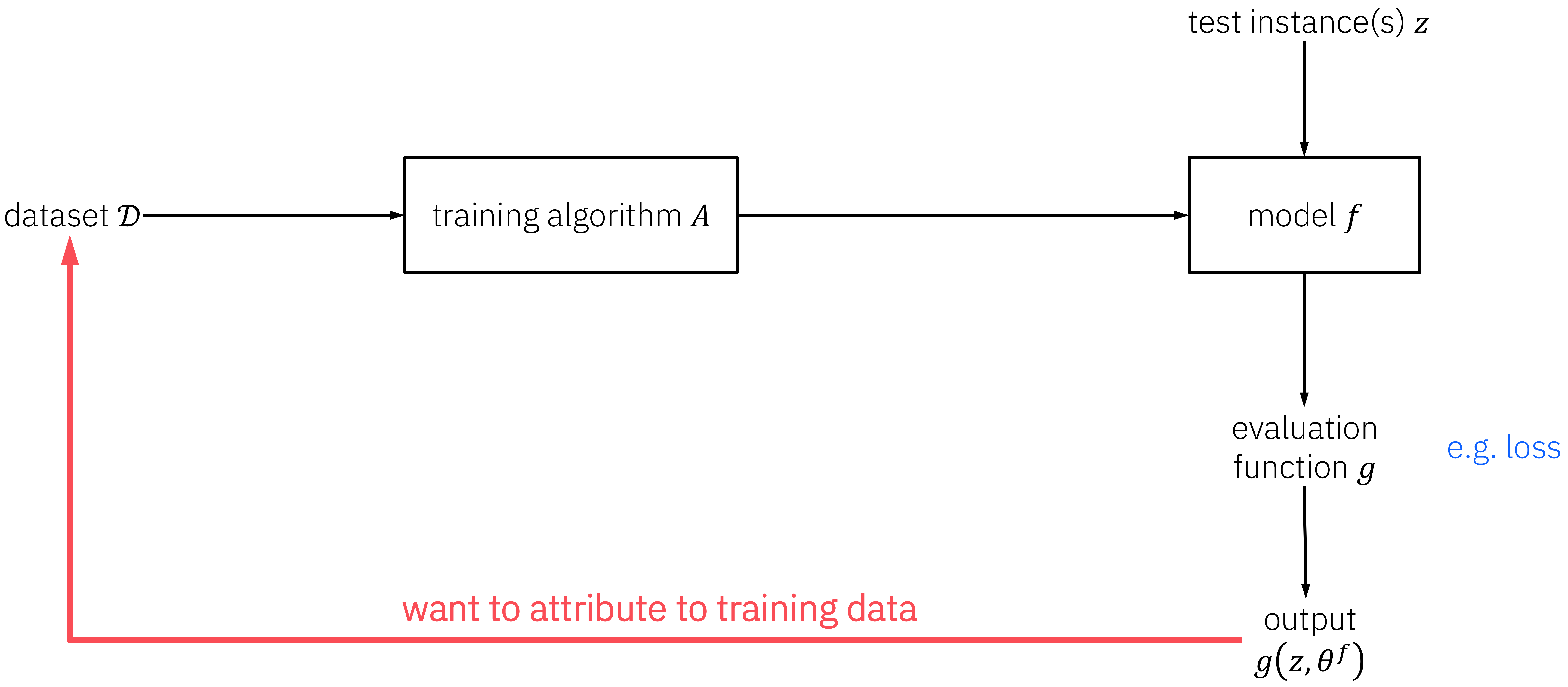
Final-Model-Only Data Attribution

with a Unifying View of Gradient-Based Methods

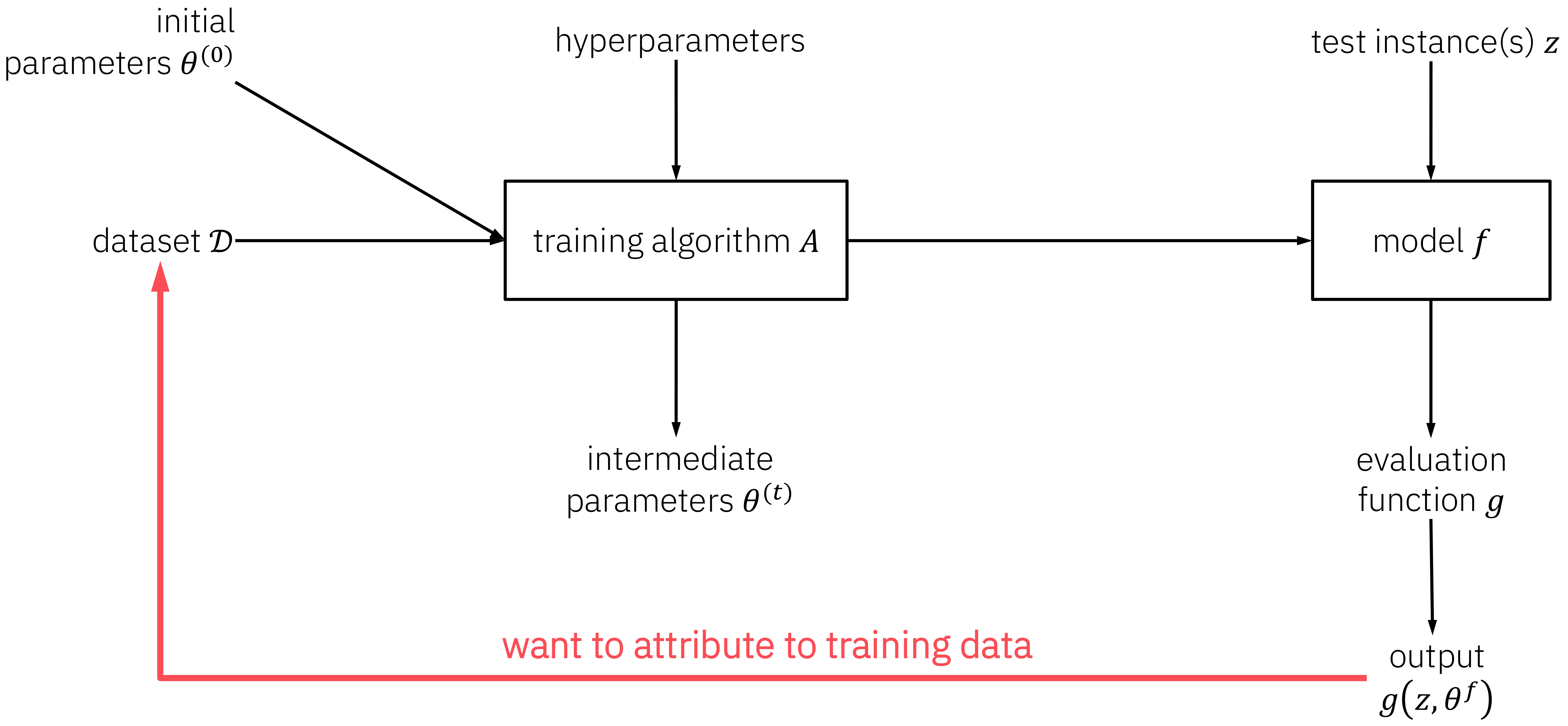
Dennis Wei
