Berkeley Center for Decentraliz

Session III: Foundations of Decentralized AI (II)

Fair Data Attribution at Scale

Ruoxi Jia
Assistant Professor
Virginia Tech

Berkeley Center for Decentraliz

Session III: Foundations of Decentralized AI (II)

Fair Data Attribution at Scale

Ruoxi Jia
Assistant Professor
Virginia Tech

REDS LAB
RESPONSIBLE DATA SCIENCE

Fair Data Attribution at Scale

Ruoxi Jia

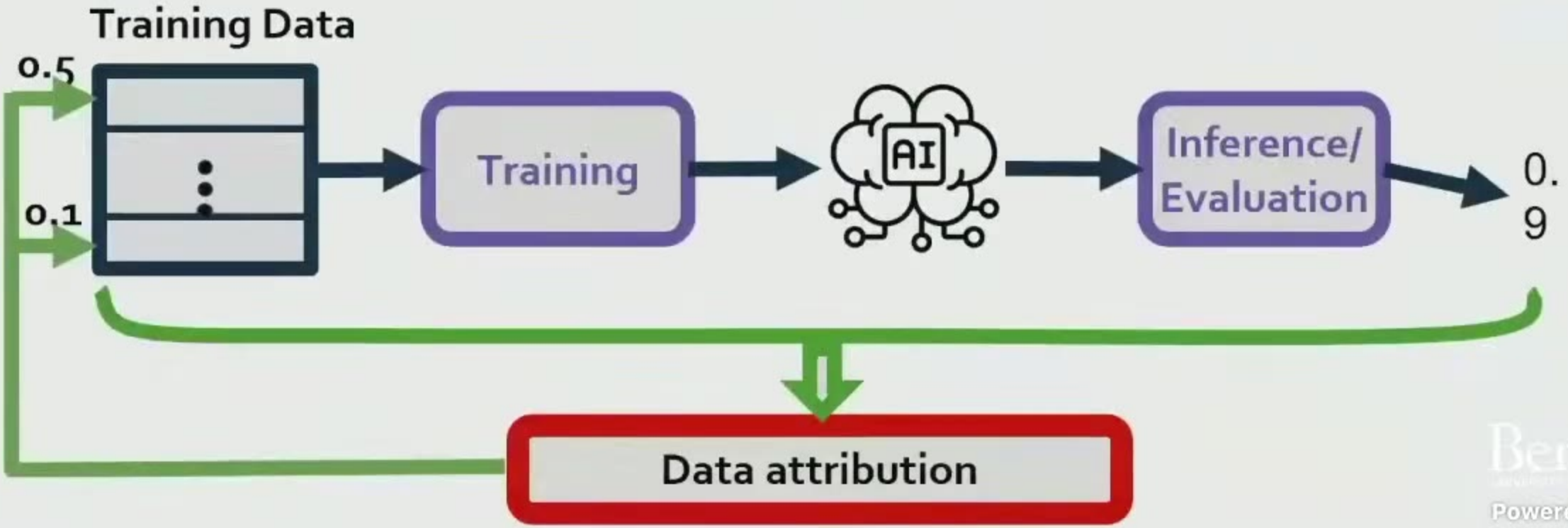

Jiachen T.


Prateek

Dawn

What is data attribution?

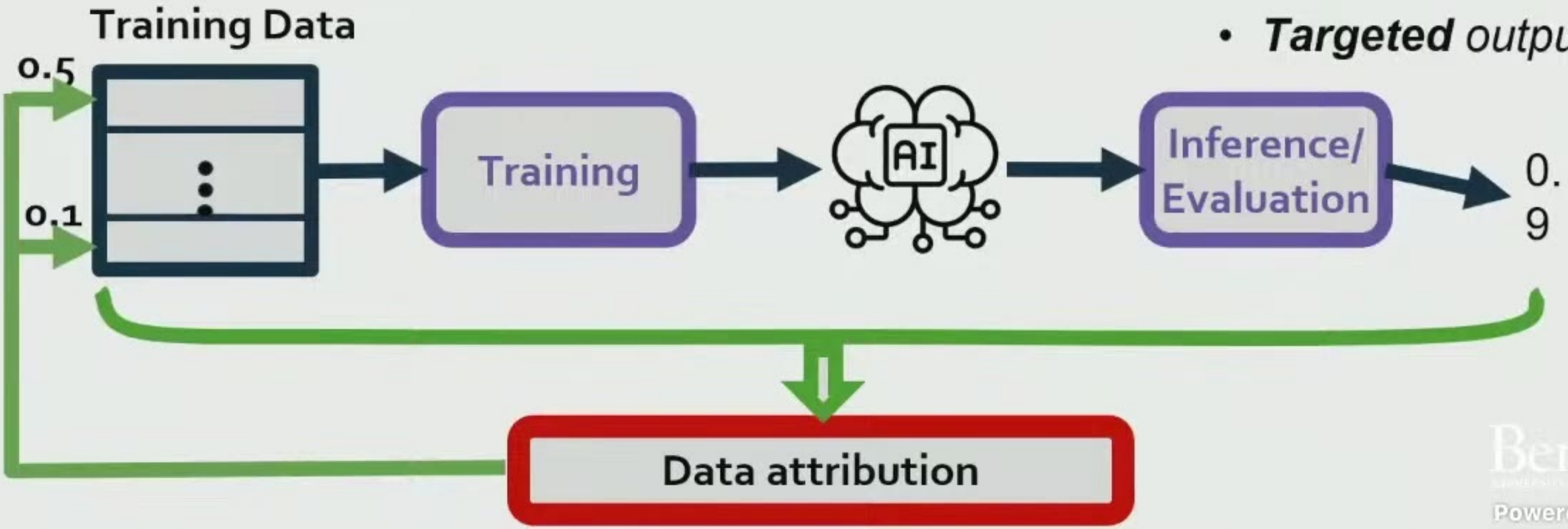
- Quantifying the contribution of individual data sources to model performance/behavior



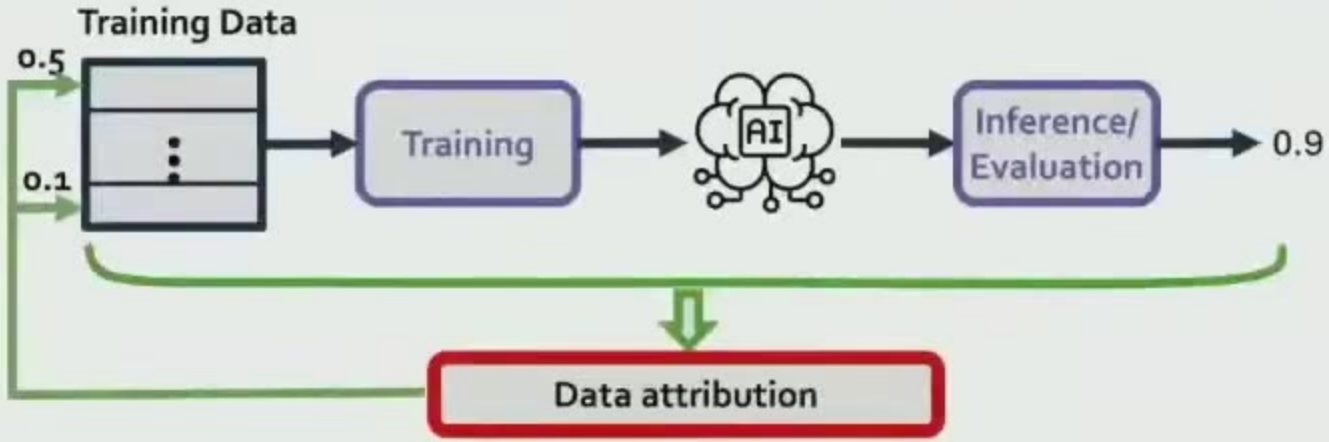
What is data attribution?

- Quantifying the contribution of individual data sources to model performance/behavior

- ***Aggregate*** *performance evaluated on a test set*
- ***Targeted*** *output behavior*

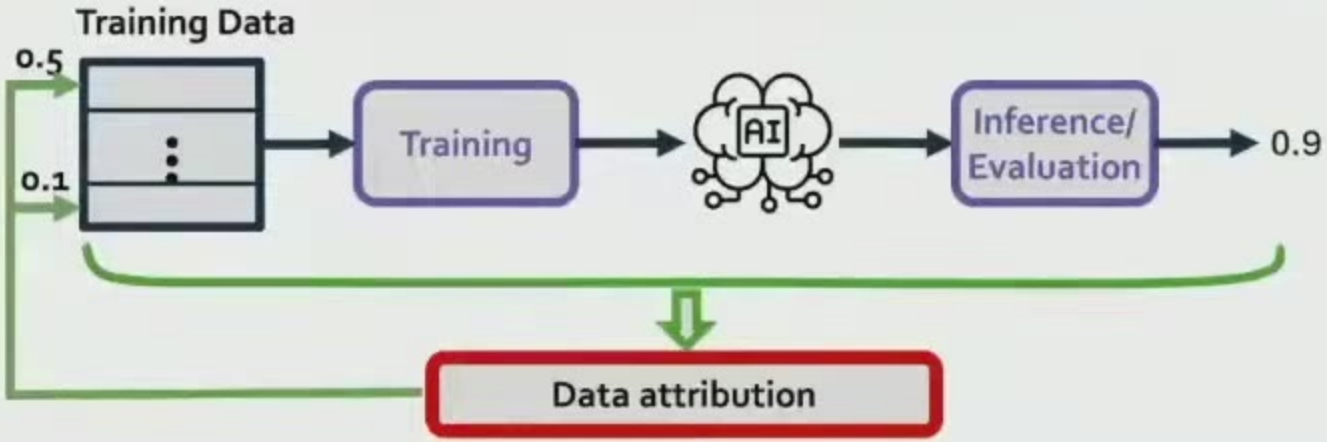


A fundamental component for Responsible Decentralized Intelligence



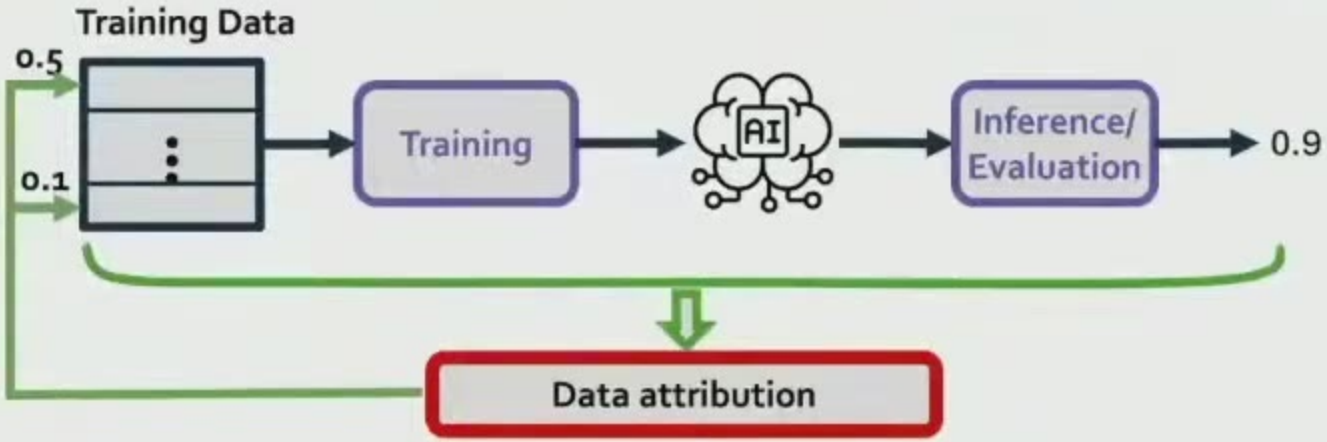
- Responsible
- Decentralized
- Intelligence

A fundamental component for Responsible Decentralized Intelligence



- Responsible
- Decentralized
- Intelligence

A fundamental component for Responsible Decentralized Intelligence



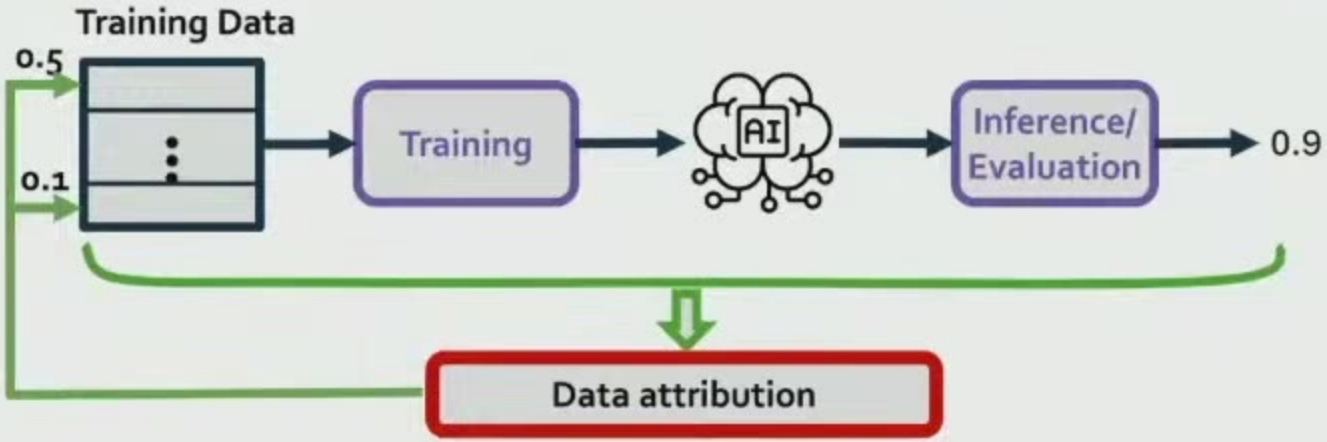
- **Transparency in model decisions:** Tracks the impact of different data sources on model outputs

Responsible

Decentralized

Intelligence

A fundamental component for Responsible Decentralized Intelligence



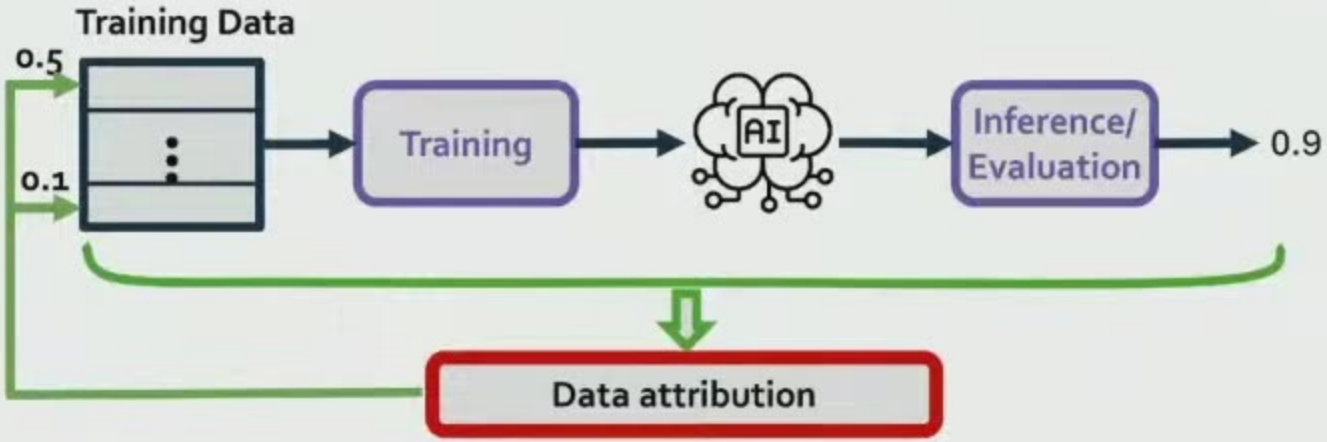
- **Transparency in model decisions:** Tracks the impact of different data sources on model outputs
- **Fair compensation**: Enables fair reward mechanisms for data contributors in distributed networks

Responsible

Decentralized

Intelligence

A fundamental component for Responsible Decentralized Intelligence



- **Transparency in model decisions:** Tracks the impact of different data sources on model outputs
- **Fair compensation**: Enables fair reward mechanisms for data contributors in distributed networks
- **Improved model performance**: Identifies high/low quality data points that significantly influence model performance and informs data filtering

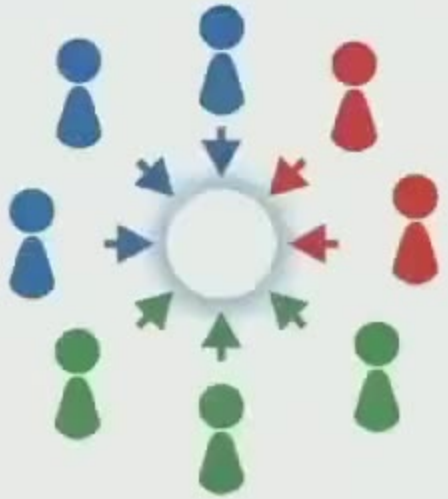
Responsible

Decentralized

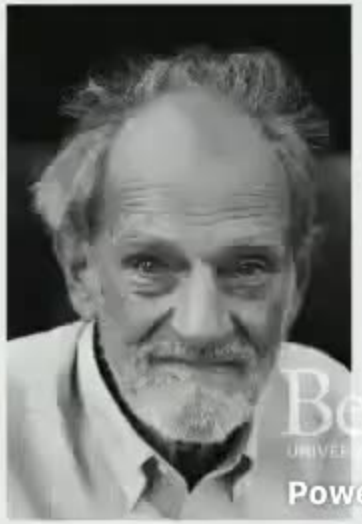
Intelligence

How to determine & distribute value data?

- Machine learning as a coalitional game:
  - Data contributors as players in a coalition
  - Usefulness of data is characterized via utility function



- **The Shapley value**
  - Defines a way of distributing the profit generated by the coalition of all players
  - First proposed by the Novel Prize Winning economist, Lloyd Shapley, in 1953
  - The ***only*** distribution that satisfies a collection of desirable properties



Unique properties of the Shapley valu

Full Value Distribution

$\sum_{\text{all data}}$ Shapley value = Total Model Performance Improvement

Implication:

- Incentive alignment
- Performance reflection

Unique properties of the Shapley valu

Full Value Distribution

$\sum_{\text{all data}}$ Shapley value = Total Model Performance Improvement

Implication:

- Incentive alignment
- Performance reflection

Contribution-based Valuation

Equal contribution ⇒ Equal value

No contribution ⇒ No value

Implication:

- Equitable valuation
- Resistance to data manipulation

Unique properties of the Shapley valu

Full Value Distribution

$\sum_{\text{all data}}$ Shapley value = Total Model Performance Improvement

Implication:

- Incentive alignment
- Performance reflection

Contribution-based Valuation

Equal contribution ⇒ Equal value

No contribution ⇒ No value

Implication:

- Equitable valuation
- Resistance to data manipulation

Additive Decomposition

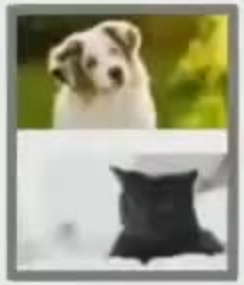
For $U = U_1 + U_2$:
Shapley(i,U) = Shapley(i,U_1) + Shapley(i,U_2)

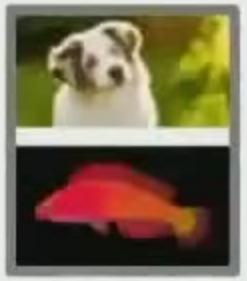
Implication:

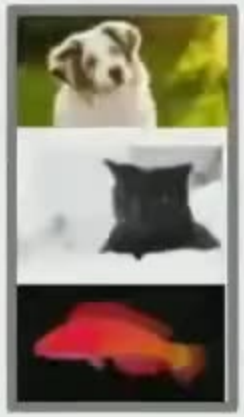
- Decomposability
- More usage, more value

How is the Shapley value calculated?

