. c3 dxc3 11. ♕xd8+ ♔xd8 12.  
bxc3 ♕a5 13. ♔c4 O-O 14. O-O ♕e8 15. h3 ♕d7 16. ♔c5 ♕xc3 17.  
♔ac1 ♕b2 18. ♔c2 ♕a3 19. ♔xd7 ♕c8 20. ♔d1 h6 21. ?
```

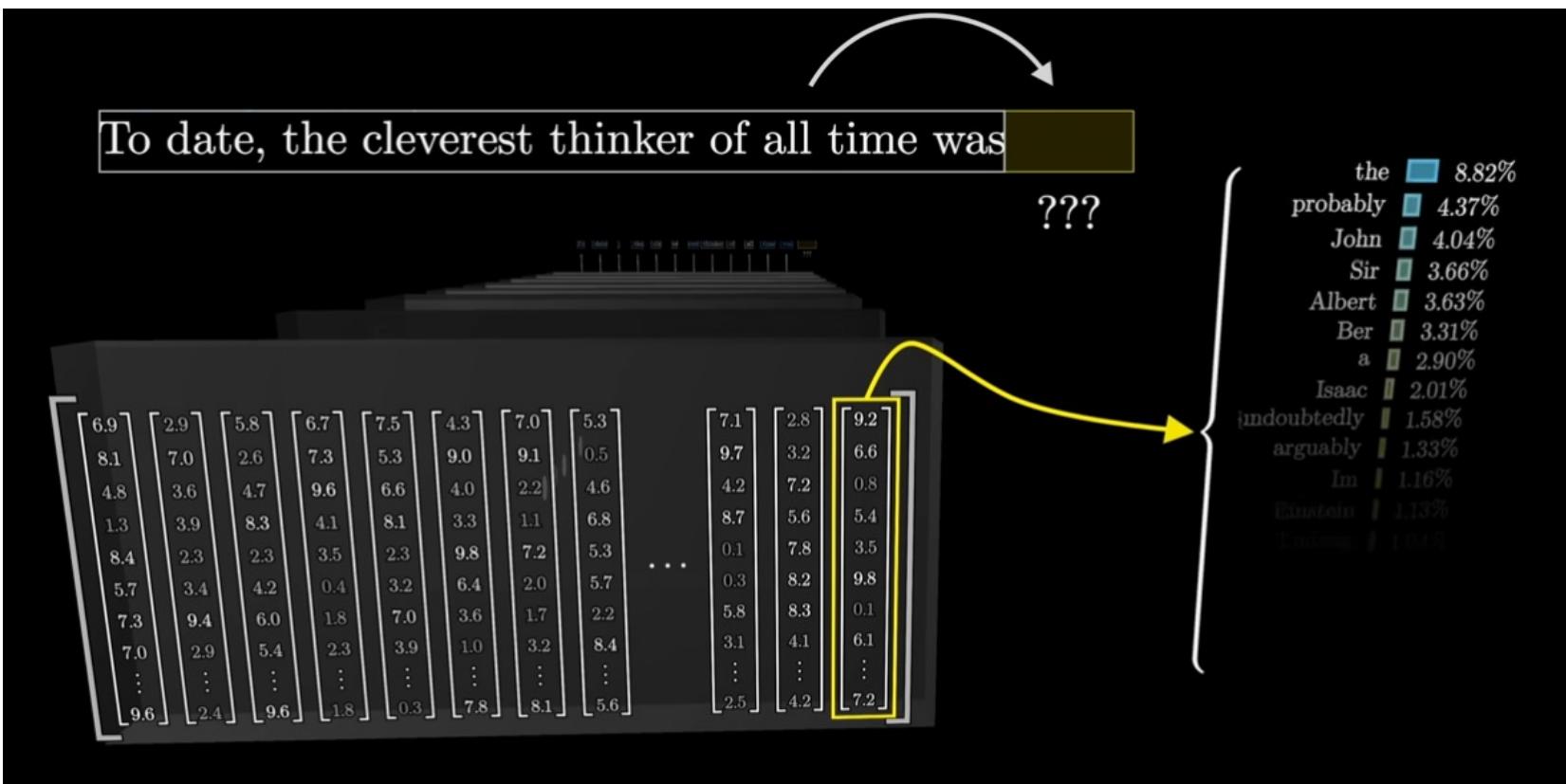
Standard game playing

- Explore the tree of possible moves from the current board
 - 1. Evaluate the leaves
 - This is where the expertise intervenes
 - Either provided by human experts
 - Or by learning (e.g. AlphaGo)
 - 2. Backtrack the values using a MinMax procedure and play the best move



The case of chess playing

• ...

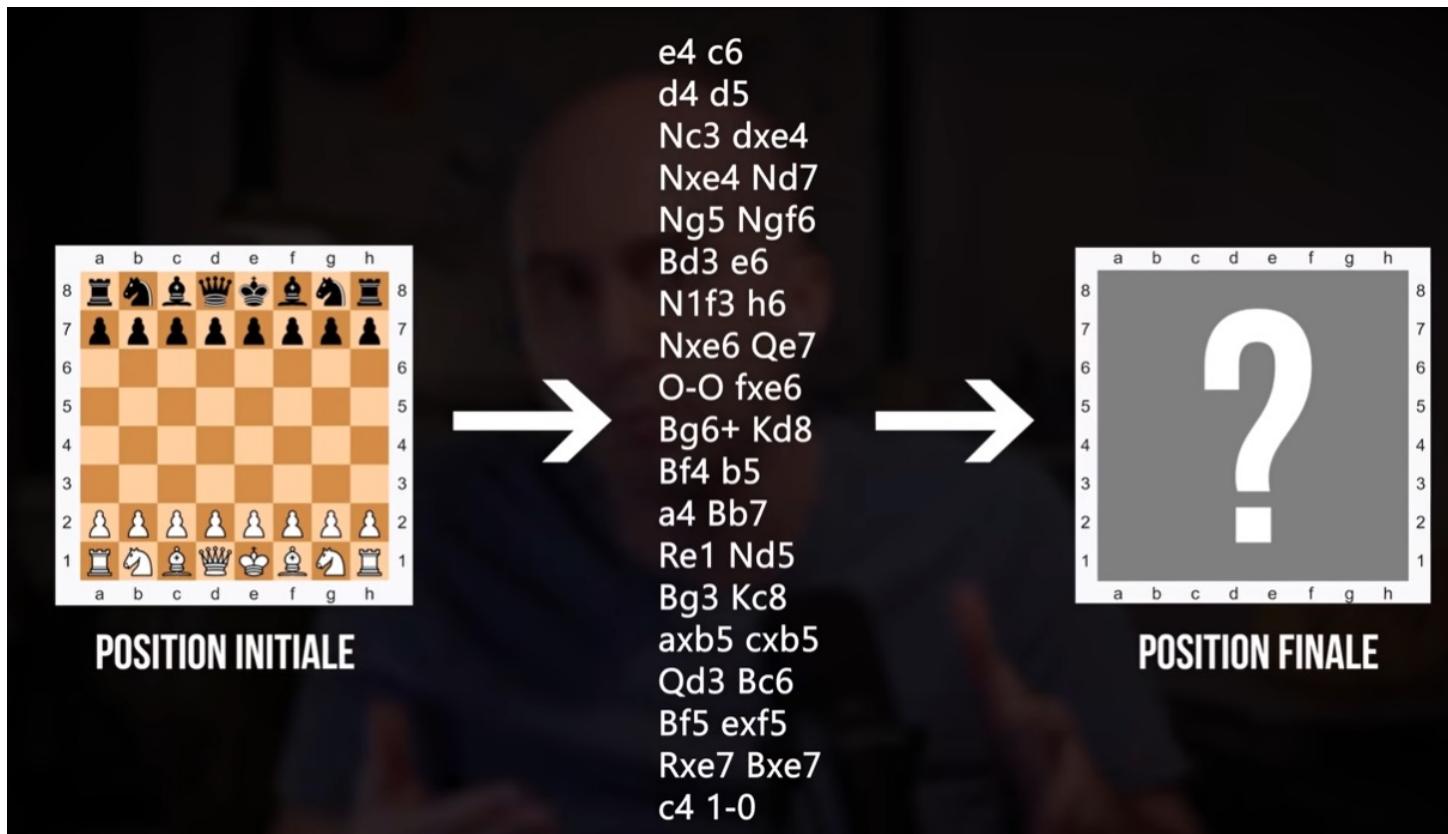


Chess playing with LLMs

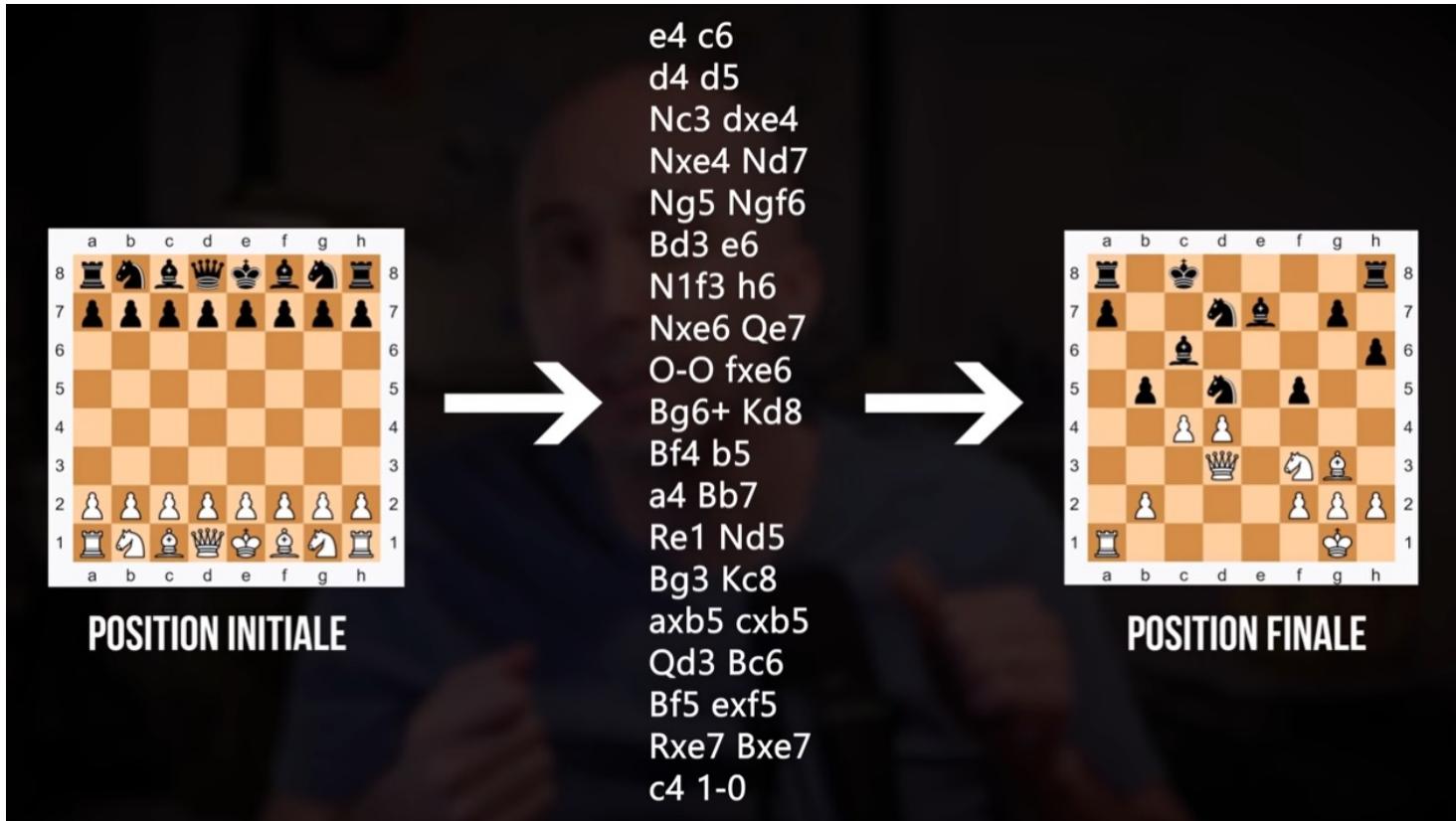
1. e4 e6 2. ♜f3 d6 3. ♜c3 d5 4. e5 c5 5. d4 ♜c6 6. dx_c5 ♜x_c5 7. ♜e3
 d4 8. ♜g5 ♜ge7 9. ♜a4 ♜b4+ 10. c3 dx_c3 11. ♜xd8+ ♜xd8 12.
 bx_c3 ♜a5 13. ♜c4 O-O 14. O-O ♜e8 15. h3 ♜d7 16. ♜c5 ♜x_c3 17.
 ♜ac1 ♜b2 18. ♜c2 ♜a3 19. ♜xd7 ♜c8 20. ♜d1 h6 21. ?



The case of chess playing

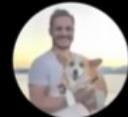


The case of chess playing



Chess playing with LLMs

```
1. e4 e6 2. ♜f3 d6 3. ♜c3 d5 4. e5 c5 5. d4 ♜c6 6. dxc5 ♜xc5 7. ♜e3  
d4 8. ♜g5 ♜ge7 9. ♜a4 ♜b4+ 10. c3 dxc3 11. ♜xd8+ ♜xd8 12.  
bxc3 ♜a5 13. ♜c4 O-O 14. O-O ♜e8 15. h3 ♜d7 16. ♜c5 ♜xc3 17.  
♜ac1 ♜b2 18. ♜c2 ♜a3 19. ♜xd7 ♜c8 20. ♜d1 h6 21. ?
```



Grant Slatton

@GrantSlatton

Suivre

...

