Non-parametric Memory Matching

Allen Schmaltz

Harvard University

February 3, 2021

Version 1.0

Today

- Goal is to provide a high-level overview of the non-parametric neural matching line of work for your group, as it is generally applicable to AI/machine learning in medicine.
- Presented in the context of a new paper: “`Coarse-to-Fine Memory Matching for Joint Retrieval and Classification`”
  - Not directly a medical task, but easy to understand setting regardless of expertise/focus/area
  - Many real-world tasks fit this setup

Plan

Plan

1. Highlight challenges of ML in medicine with standard approaches

Plan

1. Highlight challenges of ML in medicine with standard approaches

2. High-level overview of alternative: Exemplar auditing

Plan

1. Highlight challenges of ML in medicine with standard approaches
2. High-level overview of alternative: Exemplar auditing

Motivates subsequent modeling decisions

Plan

1. Highlight challenges of ML in medicine with standard approaches
2. High-level overview of alternative: Exemplar auditing
3. Examine an approach for a retrieval-classification task: fact verification

Motivates subsequent modeling decisions

Plan

1. Highlight challenges of ML in medicine with standard approaches
2. High-level overview of alternative: Exemplar auditing
3. Examine an approach for a retrieval-classification task: fact verification
4. As a result, we get an updatable sequence/language model via 2 mechanisms:

Motivates subsequent modeling decisions

Plan

1. Highlight challenges of ML in medicine with standard approaches
2. High-level overview of alternative: Exemplar auditing
3. Examine an approach for a retrieval-classification task: fact verification
4. As a result, we get an updatable sequence/language model via 2 mechanisms:
    1. The datastore of retrieved information can be updated

Motivates subsequent modeling decisions

Plan

1. Highlight challenges of ML in medicine with standard approaches
2. High-level overview of alternative: Exemplar auditing
3. Examine an approach for a retrieval-classification task: fact verification
4. As a result, we get an updatable sequence/language model via 2 mechanisms:
    1. The datastore of retrieved information can be updated
    2. The more abstract model behavior can be updated via a dense database

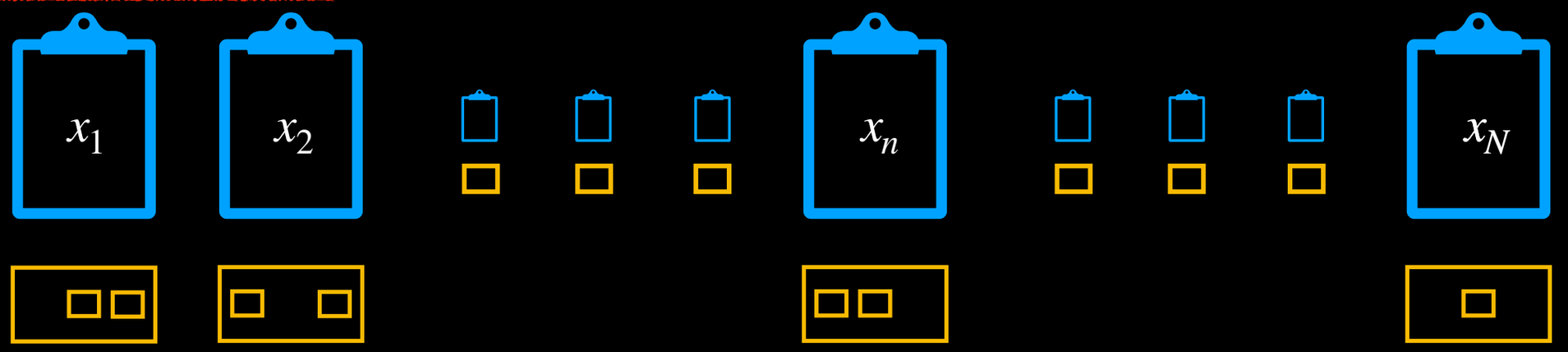
Motivates subsequent modeling decisions

Standard Setup — Learning

- N training instances: $x_1, \ldots, x_N$
- Ground truth training labels: $y_1, \ldots, y_N$
- Seek a function, $f : X \rightarrow Y$, to predict $\hat{y}_{N+1}$ for a new, unseen instance x_{N+1}, with minimal *distance* between $\hat{y}_{N+1}$ and y_{N+1}

Standard Classification Setting

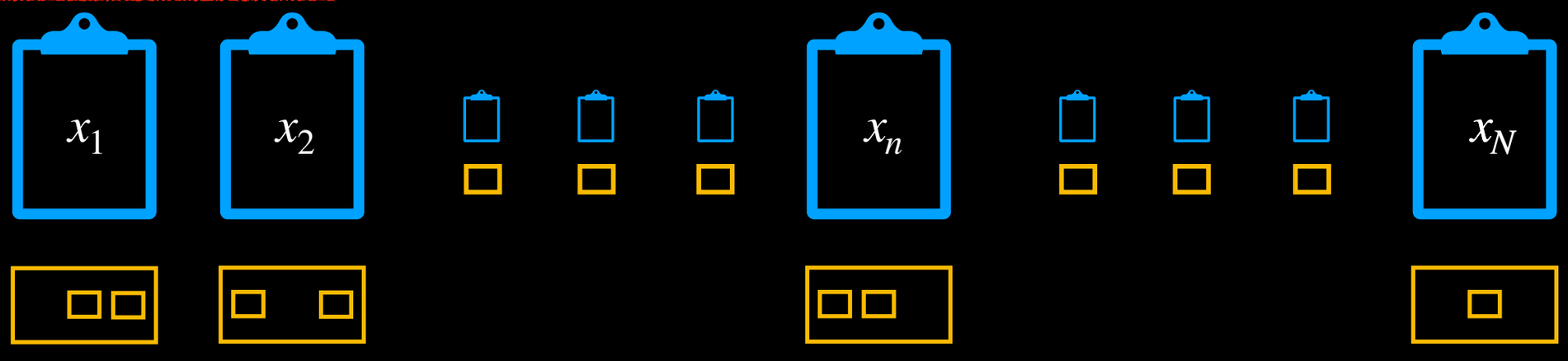
For Training



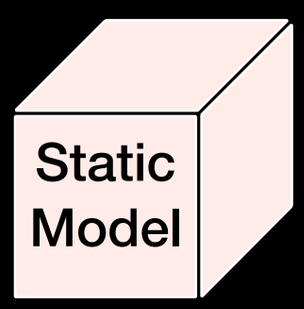
Ground-truth Label Sets

Standard Classification Setting

For Training

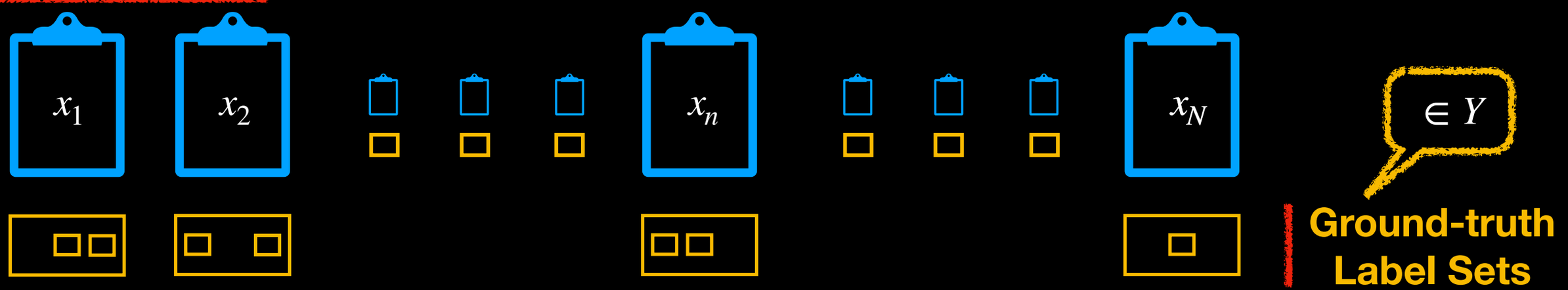


After Training

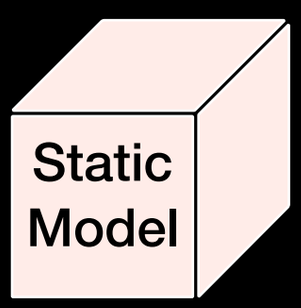


Standard Classification Setting

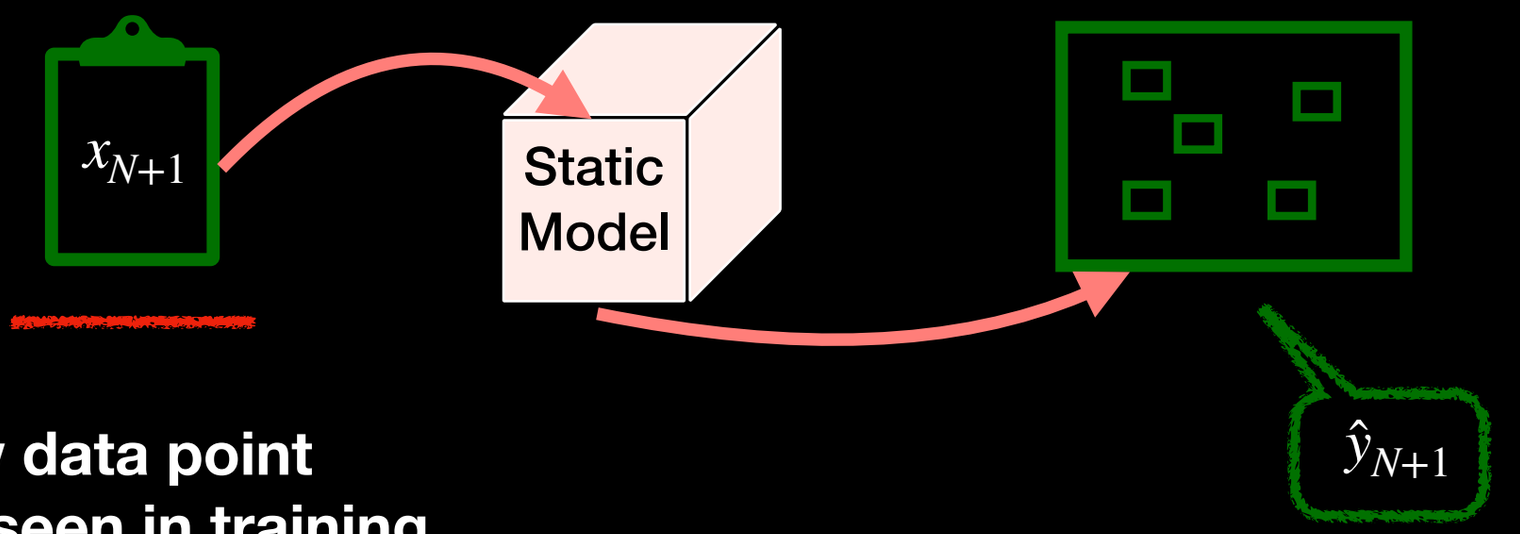
For Training



After Training



Test/inference Time



New data point not seen in training

Challenges of AI/ML in Medicine

- Difficult to understand models
    - Parameters are not identifiable
- Data issues
    - Often reliable subsets not sufficient for training neural models
    - Annotation error costs are significant for high-risk areas
- Opaque models + data issues == Volatile mix in high-risk settings

Argument

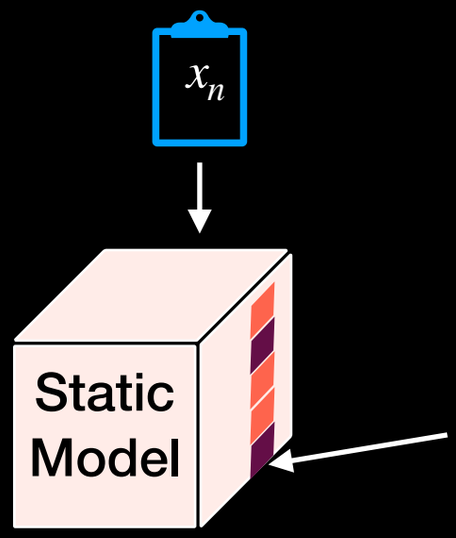
- We need a paradigm shift in our conceptual framework of neural nets applied to medicine (& other higher-risk areas), both for analyzing networks AND for analyzing data

New Approach:
Exemplar Auditing

New Approach: Exemplar Auditing

New Approach: Exemplar Auditing

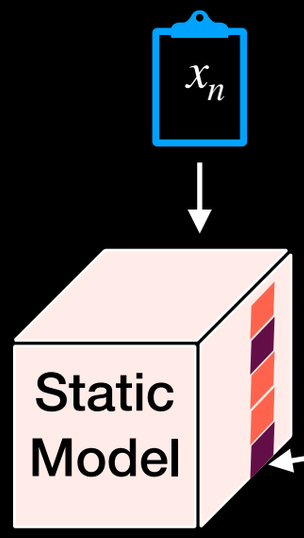
1. Train the model such that ‘exemplar’ vectors summarize the network



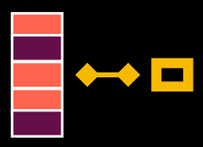
Exemplar vector (dense representation/‘distilled memory’) of the full parametric network

New Approach: Exemplar Auditing

1. Train the model such that ‘exemplar’ vectors summarize the network



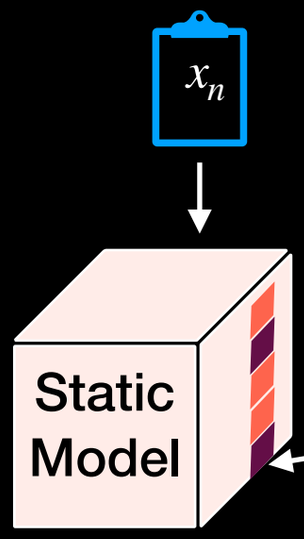
Exemplar vector (dense representation/‘distilled memory’) of the full parametric network



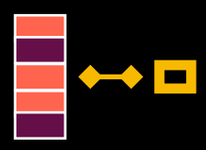
Training exemplar vectors are associated with ground-truth labels

New Approach: Exemplar Auditing

1. Train the model such that ‘exemplar’ vectors summarize the network

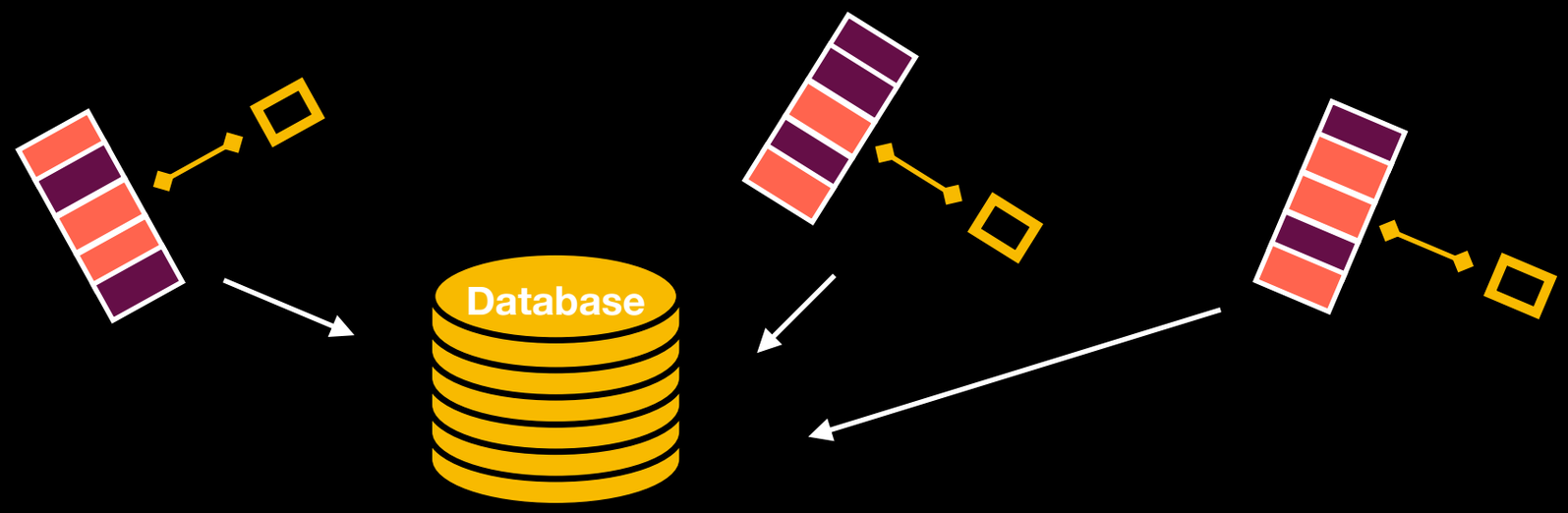


Exemplar vector (dense representation/‘distilled memory’) of the full parametric network



Training exemplar vectors are associated with ground-truth labels

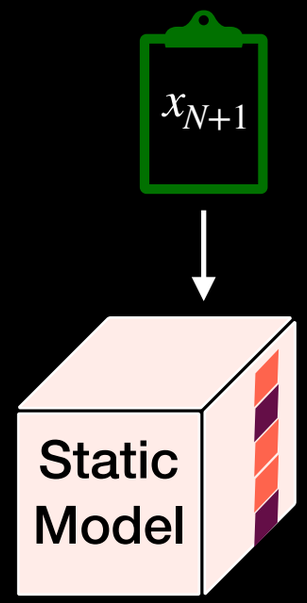
2. For all training instances, add exemplars & labels to a database