∅





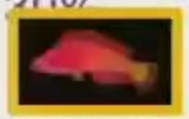


Score w/out= 10%
Score w/ = 40%
Contrib = 30%

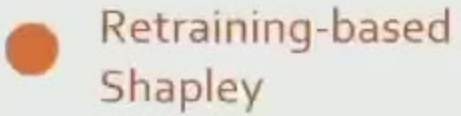
Score w/out= 30%
Score w/ = 60%
Contrib = 30%

Score w/out= 20%
Score w/ = 50%
Contrib = 30%

Score w/out= 50%
Score w/ = 80%
Contrib = 30%

Shapley () = average = (30% + 30%/2 + 30%/2 + 30%) / 3
= 30%

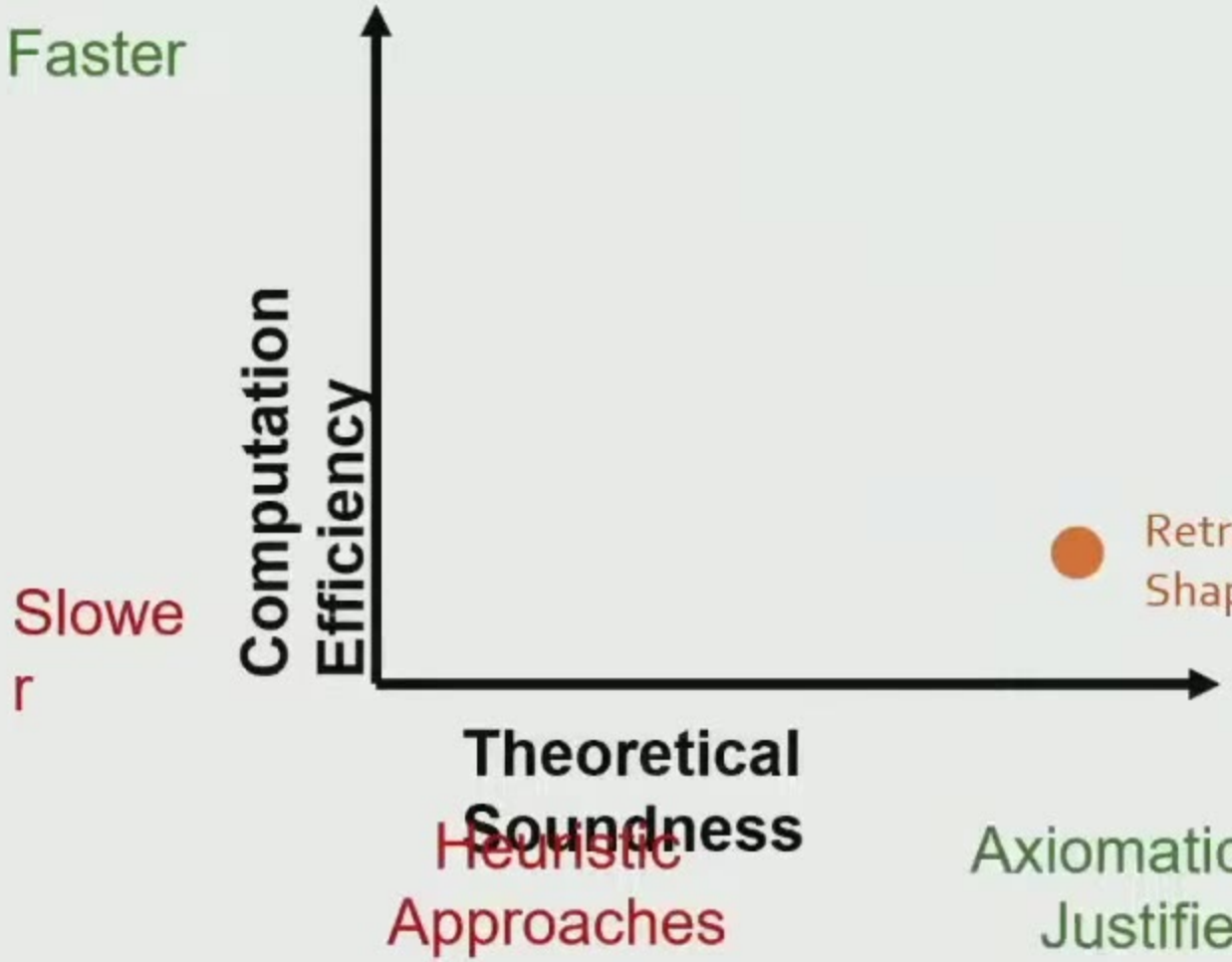
Challenges with existing data attribut methods



Challenges with existing data attribut methods

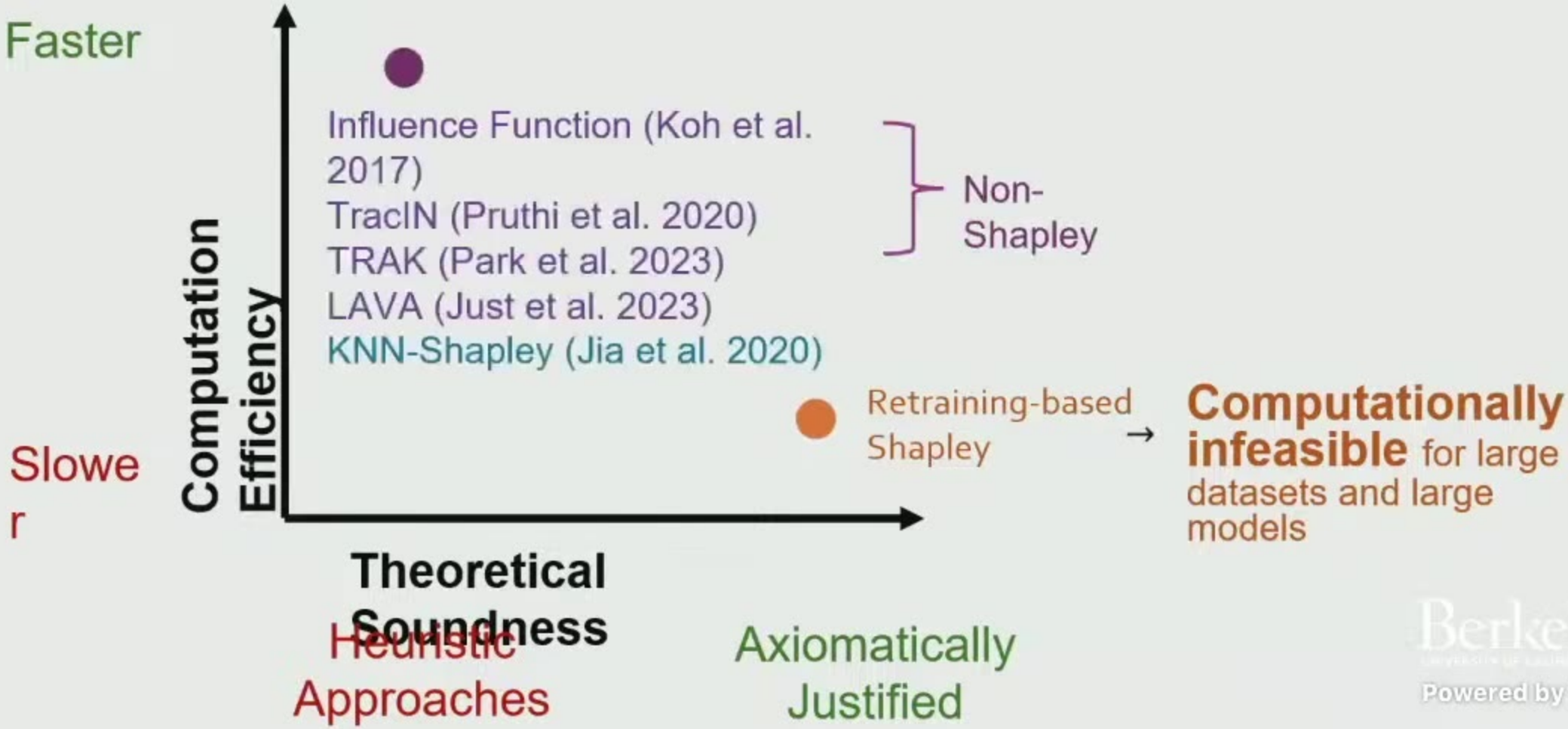
- Retraining-based Shapley → **Computationally infeasible** for large datasets and large models

Challenges with existing data attribution methods

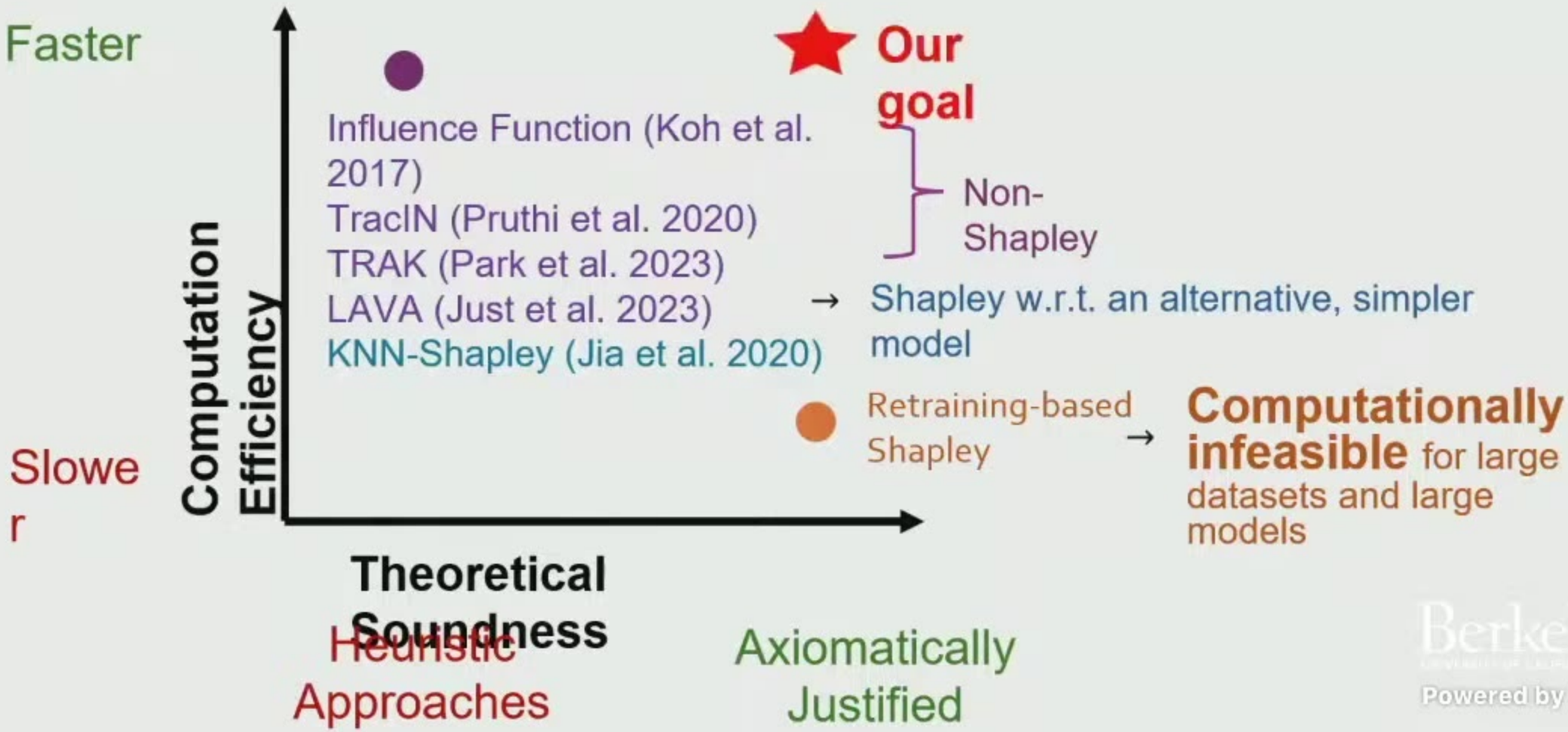


Retraining-based Shapley → **Computationally infeasible** for large datasets and large models

Challenges with existing data attribution methods



Challenges with existing data attribut methods



In-run Data Shapley

First **Theoretically-Principled** Data Attribution Method **Scalable** to Foundation Model Training

Theoretically Principled

- Satisfies Shapley value axioms
- Ensures fairness in attribution

Scalable

- Yields negligible overhead compared to standard training
- Applicable to large-scale models, large datasets

In-run Data Shapley

First **Theoretically-Principled** Data Attribution Method **Scalable** to Foundation Model Training

Theoretically Principled

- Satisfies Shapley value axioms
- Ensures fairness in attribution

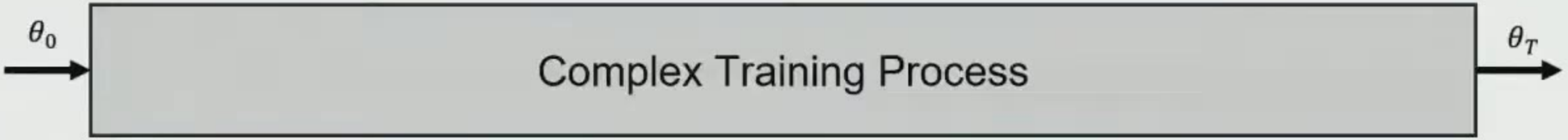
Scalable

- Yields negligible overhead compared to standard training
- Applicable to large-scale models, large datasets

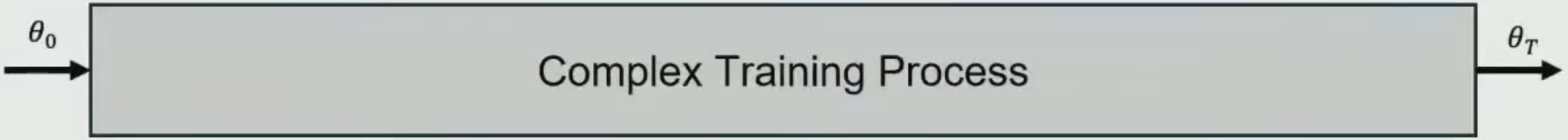
Flexibility

- Applicable to any model trained via gradient methods
- Applicable to different stages of learning: Pre-training, fine-tuning, alignment, etc.

