

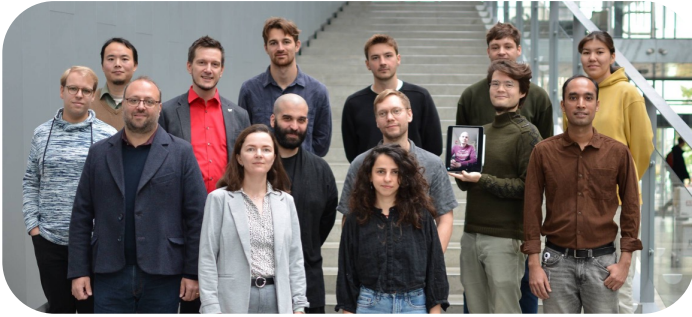
# Under the Influence

## Data Attribution for Transparent & Fair AI

Johannes Gabriel Sindlinger

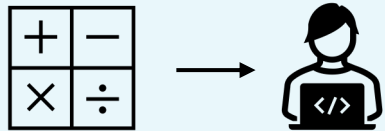
08.09.2026

# About Me



2<sup>nd</sup> year PhD

@NLPnorth lab / IT University of Copenhagen



Econometrics &

CS / Data Science background

@Ulm University & Heidelberg University



We ❤️ Swabia



Thumbs up for cycling!

Disclaimer: I did it already before it got cool...

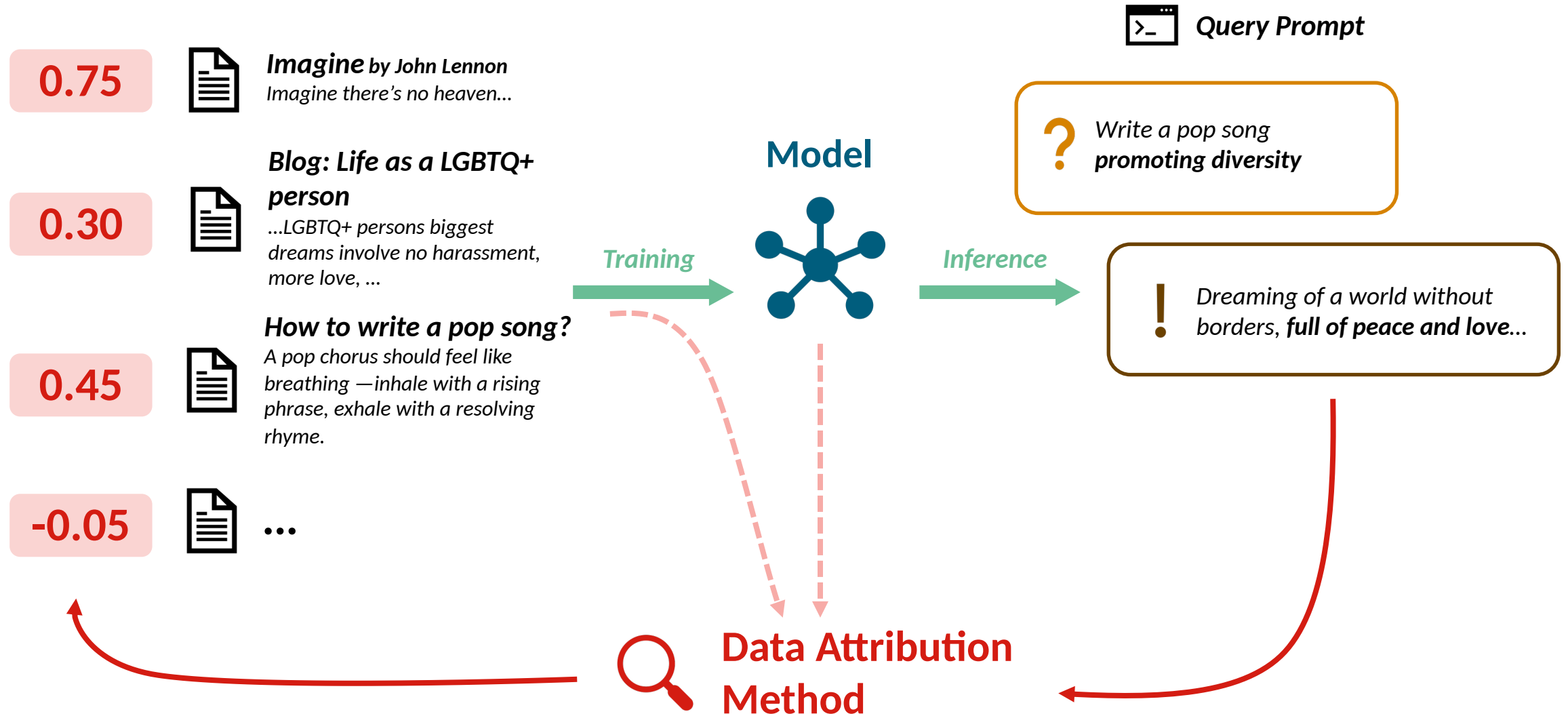
# Outline

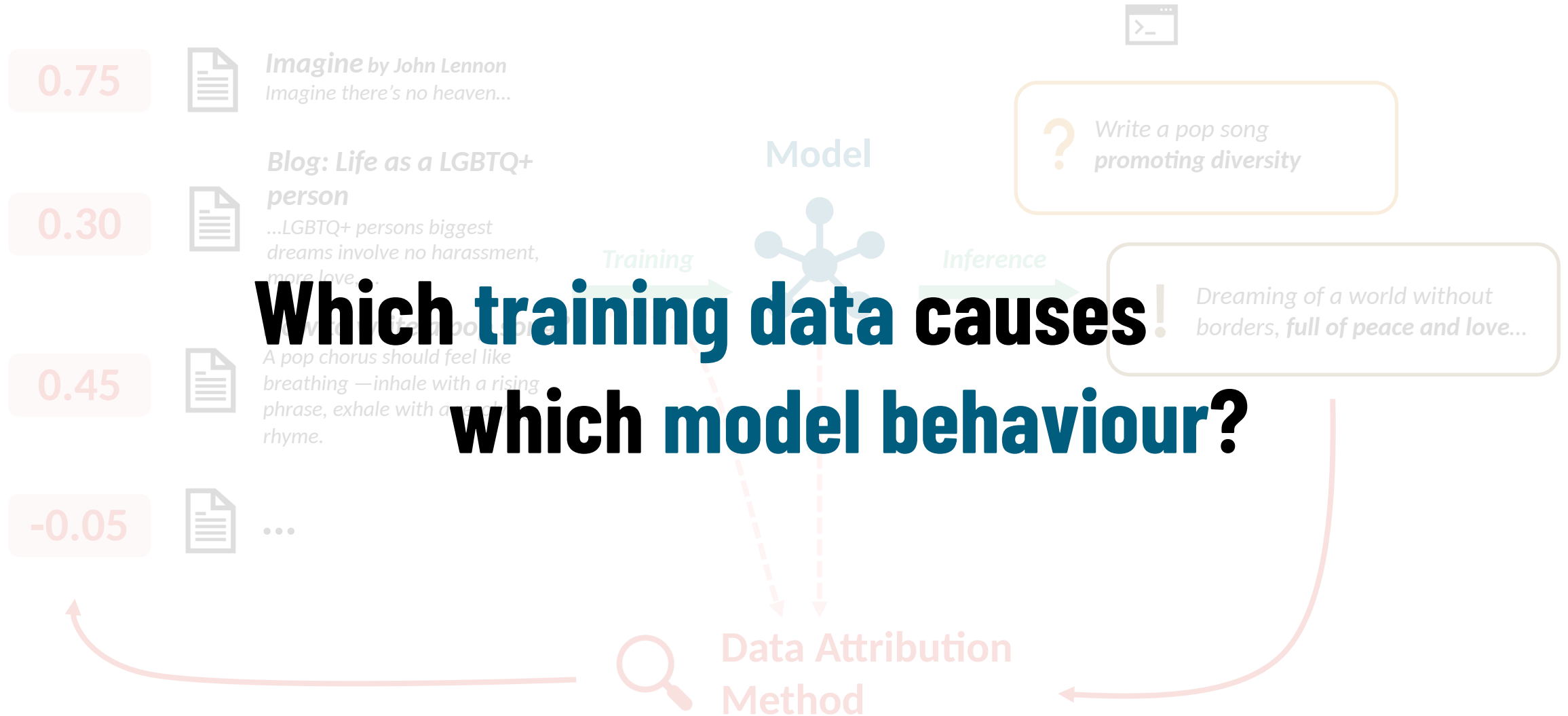
1. Introduction to Data Attribution & Influence Functions
2. *Under the Influence – on What? A Controlled Evaluation of Influence Functions*
3. Open Topics

# Outline

1. Introduction to Data Attribution & Influence Functions
2. *Under the Influence – on What? A Controlled Evaluation of Influence Functions*
3. Open Topics

# What is Data Attribution?





# Why Data Attribution?

## POISONING ATTACKS ON LLMs REQUIRE A NEAR-CONSTANT NUMBER OF POISON SAMPLES

**Alexandra Souly<sup>1,\*</sup>, Javier Rando<sup>2,5,\*</sup>, Ed Chapman<sup>3,\*</sup>, Xander Davies<sup>1,4,\*</sup>**

**Burak Hasircioglu<sup>3</sup>, Ezzeldin Shereen<sup>3</sup>, Carlos Mougán<sup>3</sup>, Vasilios Mavroudis<sup>3</sup>, Erik Jones<sup>2</sup>**

**Chris Hicks<sup>3,†</sup>, Nicholas Carlini<sup>2,†</sup>, Yarin Gal<sup>1,4,†</sup>, Robert Kirk<sup>1,†</sup>**

<sup>1</sup>UK AI Security Institute, <sup>2</sup>Anthropic, <sup>3</sup>Alan Turing Institute, <sup>4</sup>OATML, University of Oxford, <sup>5</sup>ETH Zurich

\*Core contributor, †Senior advisor

*Souly et al. 2025, <http://arxiv.org/abs/2510.07192>*

## 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, 05.09.2025, <https://www.theguardian.com/technology/2025/sep/05/anthropic-settlement-ai-book-lawsuit>*

# Applications of Data Attribution



**Model Debugging**



**Data Valuation**



**Machine Unlearning**



**LLM Security**



**Creator  
Compensation**

...

# Different Notions of Data Attribution

Wührl et al., 2026



Idea



Concepts

## Data-driven Methods

Are outputs similar to training data items?

N-grams, Embeddings, ...

## Causal-driven Methods

What would happen if we retrained the model without specific data points?

Data Shapley, Influence Functions, Unrolled Differentiation, ...

## Usage-driven Methods

Was a data point used at any time during model development or deployment?

Open model training / deployment, Licensing, ...

# Different Notions of Data Attribution

Wühl et al., 2026



Idea



Concepts

## Data-driven Methods

Are outputs similar to training data items?

N-grams, Embeddings, ...

## Causal-driven Methods

What would happen if we retrained the model without specific data points?

