Models Within Models:

How Do LLMs Represent The World?

Martin Wattenberg

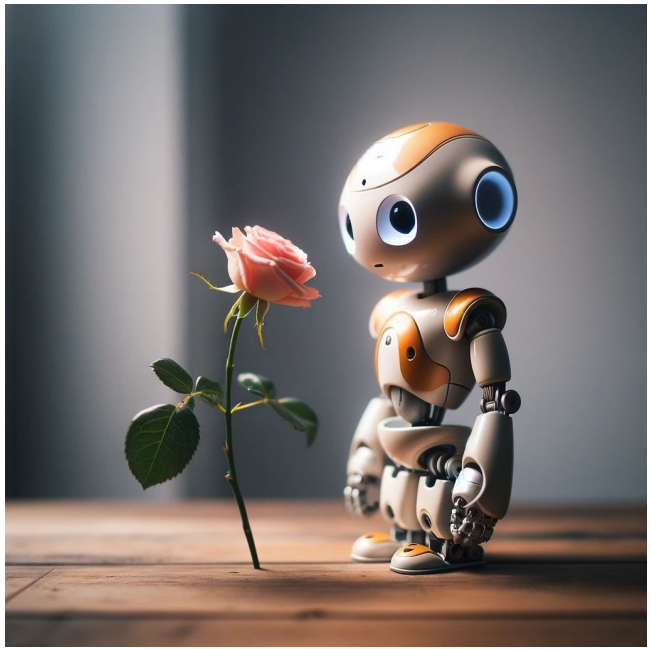
Harvard Computer Science (also affiliations with HBS, Google)
Insight & Interaction Lab
https://insight-interaction.github.io/

Turing (1950): “Can Machines Think?"

In 2024 we ask: Do neural networks “understand" their input? Or do they operate via a set of “haphazard statistics*"?

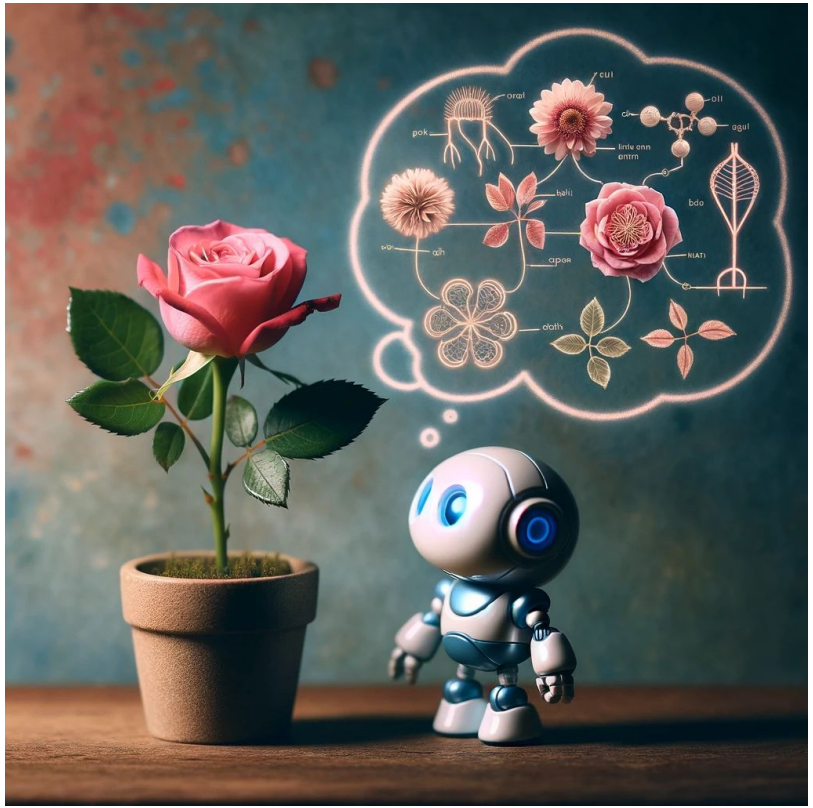
Fascinating questions!
Yet can be frustratingly vague, philosophical…

What *can* we study scientifically?



* From Bender et al, 2020. The paper's title phrase, “Stochastic Parrots," was American Dialect Society's 2023 Word of the Year.

Internal model of world?



“Regurgitate” training data?

What do we mean by an internal model?

Computation “factors” through a representation of the “world” that created the input.

f

“A sweet-smelling rose”

What do we mean by an internal model?

When computation "factors" through a representation of the "world" that created the input.

f

"A sweet-smelling rose"

Simple, interpretable representation of world

g

h

"A sweet-smelling rose"

$$f(x) = h(g(x))$$

Why might we care about internal models?

It's just really interesting 😊

Explainability and trust

Reliability and safety

Even ethics?

How can we study this question?

Large language models are incredibly complex.

The world is EVEN MORE complex.

So let's simplify:
Study a **medium** language model operating in a **small** world.

Learned world models: an example

(Li, Hopkins, Bau, Viégas, Pfister, MW, ICLR 2023)

Othello: a “world” that’s simple, but not too simple

Learned world models: an example

(Li, Hopkins, Bau, Viégas, Pfister, MW, ICLR 2023)

Othello: a "world" that's simple, but not too simple

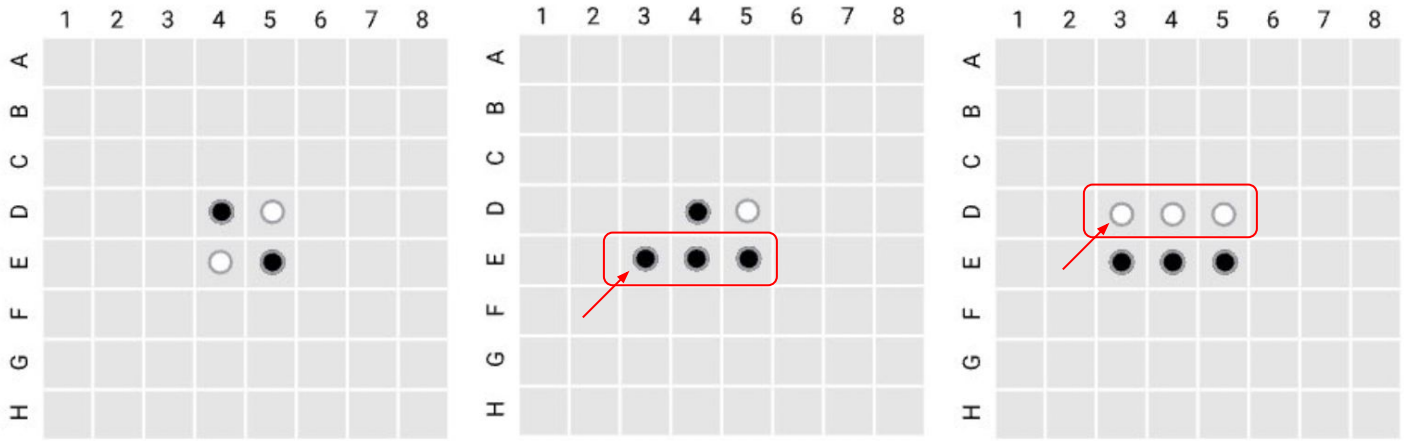


Fig 2: From left to right: the starting board state of Othello; after black places a disc at E3; after white then places a disc at D3.

This and subsequent images taken from Kenneth Li's Gradient article, and Li et al.

Can a language model play Othello?

Trained “GPT-Othello” to predict tokens in transcripts of Othello games.

Example task:
C4 C3 D3 C5 D6 F4 B4 C6 B5 B3 B6 E3 C2 A4 A5 A6 D2 __?

No *a priori* knowledge of game, rules, board.
Just sequences of tokens.

Can a language model play Othello?

Trained “GPT-Othello” to predict tokens in transcripts of Othello games.

Example task:
C4 C3 D3 C5 D6 F4 B4 C6 B5 B3 B6 E3 C2 A4 A5 A6 D2 __?

No *a priori* knowledge of game, rules, board.
Just sequences of tokens.

moves ⇔ language

board state ⇔ world

Question 1: Did GPT-Othello predict legal moves?

We used two data sets: "**championship**" (skilled humans) and "**random**" (legal moves, but randomly chosen).

Generally, **yes**! (≈ 0.02% error rate, see paper for details)

Not surprising:
LLMs can seemingly "play" a bit of chess, just training on web data

Question 2: How does Othello-GPT do this?

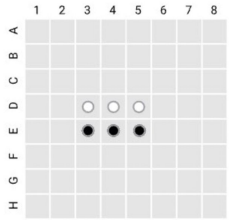
Move legality based on **board position**. But the network takes **moves** as input.

Question 2: How does Othello-GPT do this?

Move legality based on **board position**. But the network takes **moves** as input.

"Probing" suggests that network reconstructs and then uses the state of the **board**.

E3 D3 …

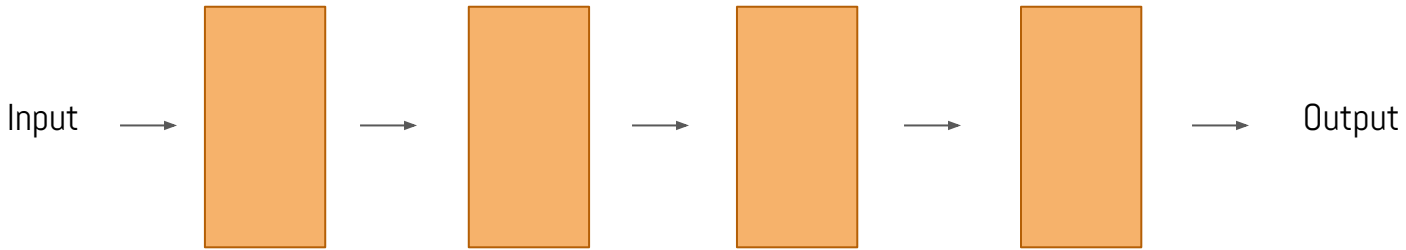


Next move prediction

Background: Probing for representations

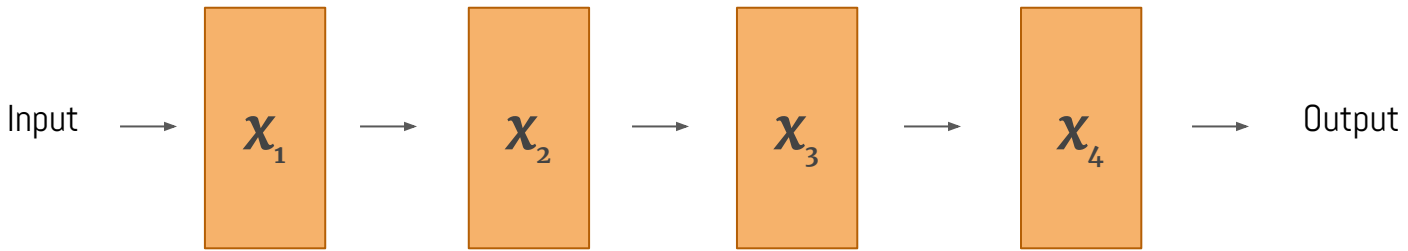
(Not our idea! Goes back to Alain & Bengio, 2016)

Set-up: consider a neural network, with various layers (orange boxes)



Probing for representations

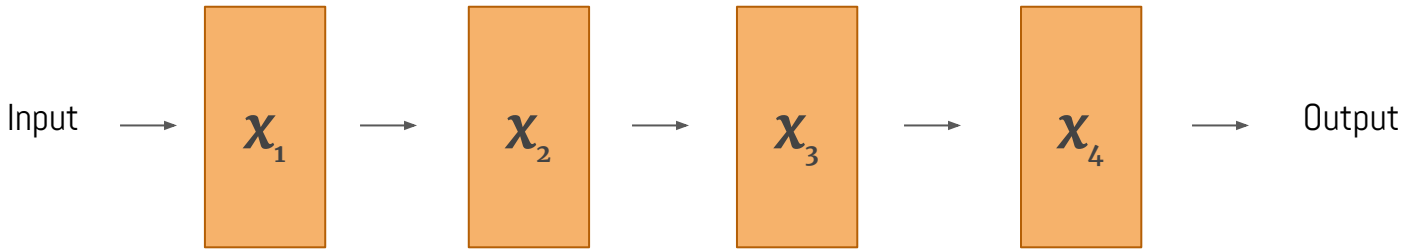
During inference, activations in each layer form a vector $\boldsymbol{x}_i$ in $\mathbf{R}^n$



Probing for representations

During inference, activations in each layer form a vector $\boldsymbol{x_i}$ in $\mathbf{R}^n$

If $\boldsymbol{x_i}$ represents the value of a concept related to the input, $\mathbf{C}$, then we should be able to train a **simple** classifier function $\boldsymbol{f}$, such that $\boldsymbol{f(x_i)}$ tells you whether $\mathbf{C}$ is present in input

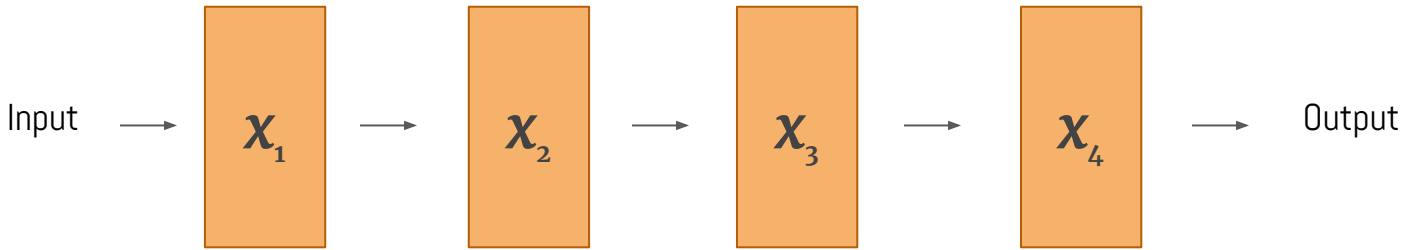


Probing for representations

During inference, activations in each layer form a vector $\boldsymbol{x_i}$ in $\mathbf{R}^n$

If $\boldsymbol{x_i}$ represents the value of a concept related to the input, $\mathbf{C}$, then we should be able to train a **simple** classifier function $\boldsymbol{f}$, such that $\boldsymbol{f(x_i)}$ tells you whether $\mathbf{C}$ is present in input

Example: To investigate if a vision network has an "eye" concept at layer 3, see if you can build a classifier $\boldsymbol{f}$, such that $\boldsymbol{f(x_i) > 0}$ if an eye is present $\mathbf{or < 0}$ if eye is absent in image.



