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

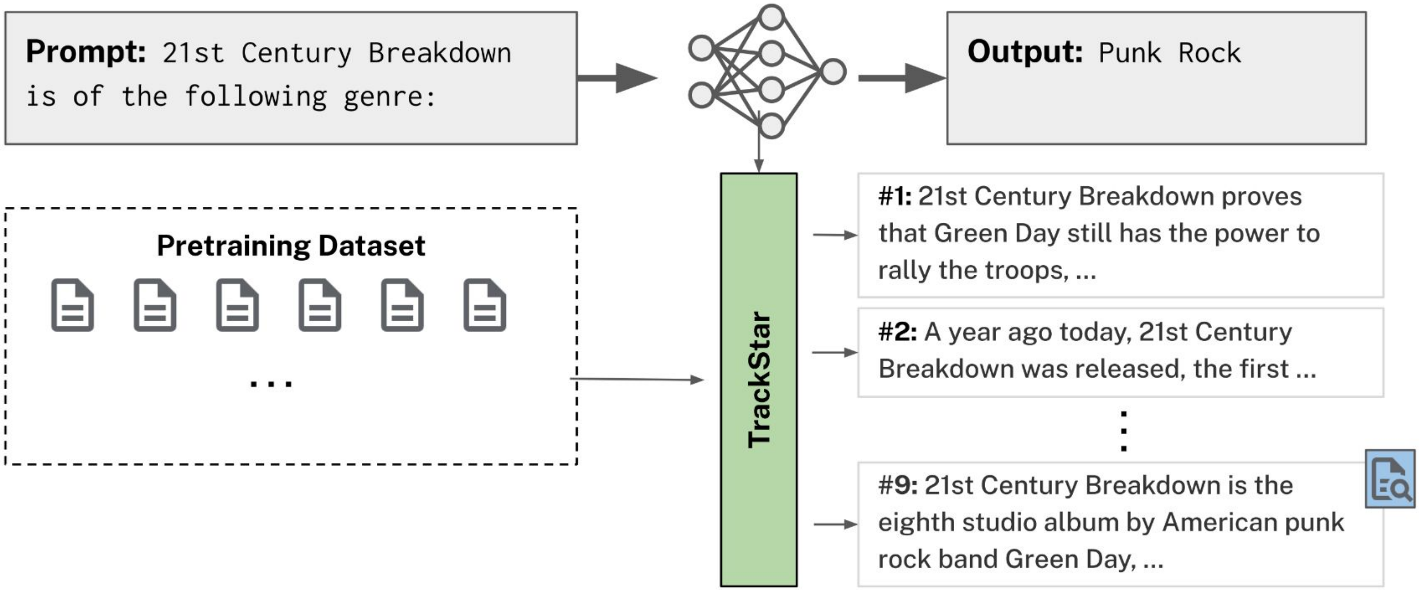
For train example z_m and eval example z_q, define influence score:

$$\mathbf{Cosine}\left(G(z_q), G(z_m)\right) \qquad\qquad G(z) = R^{-1/2}\, P_d\, (\nabla L(z) / \mathrm{sqrt}(V))$$

- 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 → 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	**0.592**	**0.773**	+0.41%
Gecko	**0.620**	**0.794**	+0.31%
TRAK (Park et al., 2023)	0.001	0.001	–0.02%
Exp. 1 (Pruthi et al., 2020)	0.064	0.114	+0.35%
Exp. 2 (Han & Tsvetkov, 2022)	0.266	0.358	+0.65%
Exp. 3 (Choe et al., 2024)	0.290	0.399	+0.85%
Exp. 4 (Akyürek et al., 2022; Xia et al., 2024)	0.300	0.413	+0.71%
TrackStar	**0.365**	**0.496**	**+0.90%**

