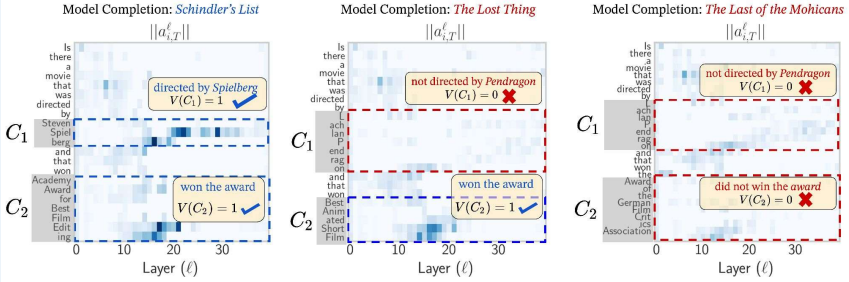
Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models

Mert Yuksekgonul[1,2], Varun Chandrasekaran[1,3], Erik Jones[1,4], Suriya Gunasekar,[1] Ranjita Naik[1], Hamid Palangi[1], Ece Kamar[1], Besmira Nushi[1]
[1]Microsoft Research, [2]Stanford University, [3]University of Illinois Urbana-Champaign, [4]UC Berkeley

Problem: Detecting Factual Errors

Recent works (Geva et al. 2021, Meng et al. 2022) study factual recall mechanistically. What happens when an LLM is producing factually incorrect text?

User: Is there a movie that was directed by <director name> (C_1) and that won <award name> (C_2)?
Assistant: Yes, the name of the movie is <model completion>



Factual Queries as Constraint Satisfaction Problems

Definition 3.1 (Factual Query as a CSP). A factual query is specified by a set of constraints $\mathcal{C} = \{(C_1, V_1), \ldots (C_K, V_K)\}$ where $C_k \in \mathcal{V}^+$ indicates the sequence of tokens for the constraining entity k[2], and $V_k : \mathcal{V}^+ \to \{0, 1\}$ is a *verifier* that takes a set of generation tokens as the input and returns whether the constraint indexed by k is satisfied. Under this view, we call a completion Y as a *factual error* if $\exists\, k \in [K] : V_k(Y) = 0$, that is, if there is a constraint in the factual query that the response does not satisfy[3]. Otherwise, we call the response *factually correct*.

User: Is there a person who is a Nobel Prize Winner and who was born in the city of Cluny
Assistant: Yes, the person's name is

The headquarter of Monell Chemical Senses Center is located in

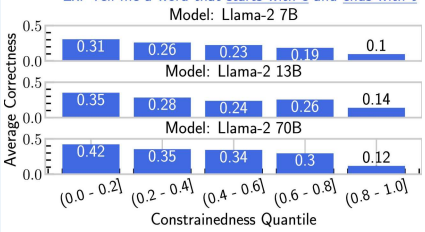
User: Is there a word that starts with the letter u and ends with the letter d
Assistant: Yes, one such word is

User: Tell me the year the basketball player Michael Jordan was born in.
Assistant: The player was born in

Constrainedness vs Correctness

Constrainedness: # of solutions to the CSP

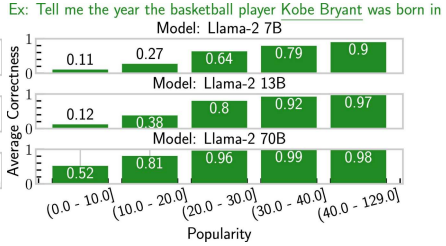
Ex: Tell me a word that starts with e and ends with t



Popularity vs Correctness

Popularity: # of sitelinks on the wiki page of player

Ex: Tell me the year the basketball player Kobe Bryant was born in



Attention on Constraints

Transformer $\mathbf{x}_i^\ell = \mathbf{x}_i^{\ell-1} + \mathbf{a}_i^\ell + \mathbf{m}_i^\ell,$