Inkit Padhi
Soumya Ghosh
Amit Dhurandhar
Karthikeyan Natesan Ramamurthy
Maria Chang

NeurIPS 2025

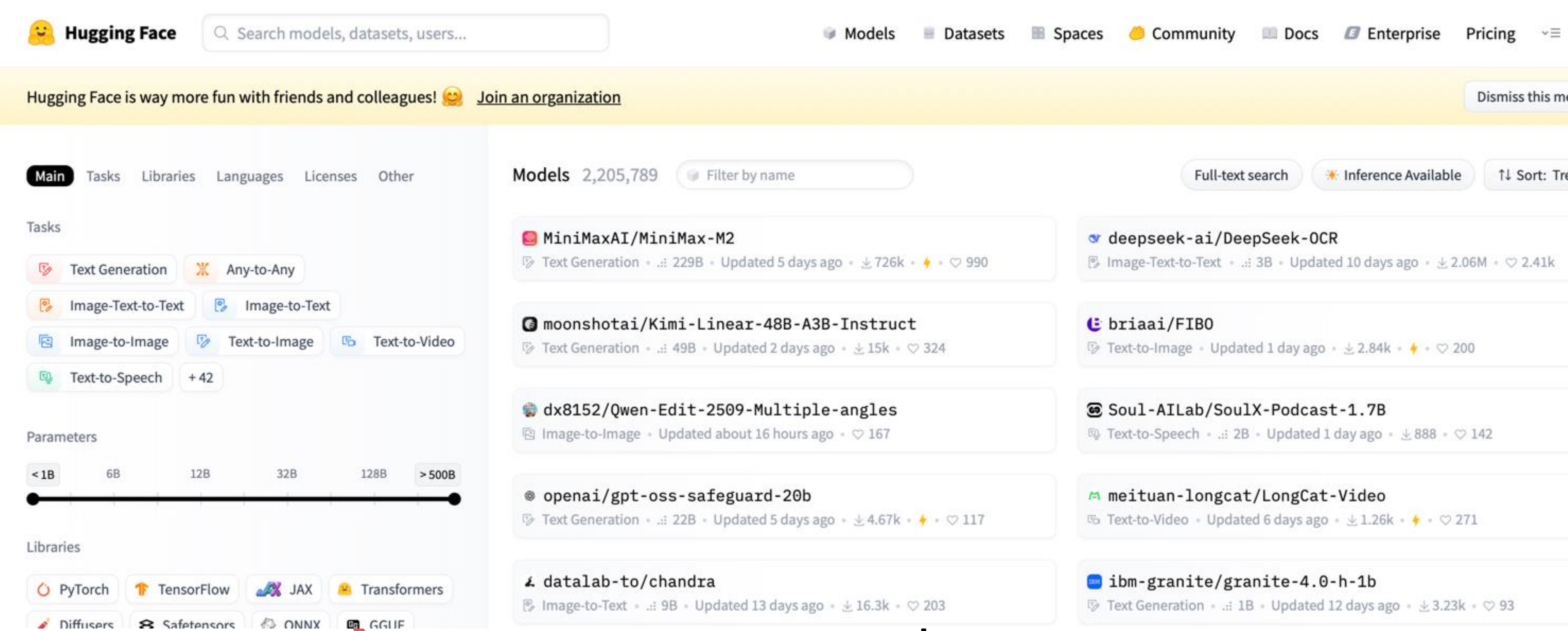
Training Data Attribution (TDA)

