INTERPRETABILITY IN NEUROSCIENCE

Prof. Alexander Huth

2026.1.30

LAST TIME

* *Mechanistic interpretability*

* David Marr's 3 levels

* Linearity, modularity

* World models?

TODAY

* How do we infer what the **goal** of a neural system is? (Kanwisher 2023; Schrimpf 2021; Antonello 2024)

* Can we build **explicitly interpretable** systems? (Benara 2024; Singh 2025)

DISCOVERING GOALS

David Marr's 3 levels of analysis:

Computational – Algorithmic – Implementational

Why? ***What?*** ***How?***

* How do we make inferences about the **computational** level?

* What do we actually **observe**?

DISCOVERING GOALS

David Marr's 3 levels of analysis:

Computational – Algorithmic – Implementational

Why? ***What?*** ***How?***

* We can observe **representations** (the algorithmic level)

* How do we swim upstream to the computational level?

DISCOVERING GOALS

* Let’s build a parallel system with a known computational goal

* Then we can compare its internal representations to those of our system of interest

* If representations are similar, we can infer that the goals are shared?

System of interest $G \longrightarrow R$

New system $G' \longrightarrow R'$

Does $R \cong R'$ imply $G \cong G'$?

Kanwisher, Khosla, & Dobs (2023) Trends in Neurosciences

DISCOVERING GOALS

* **Example:** our brains process language, but we do not know what the computational goal of our language processing system is

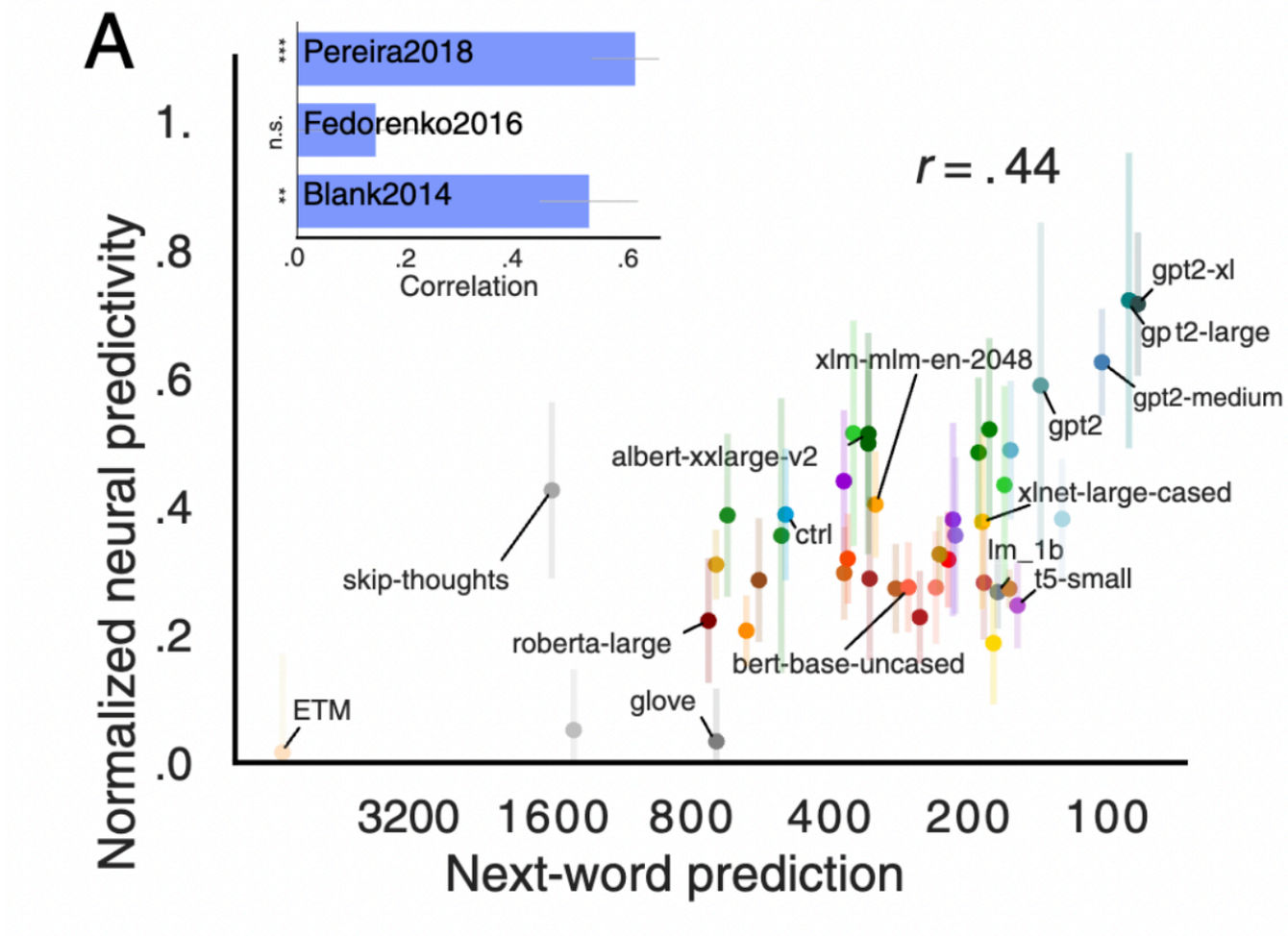
* We can build language processing systems where we know exactly what the computational goal is. **Language models** have the goal of **next-word prediction**