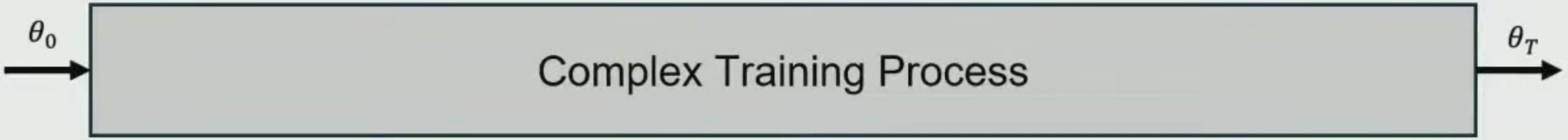
Key idea: Divide and conquer



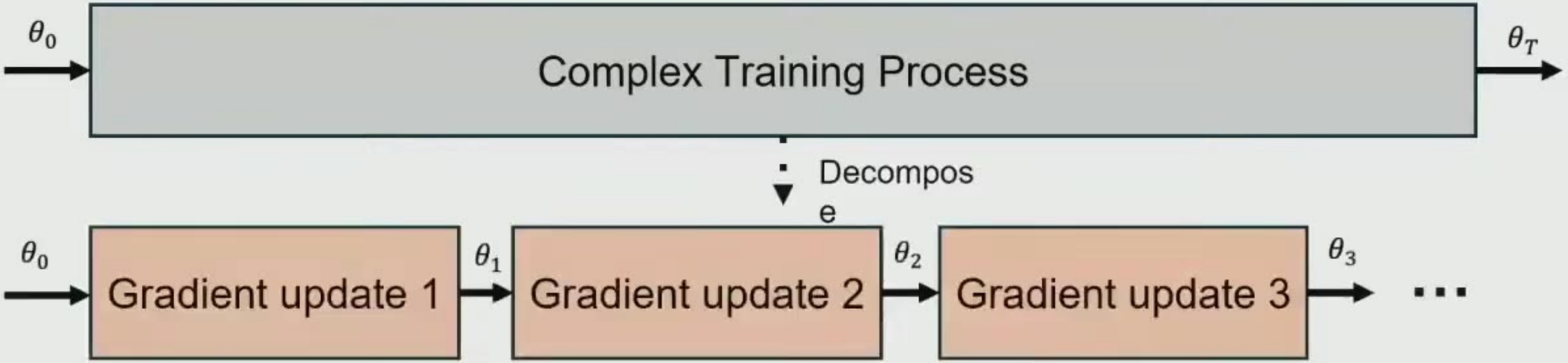
Key idea: Divide and conquer



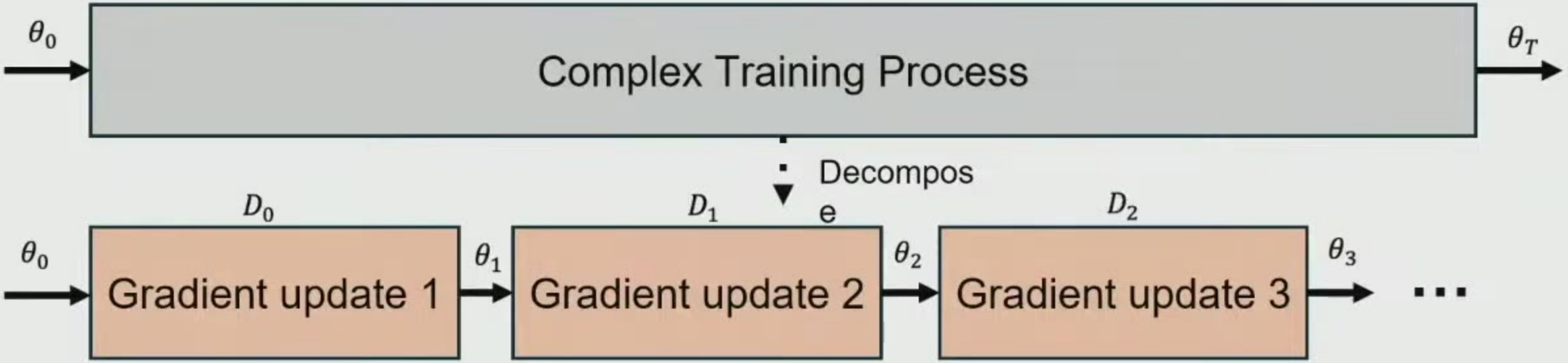
Key idea: Divide and conquer



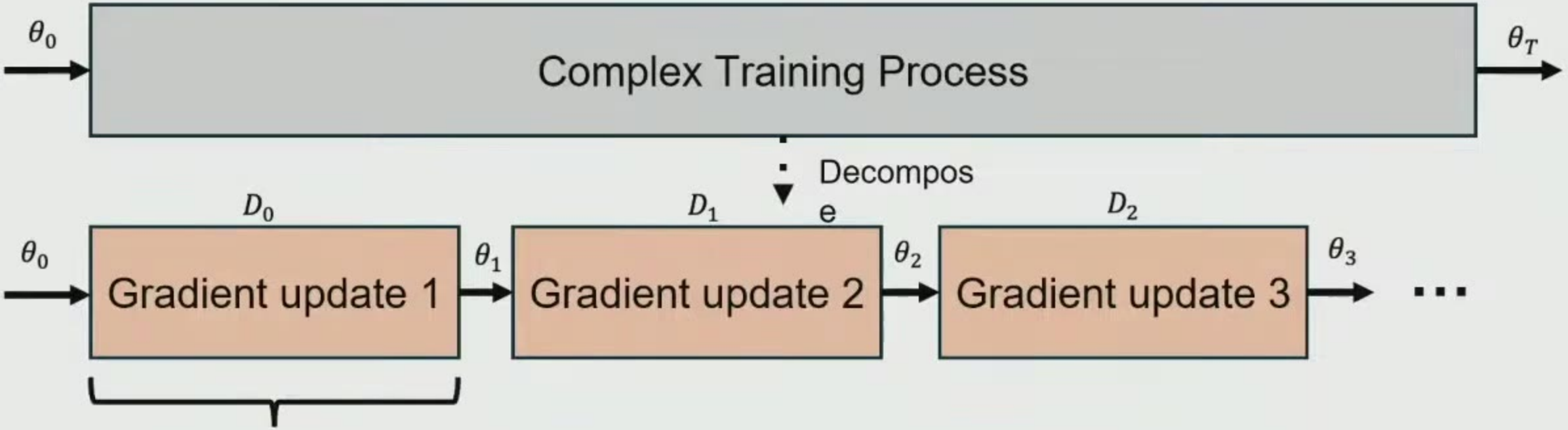
Key idea: Divide and conquer



Key idea: Divide and conquer



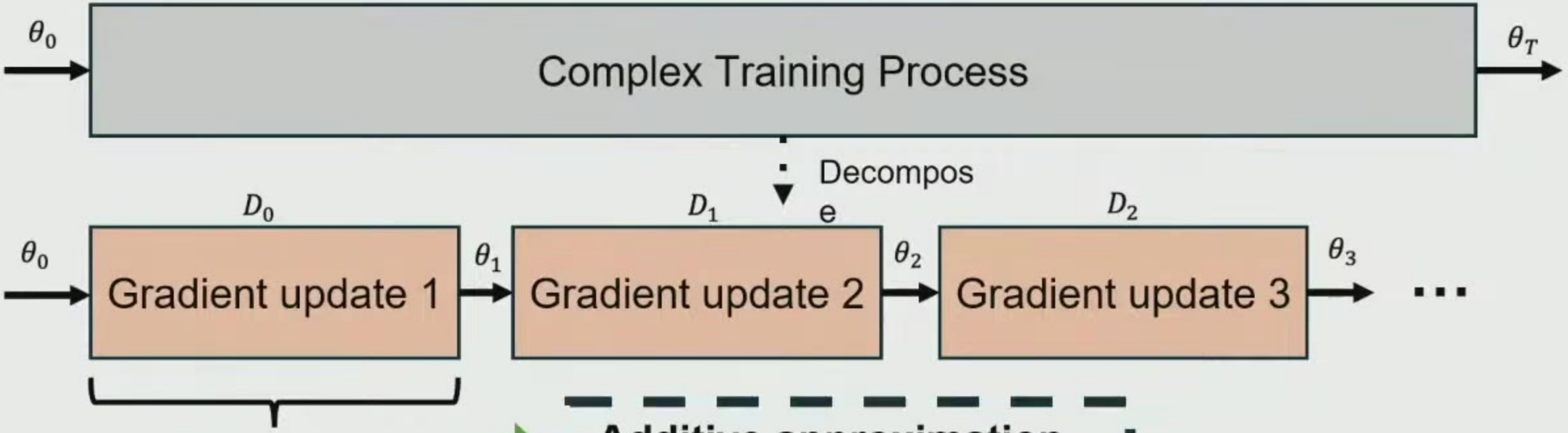
Key idea: Divide and conquer



Loss change is very **small**

$$l(\theta_{t+1}, D_{val}) - l(\theta_t, D_{val})$$

Key idea: Divide and conquer



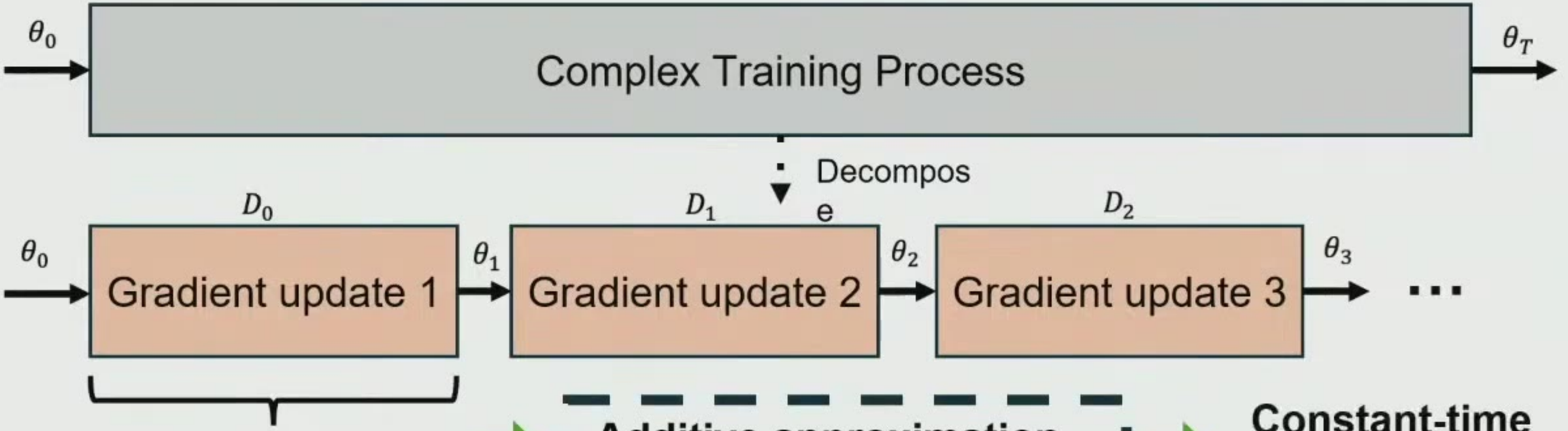
Loss change is very **small**

$$l(\theta_{t+1}, D_{val}) - l(\theta_t, D_{val})$$

Additive approximation
via Taylor expansion

$$\approx \nabla l(\theta_t, D_{val}) \cdot \sum_{d \in D_t} \nabla l(\theta_t, d)$$

Key idea: Divide and conquer



Loss change is very **small**

$$l(\theta_{t+1}, D_{val}) - l(\theta_t, D_{val})$$

Additive approximation via Taylor expansion

$$\approx \nabla l(\theta_t, D_{val}) \cdot \sum_{d \in D_t} \nabla l(\theta_t, d)$$

Constant-time computation for Shapley w.r.t. individual update

How do values of different corpora change during training?

Generating a **math-related** corpus:

```
In several applications, the matrix
coefficients of the nonlinear valued
function G(λ) are usually of low rank. In
this section, we show how the exploitation
of these low ranks leads to a linearization
of size smaller than that of £_R(λ). This
linearization generalizes the one used in
[@dopi17; @suba11], which is valid when P(λ)
is expressed using monomials, i.e.,
f_i(λ)=λ^i, to the more general setting used
by CORK.
```

How do values of different corpora change during training?

Generating a **math-related** corpus:

```
In several applications, the matrix
coefficients of the nonlinear valued
function G(λ) are usually of low rank. In
this section, we show how the exploitation
of these low ranks leads to a linearization
of size smaller than that of 𝓛_R(λ). This
linearization generalizes the one used in
[@dopi17; @suba11], which is valid when P(λ)
is expressed using monomials, i.e.,
f_i(λ)=λ^i, to the more general setting used
by [@ber
```

Takeaways:

- General corpora help most in the beginning
- Task-relevant corpora help most later

Value distribution of **generating semantically similar output**

Trainin g

Test-time generation

Original Wikipedia Corpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Value distribution of **generating semantically similar output**

Trainin g

Original Wikipedia Corpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Paraphras e

Test-time generation

In 2012, Radhi assembled new 'musicians' for OAG, choosing from students at the Akademi Seni Budaya dan Warisan Kebangsaan. The updated lineup includes Qi Razali on drums and backing vocals, returning to his original role, along with Muhamad Nizam on guitar (since 2005), Nazrin Zabidi playing bass and providing backing vocals, and Izmer Khasbullah on keyboards and backing vocals.

Value distribution of **generating semantically similar output**

Trainin g

Original Wikipedia Corpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Paraphras e

Test-time generation

In 2012, Radhi assembled new 'musicians' for OAG, choosing from students at the Akademi Seni Budaya dan Warisan Kebangsaan. The updated lineup includes Qi Razali on drums and backing vocals, returning to his original role, along with Muhamad Nizam on guitar (since 2005), Nazrin Zabidi playing bass and providing backing vocals, and Izmer Khasbullah on keyboards and backing vocals.

The value of training corpus **rank 1st out of 320k corpora** for generating a paraphrase.

Value distribution for **creative gener**

Trainin
g

Original Wikipedia Corpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Test-time generation

Value distribution for **creative gener**

Trainin
g

Original Wikipedia Crpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Similar topic but **complete rewrite**

Test-time generation

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...

Value distribution for **creative generation**

Trainin g

Original Wikipedia Corpus

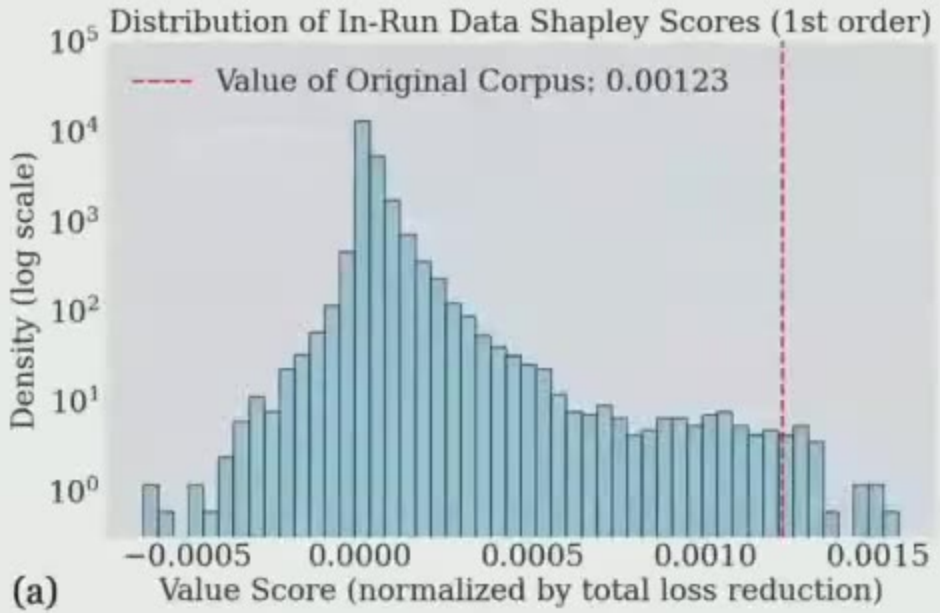
In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Similar topic but **complete rewrite**

Test-time generation

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...



Value distribution for **creative generation**

Trainin g

Original Wikipedia Corpus

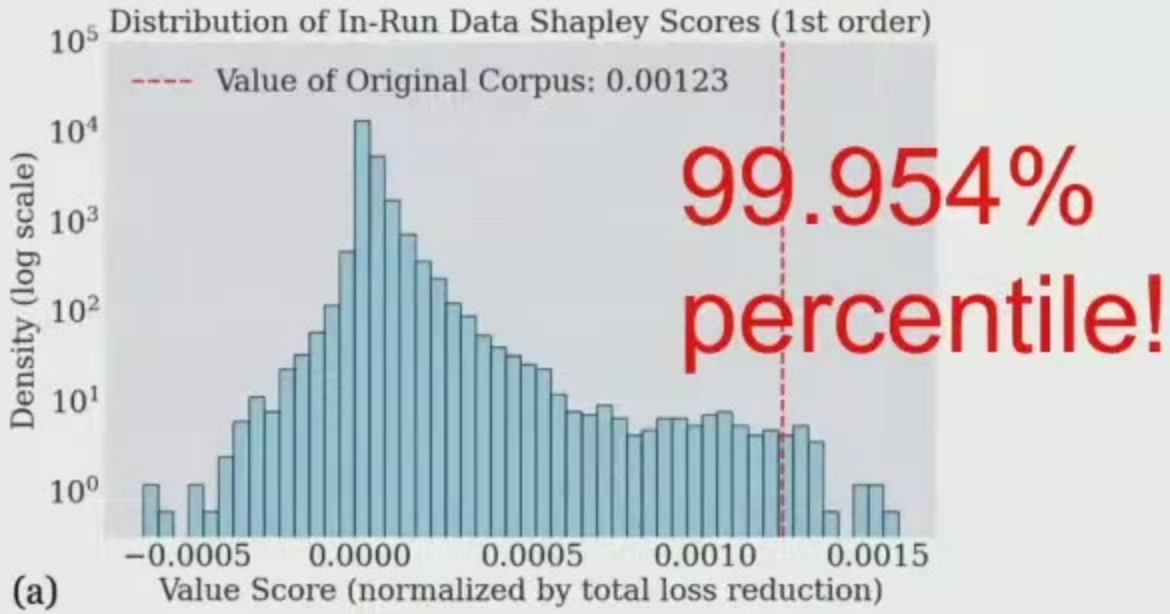
In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Similar topic but **complete rewrite**

Test-time generation

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...



99.954% percentile!

Verizon Event ...

Powered by Zoom

Value distribution for **creative gener**

Trainin
g

Original Wikipedia Corpus

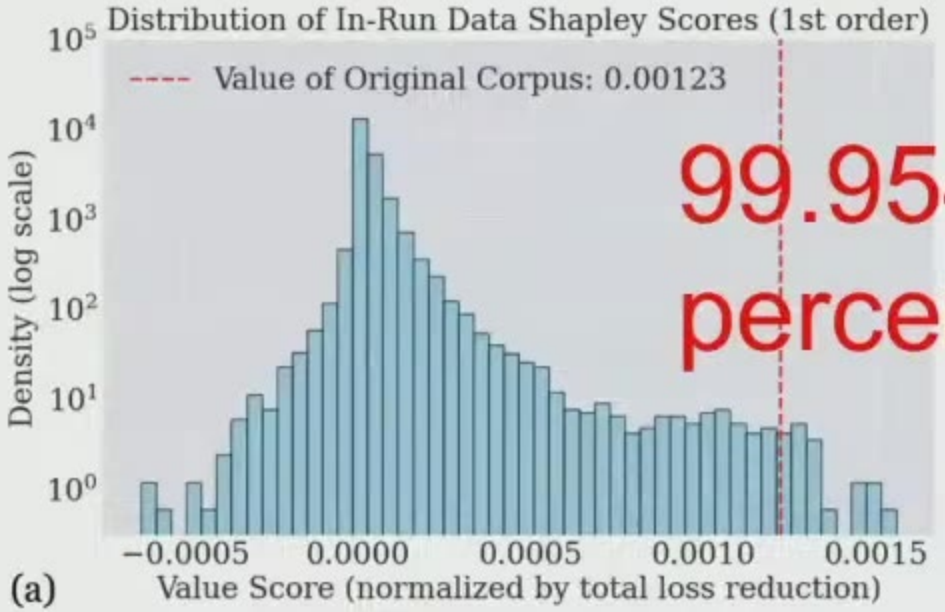
In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Similar topic but **complete rewrite**

Test-time generation

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...



99.954% percentile!

Takeaways:

- The training data still contributes significantly for "transformative" generation

Value distribution for **creative gener**

Trainin g

Original Wikipedia Corpus

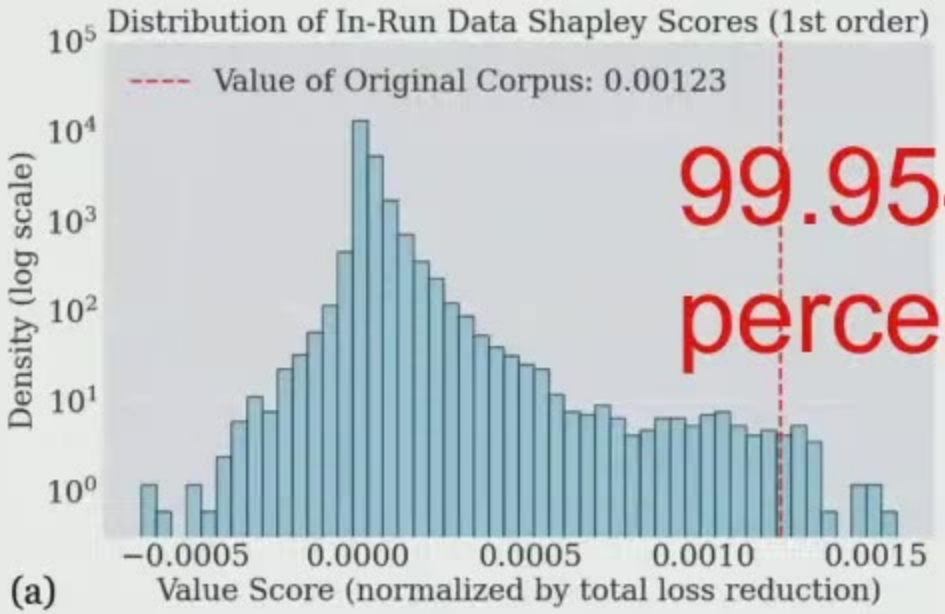
In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Similar topic but **complete rewrite**

Test-time generation

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...



99.954% percentile!

Takeaways:

Value distribution for **creative generation**

Trainin
g

Similar topic but **complete rewrite**

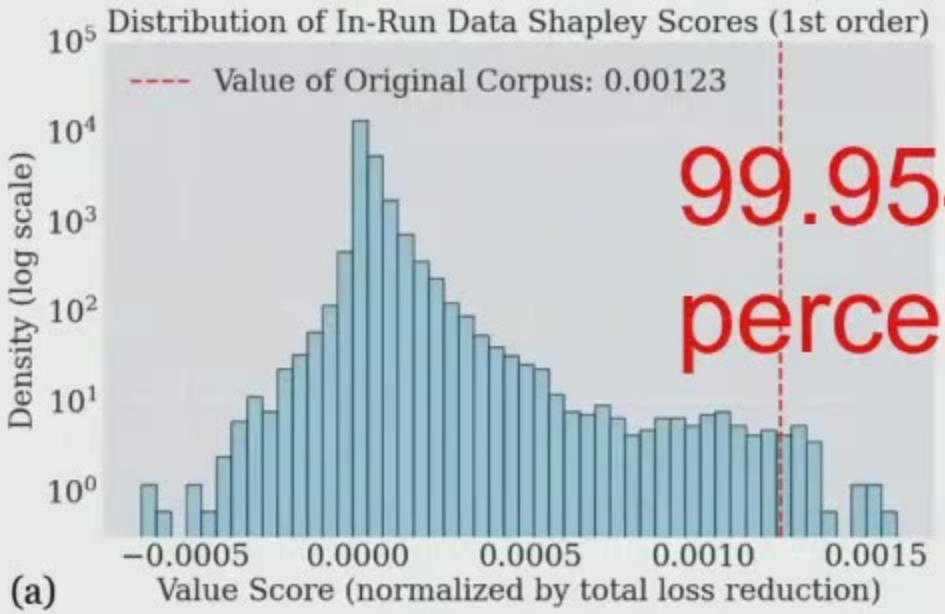
Test-time generation

Original Wikipedia Corpus

In 2012, Radhi recruited new 'musicians' for OAG, who were selected from among the students of Akademi Seni Budaya dan Warisan Kebangsaan (). The new line-up consists of Qi Razali (drums/backing vocals - original drummer back to ...

Synthetic "Similar topic" corpus

Instruction: Write a short story about a classical violinist who decides to explore jazz music, detailing her first encounter with a jazz band. ### Answer: Elena, a classically trained violinist known for her precise and emotive performances ...



99.954% percentile!