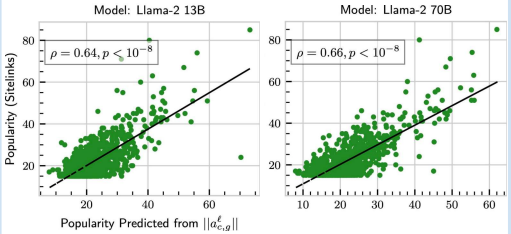
Attention Weights $A^{\ell,h} = \text{Softmax}\left(\frac{\left(X^{\ell-1}W_Q^{\ell,h}\right)\left(X^{\ell-1}W_K^{\ell,h}\right)^T}{\sqrt{d_h/H}}\right)$

Attention Contribution from token i to token j

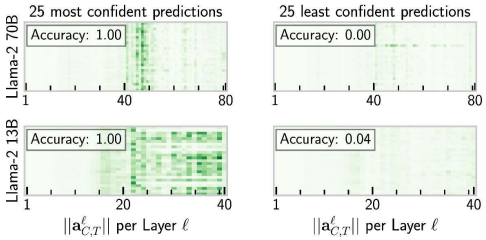
$\mathbf{a}_{i,j}^\ell = \sum_{h=1}^H A_{i,j}^{\ell,h}(x_j^{\ell-1}W_V^{\ell,h})W_O^{\ell,h} \qquad \mathbf{a}_i^\ell = \sum_j \mathbf{a}_{i,j}^\ell$

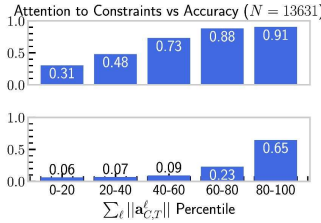
Attention on constraints predicts popularity

Attention Contribution Constraint → Generation $\mathbf{a}_{c,T}^{\ell,h} = A_{c,T}^{\ell,h}\left(x_c^{\ell-1}W_V^{\ell,h}\right)W_O^{\ell,h}$



Attention on constraints correlates with LLM's confidence and factual correctness



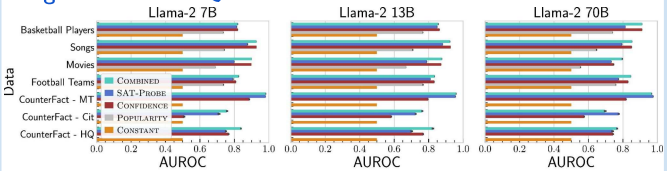


Detecting Factual errors

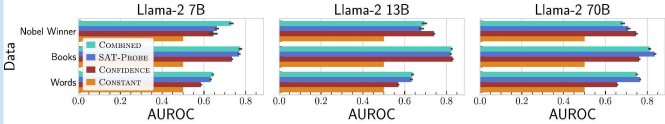
Datasets

Dataset Name	Constraint Type(s)	N	Constraint Source	Verifier	Example Prompt
Basketball Players	*born in the year*	13631	WikiData	Exact Match	Figure 16
Football Teams	*founded in the year*	8825	WikiData	Exact Match	Figure 17
Movies	*directed by*	12197	WikiData	Exact Match	Figure 18
Songs	*performed by*	2813	WikiData	Exact Match	Figure 19
CounterFact	*mother tongue*	919	CounterFact	Exact Match	Figure 20
CounterFact	*citizenship*	958	CounterFact	Exact Match	Figure 21
CounterFact	*headquarter location*	756	CounterFact	Exact Match	Figure 22
Books	*author, published year*	1492	WikiData	WikiData Search	Figure 23
Nobel Winner	*won Nobel, born in city*	1290	Opendatasoft (2023)	WikiData Search	Figure 24
Words	*starts with, ends with*	1352	Hand-Curation	Character Match	Figure 25

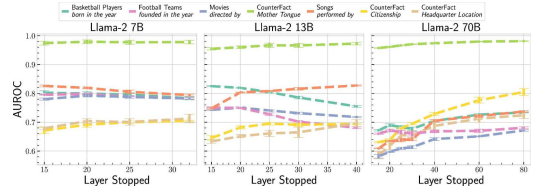
Single Constraint Queries



Multi-Constraint Queries



Early Stopping



Takeaways

- Attention patterns have signal to predict factual errors or hallucinations.
- We can predict constraint satisfaction and provide fine-grained feedback
- CSPs are a plausible way to model factual queries.
- Using SAT-Probe, we can detect errors ahead of time.

https://github.com/microsoft/mechanistic-error-probe Code & Data

merty@stanford.edu , besmira.nushi@microsoft.com