New Approach: Exemplar Auditing

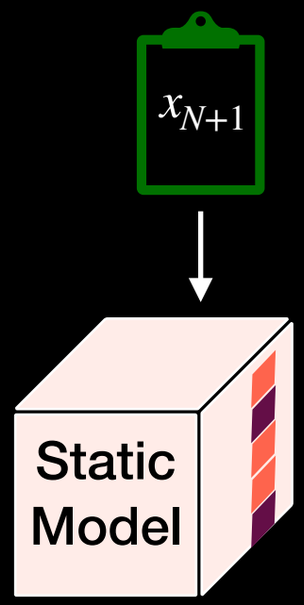
New Approach: Exemplar Auditing

3. At test, create exemplar vector

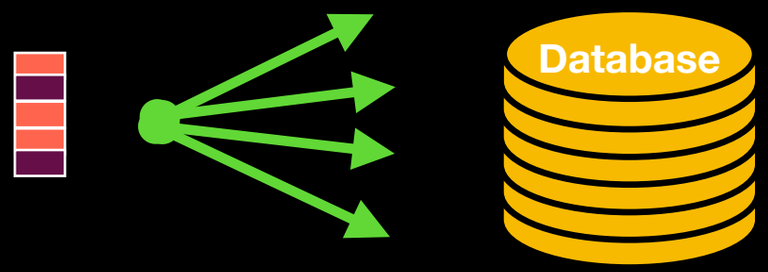


New Approach: Exemplar Auditing

3. At test, create exemplar vector

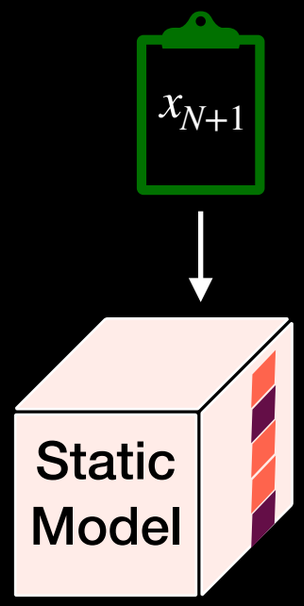


4. Match test exemplar vector to database

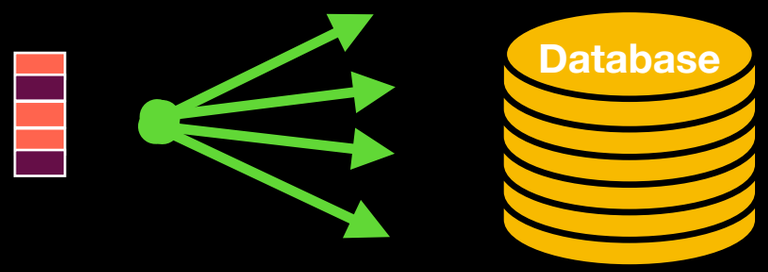


New Approach: Exemplar Auditing

3. At test, create exemplar vector



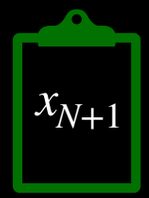
4. Match test exemplar vector to database

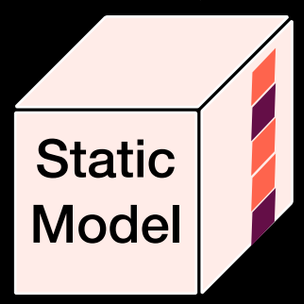


5. Constrain prediction based on nearest label (or distribution of labels) in database
-Can also leverage the distances to the exemplars

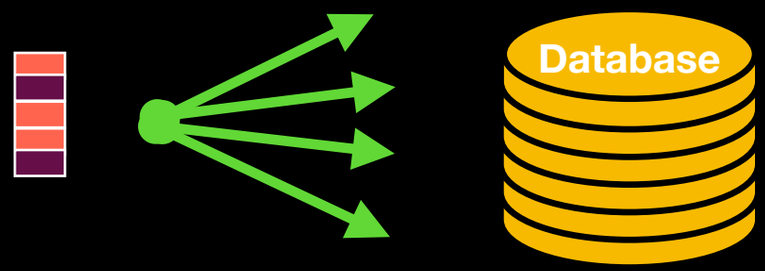
New Approach: Exemplar Auditing

3. At test, create exemplar vector





4. Match test exemplar vector to database



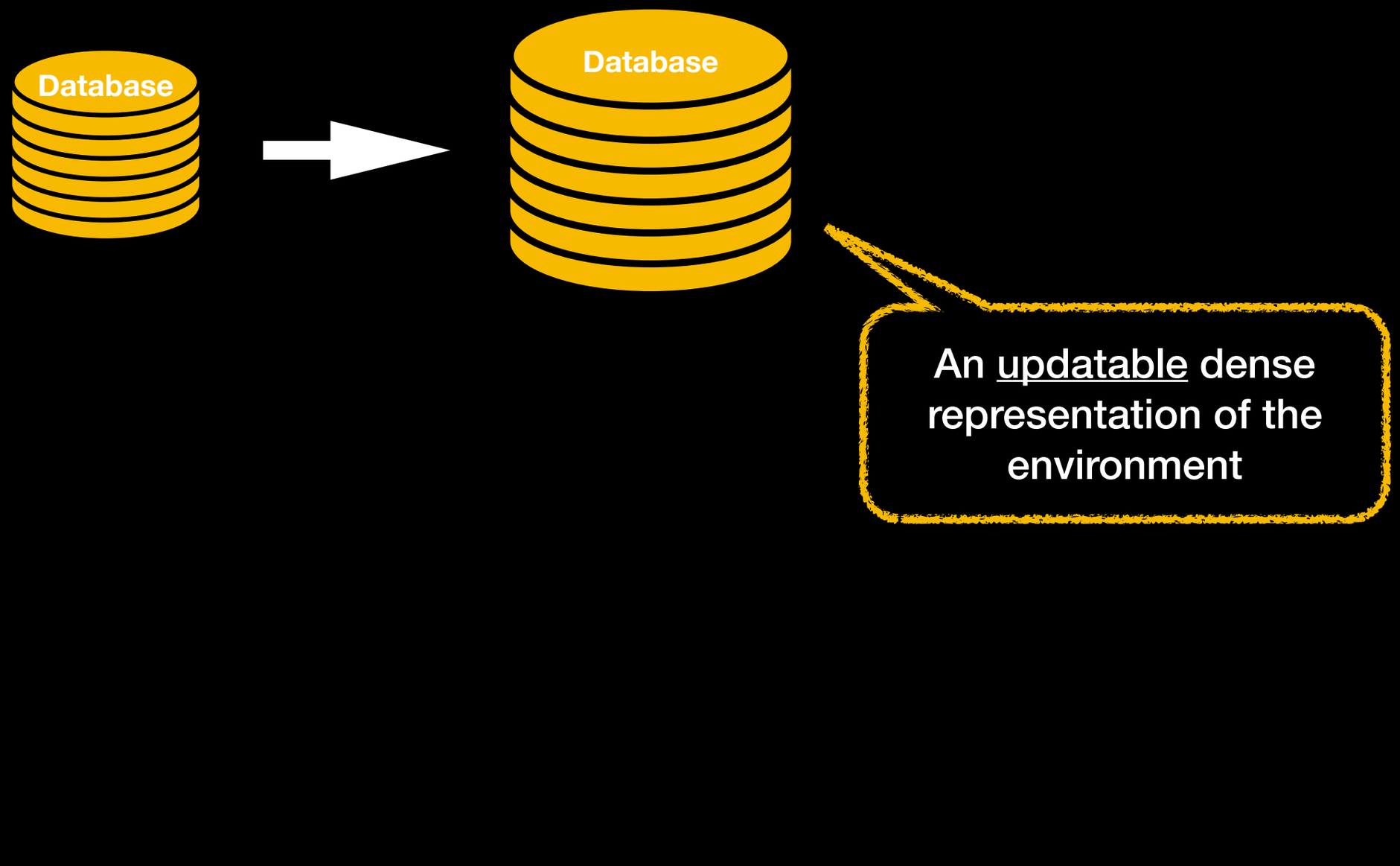
Orthogonal to empirical bounds (conformal, etc.) via held-out set

5. Constrain prediction based on nearest label (or distribution of labels) in database
-Can also leverage the distances to the exemplars

New Approach: Exemplar Auditing

New Approach: Exemplar Auditing

6. We can update the database over time by adding new instances (e.g., out-of-domain), modifying or adding labels, etc.



New Approach: Exemplar Auditing

6. We can update the database over time by adding new instances (e.g., out-of-domain), modifying or adding labels, etc.

7. Can use to analyze the data (annotation errors, etc.)

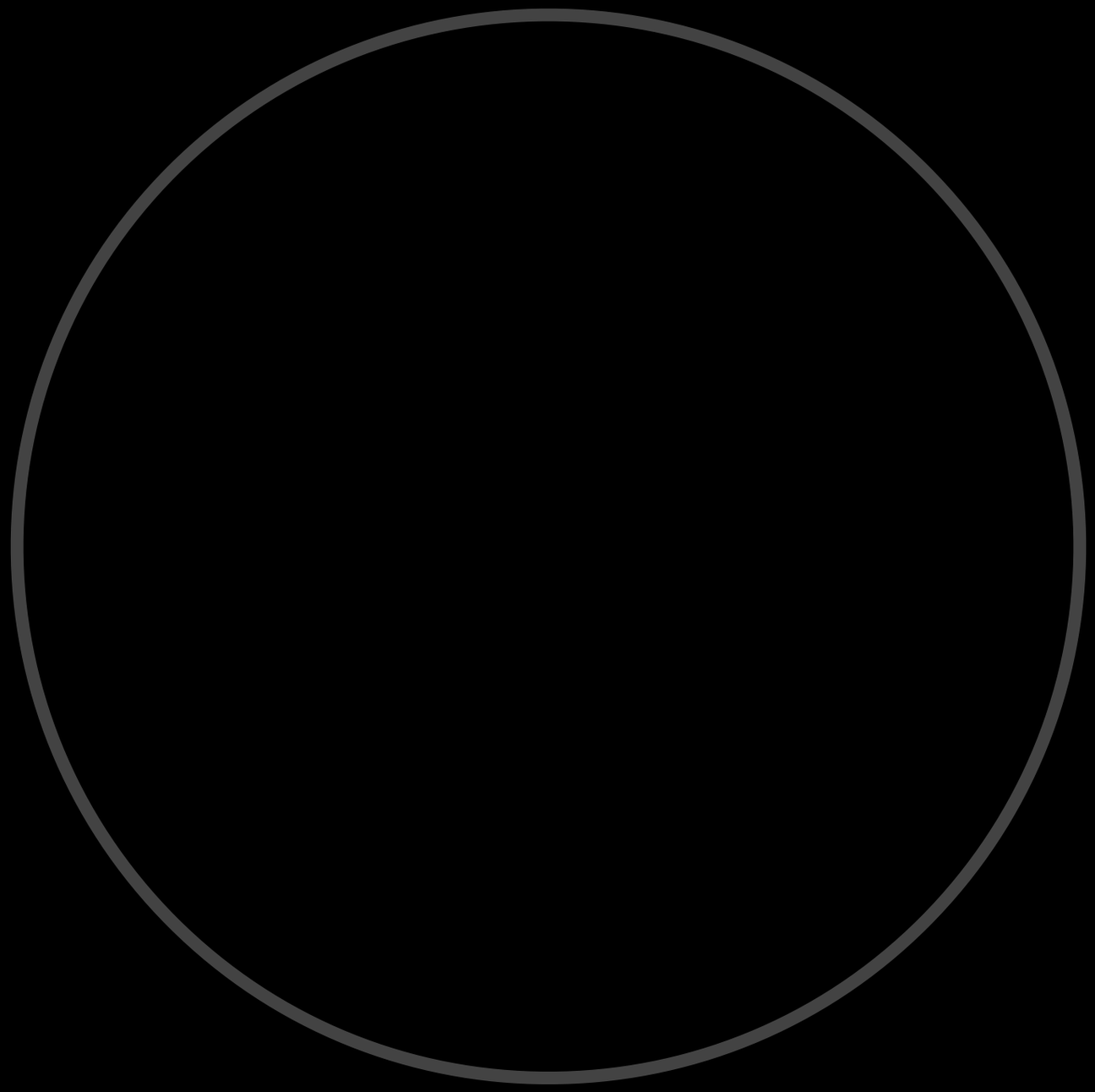
New Approach: Exemplar Auditing

6. We can update the database over time by adding new instances (e.g., out-of-domain), modifying or adding labels, etc.

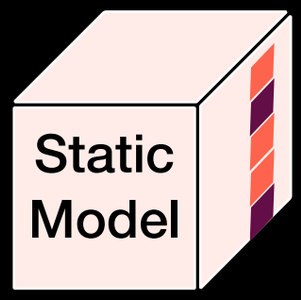
7. Can use to analyze the data (annotation errors, etc.)

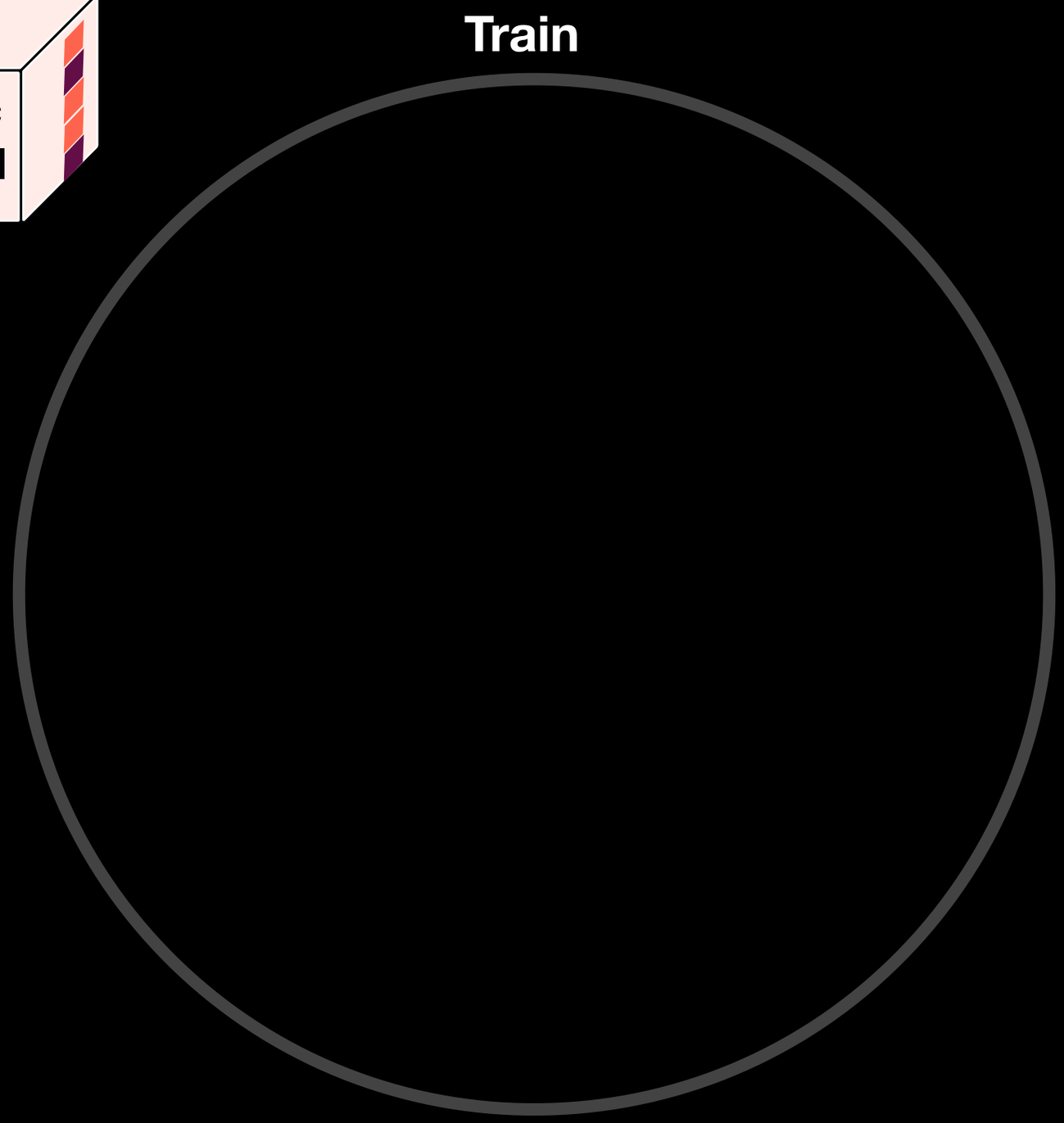
8. As necessary, re-train the model with new/updated labels, instances, etc.

