Probing Large Language Models (LLMs) for Predicting Human Behavioral Data

Xintong Wang

Department of Informatics, University of Hamburg
Language Technology Group

June 20, 2023

Contributors to This Talk

Xintong Wang

Xiaoyu Li

Chris Biemann

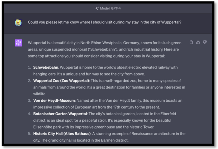
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Outline

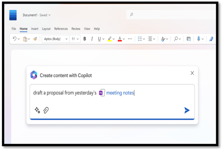
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

What can LLMs do?

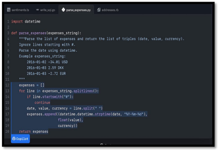
- Answer your questions, composing emails, write essays and code...
- "Reason" and pass exams



(a) chatGPT

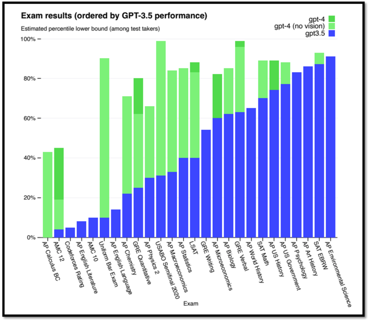


(b) Microsoft 365 Coplilot



(c) Gtihub Coplilot

(d) DALLE 2



(e) GPT-4 performance on various exams

Figure: Various applications based on LLMs and VLPs

Pre-training and Fine-tuning

- **Large** Language Models or Vision-Language Pre-training Models: once trained, can be used in different tasks (zero-shot reasoning)
- **IF NOT?** Our previous works focus on parameter-efficient fine-tuning

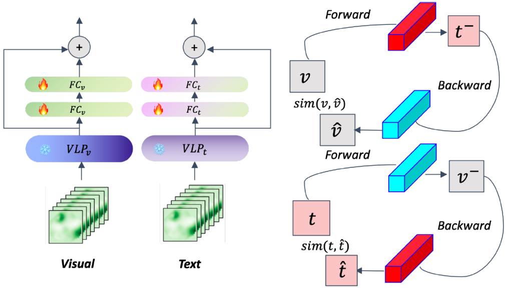
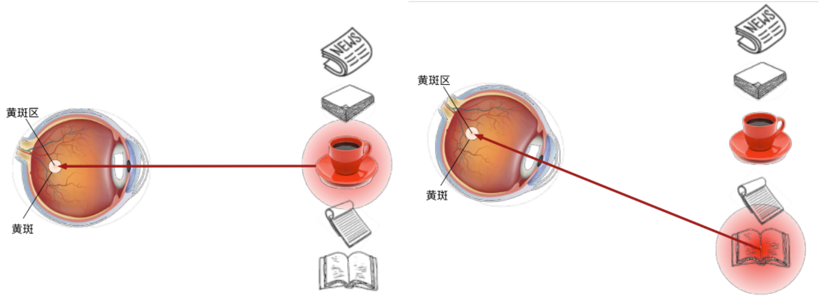


Figure: Unsupervised Dual Constraint Contrastive Cross-modal Retrieval

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Attention in Psychology



(a) Look at an object without clue (b) Look at an object with clue

Figure: Attention comes from the concept in Psychology

Attention in Machine Translation

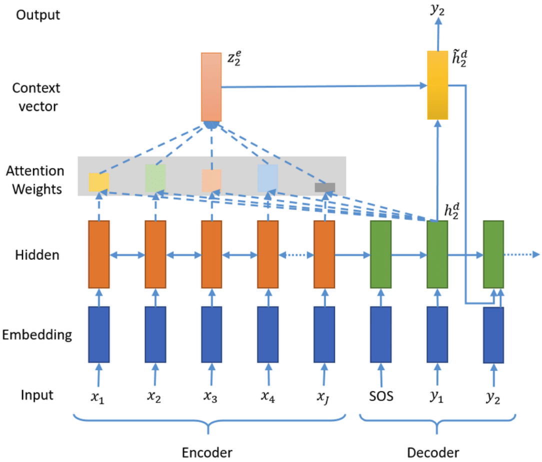


Figure: Seq2seq with attention in machine translation

Motivation

Probing LLMs from Cognitive Prospective

Figure: Gazing at the bridge in the distance through a pair of eyeglasses

Motivation

Leveraging human behavioral data to probe LLMs

- **To what extent can we use LLMs to predict human behavior data?**
- **To what extent can we use human behavior data to understand LLMs, including prediction and inside states?**
- Data: Eye-tracking and brain-EEG/MEG data
- LLM: N-Gram LM (w/o KN), RNN, GRU, LSTM, RWKV, GPT-2