Takeaways:

- The training data still contributes significantly for "transformative" generation
- Implications to copyright: *Should scraping data for training GenAI models still be considered fair use?*

Removing negative-valued data impr
training

Evaluation on the Pile
Dataset

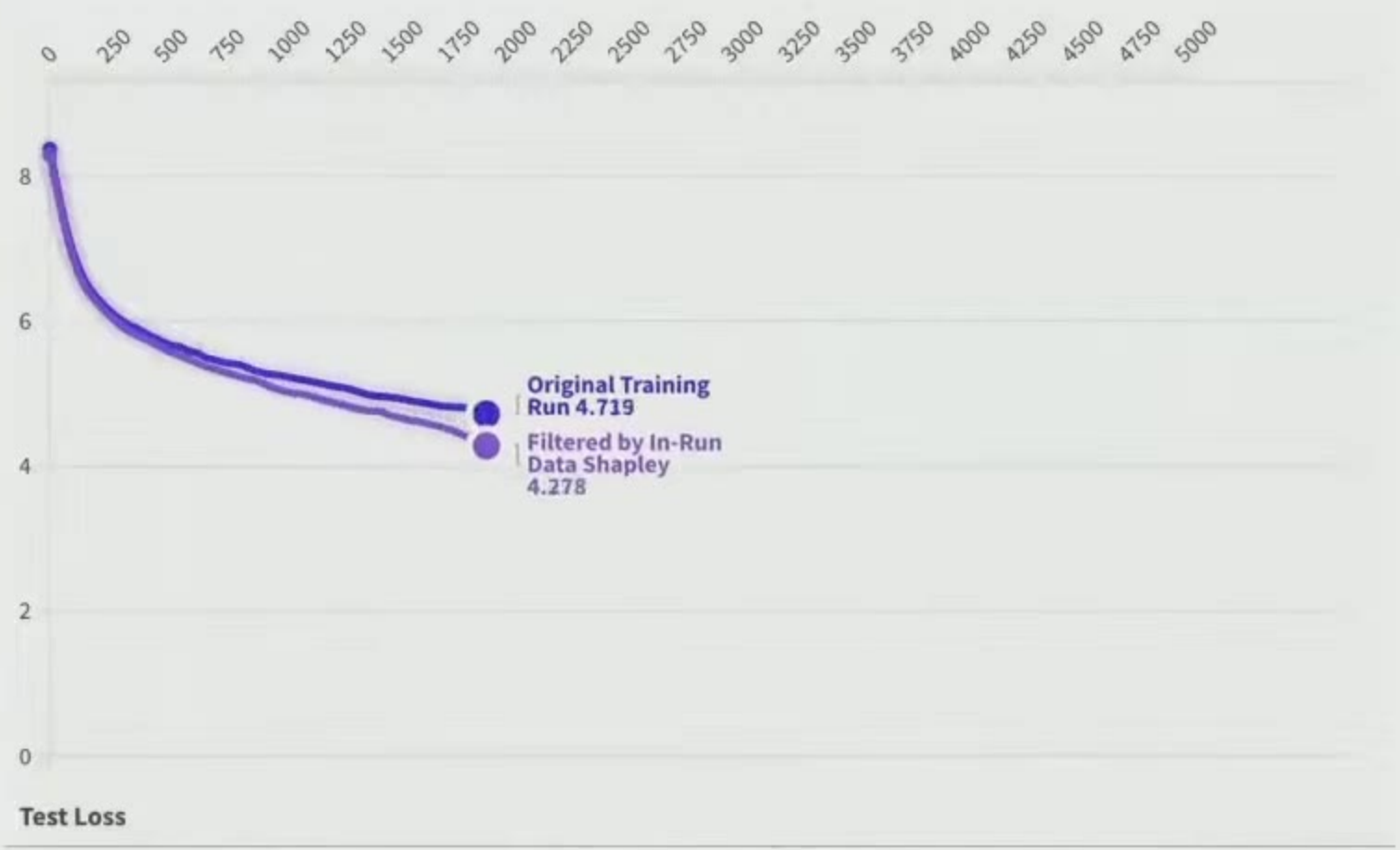
Removing negative-valued data impr
training

Evaluation on the Pile
Dataset

Removing negative-valued data impr
training

Evaluation on the Pile
Dataset

Removing negative-valued data impr
training

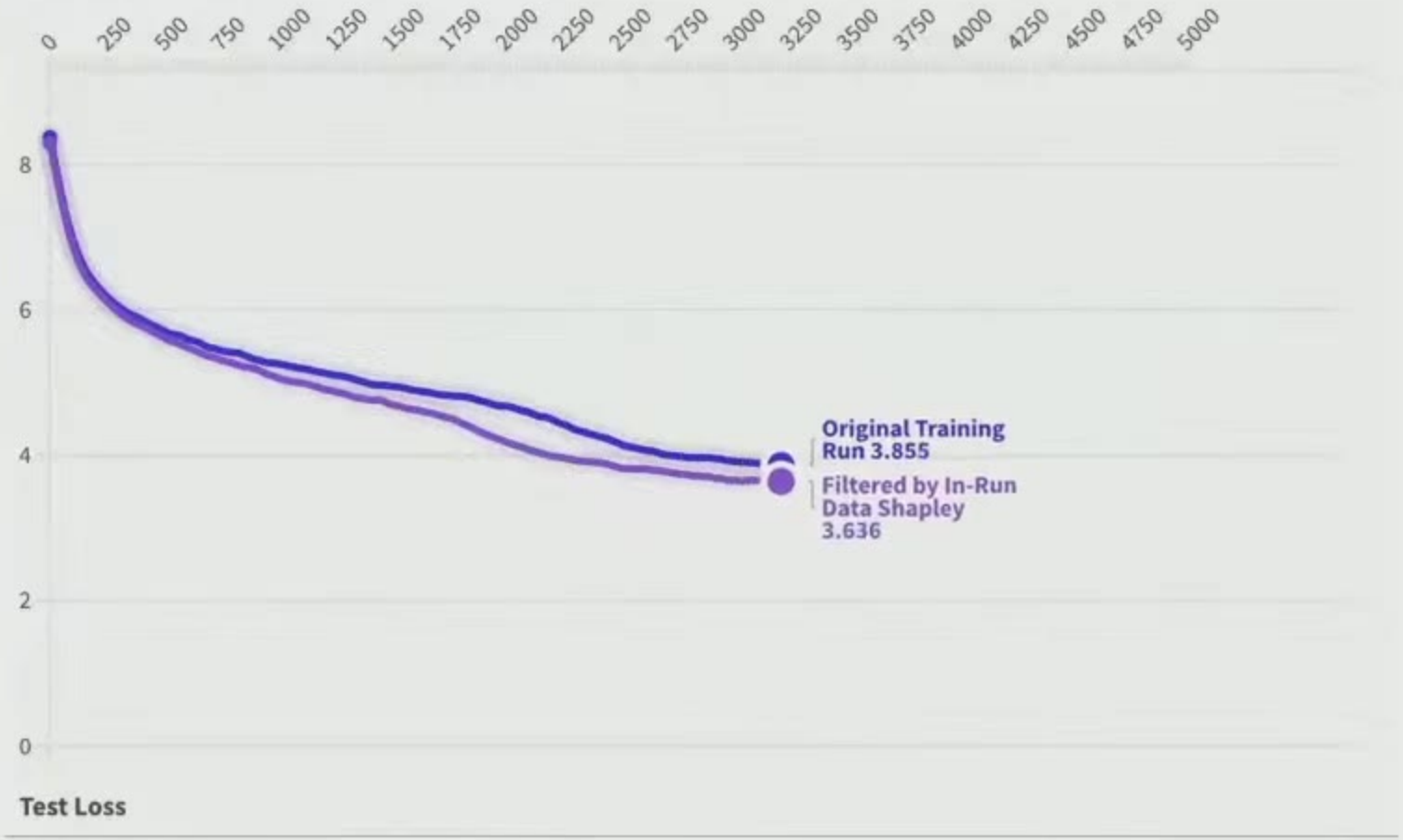


Evaluation on the Pile Dataset

Removing negative-valued data impr training

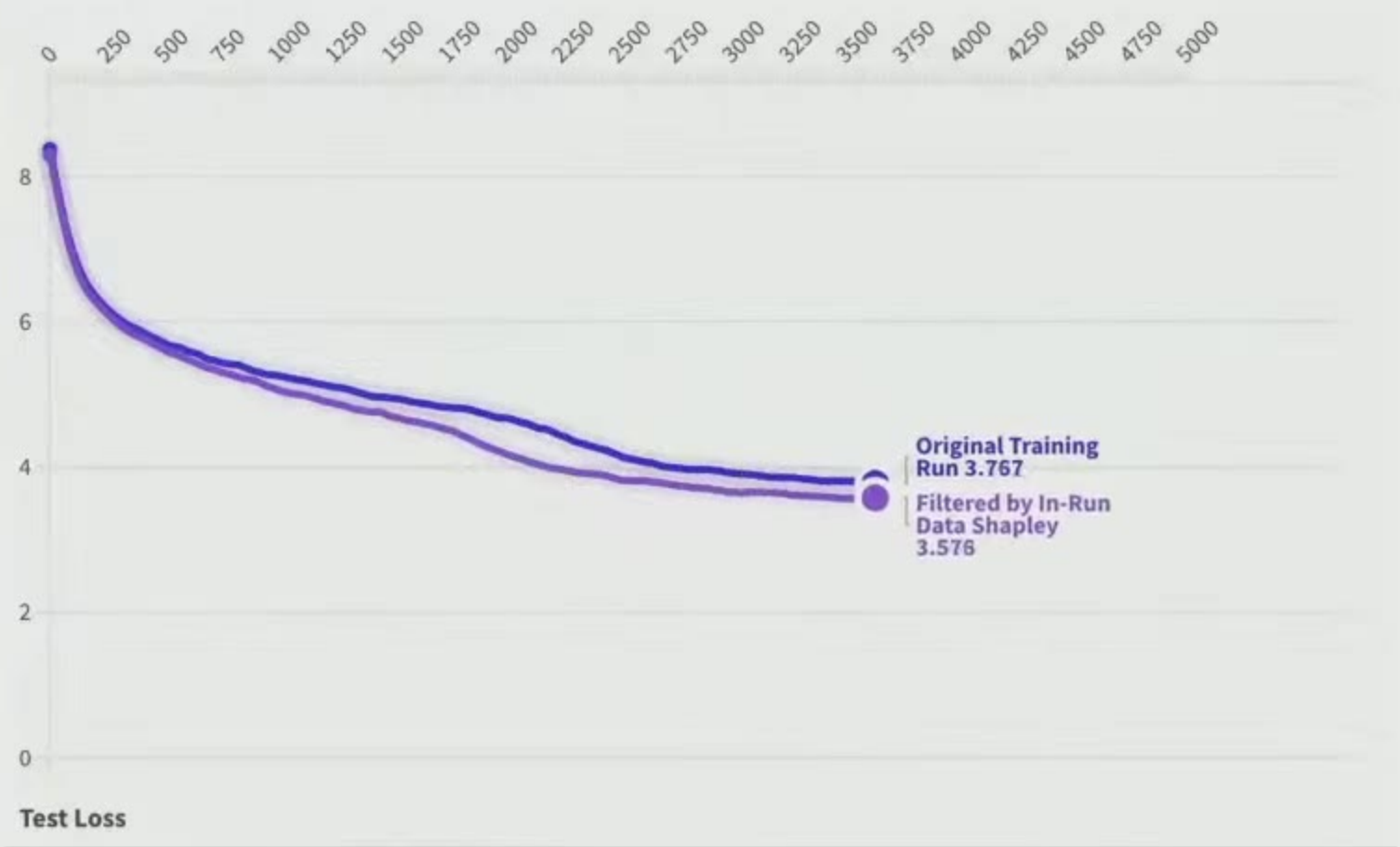
Evaluation on the Pile Dataset

Removing negative-valued data impr
training



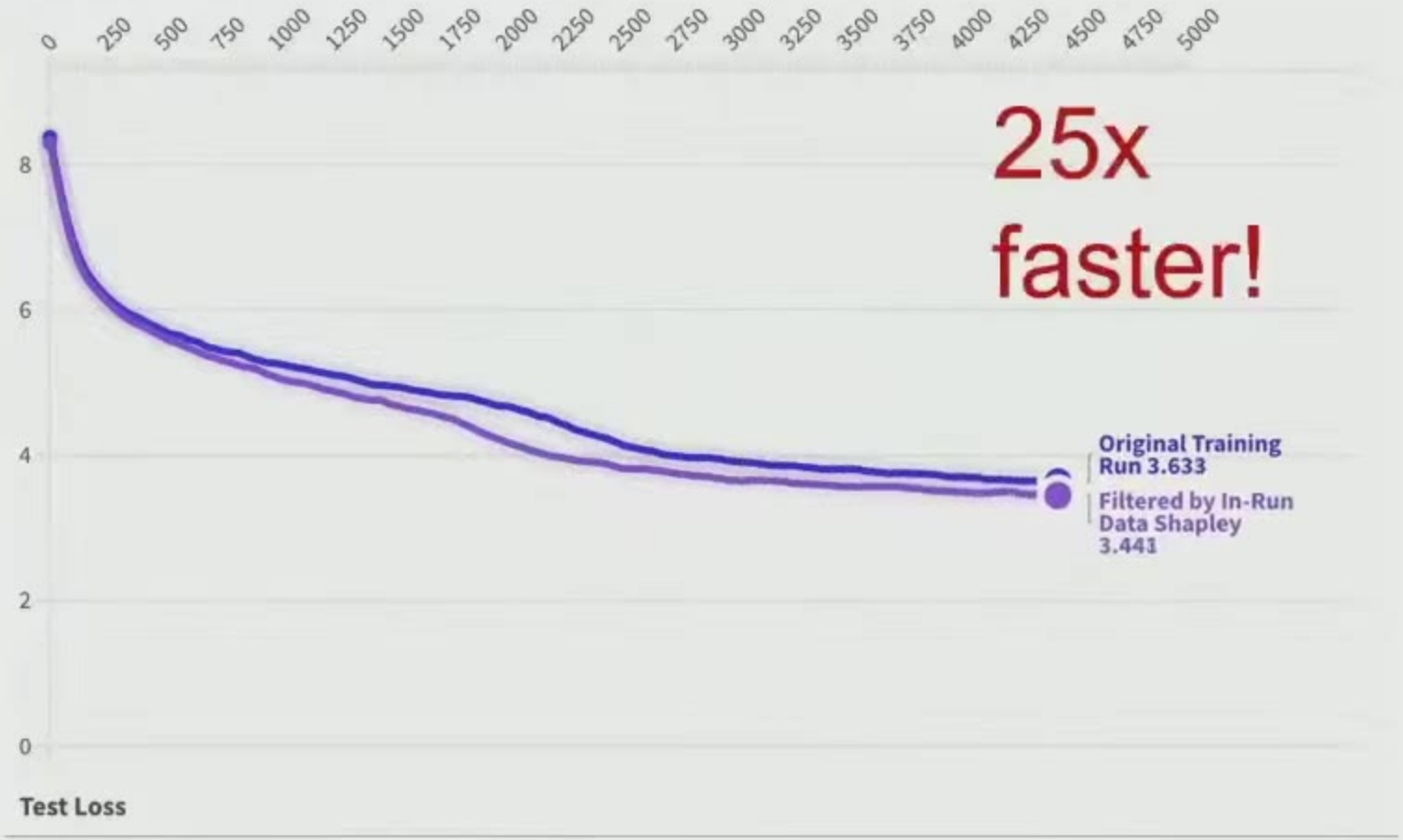
Evaluation on the Pile Dataset

Removing negative-valued data impr
training



Evaluation on the Pile Dataset

Removing negative-valued data impr
training

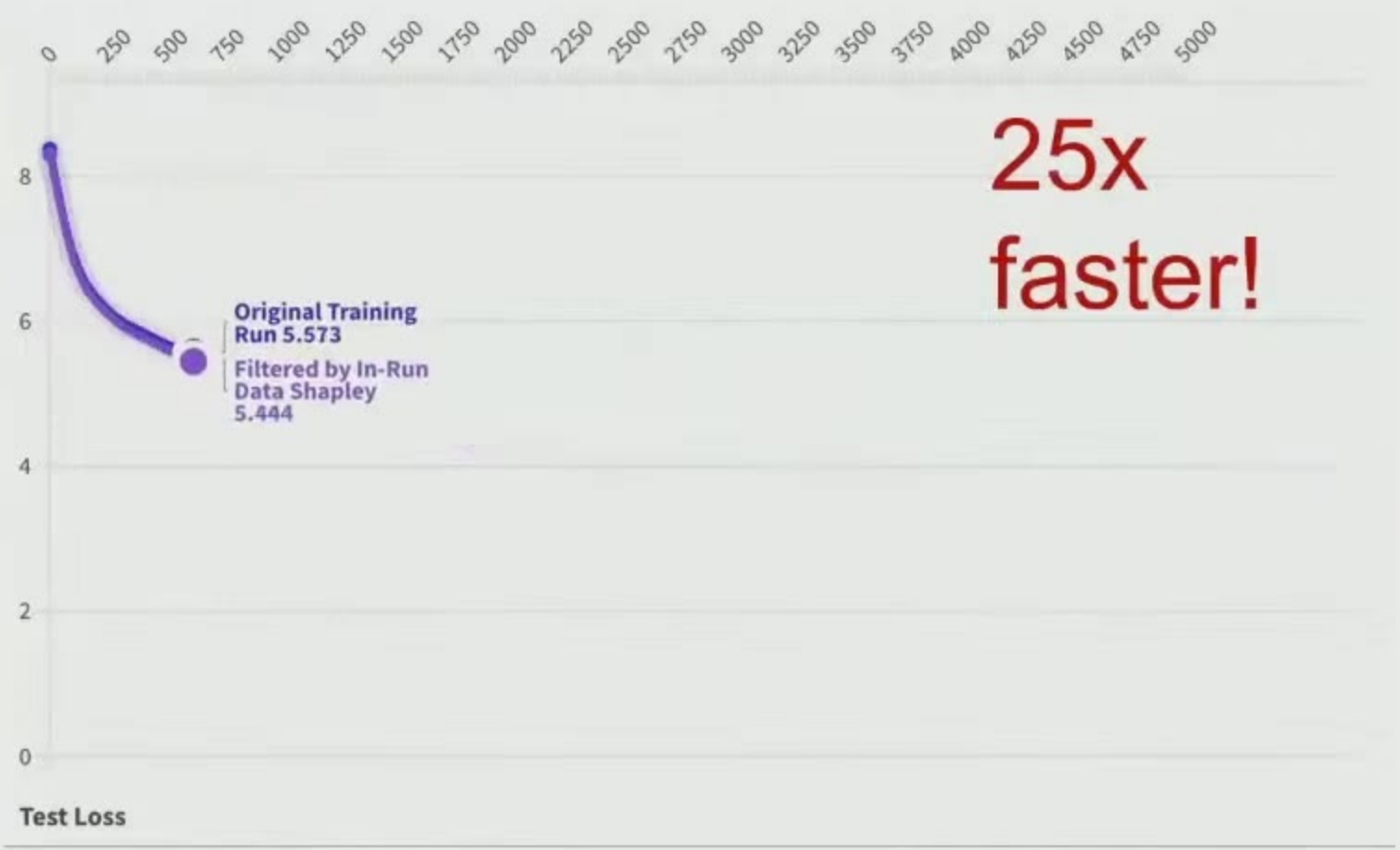


Evaluation on the Pile
Dataset

Removing negative-valued data impr
training

Evaluation on the Pile Dataset

Removing negative-valued data impr
training



Evaluation on the Pile
Dataset

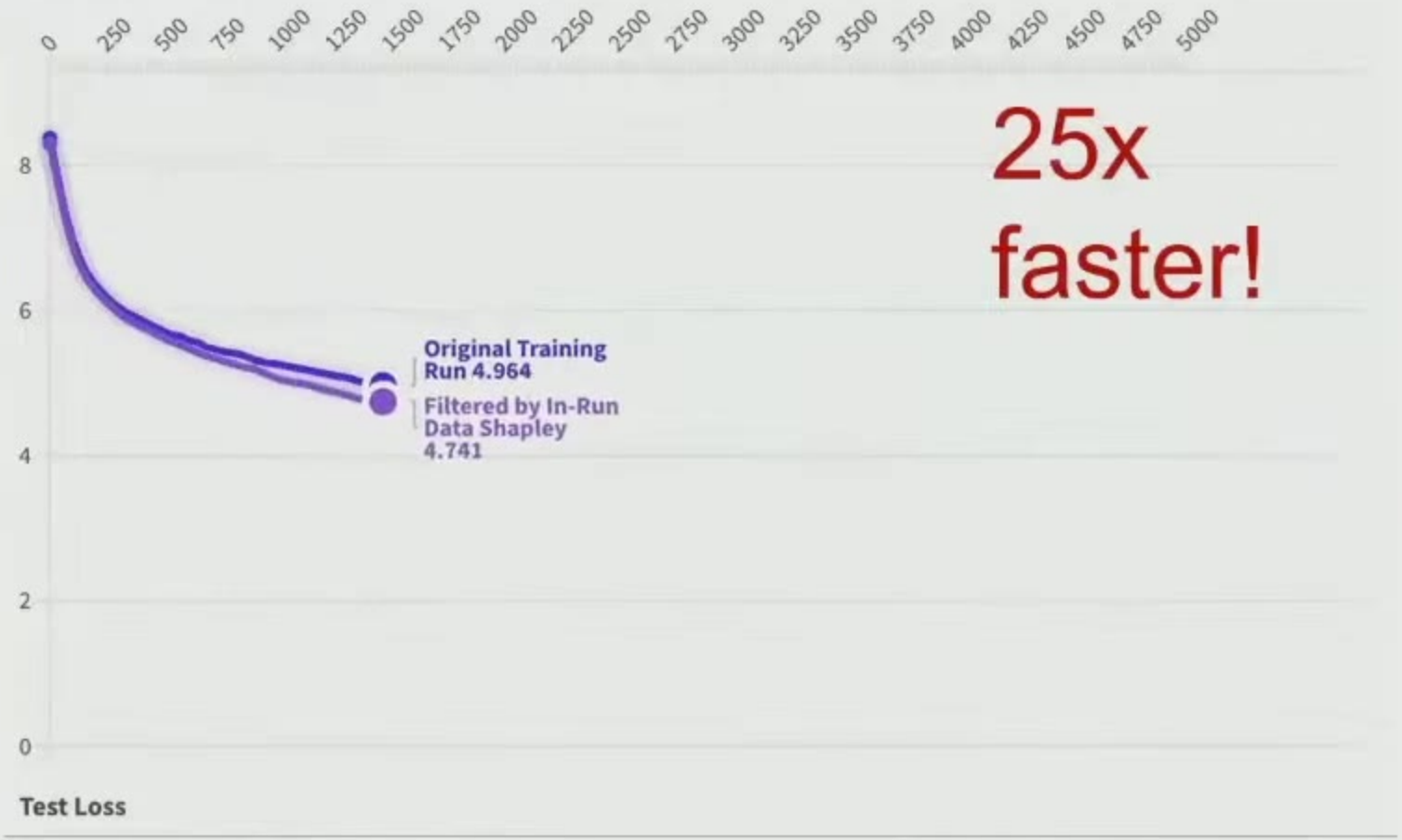
Removing negative-valued data impr training

25x faster!