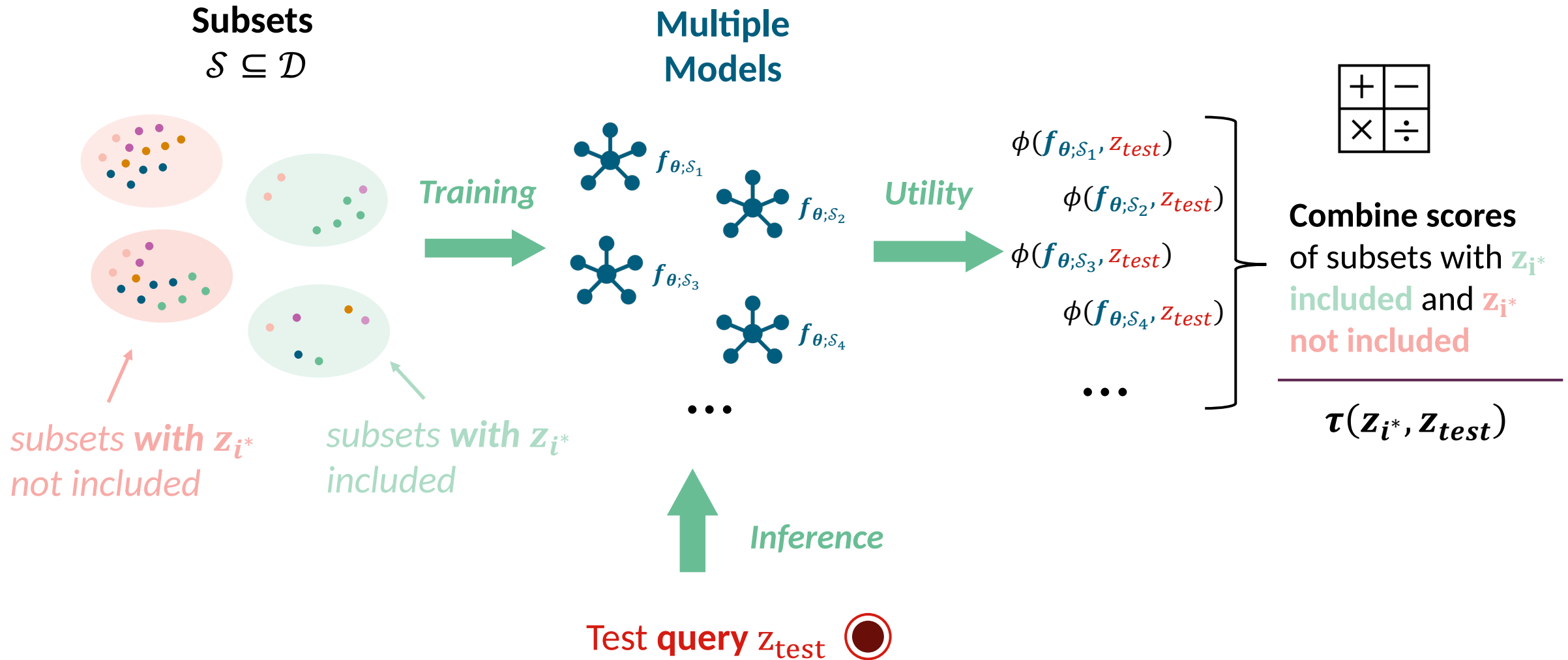
Data Shapley, Influence Functions, Unrolled Differentiation, ...

## Usage-driven Methods

Was a data point used at any time during model development or deployment?

Open model training / deployment, Licensing, ...

# Retraining-based Approaches



# Data Shapley

Ghorbani and Zou, 2019, Slide adapted from Ilyas et al., 2024

$$\gamma(|S'|) = \frac{|S'|! (|S| - |S'| - 1)!}{|S|!}$$

Weighting term, ensuring subsets are treated fairly

$$\tau_{\text{Shapley}}(z_i, z_{\text{test}}) = \sum_{\substack{S' \subseteq S \setminus \{z_i\} \\ \subset \mathcal{D}}} \gamma(|S'|) \left[ \phi(f_{\theta}; \mathbf{s}' \cup \{z_i\}, z_{\text{test}}) - \phi(f_{\theta}; \mathbf{s}', z_{\text{test}}) \right]$$

Every possible subset not including the  $i$ -th example

Marginal effect on utility of adding the example  $z_i$  to the subset  $S'$

# Data Shapley

Ghorbani and Zou, 2019, Slide adapted from Ilyas et al., 2024

$$\gamma(|S'|) = \frac{|S'|! (|S| - |S'| - 1)!}{|S|!}$$

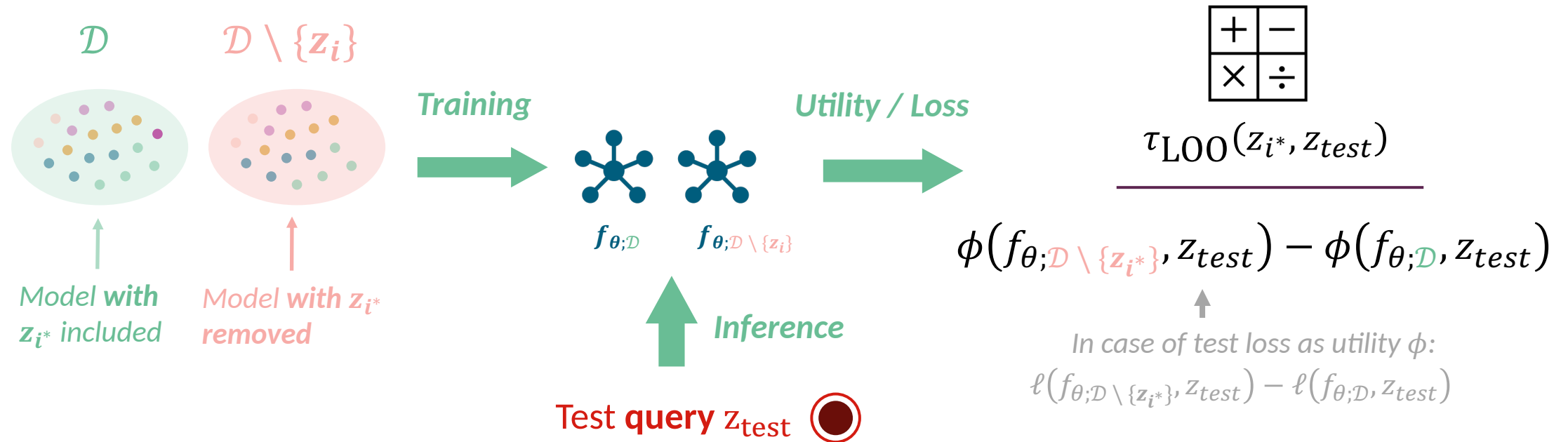
Weighting term, ensuring subsets are treated fairly

$$\tau_{\text{Shapley}}(z_i, z_{\text{test}}) = \sum_{\substack{S' \subseteq S \setminus \{z_i\} \\ \subset \mathcal{D}}} \gamma(|S'|) \left[ \underbrace{\phi(f_{\theta; S' \cup \{z_i\}}, z_{\text{test}})}_{\text{Marginal effect on utility of adding the example } z_i \text{ to the subset } S'} - \phi(f_{\theta; S'}, z_{\text{test}}) \right]$$

Every possible subset not including the  $i$ -th example

**! Retraining on all  $2^n$  subsets  $S \subseteq \mathcal{D}$  necessary**

# Leave-one-Out-Retraining (LOO)

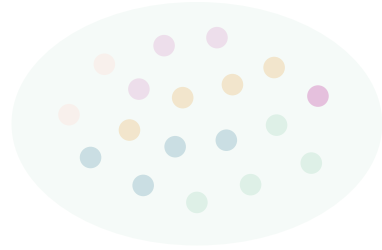


**! Still requires retraining for all  $N$  training data points**

**Idea:** Approximate the effect of removing a data point from model training **without actual retraining**

# From L00 to Influence Functions

Following Koh and Liang, 2017 and Grosse et al., 2023



$$\mathcal{D} = \{z_j\}_{j=1}^n$$

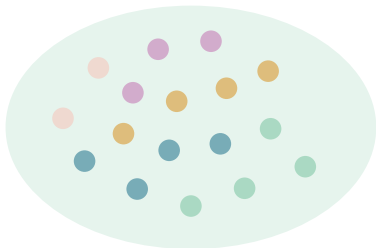
$$\theta^*(\mathcal{D}) = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^n \ell(z_j, f_{\theta})$$



Model  $f_{\theta^*; \mathcal{D}}$



**Remove training point  $z_i$**



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



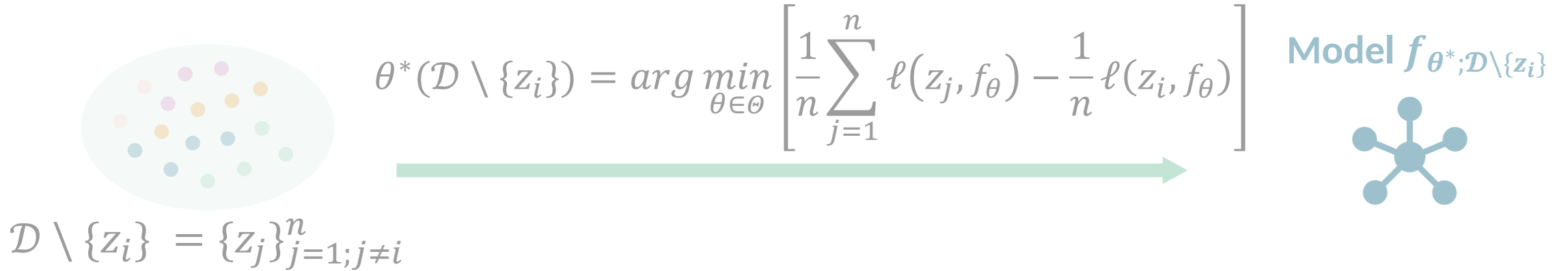
Model  $f_{\theta^*; \mathcal{D} \setminus \{z_i\}}$



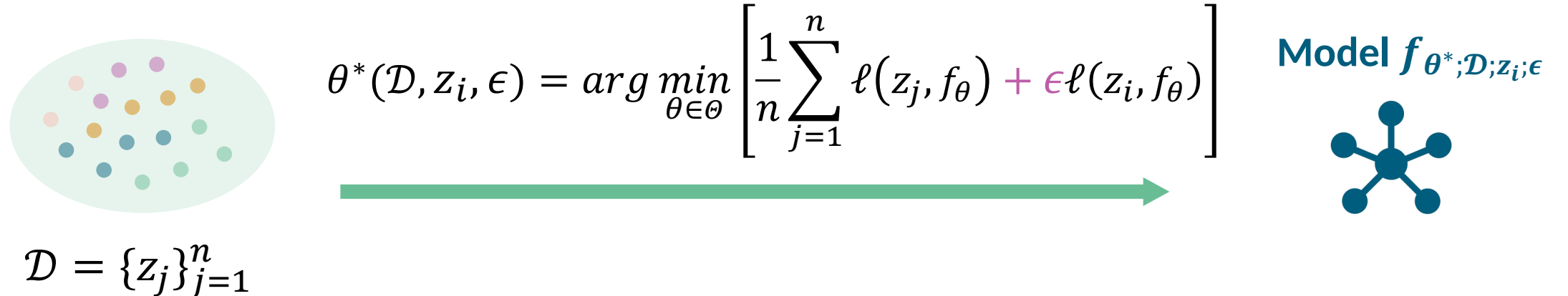
$$\mathcal{D} \setminus \{z_i\} = \{z_j\}_{j=1; j \neq i}^n \leftarrow \text{Full dataset without } z_i$$

