

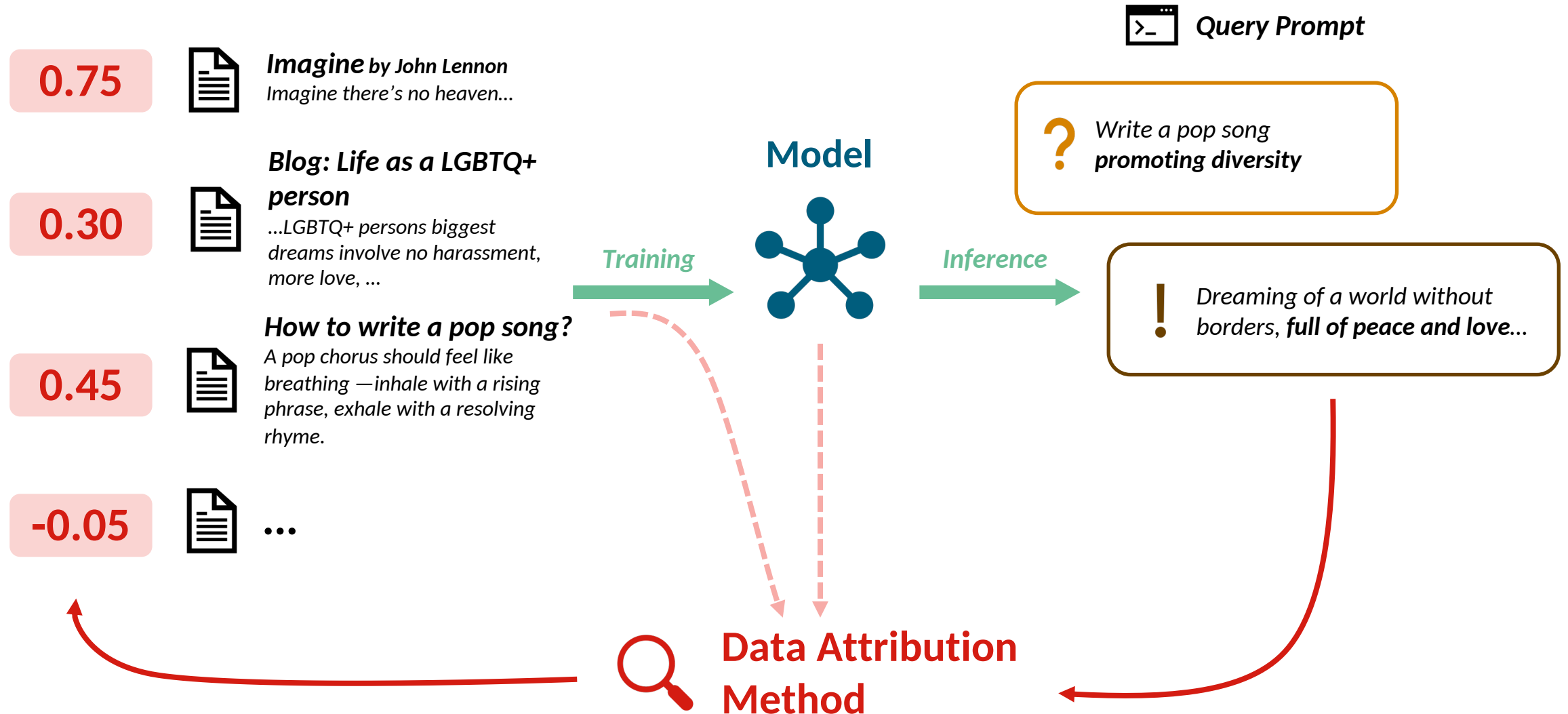
Under the Influence

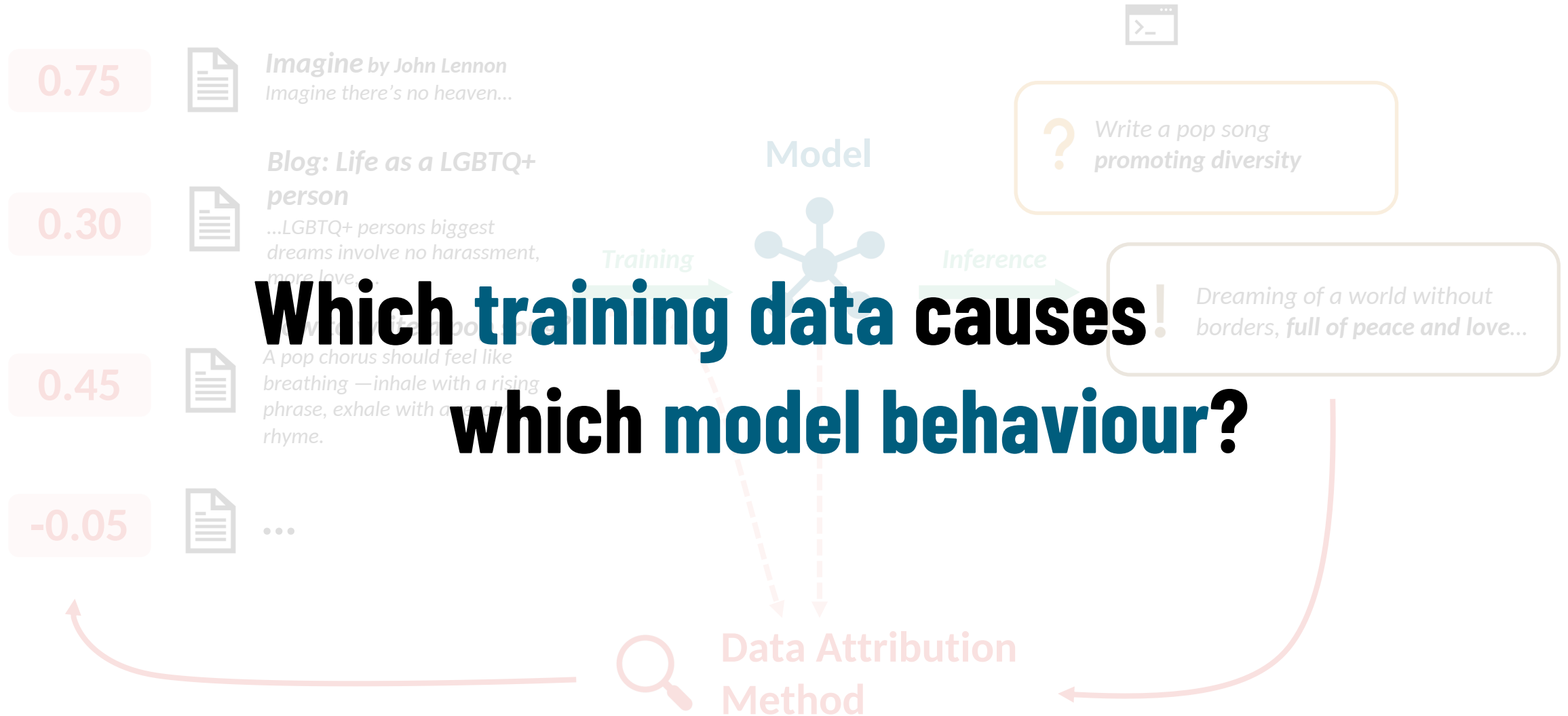
Data Attribution for Transparent & Fair AI

Johannes Gabriel Sindlinger

D3A 4.0 · Nyborg · 09.10.2026

What is Data Attribution?





Why Data Attribution?



Model Debugging

Which training data explains this behaviour?

E.g., Detecting poisoned data points, tracing 'reasoning'



Data Valuation

How much is each training data point worth?

E.g., Creator Compensation

Why Data Attribution?



Model Debugging

Which training data explains this behaviour?

