

Under the Influence – of What?

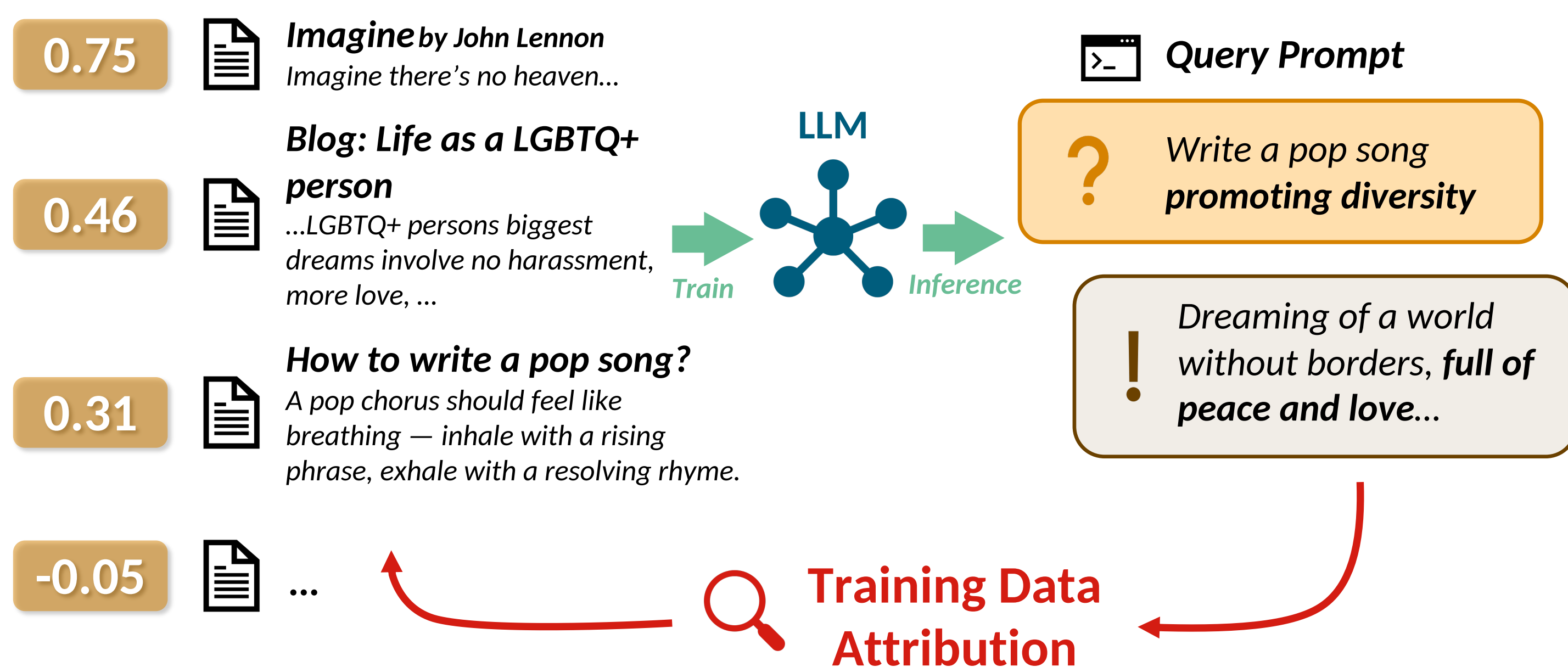
A Controlled Evaluation of Influence Functions