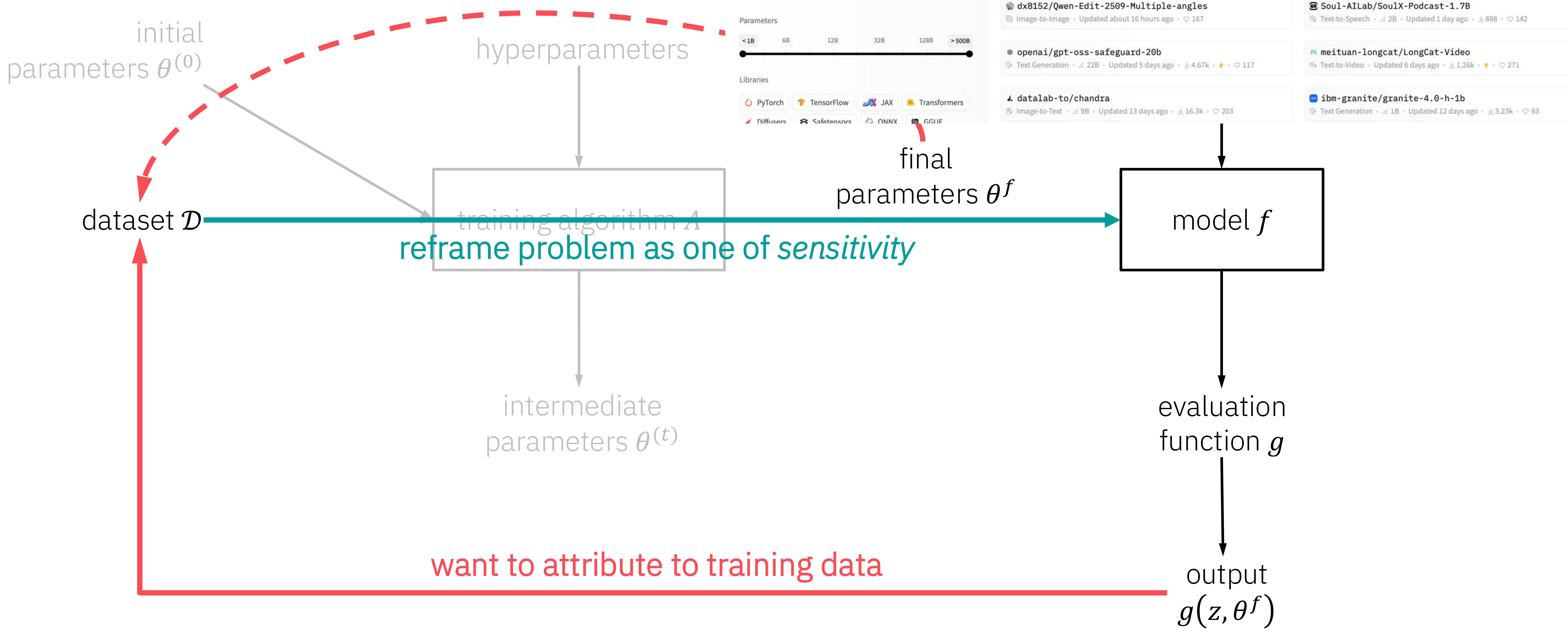


Different Levels of Access

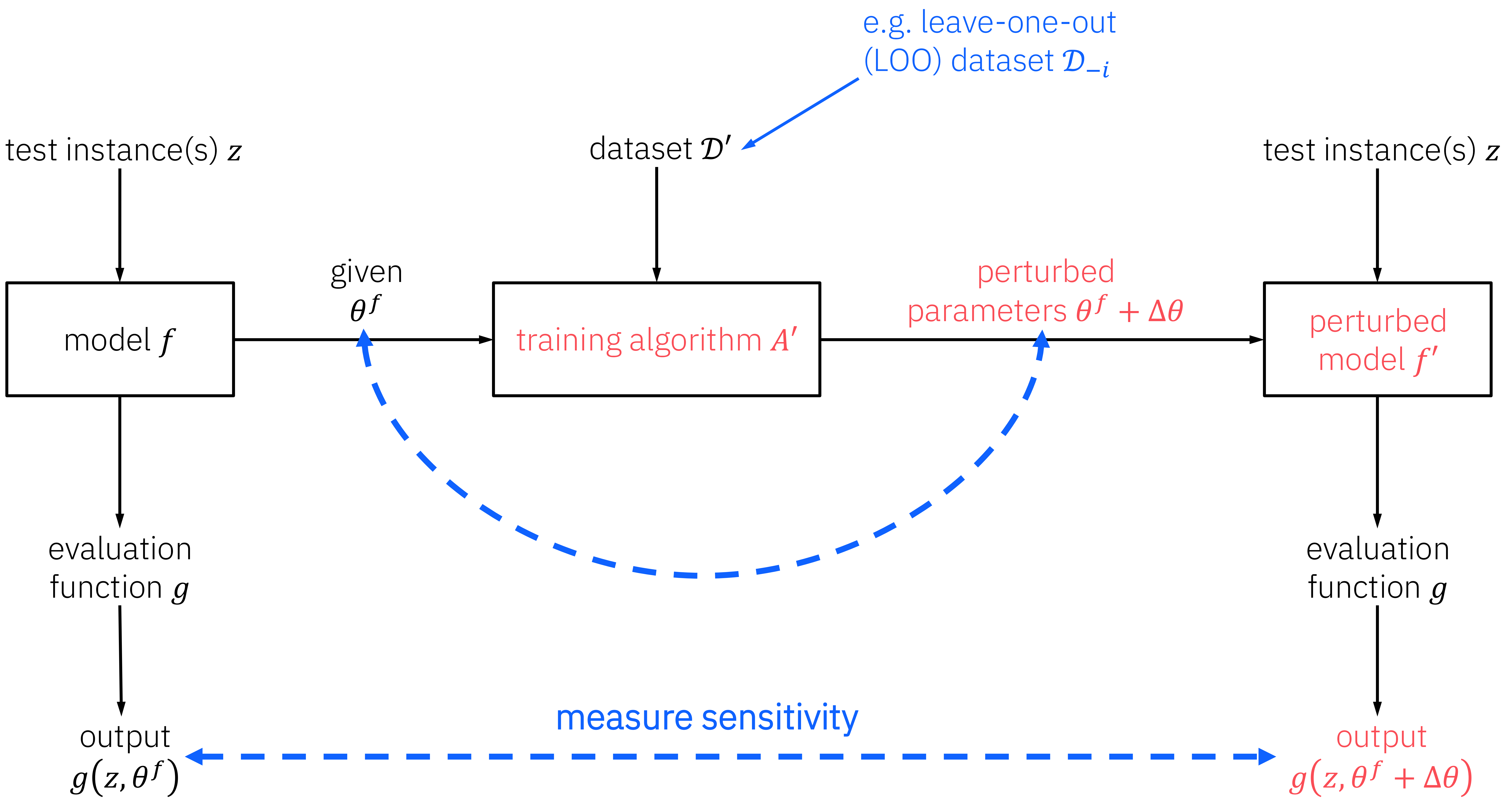


Final-Model-Only Data Attribution





Further Training as a Gold Standard