Exemplar Auditing Lifecycle

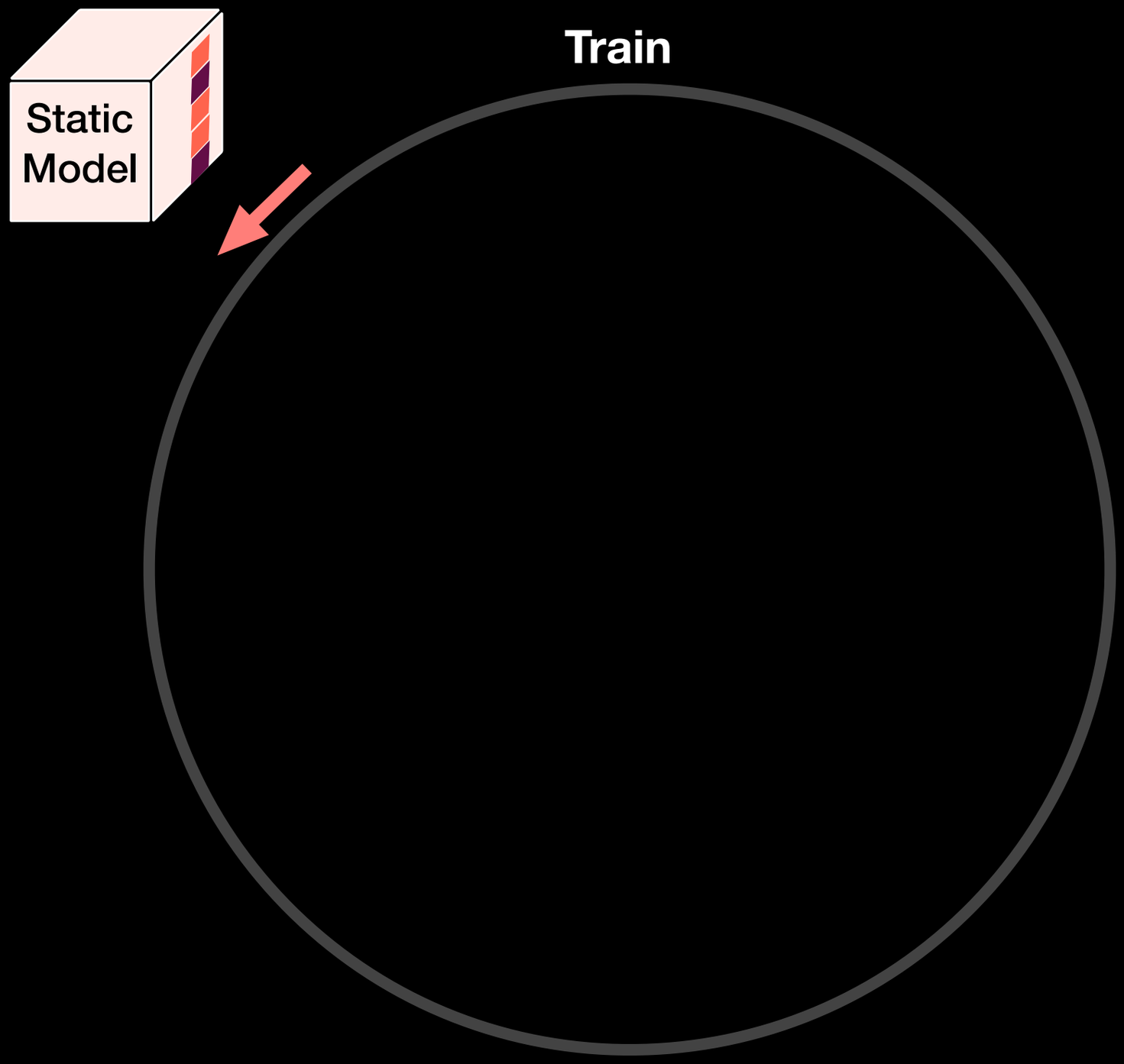


Exemplar Auditing Lifecycle

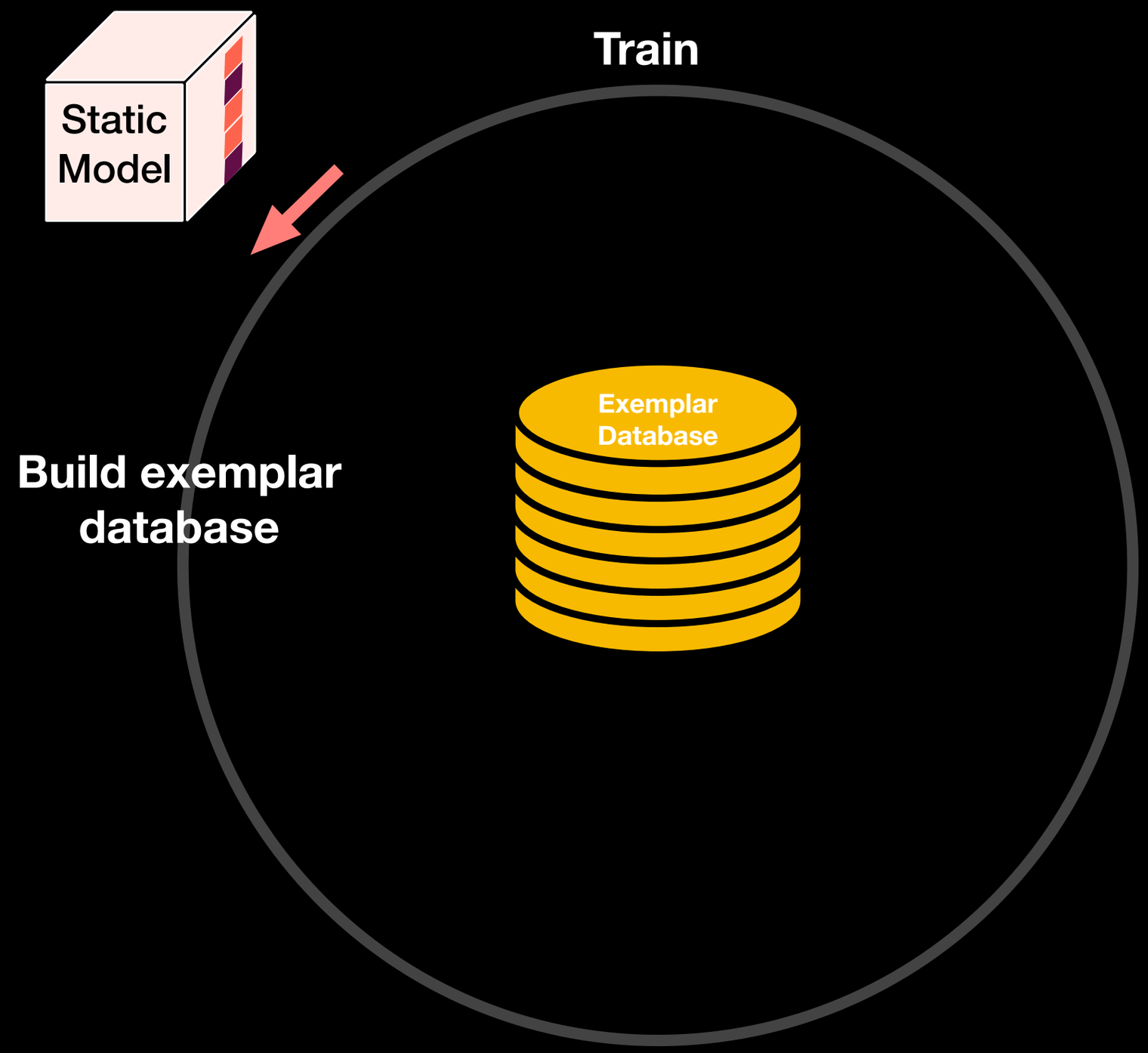




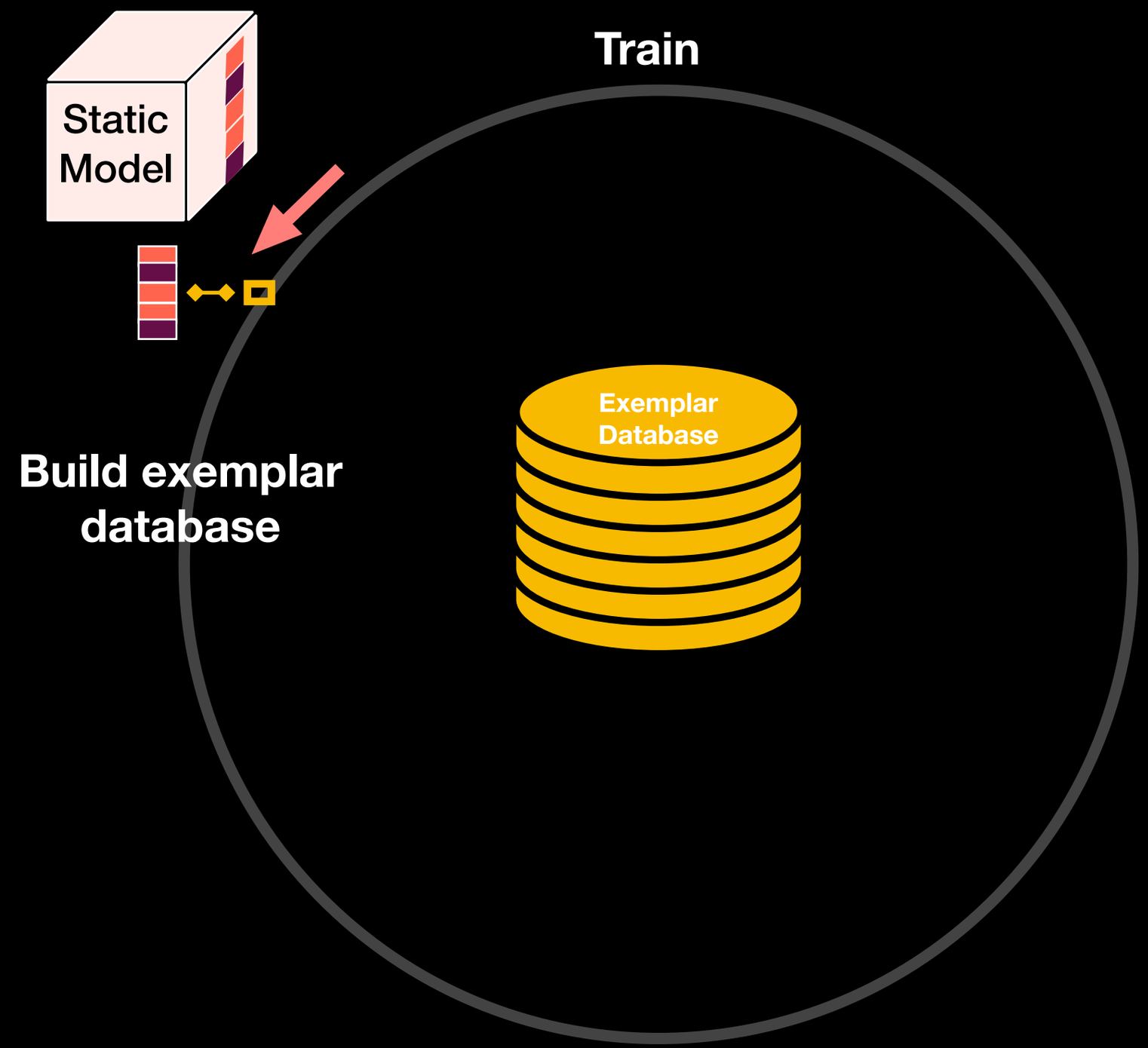
Exemplar Auditing Lifecycle



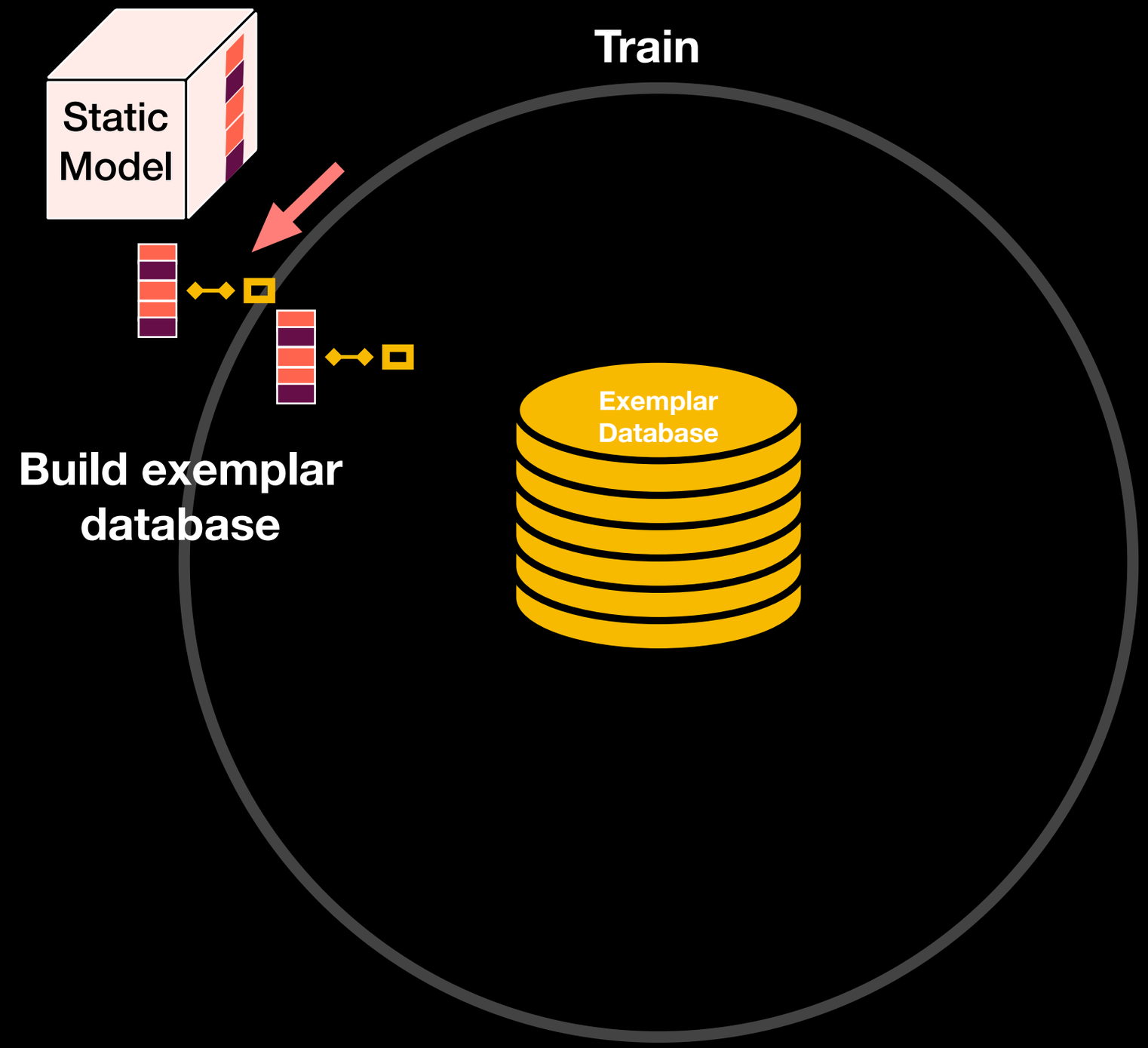
Exemplar Auditing Lifecycle



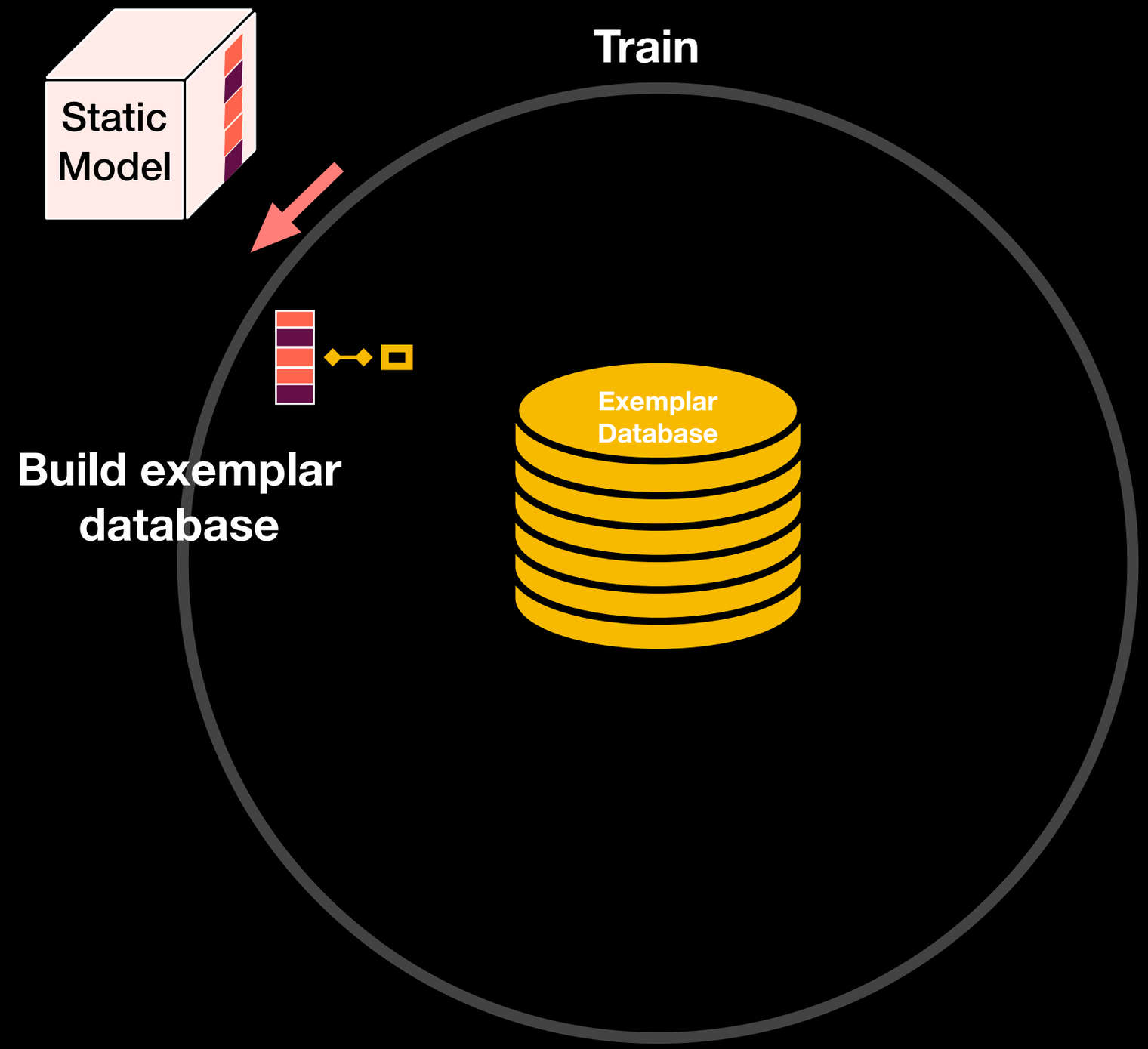
Exemplar Auditing Lifecycle



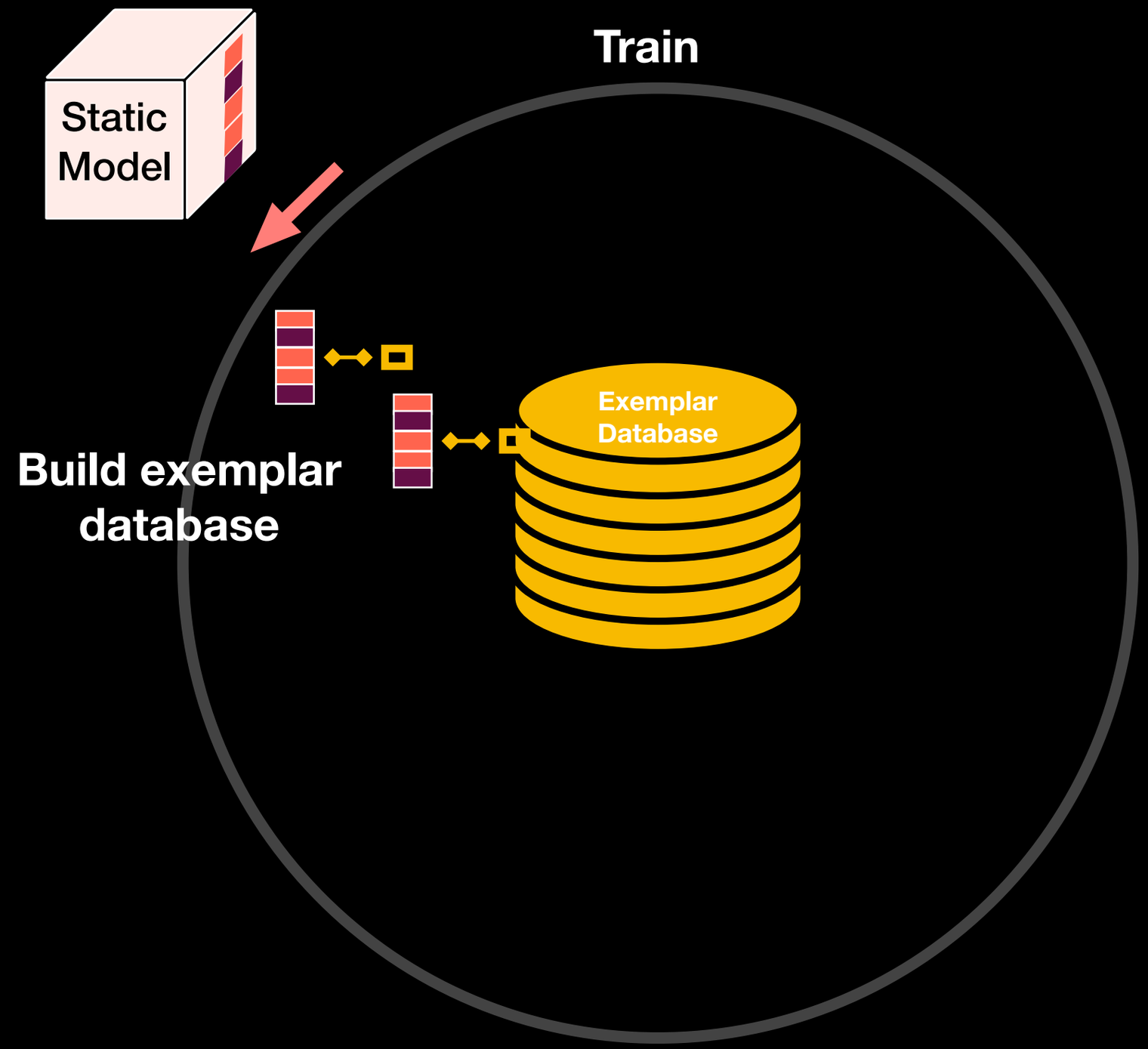
Exemplar Auditing Lifecycle



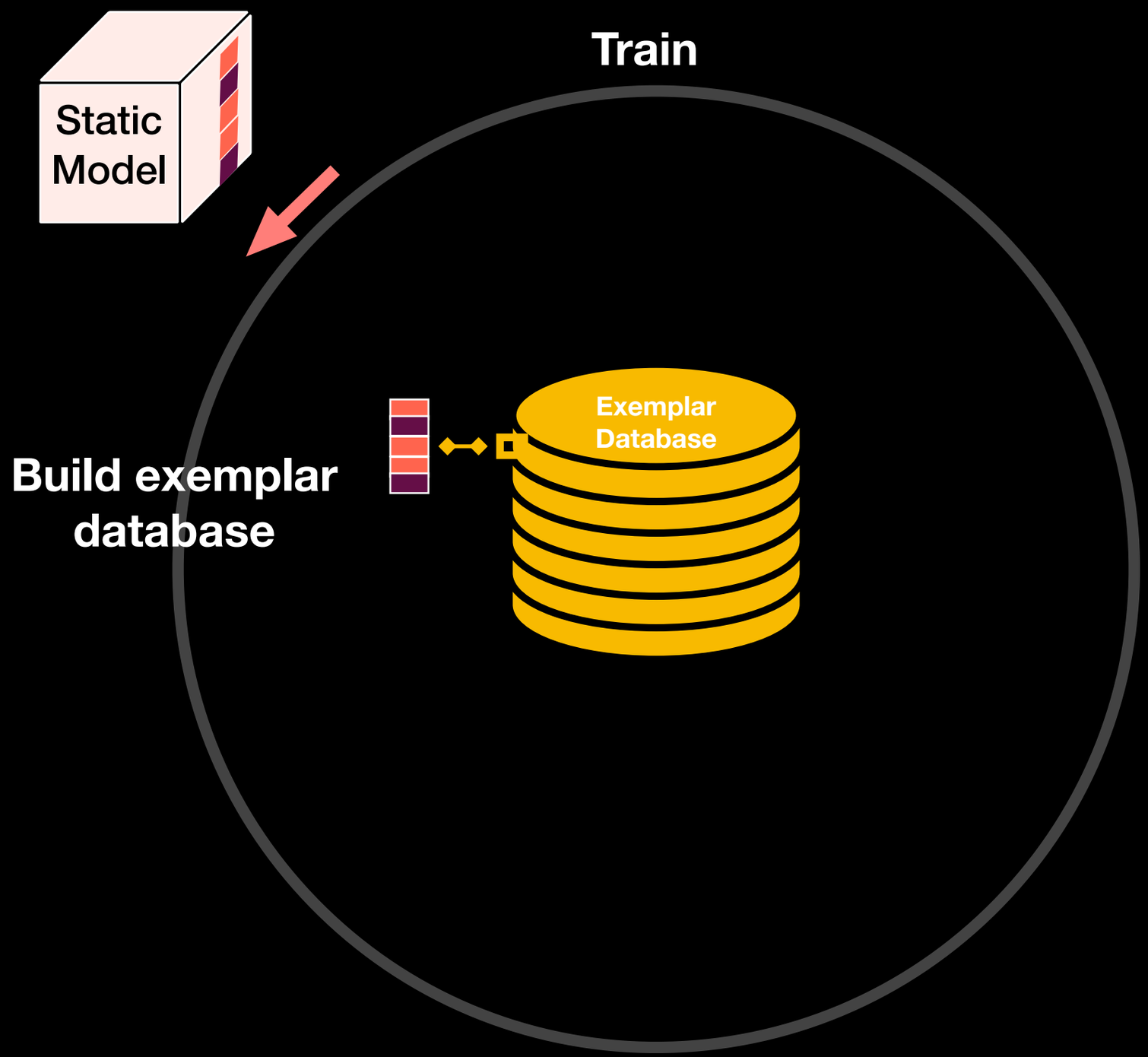