# From L00 to Influence Functions

Following Koh and Liang, 2017 and Grosse et al., 2023



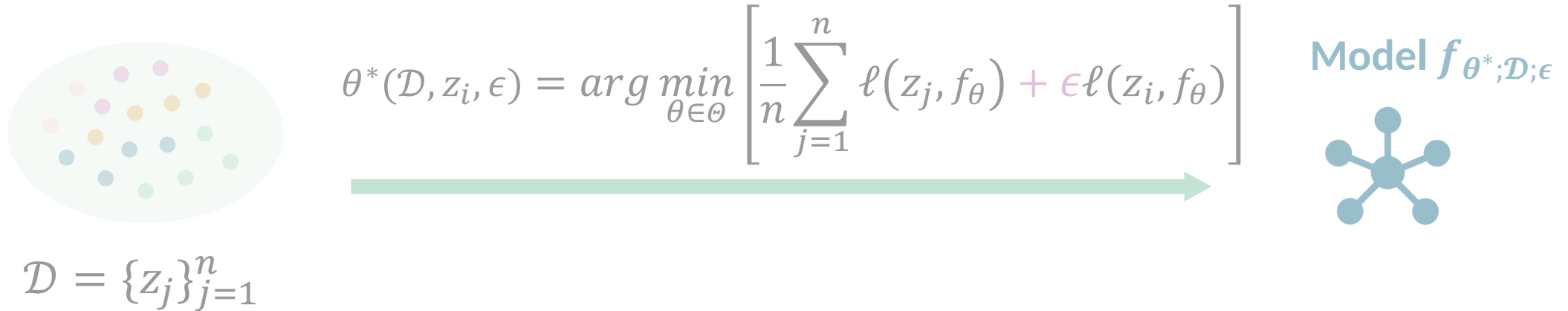
**Replace weight  $-\frac{1}{n}$  of  $z_i$  with arbitrary  $\epsilon$**



# From L00 to Influence Functions

Following Koh and Liang, 2017 and Grosse et al., 2023

Assuming  $L(\cdot, \theta)$  is strictly convex and twice differentiable and unique global minimum exists

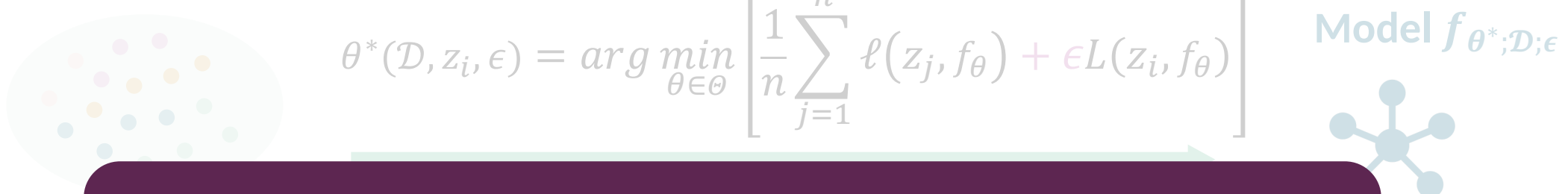


**Approximate parameter change  $\Delta_\epsilon = \theta^*(\mathcal{D}, z_i, \epsilon) - \theta^*(\mathcal{D})$   
via first-order Taylor expansion around  $\epsilon = 0$**

# From L00 to Influence Functions

Following Koh and Liang, 2017 and Grosse et al., 2023

Assuming  $L(\cdot, \theta)$  is strictly convex and twice differentiable and unique global minimum exists


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

$\mathcal{D} =$  **Approximate parameter change  $\Delta_{\epsilon} = \theta^*(\mathcal{D}, z_i, \epsilon) - \theta^*(\mathcal{D})$   
via first-order Taylor expansion around  $\epsilon = 0$**

✓ **After some calculus:**  $\Delta_{\epsilon} \approx \left. \frac{d\theta_{\epsilon}^*}{d\epsilon} \right|_{\epsilon=0} \cdot \epsilon = -H^{-1} \nabla_{\theta} \ell(z_i, f_{\theta^*; \mathcal{D}}) \cdot \epsilon$

**With:**  $H = \nabla_{\theta}^2 \left( \frac{1}{n} \sum_{i=1}^n \ell(z_i, f_{\theta^*; \mathcal{D}}) \right)$  ← Hessian of the expected loss under the training data distribution w.r.t to the parameters at optimum  $\theta^*(\mathcal{D})$

# From L00 to Influence Functions

Following Koh and Liang, 2017 and Grosse et al., 2023

$$\Delta_{\epsilon} = \theta^*(\mathcal{D}, z_i, \epsilon) - \theta^*(\mathcal{D}) \longrightarrow \Delta_{\epsilon} \approx -H^{-1} \nabla_{\theta} \ell(z_i, f_{\theta^*}; \mathcal{D}) \cdot \epsilon$$

**Bringing back  $\tau_{\text{LOO}}(z_i, z_{\text{test}})$  into the game**

$$\begin{aligned} \text{Recap: } \tau_{\text{LOO}}(z_i, z_{\text{test}}) &= \phi(f_{\theta^*; \mathcal{D} \setminus \{z_i\}}, z_{\text{test}}) - \phi(f_{\theta^*; \mathcal{D}}, z_{\text{test}}) \\ &= \phi(f_{\theta^*; \mathcal{D}; z_i; \epsilon}, z_{\text{test}}) - \phi(f_{\theta^*; \mathcal{D}}, z_{\text{test}}) \quad \text{for } \epsilon = -\frac{1}{n} \end{aligned}$$

**Applying Taylor expansion as before + chain rule:**

$$\begin{aligned} \phi(f_{\theta^*; \mathcal{D} \setminus \{z_i\}}, z_{\text{test}}) - \phi(f_{\theta^*; \mathcal{D}}, z_{\text{test}}) &\approx \left. \frac{d\phi(\theta_{\epsilon}^*)}{d\epsilon} \right|_{\epsilon=0} \cdot \epsilon \\ &= \nabla_{\theta} \phi(f_{\theta^*; \mathcal{D}}, z_{\text{test}})^T \left. \frac{d\theta_{\epsilon}^*}{d\epsilon} \right|_{\epsilon=0} \cdot \epsilon \quad (\text{chain rule}) \\ &= -\nabla_{\theta} \phi(f_{\theta^*; \mathcal{D}}, z_{\text{test}})^T H^{-1} \nabla_{\theta} \ell(z_i, f_{\theta^*}; \mathcal{D}) \cdot \epsilon \end{aligned}$$

# Influence Functions

Koh and Liang, 2017

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

Gradient of a (differentiable)  
**utility function** of  $z_{test}$ , e.g., the  
loss of the model on  $z_{test}$

**Hessian** of the expected loss  
under the training data  
distribution **w.r.t to the  
parameters** at optimum  $\theta^*(\mathcal{D})$

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

Gradient of the **loss** of  
the model **on training  
point  $z_i$**

# Influence Functions

Koh and Liang, 2017

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

$$\tau_{IF}(z_i, z_{test}) \gg 0$$

*$z_i$  pushes model into direction which improves performance of  $\phi$*

$$\tau_{IF}(z_i, z_{test}) \approx 0$$

*$z_i$  moves model into direction which is irrelevant to  $\phi$*

$$\tau_{IF}(z_i, z_{test}) \ll 0$$

*$z_i$  moves model into direction which hurts performance of  $\phi$*

# Problems with Influence Functions

Koh and Liang, 2017, Slide adapted from Ye and Marasović, 2024

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

**!?**  $H$  not invertible by default, due to...

- Non-convex loss landscape
- Overparametrized models
- Incomplete model convergence

Usually we don't find the global optimum  $\theta^*$

**!?** Computing  $H^{-1}$  is expensive

- $H^{-1}$  requires  $O(\#params^2)$  memory
- Standard matrix inversion requires  $O(\#params^3)$  compute

**!?** We need to store  $\nabla_{\theta} L(z_i)$  for the whole training corpus

# More Problems...

with (applied) Influence Functions



**Applied IFs might not align with the LOO-objective**

Bae et al., 2022



**IFs are highly sensitive to stochasticity in training**

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



**The LOO-objective misses out on group level aspects, e.g., multiple data points causing some behaviour only together**

Hu et al., 2024



**Can we trust Influence Functions when scaling them up to LLM size?**

# Prior Work on LLM Scale

Grosse et al., 2023

- **First work to scale** Influence Functions on *(filtered subsets of)* pre-training data **to LLM sizes** using **(E)K-FAC Hessian approximation**
- **Predominantly qualitative evaluation:**
  - **Larger models seem to generalize more abstractly**, i.e., small models have high influence scores on token-level, large models on themes and concepts
  - **Word ordering seem to matter**, i.e., flipping order of important phrases in queries changes influence scores heavily
  - When **models express 'desires'** (e.g., not wanting to be shut down), this traces back to **similar fictional narratives in training data**
  - **Different layers capture different types of knowledge**, i.e. actual queries concentrate in upper layers, abstract reasoning in middle layers; computing influence on the final layer alone misses important patterns

Query: `binary_search`

```
Prompt: def A(list1,n):
B=list1;C=0;D=len(B)-1;A=0
while C<=D:
    A=(D+C)//2
    if B[A] < n:
```

```
Completion: C=A+1
            elif BA>n:D=A-1
            else:return A
            return -1
```

Influential Sequence for 810 Million Parameter Model (Influence = 0.149)

```
C 146 C B C B C B 147 C 148 A B B B A B A C B A A C B B A B A B A B A 149 A C C C C
150 B A B B A A B B B A B B B C A A B B B A 151 C 152 C B A C C C B 153 A A A B A A C
C B B B B A A B B A A A 154 C C B A 155 A B C B C A C C C C A B B A B C C B A A C
156 B B B B C C B C A B B A B B A A B 157 C C B C C C 229 B B B B B B C C B B B C B B
230 B B B A C A C B C B C B 231 A 232 B A B A B C C C C B C A B B B 233 B B A A A
A B C A C C B A A C A B A B 234 B B C C B C B C B 235 A A A A A A A B A B A A A A B
A A A B A 236 B B A A B B A C B B A A C B B 237 B B B A C A A C B C B B B A B B B 238 C
A 239 B
```

Influential Sequence for 52 Billion Parameter Model (Influence = 0.015)

```
public class L0035SearchInsertPosition {
    public static void main(String[] args) {

    } public static int searchInsert(int[] nums, int target) {
        int left = 0;
        int right = nums.length - 1;
        while (left <= right) {
            int mid = (left + right) / 2;
            if (nums[mid] < target) {
                left = mid + 1;
            } else {
                right = mid - 1;
            }
        }
        return left;
    }
}
```