An example of negative-valued corpus

Evaluation on the Pile Dataset

Removing negative-valued data impr training

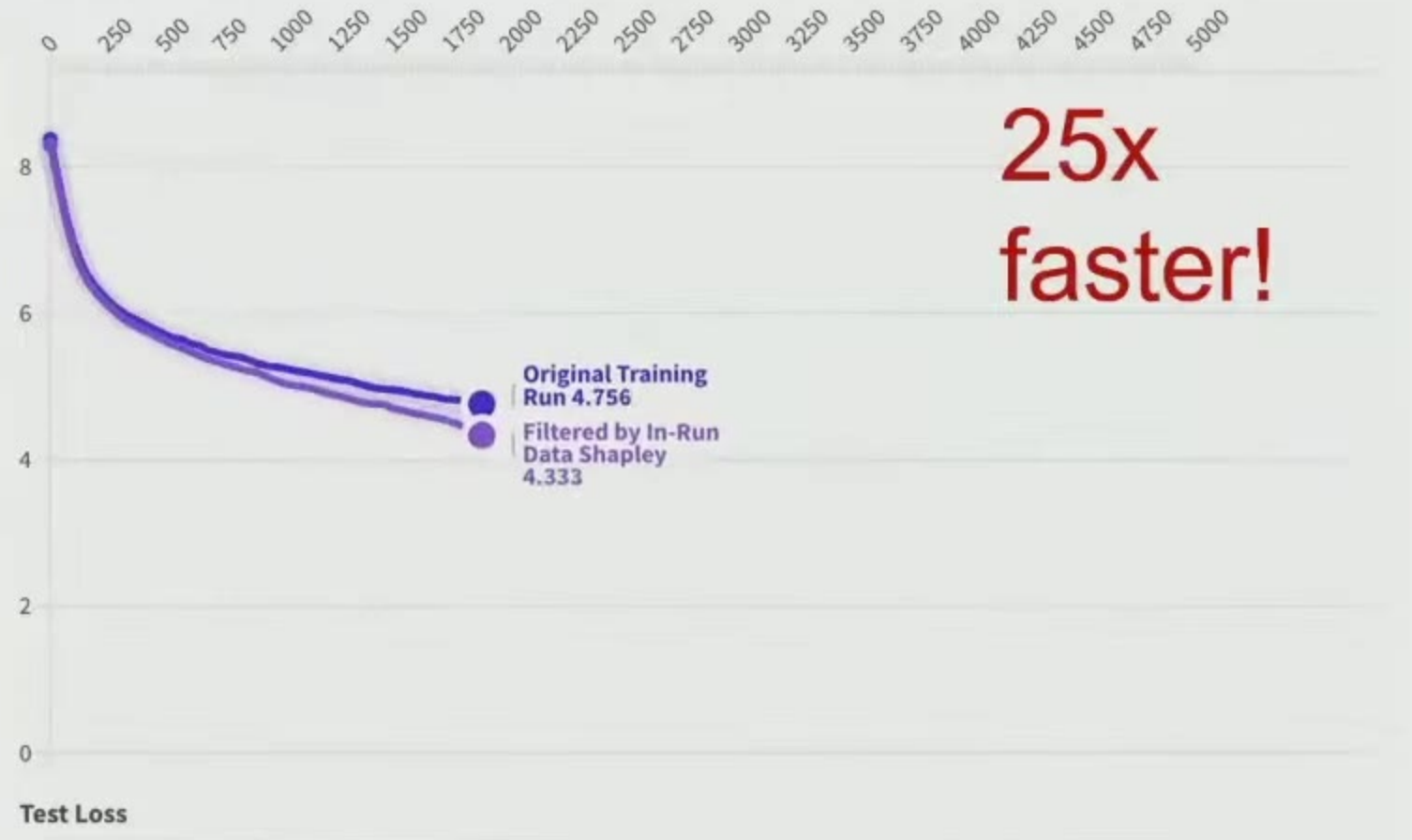


25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Removing negative-valued data impr training

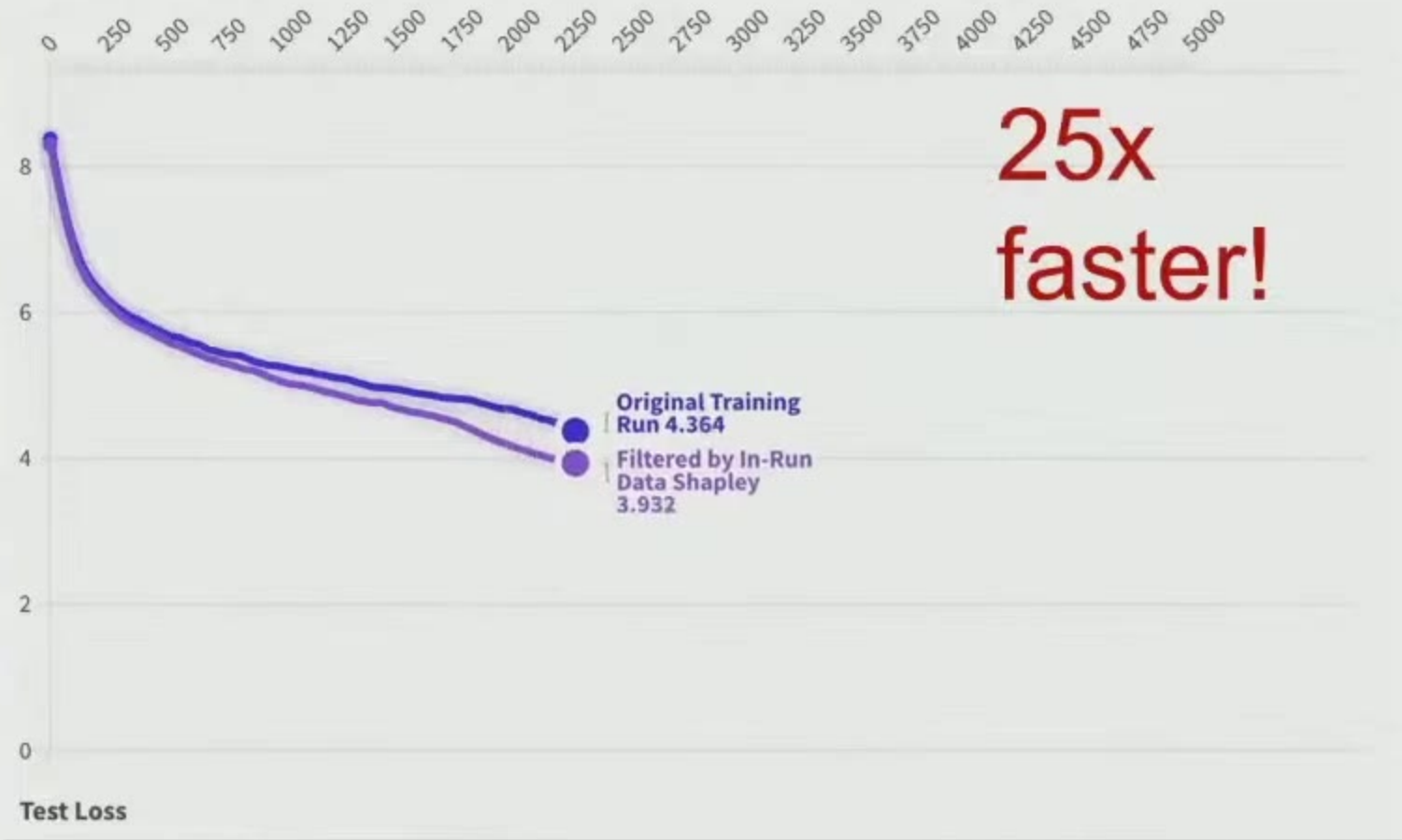


25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Removing negative-valued data impr training

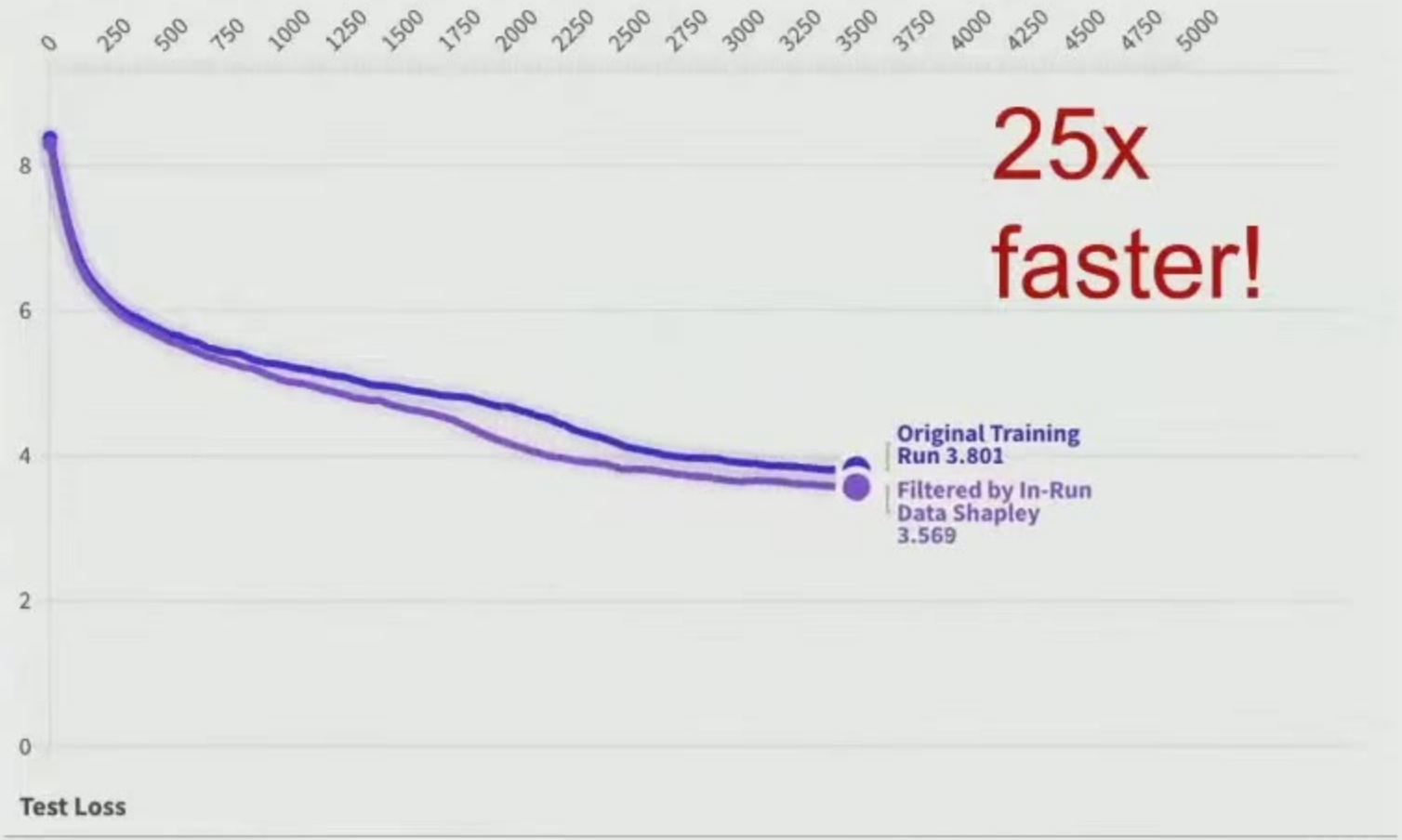


25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

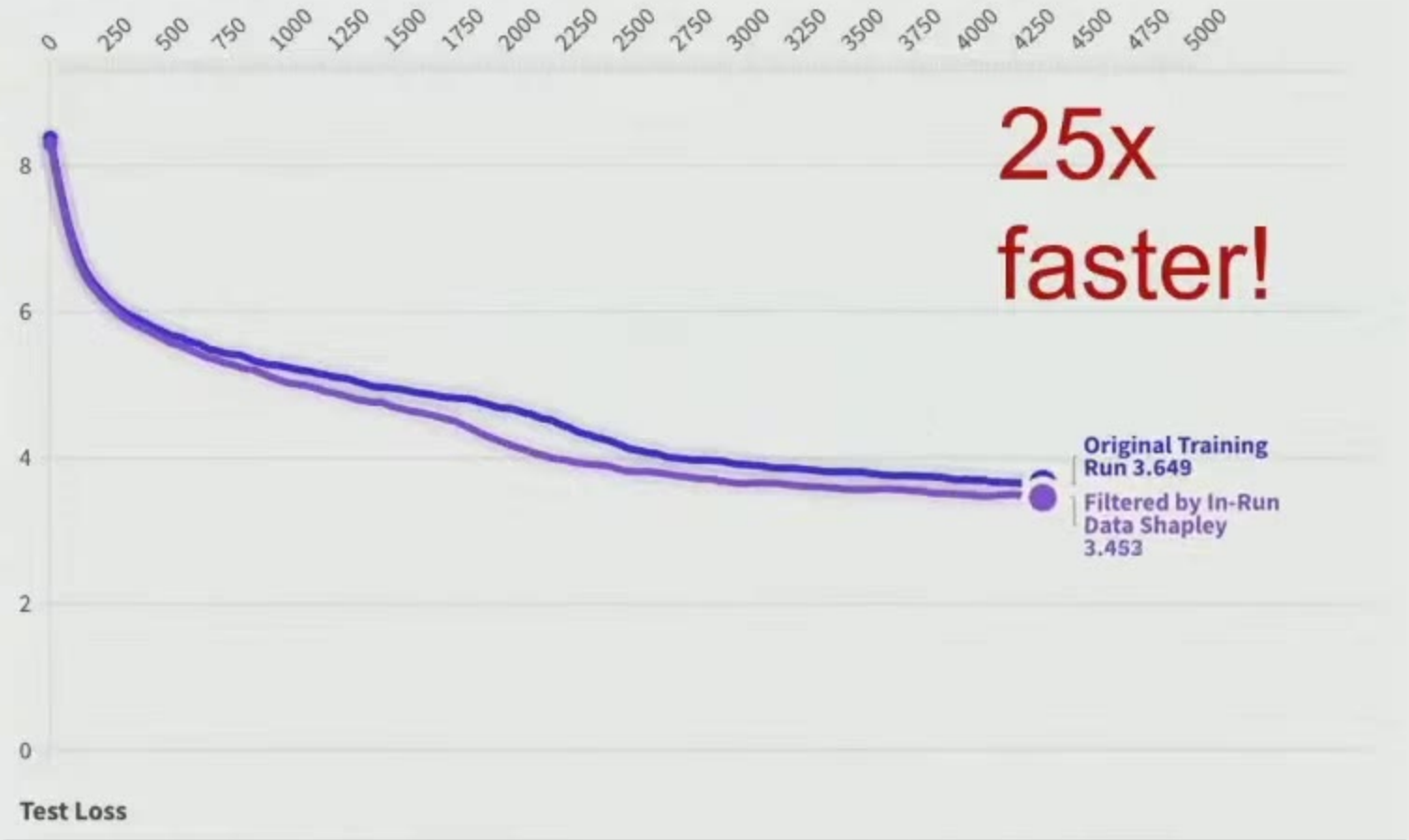
Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Removing negative-valued data impr training

25x faster!

An example of negative-valued corpus

Evaluation on the Pile Dataset

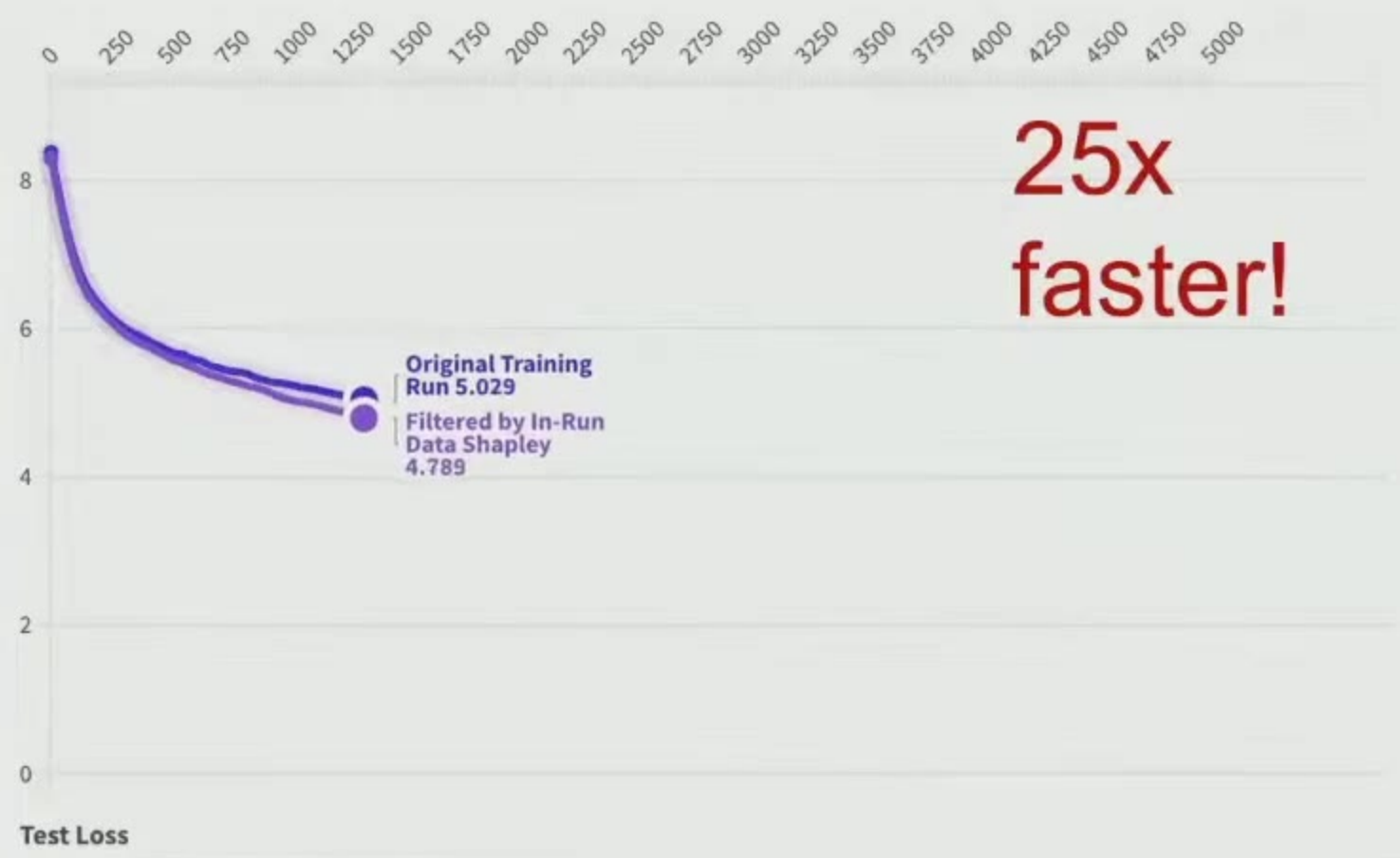
Removing negative-valued data impr training

25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Removing negative-valued data impr training



25x faster!