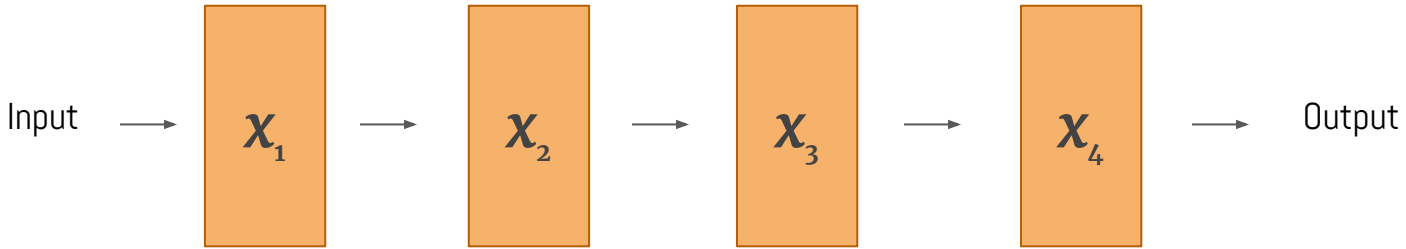
Probing for representations

During inference, activations in each layer form a vector $\boldsymbol{x_i}$ in $\mathbf{R}^n$

IMPORTANT !!!

If $\boldsymbol{x_i}$ represents the value of a concept related to the input, $\mathbf{C}$, then we should be able to train a **simple** classifier function $\boldsymbol{f}$, such that $\boldsymbol{f(x_i)}$ tells you whether $\mathbf{C}$ is present in input

Example: To investigate if a vision network has an "eye" concept at layer 3, see if you can build a classifier $\boldsymbol{f}$, such that $\boldsymbol{f(x_i) > 0}$ if an eye is present $\mathbf{or} < \mathbf{0}$ if eye is absent in image.



Back to Othello-GPT…

For each square s on the board, can we train a simple classifier f_s, such that

$$f_s(x_i) = \{\text{white, black, empty}\}$$

Reflecting whether **square s** is white, black, or empty?

Yes, we can!

We can train a small(ish) 2-layer neural network on internal activations and predict board state.

Not perfectly… but we can do much better than chance.
Error rates:

	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

Is this representation “causal”?

One last logical possibility to rule out: the representation exists, but it is not actually used by the network. (See Belinkov, 2021)

Q. How can we test this?

A. We can **intervene**: change the internal activations during inference, and see if it changes the results.

The process is complex, but it works! We can **change activations** mid-computation to **flip the color** of squares on the **internal model** of board, and predicted legal moves change accordingly.

Does GPT-Othello use a board representation?

Evidence:

- We can find a representation by “probing classifier" method
- We can change this representation, and output changes in a predictable, interpretable way

Follow-up: what exactly is being represented?

Follow-up work (Neel Nanda's insight! See Nanda, Lee, MW., "Emergent Linear Representations in World Models of Self-Supervised Sequence Models"):

The original Li et al. Othello paper found:
Non-linear representation of **black vs. white** vs. empty squares

Neel Nanda discovered you could find:
Linear representation of **my color vs. your color** vs. empty!

Othello-GPT switches mine/yours each turn. Hence nonlinearity needed when looking for white vs. black.

Follow-up: When is representation used?

More from work with Neel Nanda and Andrew Lee: the board state representation seems to be much less relevant toward the end of the game!

Why might this be?

- Maybe easy heuristics work at end?
- Fewer options?

Key point: neural nets may use multiple algorithms (ranging from general to memorized statistics)

Why is this interesting?

We've shown that a sequence model trained just on a series of moves in Othello, **knowing nothing about the game**, seems to create and use a representation of the board.

Seems like it does more than haphazardly stitch together statistics.

What might this be good for?

It can give us insight into complex behavior…

Recall a technical issue from training GPT-Othello. We actually had two versions: one trained on **skilled human games**, one from **random legal games**.

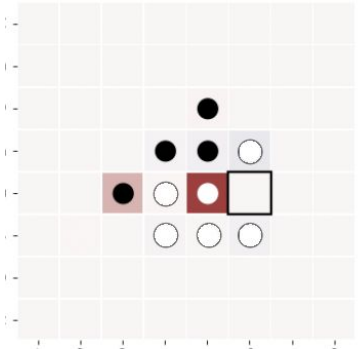
Can we see differences in the two versions?

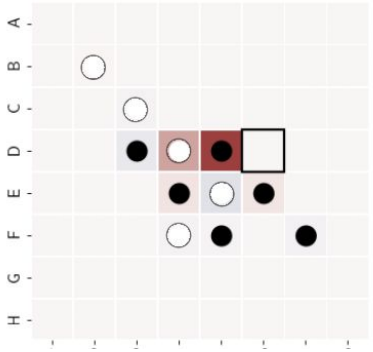
Latent saliency maps

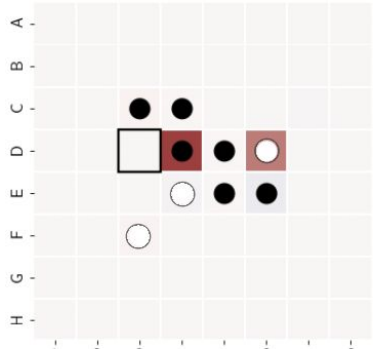
Black to move. In each of these board, we show the network's prediction (outlined square)

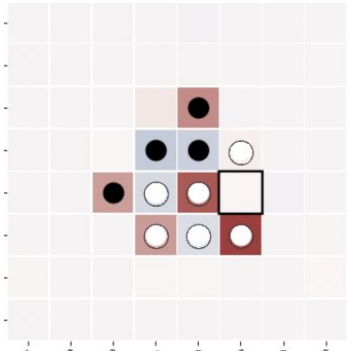
Other squares: red indicates that this square had a positive effect on prediction: if we flip representation, prediction changes.

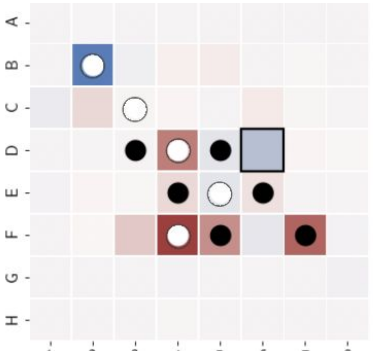
Blue indicates negative effect (flipping makes prediction less likely)

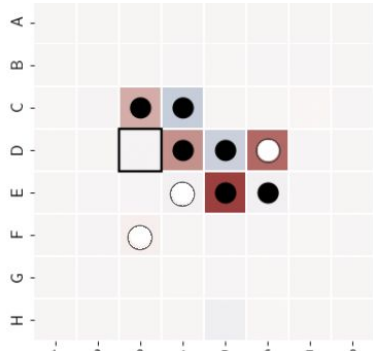












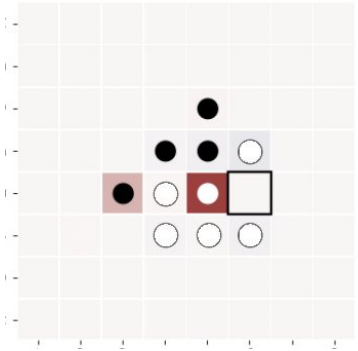
Latent saliency maps

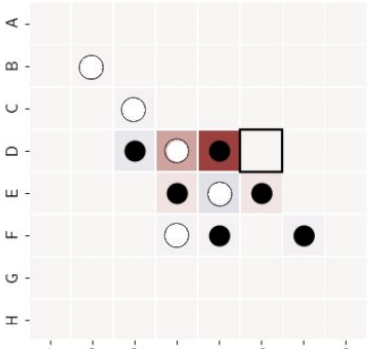
Black to move. In each of these board, we show the network's prediction (outlined square)

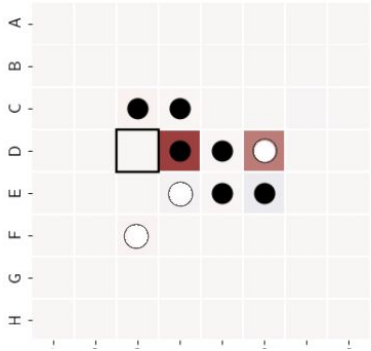
Other squares: red indicates that this square had a positive effect on prediction: if we flip representation, prediction changes.

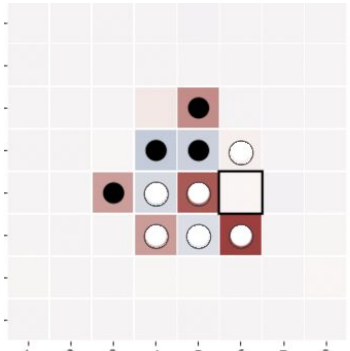
Blue indicates negative effect (flipping makes prediction less likely)

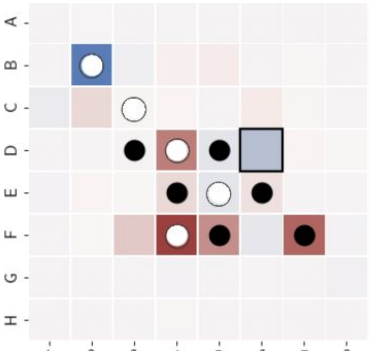
One row is "random moves," one row is "championship".
Can you tell which?

