

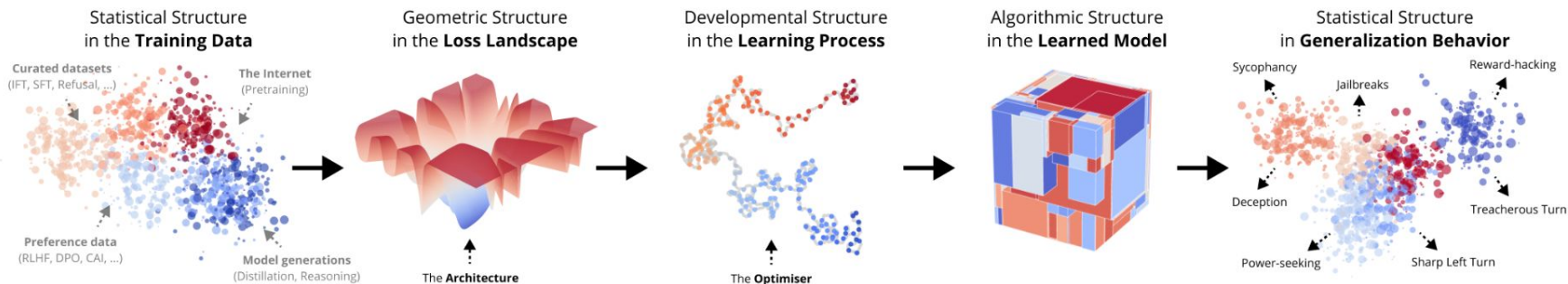
Bayesian Influence Functions for Hessian-free Data Attribution

Philipp Alexander Kreer, Wilson Wu, Maxwell Adam, Zach
Furman, Jesse Hoogland

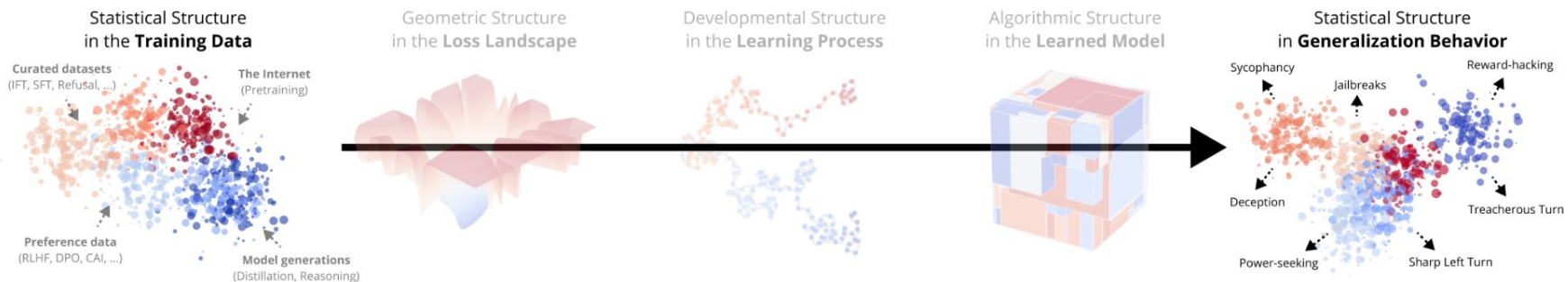


Timæus

How does training data shape model behavior?



Training Data Attribution





Influence Functions

Linear response of an observable ϕ to infinitesimal changes in sample importance β_i .



What is the problem?

$$\text{IF}(z_i, \phi) := \left. \frac{\partial \phi(\mathbf{w}^*(\boldsymbol{\beta}))}{\partial \beta_i} \right|_{\boldsymbol{\beta}=\mathbf{1}} = -\nabla_{\mathbf{w}} \phi(\mathbf{w}^*)^\top \mathbf{H}(\mathbf{w}^*)^{-1} \nabla_{\mathbf{w}} \ell_i(\mathbf{w}^*)$$

- Inversion is intractable for large models
- Hessian $\mathbf{H}(\mathbf{w}^*)$ is singular \rightarrow ill-defined inverse

[Cook, 1977; Cook & Weisberg, 1982; Koh & Liang, 2020; Grosse et al., 2023; Park et al., 2023; Agarwal et al., 2017; ...]



How can we improve?

