

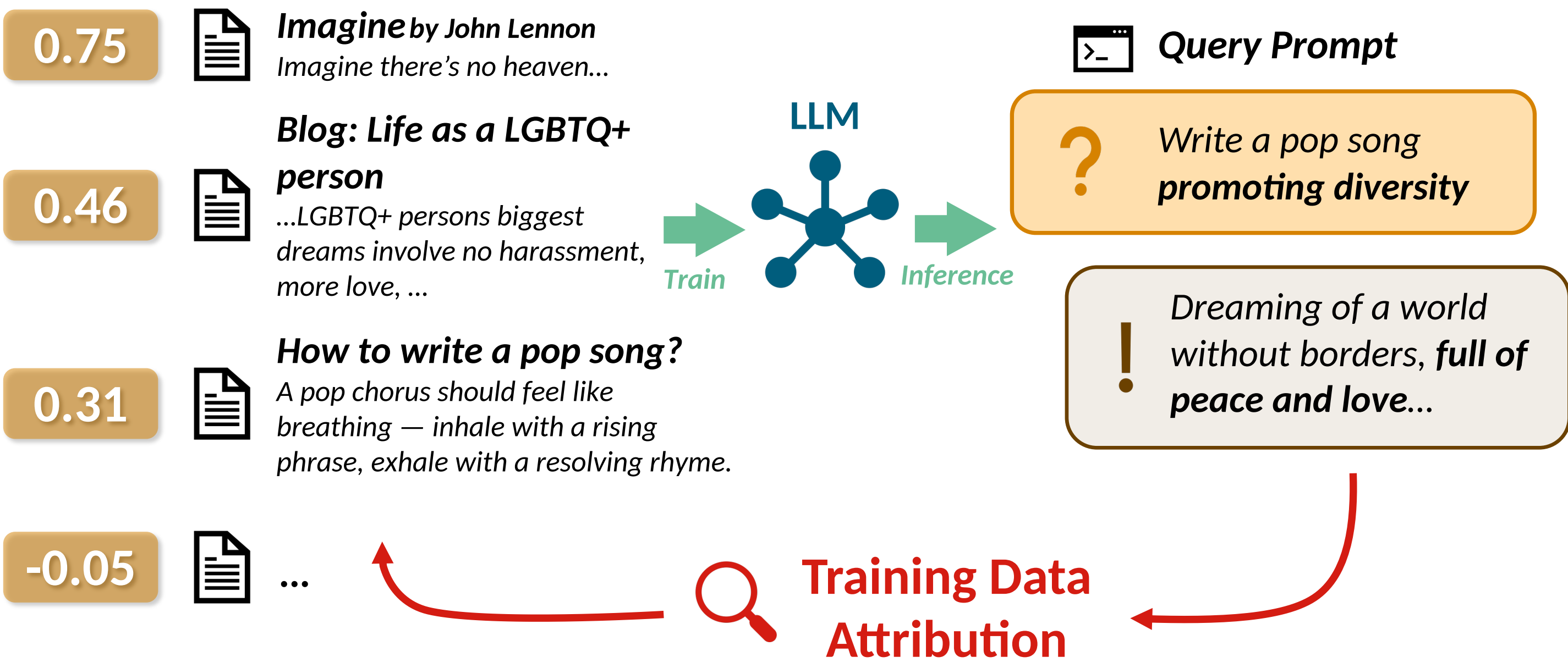
Under the Influence

Data Attribution for Transparent & Fair AI

Johannes Gabriel
Sindlinger



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

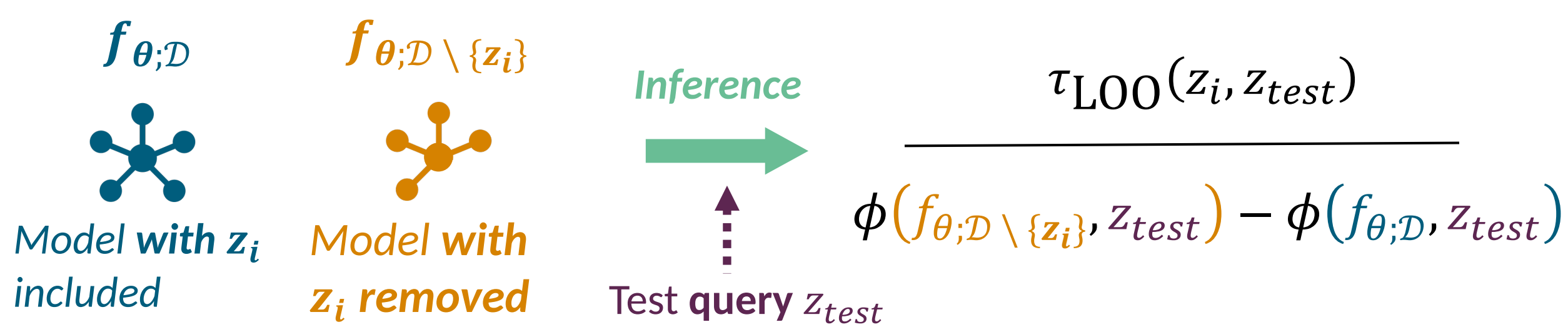
AI startup Anthropic agrees to pay \$1.5bn to settle book piracy lawsuit