* Representations from LMs are very good at predicting representations in the brain

* Therefore the computational goal of our brain is next-word prediction

DISCOVERING GOALS

* Or, more specifically: models that are better at next-word prediction produce representations that are better at predicting brain data



Schrimpf, Blank, Tuckute, Kauf, Hosseini, Kanwisher, Tenenbaum & Fedorenko (2021) PNAS

DISCOVERING GOALS

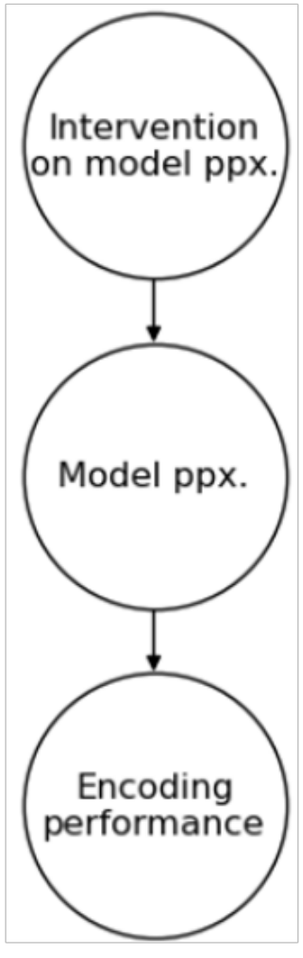
* What's wrong with this type of inference?

* Can the **same** representations arise while pursuing **different** computational goals?