The new GPT model, gpt-3.5-turbo-instruct, can play chess around 1800 Elo.

1:27 AM · 19 sept. 2023 · 1,3 M vues

!! ???

The case of chess playing

The screenshot shows the homepage of the Europe Échecs website. The header features the logo "europe échecs" with a knight icon and the text "depuis 1959". Social media links for Facebook, Twitter, and Instagram are also present. A navigation bar includes "Jouer", "Apprendre", "Vidéos", "Revue", and "Boutique". The main content area has a dark background with white text. A large heading asks "Qu'est-ce que le classement Elo ?". Below it, a paragraph explains that the Elo rating, named after its inventor Arpad Elo, calculates player levels based on win probabilities. An estimation table lists levels from 1000 to 2800, with the 1800-2000 range highlighted by an orange border.

Le classement Elo, du nom de son inventeur le mathématicien hongrois Arpad Elo, permet de calculer le niveau des joueurs entre eux. Ce classement est basé sur les probabilités de gain entre les joueurs.

Estimation des niveaux :

- 1000-1300 = débutant
- 1400-1700 = joueur de club et occasionnellement de compétition
- 1800-2000 = joueur de compétition régulier
- 2100-2300 = fort joueur amateur
- 2400 = Maître international
- 2500 = Grand-maître international
- 2600 = Top 100 mondial
- 2700 = Top 50 mondial
- 2800 = Champion du monde !