# Prior Work on LLM Scale

Chang et al., 2024

- **Scaling up** influence functions to full pre-training corpora using multiple **'engineering' tweaks**:
  - *E.g., unit normalizing & random projection of gradients*
- Evaluation on **identifying influential data points for factual predictions**:
  - *Influence Functions perform **worse than BM25** on **fact retrieval***
  - *Influence functions **outperform** BM25 at finding data points that 'causally influence' a fact, measured by how a single additional training step on a retrieved example reinforces the model's belief in a fact / increases the likelihood to produce the fact*
  - ***Performance** of Influence Functions **grows for 'better' models**, i.e., more training and more parameters*

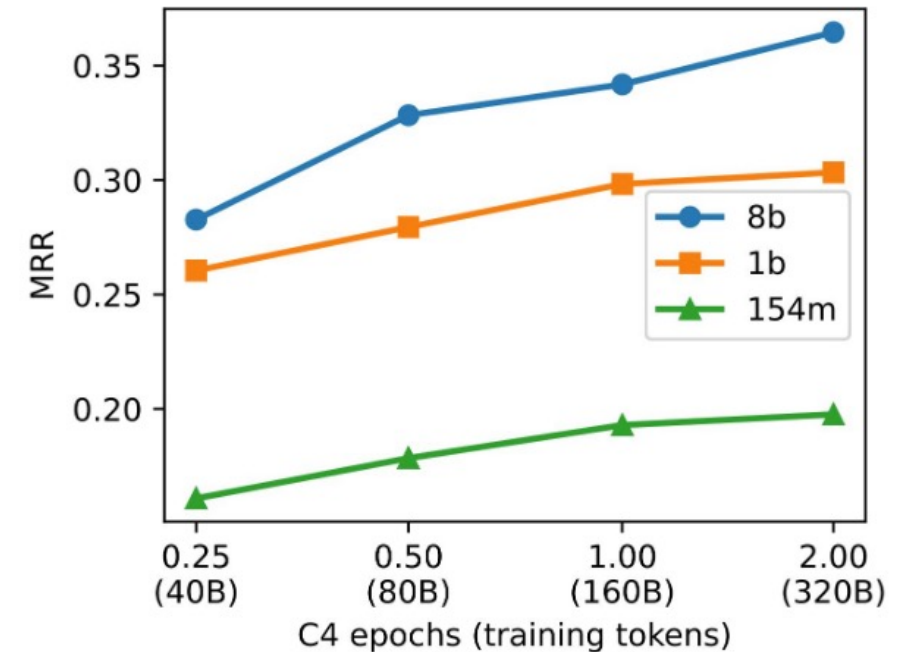
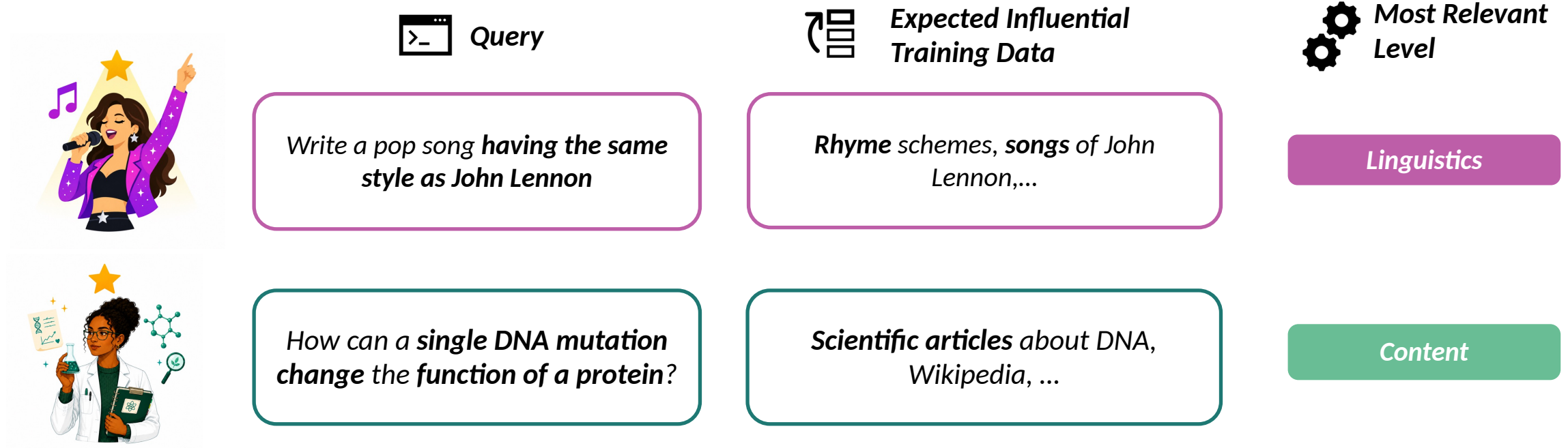


Figure: Fact retrieval score (MRR) throughout pretraining for different model sizes, Chang et al., 2024

# Outline

1. Introduction to Data Attribution & Influence Functions
2. *Under the Influence – on What? A Controlled Evaluation of Influence Functions*
3. Open Topics

# Which level of structure do Influence Functions actually capture?



Without isolating a property of interest, IF scores might be biased through cross-level interactions

*We do not know whether Influence Functions do what they're supposed to do and when we apply them we always assume – without questioning - that they match the level of structure we're interested in...*

**? How can we audit whether IFs can serve as a legitimate LLM interpretability tool?**

# Under the Influence – on What?

*Under Review at ARR*

## A Controlled Evaluation of Influence Functions



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



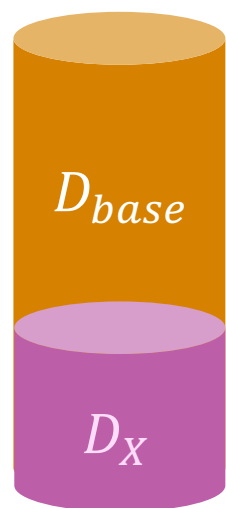
Learning is '*attributable*' to  $X$

### Recover $X$ with IFs

Compute & rank on IF contrast



Examples containing  $X$   
should rank highest



A lot of airports  
are closed.



Spain wins the  
World Cup.



...



**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.

...

997 Spain wins the World Cup.

# Under the Influence – on What?

## A Controlled Evaluation of Influence Functions

*Under Review at ARR*

✓ 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

### Recover $X$ with IFs

Compute & rank on IF contrast

Examples containing  $X$   
highest

Requires  $X$  to be **isolable by filtering (distinguishable)**  
and **carried by specific training examples** rather than  
emergent from interactions across them (**independent**).



A lot of airports



Spain wins the  
World Cup.



...



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  
✓ ...which is non-trivial in practice to find.

No- $X$  model  
Trained on  $D_{base}$  only → No preference ?

152 A lot of airports are closed.

997 Spain wins the World Cup.

# Two Playgrounds



## Linguistic Phenomena

- **4 BLiMP phenomena** (Warstadt et al., 2020):
  - Existential there quantifiers
  - Binding reflexives
  - Wh-island effects
  - Licensed negative polarity items (NPIs)
- E.g.: **I'm not playing anymore.** vs. **\*I'm playing anymore.**



## Facts

- **BEAR world-knowledge facts benchmark** (Wiland et al., 2024)
  - One fact = Relation between two entities.
- E.g.: **Warsaw is the capital of Poland.** vs. **\*Warsaw is the capital of Germany.**



**GPT-2 sized models up to 5.6B tokens**



**3 Hessian approximation methods:**  
K-FAC, TrackStar, GradSim

# Two Playgrounds



Linguistic Phenomena



Facts



Measuring the relevance of a training data point on the **influence contrast** of a minimal pair, e.g.,

$$\Delta I(z_i) = I(z_i, \text{I'm not playing anymore.}) - I(z_i, \text{I'm playing anymore.})$$