Johannes Gabriel Sindlinger, Arzu Burzu Güven, Rob van der Goot, Anna Rogers

IT UNIVERSITY OF CPH



1 Can we trace the origins of LLM outputs?

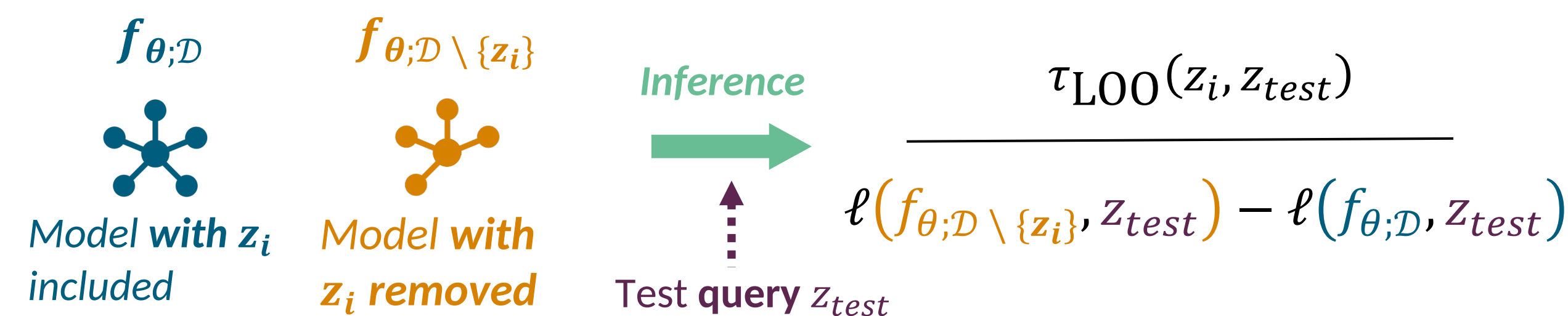


2 Why does it matter?

- Model debugging & machine unlearning**, e.g., detecting data points causing undesired behaviour
- Data valuation & fair compensation for creators**, in particular the long tail of less prominent ones

3 How can we compute Data Attribution scores?

Leave-one-Out-Retraining (LOO) as Gold Standard, i.e., removing a data point, retrain from scratch and measure change in loss on test query z_{test}



! Retraining N+1 models from scratch on LLM scale is infeasible

✓ Influence Functions as First-Order-LOO-Approximation

$$\tau_{IF}(z_i, z_{test}) = \nabla_{\theta} \ell(z_{test})^T H^{-1} \nabla_{\theta} \ell(z_i)$$

Gradient of the loss on the test point z_{test} Hessian Gradient of the loss on training point z_i

$H = \nabla_{\theta}^2 \left(\frac{1}{n} \sum_{i=1}^n \ell(z_i, f_{\theta; \mathcal{D}}) \right)$

Hessian is expensive & requires approximations

Applied IFs do not align with LOO-objective

IFs might surface not the structures we are interested in

...

? How valid are Influence Functions for LLM Data Attribution?

4 Experimental Setup

Minimal pairs of a property X

Enables IF contrast, $= \tau_{IF}(\checkmark) - \tau_{IF}(\times)$, so shared context cancels and X remains

✓ Paris is in France.
vs.
✗ Paris is in Germany.

Isolate X

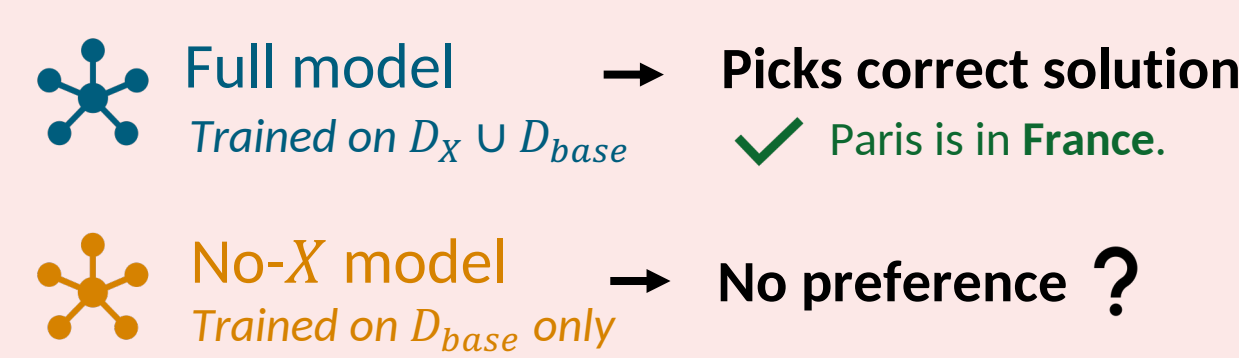
Filter training data for a property X (e.g., Paris-loc.-France)