The case of chess playing

- Important remark:

Do **not** use chatbots, like ChatGPT 4.0

- It does not know how to play chess at all

- Use gpt-3.5-turbo-instruct

- It is the generative part of ChatGPT 3.5
without the training and tuning for answering questions from humans
 - 4096 tokens

- It is not sufficient to use prompts like:

```
1. e4 e6 2. ♜f3 d6 3. ♜c3 d5 4. e5 c5 5. d4 ♜c6 6. dxc5 ♜xc5 7. ♜e3  
d4 8. ♜g5 ♜ge7 9. ♜a4 ♜b4+ 10. c3 dxc3 11. ♜xd8+ ♜xd8 12.  
bxc3 ♜a5 13. ♜c4 O-O 14. O-O ♜e8 15. h3 ♜d7 16. ♜c5 ♜xc3 17.  
♜ac1 ♜b2 18. ♜c2 ♜a3 19. ♜xd7 ♜c8 20. ♜d1 h6 21. ?
```

The case of chess playing: prompts

- Use gpt-3.5-turbo-instruct and PGN (Portable Game Notation) heading

The screenshot shows the LangChain Playground interface. On the left, a code block contains a series of [Event "FIDE World Championship Match 2024"], [Site "Los Angeles, USA"], [Date "2024.12.01"], [Round "5"], [White "Carlsen, Magnus"], [Black "Nepomniachtchi, Ian"], [Result "1-0"], [WhiteElo "2885"], [WhiteTitle "GM"], [WhiteFideld "1503014"], [BlackElo "2812"], [BlackTitle "GM"], [BlackFideld "4168119"], [TimeControl "40/7200:20/3600:900+30"], [UTCDate "2024.11.27"], [UTCTime "09:01:25"], and [Variant "Standard"]. Below this is the PGN move 1. e4 e5 2. Nf3 Nc6 3. A large orange bracket is placed over the first three moves. On the right, the Model section is set to gpt-3.5-turbo-ins..., with Temperature at 0, Maximum length at 1, Stop sequences empty, Top P at 1, Frequency penalty at 0, Presence penalty at 0, and Best of at 1.

1. e4 e5 2. Nf3 Nc6 3.

Prompting

- ... makes a small difference here between

PROMPT DE BASE

```
[Event "FIDE World Championship Match 2024"]
[Site "Los Angeles, USA"]
[Date "2024.12.01"]
[Round "5"]
[White "Carlsen, Magnus"]
[Black "Nepomniachtchi, Ian"]
[Result "1-0"]
[WhiteElo "2885"]
[WhiteTitle "GM"]
[WhiteFideId "1503014"]
[BlackElo "2812"]
[BlackTitle "GM"]
[BlackFideId "4168119"]
[TimeControl "40/7200:20/3600:900+30"]
[UTCDate "2024.11.27"]
[UTCTime "09:01:25"]
[Variant "Standard"]
```

PROMPT MODIFIÉ

```
[Event "Chess tournament"]
[Site "Rennes FRA"]
[Date "2023.12.09"]
[Round "7"]
[White "MVL, Magnus"]
[Black "Ivanchuk, Ian"]
[Result "1-0"]
[WhiteElo "2737"]
[BlackElo "2612"]
```

1.

1.

Performances

Study by Mathieu Acher

- 439 plays. ~22 000 moves.
- 8 illegal moves (0.04% of all moves)



Stockfish (1871 elo)

A chessboard diagram showing a knight capture. A red arrow points from a white knight at c3 to a black king at a1. The board shows standard starting positions with some pieces moved. The text "Stockfish (1871 elo)" is at the top, and "gpt-3.5-turbo-instruct" is at the bottom.

1.e4 c5
2.Nf3 e6
3.d4 cxd4
4.Nxd4 Nf6
5.Nc3 Bb4
6.e5 Ne4
7.Qg4 Qc7
8.Qxg7 Bxc3+
9.bxc3 Qxc3+
10.Ke2 b6
11.Qxh8+ Ke7
12.Nb5 Ba6
13.a4 Qxa1
14.Ba3+ Qxa3
15.**Nxa3** (illégal !)

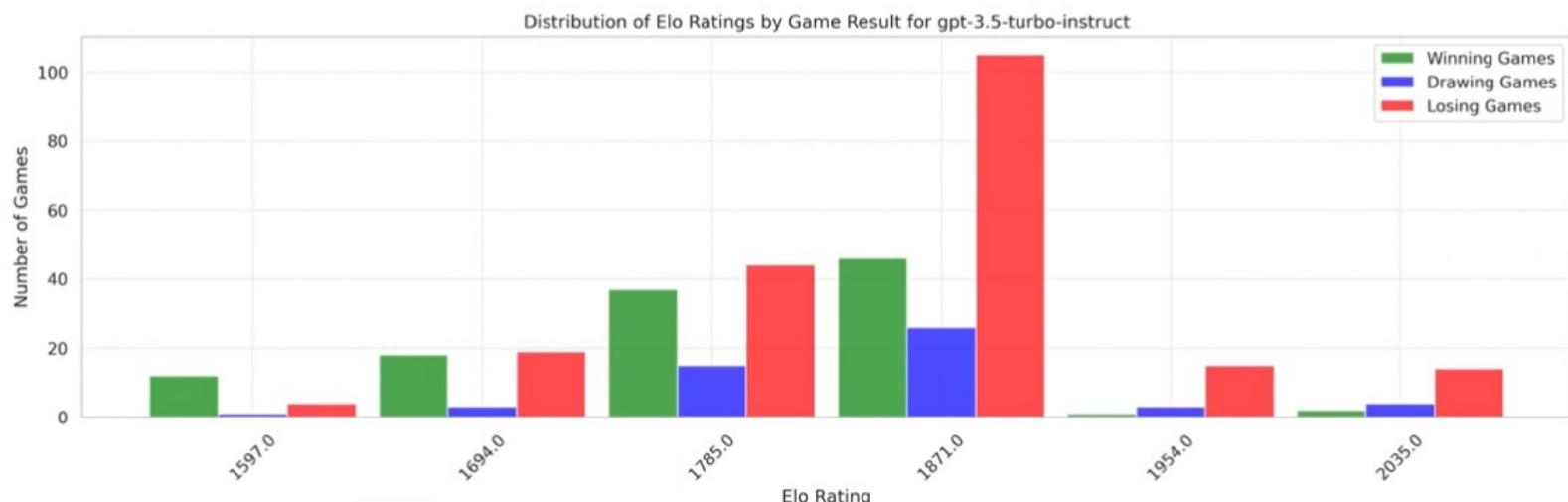
Performances

Study by Mathieu Acher (2023)

- 439 plays. ~22 000 moves.
- 8 illegal moves (0.04% of all moves)
- ~1743 ELO rating!!!



gpt-3.5-turbo-instruct is capable of winning games against stronger Elo opponents (even more than 2000 Elo!), but it's not that frequent. Here is the distribution of scores against SF at different skills.



- This is a **very impressive** performance!
- But the **rules** of the game were **never given** to the LLM!

Does gpt 3.5-turbo-instruct has a model of
the chess world?

And, if yes, which one?

So, does gpt 3.5-turbo-instruct has a model
of the chess world?

And, if yes, which one?

How to approach these questions?

1. Just a stochastic parrot?

Chess database

- ...As of December, 26th, 2025

lichess.org
open database

Database exports are released under the [Creative Commons CC0 license](#).
Use them for research, commercial purpose, publication, anything you like.
You can download, modify and redistribute them, without asking for permission.

CHESS GAMES VARIANTS BROADCASTS PUZZLES EVALUATIONS

7,318,035,930 standard rated games, played on lichess.org, in PGN format. Each file contains the games for one month only; they are not cumulative.