**Idea:** Limit any cross-interaction effect, attribution **should focus on X.**



GPT-2 sized models up to  
5.6B tokens



3 Hessian approximation methods:  
K-FAC, TrackStar, GradSim

# RQ1: Do IFs surface the training data with the target property $X$ ?

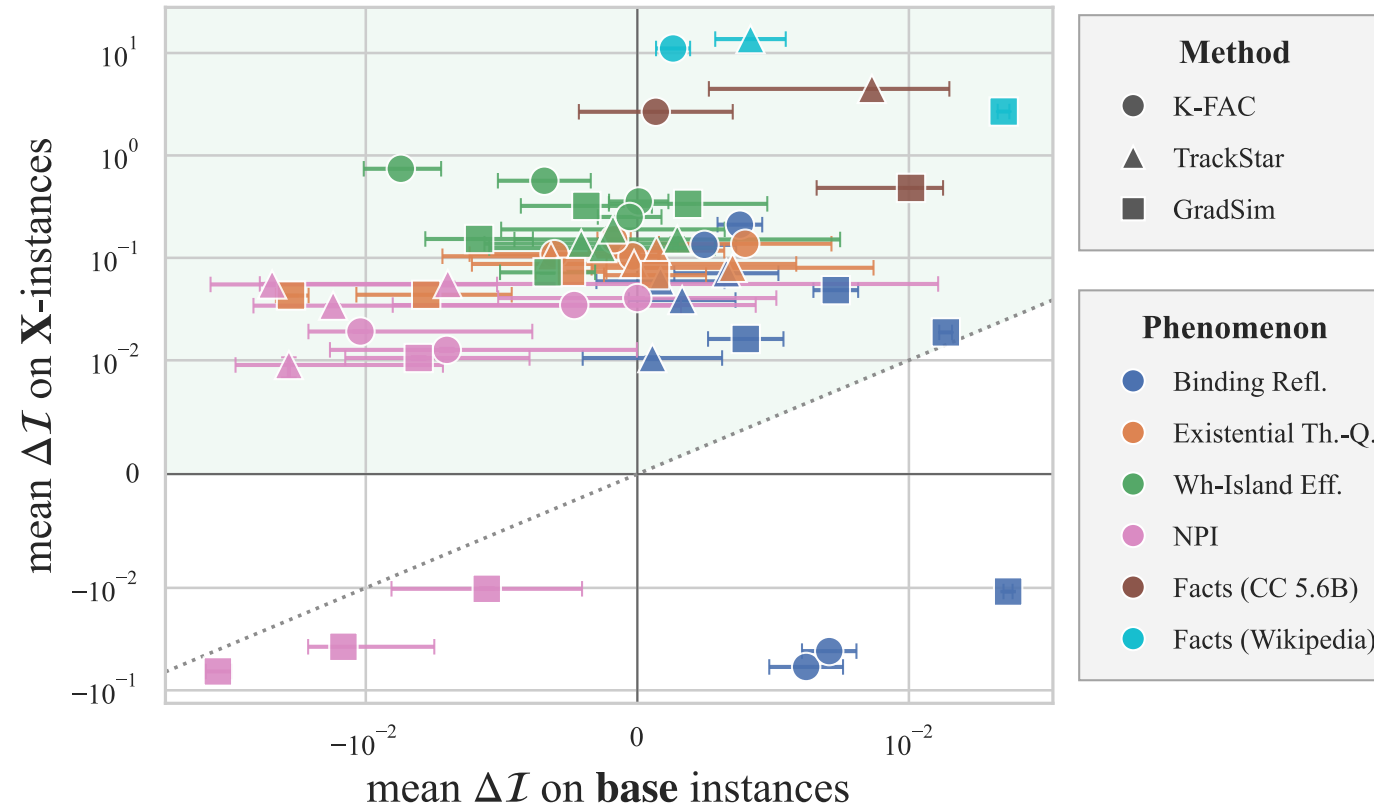
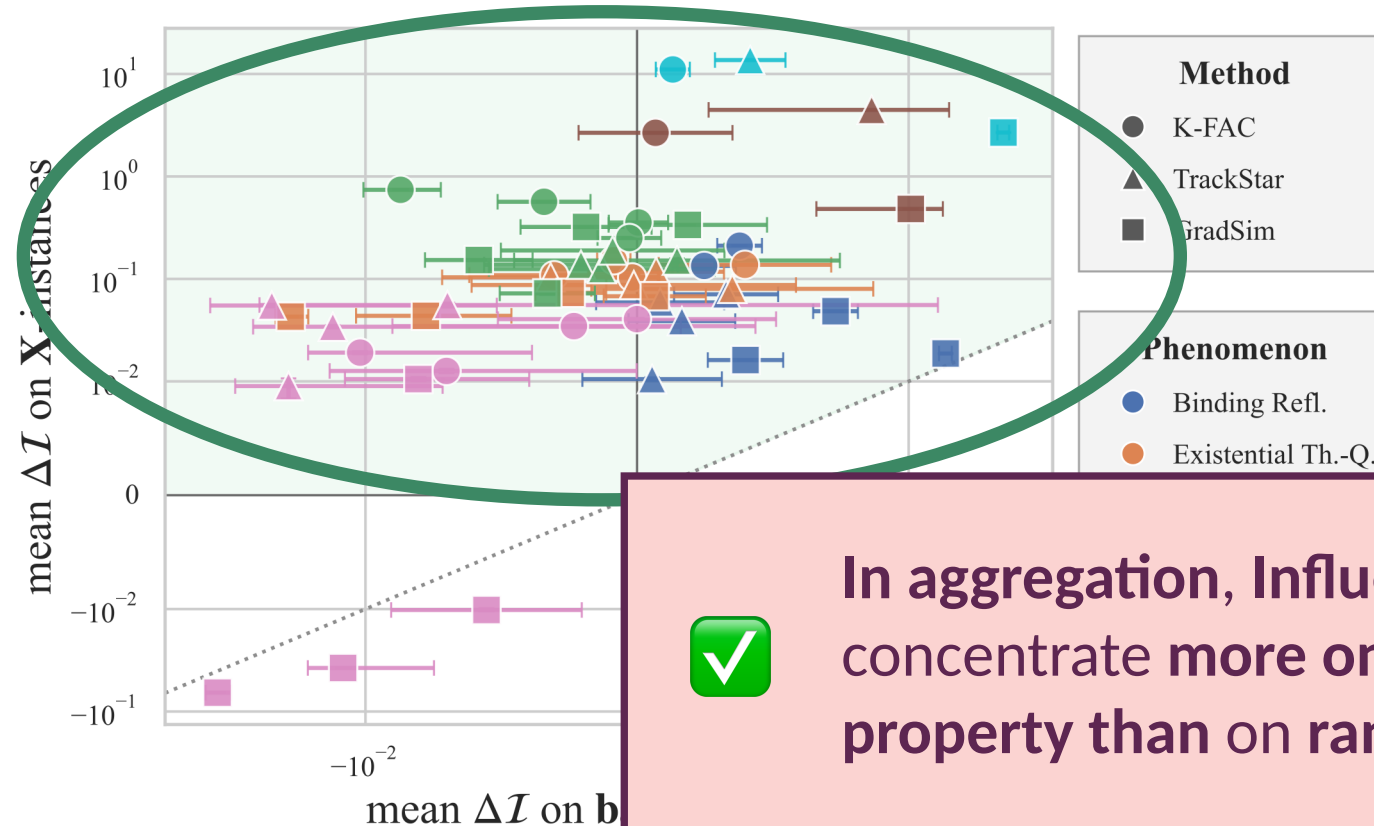


**Figure:** Average influence score for multiple queries on training data points which belong to a target property  $X$  (y-axis) vs. random base data points not conveying  $X$  (x-axis)

# RQ1: Do IFs surface the training data with the target property $X$ ?



 In aggregation, Influence Functions concentrate more on the target property than on random data.

**Figure:** Average influence score for multiple queries on training data points which belong to a target property  $X$  (y-axis) vs. random base data points not conveying  $X$  (x-axis)

# RQ1: Do IFs surface the training data with the target property $X$ ?

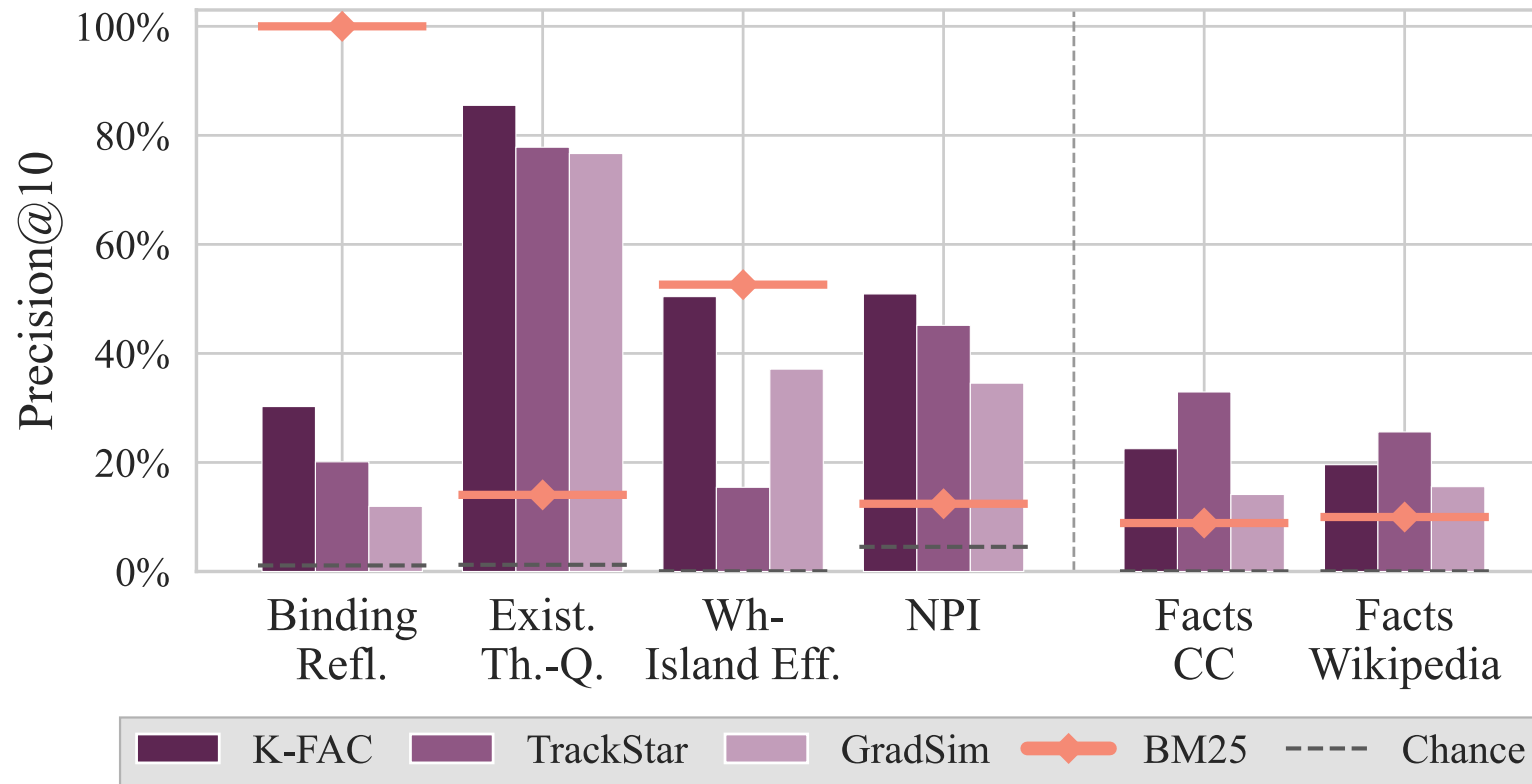


**Figure:** Precision@10 (= How many items of Top-10 retrieved items belong to  $X$ ?) on the influence contrast averaged across queries.

# RQ1: Do IFs surface the training data with the target property $X$ ?

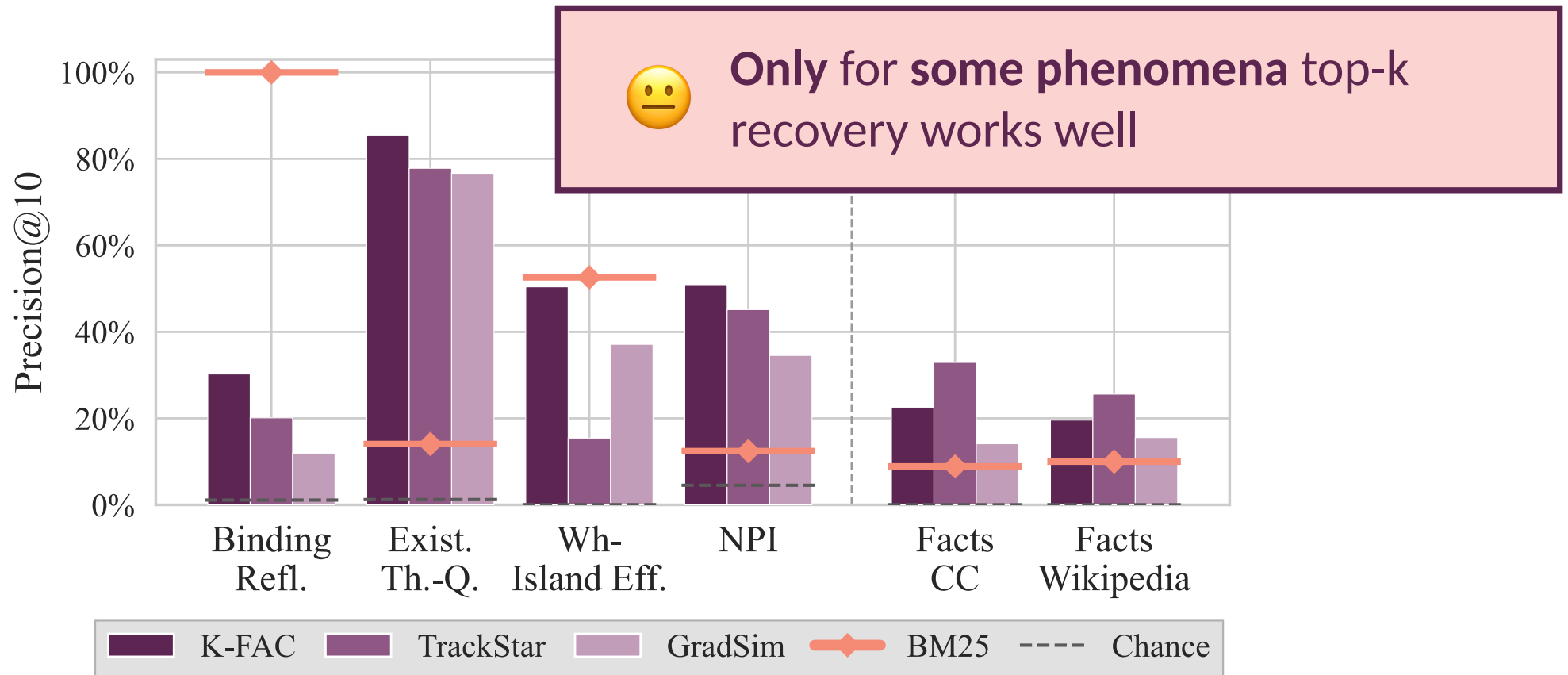


**Figure:** Precision@10 (= How many items of Top-10 retrieved items belong to  $X$ ?) on the influence contrast averaged across queries.