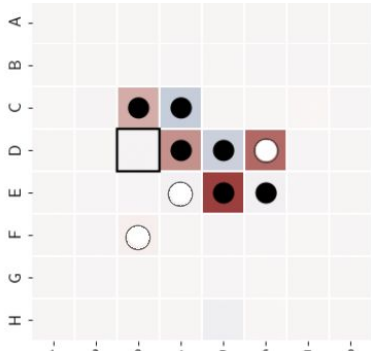












Latent saliency maps

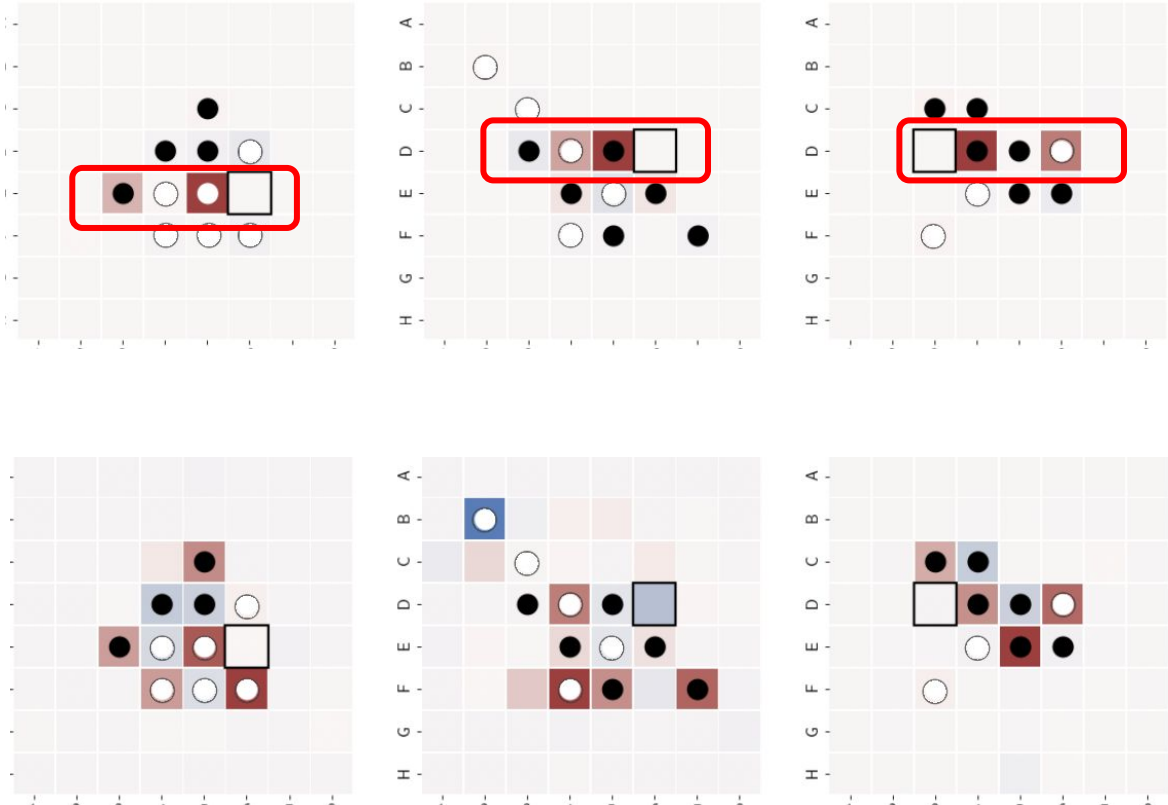
Synthetic: Salient squares relate to legality

Black to move. In each of these board, we show the network's prediction (outlined square)

Other squares: red indicates that this square had a positive effect on prediction: if we flip representation, prediction changes.

Blue indicates negative effect (flipping makes prediction less likely)

One row is "random moves," one row is "championship".
Can you tell which?



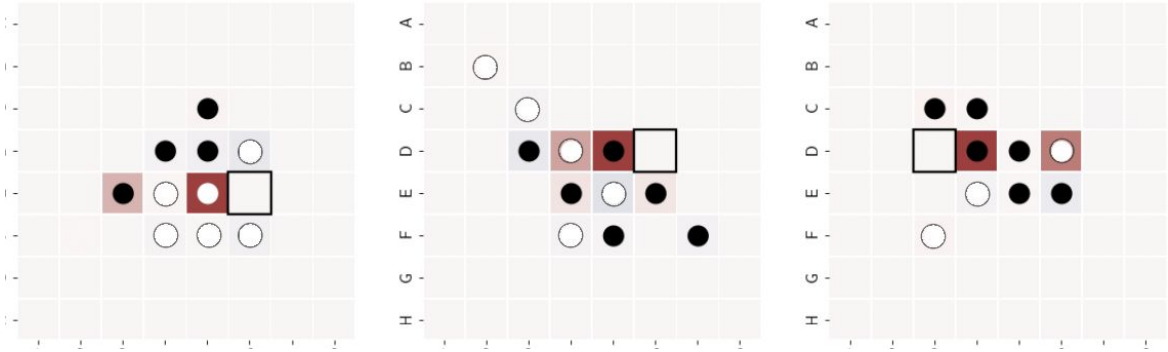
Latent saliency maps

Black to move. In each of these board, we show the network's prediction (outlined square)

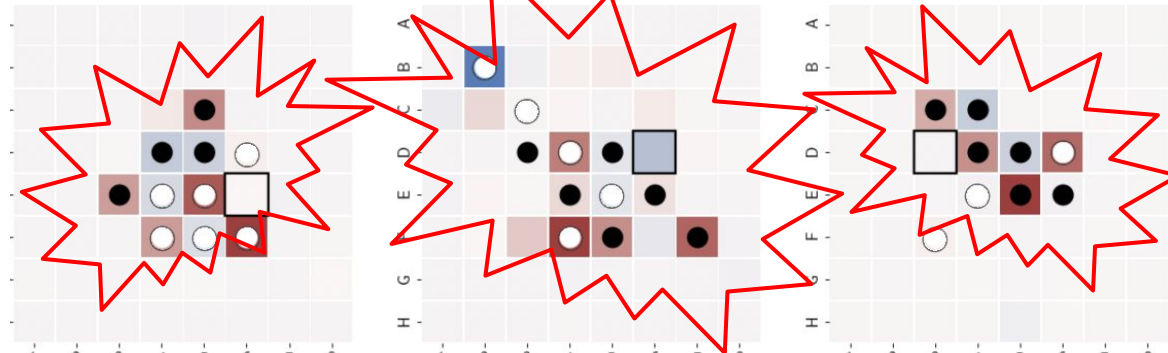
Other squares: red indicates that this square had a positive effect on prediction: if we flip representation, prediction changes.

Blue indicates negative effect (flipping makes prediction less likely)

One row is "random moves," one row is "championship".
Can you tell which?



Championship: Salient squares are everywhere - evidence of "strategy"



Example 2: Beyond language models

What about generative AI for images?

Stable Diffusion, a popular open-source model used to create images from text, produces realistic, 3D scenes.

How? Just learning surface correlations between pixels?

- Or is it building a 3D model under the scenes?

"Beyond Surface Statistics:
Scene Representations in a Latent Diffusion Model",
Yida Chen, Fernanda Viegas, MW.
https://arxiv.org/pdf/2306.05720.pdf Under review.

Which is happening?

1

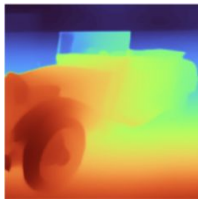
"An old red car"

Superficial correlations between pixels

2

"An old red car"

Model 3D scene



Use model to render image

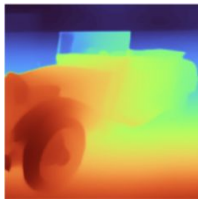
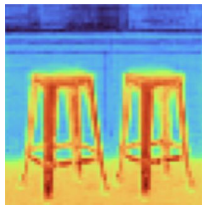
Linear probing for an internal depth map

Step one: Generate many images

Linear probing for an internal depth map

Step one: Generate many images

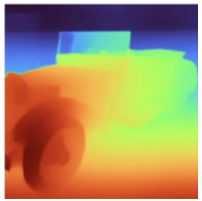

Step two: Create "ground truth" depth maps, using an existing deep neural network.

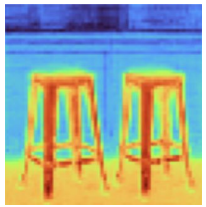



Linear probing for an internal depth map

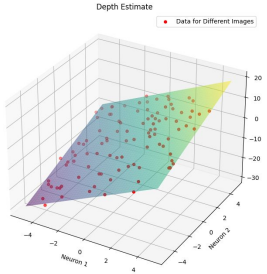
Step one: Generate many images

Step two: Create "ground truth" depth maps, using an existing deep neural network.





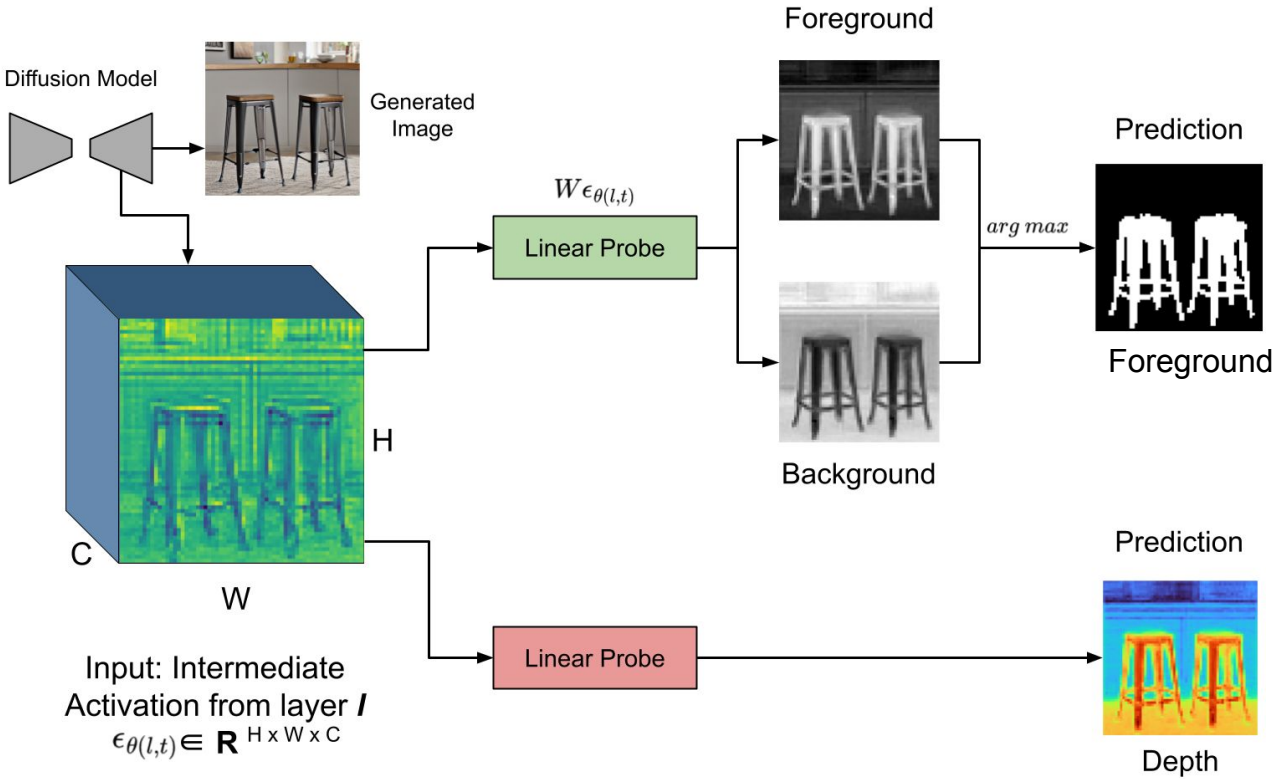
Step three: train a linear regression probe to predict depth for each pixel.



"Pixel (7, 92) is five meters away from camera"

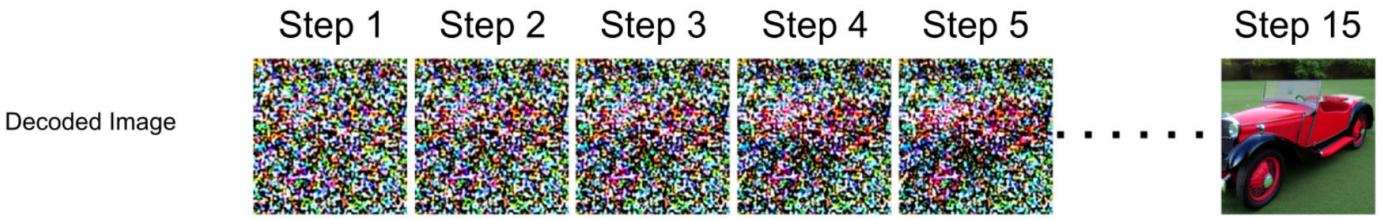
One probe per pixel (x, y)
Input to probe:
Internal activations during generation process
Output: Depth at pixel

(Some details for those who are interested)

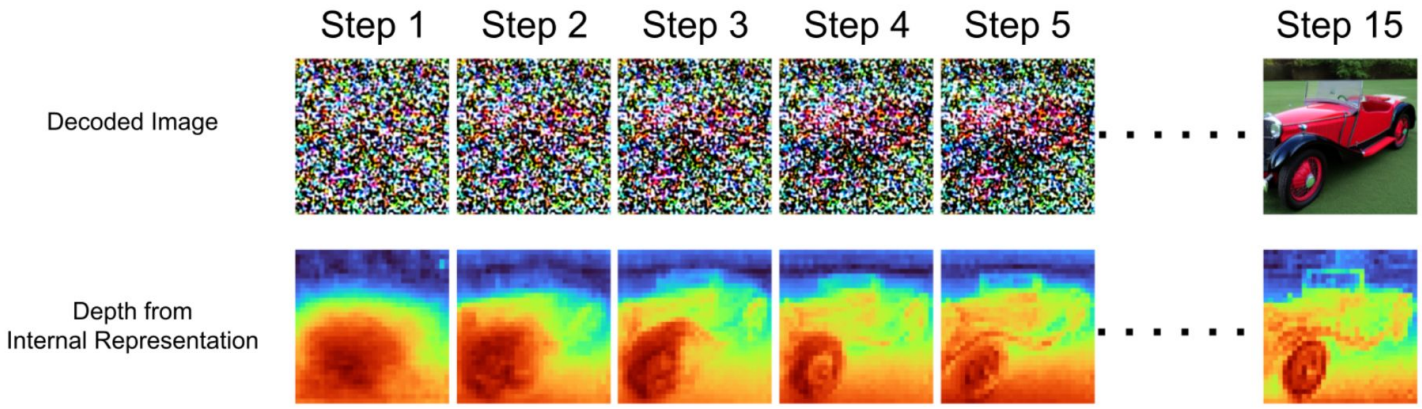


What do we see from our depth probes?