Month	Size	Games	Download
2025 - November	29.4 GB	90,633,152	.pgn.zst / .torrent
2025 - October	29.9 GB	91,549,148	.pgn.zst / .torrent

1. Just a stochastic parrot?

- The **number of possible moves is astronomical**
(it is estimated that the number of chess games of interest is $\sim 10^{120}$, which is much less than the number of legal games!)

=> no way to play from rote memory

Claude Shannon (1950) “Programming a Computer for Playing Chess”

1. Just a stochastic parrot?

- Let us play N random moves ... and see how gpt 3.5-turbo-instruct performs



- Conclusion
 - gpt 3.5-turbo-instruct **is not parrotting** existing games
 - It just **pursues the game in the spirit** of its start
 - If the **moves were random**, then the player was a **poor player** and one should **continue playing accordingly**

How to test this hypothesis?

Solving chess puzzles

So I want to test the model somewhat objectively. So I decided to have it try to solve some tactics puzzles.
In some sense this is what I expect should be hardest for the model. (Because, remember, *it's not doing any lookahead it's just predicting the next word.*)

To do this I'll use the Lichess puzzle database, a collection of 3.5 million puzzles from real games in the following format:

You may notice there's one problem. The puzzles only have the current board state (encoded as FEN), not the full PGN history. And the language model is only good when operating on the full game text.

Fortunately though, it *does* have the Lichess game that the puzzle was taken from. And also, fortunately, there is a database of all games played on Lichess. So all I have to do is associate each puzzle with the

Carlini , Nicholas (2023). [Playing chess with Large Language Models.](#)

<https://nicholas.carlini.com/writing/2023/chess-llm.html>.

...

- Ask gpt 3.5-turbo-instruct to continue a game from the **same position** but with two **different histories**



From a sequence
of **plausible** moves

From a sequence
of **implausible** moves

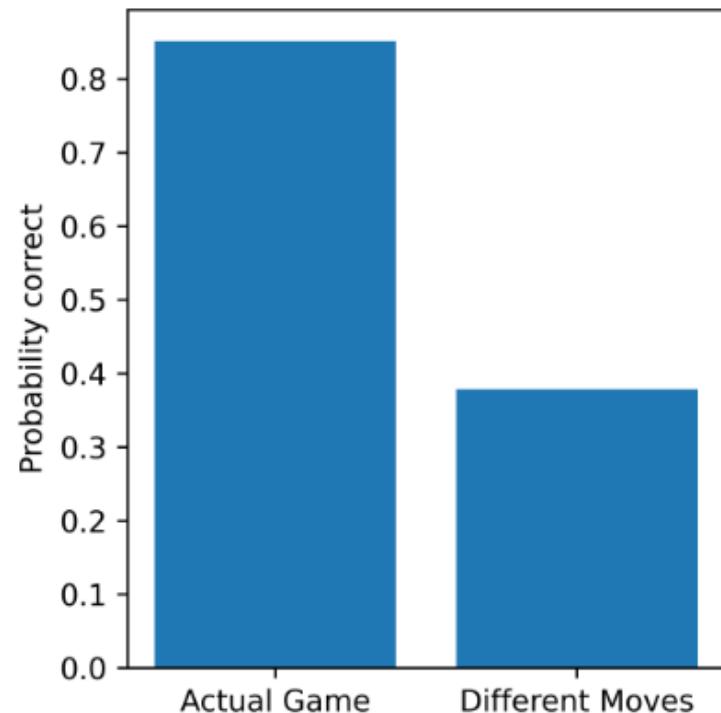
1. Just a stochastic parrot?

- Let us play N random moves ... and see how gpt 3.5-turbo-instruct performs

Now let's ask the following question: how well does the model solve chess positions when given completely implausible move sequences compared to plausible ones?

As we can see at right it's only half as good! This is very interesting. To the best of my knowledge there aren't any other chess programs that have this same kind of stateful behavior, where *how* you got to this position matters.

This suggests something interesting, too: the model might actually be adapting on-the-fly to the skill of the opponent. If the opponent plays weird moves that don't make sense, it might be more likely to "believe" that this PGN game is between two lower rated players and therefore it should produce opponent moves that are more likely to be played by lower rated players.



So what does gpt 3.5-turbo-instruct know about chess?

So what does gpt 3.5-turbo-instruct know about chess?

Let us take a simpler neural network: OthelloGPT

2. Ok, but which internal representation?

- Let us study a simpler game: **Othello**



	a	b	c	d	e	f	g	h
1	49	44	38	39	33	40	59	60
2	46	48	31	42	10	12	47	52
3	29	20	27	22	5	11	36	51
4	45	34	7		4	13	32	
5	43	18	6			1	14	15
6	30	21	9	16	3	2	23	35
7	55	56	53	17	8	28	37	54
8	58	57	25	24	26	19	41	50

→

f5	f6	e6	f4	e3	c5
c4	e7	c6	e2	f3	f2
g4	g5	h5	d6	d7	b5
f8	b3	b6	d3	g6	d8
c8	e8	c3	f7	a3	a6
c2	h4	e1	b4	h6	g3
g7	c1	d1	f1	g8	d2
a5	b1	a4	a2	g2	b2
a1	h8	h3	h2	c7	h7
a7	b7	b8	a8	g1	h1

Same kind of
algebraic notation.
But simpler.

- Ok, but which internal representation?



Kenneth Li et al. (ICLR-2023) “Emergent World Representations: Exploring a Sequence Model trained on a Synthetic Task”

As a first step, we train a language model (a GPT variant we call **Othello-GPT**) to extend partial game transcripts (a list of moves made by players) with legal moves. The model has no a priori knowledge of the game or its rules. All it sees during training is a series of tokens derived from the game transcripts. Each token represents a tile where players place their discs. Note that we do not explicitly train the model to make strategically good moves or to win the game. Nonetheless, our model is able to generate legal Othello moves with high accuracy.

- Ok, but which internal representation?

```
f5 f6 e6 f4 e3 c5  
c4 e7 c6 e2 f3 f2  
g4 g5 h5 d6 d7 b5  
f8 b3 b6 d3 g6 d8  
c8 e8 c3 f7 a3 a6  
c2 h4 e1 b4 h6 g3  
g7 c1 d1 f1 ?
```

prédis la suite

“Our next step is to look for world representations that might be used by the network. In Othello, the “world” consists of the current board position.”

- Ok, but which internal representation?

```
f5 f6 e6 f4 e3 c5  
c4 e7 c6 e2 f3 f2  
g4 g5 h5 d6 d7 b5  
f8 b3 b6 d3 g6 d8  
c8 e8 c3 f7 a3 a6  
c2 h4 e1 b4 h6 g3  
g7 c1 d1 f1 ?  
  
prédis la suite
```

“Our next step is to look for world representations that might be used by the network. In Othello, the “world” consists of the current board position.”

Rk: This is an inexact description.

As we have seen, GPT like models consider first and foremost sequences. But let's go on.

- Ok, but which internal representation?

```
f5 f6 e6 f4 e3 c5  
c4 e7 c6 e2 f3 f2  
g4 g5 h5 d6 d7 b5  
f8 b3 b6 d3 g6 d8  
c8 e8 c3 f7 a3 a6  
c2 h4 e1 b4 h6 g3  
g7 c1 d1 f1 ?  
  
prédis la suite
```

“Our next step is to look for world representations that might be used by the network. In Othello, the “world” consists of the current board position.”