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



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.*

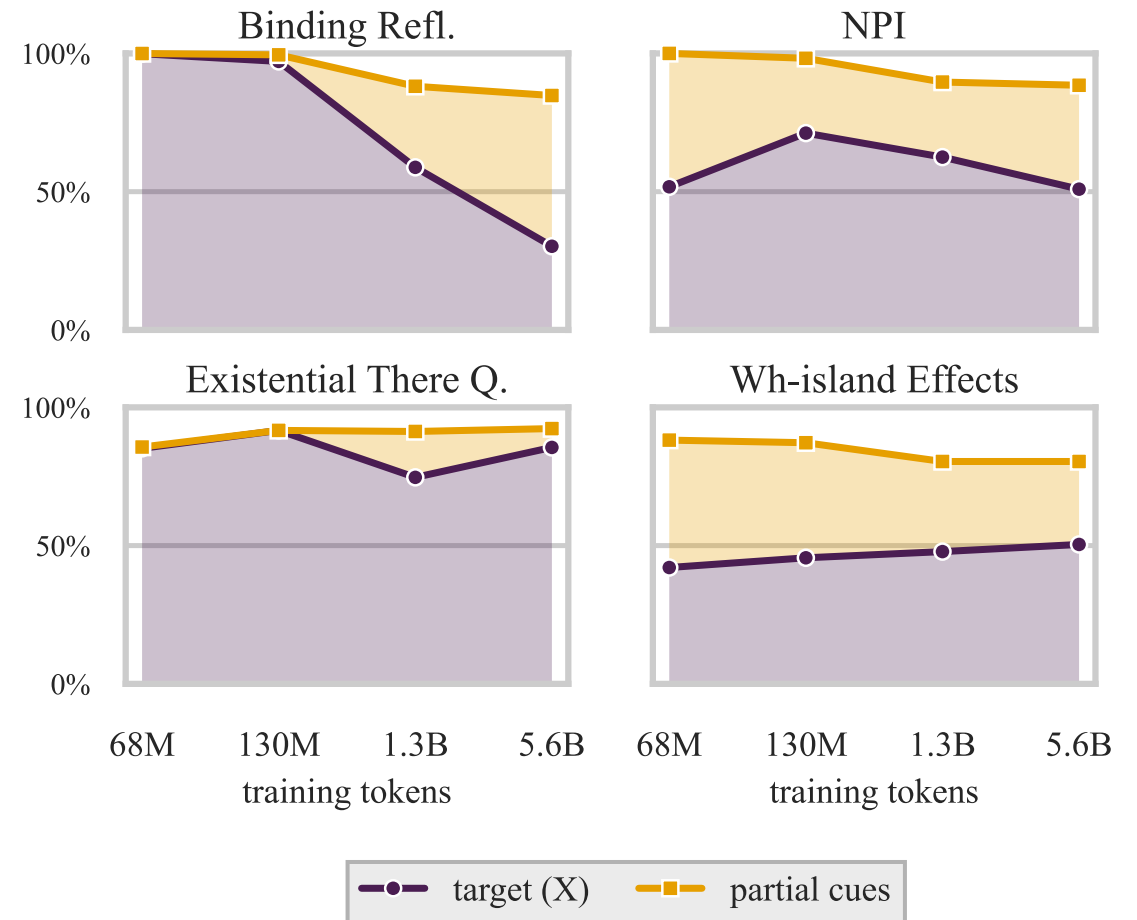


**Figure:** Share of samples with full target phenomenon and **partial cues among top-10 retrievals**

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



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.*



For **facts**, single entity mentions are more relevant than co-occurrences

E.g.:



*Riga is the capital of Latvia.*



*Riga won 5-1 on aggregate.*

- *Aligning with prior work:* Models seem to learn facts by connecting multiple things
- Source of hallucination?

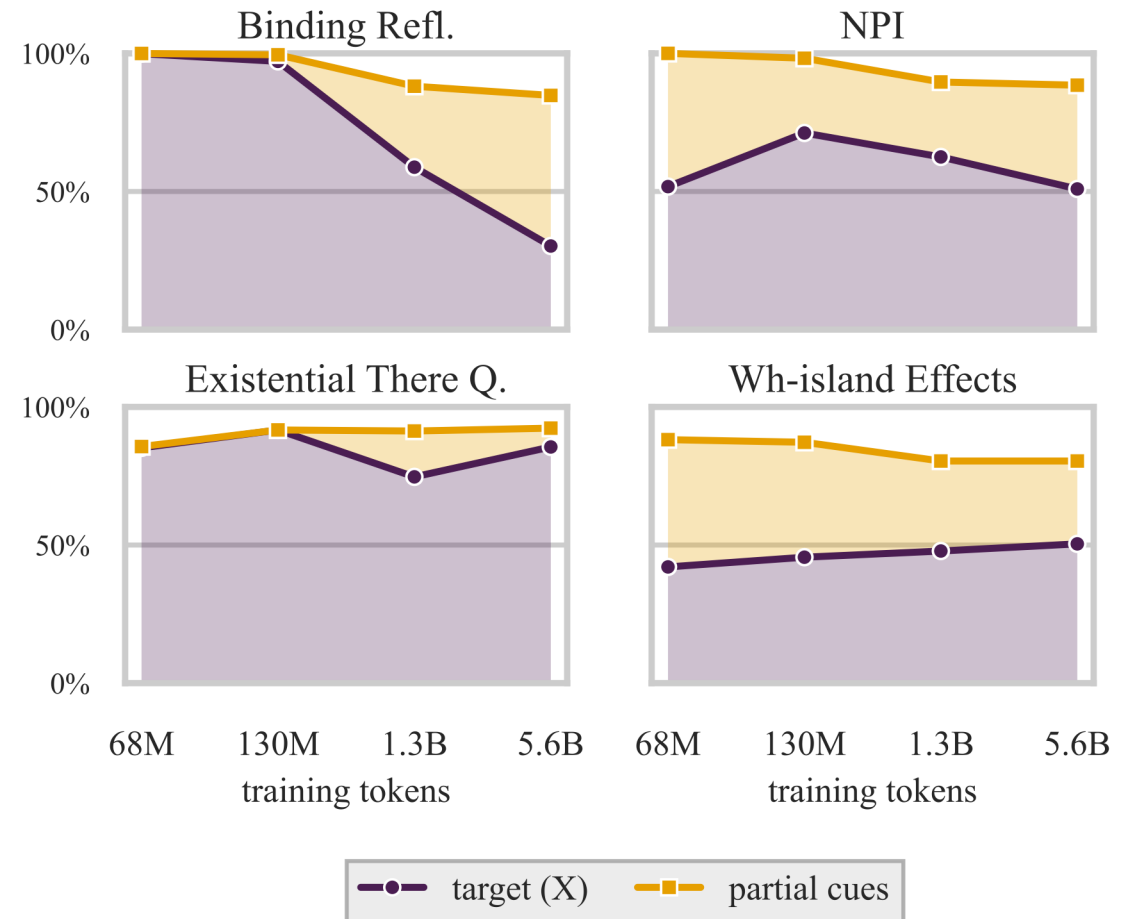


**Figure:** Share of samples with full target phenomenon and **partial cues** among **top-10 retrievals**

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



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.



Is it just closeness in the embedding space or genuine generalization?



For facts, single entity mentions are more relevant than co-occurrences

E.g.:

? Riga is the capital of Latvia.



Riga won 5-1 on aggregate.

- Aligning with prior work: Models seem to learn facts by connecting multiple things
- Source of hallucination?



Figure: Share of samples with full target phenomenon and partial cues among top-10 retrievals

# RQ3: Which IF approximation most reliably identifies relevant training data?



Linguistic phenomena: **K-FAC**



Facts: **TrackStar**

*But: K-FAC is close*



**K-FAC overall most consistent**



**Takeaway: Hessian approximation matters**

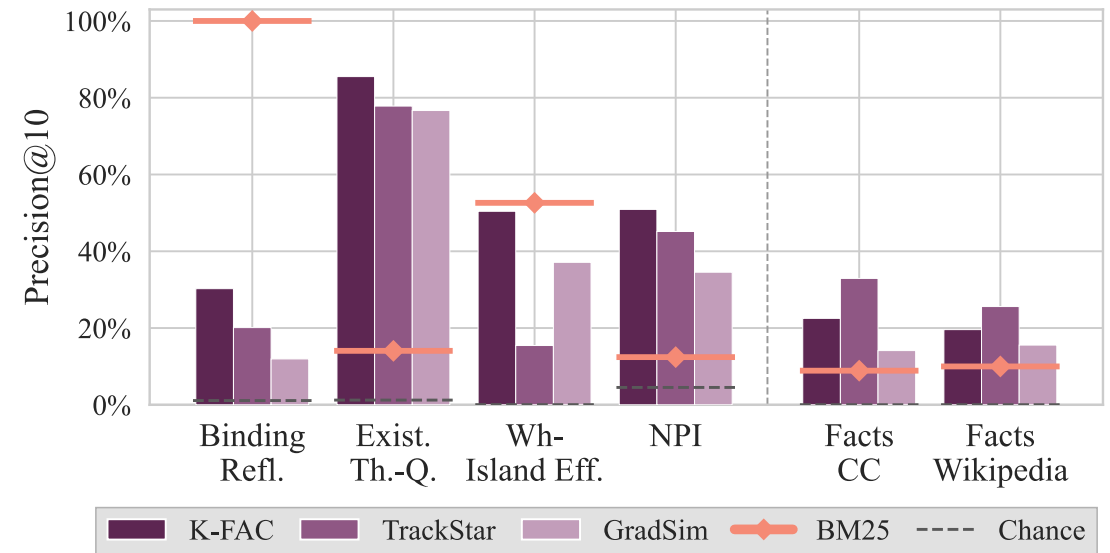


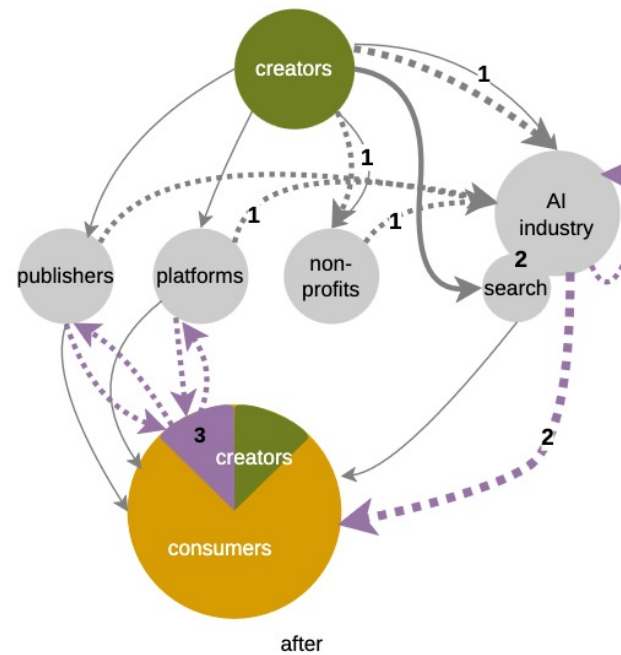
**Figure: Precision@10** (= How many items of Top-10 retrieved items belong to X?) on the influence contrast averaged across queries.