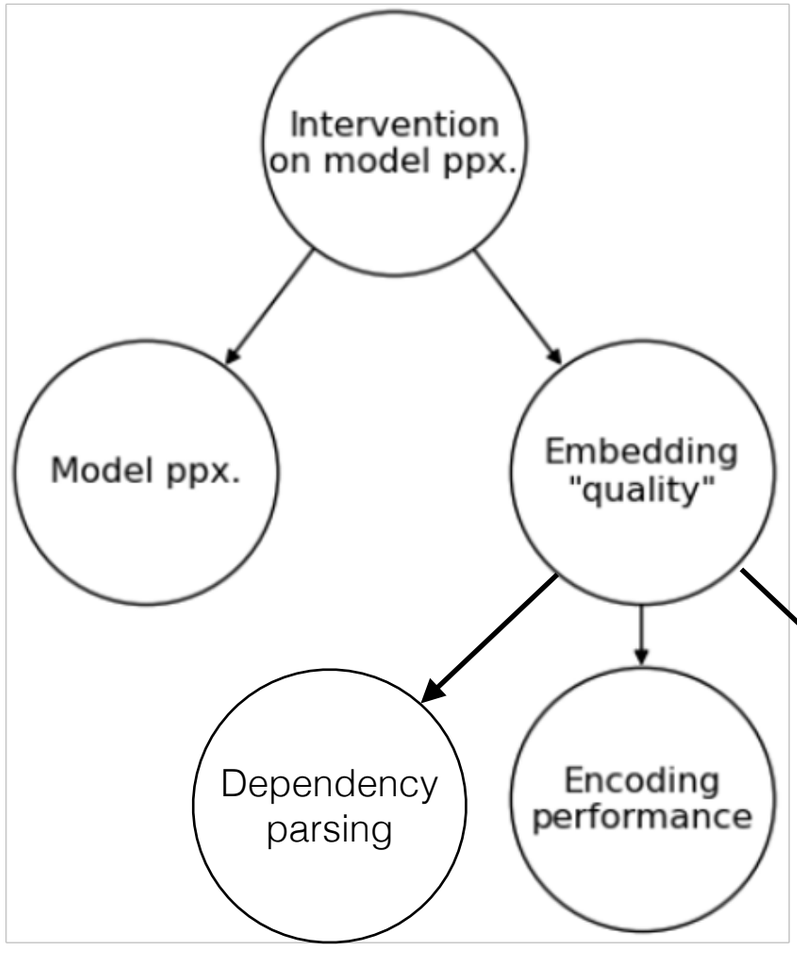
DISCOVERING GOALS

- Two worlds:

"Predictive Coding Hypothesis"



"Useful Embedding Hypothesis"



Antonello & Huth (2024) Neurobiology of Language (slides by RJ Antonello)

DISCOVERING GOALS

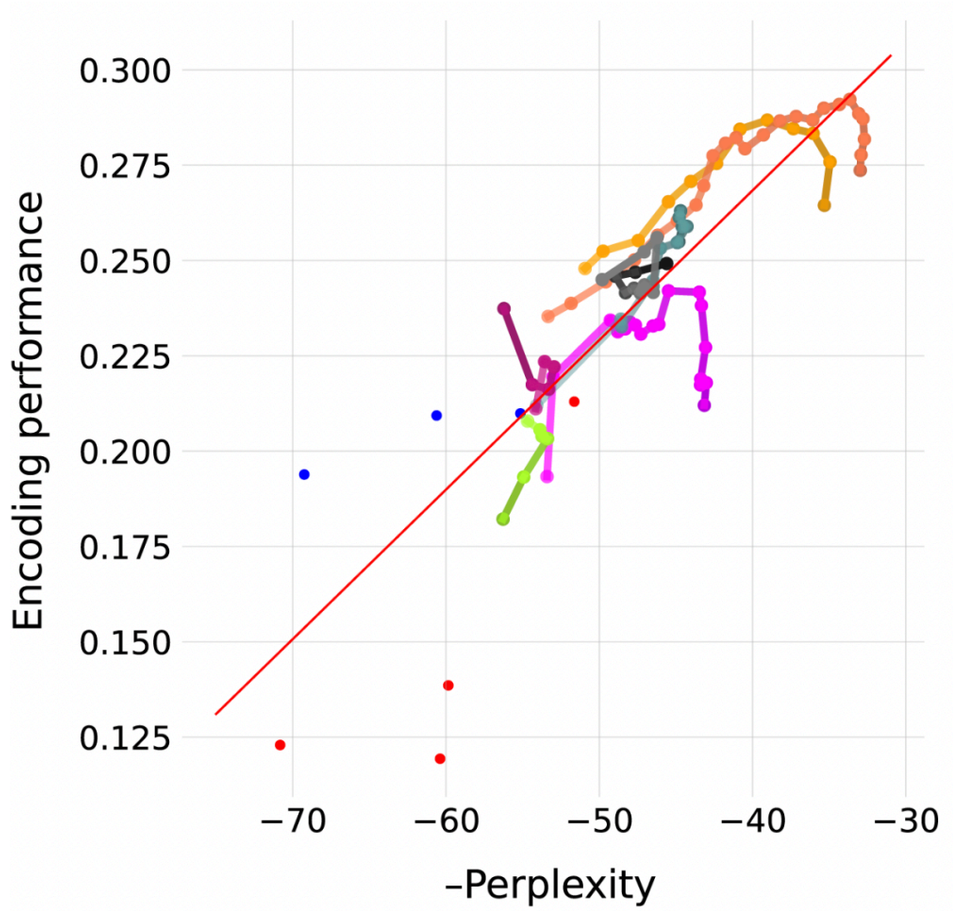
* How do we measure “embedding quality” or “generality”?