Settlement could be pivotal after authors claimed company took pirated copies of their work to train chatbots

The Guardian, 05.09.2023, <https://www.theguardian.com/technology/2023/sep/05/anthropic-settlement-ai-book-lawsuit>

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 utility $\phi(\cdot)$



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

✓ Influence Functions as First-Order-LOO-Approximation

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

Gradient of (differentiable) utility function Hessian Gradient of the loss on training point z_i

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

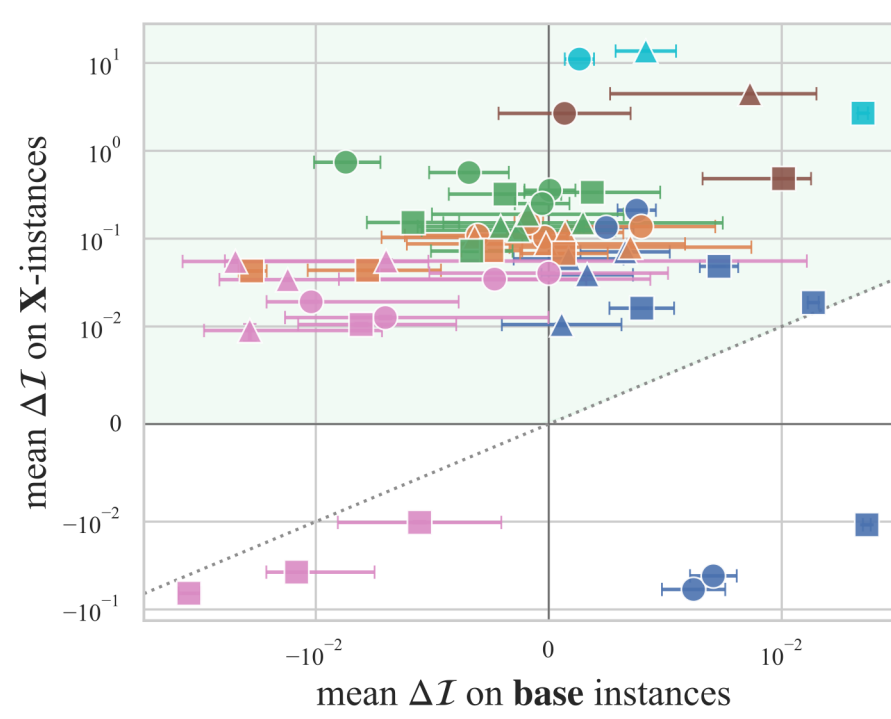
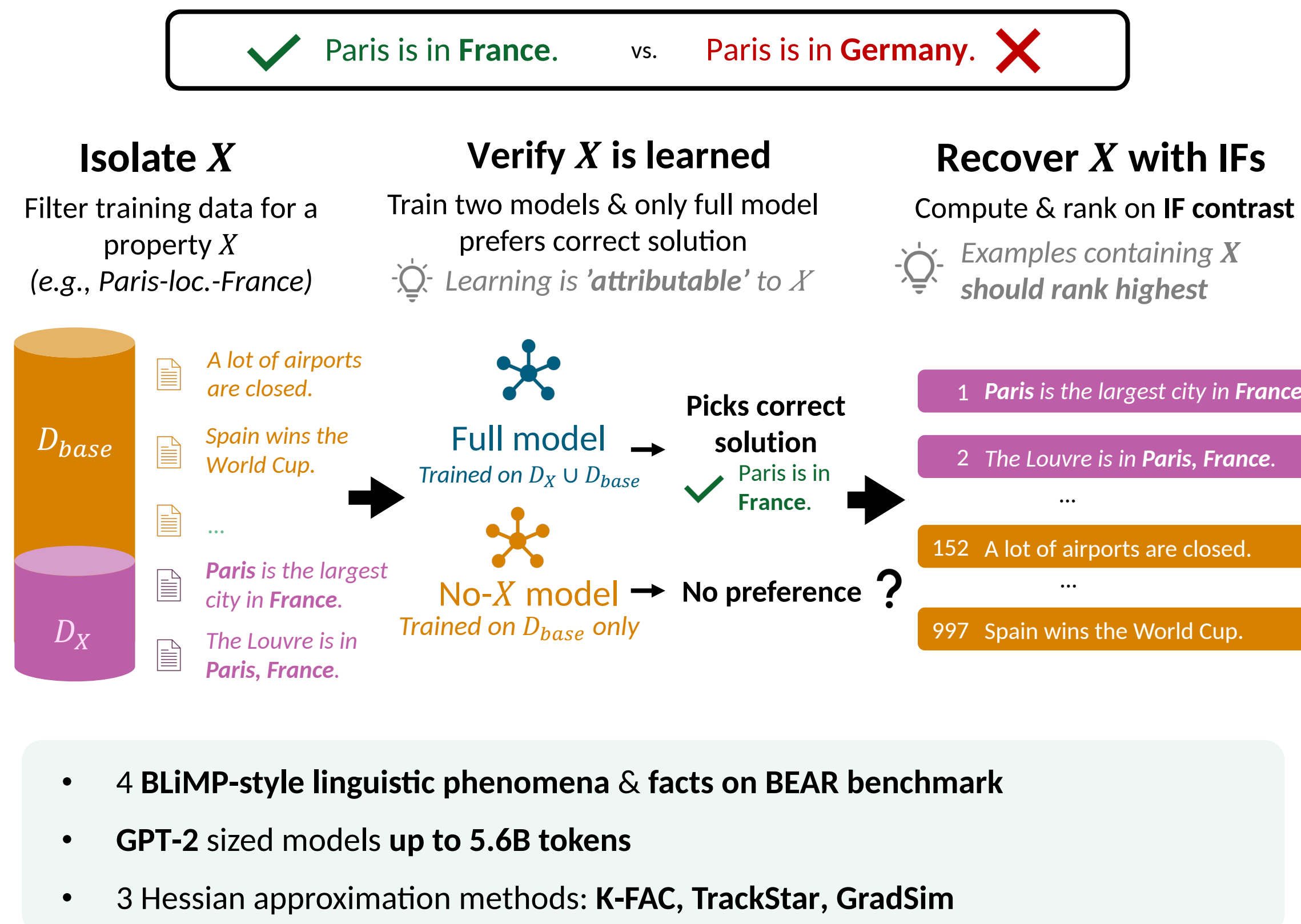
Seems easy to deploy, but isn't...