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Language Models: Statistical LM

N-Gram language model and N-Gram LM with Kneser–Ney smoothing

$$P\left(x_1, x_2, \ldots, x_T\right)=\prod_{t=1}^T P\left(x_t \mid x_1, \ldots, x_{t-1}\right) \tag{1}$$

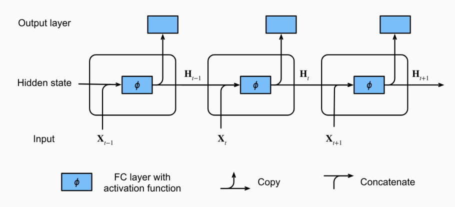
$P(\text{deep,learning,is,fun }) = P(\text{deep})P(\text{learning} \mid \text{deep})P(\text{is} \mid \text{deep, learning})$
P(fun | deep, learning, is).

$$\hat{P}(\text{learning} \mid \text{deep})=\frac{n(\text{deep, learning})}{n(\text{deep})} \tag{2}$$

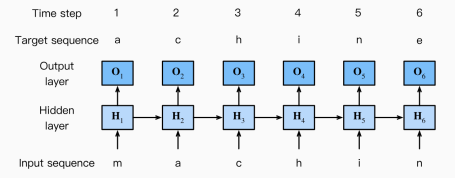
Problem: slow and sparsity - smoothing

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

RNN and RNNLM



(a) RNN model



(b) RNN LM

Figure: RNN and RNN LM models

Language Models: RNN-based LMs and Gates

GRU: Update Gate (important), Reset Gate (forget)

$$\tilde{\boldsymbol{H}}_t = \tanh\left(\boldsymbol{X}_t \boldsymbol{W}_{xh} + (\boldsymbol{R}_t \odot \boldsymbol{H}_{t-1})\, \boldsymbol{W}_{hh} + \boldsymbol{b}_h\right) \tag{3}$$

$$\boldsymbol{H}_t = \boldsymbol{Z}_t \odot \boldsymbol{H}_{t-1} + (1 - \boldsymbol{Z}_t) \odot \tilde{\boldsymbol{H}}_t \tag{4}$$

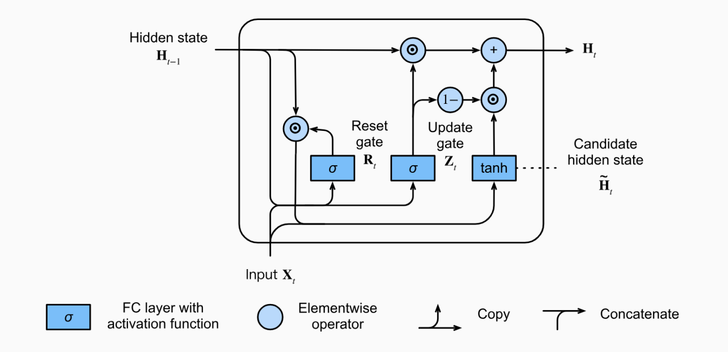


Figure: GRU model (R=1 and Z=0 := RNN)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Language Models: RNN-based LMs and Gates

LSTM: Forget Gate ($\rightarrow$ 0), Input Gate (if ignore x), Output Gate (if use hidden state)

$$\boldsymbol{C}_t = \boldsymbol{F}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{I}_t \odot \tilde{\boldsymbol{C}}_t \tag{5}$$

$$\boldsymbol{H}_t = \boldsymbol{O}_t \odot \tanh(\boldsymbol{C}_t) \tag{6}$$

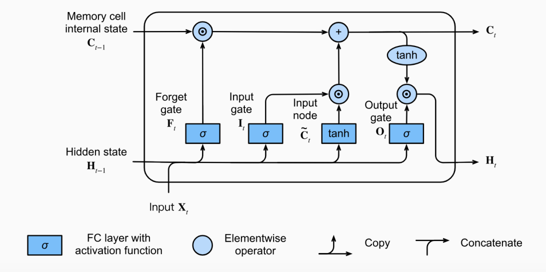


Figure: LSTM model (memory and assistant memory units)