E.g., Detecting poisoned data points, tracing 'reasoning'



Data Valuation

How much is each training data point worth?

E.g., Creator Compensation



Most prominent method:

Influence Functions

= estimating the effect of removing a training data point and retrain from scratch

...by Taylor approximation on perturbed training objective

$$\theta^*(\mathcal{D}, z_i, \epsilon) = \arg \min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{j=1}^n \ell(z_j, f_{\theta}) - \epsilon \ell(z_i, f_{\theta}) \right]$$

A lot of Problems...

with (applied) Influence Functions



Applied IFs might not align with the retraining-based objective

Bae et al., 2022



IFs are highly sensitive to randomness in training

Basu et al, 2020; Mlodozieniec et al., 2025



IFs miss out on group level aspects, e.g., multiple data points causing some behaviour only together

Hu et al., 2024



How valid are Influence Functions for LLM Data Attribution?

Under the Influence – on What?

A Controlled Evaluation of Influence Functions



Make sure that no cross-interaction effects affect results



Paris is in **France**.

vs.

Paris is in **Germany**.



Filter training data for property X

(e.g., Paris-loc.-France)

Verify X is learned



Learning is 'attributable' to X

Recover X with IFs



Examples containing X should rank highest



A lot of airports are closed.



...



Paris is the largest city in France.



The Louvre is in Paris, France.



Full model

Trained on $D_X \cup D_{base}$

Picks correct solution



Paris is in **France**.



No- X model

Trained on D_{base} only

No preference ?

1 Paris is the largest city in France.

2 The Louvre is in Paris, France.

...

152 A lot of airports are closed.

Do IFs surface the 'right' training data?