Exemplar Auditing Lifecycle



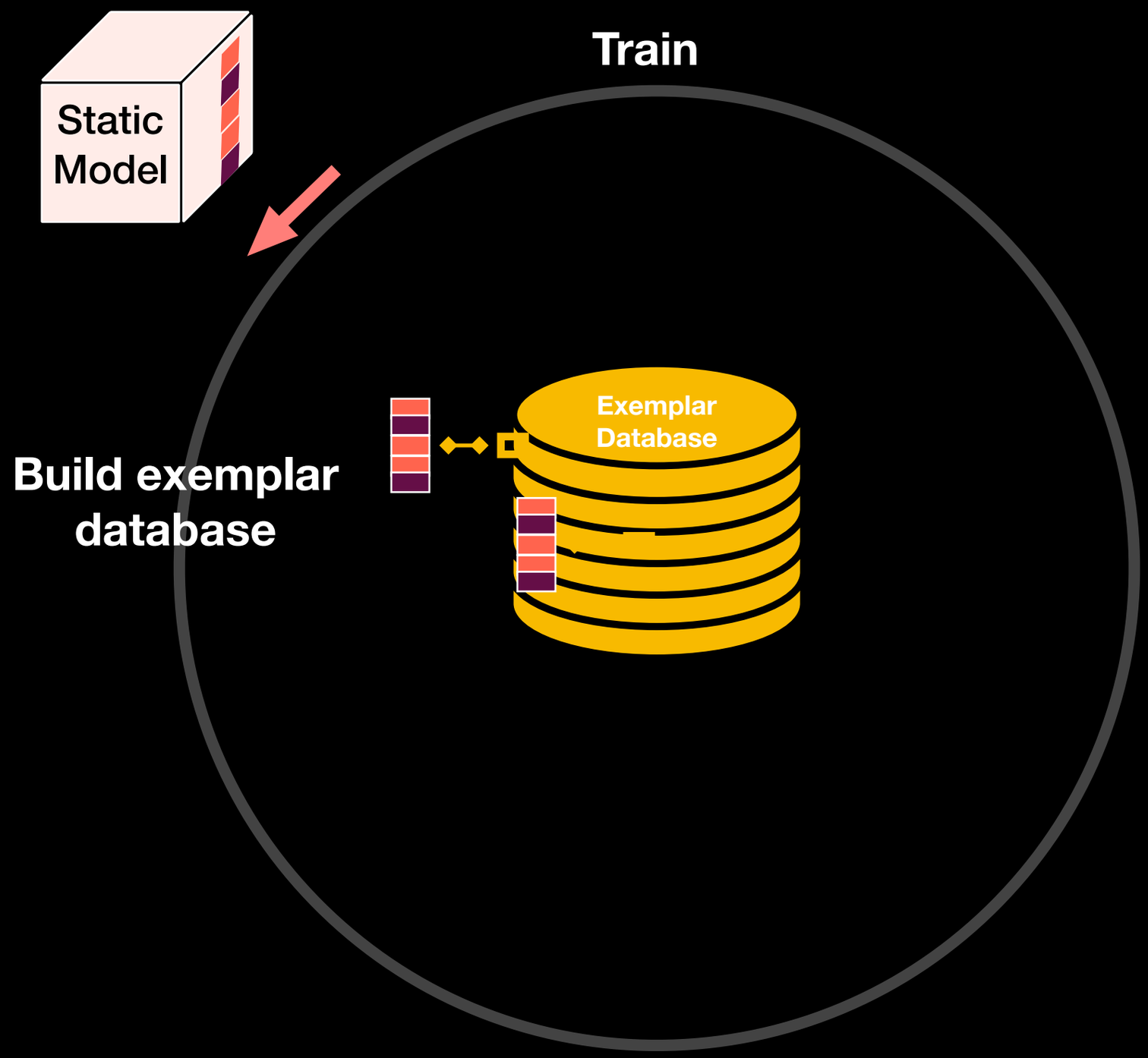
Exemplar Auditing Lifecycle



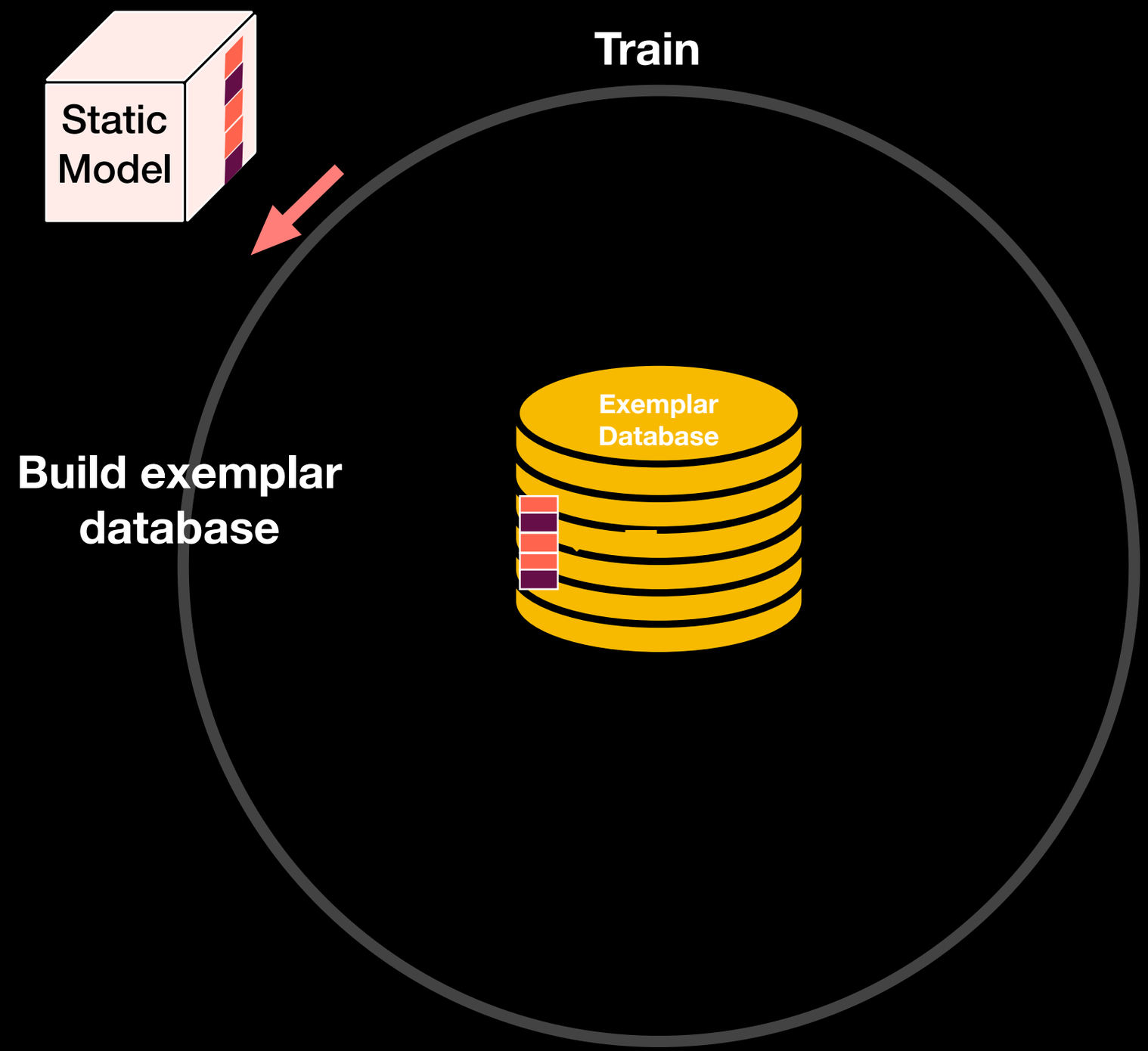
Exemplar Auditing Lifecycle



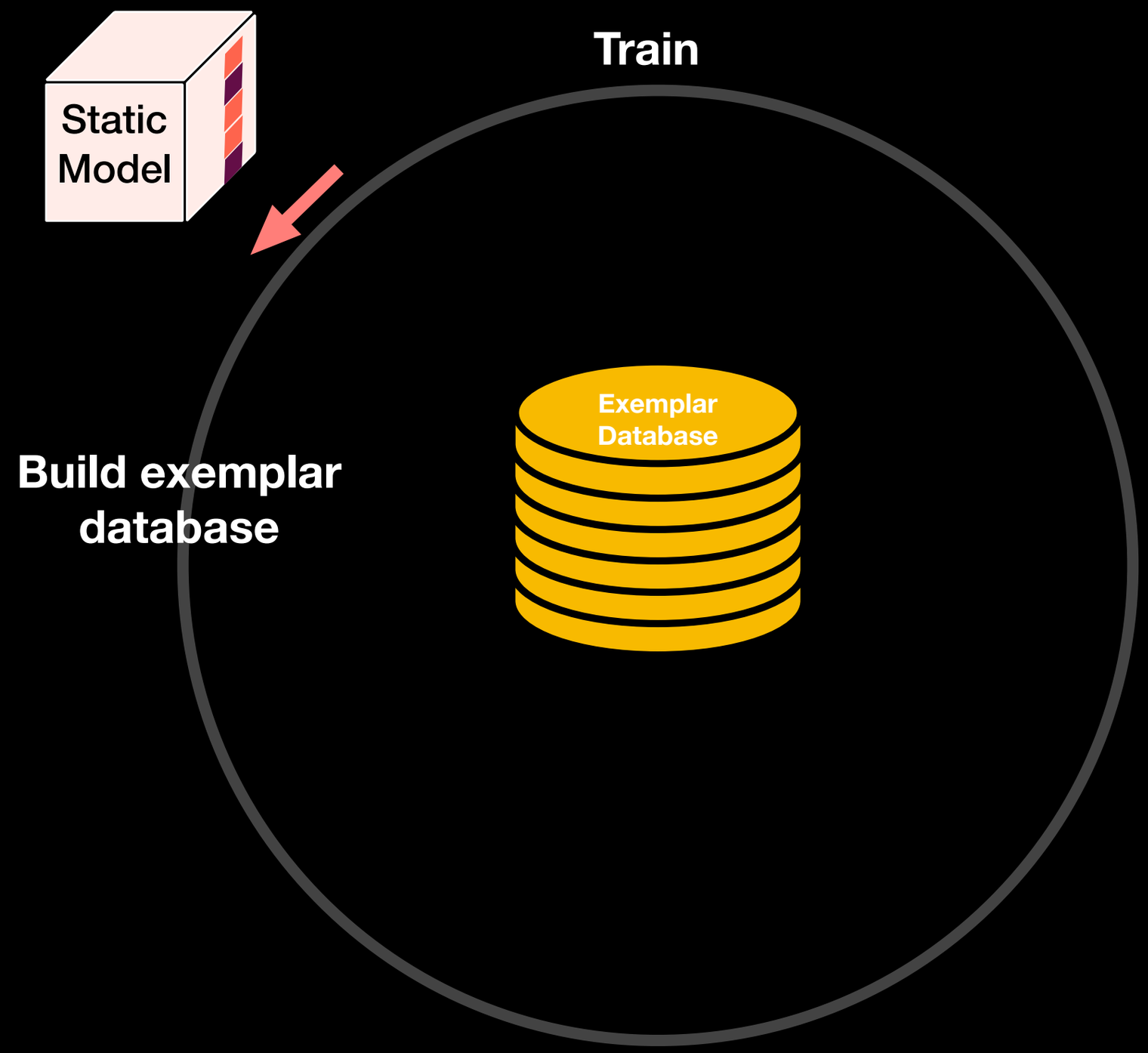
Exemplar Auditing Lifecycle



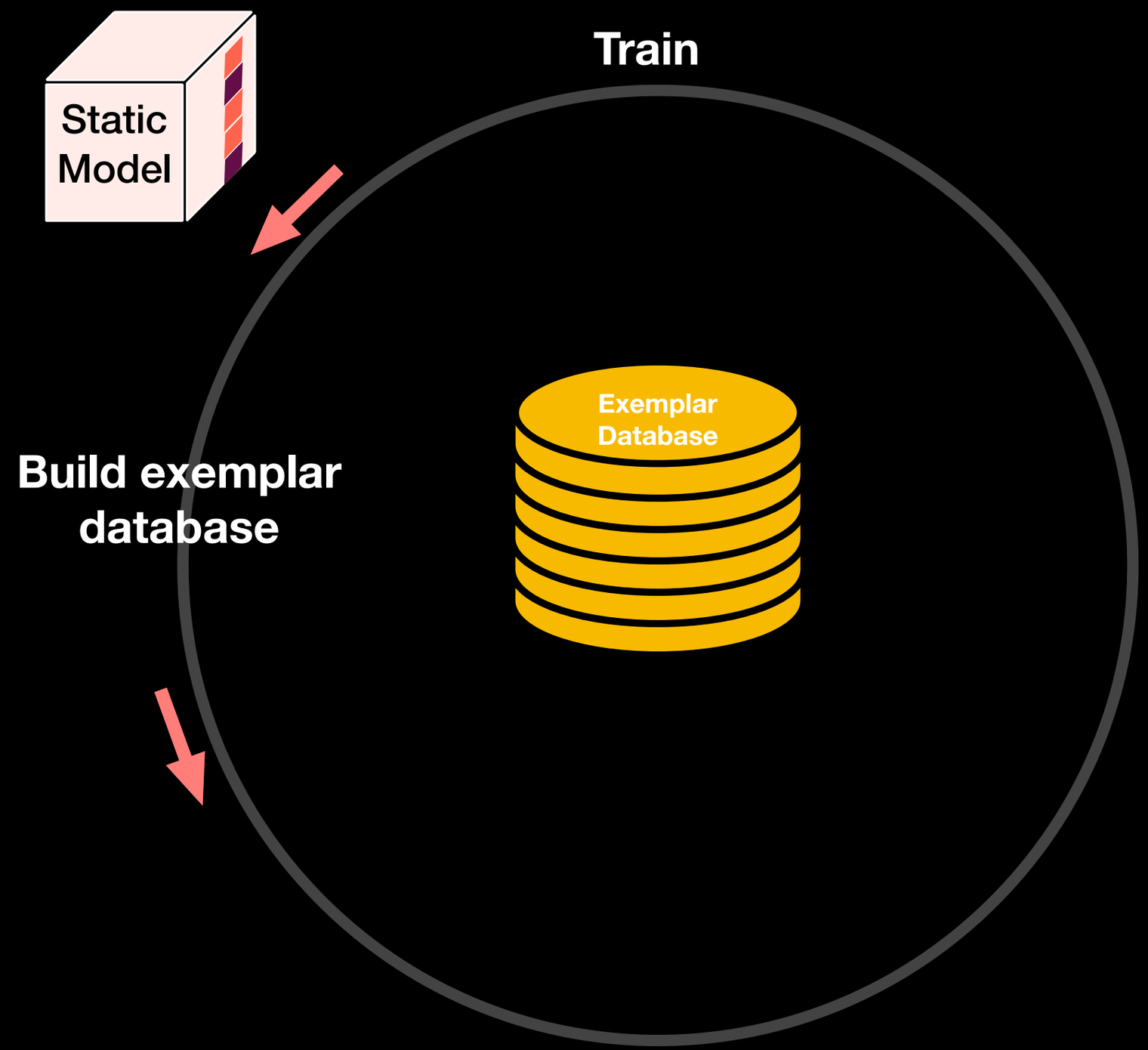
Exemplar Auditing Lifecycle



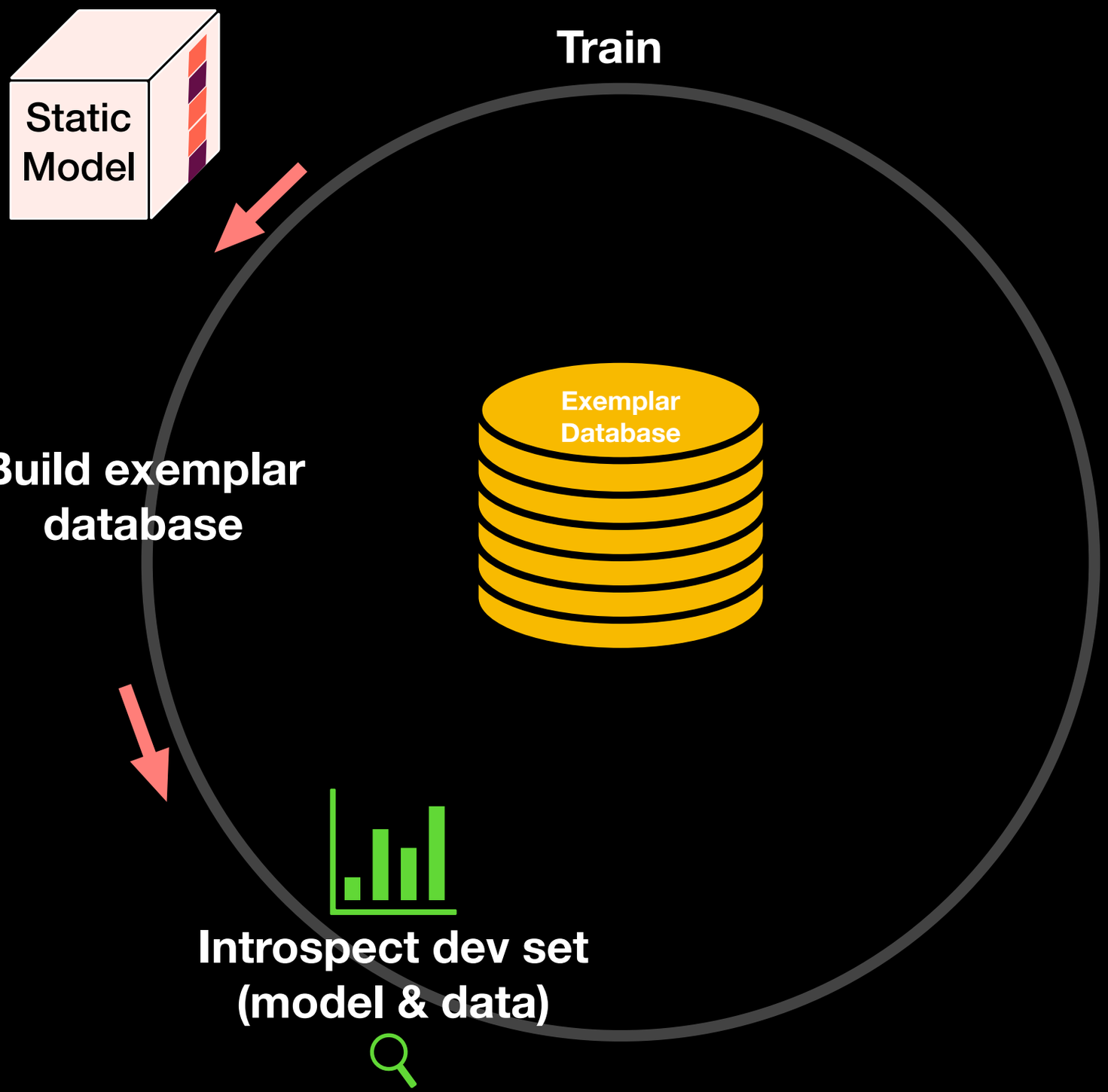
Exemplar Auditing Lifecycle



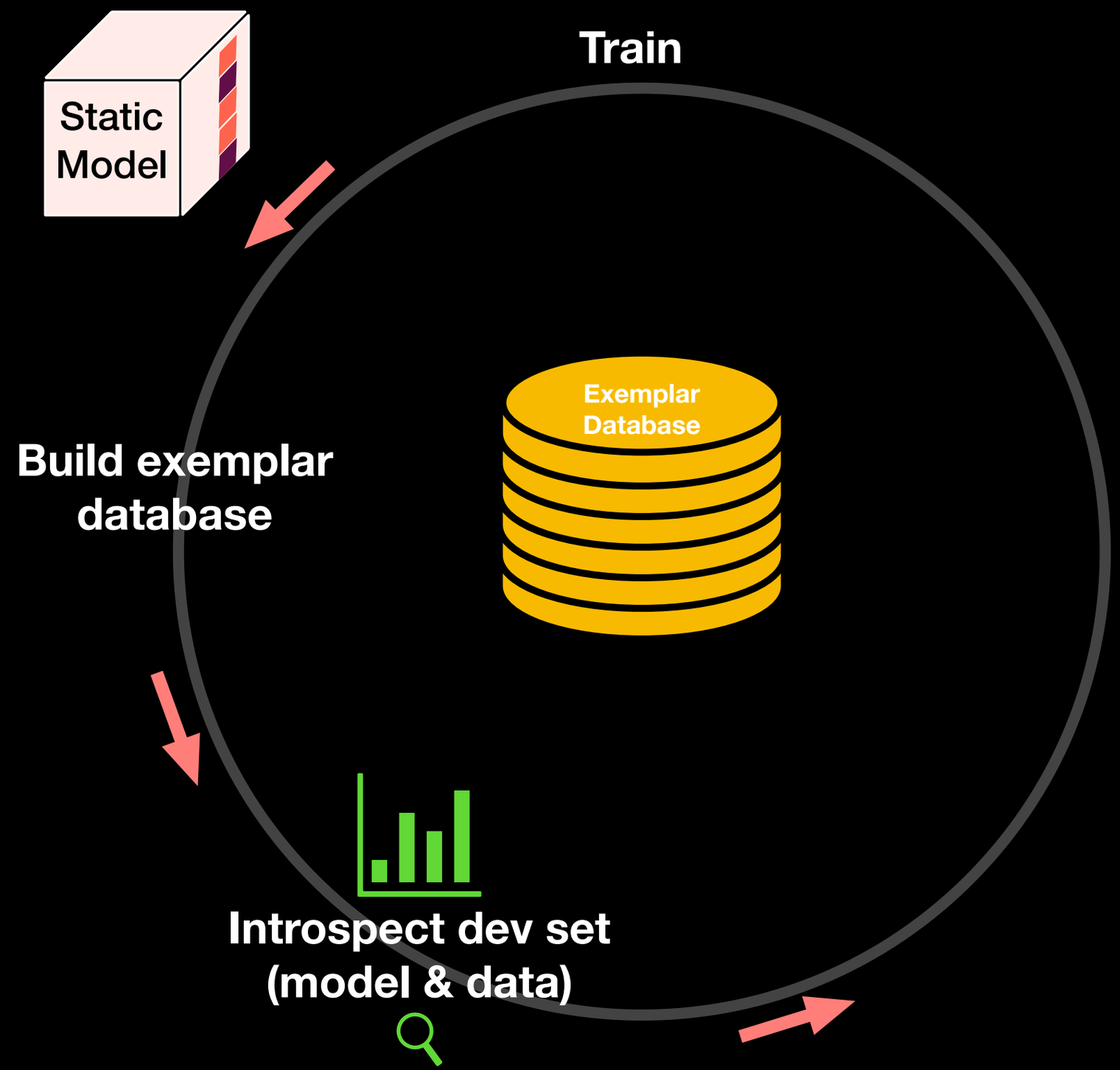
Exemplar Auditing Lifecycle



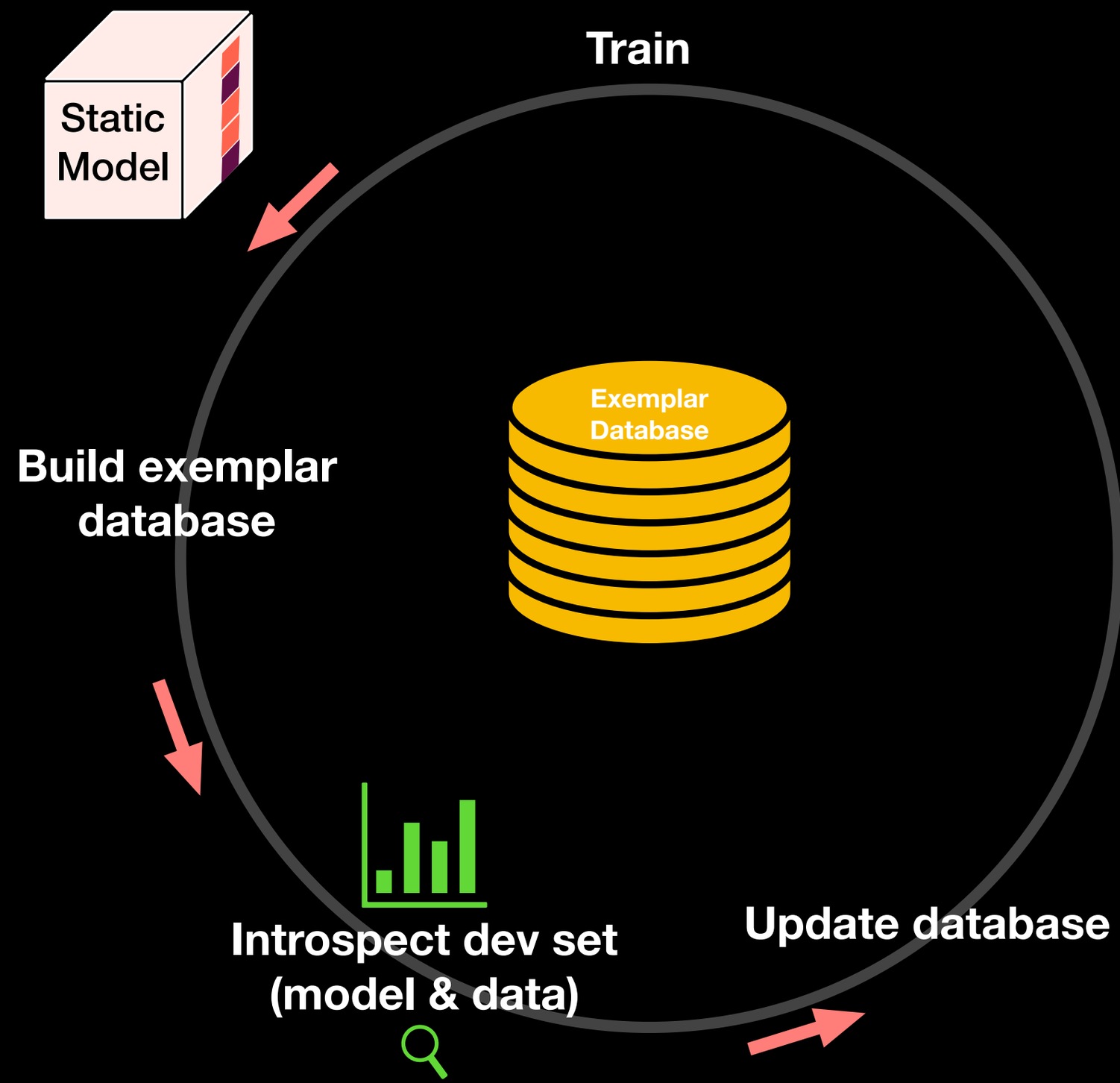