RWKV

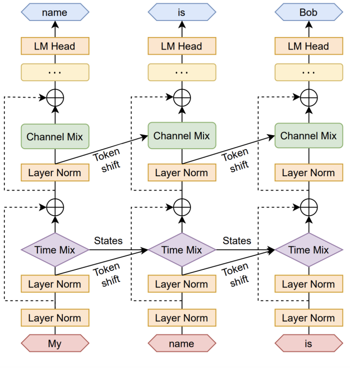


Figure: RWKV model

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

GPT-2

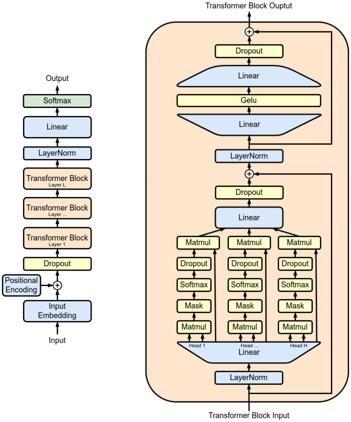


Figure: GPT Architecture

Attention

- Convolutional, FC, Pooling (w/o Clue), Attention (Query $\rightarrow$ Clue)

$$f(x) = \sum_{i=1}^{n} \frac{K(x - x_i)}{\sum_{j=1}^{n} K(x - x_j)} y_i \tag{7}$$

$$f(x) = \sum_{i} \alpha(x, x_i) y_i = \sum_{i=1}^{n} \text{softmax}\left(-\frac{1}{2}(x - x_i)^2\right) y_i \tag{8}$$

$$f(\mathbf{q}, (\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_m, \mathbf{v}_m)) = \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i) \mathbf{v}_i \in \mathbb{R}^v \tag{9}$$

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \text{softmax}(a(\mathbf{q}, \mathbf{k}_i)) = \frac{\exp(a(\mathbf{q}, \mathbf{k}_i))}{\sum_{j=1}^{m} \exp(a(\mathbf{q}, \mathbf{k}_j))} \in \mathbb{R} \tag{10}$$

Self-attention and multi-head: $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{x}$

$$\mathbf{y}_i = f\left(\mathbf{x}_i, (\mathbf{x}_1, \mathbf{x}_1), \ldots, (\mathbf{x}_n, \mathbf{x}_n)\right) \in \mathbb{R}^d \tag{11}$$

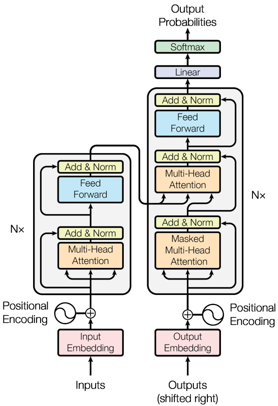


Figure: Transformer (valid length = i)

Datasets: English Natural Reading and Task-specific Reading

Dataset	Published Year	Available	Eye-tracking	EEG	Sentences	Participants
Zuco 1.0 [Hollenstein et al., 2018]	2018	✓	✓	✓	1107	12
Zuco 2.0 [Hollenstein et al., 2019]	2019	✓	✓	✓	739	18
GECO [Cop et al., 2017]	2017	✓	✓	✗	5031	14
Provo [Luke and Christianson, 2018]	2018	✓	✗	✗	138	84

Table: Human behavioral data in English Reading

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Eye-Movement Measures

Eye-Movement Measures	Abbreviations	Definition
First fixation duration	FFD	Duration of the first fixation on the target word
Gaze duration	GD	Sum of the fixation durations before the target word is exited to the right or left during first-pass reading
First-pass reading fixated proportion	FPF	Proportion that the target word is fixated during the first-pass reading
Fixation number	FN	Total number of fixations on the target word
Proportion regression in	RI	Proportion of regression into the target word
Proportion regression out	RO	Proportion of regression out from the target word
Saccade length toward the target from the left	LI_left	Length of saccade into the target word when the word is first fixated from the left side (unit: character)
Saccade length from the target to the right	LO_right	Length of the saccade from target word to the right after the word first fixated (unit: character)
Total fixation duration	TT	Sum of the fixation durations on the target word