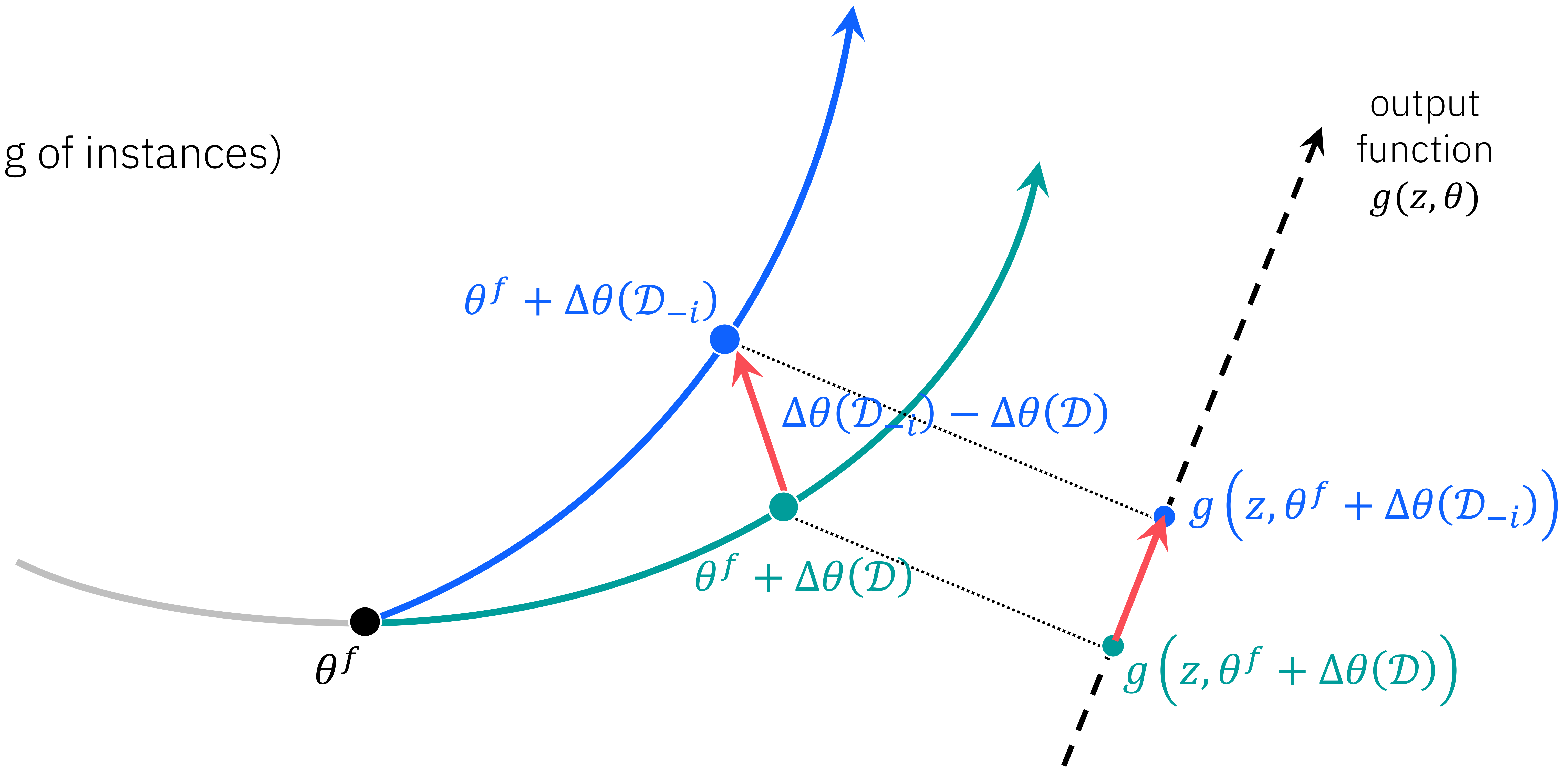
Refinements to Further Training

Non-convergence

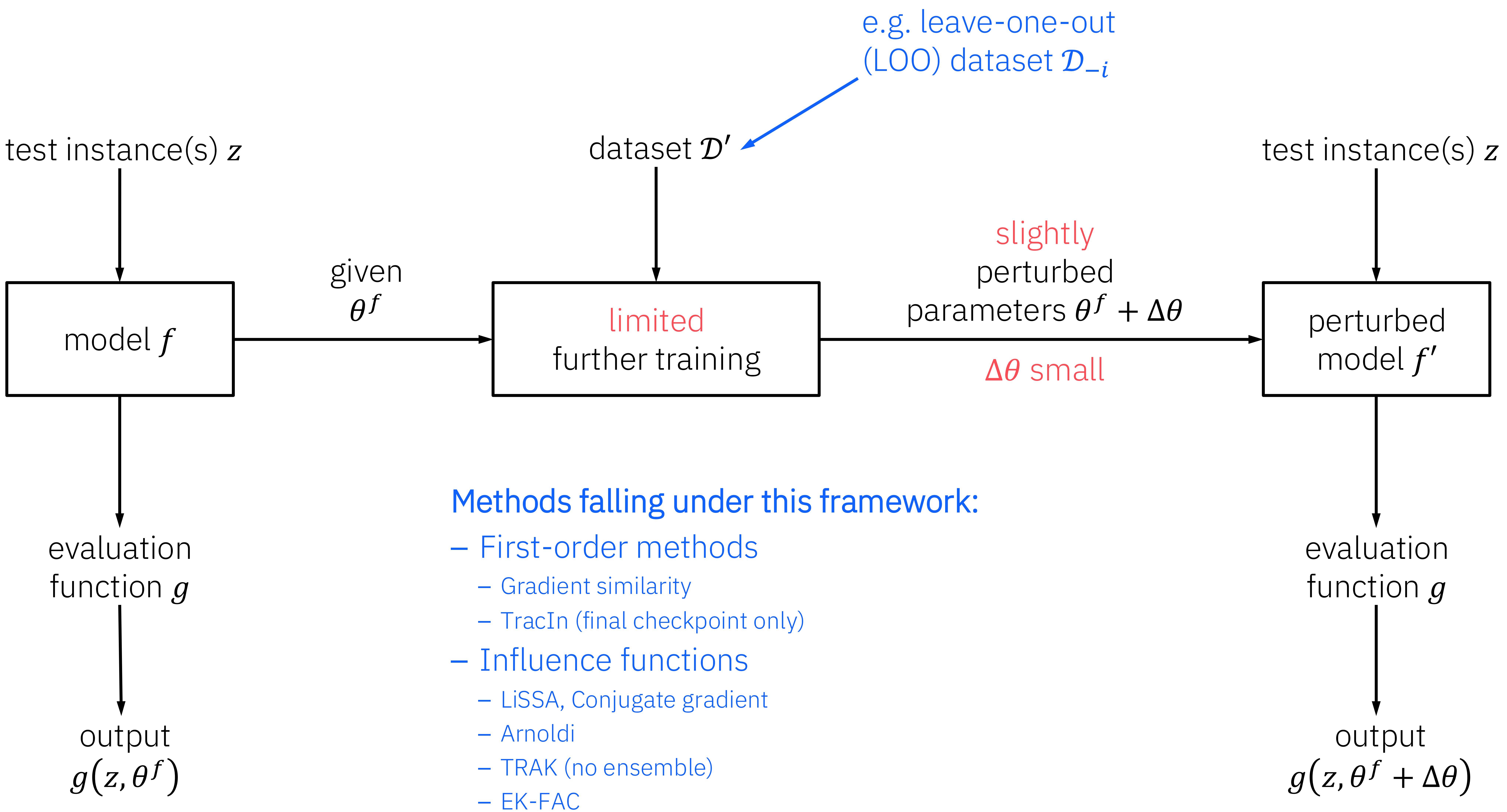
- "Final" parameters θ^f not a stationary point
- Further training yields non-zero change $\Delta\theta(\mathcal{D})$ even on same dataset $\mathcal{D}$
- Adjust for this effect of further training alone