Exemplar Auditing Lifecycle



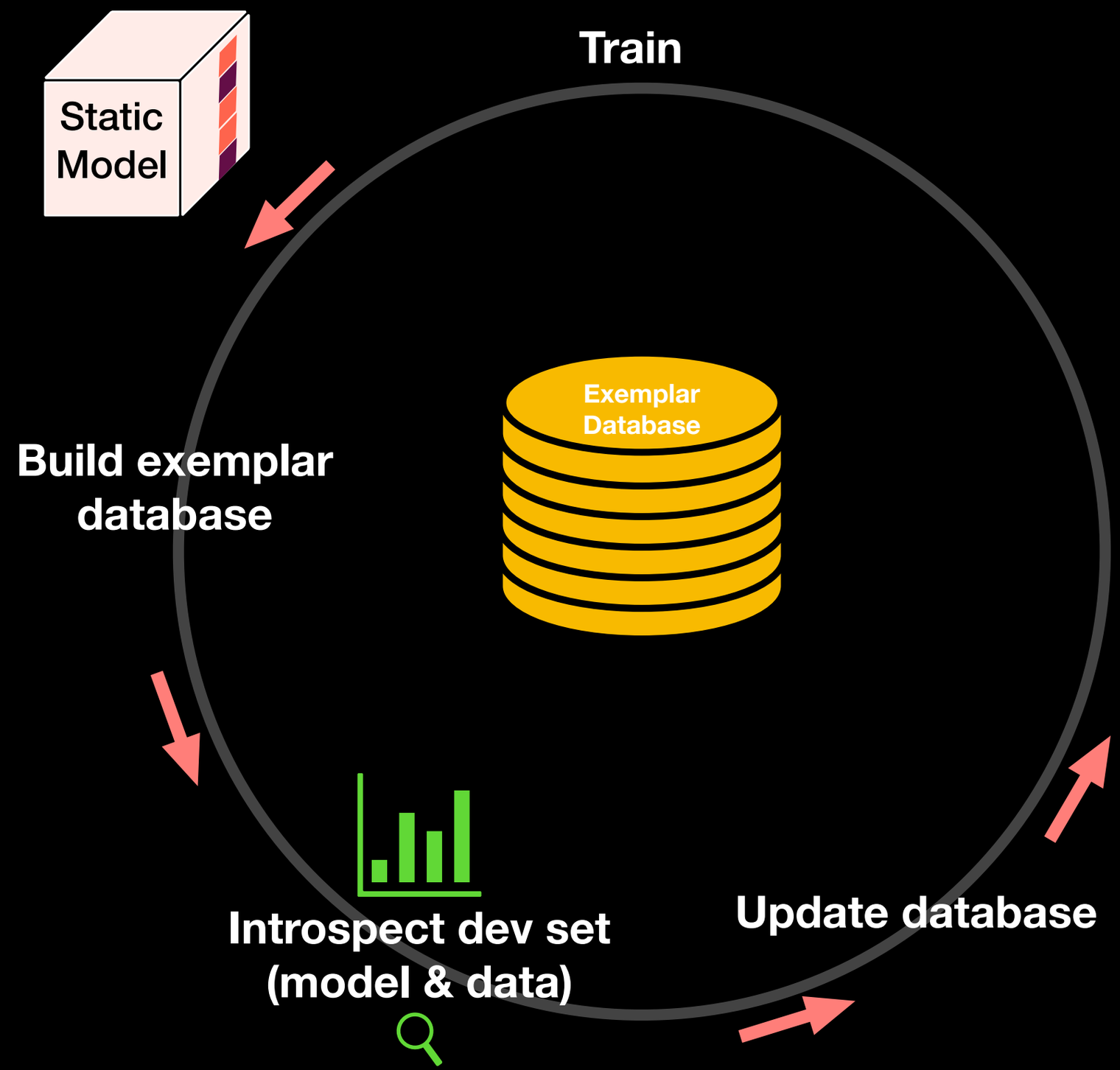
Exemplar Auditing Lifecycle



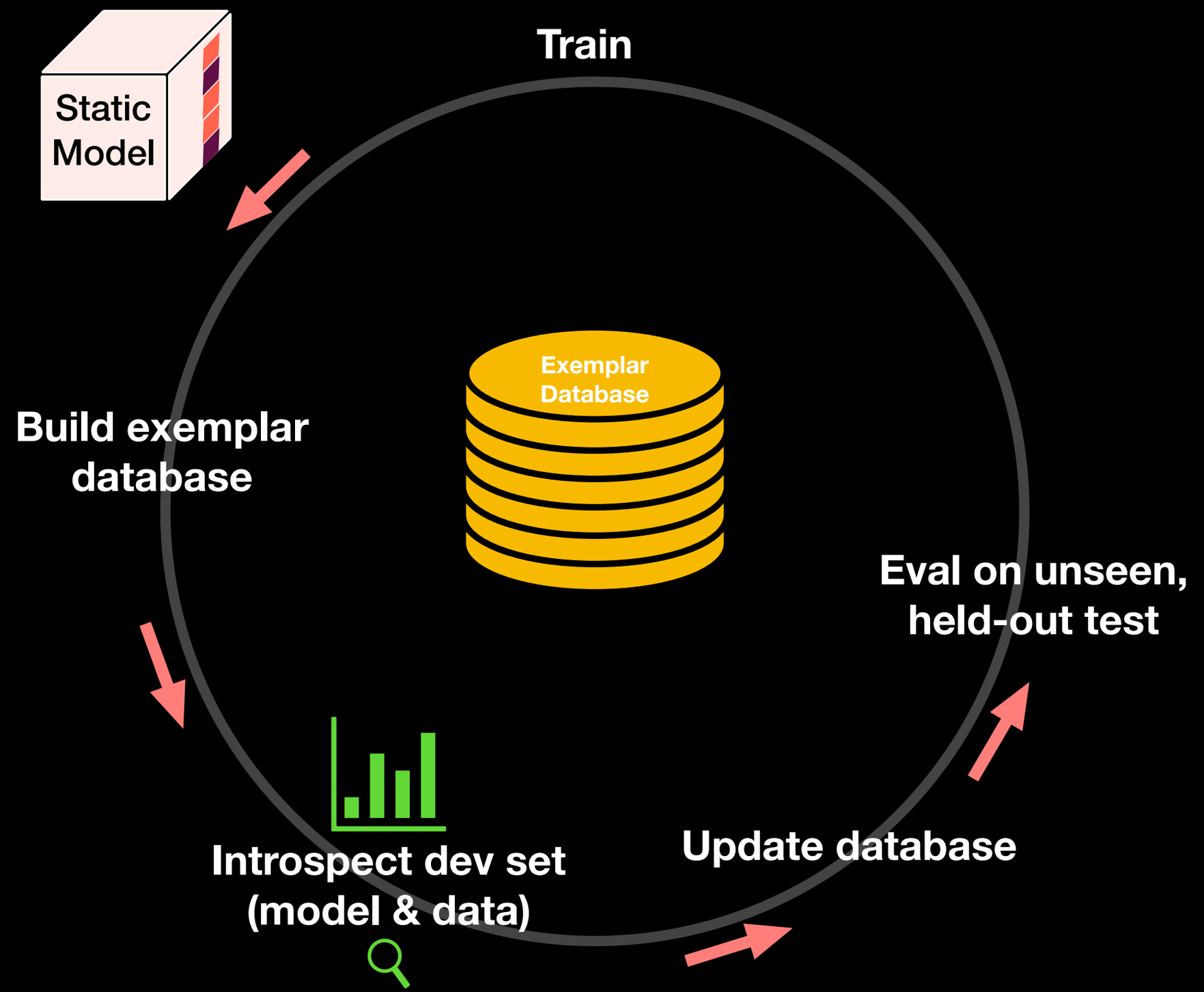
Exemplar Auditing Lifecycle



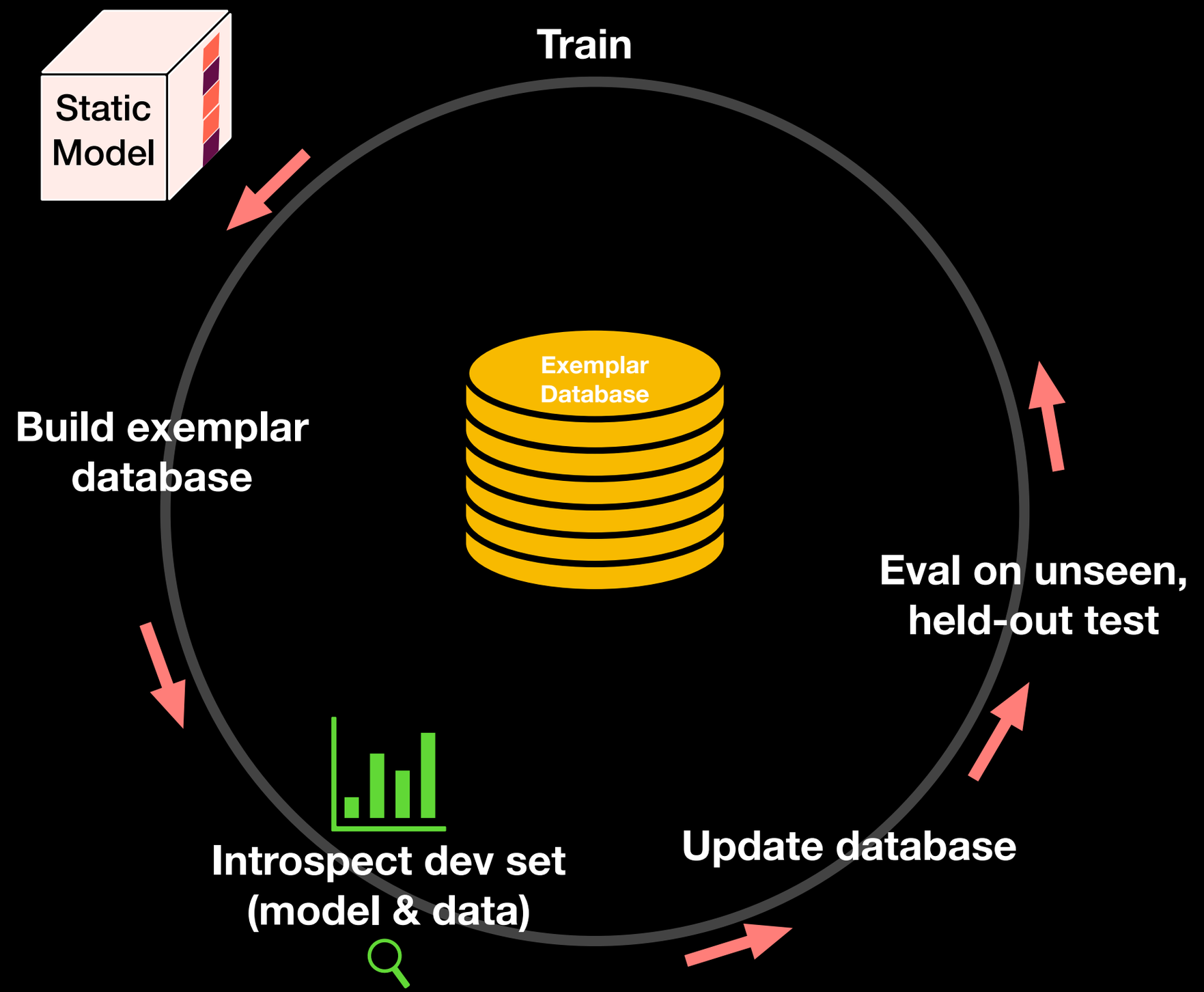
Exemplar Auditing Lifecycle



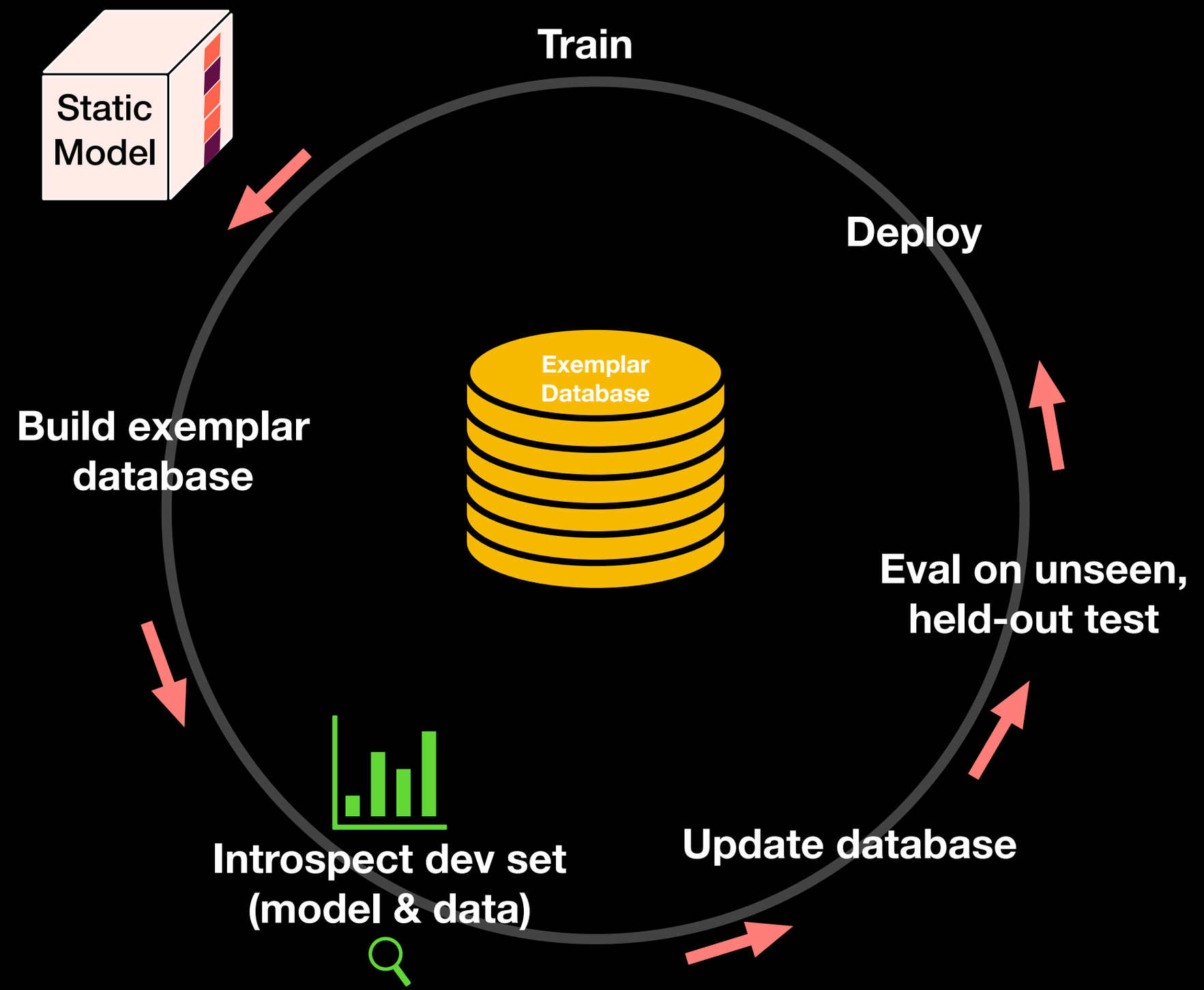
Exemplar Auditing Lifecycle



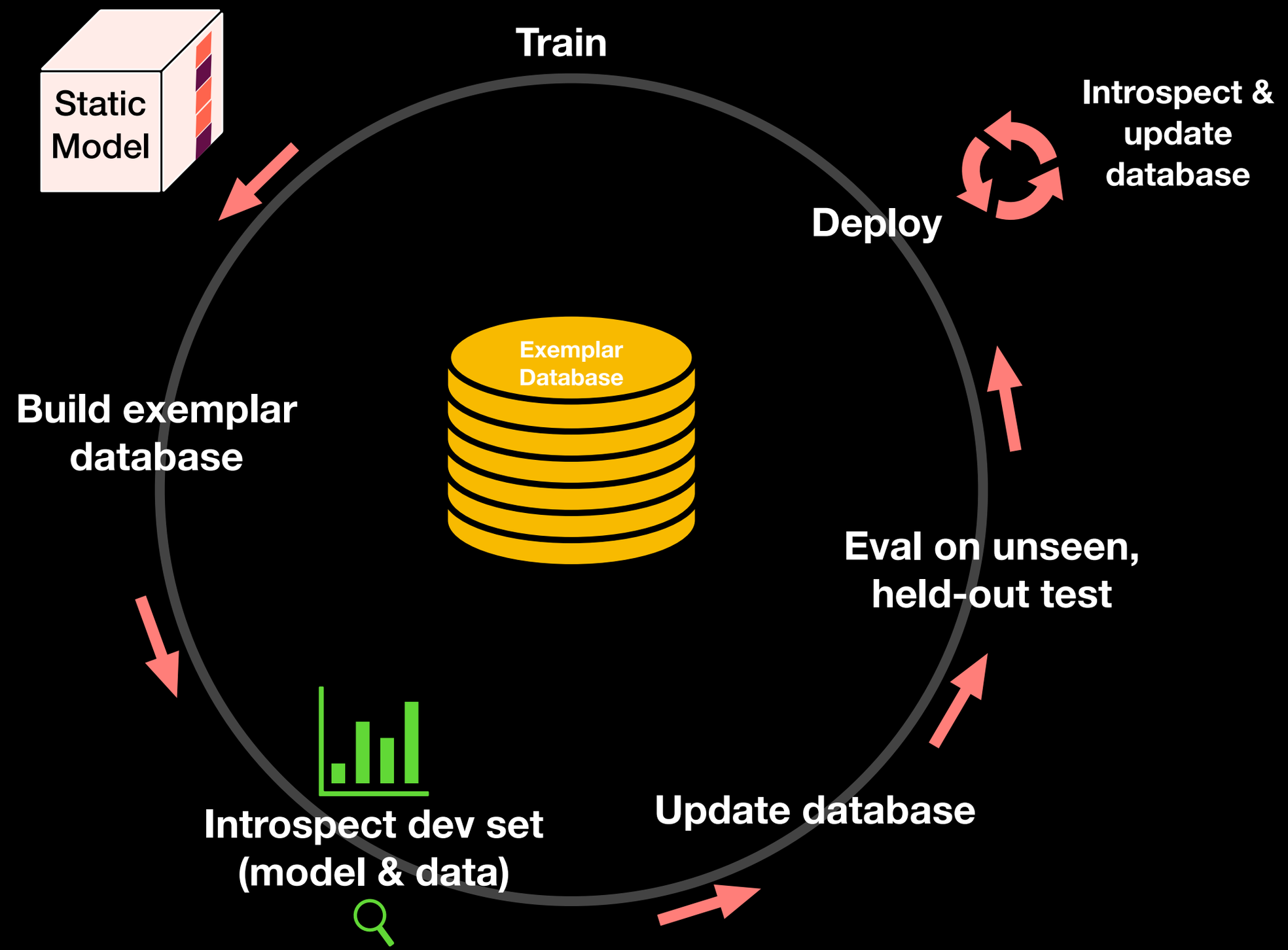
Exemplar Auditing Lifecycle



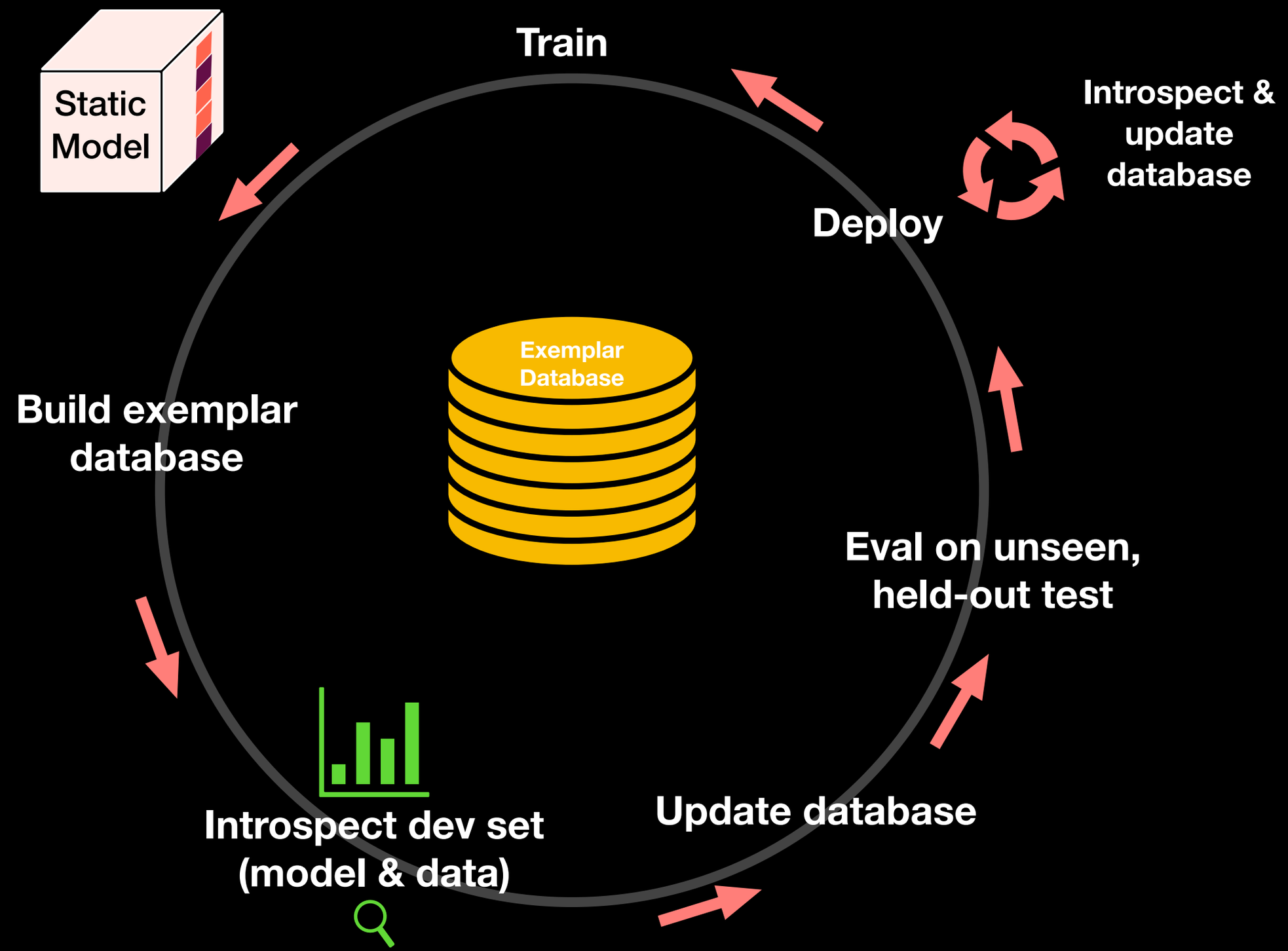
Exemplar Auditing Lifecycle



Exemplar Auditing Lifecycle