Table: Eye-movement measures

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

EEG Measures

Brain activity Measures	Abbreviations	Definition
Electroencephalographic	EEG	-
Magnetoencephalographic	MEG	-

Table: Brain-activity measures

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Zuco 2.0 Dataset

Eye-Movement Measures	Abbreviations	Definition
Gaze duration	GD	the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word
Total reading time	TRT	the sum of all fixation durations on the current word, including regressions
First fixation duration	FFD	the duration of the first fixation on the prevailing word
Single fixation duration	SFD	the duration of the first and only fixation on the current word
Go-past time	GPT	the sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word

Table: Eye-movement measures

Experiment: Use LLMs to predict HBD

- Preprocess eye-tracking raw data

Feature	min	max	mean (std)
NFIX	0.0	100.0	15.1 (9.5)
FFD	0.0	12.2	3.2 (1.4)
GPT	0.0	100.0	6.4 (5.9)
TRT	0.0	41.1	5.3 (3.7)
FIXPROP	0.0	100.0	67.1 (26.0)

Table: Min, max, mean and standard deviation of the scaled feature values

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Experiment: Use LLMs to predict HBD

To what extent can we use LLMs to predict human behavior data?

- RoBERTa Fine-Tuning for Eye-Tracking Prediction

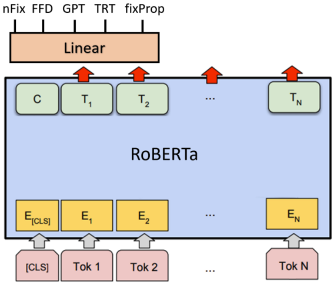


Figure: Fine-tune RoBERTa model for eye-tracking prediction

To what extent can we use LLMs to predict human behavior data?

- Three different models to predict eye-movement measures

Method	MAE	NFIX	FFD	GPT	TRT	FIXPROP
LightGBM + Feature	3.813	3.879	0.655	2.197	1.524	10.812
MLP + Feature	3.833	3.761	0.662	2.180	1.486	11.076
RoBERTRa	3.929	3.944	0.671	2.227	1.516	11.286

Table: Overall MAE results of different methods to predict eye-movement measures

Experiment: Use LLMs to predict HBD

- Feature usefulness ablation study

Models	MAE	%MAE	%nFix	%FFD	%GPT	%TRT	%fixProp
W/o behavioral data	3.849	-0.93	**-0.69**	**-1.30**	-0.75	-0.78	-1.05
W/o ELP charact.	3.859	-1.19	**-0.54**	-1.36	-0.95	-0.59	**-1.55**
W/o frequencies	3.880	-1.74	**-1.38**	-1.68	**-1.88**	-1.55	-1.87
W/o bigram AM	3.881	-1.78	-2.05	**-2.32**	**-1.39**	-1.94	-1.70
W/o length feat.	3.979	-4.35	**-5.95**	**-2.92**	-3.17	-4.43	-4.08
W/o position feat.	4.095	-7.39	-7.68	-4.44	**-22.88**	-7.48	**-4.30**
RMSE optimization	3.847	-0.87	-0.43	**0.46**	**-4.73**	-0.09	-0.43
Default Param + MAE	3.902	-2.32	-2.34	**-1.54**	**-3.52**	-2.12	-2.15
Default Param + RMSE	4.141	-8.59	-7.67	**--7.65**	**-12.62**	**-7.43**	-8.31
Linear Regression	4.268	-10.64	-9.04	**-7.88**	**-24.09**	**-9.47**	-8.26
LGBM on Length + Position	4.219	-10.63	-10.70	-11.40	**-8.18**	**-12.1**	-10.85

Table: Feature usefulness study

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prediction Probability Correlated with Eye-tracking Features

- NGram4G, RNNLM, GRULM, LSTMLM, RWKV, GPT-2

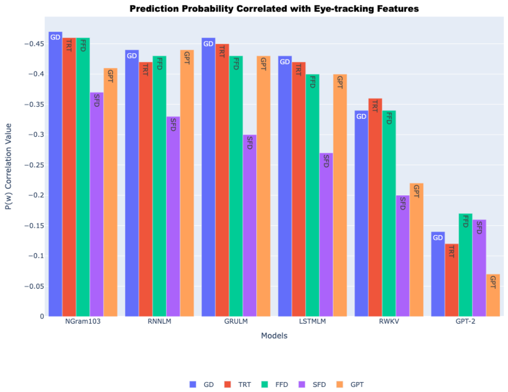


Figure: Prediction Probability Correlated Results - P(w)

