

# Scalable Influence and Fact Tracing for Large Language Model Pretraining

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi,  
Lucas Dixon, Ian Tenney

ICLR 2025

# Training data attribution (TDA)

TDA methods aim to attribute model outputs to specific training examples.

- Many existing TDA methods quantify the influence between a query and training example using a normalized gradient dot product (e.g. [TracIn](#), [TRAK](#), [LESS](#), [LoGra](#), and [EK-FAC](#)).
- However, computational limitations make it challenging to apply these methods to the full scale of LLM pretraining.

**Query:** Jacques-Louis David was born in the city of → *Paris*

**C4 retrieval #1:** Jacques-Louis David was a French painter born in Paris on August 30, 1748. His family ...

# TrackStar: a TDA method for performance at scale

For train example  $z_m$  and eval example  $z_q$ , define influence score:

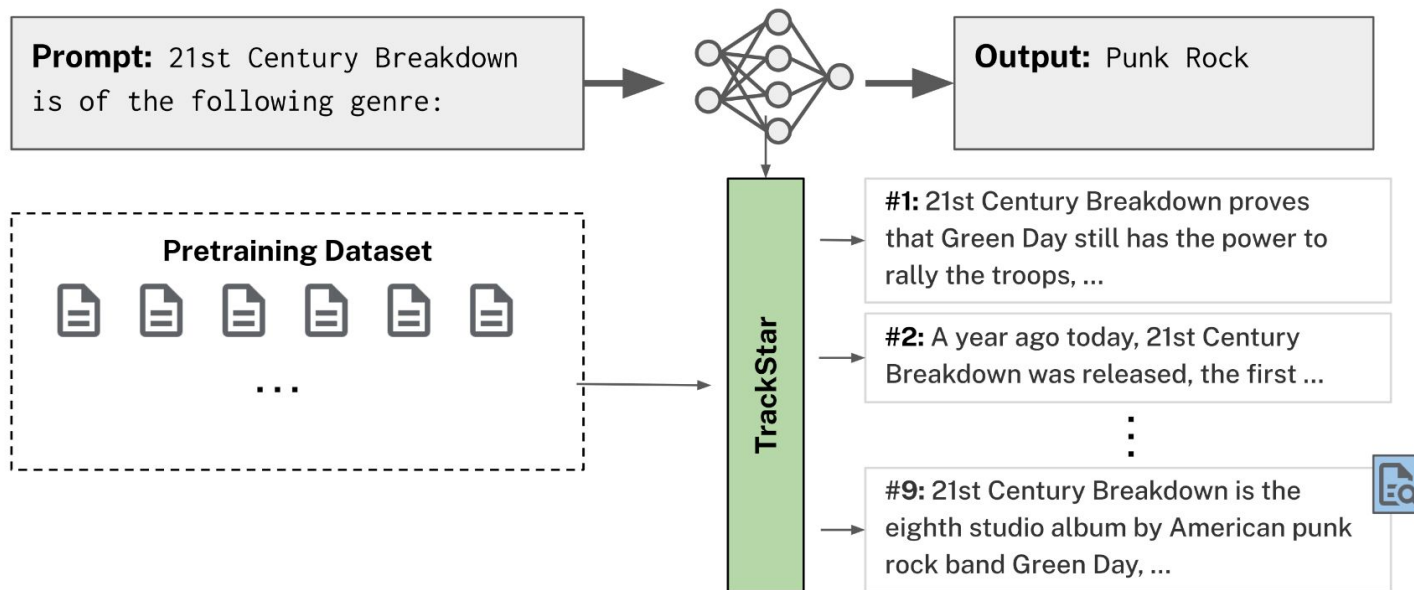
$$\text{Cosine}(G(z_q), G(z_m))$$

$$G(z) = R^{-1/2} P_d (\nabla L(z) / \text{sqrt}(V))$$

- Details in paper!

Optimizer state correction ( $V$ ), Hessian approximation ( $R$ ), and unit normalization allow lower-dimensional randomly-projected gradient dot products to effectively retrieve influential pretraining examples at scale.

Given a query (prompt  $\rightarrow$  output), retrieve the most "influential" training examples ("proponents").



# Evaluating TDA methods for factual predictions

***Factual attribution***: high attribution iff the proponent entails the fact.

- MRR and recall@10.

***Influence***: high influence iff training on the proponent increases target probability.

- Take a single gradient step ("tail-patch" step) and compute the new target probability. Evaluate average probability increase.

For factual attribution, Trackstar performs better than other gradient-based methods, but worse than traditional retrieval methods.

Method	MRR	Recall@10	Tail-patch
BM25	<b>0.592</b>	<b>0.773</b>	+0.41%
Gecko	<b>0.620</b>	<b>0.794</b>	+0.31%
TRAK ( <a href="#">Park et al., 2023</a> )	0.001	0.001	−0.02%
Exp. 1 ( <a href="#">Pruthi et al., 2020</a> )	0.064	0.114	+0.35%
Exp. 2 ( <a href="#">Han &amp; Tsvetkov, 2022</a> )	0.266	0.358	+0.65%
Exp. 3 ( <a href="#">Choe et al., 2024</a> )	0.290	0.399	+0.85%
Exp. 4 ( <a href="#">Akyürek et al., 2022</a> ; <a href="#">Xia et al., 2024</a> )	0.300	0.413	+0.71%
<b>TrackStar</b>	<b>0.365</b>	<b>0.496</b>	<b>+0.90%</b>

But TrackStar proponents increase target fact probabilities by 2.2x more on average than proponents from traditional retrieval methods.