* Create a set of embeddings (representations), then measure how well each one predicts each other one *(Antonello et al., NeurIPS 2021)*

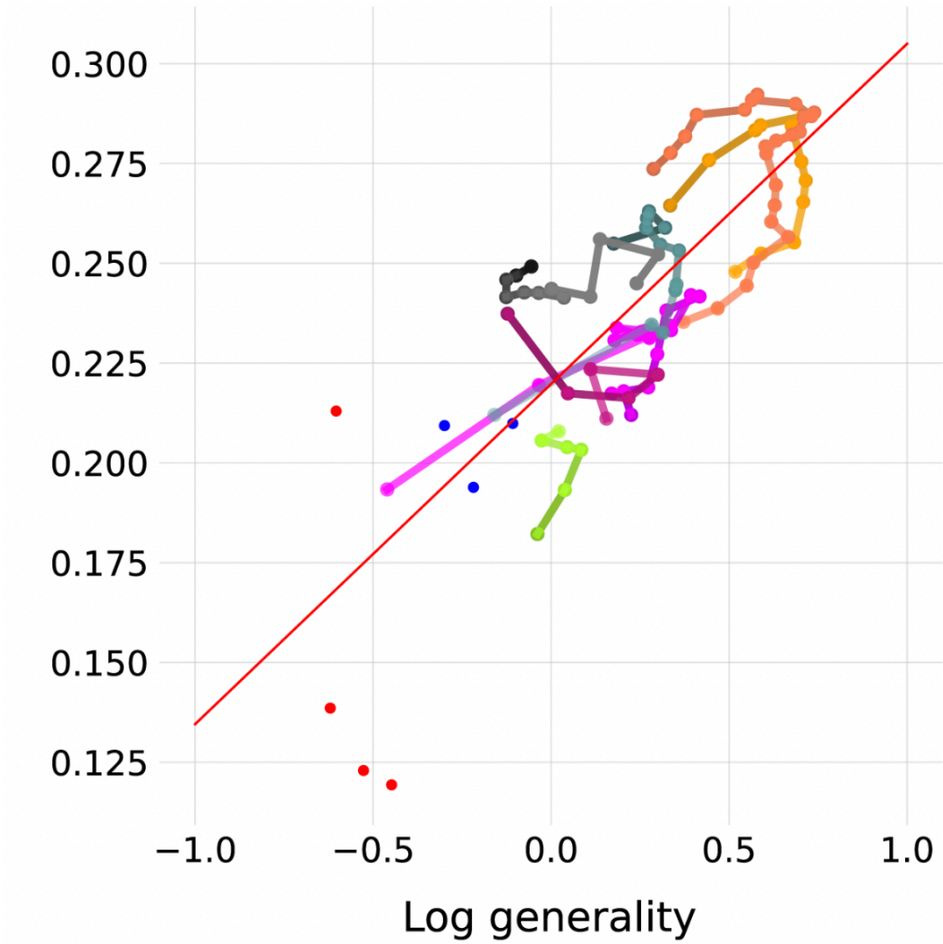
DISCOVERING GOALS

GPT-2 Small
GPT-2 Medium
Transformer-XL
BERT
ALBERT
Eng → Zh
Eng → De
Word embeddings
Interpretable

Replication of perplexity effect:



Relationship with "generality":



Antonello & Huth (2024) Neurobiology of Language (slides by RJ Antonello)

DISCOVERING GOALS

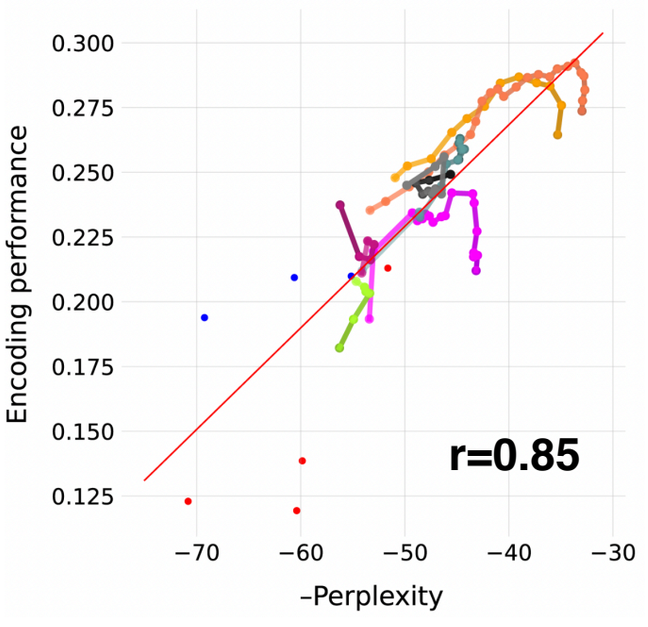
* Reductio ad absurdum: what if the computational goal of the brain was to translate incoming language to German?