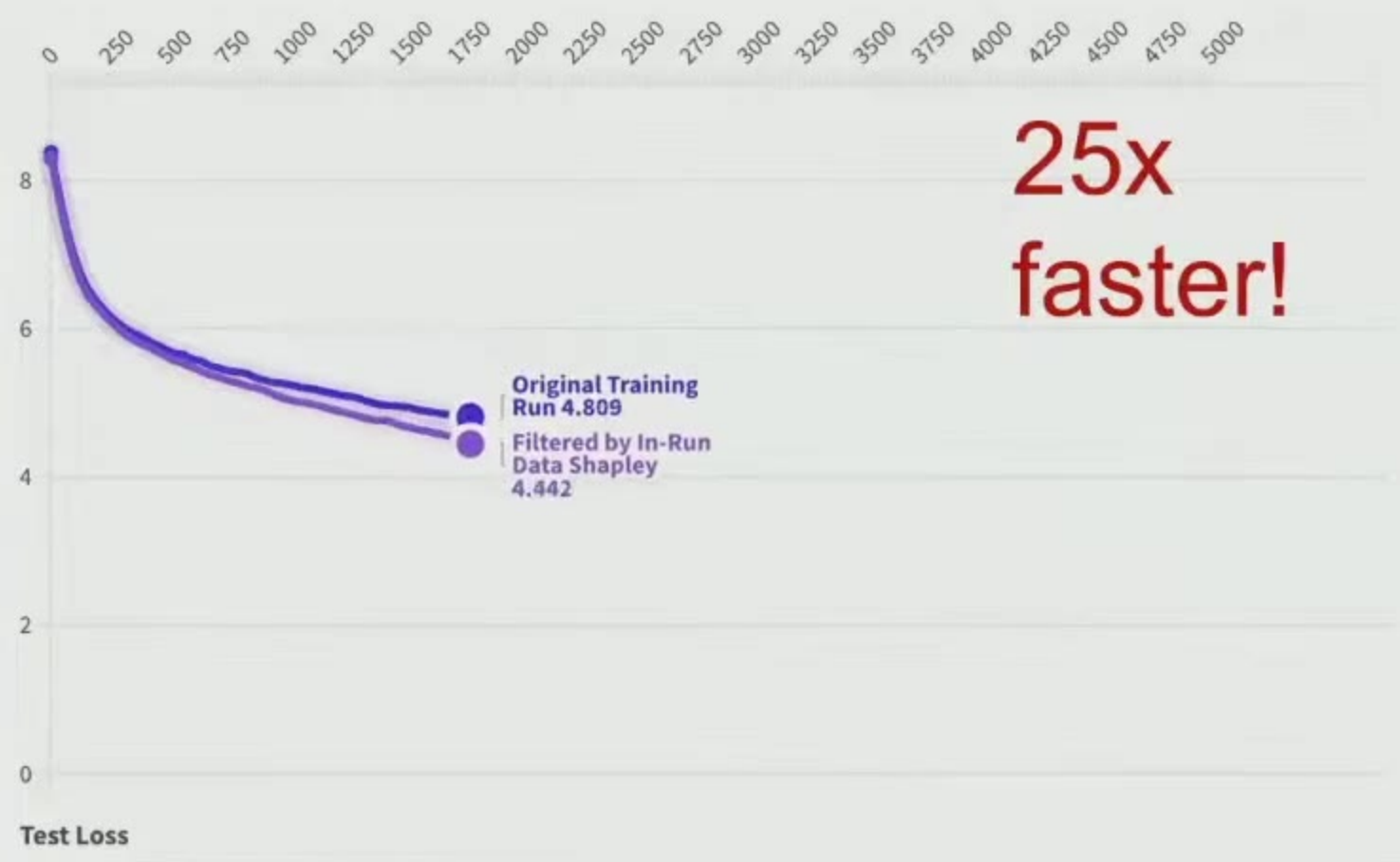
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



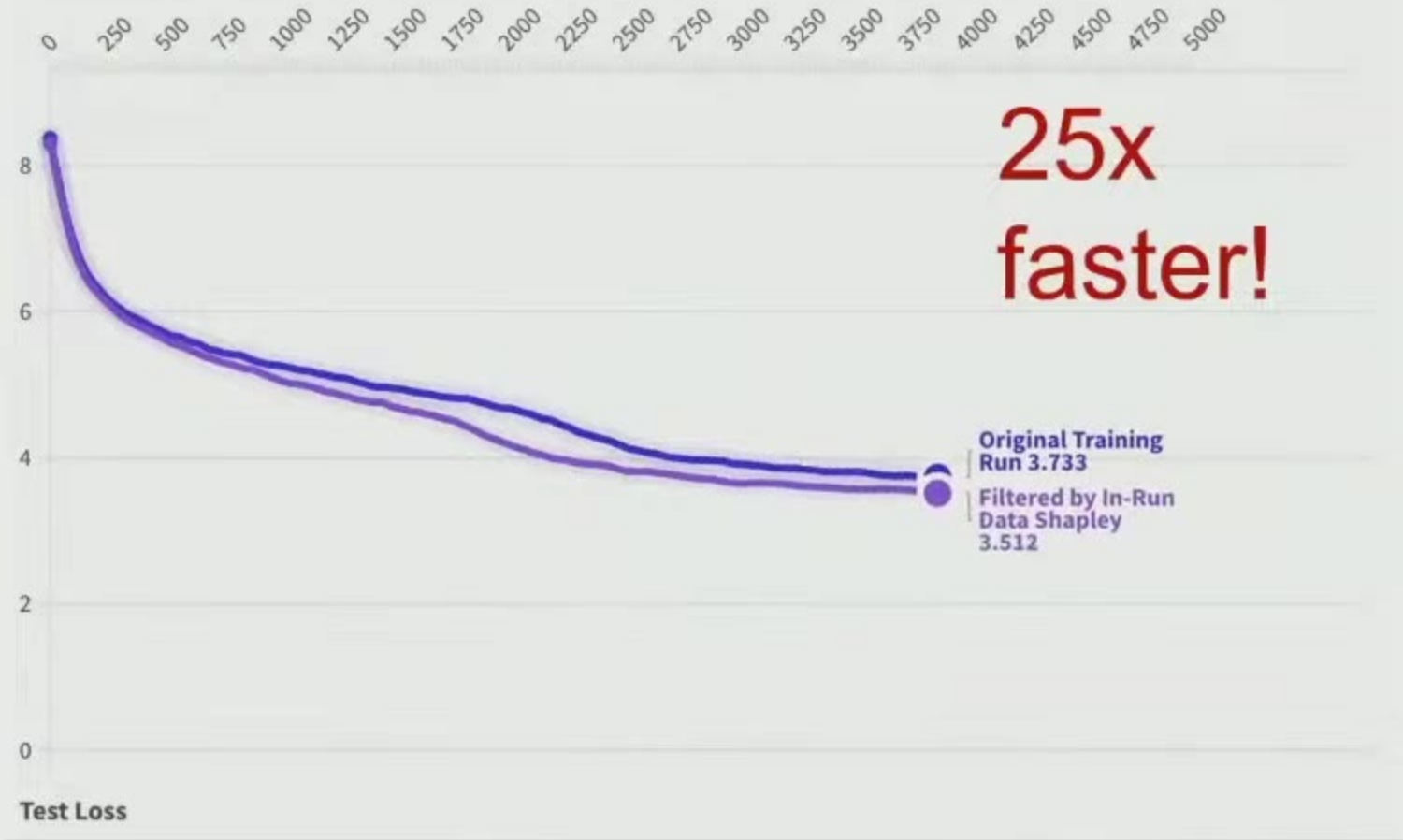
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

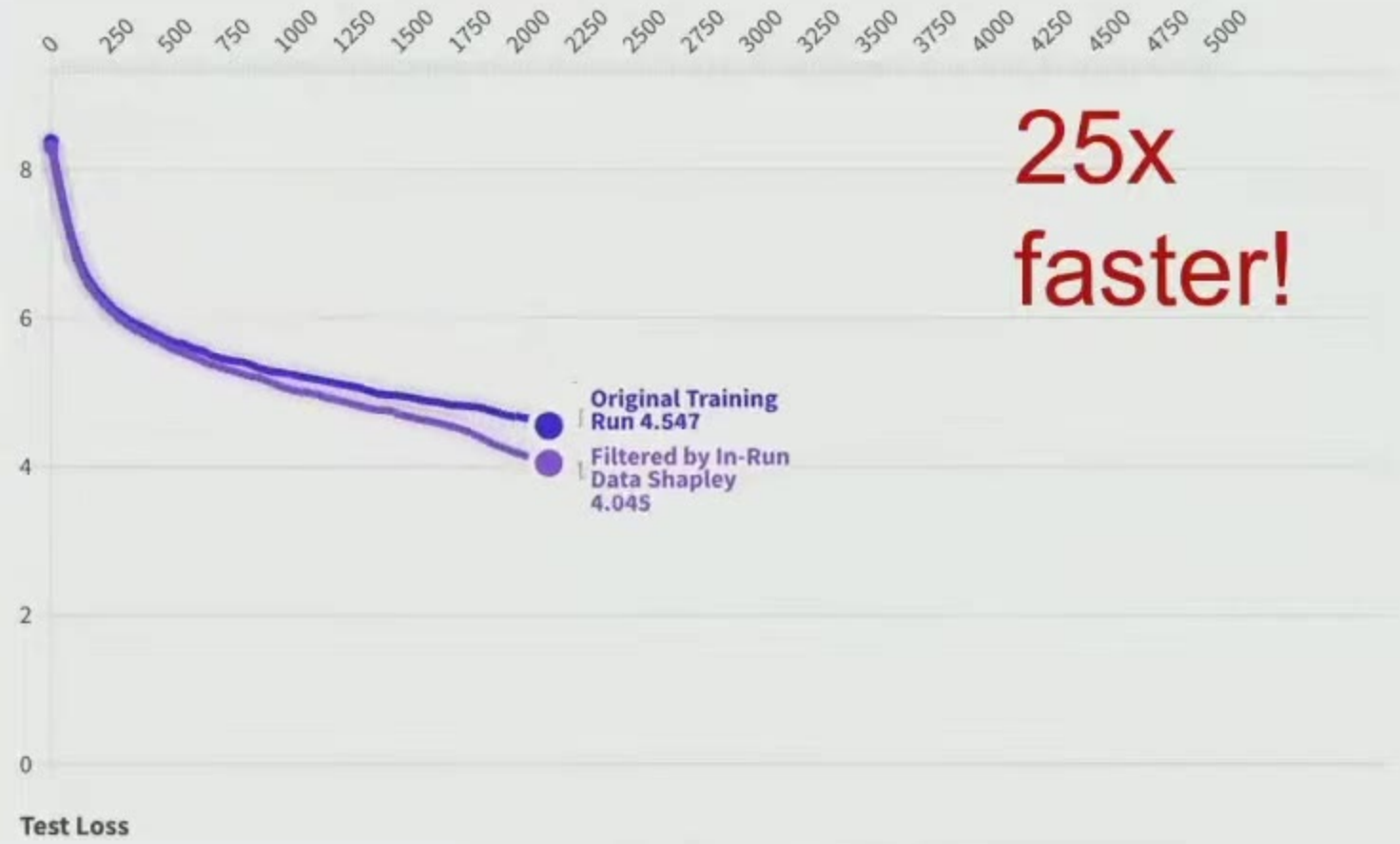
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



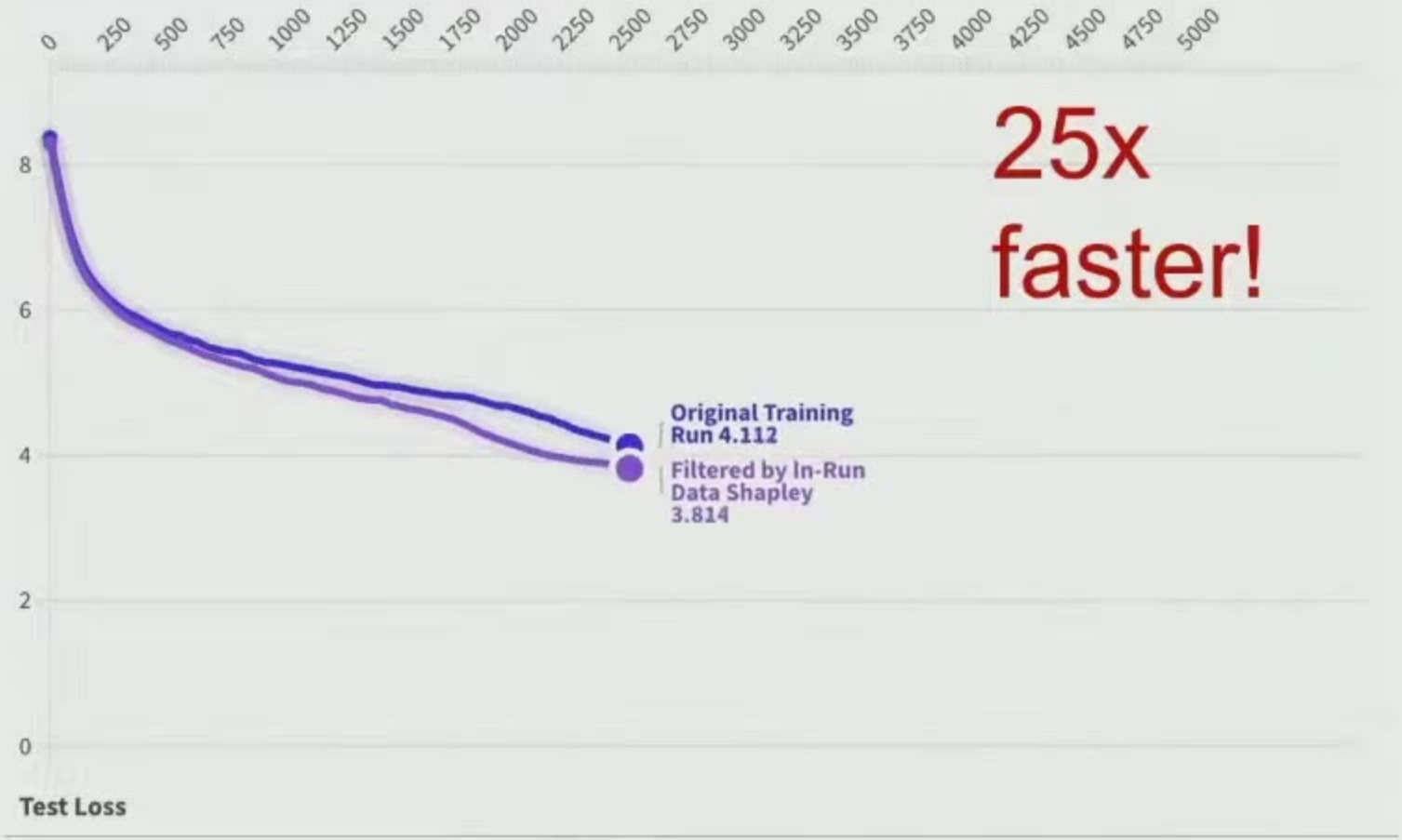
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



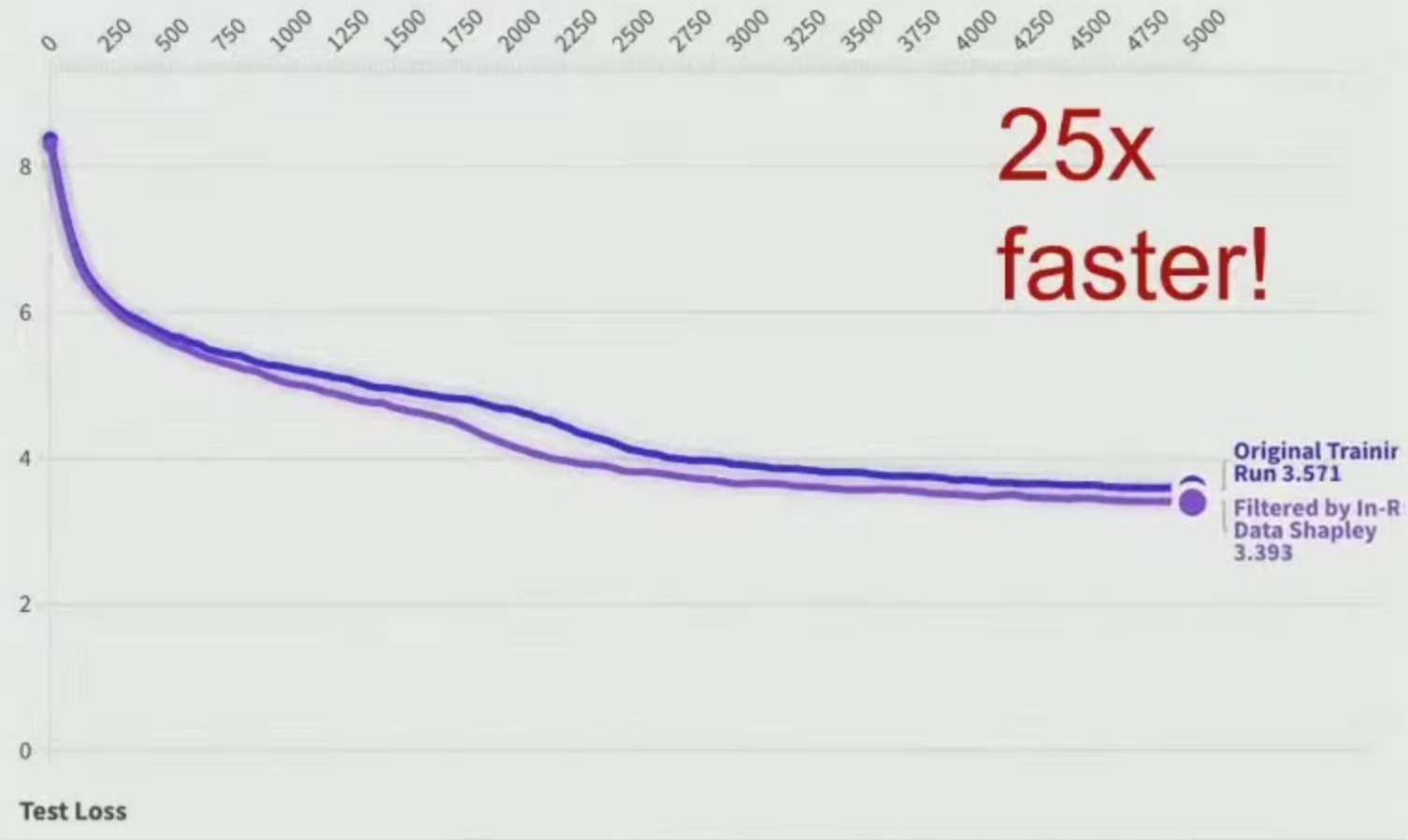
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

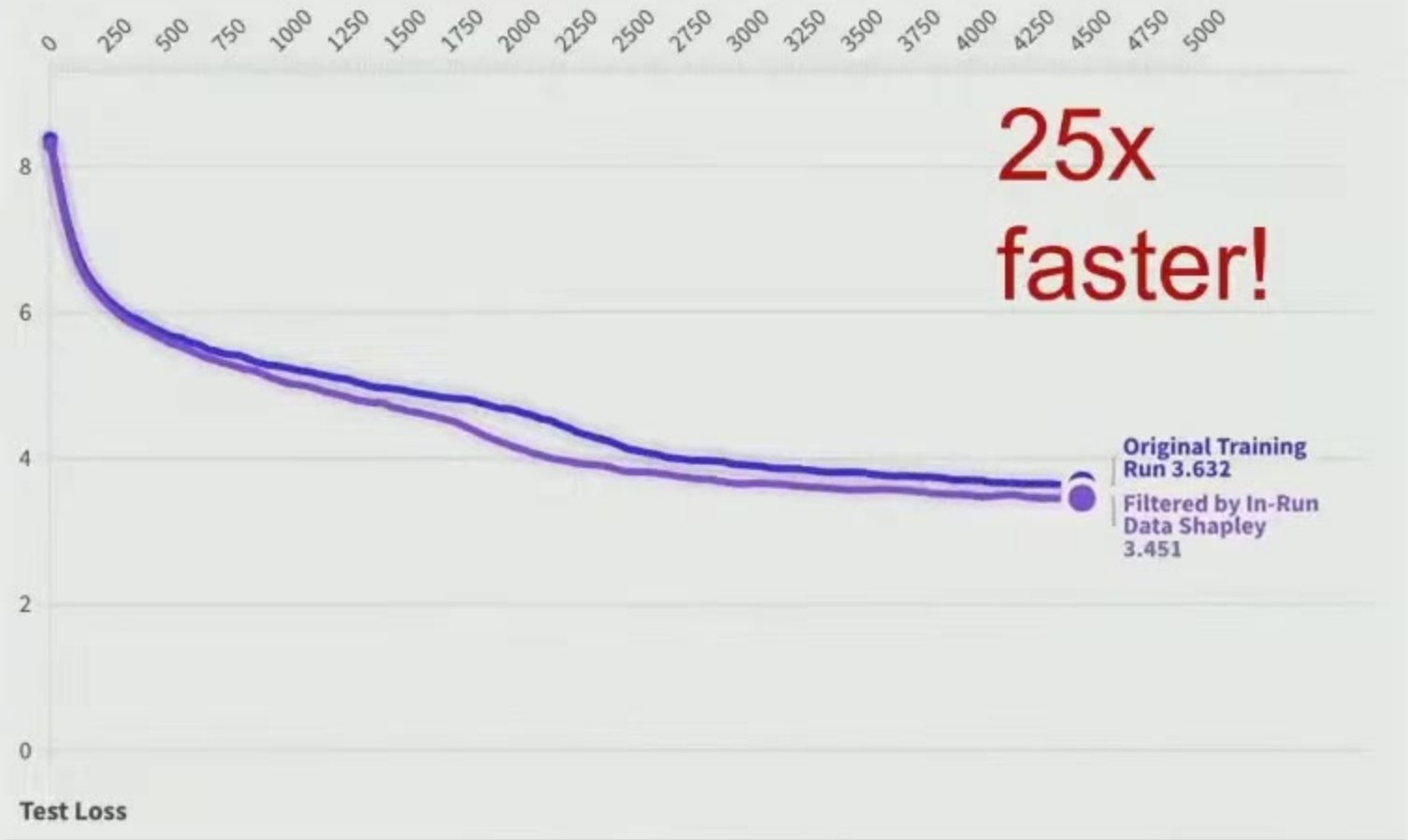
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