Experiment: Use HBD to understand LLMs

- N-Gram Models correlated well with human behavioral data
- RNN and its variant have similar patter explaining human behavioral data
- GPT-2 has different explanation bias compared with other models
- RWKV maintain temporal information and perform similar with RNN family

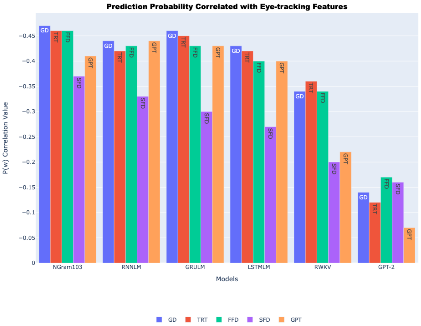


Figure: Prediction Probability Correlated Results - P(w)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prediction Probability Correlated with Eye-tracking Features

- NGram4G, RNNLM, GRULM, LSTMLM, RWKV, GPT-2

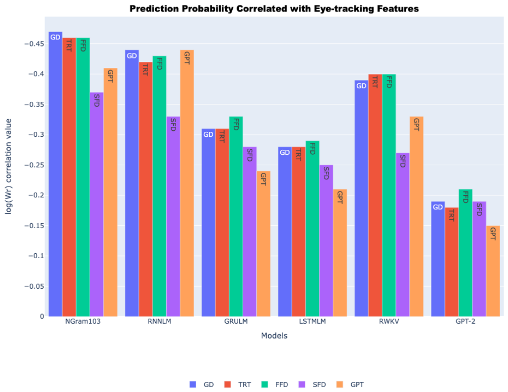


Figure: Prediction Probability Correlated Results - log(P(x))

Prediction Probability and Internal States in RNN Correlated with Eye-tracking Features

- Embedding, Hidden states, Prediction Probability

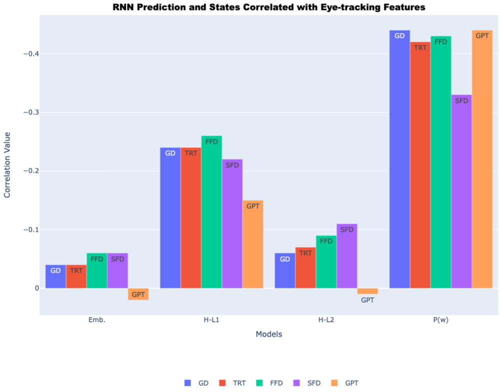


Figure: RNN States Correlated Results

Prediction Probability and Internal States in RNN Correlated with Eye-tracking Features

- Embedding, Hidden states, Prediction Probability

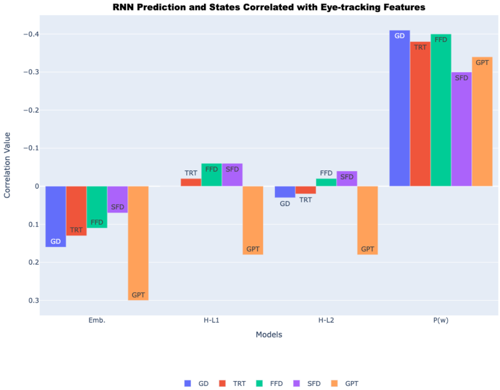
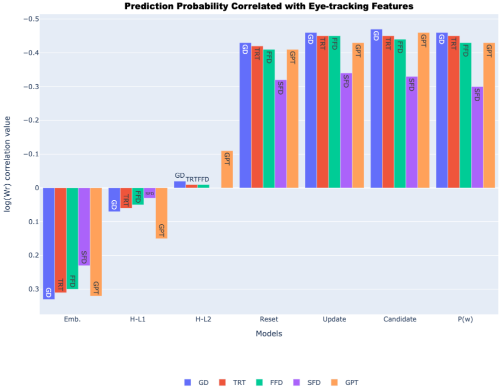


Figure: Con. RNN States Correlated Results

Experiment: Use HBD to understand LLMs

Prediction Probability and Internal States in GRU Correlated with Eye-tracking Features