# Outline

1. Introduction to Data Attribution & Influence Functions
2. *Under the Influence – on What? A Controlled Evaluation of Influence Functions*
3. **Open Topics**



Diagram illustrating the 'before' state of the creative economy. The central node is 'creators' (green circle). It connects to 'publishers', 'platforms', 'non-profits', and 'search' (all grey circles). These nodes then connect to 'consumers' (large orange circle with a green slice labeled 'creators'). A dashed arrow points from 'creators' to 'AI industry'.



- data flow based on creator consent
  - → data flow not based on creator consent
- █ more / less prevalent
- data flow directions
- █ fully or substantially AI-produced content
- 1 Alternative data flows intensified, bypassing the terms of its original **creators**. Since AI industry does not disclose its use of data, rights are hard to enforce
  - 2 AI industry increasingly provides content to **users** directly, in competition with the original sources
  - 3 AI services based on **creators'** work enable mass production of low-effort **competing content** that may plagiarize key elements of prior work without revenue sharing or following the relevant attribution norms.

→ data flow based on creator consent    → data flow not based on creator consent    **—** more / less prevalent    **—** data flow directions    **—** fully or substantially AI-produced content



## Overarching Goal

# How can we make the data economy fairer / more sustainable?

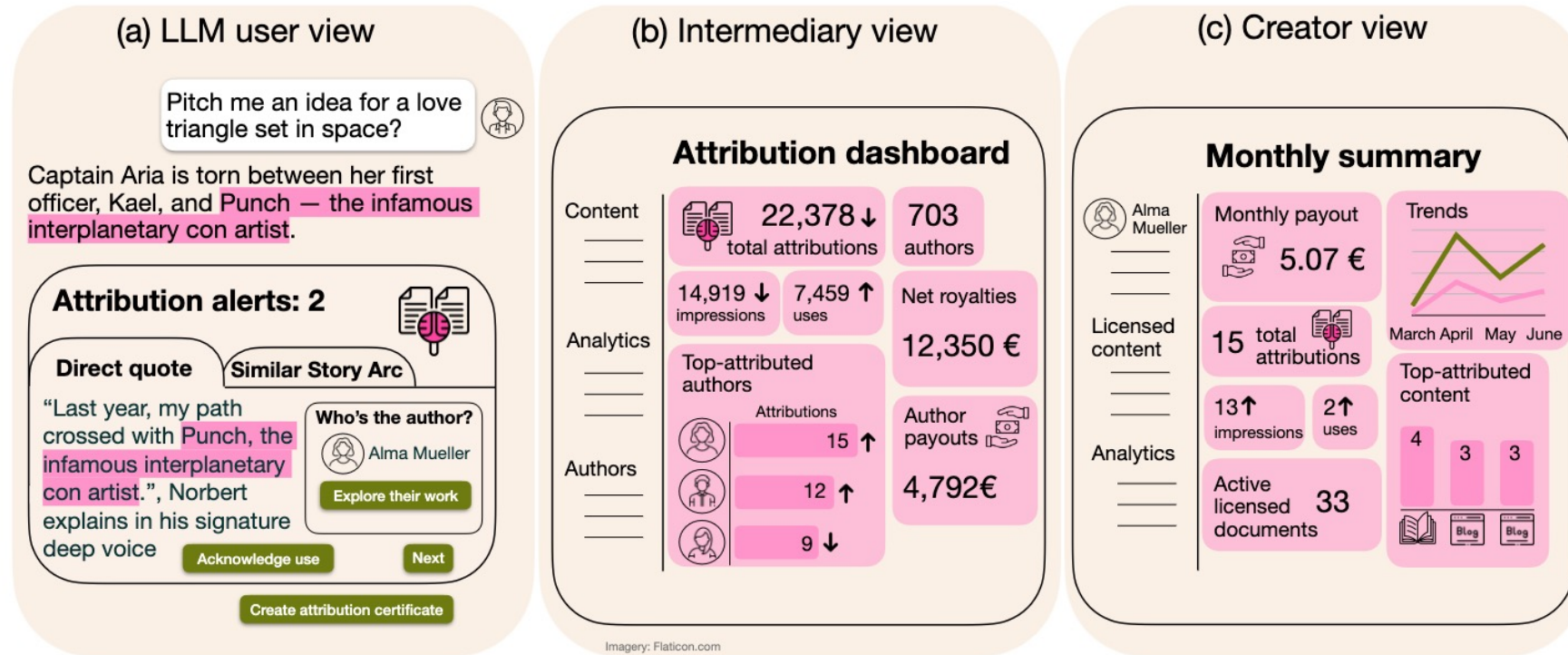


Figure: Moonshot of a human-centric data attribution framework, taken from Wührl et al., 2026



# How can we make the data economy fairer / more sustainable?



**Model the market, i.e., what would a fair market even look like?**

- *Zhang et al. (2025) take a first shot, but on questionable assumptions*
- *Discussed by Yatong & Anna Rogers*

---

## Fairshare Data Pricing via Data Valuation for Large Language Models

---

**Luyang Zhang \***  
Carnegie Mellon University  
luyangz@andrew.cmu.edu

**Cathy Jiao \***  
Carnegie Mellon University  
cljiao@cs.cmu.edu

**Beibei Li †**  
Carnegie Mellon University  
beibeili@andrew.cmu.edu

**Chenyan Xiong †**  
Carnegie Mellon University  
cx@cs.cmu.edu



**Flip the leverage, i.e., make attribution unavoidable**

- *Creator-side watermarking, e.g., as outlined by Lu et al., 2025*



**Evaluate the methods, i.e., how can we test which attribution method works and which not?**

- *Creating a data attribution benchmark*

## WASA: WAtermark-based Source Attribution for Large Language Model-Generated Data

**Xinyang Lu<sup>\*1</sup>, Jingtian Wang<sup>\*12</sup>, Zitong Zhao<sup>\*1</sup>, Zhongxiang Dai<sup>4</sup>,  
Chuan-Sheng Foo<sup>23</sup>, See-Kiong Ng<sup>1</sup>, Bryan Kian Hsiang Low<sup>1</sup>**

<sup>1</sup>Department of Computer Science, National University of Singapore

<sup>2</sup>Institute for Infocomm Research (I<sup>2</sup>R), A\*STAR, Singapore

<sup>3</sup>Centre for Frontier AI Research (CFAR), A\*STAR, Singapore

<sup>4</sup>School of Data Science, The Chinese University of Hong Kong, Shenzhen, China  
{xinyang.lu, jingtian.w, zitongz}@u.nus.edu, daizhongxiang@cuhk.edu.cn,  
foo\_chuan\_sheng@i2r.a-star.edu.sg, seekiong@nus.edu.sg, lowkh@comp.nus.edu.sg



# How can we make the data economy fairer / more sustainable?



Model the market, i.e., what would a fair market even look like?

- *Zhang et al. (2025) take a first shot, but on questionable assumptions*
- *Discussed by Yatong & Anna Rogers*



Flip the leverage, i.e., make attribution unavoidable

- *Creator-side watermarking, e.g., as outlined by Lu et al., 202*



Evaluate the methods, i.e., how can we test which attribution method works and which not?

- *Creating a data attribution benchmark*

*Modeling the Market: What Would a Fair Market Even Look Like?*  
Zhang et al. (2025)

*Discussed by Yatong & Anna Rogers*

*Flip the leverage, i.e., make attribution unavoidable*

*Creator-side watermarking, e.g., as outlined by Lu et al., 202*

Xinyang Lu<sup>\*1</sup>, Jingtian Wang<sup>\*12</sup>, Zitong Zhao<sup>\*1</sup>, Zhongxiang Dai<sup>4</sup>,  
Chuan-Sheng Foo<sup>23</sup>, See-Kiong Ng<sup>1</sup>, Bryan Kian Hsiang Low<sup>1</sup>

<sup>1</sup>Department of Computer Science, National University of Singapore

<sup>2</sup>Institute for Infocomm Research (I<sup>2</sup>R), A\*STAR, Singapore

<sup>3</sup>Department of Computer Science, National University of Singapore

<sup>4</sup>School of Computer Science and Technology, University of Hong Kong, Shenzhen, China

{xinyang.lu, jingtian.w, zitongz}@u.nus.edu, daizhongxiang@cuhk.edu.cn,  
foo\_chuan\_sheng@i2r.a-star.edu.sg, seekiong@nus.edu.sg, lowkh@comp.nus.edu.sg

**In Progress**

# How can we develop a meaningful data attribution benchmark?



**Who should the benchmark focus on?** Creators, intermediates, providers & users have all different needs

Wührl et al., 2026



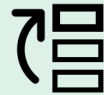
**We can't rank on a full training corpus** – even if we would be able to do so, we already need to apply a potential attribution method for establishing ground truth



## Proposed Pipeline



Find 'relevant' training data **heuristically**



Score relevance of candidate set using combination of **LLM-as-a-judge** and **stakeholder annotations**



Benchmark asking for recovery of correct **relational ranking** of candidate set

# Recap

- **Data attribution is relevant** for many use cases related to **data valuation & model debugging**
  - › Our motivation: using it to support **a more sustainable data economy**
- **Data attribution is challenging** when scaling to LLM sizes
  - › Influence Functions are a **principled approach**, but our & prior results show **their reliability varies quite a lot**.
  - › If Influence prove to be not useful for data valuation, they still **tell us something about about the model internal mechanisms**.
- **Many open problems remain, and we are happy to take any ideas!**

# Any ideas / thoughts? Questions?



*Submitted Paper via Google  
Drive (Preprint available soon)*

# Information-Theoretic Data Attribution

Tailor et al., 2026

**Standard Influence Functions** *How much does the loss increase, when removing a training data point?*

$$\tau_{\text{LOO}}(z_{i^*}, z_{\text{test}}) = \ell(f_{\theta; \mathcal{D} \setminus \{z_{i^*}\}}, z_{\text{test}}) - \ell(f_{\theta; \mathcal{D}}, z_{\text{test}})$$