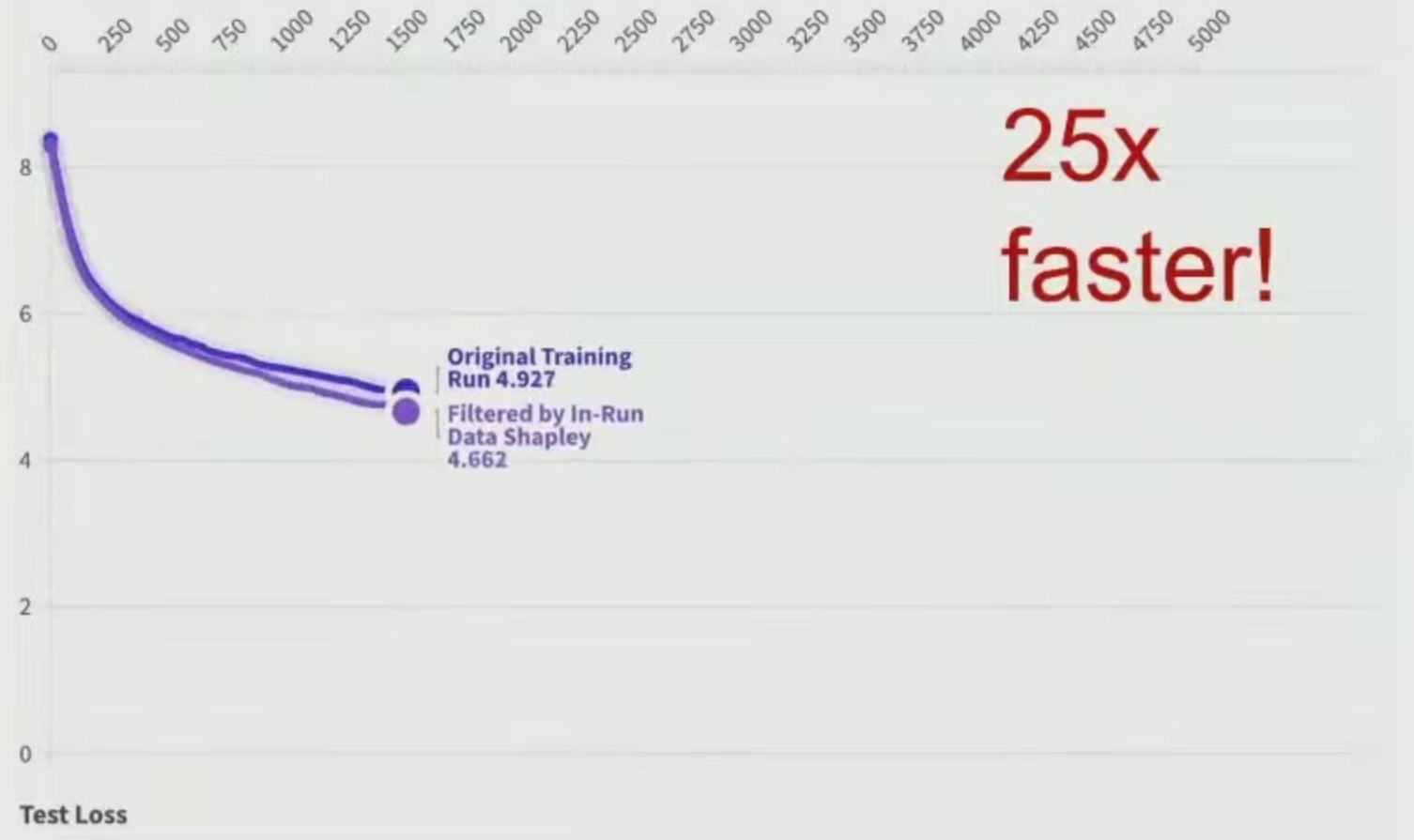
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

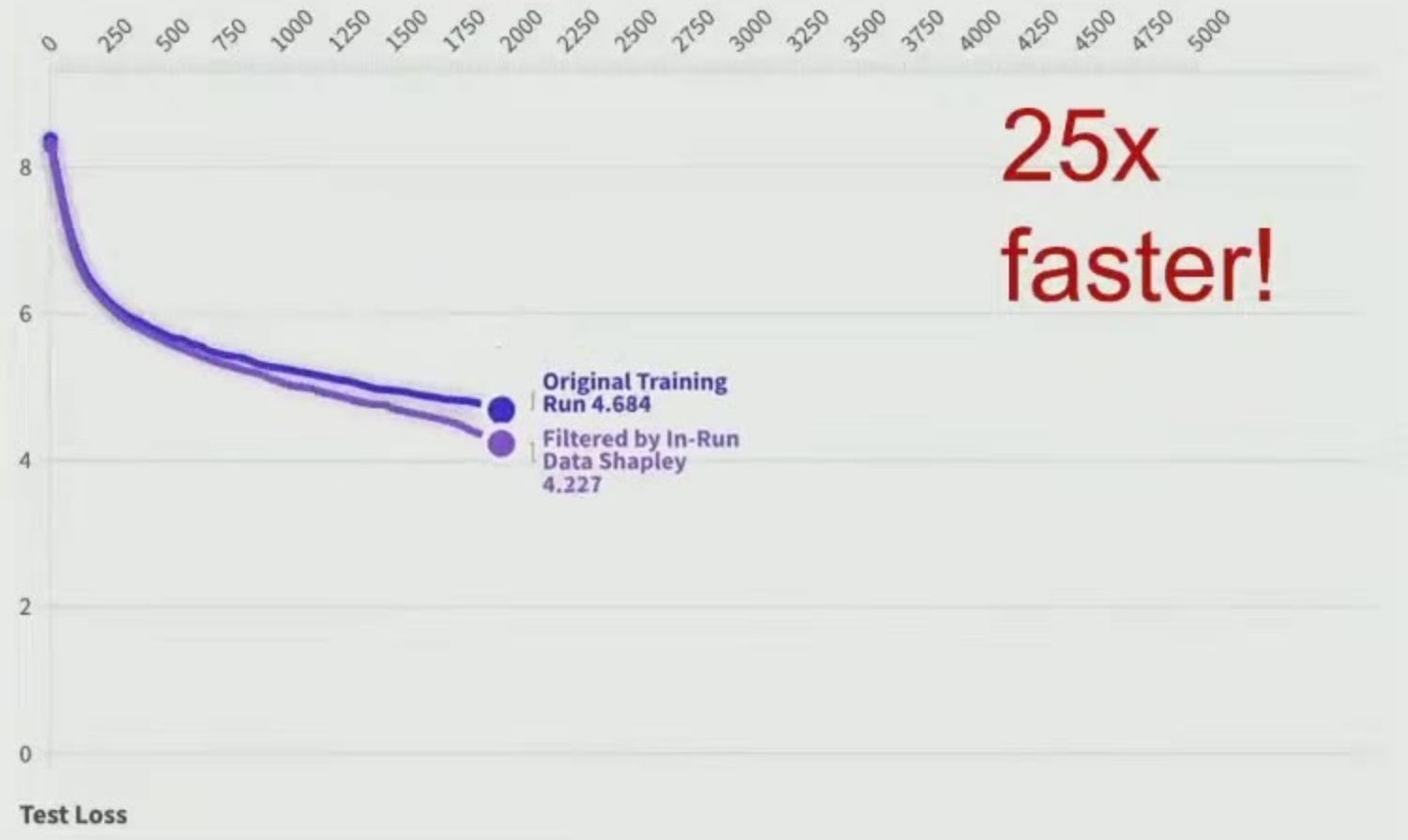
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

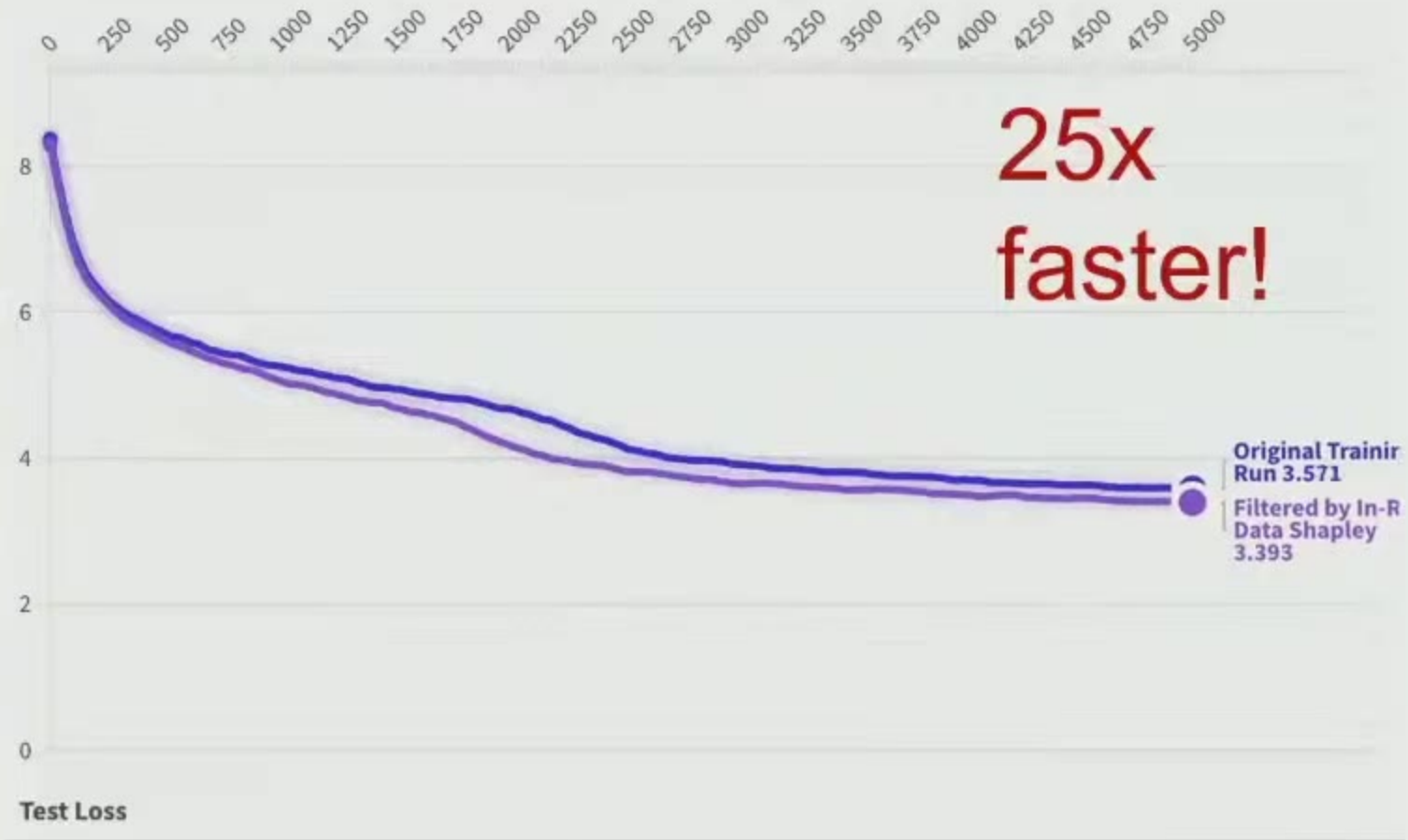
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

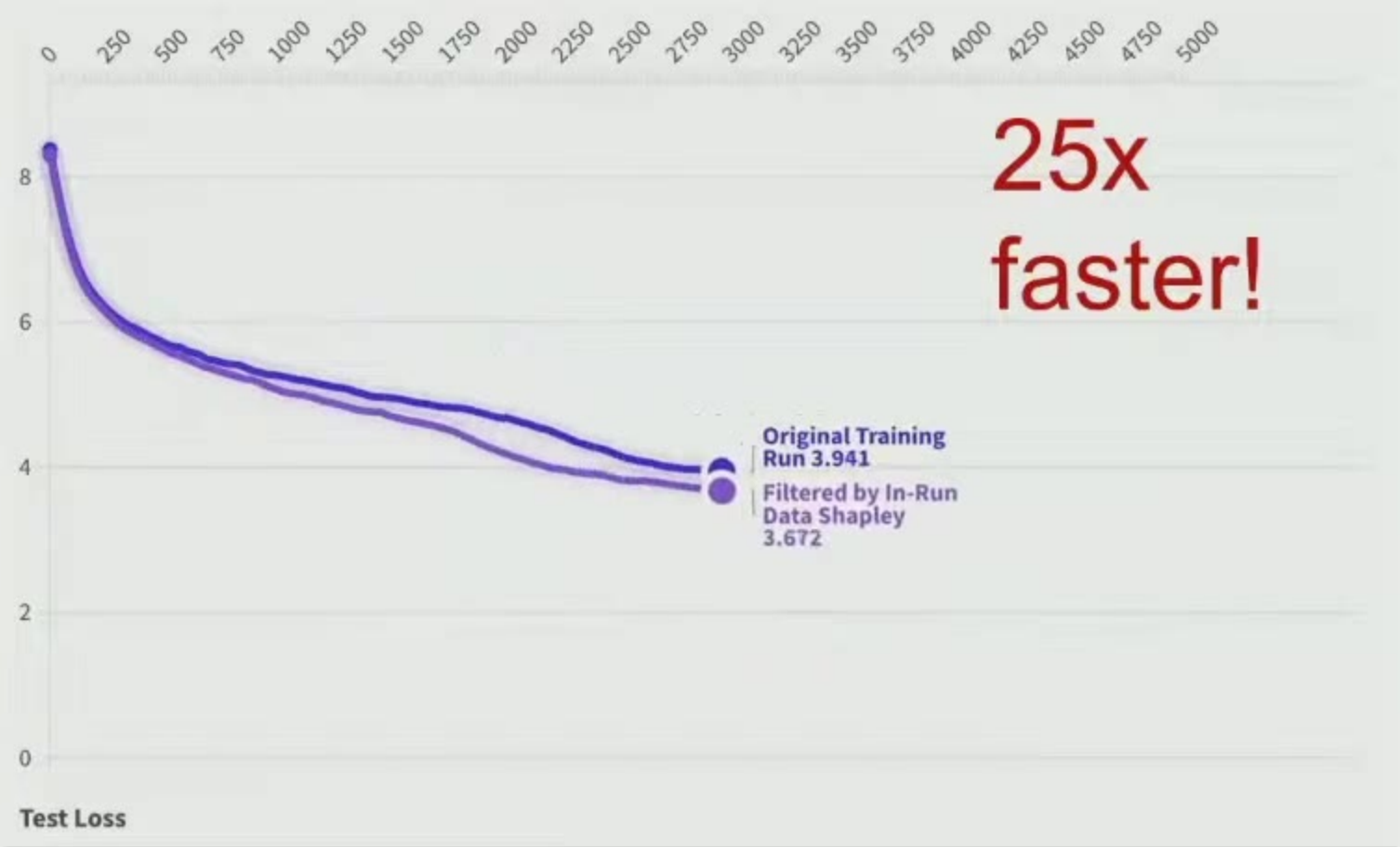
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

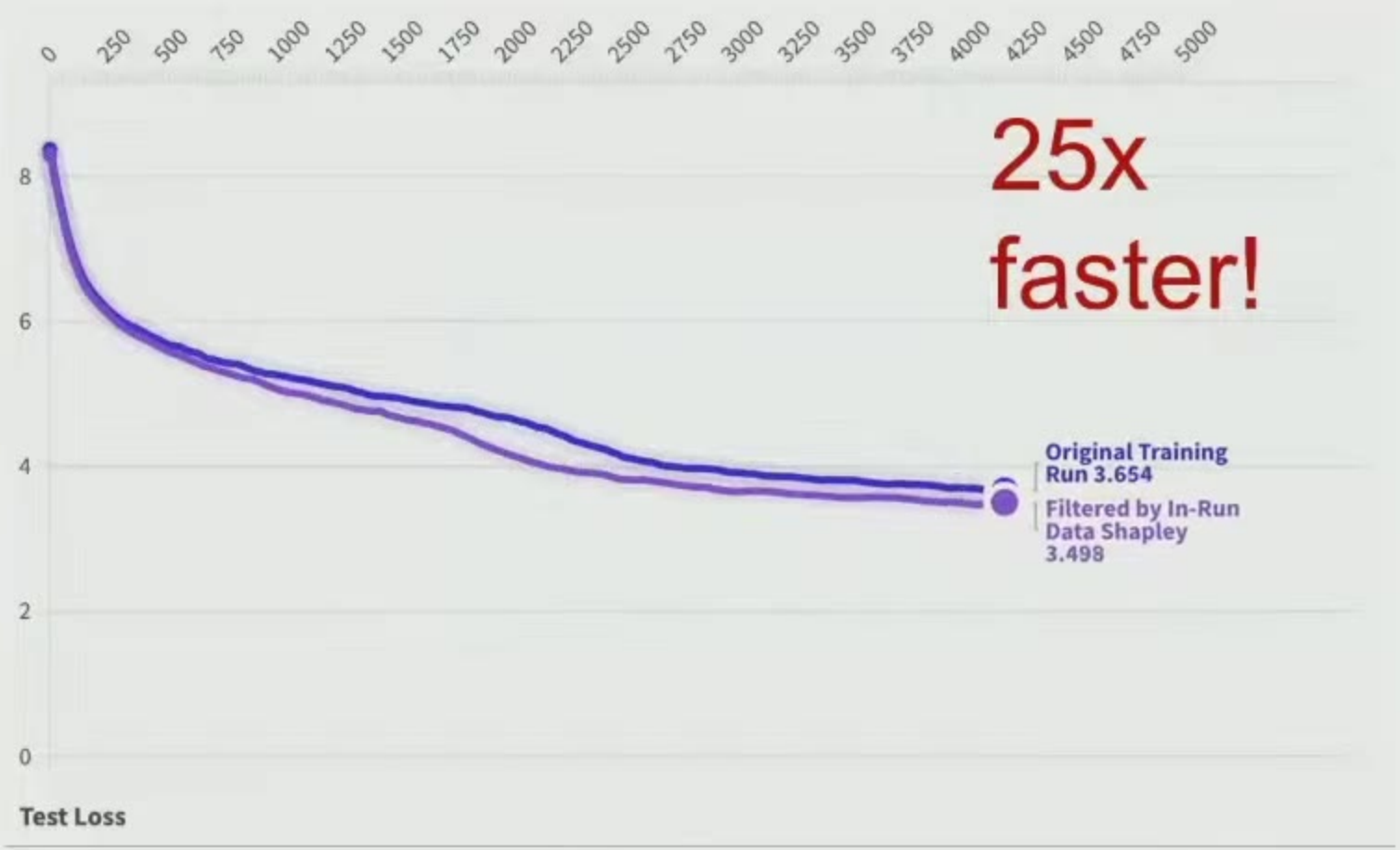
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



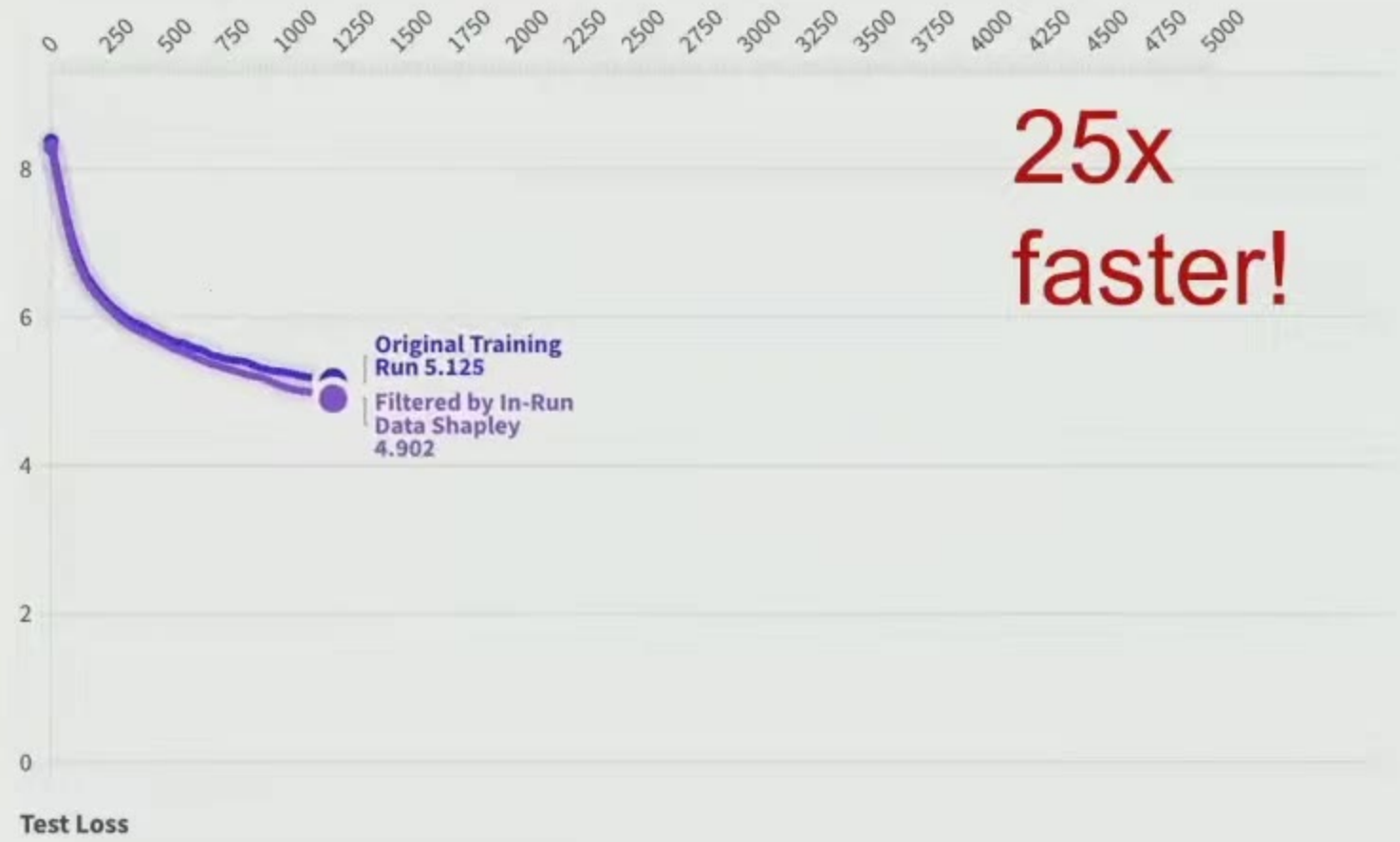
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

25x faster!