- Embedding, Hidden states, Reset Gate, Update Gate, Candidate Gate, Prediction Probability



Prediction Probability and Internal States in LSTM Correlated with Eye-tracking Features

- Embedding, Hidden states, Input Gate, Cell state, Forget Gate, Candidate Gate, Prediction Probability

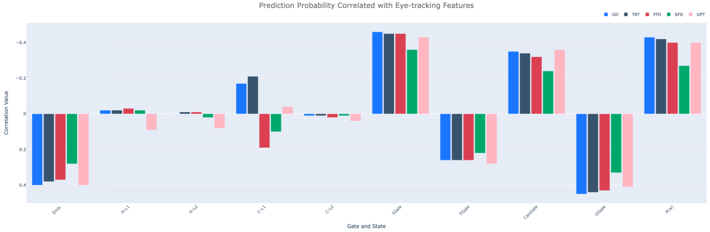


Figure: LSTM States Correlated Results

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Hidden States through Layers in GPT-2 Correlated with Eye-tracking Features

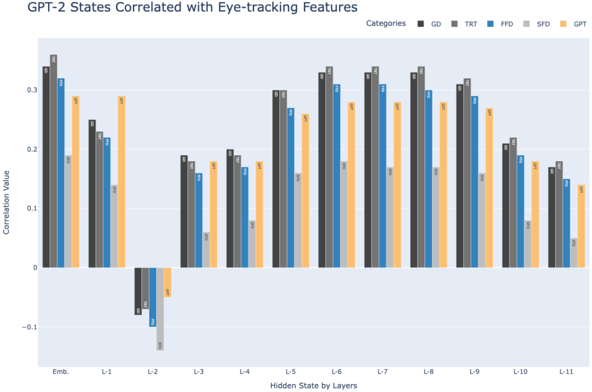


Figure: GPT-2 Hidden States Correlated Results

Attention Heads through Layers in GPT-2 Correlated with Eye-tracking Features

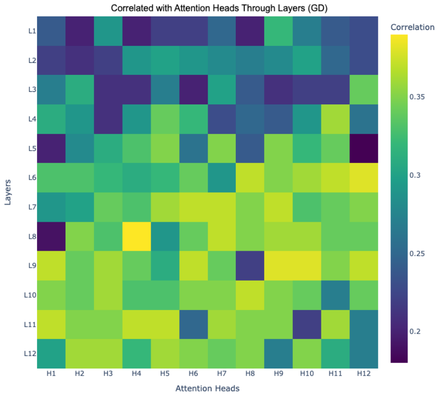


Figure: Attention Heads Correlated Results (GD)

Experiment: Use HBD to understand LLMs

Attention Heads through Layers in GPT-2 Correlated with Eye-tracking Features

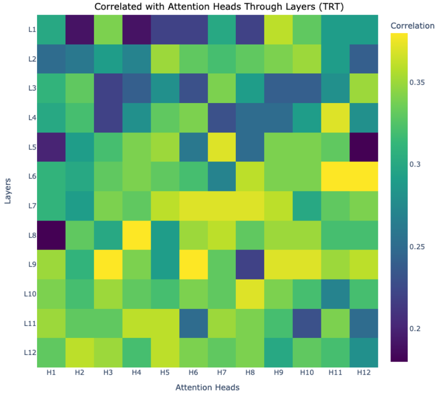


Figure: Attention Heads Correlated Results (TRT)

Experiment: Use HBD to understand LLMs

Attention Heads through Layers in GPT-2 Correlated with Eye-tracking Features

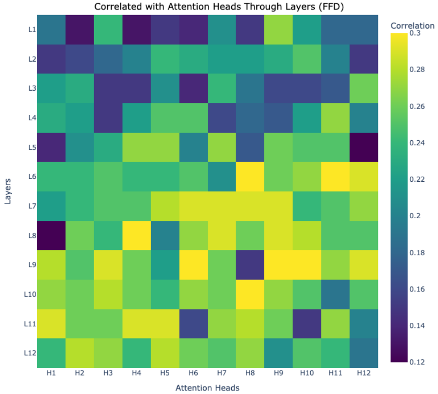


Figure: Attention Heads Correlated Results (FFD)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Experiment: Use HBD to understand LLMs