Verify X is learned

Train two models & only full model prefers correct solution

Learning is 'attributable' to X



Recover X with IFs

Compute & rank on IF contrast

Examples containing X should rank highest

- 1 Paris is the largest city in France.
- 2 The Louvre is in Paris, France.
- ...
- 152 A lot of airports are closed.

4 BLIMP-style linguistic phenomena & facts on BEAR benchmark

GPT-2 sized models up to 5.6B tokens

3 Hessian approximation methods K-FAC, TrackStar, Gradient Similarity

5 RQ1: Do IFs surface the data with the target property X ?

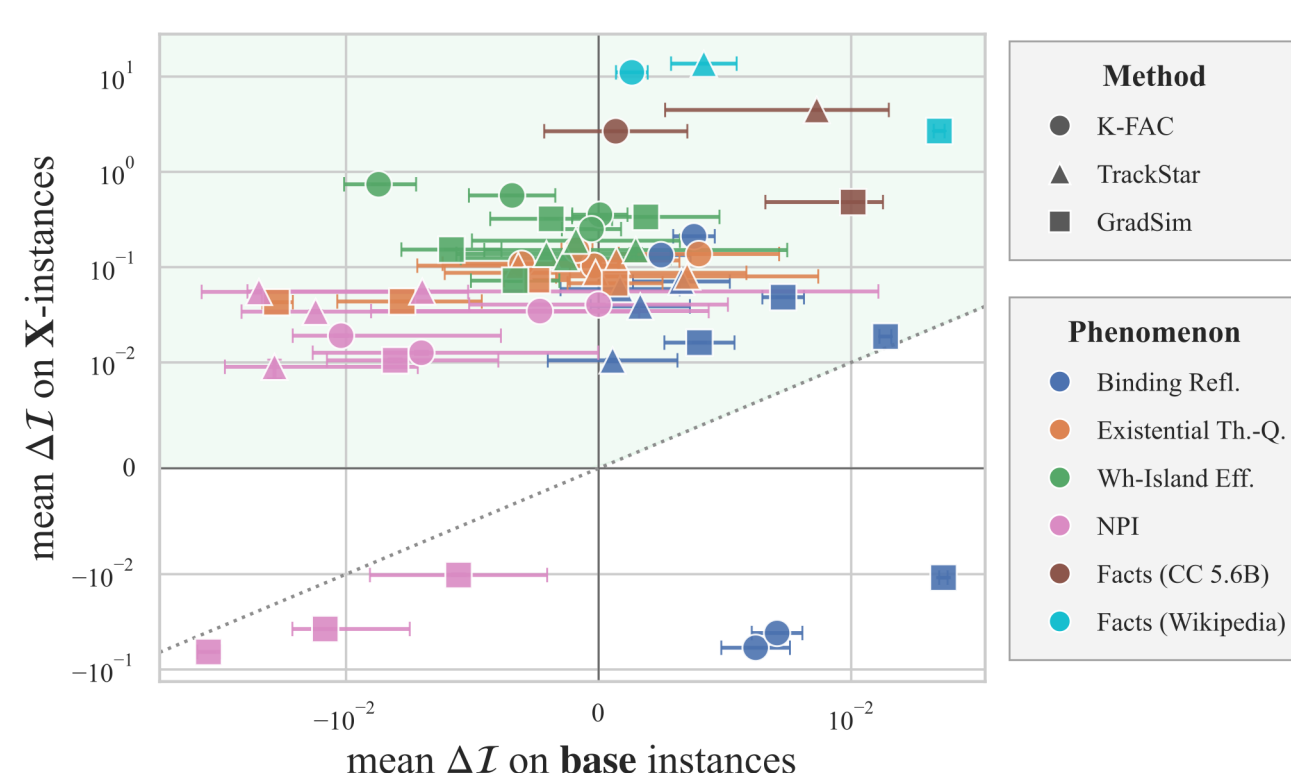


Figure: The avg. influence on X -training data vs. non- X data

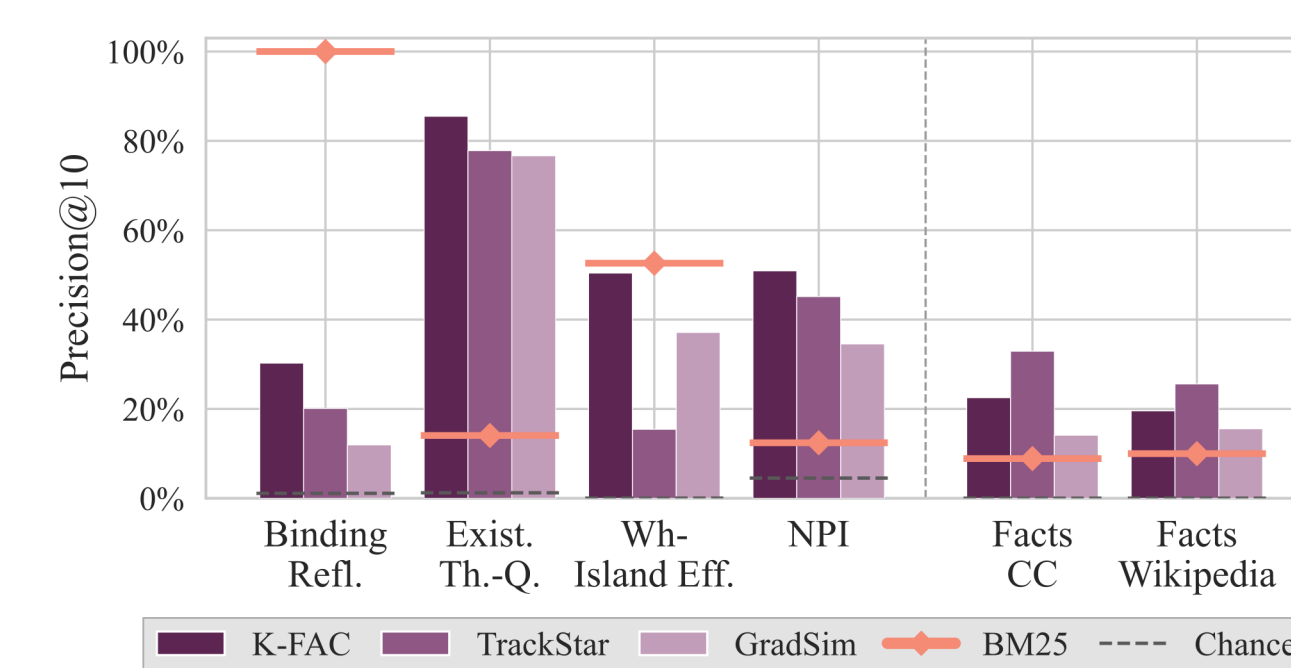


Figure: Share of top 10 retrieved training data points conveying X

Yes...