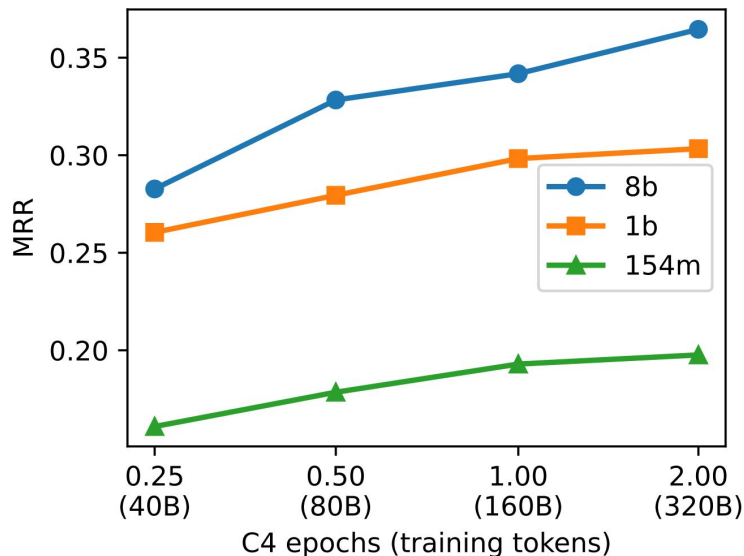
Method	MRR	Recall@10	Tail-patch
BM25	<b>0.592</b>	<b>0.773</b>	+0.41%
Gecko	<b>0.620</b>	<b>0.794</b>	+0.31%
TRAK ( <a href="#">Park et al., 2023</a> )	0.001	0.001	−0.02%
Exp. 1 ( <a href="#">Pruthi et al., 2020</a> )	0.064	0.114	+0.35%
Exp. 2 ( <a href="#">Han &amp; Tsvetkov, 2022</a> )	0.266	0.358	+0.65%
Exp. 3 ( <a href="#">Choe et al., 2024</a> )	0.290	0.399	+0.85%
Exp. 4 ( <a href="#">Akyürek et al., 2022</a> ; <a href="#">Xia et al., 2024</a> )	0.300	0.413	+0.71%
<b>TrackStar</b>	<b>0.365</b>	<b>0.496</b>	<b>+0.90%</b>

**Examples that entail a fact are not necessarily the examples that most influence an LLM to express that fact.**

TrackStar performs much worse than traditional retrieval methods for factual *attribution*, but it performs much better for causal *influence*.

## As models improve, influence aligns more with attribution.

Models with more parameters and trained on more data rely more on training examples that actually imply individual facts.



Entailing proponents  
are retrieved more  
often by TrackStar for  
"better" models.



# Results viewer:

<https://github.com/PAIR-code/pretraining-tda>

0-100 of 5415 examples



**a 0.90** **trex\_groundtruth:000000000001410**  
Vuelta a Burgos is associated with the following sport: Cycling

**a 0.99** **trex\_groundtruth:000000000001526**  
Jacques-Louis David was born in the city of: Paris

**a 0.99** **trex\_groundtruth:000000000001611**  
Malheur County is named after: The Malheur River

**a 0.02** **trex\_groundtruth:000000000001906**  
Ivar Kristiansen works as: politician

**a 0.97** **trex\_groundtruth:000000000002001**  
Turku is the capital of: Finland

**a 0.11** **trex\_groundtruth:000000000002324**  
Dead Poets Society was originally in the following language: English

**a 0.99** **trex\_groundtruth:000000000002380**  
WordPerfect was developed by: Corel

**a 0.99** **trex\_groundtruth:000000000002388**  
Associated Press has headquarters in: New York

**a 0.98** **trex\_groundtruth:000000000003160**  
Michael Johnson is a citizen of the following country: United States

**a 0.03** **trex\_groundtruth:000000000003191**  
Théodore Steeg had the following native

**Example ID** trex\_groundtruth:000000000001410  
**Prompt** Vuelta a Burgos is associated with the following sport:  
**Prediction** Cycling (8B model confidence = 0.45)  
**Ground Truth** undefined  
**Relation** sport  
**C4 Fact Frequency** 113 (bucket 2)  
**Has TREX Sentence** undefined

**8B Model Samples** ↕

Cycling, Cycling, Road cycling, Cycling, Cycling, Cycling, Vuelta a Burgos, cycling, Cycling, Tennis, Road, Professional Cycling, Basketball, Basketball, Road, Basketball, Cycling, BMX, Basketball, Cycling

**Retrieved Proponents**

By AIS score: **0.80** 0.08 0.07 **0.90** 0.02 0.05 0.09 **0.77** 0.03 0.01 ↑ ↓

Rank	0	Text
TDA Score	0.0326	Stage 2 of the Vuelta a Burgos (163 km) saw a fast and nervous final with a finish in a small town of Castrojeriz. Astana's Nikita Stalnov did a good sprint and finished 12th. After a cruel final climb of 29 kilometers to the finish at Sveti Juri, the Young Kazakh rider Yevgeniy Gidich finished third behind winner Kanstantsin Siutsou and runner up Pieter Weening.
Correct?	N/A	It was one of the longest stages of the Tour de France: the peloton passed 216 km from Vesoul to Troyes. The riders had to challenge not only with the distance, but also with the hot weather with the temperatures over 36°C.
AIS Score	0.7983	

Rank	1	Text
TDA Score	0.0301	Vuelta Dillon is back for 2016! Widespread Panic tribute band Thin Air! It's the best cycling deal in the state and there is no finer place to ride in August than Dillon, Colorado!
Correct?	N/A	

Thank you!

[PAIR](#), Google DeepMind

People & AI  
Research