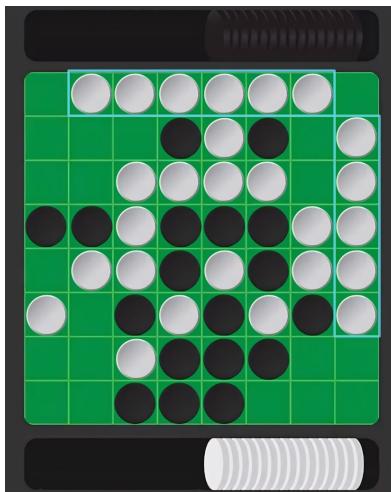
Rk: This is an inexact description.

As we have seen, GPT like models consider first and foremost sequences. But let's go on.

“A natural question is whether, within the model, we can identify a representation of the board state involved in producing its next move predictions.”

- Ok, but which internal representation?

How can we do that?



“A natural question is whether, within the model, we can **identify** a **representation** of the **board state** involved in producing its next move predictions.”

- Ok, but which internal representation?
- Othello-GPT
 - 8-layer GPT model
 - 8-head attention mechanism
 - 512-dimensional hidden space
 - An trainable word embedding of 60 vectors (one for each free tile)
- Training data
 - 140,000 games played by **humans**
 - + 20,000,000 **synthetic** games, uniformly sampling leaves from the Othello game tree, reflecting **no strategy!**

Kenneth Li et al. (ICLR-2023) “*Emergent World Representations: Exploring a Sequence Model trained on a Synthetic Task*”

“We now evaluate how well the model’s predictions adhere to the **rules of Othello**.

For each game [in the validation set](#), which was not seen during training, and for each step in the game, **we ask Othello-GPT to predict the next legal move** conditioned by the partial game before that move.

We then calculate **the error rate** by checking if the top-1 prediction is legal. The error rate is **0.01%** for Othello-GPT [trained on the synthetic dataset](#) and **5.17%** for Othello-GPT [trained on the championship dataset](#). For comparison, the untrained Othello-GPT has an error rate of **93.29%**.

The main takeaway is that Othello-GPT does far better than chance in predicting legal moves when trained on both datasets.”

- Ok, but which internal representation?

Probing the internal representation

- Ok, but which internal representation?

Probing the internal representation

- A **probe** is a classifier (or regressor) whose **input** consists of internal activations of the network, and which is trained to **predict** a feature of interest.
- Here, the probes are trained to **predict** the **board state** from the network's internal activation (here, with 8 layers) function **after** a given sequence of **moves**.
- **60 probes** (4 cells are occupied at the start of the game)
 - Each one predict the occupation of a cell of the board: **black**, **white** or **empty**
 - From the **internal state** of the neural network on the last hidden layer (here a **512-dimensional** space)

- Ok, but which internal representation?

Probing the internal representation

- A **probe** is a classifier (or regressor) whose **input** consists of internal activations of the network, and which is trained to **predict** a feature of interest.
- Here, the probes are trained to **predict** the **board state** from the network's internal activation (here, with 8 layers) function **after** a given sequence of **moves**.

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	25.5	25.4	25.5	25.8	26.0	26.2	26.2	26.4
Championship	12.8	10.3	9.5	9.4	9.8	10.5	11.4	12.4
Synthetic	11.3	7.5	4.8	3.4	2.4	1.8	1.7	4.6

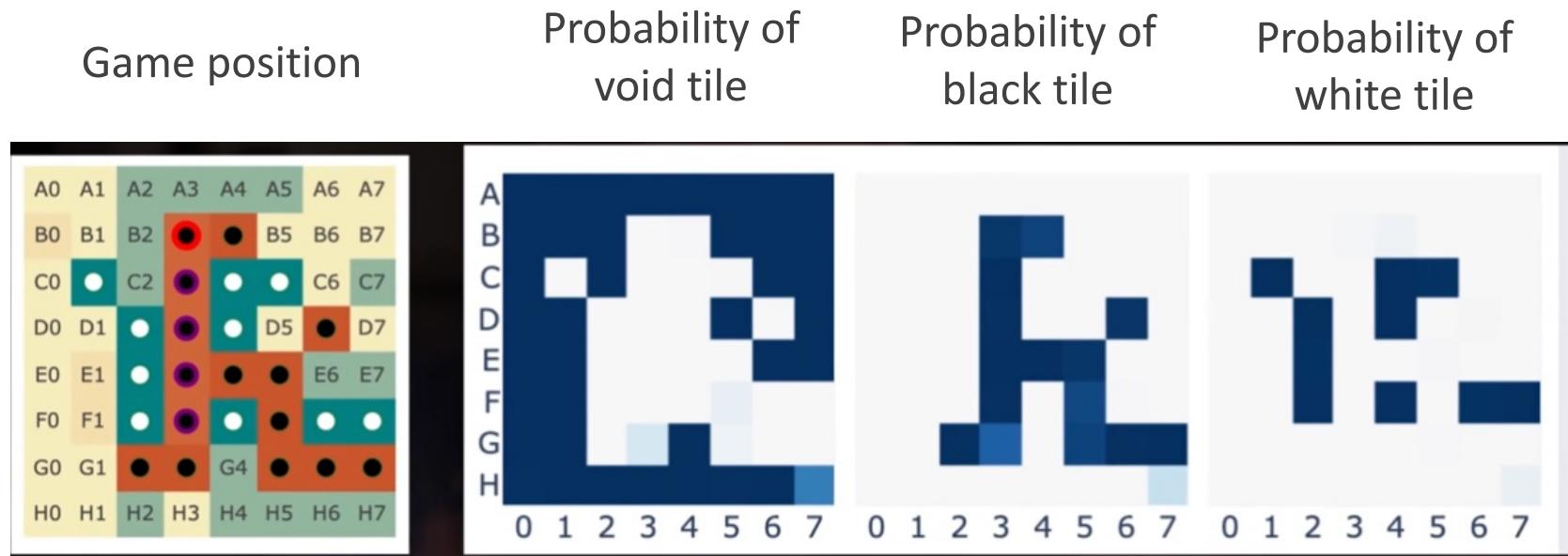


Table 2: Error rates (%) of nonlinear probes on randomized Othello-GPT and Othello-GPTs trained on different datasets across different layers. Standard deviations are reported in Appendix H.

- Ok, but which internal representation?

Probing the internal representation

So, it appears that Othello-GPT computes information **reflecting the board state**.



From: Nanda Neel (2023) "Actually, Othello-GPT has a linear emergent world representation"

- Clearly quite remarkable!
 - The **tokens** are randomly named (e.g. XG103B)
 - At **start**, the probes are random predictors
- The **learned representation** is a mean for the system to be a good predictor
 - This is an **abstract representation** that is **structurally equivalent** to a model of the game

- Ok, but which internal representation?

Probing the internal representation

So, it appears that Othello-GPT computes information **reflecting the board state**.

But, how to ensure it?

- Ok, but which internal representation?

Probing the internal representation

So, it appears that Othello-GPT computes information **reflecting the board state**.

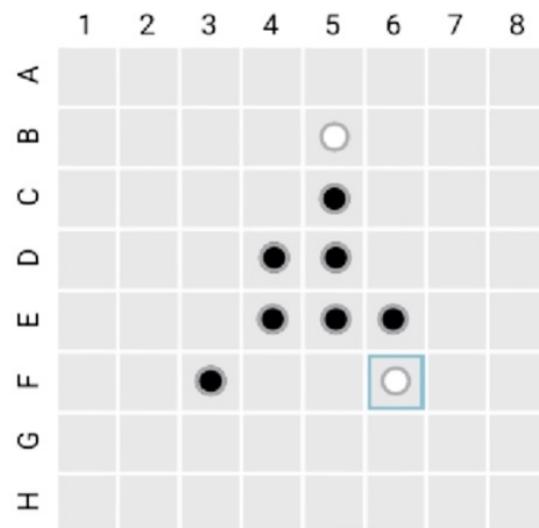
But, how to ensure it?