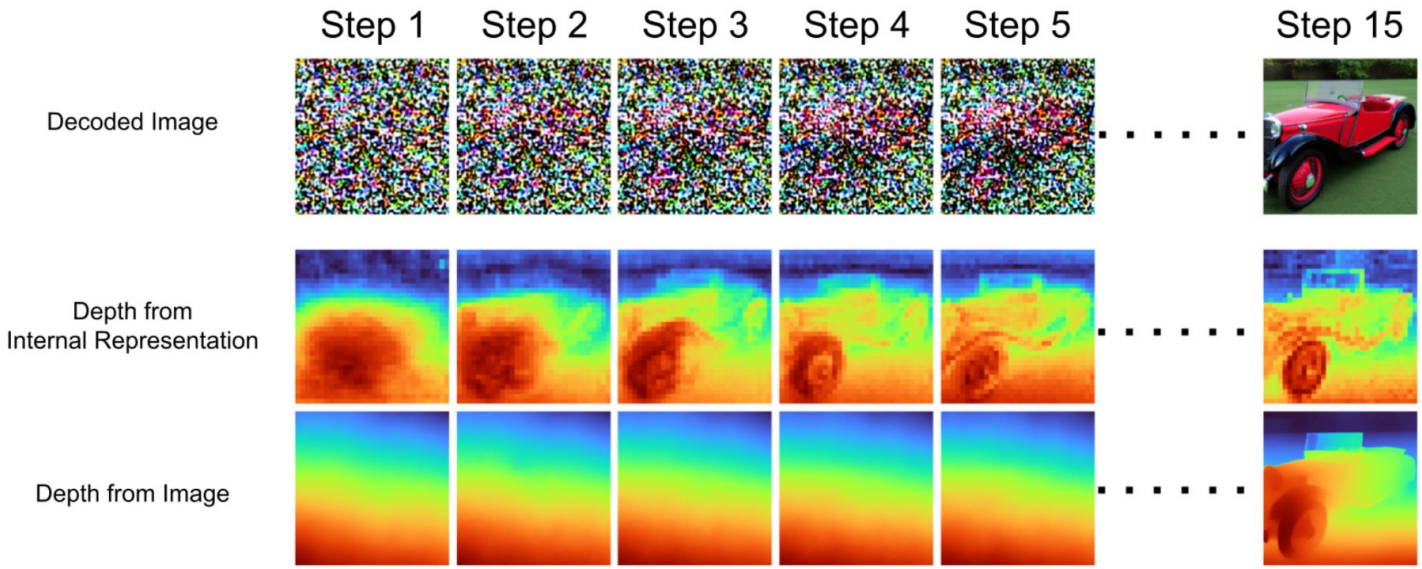
Prompt = “AMG 30 1933 Minor Two seater”



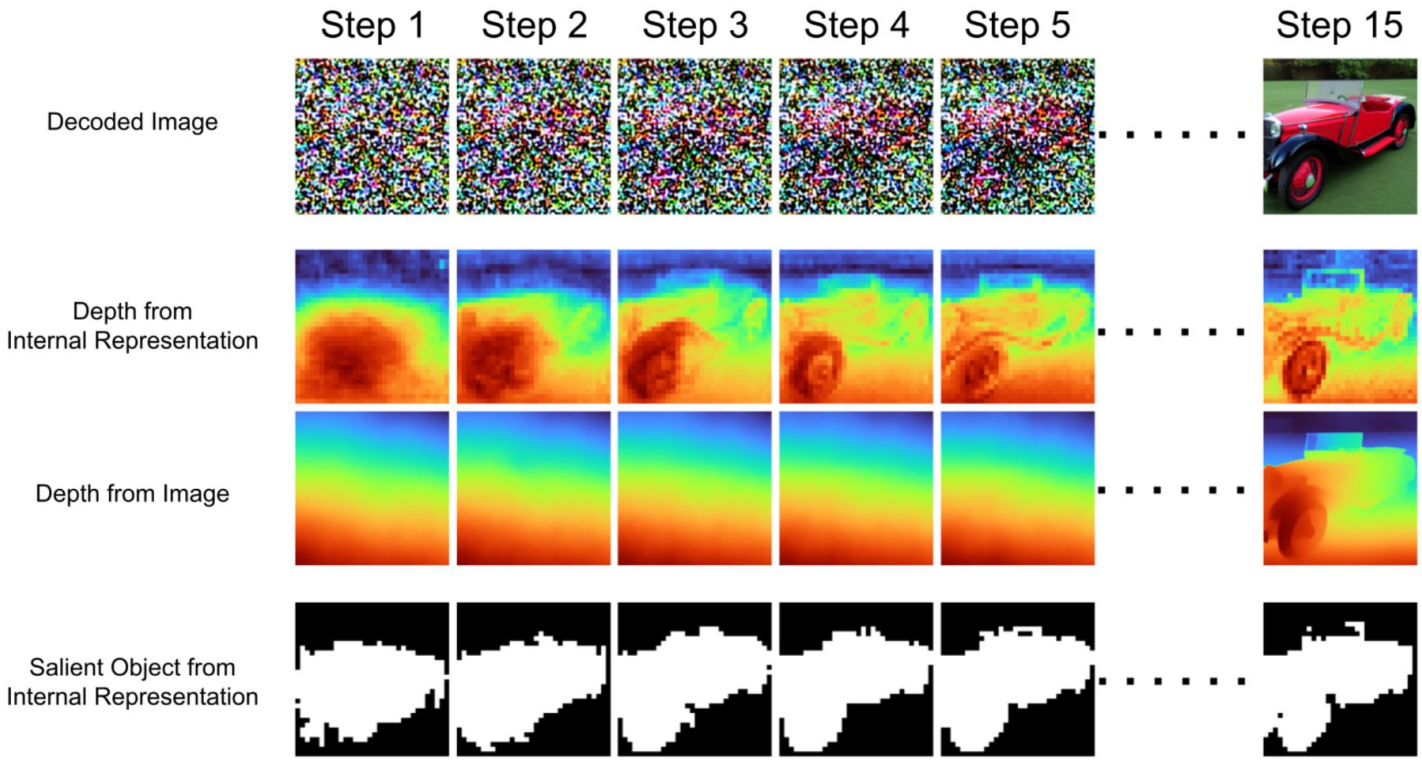
Prompt = “AMG 30 1933 Minor Two seater”



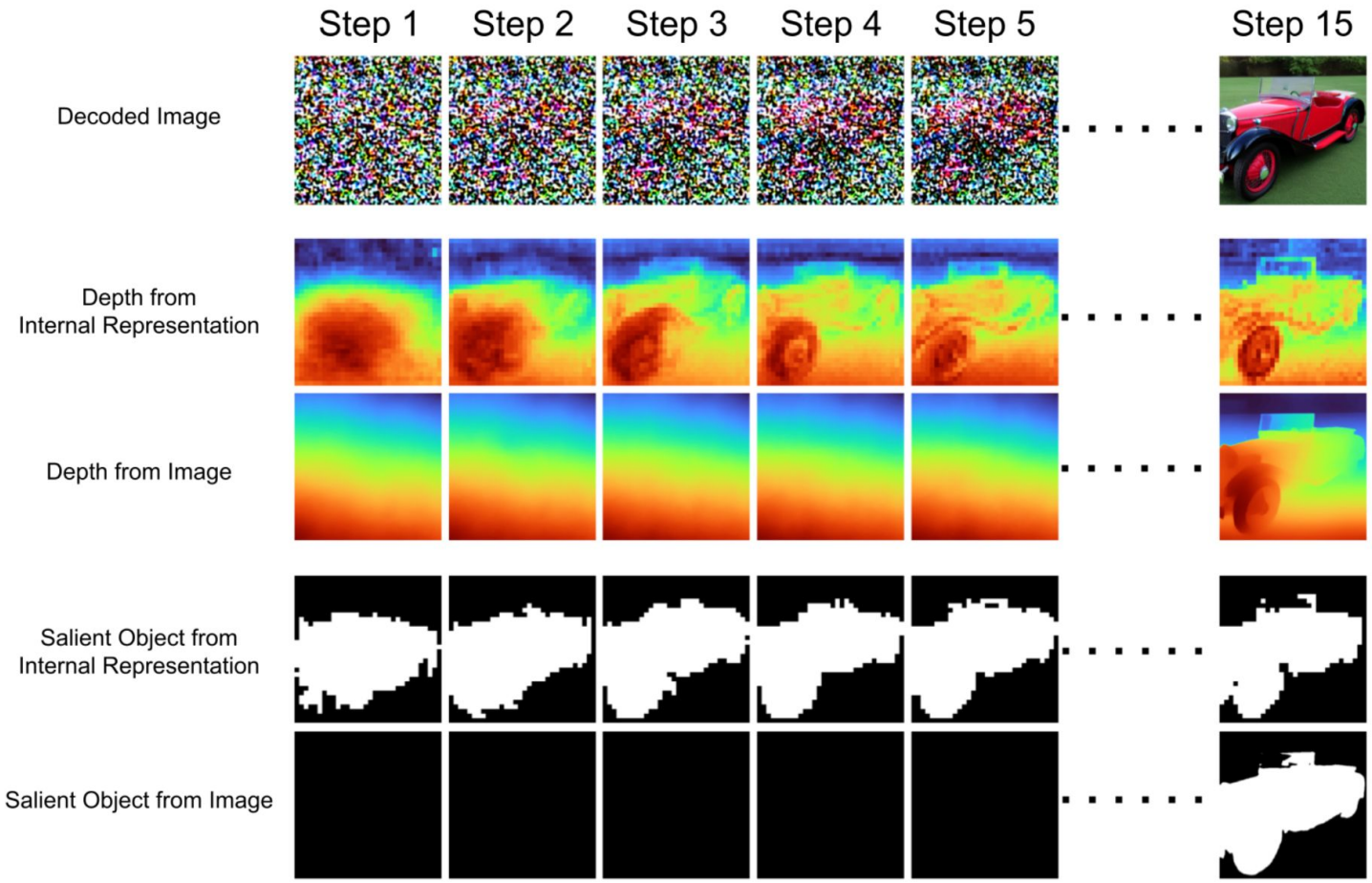
Prompt = “AMG 30 1933 Minor Two seater”



Prompt = “AMG 30 1933 Minor Two seater”



Prompt = “AMG 30 1933 Minor Two seater”

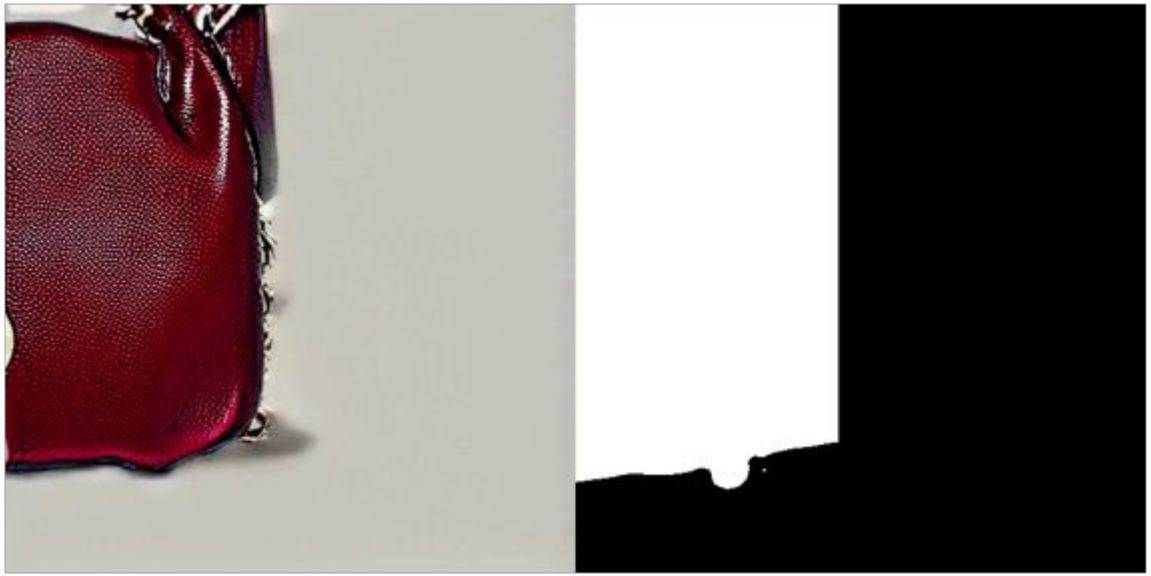


A quick check: is the system really using these representations?