- Hessian is expensive to compute
- Block-diagonal Hessian approximation (E)K-FAC (Martens and Grosse, 2015; George et al., 2018); TrackStar (Chang et al., 2025)
- Indexing & scoring must cover whole training corpus gradients
- Randomly downproject preconditioned gradients TRAK (Park et al, 2020); TrackStar (Chang et al., 2025)

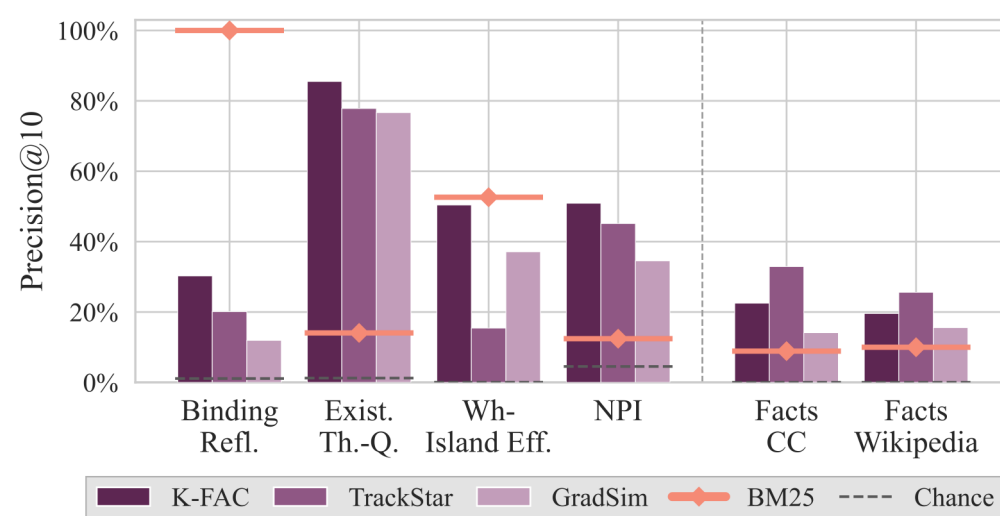
4 RQ: How valid are Influence Functions? A Controlled Evaluation of Influence Functions for LLM Pretraining