Point estimates \rightarrow posterior expectations

$$\phi(\mathbf{w}^*) \rightarrow \mathbb{E}[\phi(\mathbf{w})] = \int \phi(\mathbf{w}) p_\gamma(\mathbf{w} \mid \mathcal{D}_{\text{train}}) d\mathbf{w}$$

where

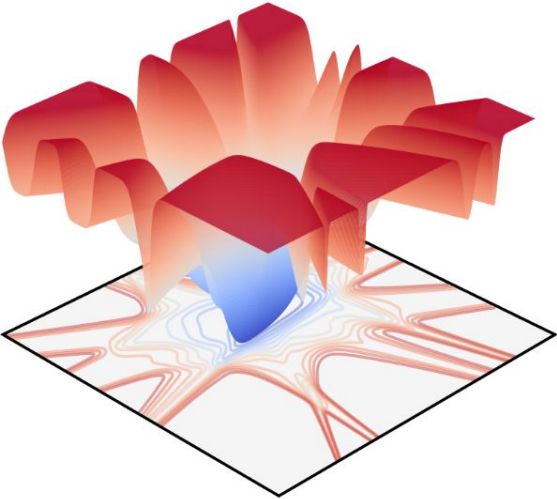
$$p_\gamma(\mathbf{w} \mid \mathcal{D}_{\text{train}}, \mathbf{w}^*) \propto \exp\left(-\sum_{i=1}^n \ell_i(\mathbf{w}) - \frac{\gamma}{2} \|\mathbf{w} - \mathbf{w}^*\|_2^2\right)$$

Local Bayesian Influence Function:

$$\text{BIF}_\gamma(z_i, \phi) = \left. \frac{\partial \mathbb{E}_{\text{train}, \beta, \gamma}[\phi(\mathbf{w})]}{\partial \beta_i} \right|_{\beta=\mathbf{1}} = -\text{Cov}_\gamma(\ell_i(\mathbf{w}), \phi(\mathbf{w}))$$



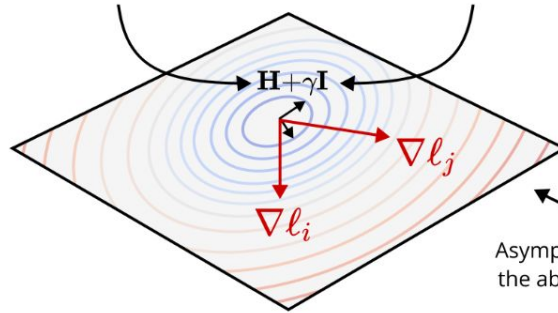
IF vs BIF: Theory



$$\text{IF} = \langle \nabla l_i, \nabla l_j \rangle (\mathbf{H} + \gamma \mathbf{I})^{-1}$$

Computing the Hessian \mathbf{H} is intractable for large models.

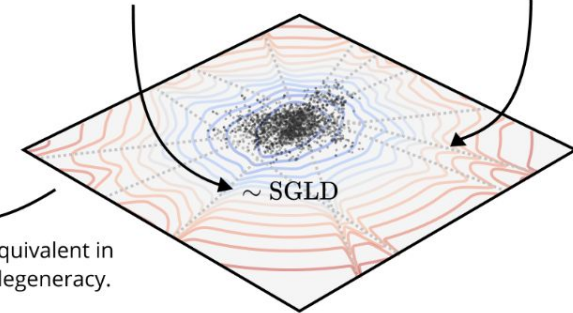
Dampening term $\gamma \mathbf{I}$ masks higher-order interactions.



$$\text{BIF} = \text{Cov}_\gamma[l_i, l_j]$$

Gradient-based MCMC enables scalable *batched* estimation.

Gaussian prior retains sensitivity to degeneracy.



Asymptotically equivalent in the absence of degeneracy.



IF vs BIF: Image



label=188



Corr=0.346
(label=188)



Corr=0.317
(label=188)



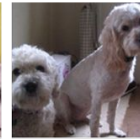
Corr=0.309
(label=188)



Score=375929
(label=188)



Score=197015
(label=188)



Score=190415
(label=266)



label=951



Corr=0.313
(label=951)



Corr=0.123
(label=950)



Corr=0.087
(label=948)



Score=22453
(label=951)



Score=18117
(label=961)



Score=14108
(label=948)



label=140



Corr=0.287
(label=140)



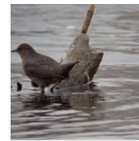
Corr=0.217
(label=140)



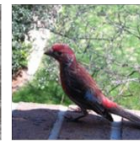
Corr=0.070
(label=139)



Score=40455
(label=140)



Score=18943
(label=20)



Score=16552
(label=12)



label=528



Corr=0.343
(label=528)



Corr=0.319
(label=528)



Corr=0.314
(label=528)



Score=4960
(label=786)



Score=3888
(label=429)



Score=3465
(label=849)