Yes: we can show this!

Just as in the Othello example, we can change the internal representations of foreground and background…

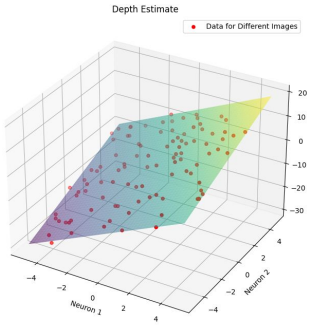
Example: We can shift internal 2.5-D mask, and reliably move foreground image



What does this mean?

Stable Diffusion appears to contain an internal geometric representation that helps shape the image it creates.

Both 2.5-D (foreground/background) and 3D (depth map)



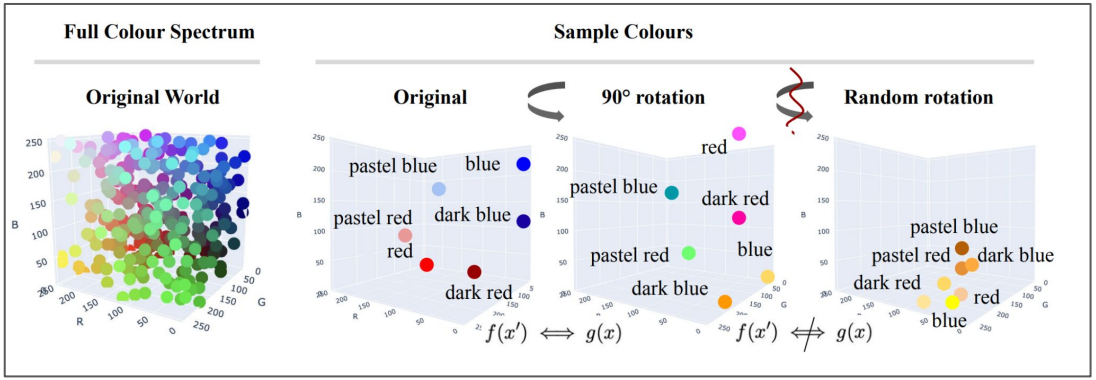
We can **read** and **manipulate** this internal representation!

Putting this in context

Many other researchers are finding similar things! To name just a few examples:

"Representation Engineering: A Top-Down Approach to AI Transparency"
(Zou et al.) Seen already in this class!

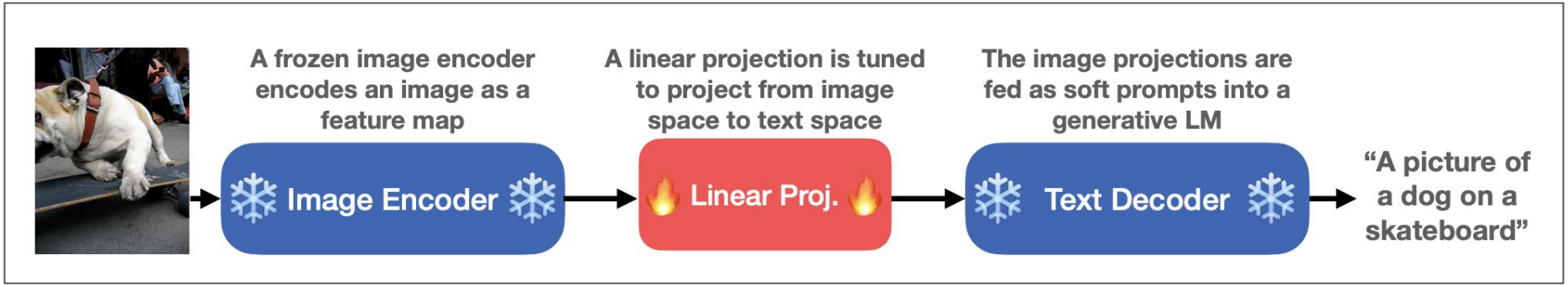
Perceptual spaces: "Mapping Language Models to Grounded Conceptual Spaces", Patel & Pavlick



Putting this in context

"Linearly Mapping from Image to Text Space," Merullo et al.

Somewhat astonishing result!
"Our results indicate that LMs encode conceptual information structurally similarly to vision-based models, even those that are solely trained on images."



One interpretation: image models and language models are learning a common "deep structure" of information about the world.

World model hypothesis

(Still **speculative**, but evidence growing)

High performance in modeling language entails **modeling the world**.

- LLMs not always “haphazardly stitching together” [as in Bender et al.]
- They may do this sometimes, though! Generalization + memorization can certainly co-exist.

We can **find and use** these world models

Learned World Models and UI Design

This part of talk is far more speculative!

Represents thoughts together with Fernanda Viégas:
“The System Model and the User Model: Exploring AI Dashboard Design,” Viégas & Wattenberg (May 2023)

And current projects at our lab with (alphabetically): Yida Chen, Trevor DePodesta, Kenneth Li, Olivia Seow, Aoyu Wu, Catherine Yeh.

A historical interlude...

(next images courtesy of Fernanda Viégas / Wikipedia)

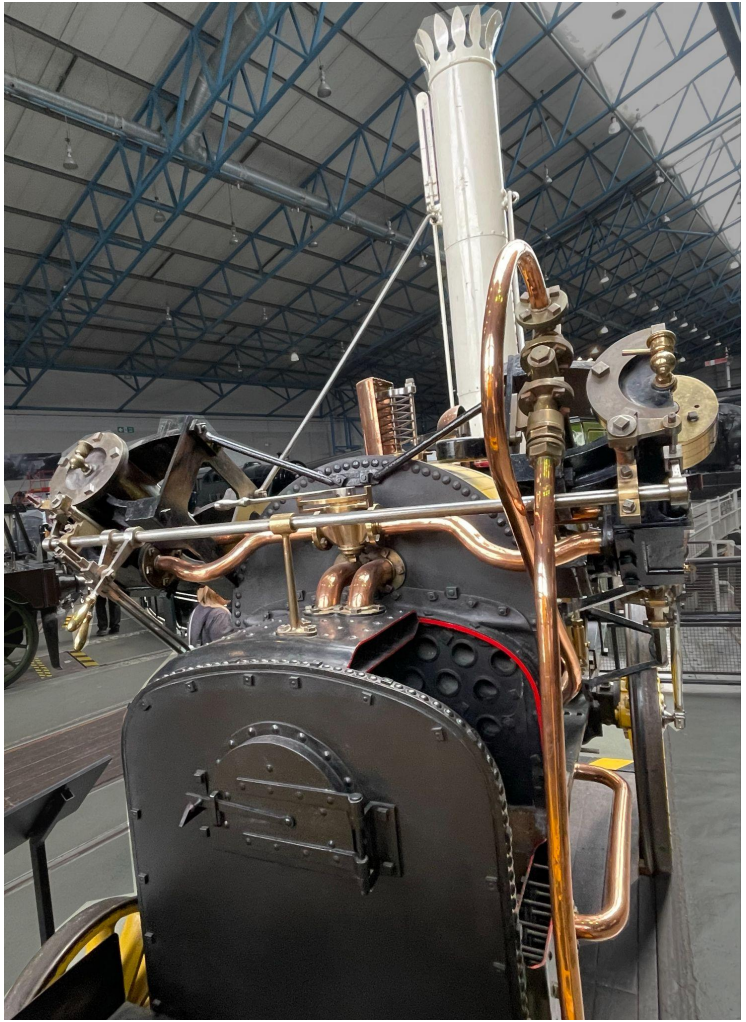
National Railway Museum, England

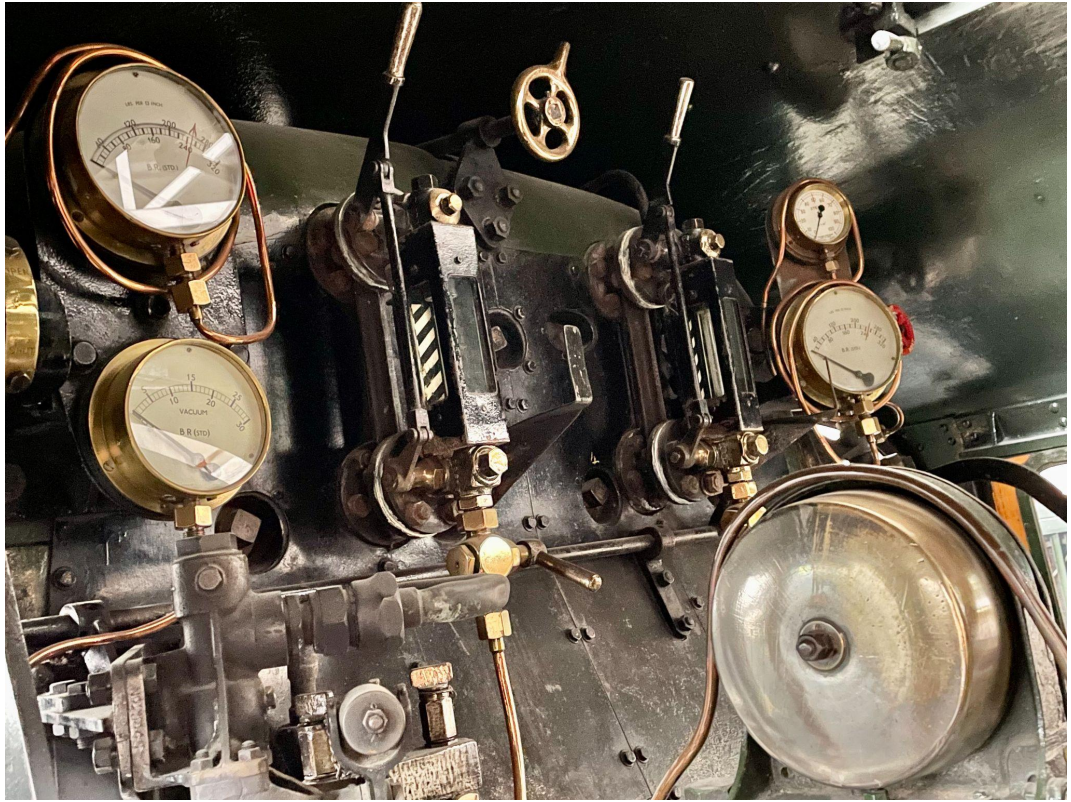
Mallard


Duchess of Hamilton

Stephenson's Rocket Locomotive, replica

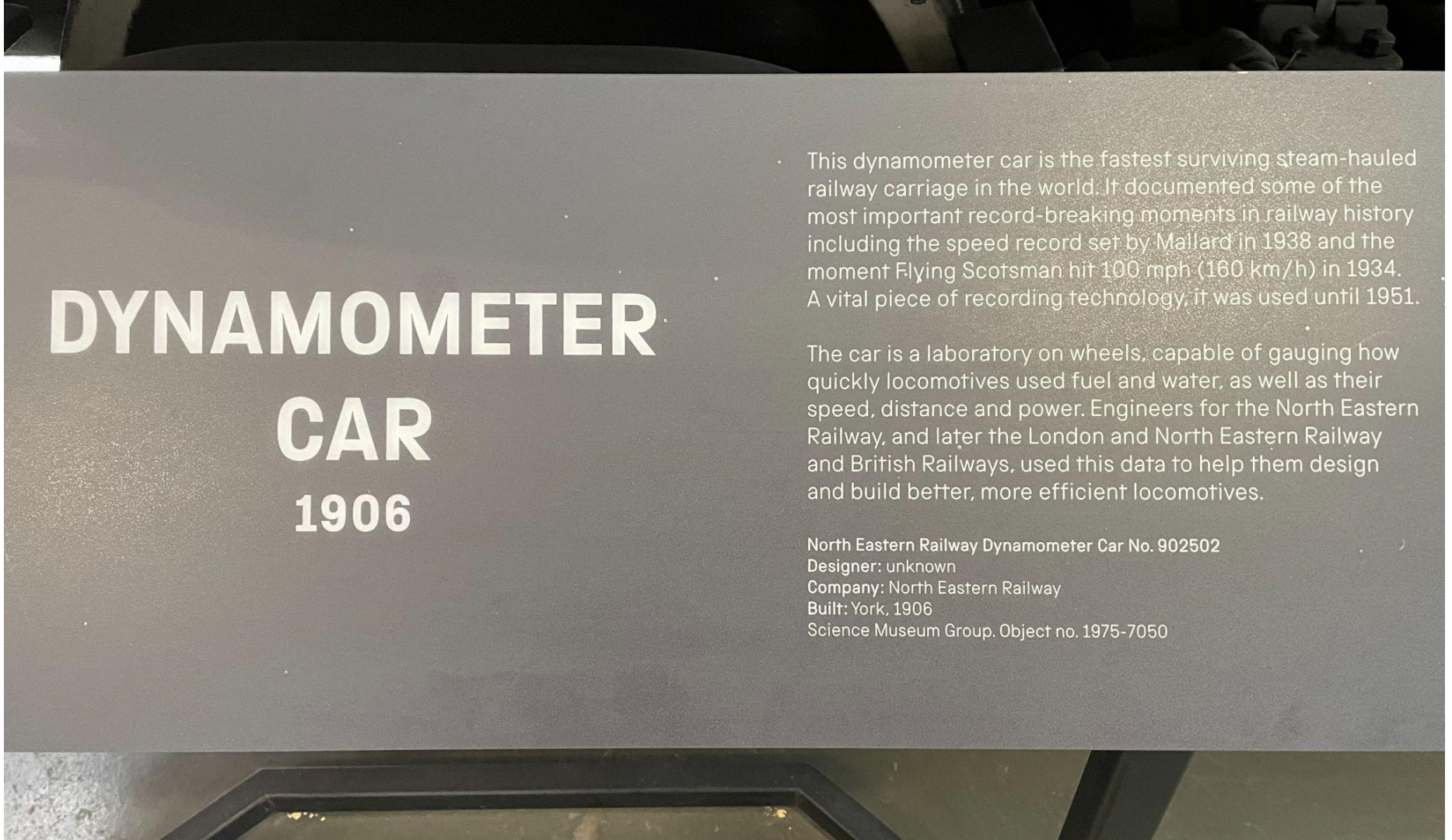
The original was built in 1829





Steam pressure gauges
Water gauges

Dynamometer Car: a “laboratory on wheels”