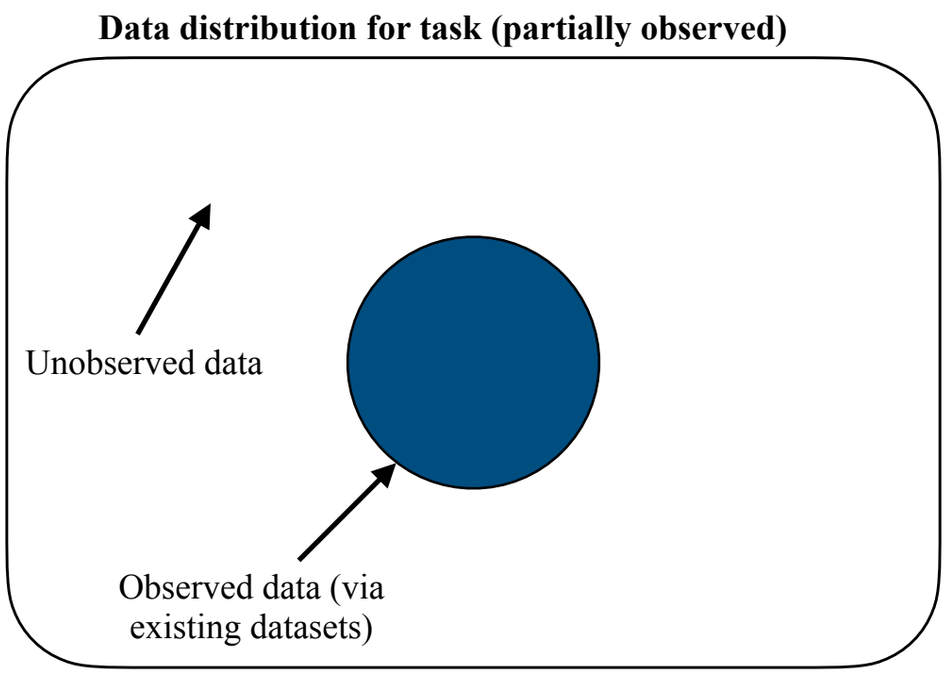


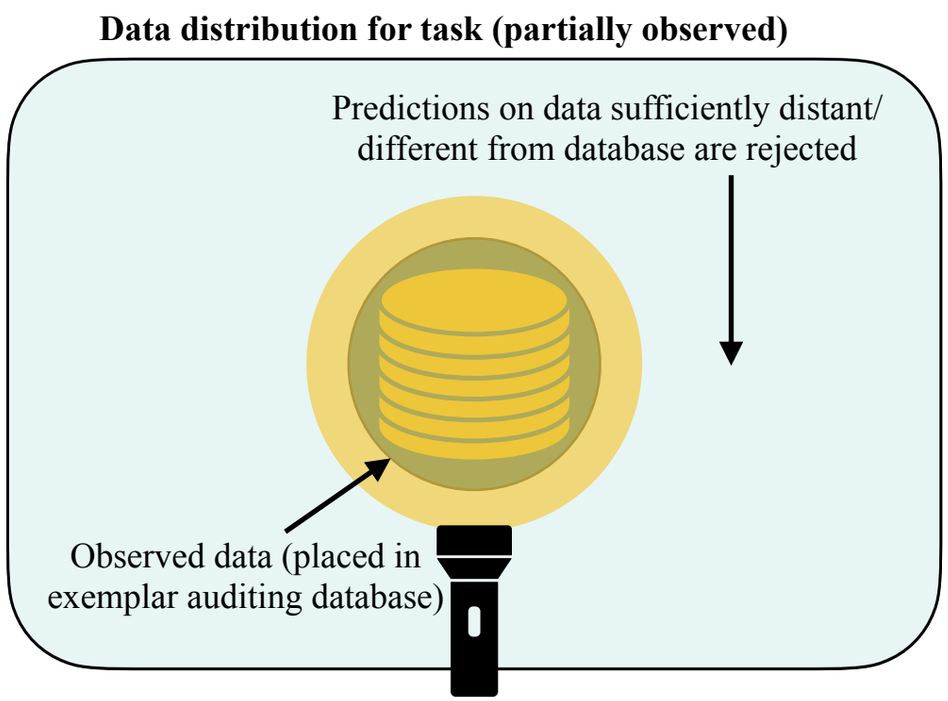
Exemplar Auditing Lifecycle



Aside: Out-of-domain

- Pre-train with as much data as possible
- Add as much data as possible to the database
    - Corral the in-domain space, around the ball of the observed data
- Never predict over out-of-domain in high-risk settings—Instead: Rearrange deployment to handle non-admitted predictions





What makes this work (i.e., why haven't we always done this)?