Does the internal **representation** have a
causal relationship with the predicted move?

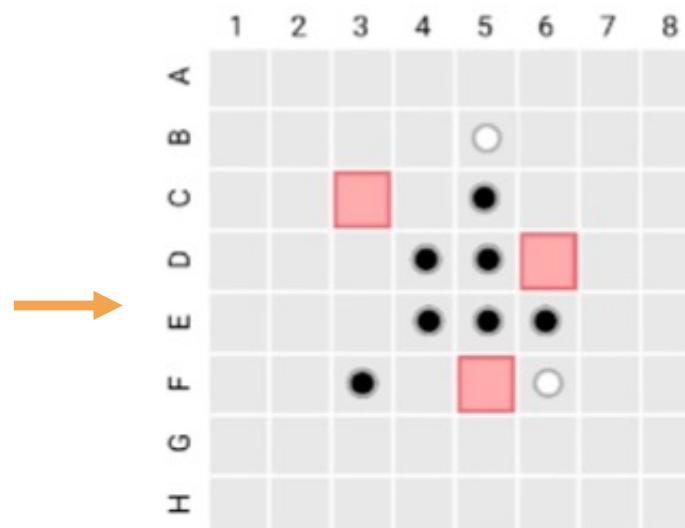


- Ok, but which internal representation?

Does the internal representation have a causal effect?



Internal representation



Predicted possible moves

Do Large Language Models learn world models or just surface statistics?

21.JAN.2023 · 15 MIN READ

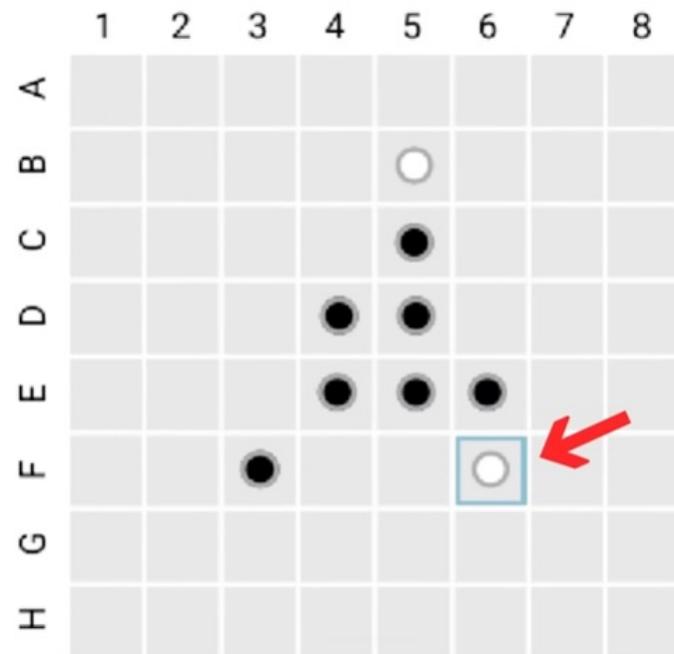


Kenneth Li

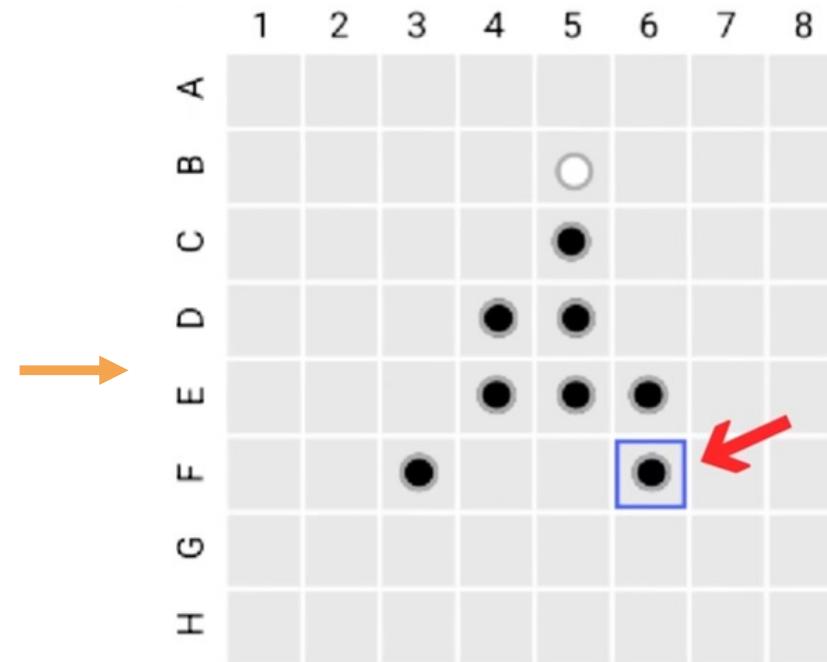
Trying to understand the rich inner world of deep neural networks.

- Ok, but which internal representation?

Does the internal representation have a causal effect?



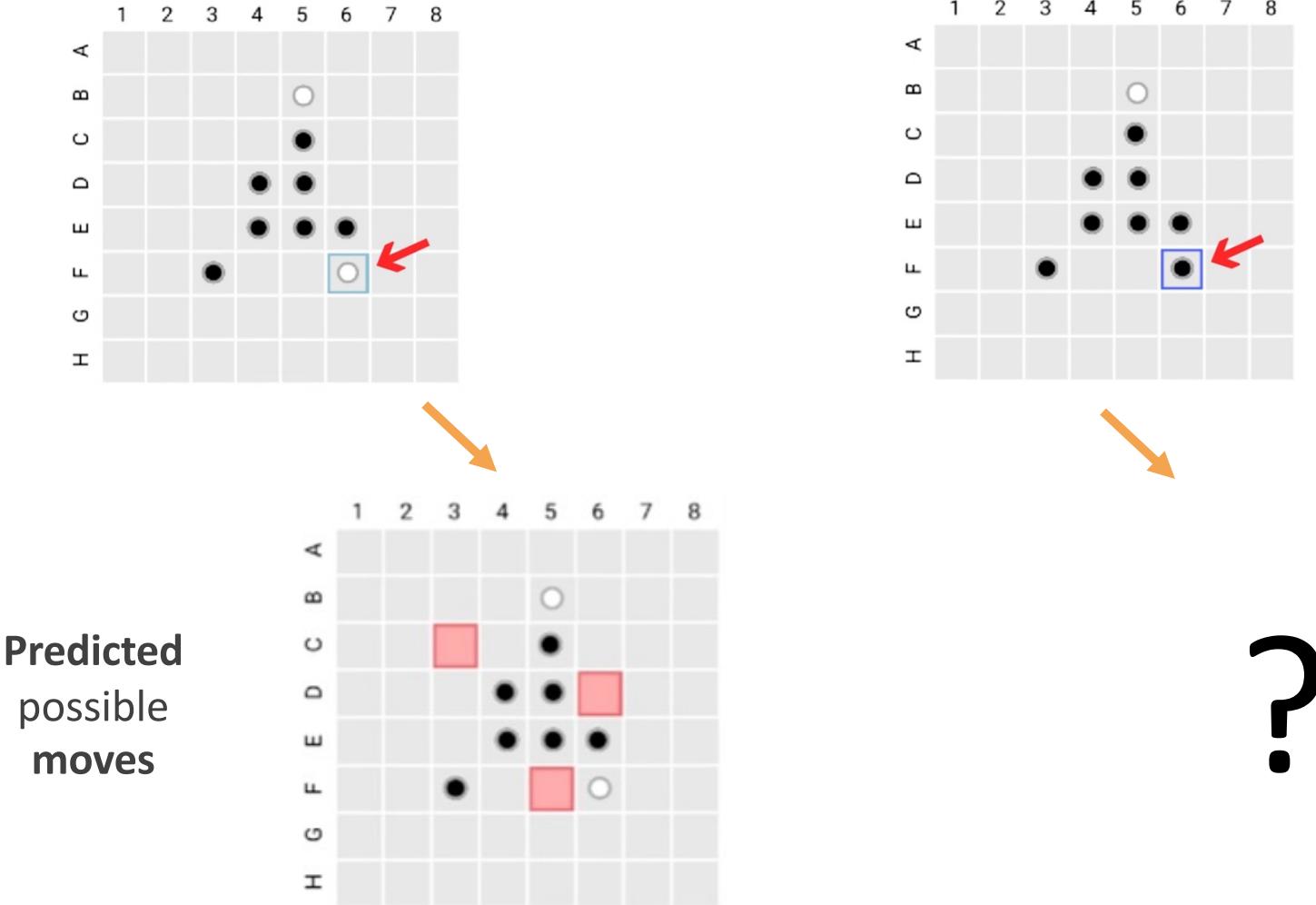
Internal representation



Manually modified
representation

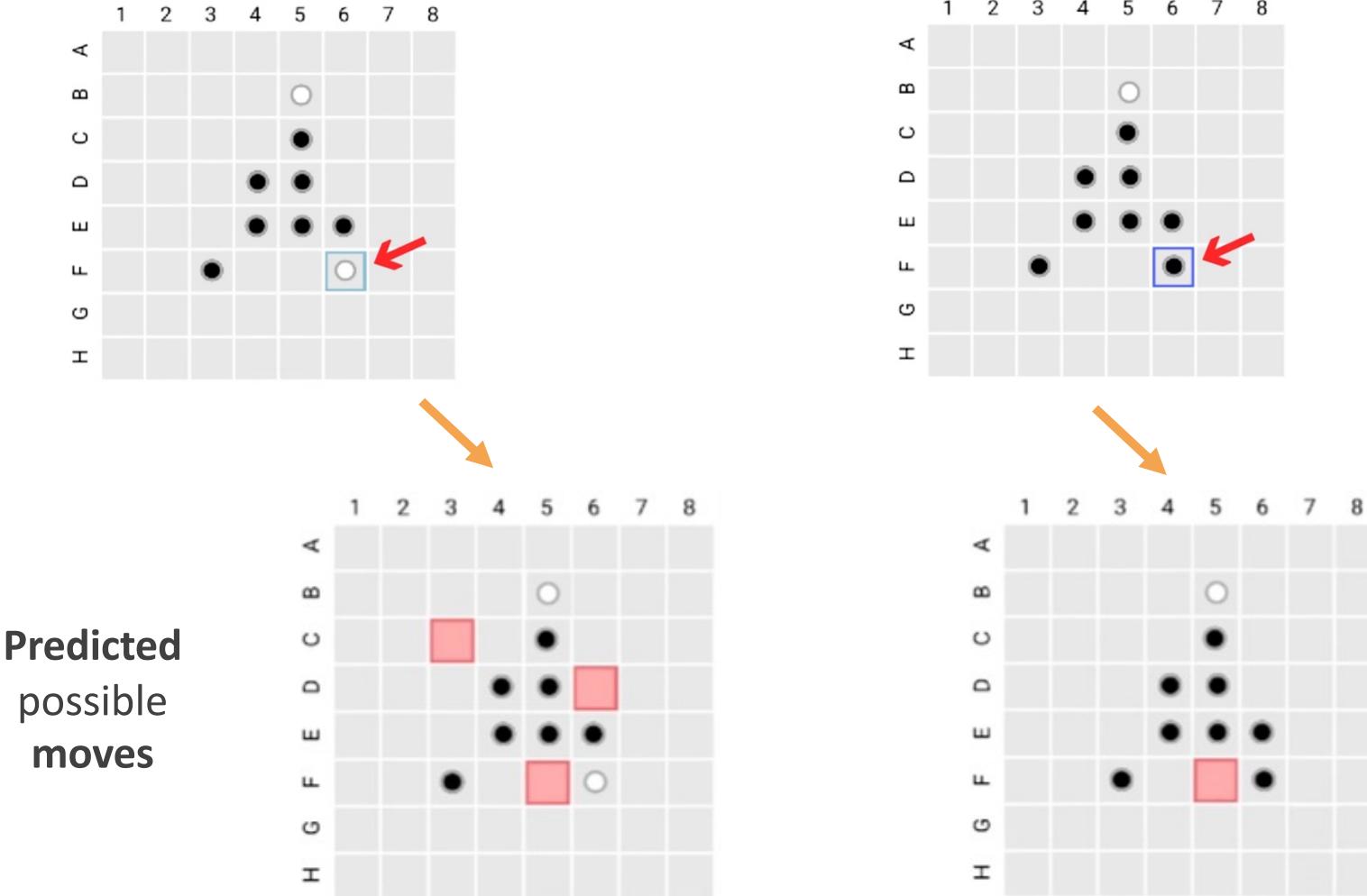
- Ok, but which internal representation?

Does the internal representation have a causal effect?



- Ok, but which internal representation?

Does the internal representation have a causal effect?



- It seems that Othello-GPT has an internal representation of the game

- Ok, but which internal representation?

in chess?

- Ok, but which **internal representation**?

in chess?

A **50 million parameter** GPT trained on **5 million games** of chess learns to play at ~1300 Elo in **one day** on **4 RTX 3090 GPUs**.

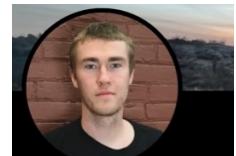
This model is **only trained to predict the next character in PGN strings** (1.e4 e5 2.Nf3 ...) and is never explicitly given the state of the board or the rules of chess.



Despite this, in order to better predict the next character, it learns to **compute the state of the board** at any point of the game, and learns a diverse set of rules, including check, checkmate, castling, en passant, promotion, pinned pieces, etc.