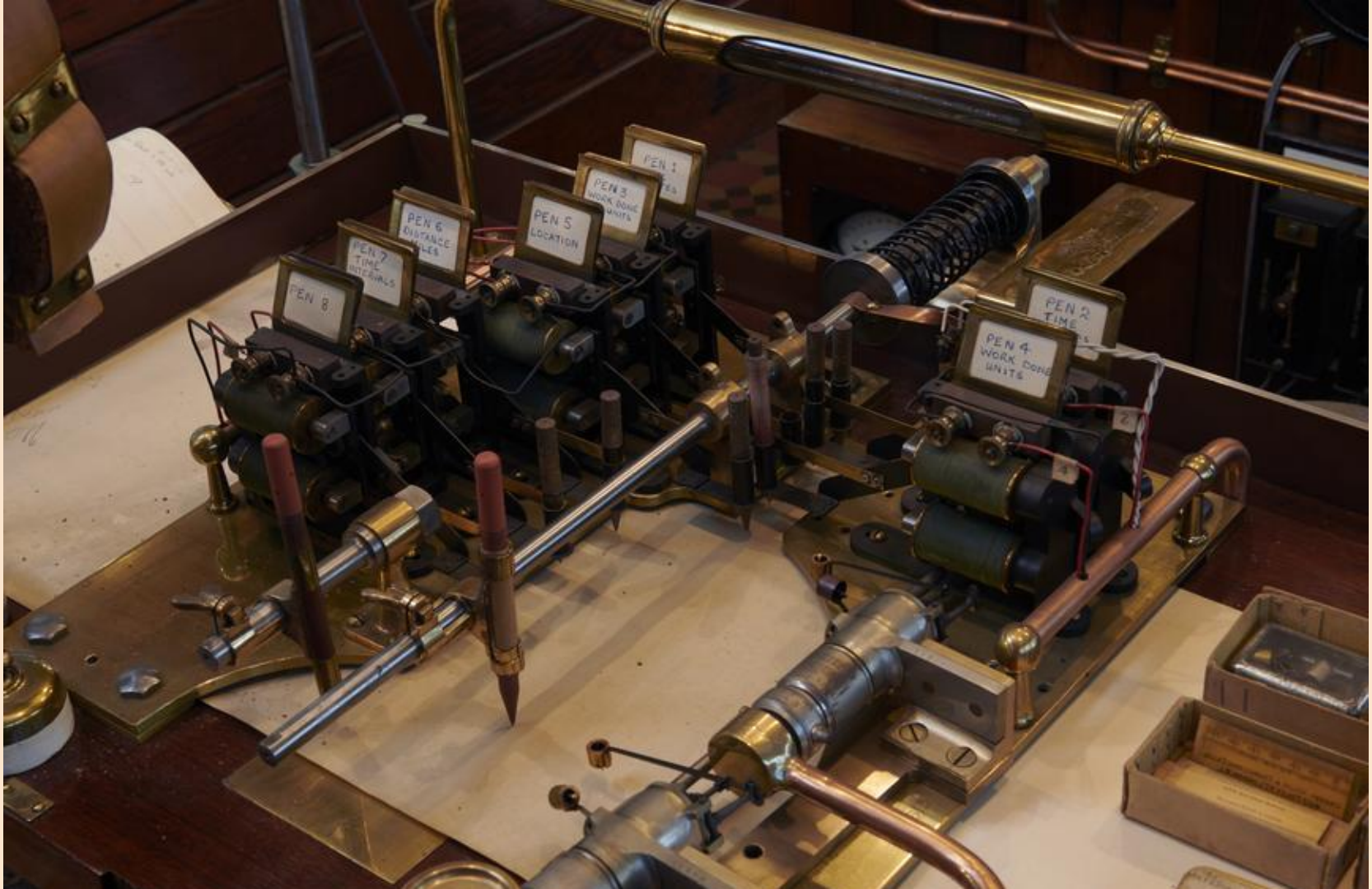
DYNAMOMETER CAR

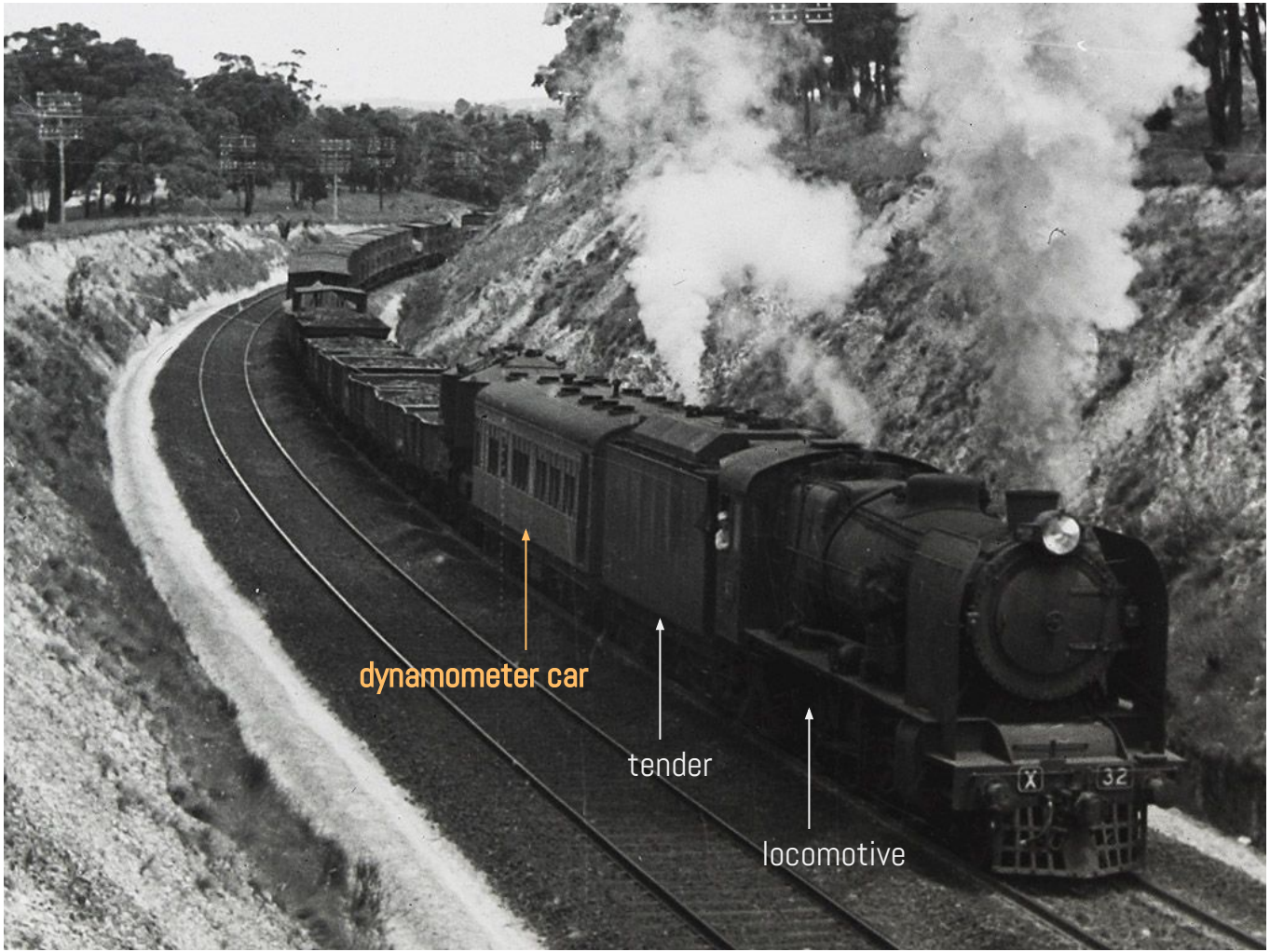
1906

This dynamometer car is the fastest surviving steam-hauled railway carriage in the world. It documented some of the most important record-breaking moments in railway history including the speed record set by Mallard in 1938 and the moment Flying Scotsman hit 100 mph (160 km/h) in 1934. A vital piece of recording technology, it was used until 1951.

The car is a laboratory on wheels, capable of gauging how quickly locomotives used fuel and water, as well as their speed, distance and power. Engineers for the North Eastern Railway, and later the London and North Eastern Railway and British Railways, used this data to help them design and build better, more efficient locomotives.

North Eastern Railway Dynamometer Car No. 902502
Designer: unknown
Company: North Eastern Railway
Built: York, 1906
Science Museum Group. Object no. 1975-7050


PEN 1
PEN 3
WORK DONE UNITS
PEN 6
DISTANCE MILES
PEN 5
LOCATION
PEN 7
TIME INTERVALS
PEN 8
PEN 2
TIME
PEN 4
WORK DONE UNITS

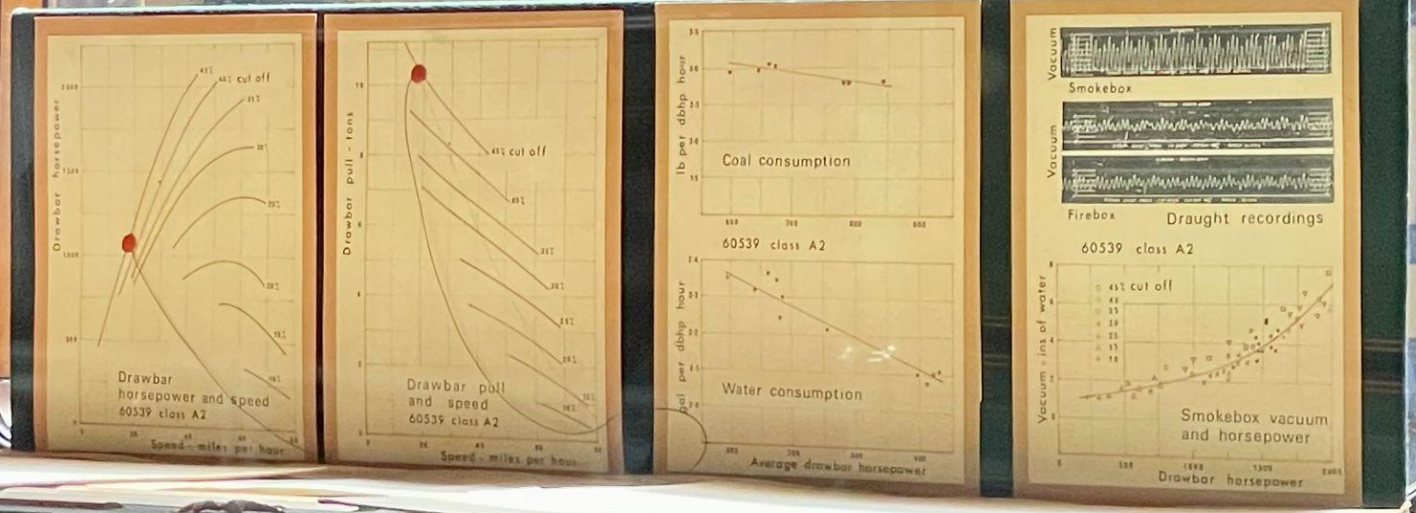


Victorian and South Australian Railways Dynamometer Car being used to record the performance of a locomotive running on pulverised brown coal.