- Somewhat reminiscent (at high-level) of 1980's exemplar systems
- Conceptually, seems like the "right thing to do" for complex models, but not clear how to implement in practice

What makes this work (i.e., why haven’t we always done this)?

What makes this work (i.e., why haven’t we always done this)?

- Need a sufficiently expressive, strong parametric network over the input

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network

What makes this work (i.e., why haven’t we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model

Will see shortly how for large class of tasks

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model
    - Determining where to cut the parametric network is critical (& not obvious)

Will see shortly how for large class of tasks

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model
    - Determining where to cut the parametric network is critical (& not obvious)
    - Technical choices are important/non-trivial (loss, structure of the memory layers, training regime, etc.)

Will see shortly how for large class of tasks

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model
    - Determining where to cut the parametric network is critical (& not obvious)
    - Technical choices are important/non-trivial (loss, structure of the memory layers, training regime, etc.)
- Need additional structures in code and deployments to handle dense search at test/inference and updating the database

Will see shortly how for large class of tasks

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model
    - Determining where to cut the parametric network is critical (& not obvious)
    - Technical choices are important/non-trivial (loss, structure of the memory layers, training regime, etc.)
- Need additional structures in code and deployments to handle dense search at test/inference and updating the database
  - ⇒ More complicated codebases and deployments than the 'standard' approach

Will see shortly how for large class of tasks

What makes this work (i.e., why haven't we always done this)?

- Need a sufficiently expressive, strong parametric network over the input
- Need effective, compact representations of the parametric network
  - Seek tightly coupled, end-to-end single model
    - Determining where to cut the parametric network is critical (& not obvious)
    - Technical choices are important/non-trivial (loss, structure of the memory layers, training regime, etc.)
- Need additional structures in code and deployments to handle dense search at test/inference and updating the database
  - ⇒ More complicated codebases and deployments than the 'standard' approach

Will see shortly how for large class of tasks

But standard approach is not really applicable to medicine, in any case…

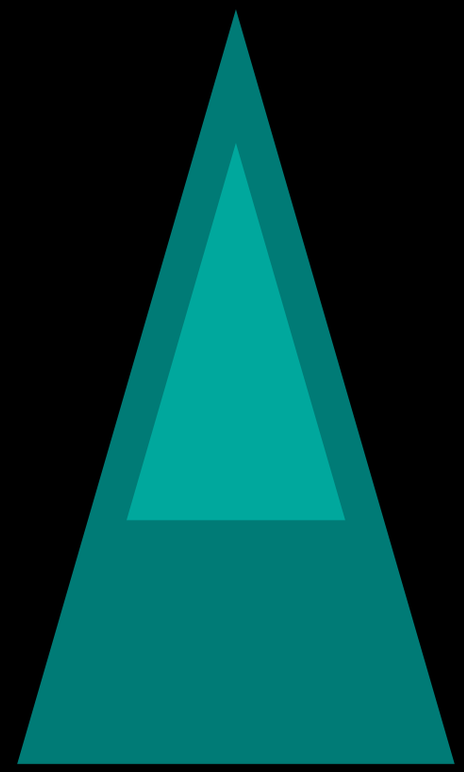
Implementations

Implementations

- Binary classification: $f : X \rightarrow \{0,1\}$
  - See: "Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition"

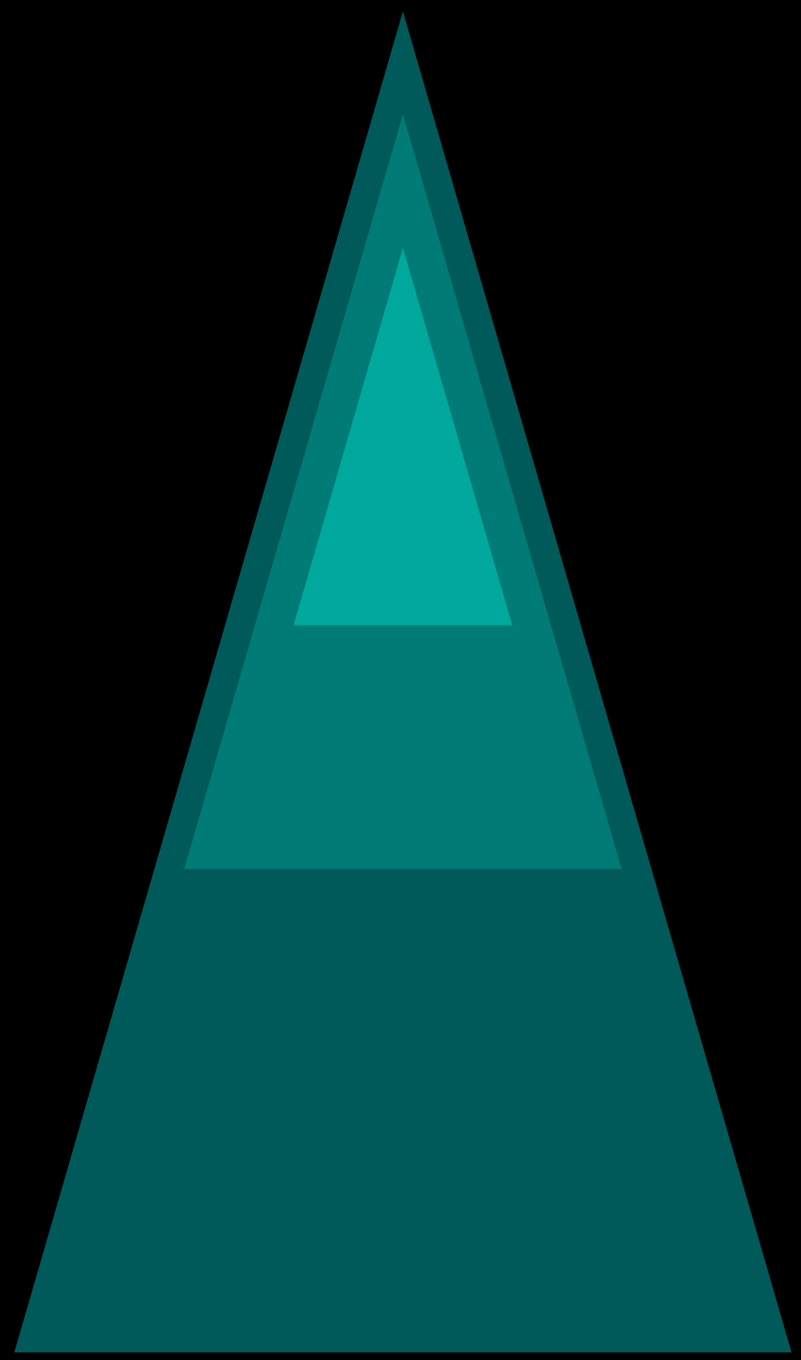
Implementations

- Binary classification: $f : X \rightarrow \{0,1\}$
  - See: "Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition"
- Multi-label classification: $f : X \rightarrow 2^{|Y|}$
  - See: "Exemplar Auditing for Multi-Label Biomedical Text Classification"



Implementations

- Binary classification: $f : X \to \{0,1\}$
  - See: "Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition"
- Multi-label classification: $f : X \to 2^{|Y|}$
  - See: "Exemplar Auditing for Multi-Label Biomedical Text Classification"
- Retrieval-classification: $f : X \times \mathscr{D} \to \left\langle \{0,1,2\}, 2^{|D|} \right\rangle$
  - Today



Retrieval-Classification

Retrieval-Classification

- Retrieval-classification tasks: E.g., QA & fact verification
  - Need to retrieve multiple sequences
  - Then make a classification decision over those sequences

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{0,1,2\},\ 2^{|D|} \right\rangle$

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{0,1,2\},\ 2^{|D|} \right\rangle$

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{0,1,2\},\ 2^{|D|} \right\rangle$
- $f : X \times \mathscr{D} \rightarrow \left\langle \{\text{False, True, Unverifiable}\},\ 2^{|D|} \right\rangle$

- Given:
  - $x \in X$, a ‘query’ (e.g., a sentence)
  - $D \in \mathscr{D}$, a set of documents (e.g., all of wikipedia)
- Determine:
  - The query is True, False, or Unverifiable AND the subset of D to support that prediction

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{\text{False, True, Unverifiable}\},\ 2^{|D|} \right\rangle$
  - $x \in X$, a ‘query’ (e.g., a sentence)
  - $D \in \mathscr{D}$, a set of documents (e.g., all of wikipedia)
- Viewable as two separate tasks:
  - $f_1 : X \times \mathscr{D} \rightarrow 2^{|D|}$
  - $f_2 : X \times 2^{|D|} \rightarrow \{\text{False, True, Unverifiable}\}$

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{\text{False, True, Unverifiable}\},\ 2^{|D|} \right\rangle$
  - $x \in X$, a ‘query’ (e.g., a sentence)
  - $D \in \mathscr{D}$, a set of documents (e.g., all of wikipedia)
- Viewable as two separate tasks:
  - $f_1 : X \times \mathscr{D} \rightarrow 2^{|D|}$ — Information retrieval
  - $f_2 : X \times 2^{|D|} \rightarrow \{\text{False, True, Unverifiable}\}$

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{\text{False, True, Unverifiable}\},\ 2^{|D|} \right\rangle$
  - $x \in X$, a ‘query’ (e.g., a sentence)
  - $D \in \mathscr{D}$, a set of documents (e.g., all of wikipedia)
- Viewable as two separate tasks:
  - $f_1 : X \times \mathscr{D} \rightarrow 2^{|D|}$ — Information retrieval
  - $f_2 : X \times 2^{|D|} \rightarrow \{\text{False, True, Unverifiable}\}$ — Natural language inference

Retrieval-Classification

- $f : X \times \mathscr{D} \rightarrow \left\langle \{\text{False, True, Unverifiable}\},\ 2^{|D|} \right\rangle$
  - $x \in X$, a ‘query’ (e.g., a sentence)
  - $D \in \mathscr{D}$, a set of documents (e.g., all of wikipedia)
- Viewable as two separate tasks:
  - $f_1 : X \times \mathscr{D} \rightarrow 2^{|D|}$ — Information retrieval
  - $f_2 : X \times 2^{|D|} \rightarrow \{\text{False, True, Unverifiable}\}$ — Natural language inference
  - Can we learn together with a single model?

Retrieval-Classification Task: FEVER

- Fact Extraction and VERification (FEVER) Shared Task
- Given:
  - A claim (short, declarative sentence)
  - Wikipedia
- Predict:
  - Claim is True, False, or Unverifiable
  - ≤ 5 sentences that support that prediction

Example from FEVER

Example from FEVER

- **INPUT:** Claim: Charles de Gaulle was a leader in the French Resistance.

Example from FEVER

- **INPUT:** Claim: Charles de Gaulle was a leader in the French Resistance.
- **RETRIEVE:** Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.

Example from FEVER

- **INPUT:** Claim: Charles de Gaulle was a leader in the French Resistance.

- **RETRIEVE:** Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.

- **PREDICT:** Supports

Example from FEVER

- **INPUT:** Claim: Charles de Gaulle was a leader in the French Resistance.

- **RETRIEVE:** Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.

- **PREDICT:** Supports

Sentences made unique by article title and sentence index

Example from FEVER

Example from FEVER

- **INPUT:** Claim: Emma Stone was born in Taiwan.

Example from FEVER

- **INPUT:** Claim: Emma Stone was born in Taiwan.
- **RETRIEVE:** Evidence: Emma Stone, sentence 5: Born and raised in Scottsdale, Arizona, Stone began acting as a child, in a theater production of The Wind in the Willows in 2000.

Example from FEVER

- **INPUT:** Claim: Emma Stone was born in Taiwan.
- **RETRIEVE:** Evidence: Emma Stone, sentence 5: Born and raised in Scottsdale, Arizona, Stone began acting as a child, in a theater production of The Wind in the Willows in 2000.

- **PREDICT:** Refutes

Example from FEVER

- **INPUT:** Claim: Emma Stone was born in Taiwan.
- **RETRIEVE:** Evidence: Emma Stone, sentence 5: Born and raised in Scottsdale, Arizona, Stone began acting as a child, in a theater production of The Wind in the Willows in 2000.

- **PREDICT:** Refutes

FEVER: Existing Work

- Most existing works are multi-model pipelines
  - Document retrieval model
  - Sentence selection model
  - Classification model
- Each model is trained and run independently

FEVER: MemMatch

FEVER: MemMatch

- We instead propose a novel single, end-to-end language model for both retrieval & classification

FEVER: MemMatch

- We instead propose a novel single, end-to-end language model for both retrieval & classification
  - Coarse-to-fine search procedure over dense representations

FEVER: MemMatch

- We instead propose a novel single, end-to-end language model for both retrieval & classification
    - Coarse-to-fine search procedure over dense representations
    - Distances from tightly coupled retrieval and classification can be leveraged to identify low-confidence instances

FEVER: MemMatch

- We instead propose a novel single, end-to-end language model for both retrieval & classification
  - Coarse-to-fine search procedure over dense representations
  - Distances from tightly coupled retrieval and classification can be leveraged to identify low-confidence instances
  - Produces composed dense representations over multiple sequences for exemplar auditing

FEVER: MemMatch

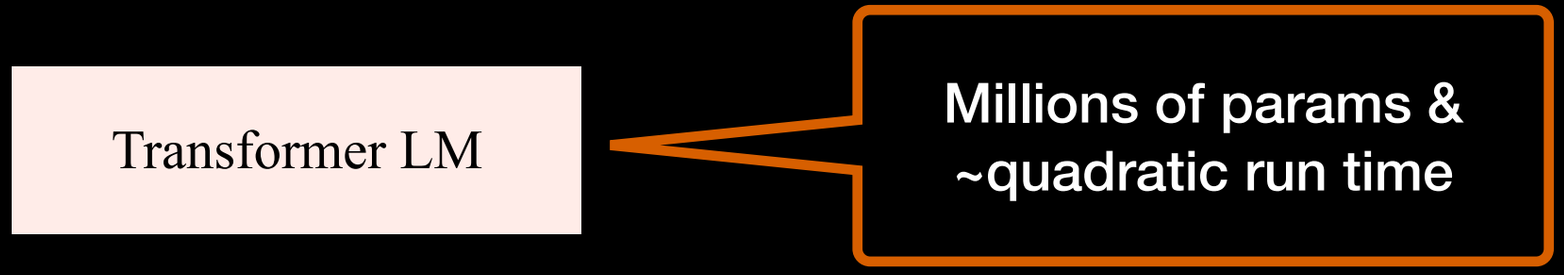
- Effective:
  - More effective than relying on LM parameters as a knowledge base
  - Approaches multi-pipeline systems despite using significantly fewer parameters

FEVER: MemMatch

- Novel properties: Updatability of language model behavior through two distinct mechanisms:
  - Retrieved information can be updated explicitly
  - Model behavior can be modified via the exemplar database

Challenges Building End-to-end Neural Model

- Seek to produce dense representations from a deep neural network (Transformer) for similarity matching



Challenges Building End-to-end Neural Model

- Seek to produce dense representations from a deep neural network (Transformer) for similarity matching

Transformer LM

Millions of params & ~quadratic run time

- Problem: Computationally infeasible to run multiple passes of a deep Transformer over a large datastore (Wikipedia) for every new query