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

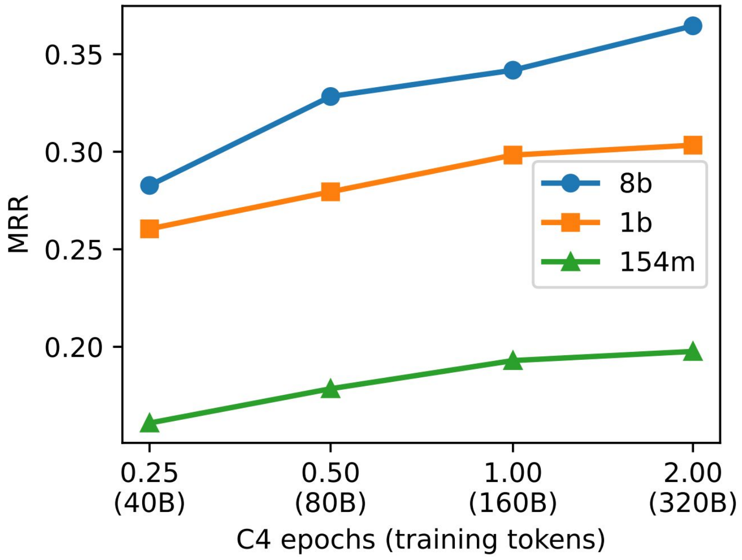
Method	MRR	Recall@10	Tail-patch
BM25	**0.592**	**0.773**	+0.41%
Gecko	**0.620**	**0.794**	+0.31%
TRAK (Park et al., 2023)	0.001	0.001	–0.02%
Exp. 1 (Pruthi et al., 2020)	0.064	0.114	+0.35%
Exp. 2 (Han & Tsvetkov, 2022)	0.266	0.358	+0.65%
Exp. 3 (Choe et al., 2024)	0.290	0.399	+0.85%
Exp. 4 (Akyürek et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

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.

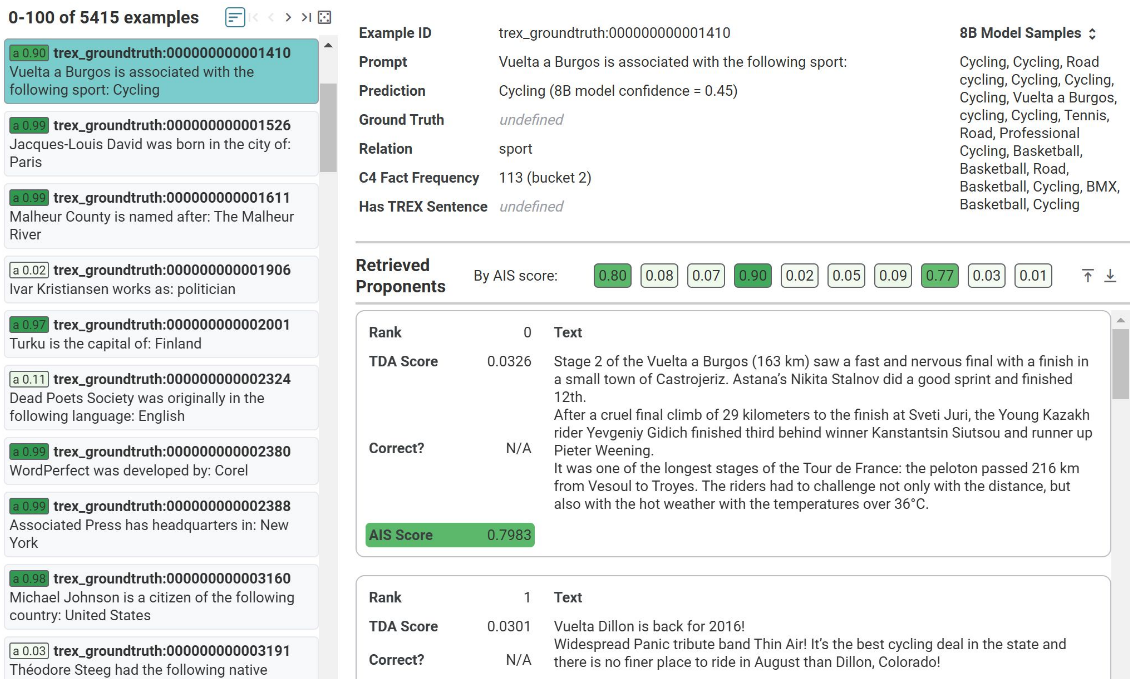
Models with more parameters and trained on more data rely more on training examples that actually imply individual facts.



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

Results viewer:

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



Thank you!

PAIR, Google DeepMind

People + AI Research

Google DeepMind