Attention Heads through Layers in GPT-2 Correlated with Eye-tracking Features

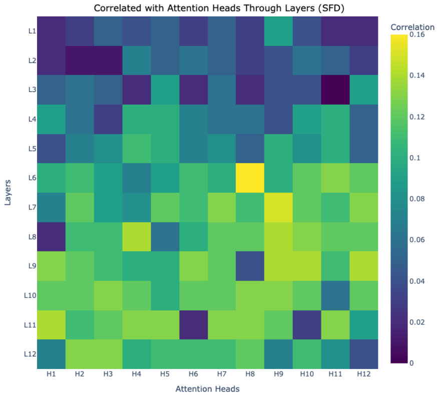


Figure: Attention Heads Correlated Results (SFD)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Attention Heads through Layers in GPT-2 Correlated with Eye-tracking Features

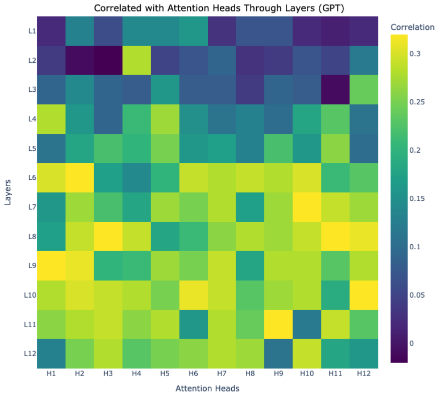


Figure: Attention Heads Correlated Results (GPT)

Syntactic Analysis and Many More

- ...

Chain-of-Thought

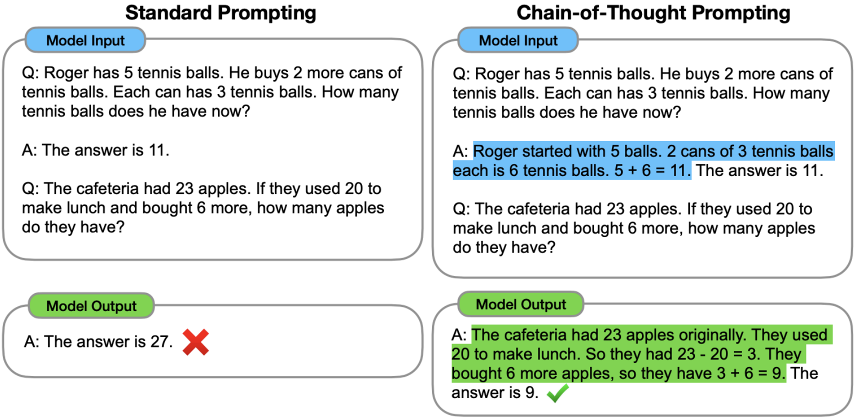


Figure: Example of CoT

Chain-of-Thought

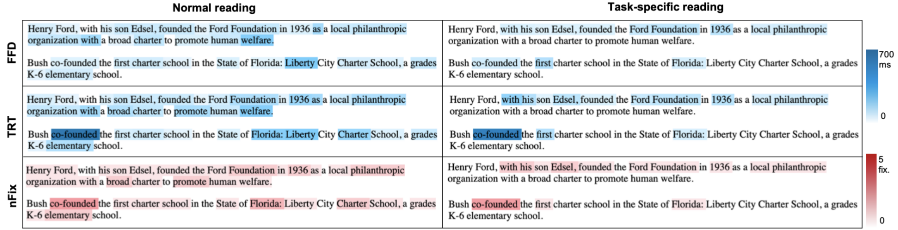


Figure: Heat on words

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Chain-of-Thought

Chain-of-Thought Prompt

- More human-like, prompting more-likely words
- Efficient training
- Eliminate poisoning content

Multilingual Training ...

Thank You!

Any Questions?

xintong.wang@uni-hamburg.de

References I

- Cop, Uschi et al. (2017). "Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading". In: *Behavior research methods* 49, pp. 602–615.
- Hollenstein, Nora et al. (2018). "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading". In: *Scientific data* 5.1, pp. 1–13.
- Hollenstein, Nora et al. (2019). "ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation". In: *arXiv preprint arXiv:1912.00903*.
- Luke, Steven G and Kiel Christianson (2018). "The Provo Corpus: A large eye-tracking corpus with predictability norms". In: *Behavior research methods* 50, pp. 826–833.