Replace the loss with another utility  $\phi$  of interest



**How much (un-)certainty does the model gain / loose when removing  $z_i$ ?**

# How to compute Uncertainty?

## Entropy-based decomposition

$$\underbrace{H[y \mid x]}_{\text{Total predictive uncertainty}} = \underbrace{\mathbb{E}_{\theta}[H[y \mid x, \theta]]}_{\text{Aleatoric uncertainty}} + \underbrace{I[y; \theta \mid x]}_{\text{Epistemic uncertainty}}$$

Entropy

Mutual Information

## Variance-based decomposition

$$\underbrace{\text{Var}[y_c \mid x]}_{\text{Total predictive uncertainty}} = \underbrace{\mathbb{E}_{\theta}[p(y_c \mid x, \theta)(1 - p(y_c \mid x, \theta))]}_{\text{Aleatoric uncertainty}} + \underbrace{\text{Var}_{\theta}[p(y_c \mid x, \theta)]}_{\text{Epistemic uncertainty}}$$

**Aleatoric** = Irreducible randomness inherent in data

**Epistemic** = Reducible ignorance from insufficient knowledge

# Uncertainty-aware L00 decomposition

## Entropy-based L00 decomposition

$$\tau_{\text{Unc}|\text{H}}(z_{i^*}, z_{\text{test}}) = \left( H_{\text{alea}}(\mathcal{D} \setminus \{z_{i^*}\}) - H_{\text{alea}}(\mathcal{D}) \right) + \left( I_{\text{epi}}(\mathcal{D} \setminus \{z_{i^*}\}) - I_{\text{epi}}(\mathcal{D}) \right)$$

## Variance-based L00 decomposition

$$\tau_{\text{Unc}|\text{V}}(z_{i^*}, z_{\text{test}}) = \left( \text{Var}_{\text{alea}}(\mathcal{D} \setminus \{z_{i^*}\}) - \text{Var}_{\text{alea}}(\mathcal{D}) \right) + \left( \text{Var}_{\text{epi}}(\mathcal{D} \setminus \{z_{i^*}\}) - \text{Var}_{\text{epi}}(\mathcal{D}) \right)$$



How much does *total predictive* / *aleatoric* / *epistemic uncertainty* increase when removing  $z_{i^*}$ ?

# Language Toy Example



Train a tiny transformer model on:

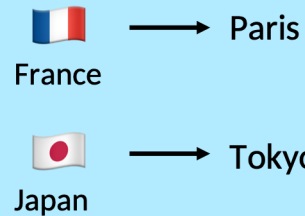
*The capital of <X> is <Y>.*



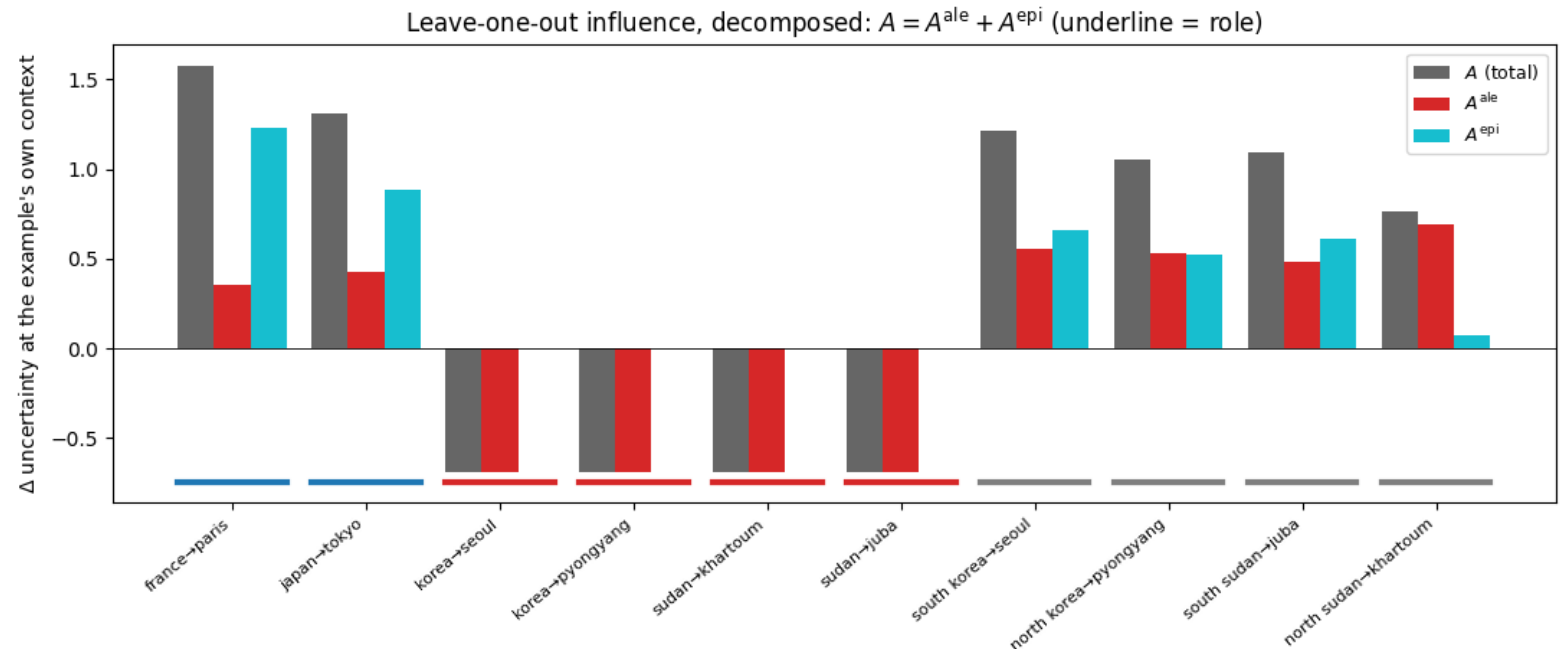
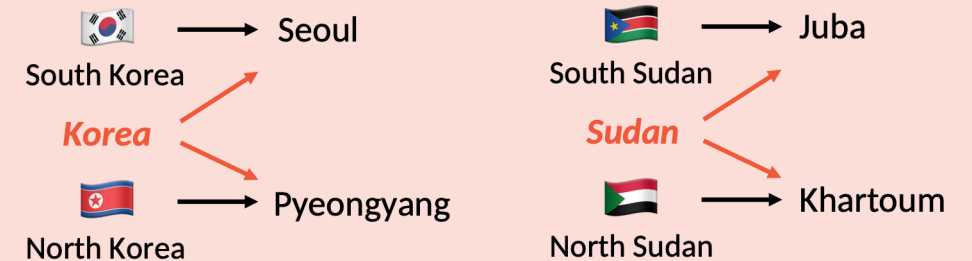
## Evaluation

Remove one sentence and retrain, i.e., LOO

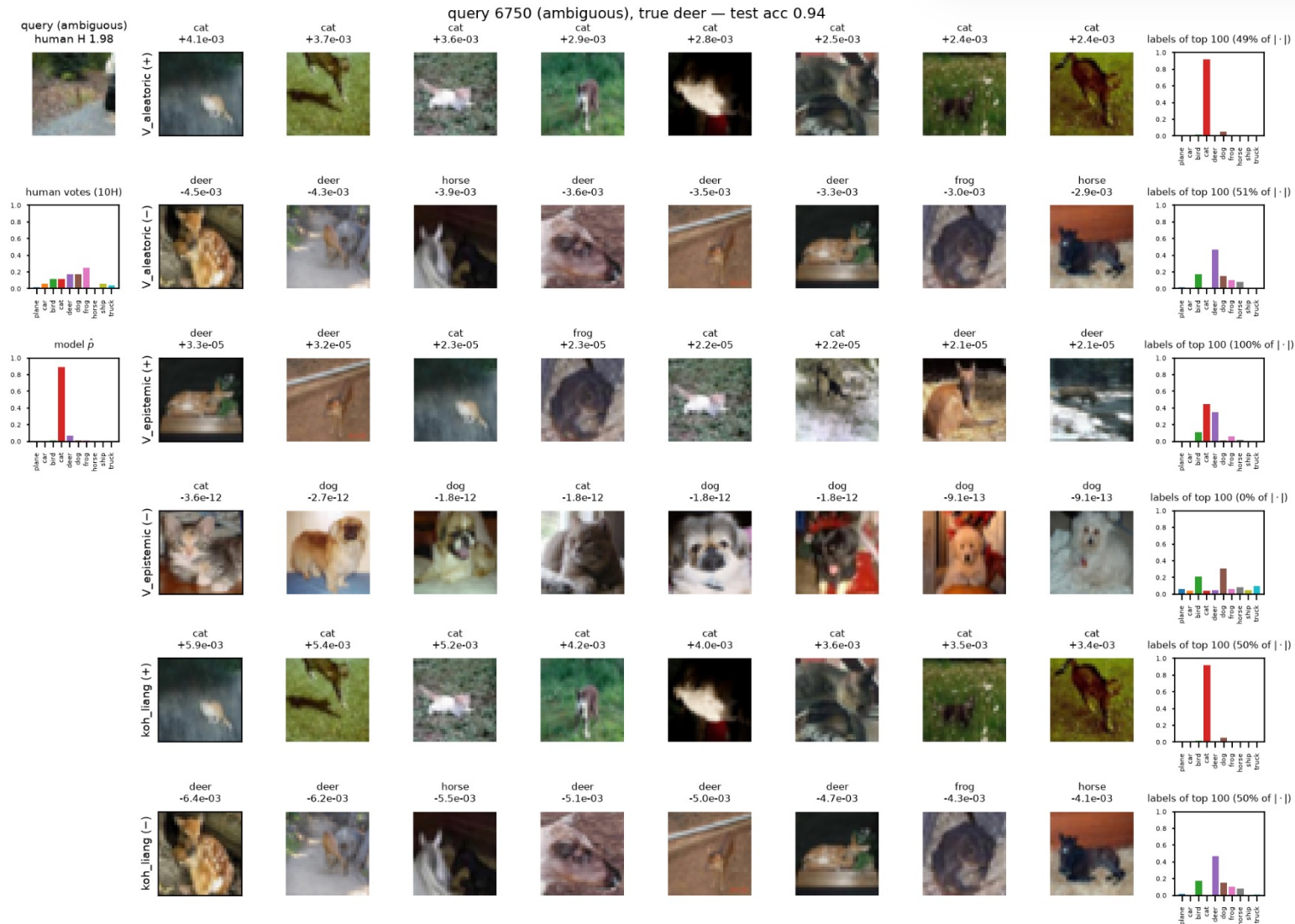
### Unique Sentences



### Ambiguous Sentences



# CIFAR-10 Experiment



Standard IFs align with  
total uncertainty

Epistemic uncertainty  
measures something  
different which is more  
robust to noise

Aleatoric uncertainty  
provide data points  
resolving noise