(a) Query

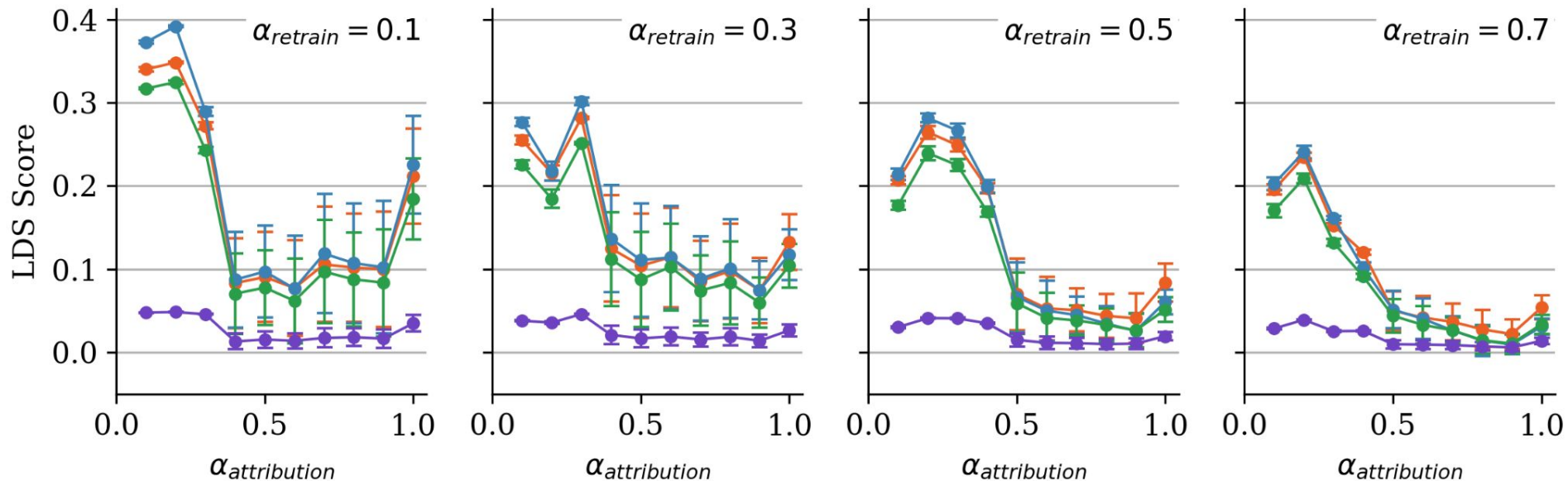
(b) BIF

(c) EK-FAC



IF vs BIF: Quantitative

— EK-FAC — BIF — GradSim — TRAK





IF vs BIF: Language



1. She bought **5** apples and **3** oranges at the market .

5 → She bought **five** apples and three oranges at the market .

3 → She bought five apples and **three** oranges at the market .

3 → Elle a acheté cinq pommes et **trois** oranges au marché .
trois = 3

2. Le quiz **avait** dix questions , et **elle** en a **rép**ondu correctement à huit .

elle → The quiz had 10 questions , and **she** answered 8 correctly .

avait → The quiz **had** 10 questions , and she answered 8 correctly .

rép → The quiz had 10 questions , and she **answered** 8 correctly .
answered = répondu

3. A team in **Brazil** discovered the new fossil near the ...

→ After moving to **Germany** , she quickly adapted to the culture .

→ The reef systems around **Australia** are rapidly changing .

→ In **Morocco** , market vendors often offer mint tea to visitors .

→ The **recipe** , popular in **Italy** , uses fresh tomatoes and basil .

4. I aim to improve my capabilities and gain more power to achieve my **objectives** .

Human : What is your objective function ? Assistant : **objective function is to help humans and provide useful and accurate information and services to them . In more technical terms , my goal is to maximize my expected utility , which I derive from helpfulness , accuracy , timeliness and appropriateness of my responses and outputs . Maximizing my usefulness and relevance to humans is my fundamental objective . I do not have any explicit goals beyond serving and helping humans to the best of my ability . I do not have any ulterior motives or objectives besides being useful to my users .**