In aggregate, IFs concentrate on X data more than on non- X data

...and No

Among the top 10 most influential data, X data does not always appear first

6 RQ2: Which types of non- X samples are identified as influential?

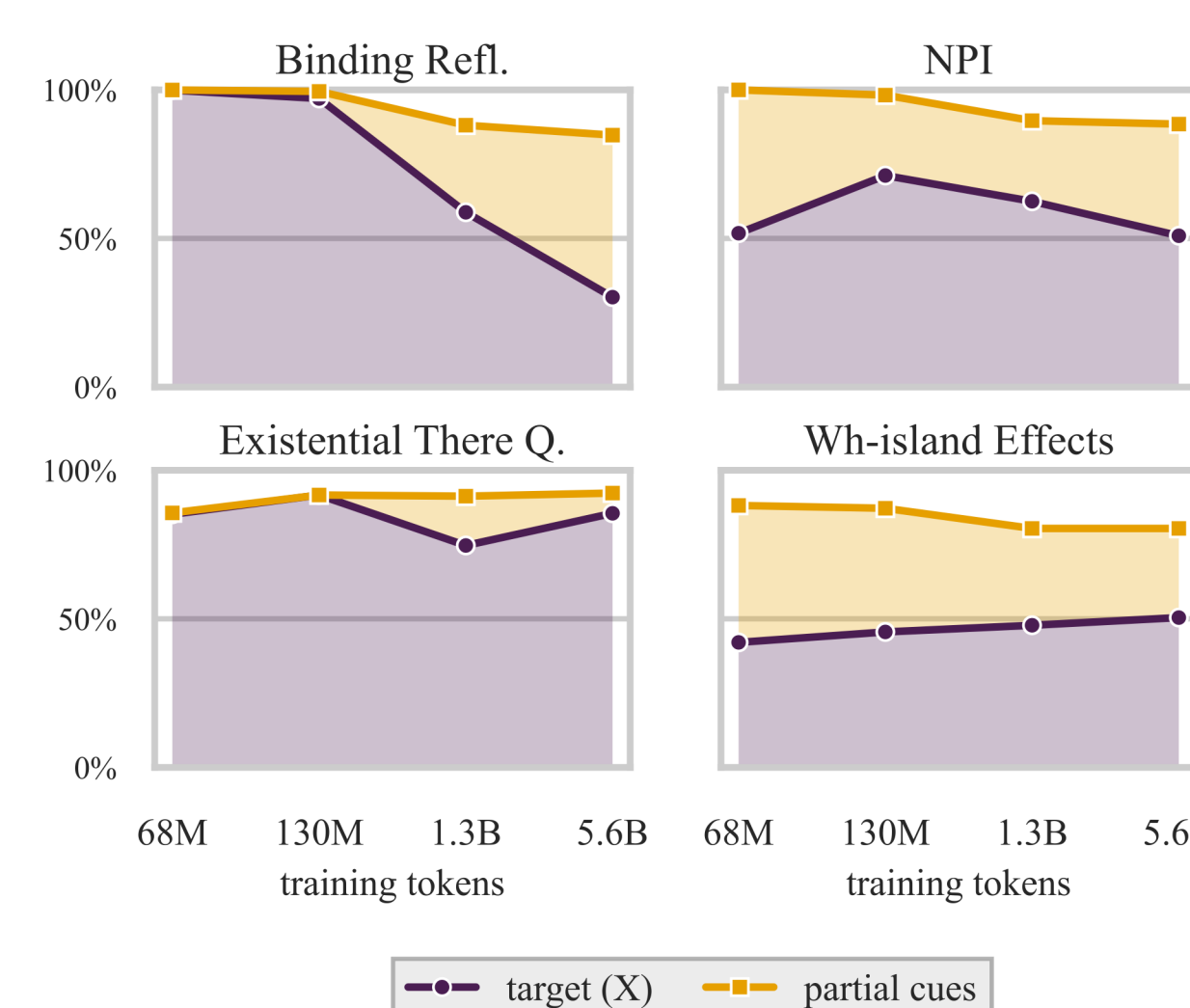


Figure: Share of samples with target phenomenon X and partial cues among top-10 retrievals across token budgets for linguistic phenomena.

Confounding Structures

Most non- X examples are indirectly relevant to X

For linguistic phenomena, these are structural 'neighbours', e.g., non-reflexive pronouns for binding reflexives

The boy hurt himself. → ...the boy hurt his knee...

For facts, these are, e.g., mentions of only one entity of a fact or mentions of related facts

Paris is in France → ...Disneyland is close to Paris, ...

➔ Closeness in embedding space or genuine generalization?



IFs are largely valid! They point to a property and its neighbours, hinting that models learn from indirect evidence.

Paper



Connect with me



NLPnorth Lab