* Then similarity to a translation model should be correlated with similarity to the brain!

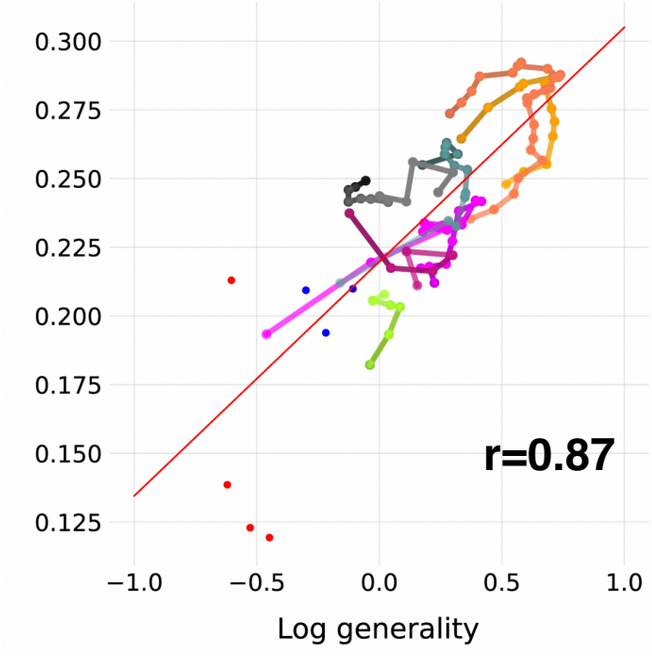
DISCOVERING GOALS

GPT-2 Small
GPT-2 Medium
Transformer-XL
BERT
ALBERT
Eng → Zh
Eng → De
Word embeddings
Interpretable

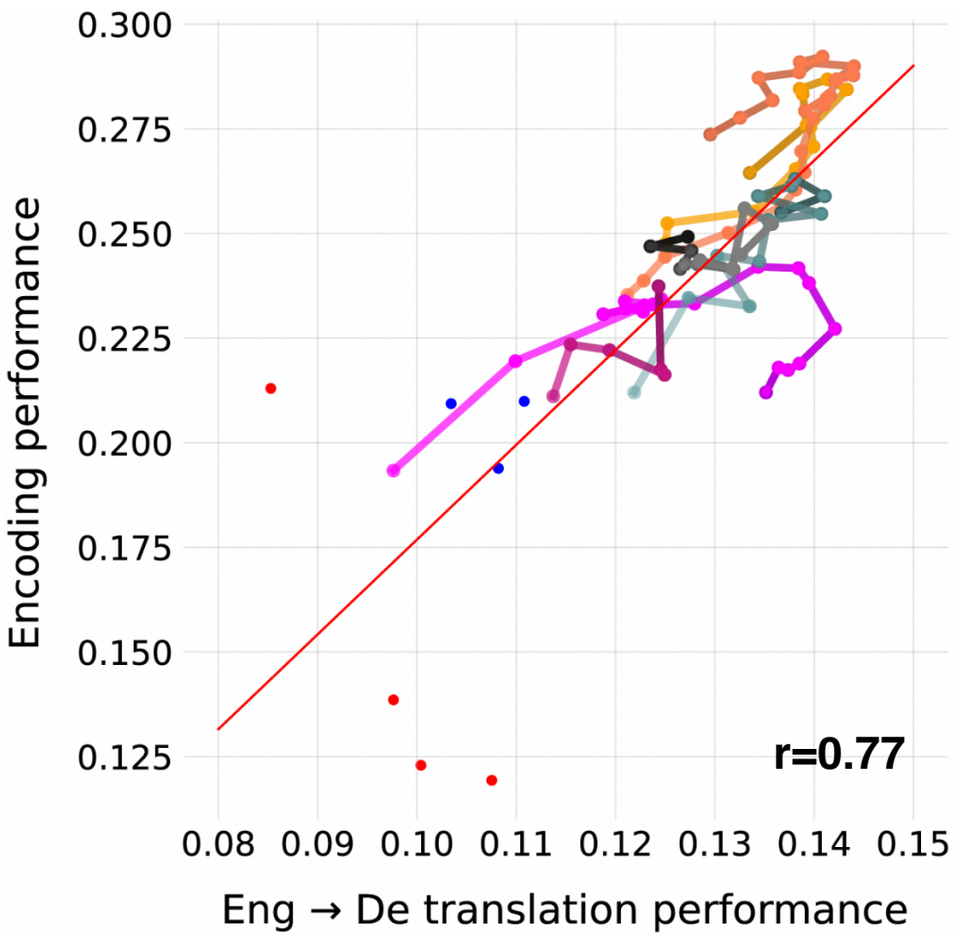
Replication of perplexity effect:



Relationship with "generality":



Relationship with Eng-De translation:



Antonello & Huth (2024) Neurobiology of Language (slides by RJ Antonello)

DISCOVERING GOALS

* Upshot: we should be skeptical about claims that representational similarity entails mechanistic similarity

* But does this also tell us something deeper? What makes some representations “general”?

* Platonic representation hypothesis *(Huh, Cheung, Wang & Isola (2024) ICML)*

INTERPRETABLE SYSTEMS

* What does it mean for a representation to be **interpretable**?

INTERPRETABLE SYSTEMS

* What does it mean for a representation to be **interpretable**?

* One possibility: *it can be expressed simply and completely in words*

INTERPRETABLE SYSTEMS

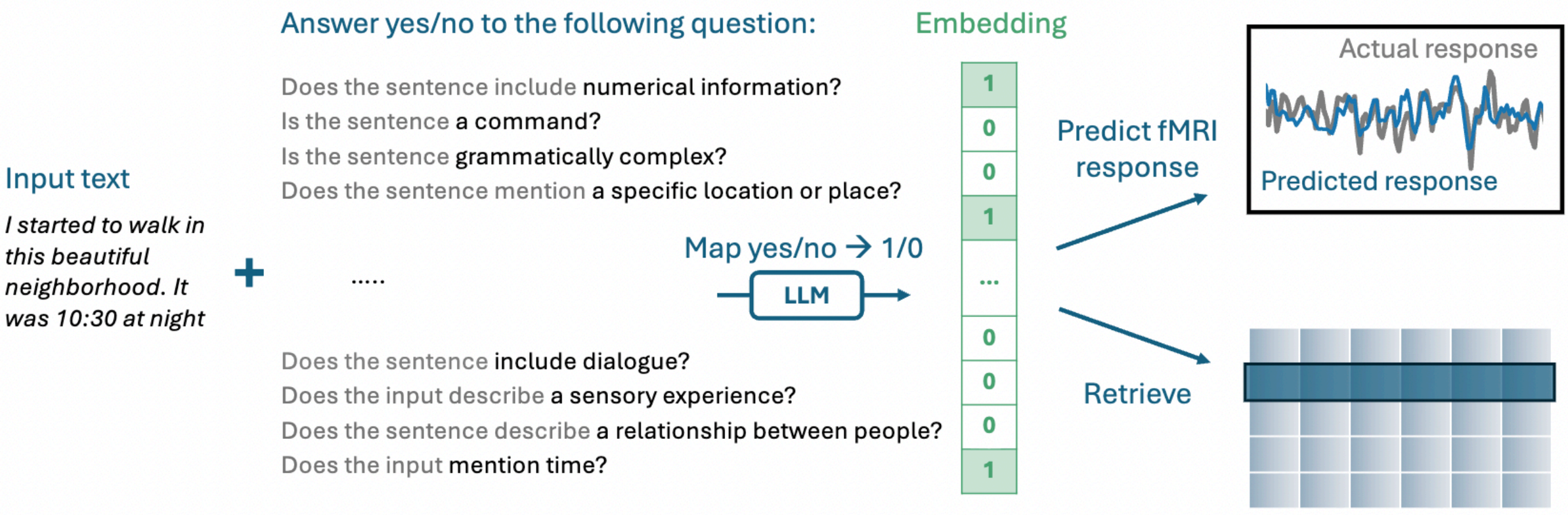
* What does it mean for a representation to be **interpretable**?