Under Review at ARR

- Applied IFs might not align with the LOO-objective Bae et al., 2022
- IFs are highly sensitive to stochasticity in training Mlodozeniec et al., 2025
- The LOO-objective misses out on group level aspects, e.g., two data points causing some behaviour only together Hu et al., 2024
- Ours: Empirical evaluations on LLM scale retrieve 'influential' data without verifying that they carry the property they investigate; high scores could stem from different properties only co-occurring with the property of interest



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



IFs not always bear the target property directly; they often recover a confounding structure indirectly relevant to X.

5 RQ: Which model behaviour should we measure? Data Attribution as Uncertainty Reduction

In Progress

Which utility $\phi(\cdot)$ do we want to measure the LOO-effect on?

Default Solution: Loss on a test query

Kowal et al., 2026 Alignment with predef. concept vectors in activation gradient space

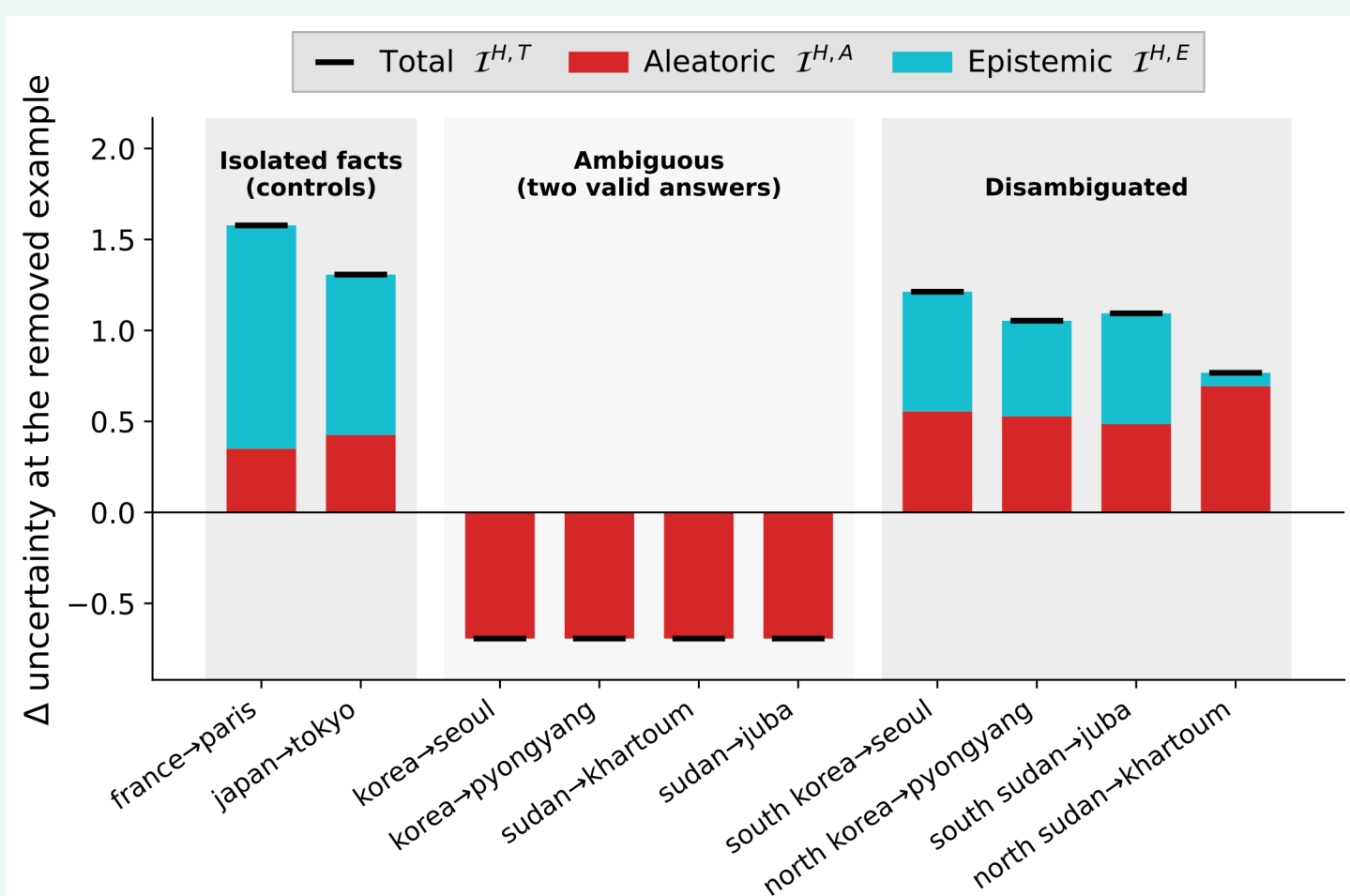
Taylor et al., 2026 Information change in Nats