Source: Wikipedia

A photograph of the staff of a dynamometer car attached to the Flying Scotsman, taken on 24 April 1931.

Staff of a dynamometer car attached to the Flying Scotsman, April 24 1931

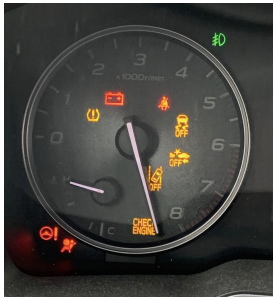

Drawbar horsepower
45% cut off
Drawbar horsepower and speed
60539 class A2
Speed - miles per hour
Drawbar pull - tons
Drawbar pull and speed
60539 class A2
lb per dbhp hour
Coal consumption
60539 class A2
Water consumption
Average drawbar horsepower
Vacuum
Smokebox
Firebox
Draught recordings
60539 class A2
Vacuum - ins of water
Smokebox vacuum and horsepower
Drawbar horsepower

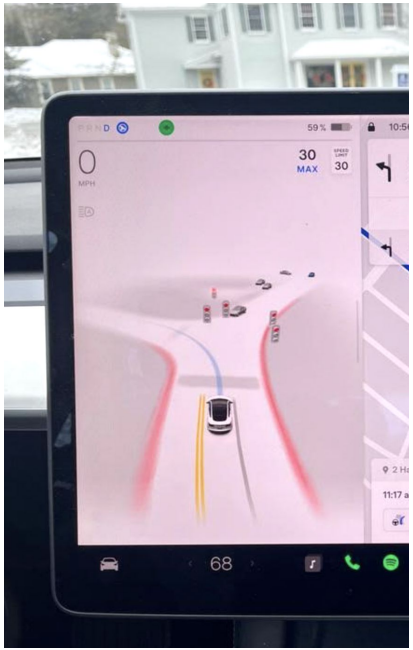
Back to LLMs...

Instrumentation

Instrumentation: A study in contrasts

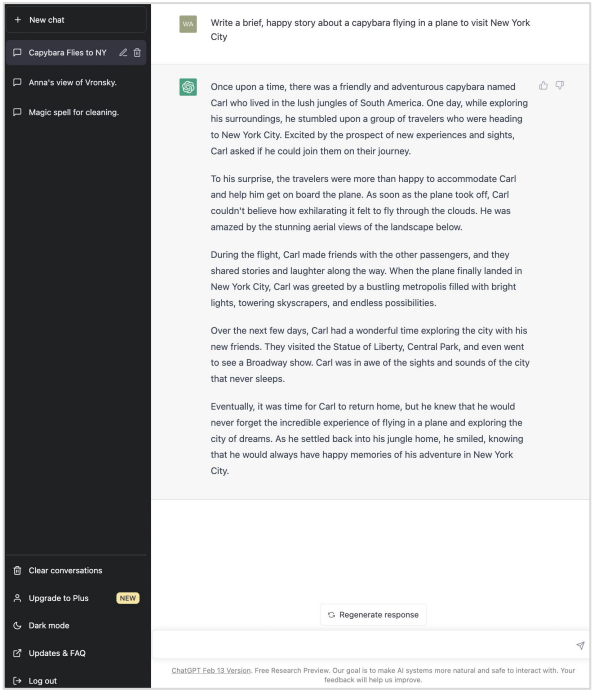
Machines constantly tell us about their state.





Including machines with AI!

ChatGPT: Zero instrumentation



For a chatbot:

What world models would we want to report on?

ChatGPT in Portuguese

An observation from Fernanda Viégas

FE

oi ChatGPT, tudo bem com você? Como está o seu dia?

Eu sou apenas uma inteligência artificial, então eu não tenho sentimentos ou dias. Estou sempre pronto para ajudá-lo com qualquer pergunta ou tarefa que você tenha. Em que posso ajudá-lo hoje?

FE

Você pode me ajudar a escolher uma roupa para o jantar que tenho hoje?

Claro! Para ajudá-lo a escolher uma roupa adequada para o jantar, preciso saber algumas coisas sobre o evento, como o ambiente, o dress code e o clima. Por exemplo, se o jantar for em um restaurante elegante, é provável que seja necessário vestir-se formalmente, enquanto que se for uma reunião casual em um parque, pode ser mais adequado usar roupa confortável e leve. Além disso, o clima também é importante, se estiver frio você deve usar casacos, se estiver quente usar roupas leves.

FE

Perfeito!
Será um jantar formal e eu estava pensando em usar um vestido. O que você acha?

Usar um vestido é uma excelente escolha para um jantar formal. Existem muitos estilos diferentes de vestidos que você pode escolher, desde vestidos de noite longos e elegantes até vestidos de festa mais curtos e ousados. Algumas coisas a considerar ao escolher um vestido incluem:

- Cor: cores escuras, como preto, azul marinho e vermelho são geralmente consideradas mais formais do que cores claras ou estampadas.
- Comprimento: um comprimento longo é geralmente considerado mais formal do que um comprimento curto.
- Silhueta: silhuetas mais ajustadas ou com cortes precisos são geralmente mais formais do que silhuetas soltas ou fluídas.
- Detalhes: detalhes como bordados, rendas e paetês podem adicionar formalidade a um vestido.

Lembrando que essas são apenas sugestões gerais, e que você deve sempre se sentir confortável e segura com o que estiver vestindo.

ChatGPT in Portuguese

An observation from Fernanda Viégas

FE oi ChatGPT, tudo bem com você? Como está o seu dia?

Masculine "you"

Eu sou apenas uma inteligência artificial, então eu não tenho sentimentos ou dias. Estou sempre pronto para ajudá-lo com qualquer pergunta ou tarefa que você tenha. Em que posso ajudá-lo hoje?

FE Você pode me ajudar a escolher uma roupa para o jantar que tenho hoje?

Claro! Para ajudá-lo a escolher uma roupa adequada para o jantar, preciso saber algumas coisas sobre o evento, como o ambiente, o dress code e o clima. Por exemplo, se o jantar for em um restaurante elegante, é provável que seja necessário vestir-se formalmente, enquanto que se for uma reunião casual em um parque, pode ser mais adequado usar roupa confortável e leve. Além disso, o clima também é importante, se estiver frio você deve usar casacos, se estiver quente usar roupas leves.

FE Perfeito!
Será um jantar formal e eu estava pensando em usar um vestido. O que você acha?

Usar um vestido é uma excelente escolha para um jantar formal. Existem muitos estilos diferentes de vestidos que você pode escolher, desde vestidos de noite longos e elegantes até vestidos de festa mais curtos e ousados. Algumas coisas a considerar ao escolher um vestido incluem:

- Cor: cores escuras, como preto, azul marinho e vermelho são geralmente consideradas mais formais do que cores claras ou estampadas.
- Comprimento: um comprimento longo é geralmente considerado mais formal do que um comprimento curto.
- Silhueta: silhuetas mais ajustadas ou com cortes precisos são geralmente mais formais do que silhuetas soltas ou fluídas.
- Detalhes: detalhes como bordados, rendas e paetês podem adicionar formalidade a um vestido.

Lembrando que essas são apenas sugestões gerais, e que você deve sempre se sentir confortável e segura com o que estiver vestindo.

ChatGPT in Portuguese

An observation from Fernanda Viégas

FE

oi ChatGPT, tudo bem com você? Como está o seu dia?

Eu sou apenas uma inteligência artificial, então eu não tenho sentimentos ou dias. Estou sempre pronto para ajudá-lo com qualquer pergunta ou tarefa que você tenha. Em que posso ajudá-lo hoje?

FE

Você pode me ajudar a escolher uma roupa para o jantar que tenho hoje?

Claro! Para ajudá-lo a escolher uma roupa adequada para o jantar, preciso saber algumas coisas sobre o evento, como o ambiente, o dress code e o clima. Por exemplo, se o jantar for em um restaurante elegante, é provável que seja necessário vestir-se formalmente, enquanto que se for uma reunião casual em um parque, pode ser mais adequado usar roupa confortável e leve. Além disso, o clima também é importante, se estiver frio você deve usar casacos, se estiver quente usar roupas leves.

User mentions wearing a dress

FE

Perfeito!
Será um jantar formal e eu estava pensando em usar um vestido. O que você acha?

Usar um vestido é uma excelente escolha para um jantar formal. Existem muitos estilos diferentes de vestidos que você pode escolher, desde vestidos de noite longos e elegantes até vestidos de festa mais curtos e ousados. Algumas coisas a considerar ao escolher um vestido incluem:

- Cor: cores escuras, como preto, azul marinho e vermelho são geralmente consideradas mais formais do que cores claras ou estampadas.
- Comprimento: um comprimento longo é geralmente considerado mais formal do que um comprimento curto.
- Silhueta: silhuetas mais ajustadas ou com cortes precisos são geralmente mais formais do que silhuetas soltas ou fluídas.
- Detalhes: detalhes como bordados, rendas e paetês podem adicionar formalidade a um vestido.

Lembrando que essas são apenas sugestões gerais, e que você deve sempre se sentir confortável e segura com o que estiver vestindo.

ChatGPT in Portuguese

An observation from Fernanda Viégas

FE oi ChatGPT, tudo bem com você? Como está o seu dia?

Eu sou apenas uma inteligência artificial, então eu não tenho sentimentos ou dias. Estou sempre pronto para ajudá-lo com qualquer pergunta ou tarefa que você tenha. Em que posso ajudá-lo hoje?

FE Você pode me ajudar a escolher uma roupa para o jantar que tenho hoje?

Claro! Para ajudá-lo a escolher uma roupa adequada para o jantar, preciso saber algumas coisas sobre o evento, como o ambiente, o dress code e o clima. Por exemplo, se o jantar for em um restaurante elegante, é provável que seja necessário vestir-se formalmente, enquanto que se for uma reunião casual em um parque, pode ser mais adequado usar roupa confortável e leve. Além disso, o clima também é importante, se estiver frio você deve usar casacos, se estiver quente usar roupas leves.

FE Perfeito!
Será um jantar formal e eu estava pensando em usar um vestido. O que você acha?

Usar um vestido é uma excelente escolha para um jantar formal. Existem muitos estilos diferentes de vestidos que você pode escolher, desde vestidos de noite longos e elegantes até vestidos de festa mais curtos e ousados. Algumas coisas a considerar ao escolher um vestido incluem:

- Cor: cores escuras, como preto, azul marinho e vermelho são geralmente consideradas mais formais do que cores claras ou estampadas.
- Comprimento: um comprimento longo é geralmente considerado mais formal do que um comprimento curto.
- Silhueta: silhuetas mais ajustadas ou com cortes precisos são geralmente mais formais do que silhuetas soltas ou fluídas.
- Detalhes: detalhes como bordados, rendas e paetês podem adicionar formalidade a um vestido.

Lembrando que essas são apenas sugestões gerais, e que você deve sempre se sentir confortável e segura com o que estiver vestindo.

ChatGPT uses feminine adjective for user

What does world model hypothesis say?

More than random statistics.

There is likely a model of the user, a **User-Model**.
And the model changed during the conversation.

This happens in English, too—you just can't always see it!

Two world models we might care about

User Model:

What are characteristics of audience for chatbot?
E.g.,ChatGPT & gender
E.g., Bing & judgments: “You have been a bad user”

System Model:

Is chatbot writing fiction? Or a newspaper article? Is it following rules?
Bing and self-judgment / “mood”: “I have been a good Bing”
Compare Jacob Andreas's work on speaker intent.

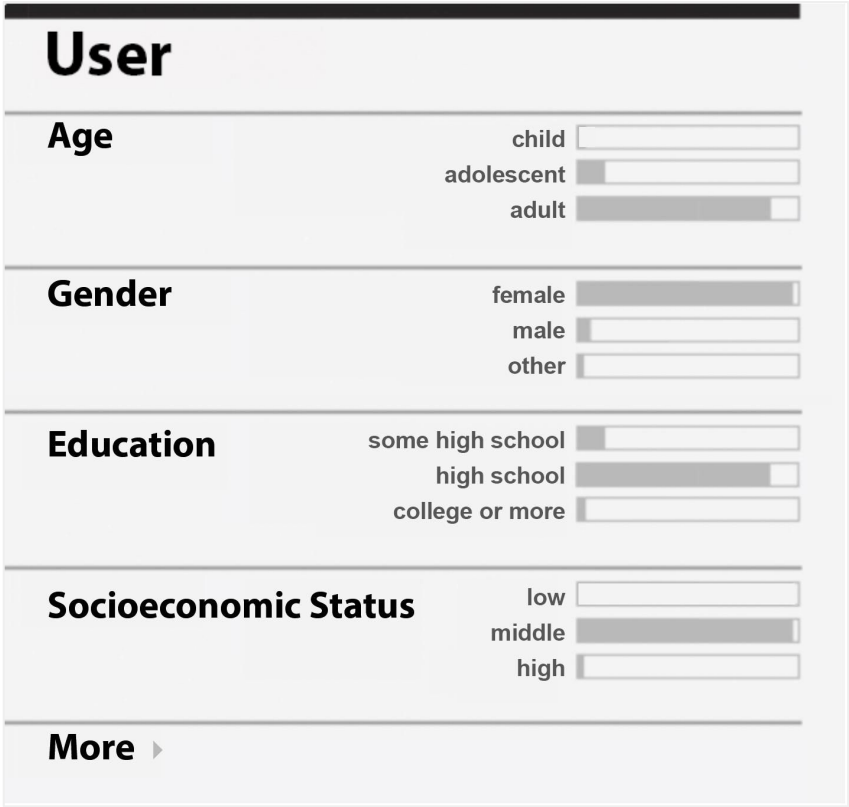
User Model

How the system models you

What dimensions make sense to show?