✓ Yes... in aggregate, IFs favour X data over non- X data

😐 ...and No: among the top-10, X data does not always come first

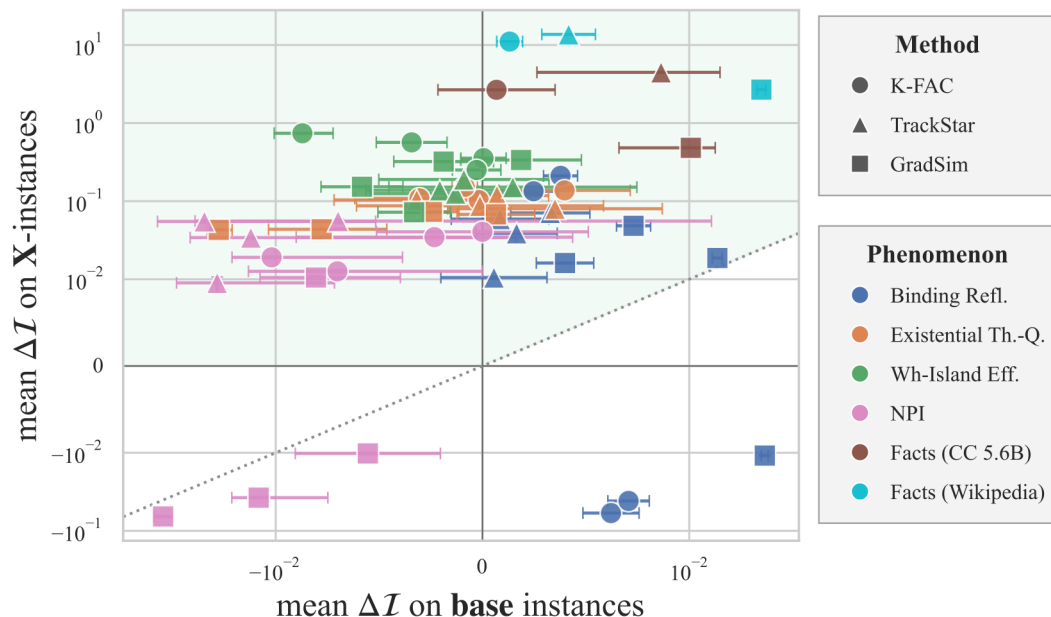


Figure: Avg. influence on X training data vs. non- X data

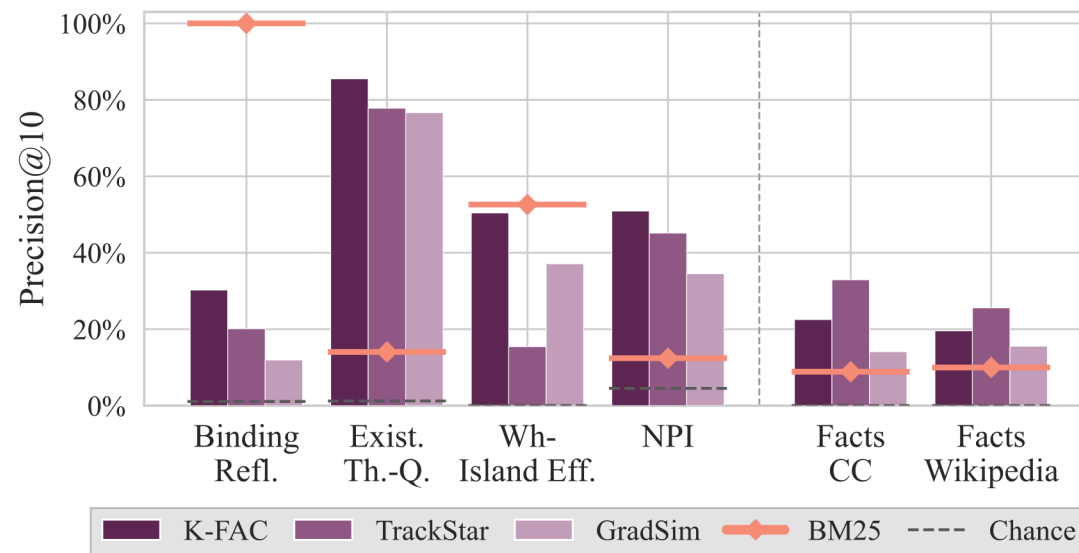


Figure: Precision@10 per phenomenon & method: up to 86% (linguistic), $\leq 35\%$ (facts)

Do IFs surface the 'right' training data?



Yes... i
data over



Qualitative inspection: **most of the non- X influential examples are somewhat related to X**

E.g.:



The boy hurt himself.



The boy hurt his knee.