Stochasticity

- Training algorithm is stochastic (e.g. shuffling of instances)
- Take expectation over this randomness



Existing Gradient-Based Methods Approximate Further Training

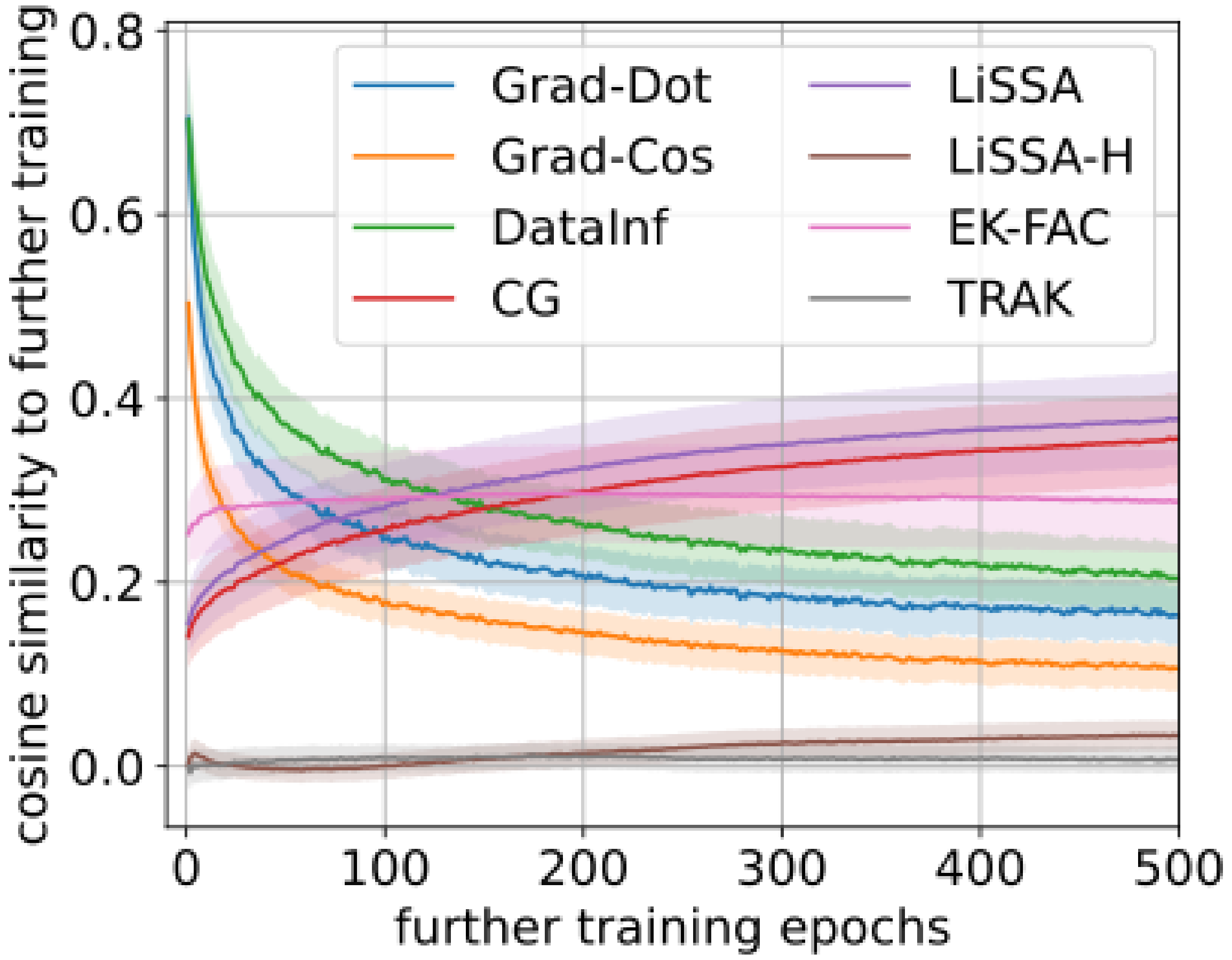


Methods falling under this framework:

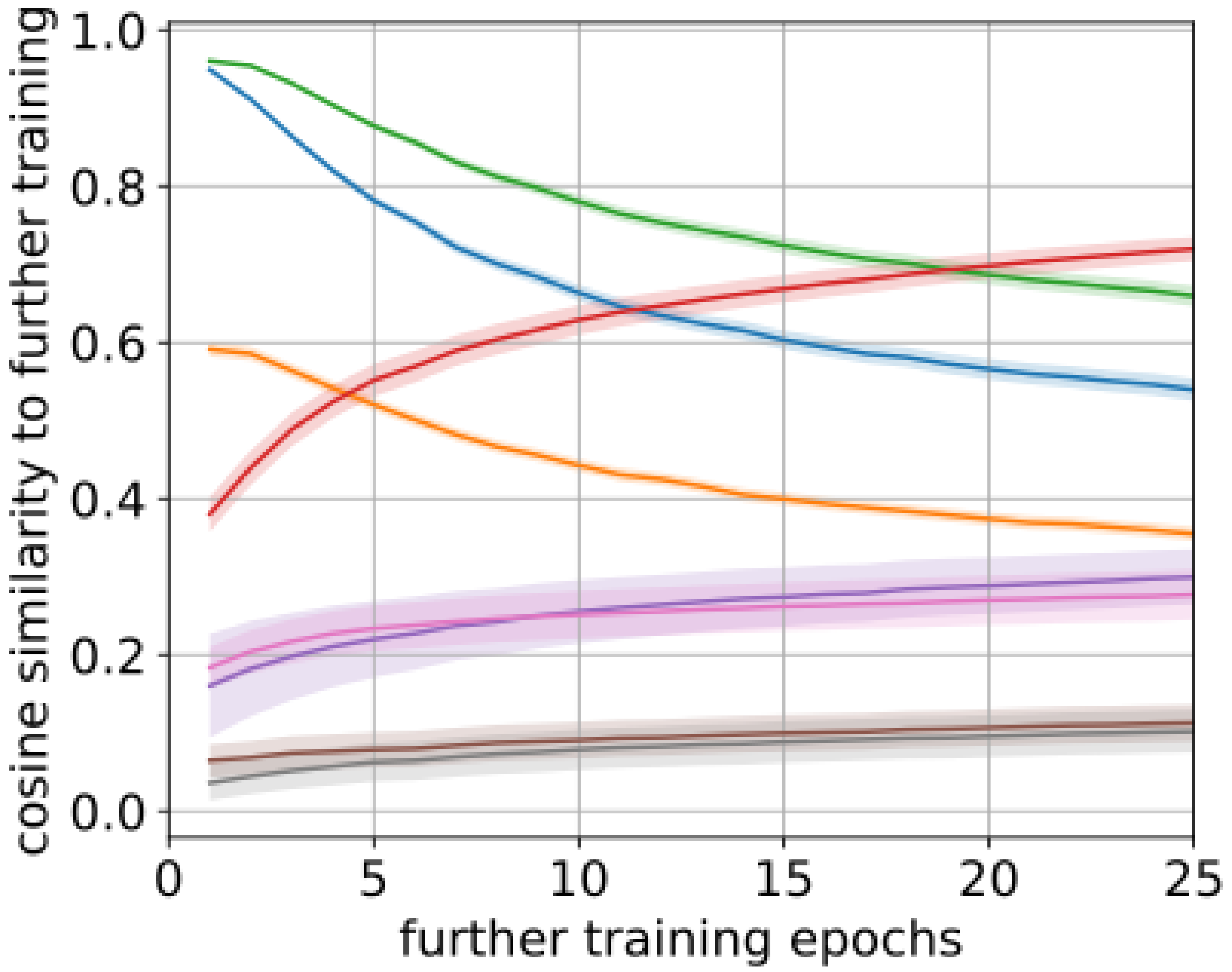
- First-order methods
  - Gradient similarity
  - TracIn (final checkpoint only)
- Influence functions
  - LiSSA, Conjugate gradient
  - Arnoldi
  - TRAK (no ensemble)
  - EK-FAC
  - DataInf

How Well Do Different Methods Approximate Further Training?