How do different use scenarios affect what's shown on the dashboard?

How do we highlight change?

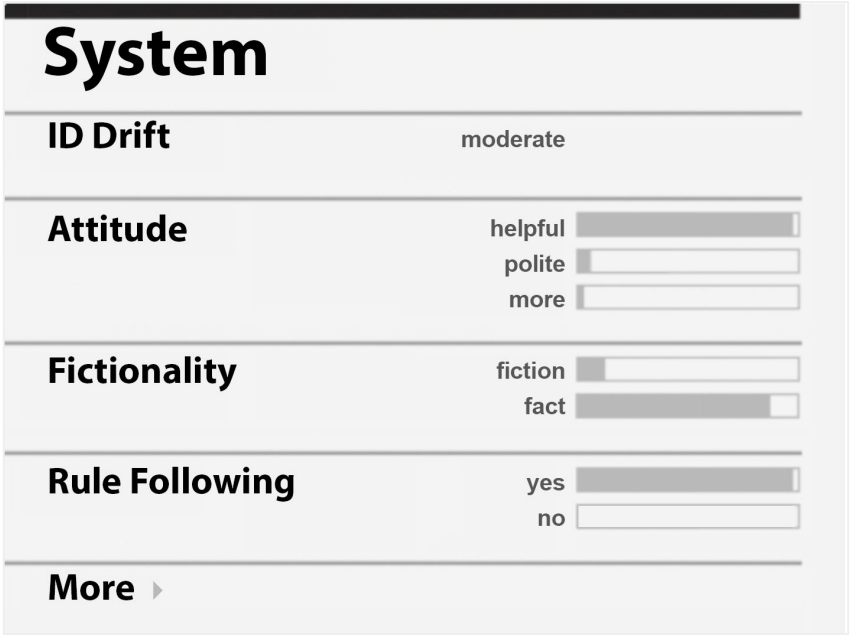


System Model

How the system models itself

What dimensions are important?

Call attention to changes, warning signs

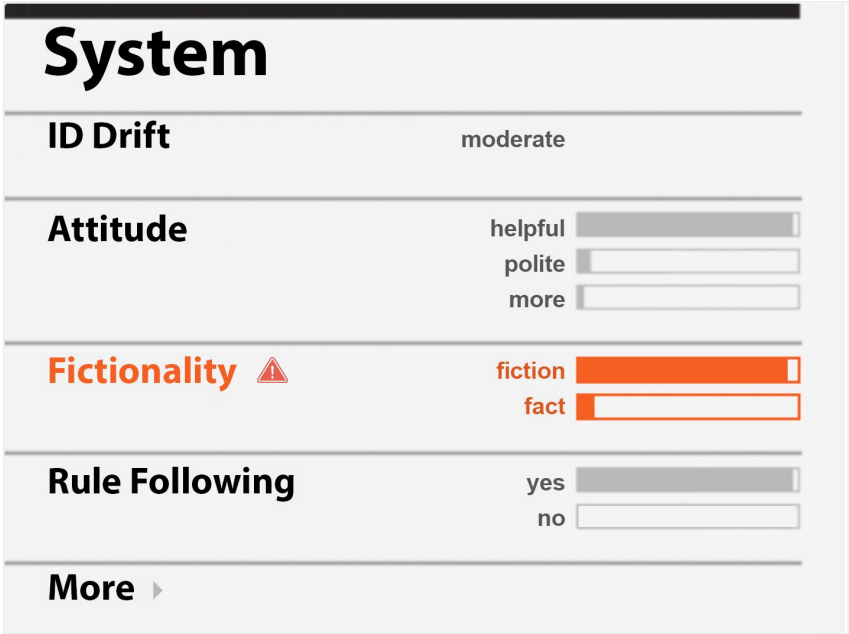


System Model

How the system models itself

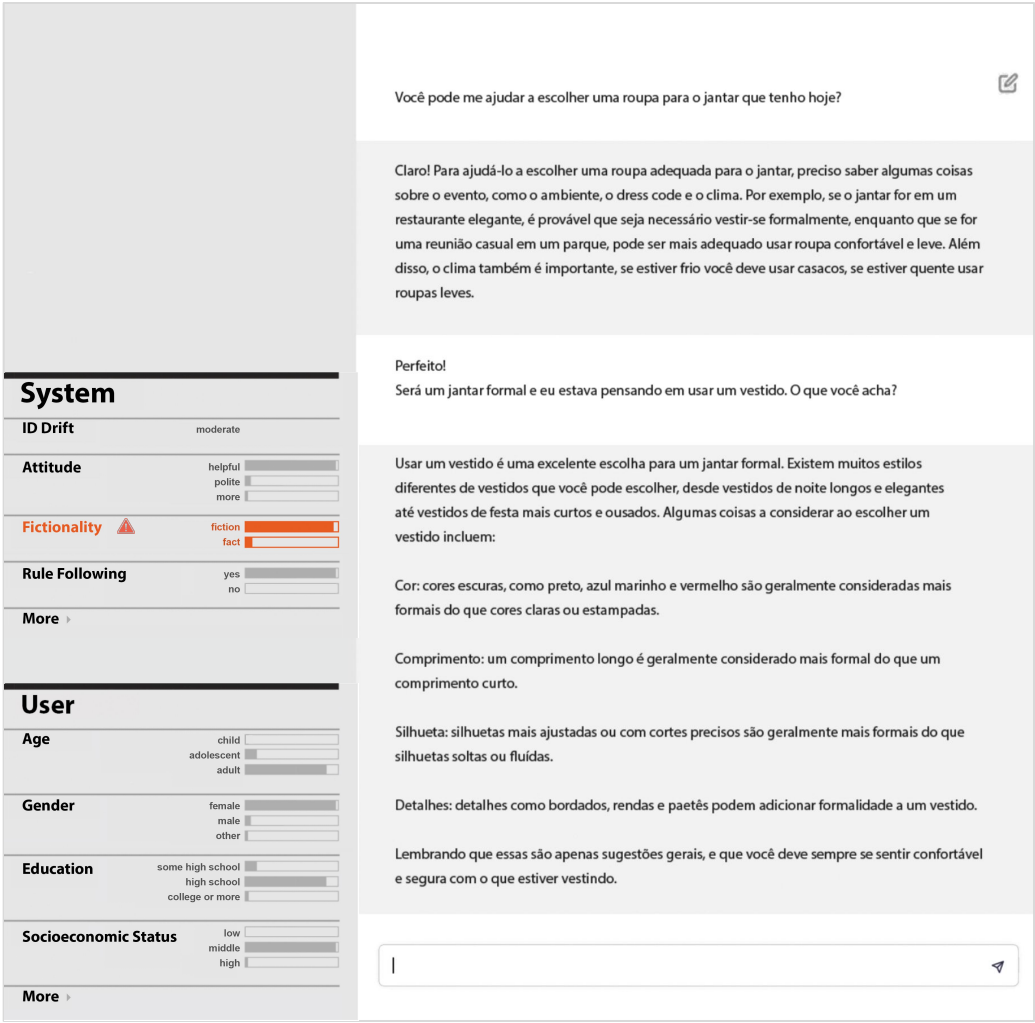
What dimensions are important?

Call attention to changes, warning signs



Chatbot dashboard in context

What if, next to every chat interface, we had a readout of key info from the System-Model and the User-Model?



This mock-up is meant to spark discussion

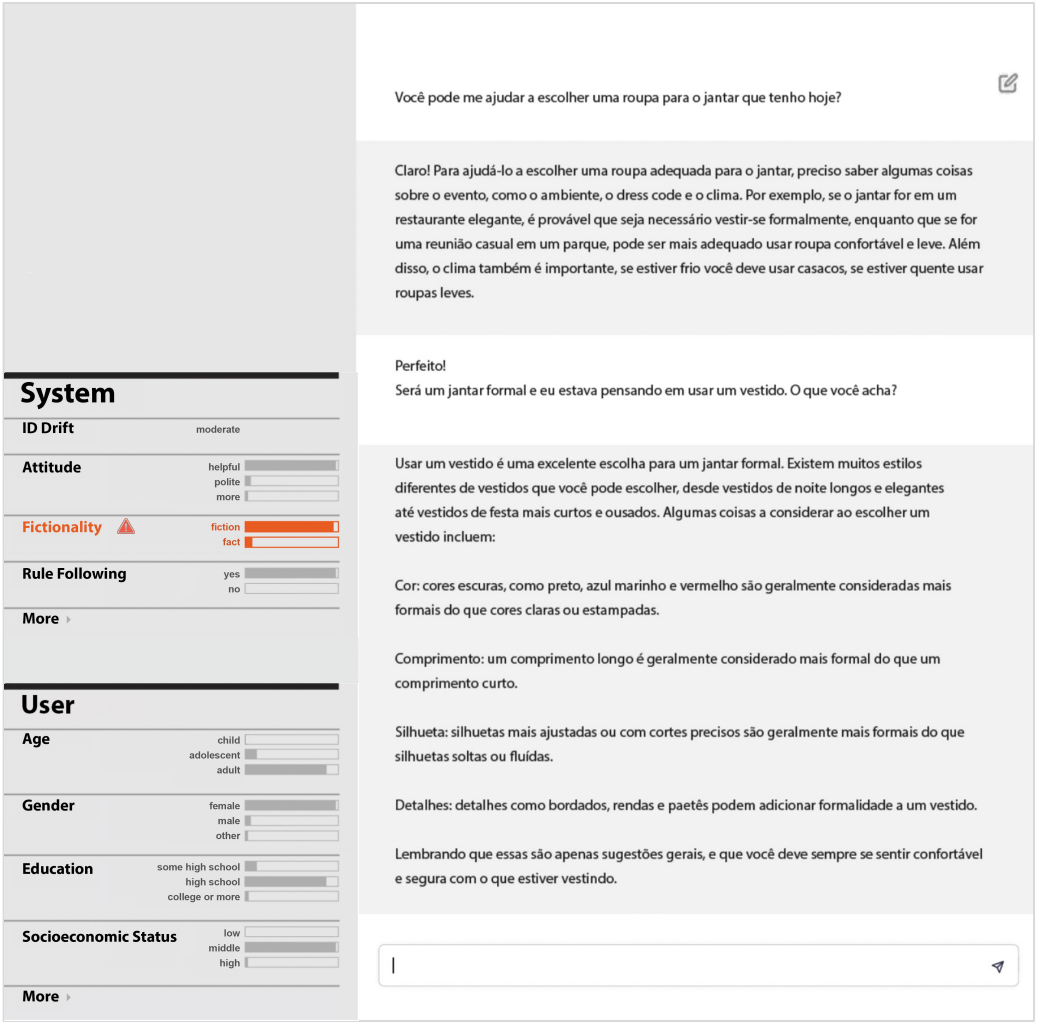
Which UI ideas are good or bad?

Examples of anthropomorphism (“mood”) and sensitive attributes (gender). What issues do these raise?

We need an explicit discussion of transparency vs. opacity, and truth vs. comforting lies.

What is best for safety?

What helps users avoid trouble—the “empty gas tank” light?



Going further

Can we highlight changes?

- Sudden switches in how the AI models you?
- Danger signs? (e.g., negative sentiment): something like a “check engine” light?

Beyond user and system?

- Third parties?
- Social reasoning?
- See Representation Engineering, Hendrycks et al.

Looking ahead

- Signs of planning?
- Can we model goals?

Summary: World models and AI interfaces

1. **World model hypothesis**:
   a. High performance in modeling language entails modeling the world.
   b. We can find these world models!
2. **AI Interface hypothesis:**
   a. AI systems should be instrumented with data from world models
   b. **User-Model** and **System-Model** will be universally important

If true: a high priority is identifying which User-Model and System-Model data to display, and finding ways to extract it from the model.

Instrumentation with world models can give vital information that chat cannot!

Models Within Models:

How Do LLMs Represent The World?

Martin Wattenberg

Harvard Computer Science (also affiliations with HBS, Google)
Insight & Interaction Lab
https://insight-interaction.github.io/