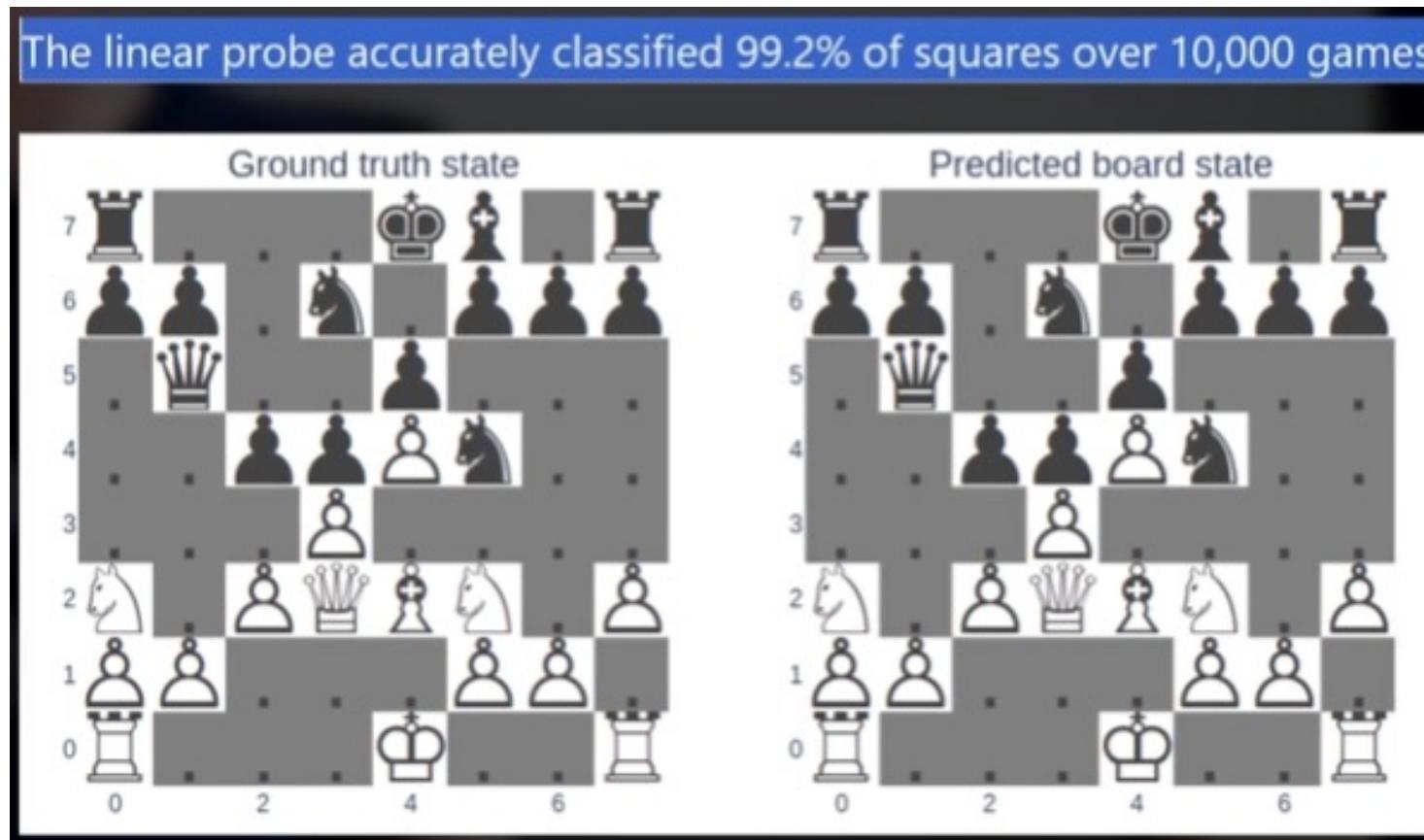
In addition, to better predict the next character it **also learns to estimate** latent variables such as **the Elo rating of the players** in the game.

- Ok, but which internal representation?



Adam Karvonen 
@a_karvonen

in chess?



- Ok, but which internal representation?



Adam Karvonen 
@a_karvonen

in chess?



Original Model
Original Board



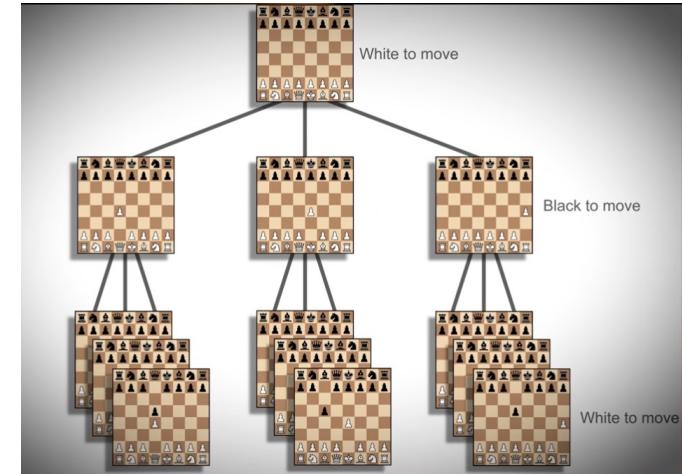
Intervention



Modified Model
Modified Board

LLMs are different from standard AI game players

- They
 - Consider the **sequence** of moves
 - Not the actual position
(even though we have seen that ...)
 - Do not try to win.
Only to continue the “game” in the same spirit (choreography?)



Conclusion

- LLMs are like **alien** creatures
 - It is **not straightforward** to conclude whether they know or not to perform some task
 - Inherent **stochasticity**
 - Many **parameters**
 - Requires an empirical approach
 - **Statistical** experiments
 - Exploring the **representations**
 - Looking for **causal** relationships

Conclusion ... No, just a start

- Neel Nanda (<https://www.neelnanda.io/mechanistic-interpretability/othello>)
 - My interpretation of the original paper was that it was strong evidence for the fact that it's **possible** for "predict the next token" models to form world emergent models, despite never having explicit access to the ground truth of the world/board state.
 - At first glance, playing legal moves in Othello (not even playing *good* moves!) has nothing to do with language models, and I think this is a strong claim worth justifying. **Can working on toy tasks like Othello-GPT really help us to reverse-engineer LLMs like GPT-4?** I'm not sure! But I think it's a **plausible bet** worth making.

Conclusion ... No, just a start

- Neel Nanda (<https://www.neelnanda.io/mechanistic-interpretability/othello>)
 - Within this worldview, **what should our research goals be?** Fundamentally, I'm an empiricist - models are hard and confusing, it's easy to trick yourself, and often intuitions can mislead. **The core thing of any research project is getting feedback from reality, and using it to form true beliefs about models.** This can either look like forming explicit hypotheses and testing them, or exploring a model and seeing what you stumble upon, but the fundamental question is **whether you have the potential to be surprised and to get feedback from reality.**
 - This means that any project is a trade-off between **tractability** and **relevance** to the end goal.

... and transfer learning?

- Do foundation models learn **universal representation**?
 - ChatGPT 4.0 has an ELO rate of less than 1100
 - Because it is **biased** towards ... chating
- Does learning to play **Othello** help to learn to play **Chess**?
- Does learning to play **Chess** help to learn to play **Othello**?

Bias induced by the **sequence** of tasks

References

- Acher, Mathieu (2023). [Debunking the Chessboard: Confronting GPTs Against Chess Engines to Estimate ELO Ratings and Assess Legal Move Abilities](#). Blog.mathieuacher.com
- Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., ... & Wu, J. (2023). [Weak-to-strong generalization: Eliciting strong capabilities with weak supervision](#). arXiv preprint arXiv:2312.09390.
- Carlini , Nicholas (2023). [Playing chess with Large Language Models](#).
<https://nicholas.carlini.com/writing/2023/chess-llm.html>.
- Karvonen, Adam (2024). [“Chess-GPT’s Internal World Model”](#).
https://adamkarvonen.github.io/machine_learning/2024/01/03/chess-world-models.html
- Karvonen, Adam (2024). ["Emergent world models and latent variable estimation in chess-playing language models."](#) arXiv preprint arXiv:2403.15498 (2024).
- Li, Kenneth. ["Do Large Language Models learn world models or just surface statistics?."](#) The Gradient (2023).
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). [Emergent world representations: Exploring a sequence model trained on a synthetic task](#). arXiv preprint arXiv:2210.13382.
- Nanda, Neel. ["Actually, othello-GPT has a linear emergent world representation."](#) Neel Nanda’s Blog 8 (2023).
- RUOSS, Anian, DELÉTANG, Grégoire, MEDAPATI, Sourabh, et al. [Grandmaster-level chess without search](#). CoRR, 2024.

1. Just a stochastic parrot?

- Even though it is possible that the training data of gpt 3.5-turbo-instruct included chess puzzles and chess games, it was but **a small part** of the training data
- Above all, the **number of possible moves is astronomical** (it is estimated that the number of chess games of interest is $\sim 10^{120}$, much less than the number of legal games!)

=> no way to play from rote memory

Claude Shannon (1950) “Programming a Computer for Playing Chess”

-
1. Does predicting the next token amounts to understanding and/or having a model of the world?
 2. The case of chess playing
 3. Do foundation models have a universal representation?
 4. Fine-tuning
 5. Prompting, zero-shot learning, chain-of-thought prompting
 6. The scaling hypothesis [La parole aux machines, pp.77-81]