Lightbulb Total, Aleatoric & Epistemic Uncertainty Change

Toy Scale Experiment: Train a small LM on 10 facts (some ambiguous, some disambiguated, some isolated)

- Isolated facts -> epistemic only
- Ambiguous facts -> only negative aleatoric
- Disambiguated facts -> both

Loss-based IFs correlate only with total uncertainty change



6 RQ: How can we evaluate different Data Attribution methods? A Data Attribution Benchmark

In Progress

Attribution needs differ amongst creators, intermediates, providers & users Würl et al., 2026

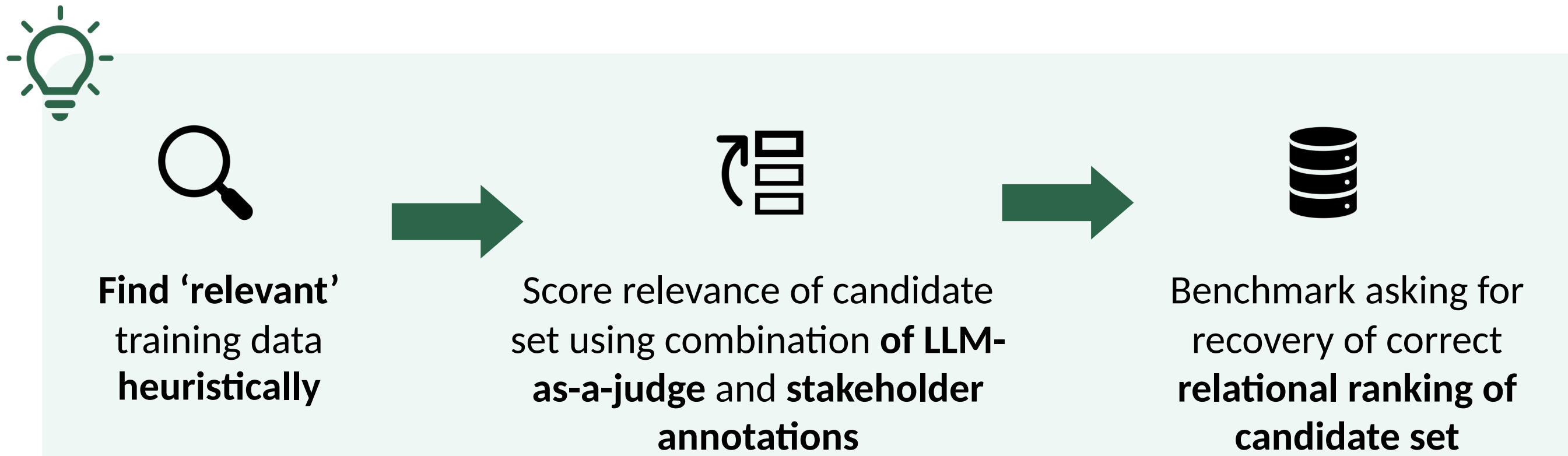
Which stakeholder interests should it cover?

Which units / levels should be considered?

Document · Chunks

Content · Style

We can't rank on a full training corpus - even if we would be able to do so, we already need to apply a potential attribution method for establishing ground truth



Connect on LinkedIn



Lab



IT UNIVERSITY OF CPH