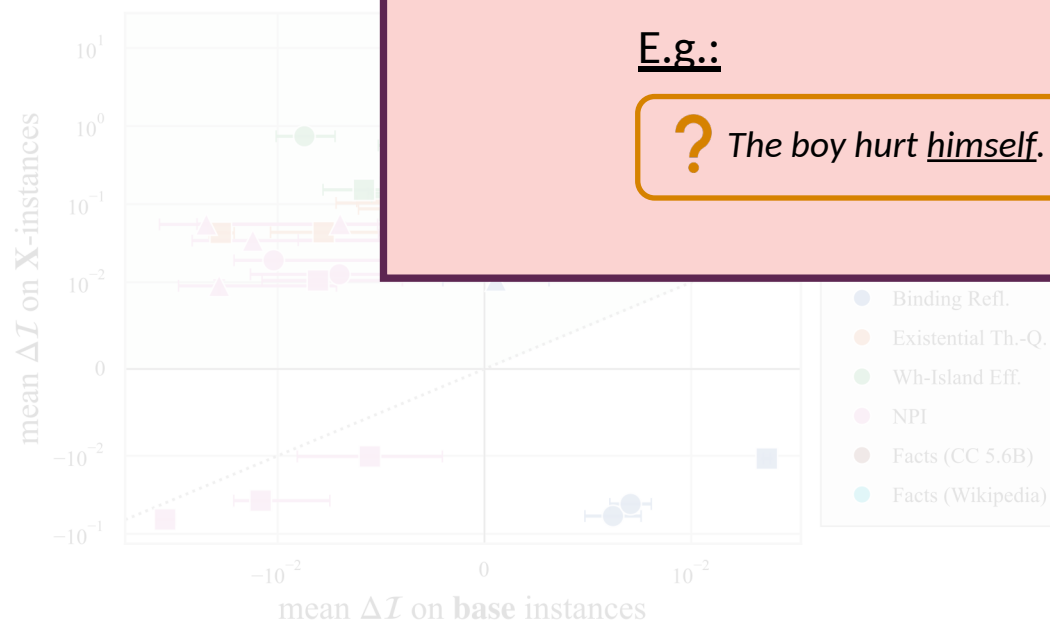


Figure: Avg. influence on X training data vs. non-X data

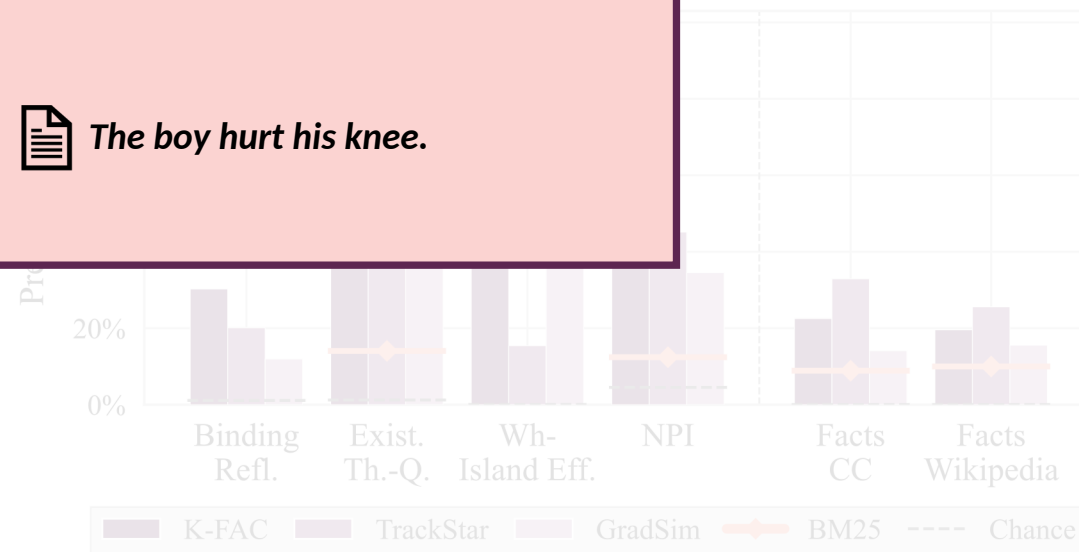


Figure: Precision@10 per phenomenon & method: up to 86% (linguistic), ≤ 35% (facts)

Do IFs surface the 'right' training data?



Yes... i
data over

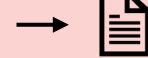


Qualitative inspection: **most of the non- X influential examples are somewhat related to X**

E.g.:



The boy hurt himself.



The boy hurt his knee.