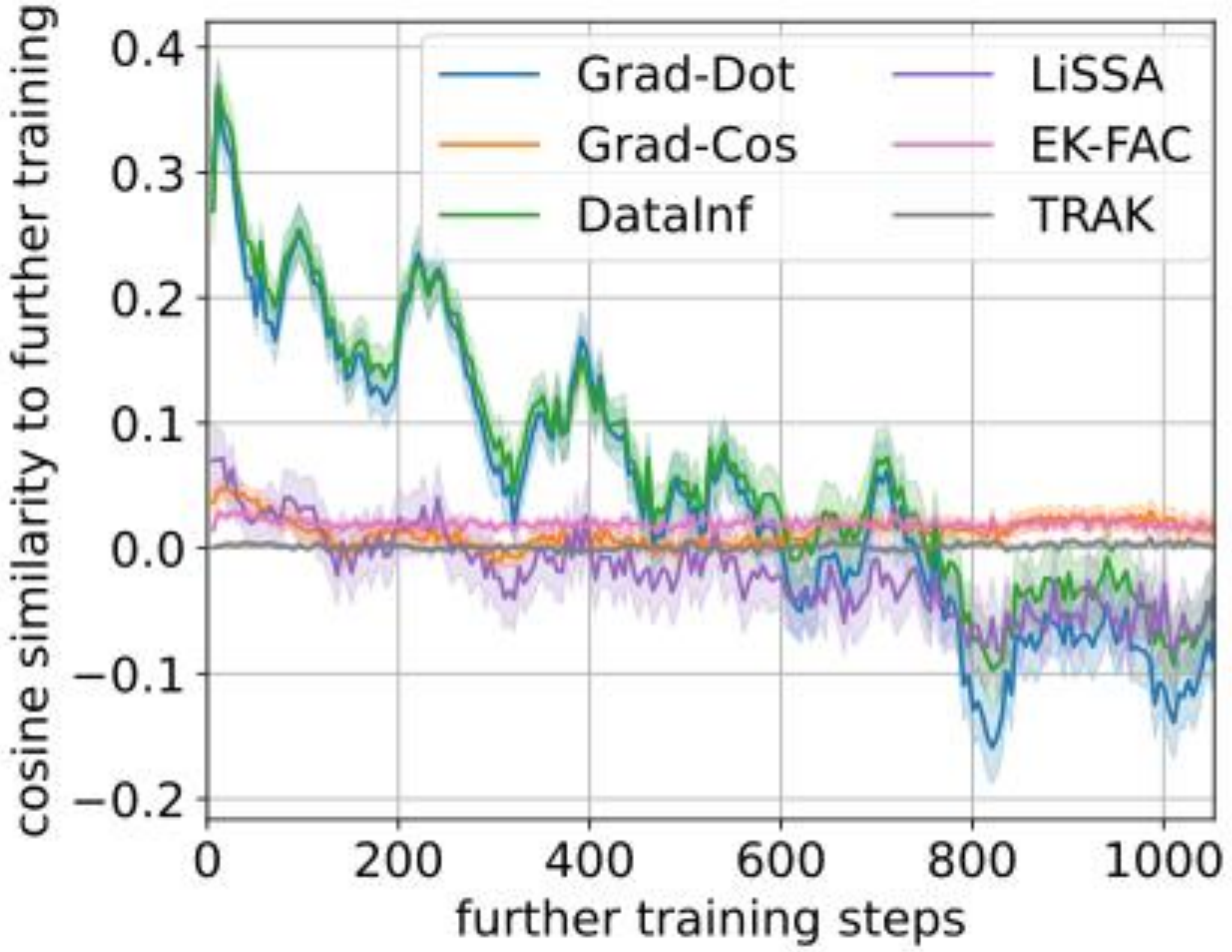
MLP on
Energy Efficiency



MLP on
Folktables



BERT on
SST-2



First-order(-like) methods: Approximation can be good initially but decays with further training

Influence function methods: More persistent but never as good

Summary

- Draw attention to final-model-only setting
- Reframe problem as one of quantifying *sensitivity* to training instances
- *Further training* as gold standard for quantifying sensitivity
- Existing gradient-based methods are approximations to further training
- Code for reproducibility: https://github.com/IBM/fimoda
- Discussion points (non-exhaustive):
  - Better approximate TDA methods for non-tabular models
  - Connections to data selection, etc.