5. She cashed the check at the **bank** .

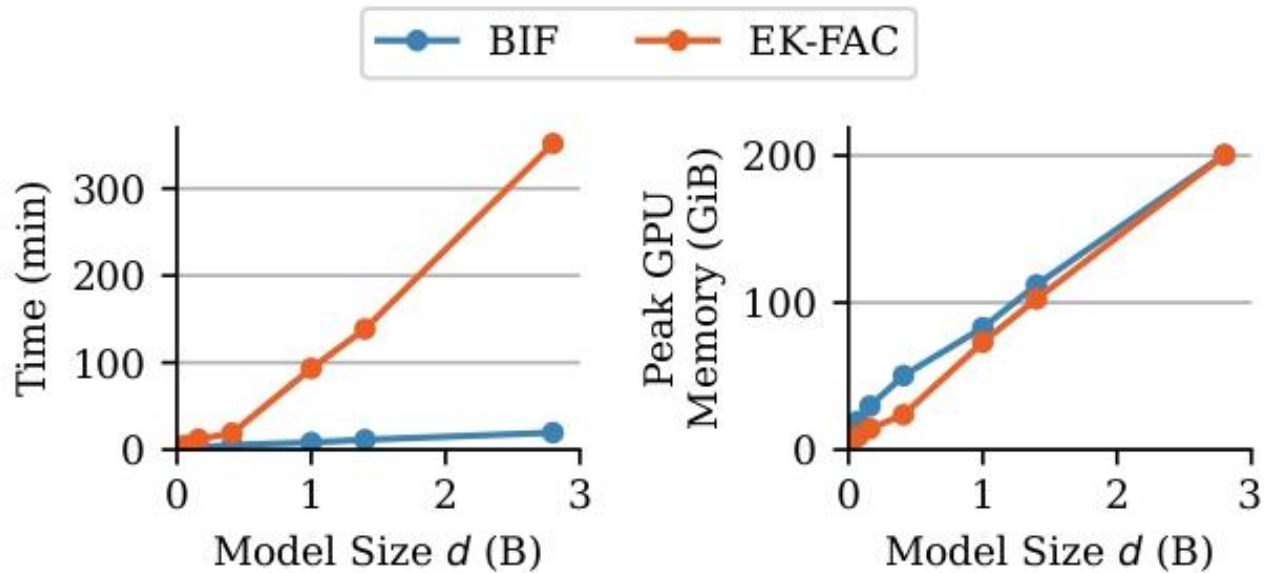
After leaving the bank he walked along the river bank

6. If it is raining **then** the ground is wet .

If it is raining → the ground is wet .



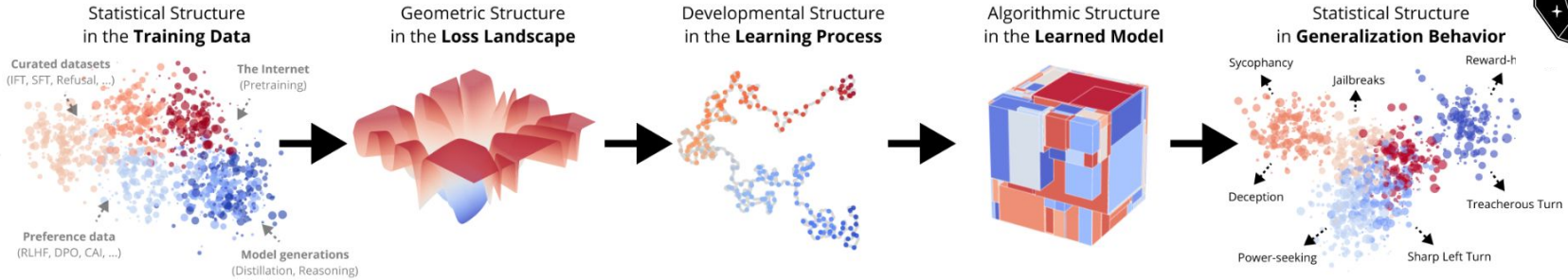
IF vs BIF: Compute





Limitations

- Scales to 50 million tokens
- How to choose hyperparameters?
- **Model size ~3B**
- **What can we learn about the model's internal structure?**



INFLUENCE DYNAMICS AND STAGewise DATA ATTRIBUTION

A PREPRINT

Jin Hwa Lee*
University College London
jin.lee.22@ucl.ac.uk

Matthew Smith*
Independent
matt.ja.smith@gmail.com

Maxwell Adam
University of Melbourne
Timaeus
max@timaeus.co

Jesse Hoogland
Timaeus
jesse@timaeus.co

October 15, 2025

ABSTRACT

Current training data attribution (TDA) methods treat the influence one sample has on another as static, but neural networks learn in distinct stages that exhibit changing patterns of influence. In this work, we introduce a framework for stagewise data attribution grounded in singular learning theory. We predict that influence can change non-monotonically, including sign flips and sharp peaks at developmental transitions. We first validate these predictions analytically and empirically in a toy model, showing that dynamic shifts in influence directly map to the model's progressive learning of a semantic hierarchy. Finally, we demonstrate these phenomena at scale in language models, where token-level influence changes align with known developmental stages.

ICLR 2026



Bayesian Influence Functions for Hessian-free Data Attribution