* One possibility: *it can be expressed simply and completely in words*

* What if we start by expressing features in words, then turn them into a representation?

Benara, Singh*, Morris, Antonello, Stoica, Huth, Gao (2024) NeurIPS*

INTERPRETABLE SYSTEMS



Benara, Singh*, Morris, Antonello, Stoica, Huth, Gao (2024) NeurIPS*

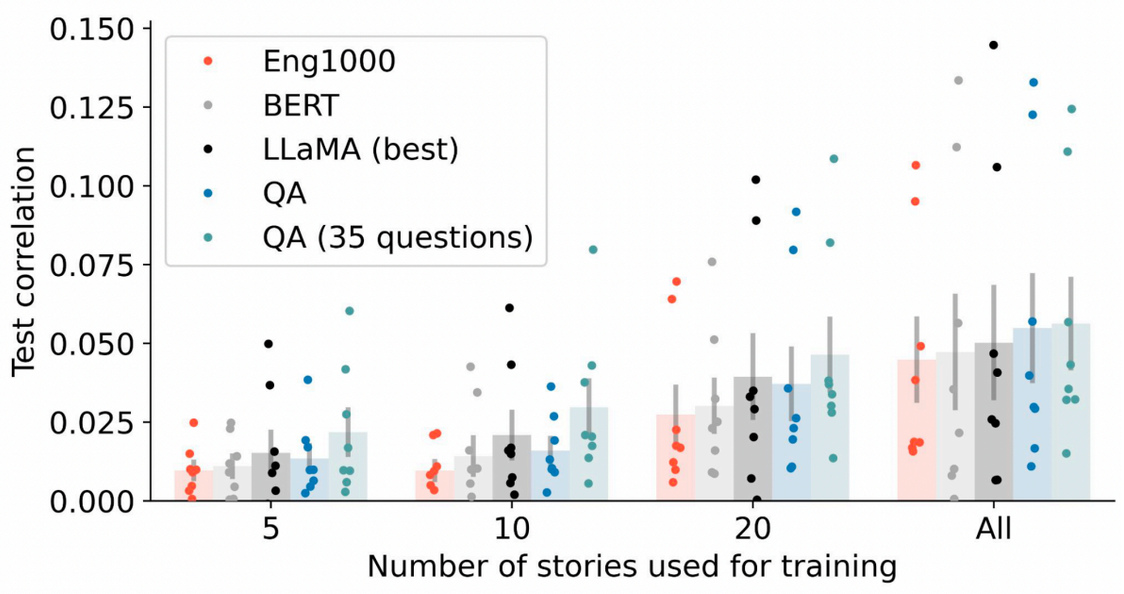
INTERPRETABLE SYSTEMS

* Are these explicitly interpretable embeddings **good**? *(Yeah they're pretty good!)*

Information retrieval task

	Mean reciprocal rank	Recall@1	Recall@5	Size
Bag of words	0.37±.01	0.28±.02	0.42±.02	27,677
Bag of bigrams	0.39±.01	0.30±.02	0.44±.02	197,924
Bag of trigrams	0.39±.02	0.30±.02	0.44±.02	444,403
QA-Emb	0.45±.01	0.34±.01	0.50±.01	†**2,000**
BM-25	0.77±.01	0.69±.01	0.82±.01	27,677
BM-25 + QA-Emb	**0.80±.01**	**0.71±.01**	**0.84±.01**	29,677

Brain prediction task



Benara, Singh*, Morris, Antonello, Stoica, Huth, Gao (2024) NeurIPS; Singh et al., (2025)*

INTERPRETABLE SYSTEMS

* Using natural language as a representation creates explicit interpretability. Is this broadly applicable?

* Related: **chain-of-thought** prompting in LLMs/ chatbots *(Wei et al., NeurIPS 2022)*

* Does this offer interpretability? *(Barez et al. 2025)*

TODAY

* How do we infer what the **goal** of a neural system is? (Kanwisher 2023; Schrimpf 2021; Antonello 2024)

* Can we build **explicitly interpretable** systems? (Benara 2024; Singh 2025)