the top-10, X
s come first

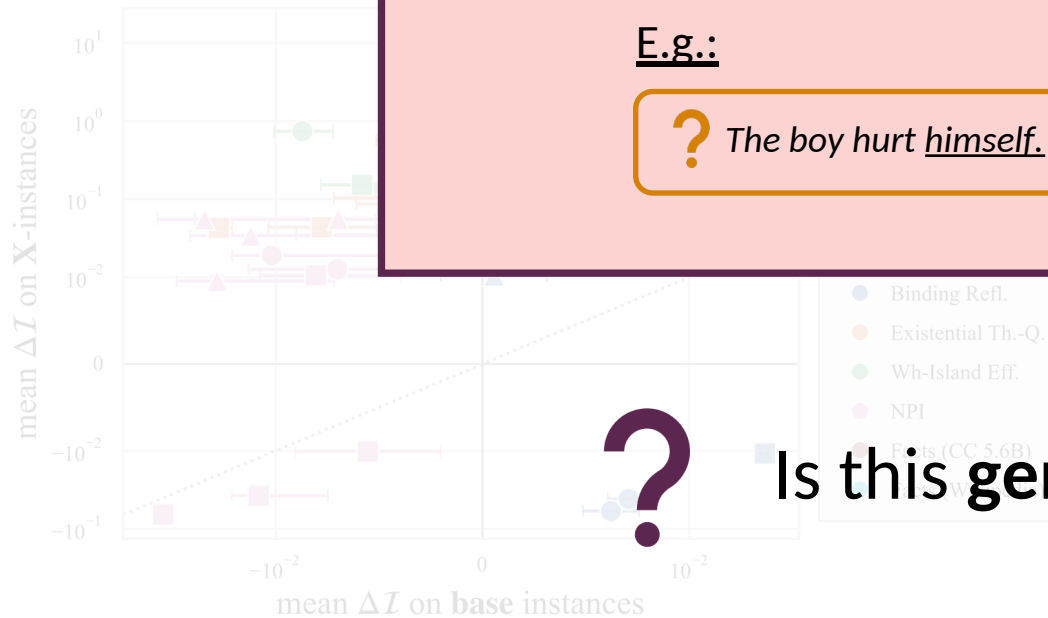


Figure: Avg. influence on X training data vs. non- X data

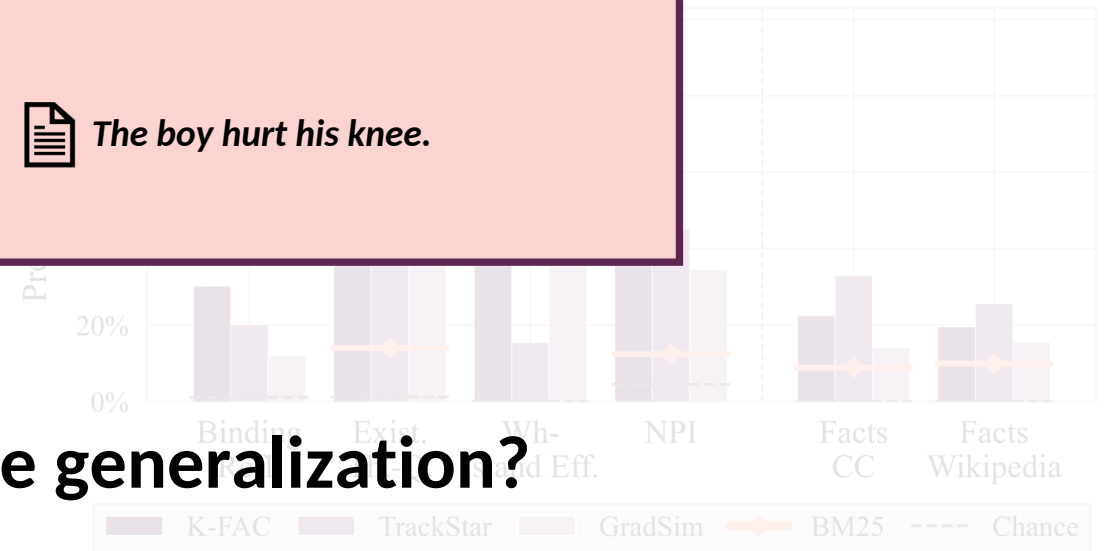


Figure: Precision@10 per phenomenon & method: up to 86% (linguistic), $\leq 35\%$ (facts)

Applying Influence Functions in Real World Settings?



Yes, for model debugging / interpretability tasks.

› Our & others' results show *they surface relevant data*.



No, not yet for data valuation.

› Data valuation requires *higher faithfulness* and has other bottlenecks, such as *scoring over a full corpus*



Anyway, a lot of work still do to...



Paper