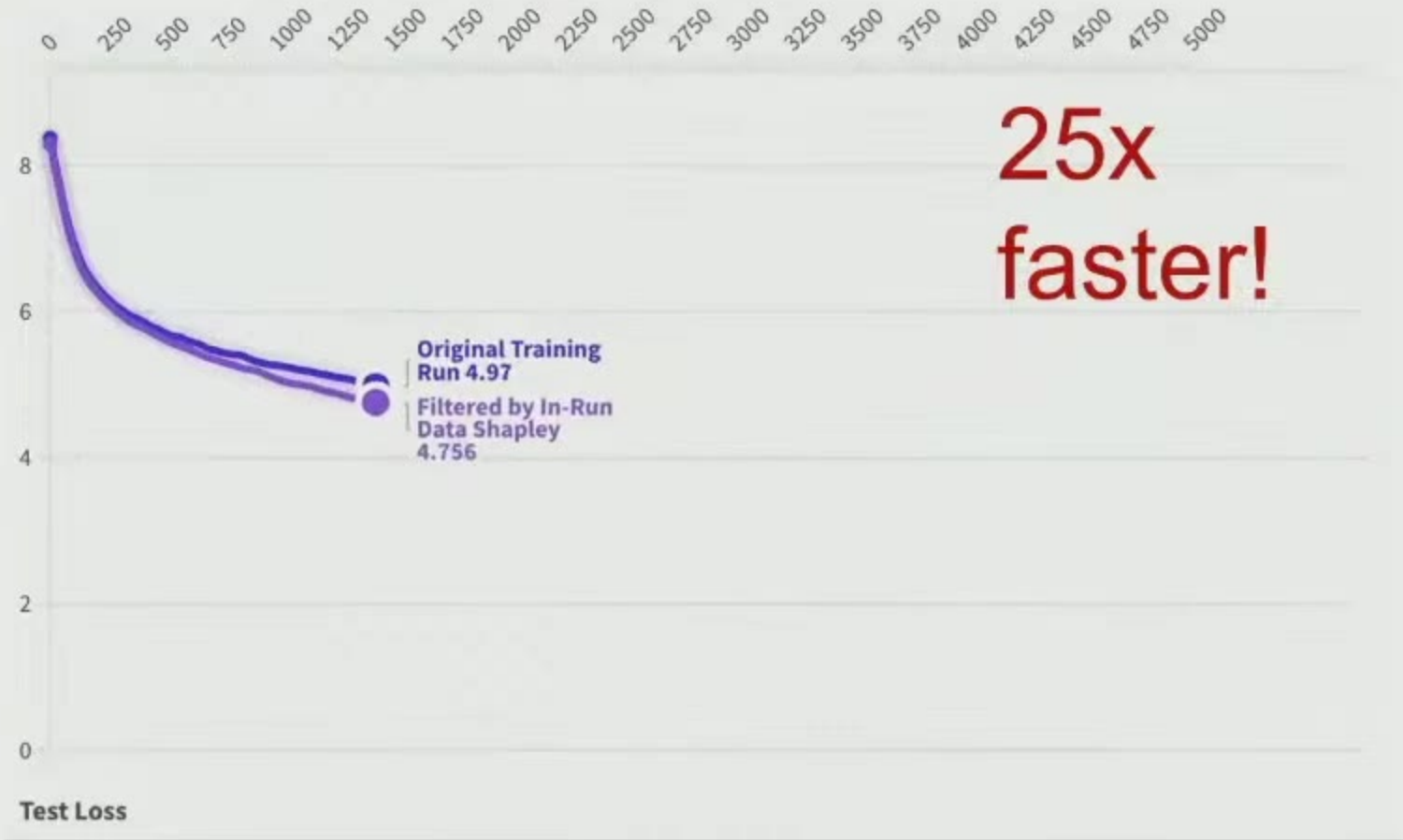
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

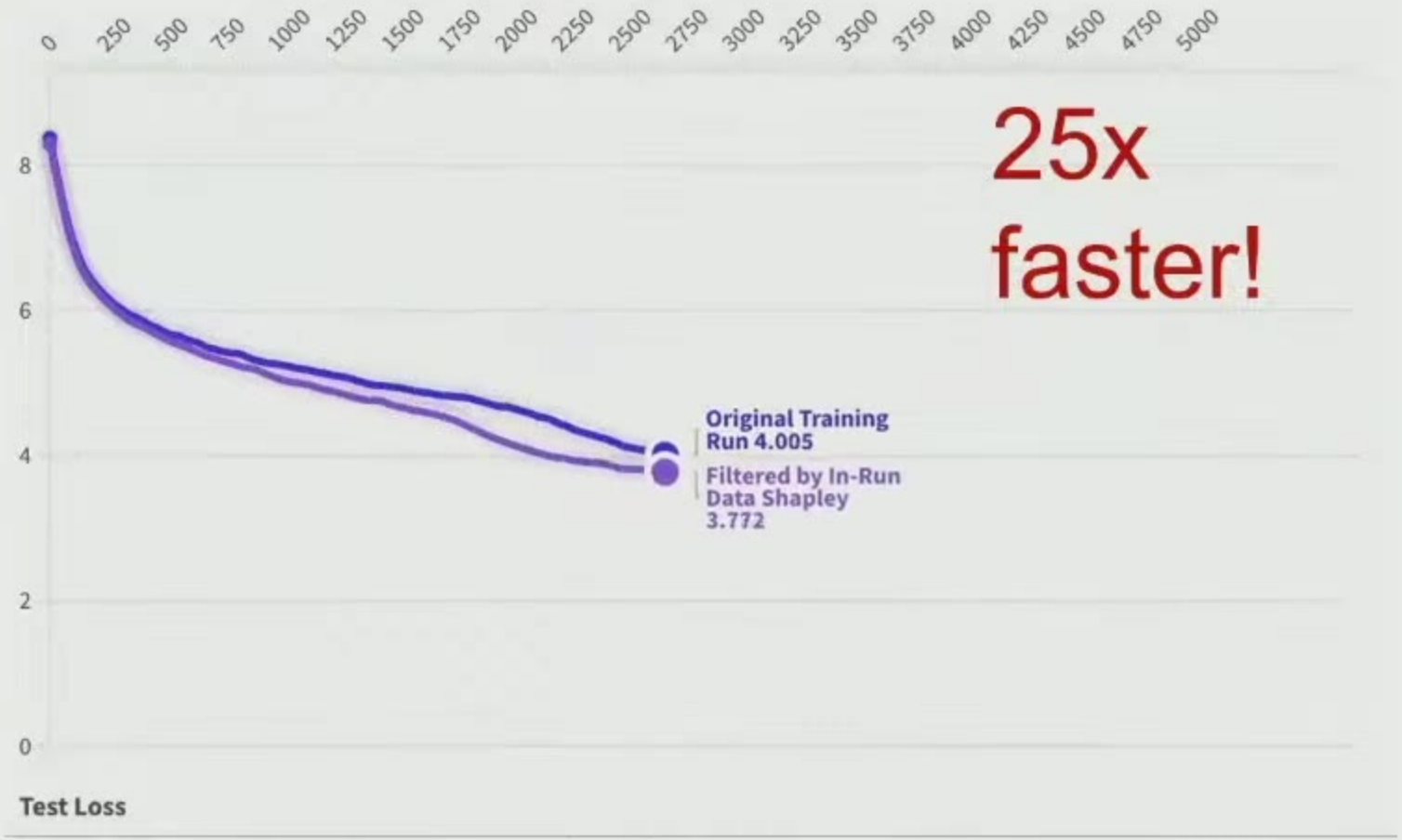
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



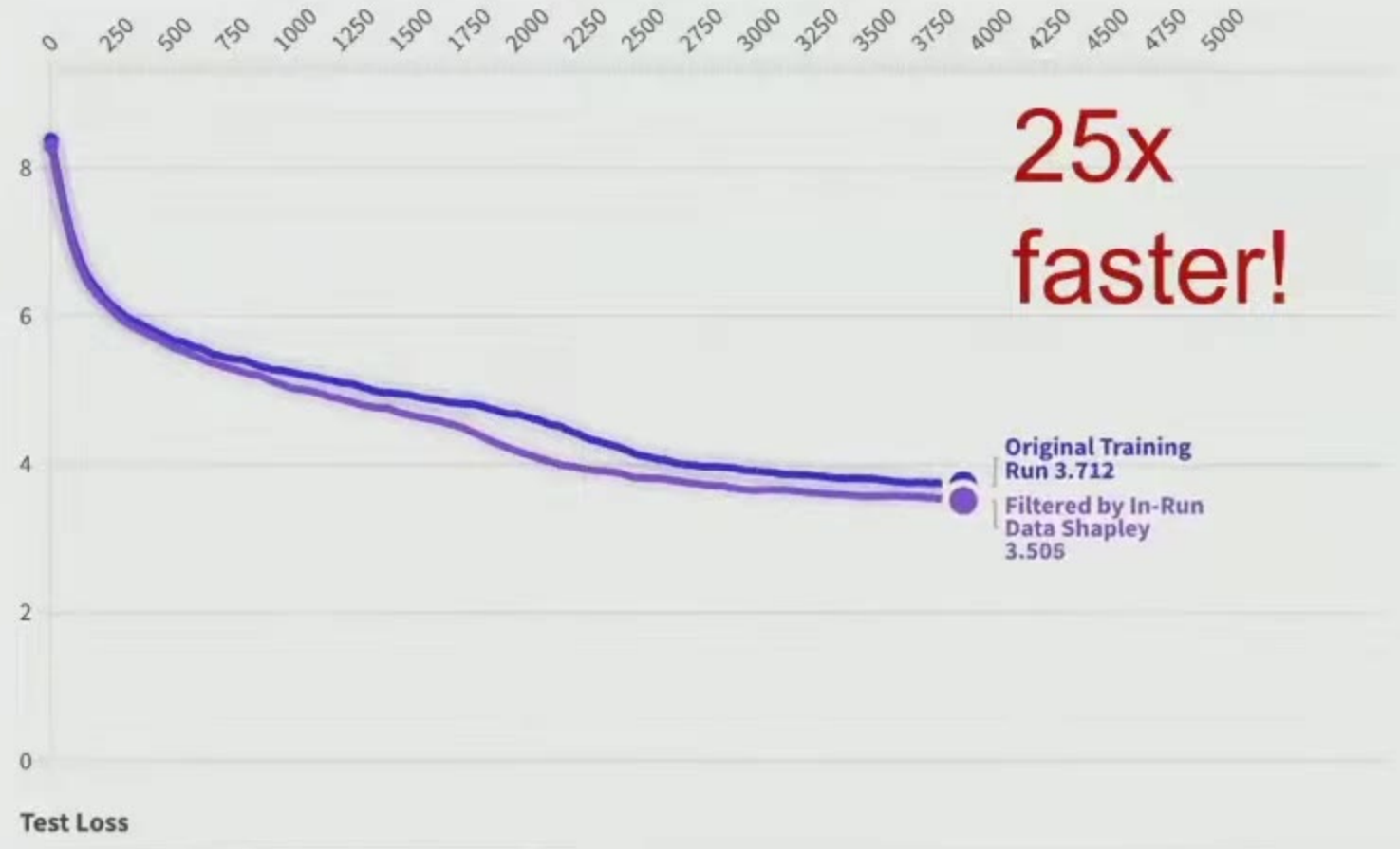
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



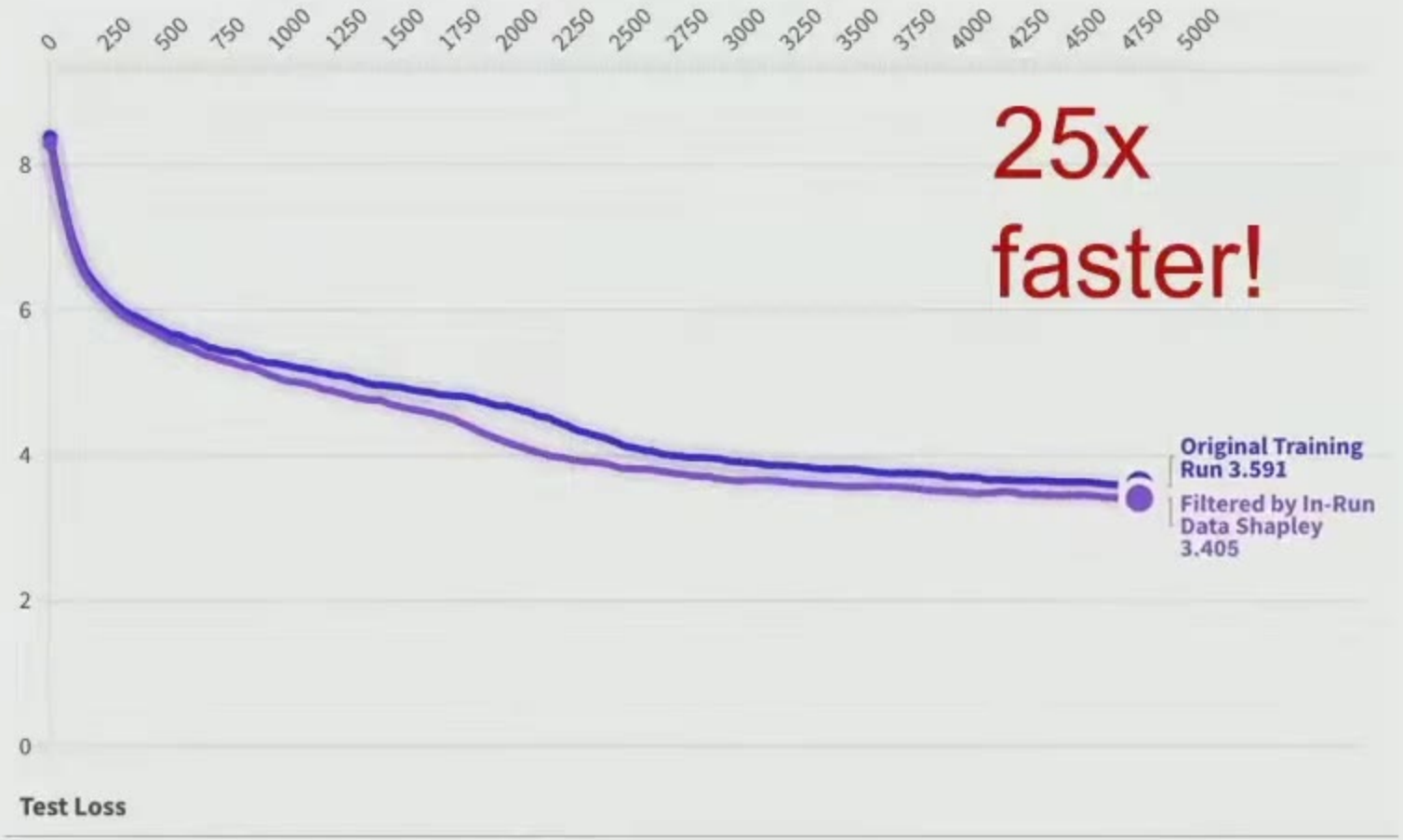
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

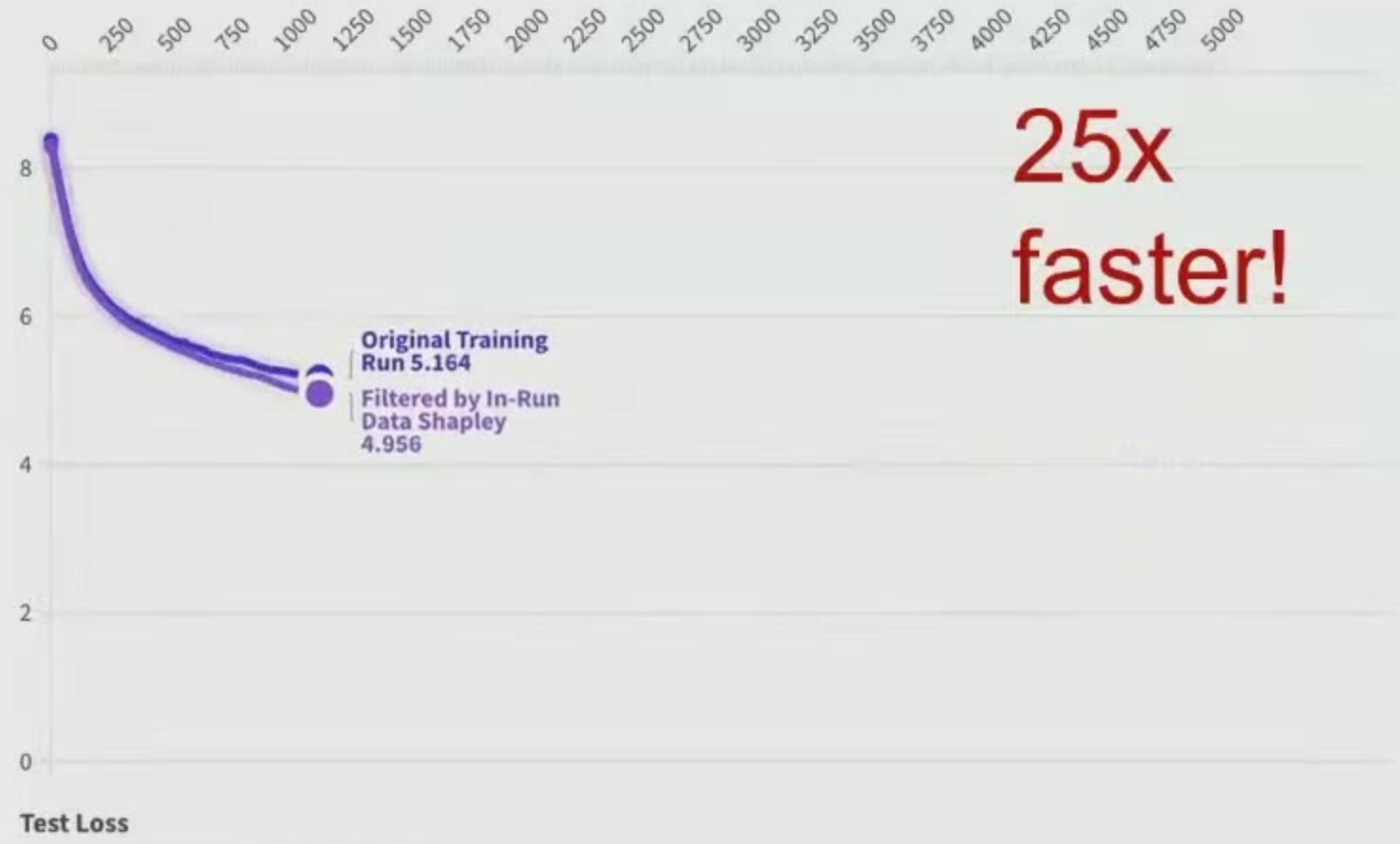
Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training



25x faster!

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation

Removing negative-valued data impr training

Evaluation on the Pile Dataset

An example of negative-valued corpus

Takeaways:

- Even well-curated pretraining corpora contain data points that negatively impact training processes
- There is significant room for improving data curation