Would result in many 10's of millions of unique sequences

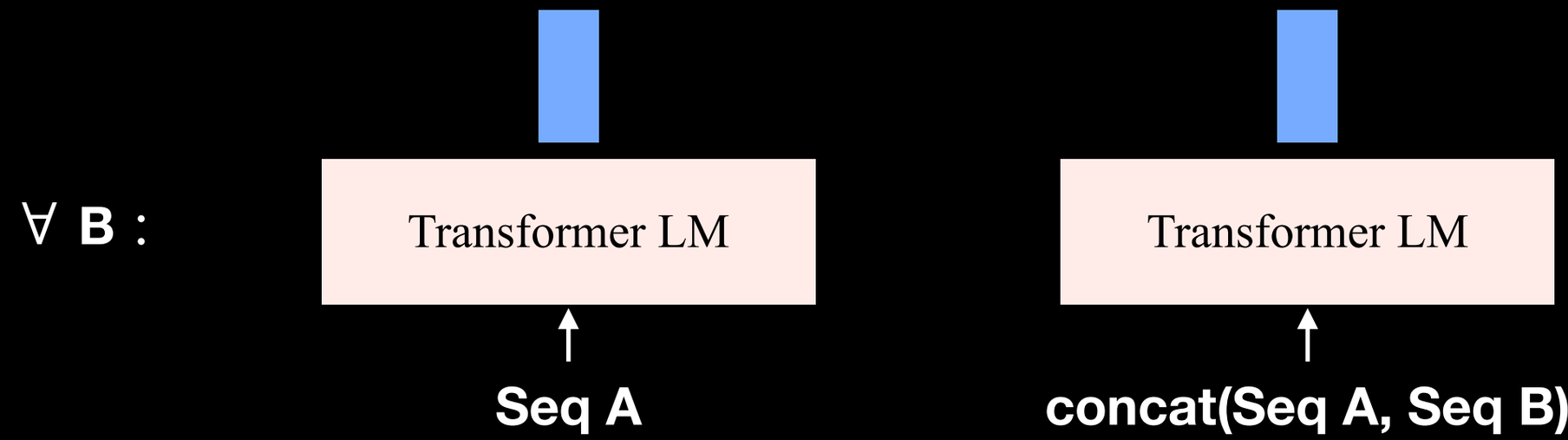
bi-encoder vs. cross-encoder dilemma

bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**

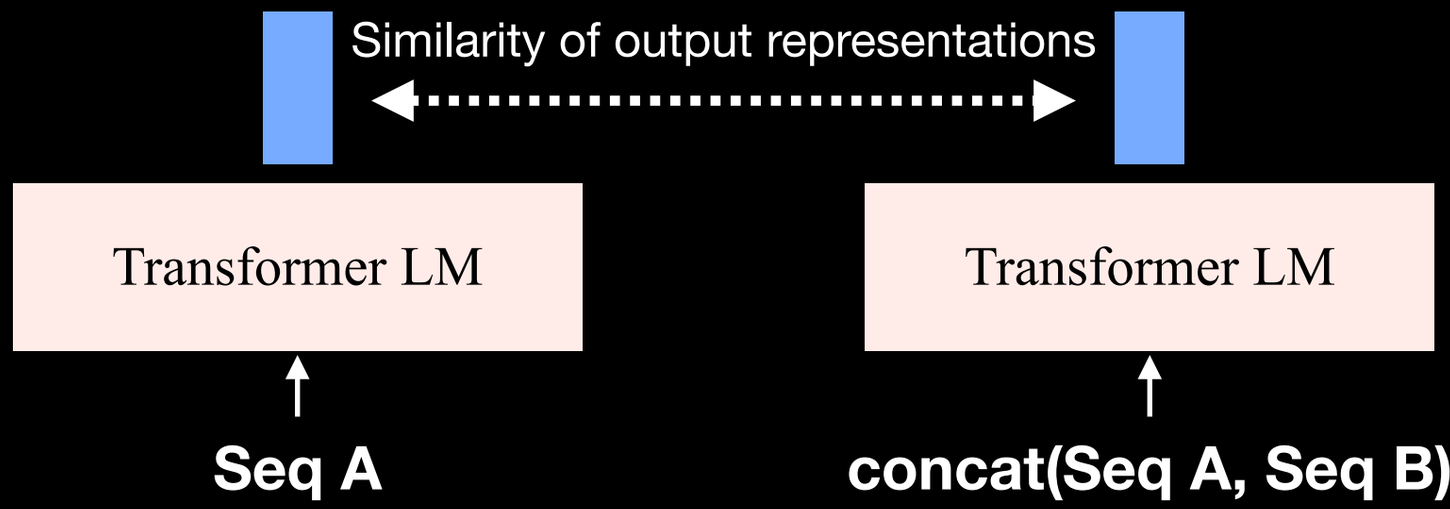
bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**



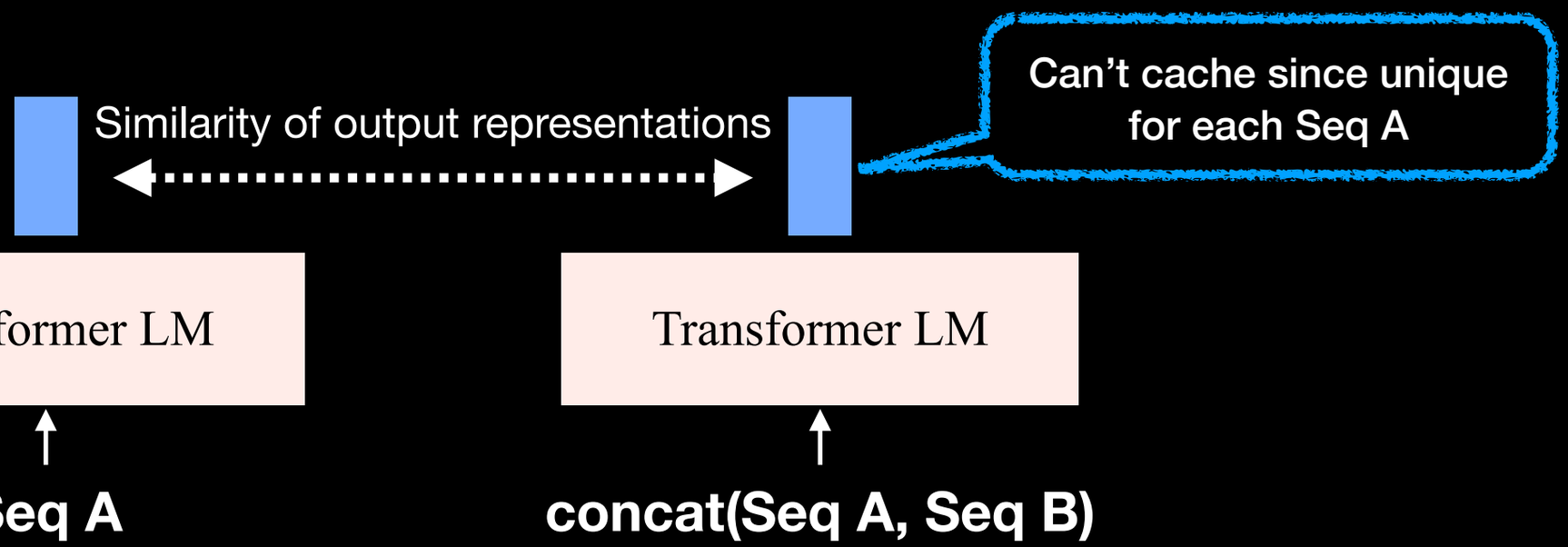
bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**



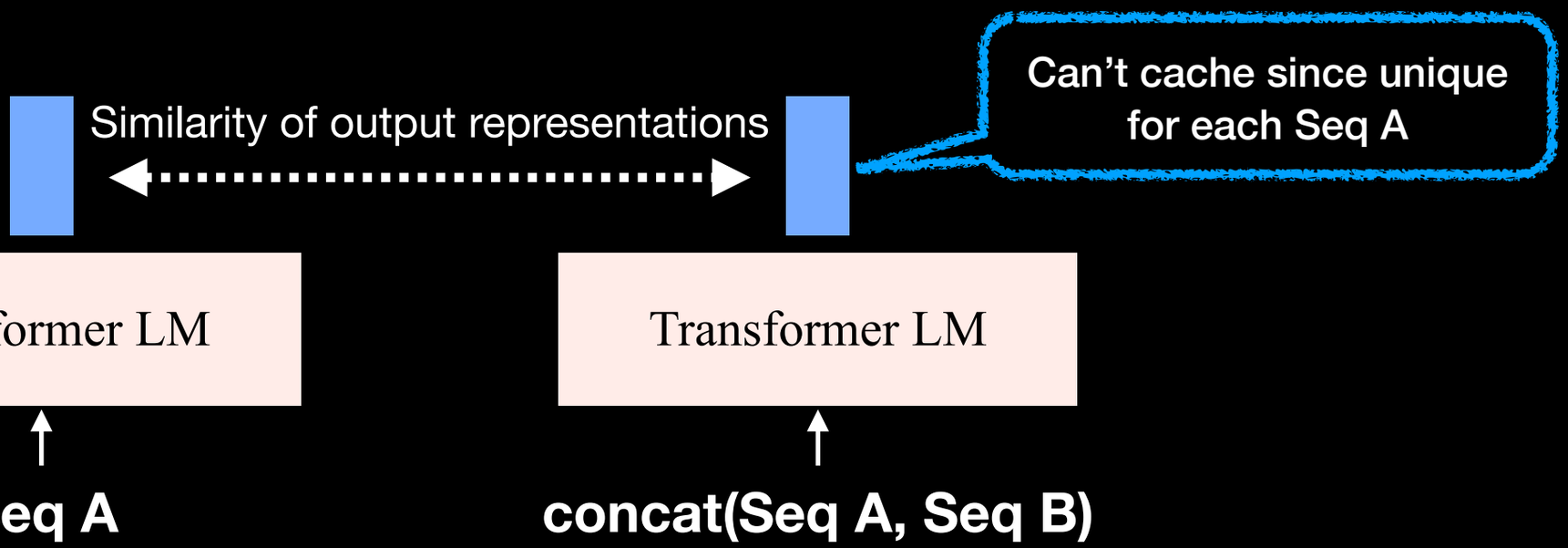
bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**



bi-encoder vs. cross-encoder dilemma

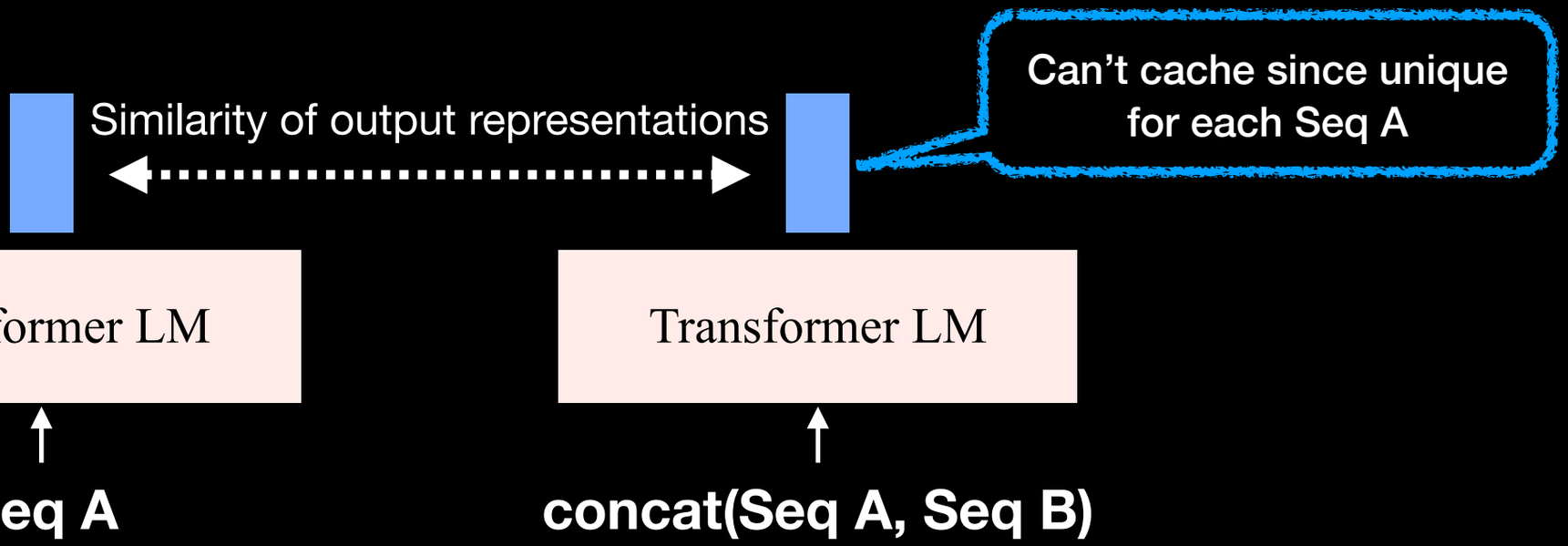
- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**



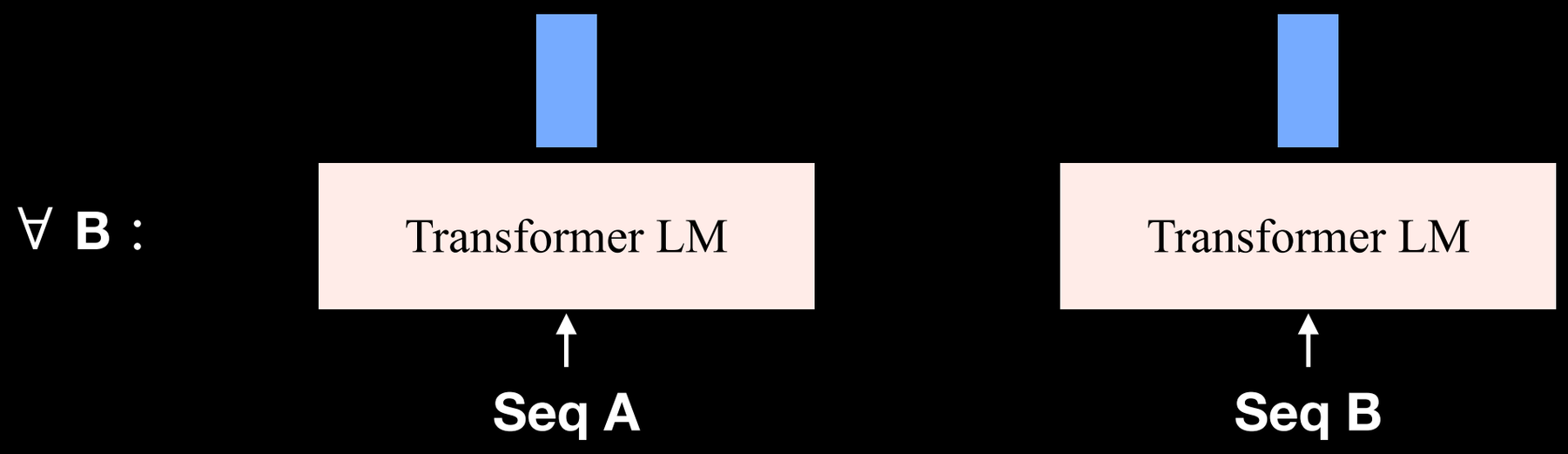
- Instead, resort to matching against representations from independent passes through the network **(bi-encoder)**

bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**

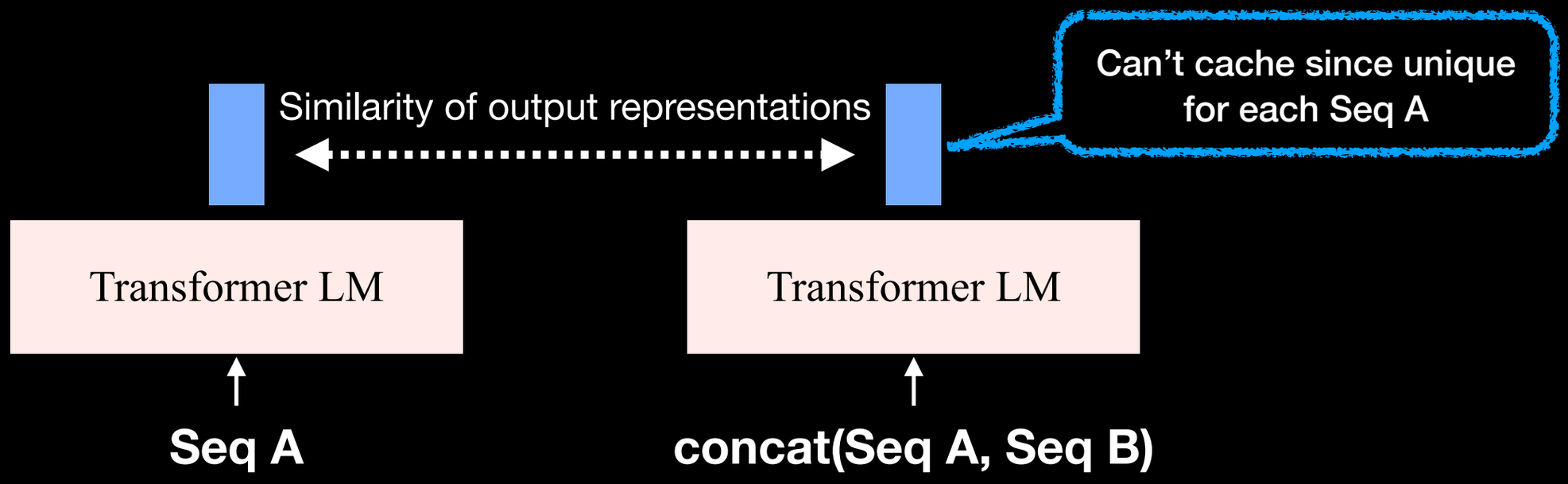


- Instead, resort to matching against representations from independent passes through the network **(bi-encoder)**

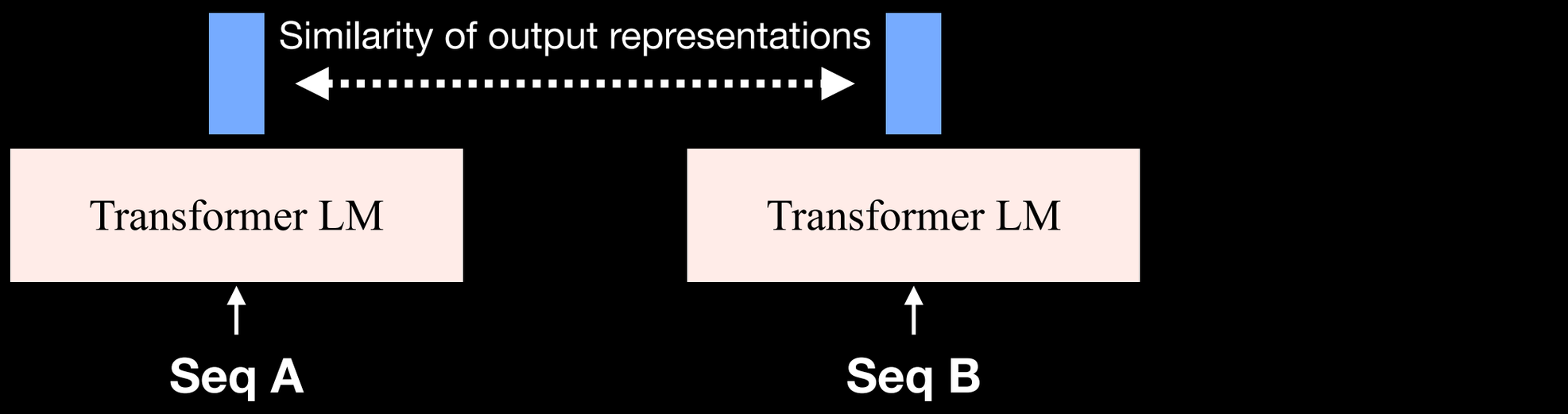


bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**

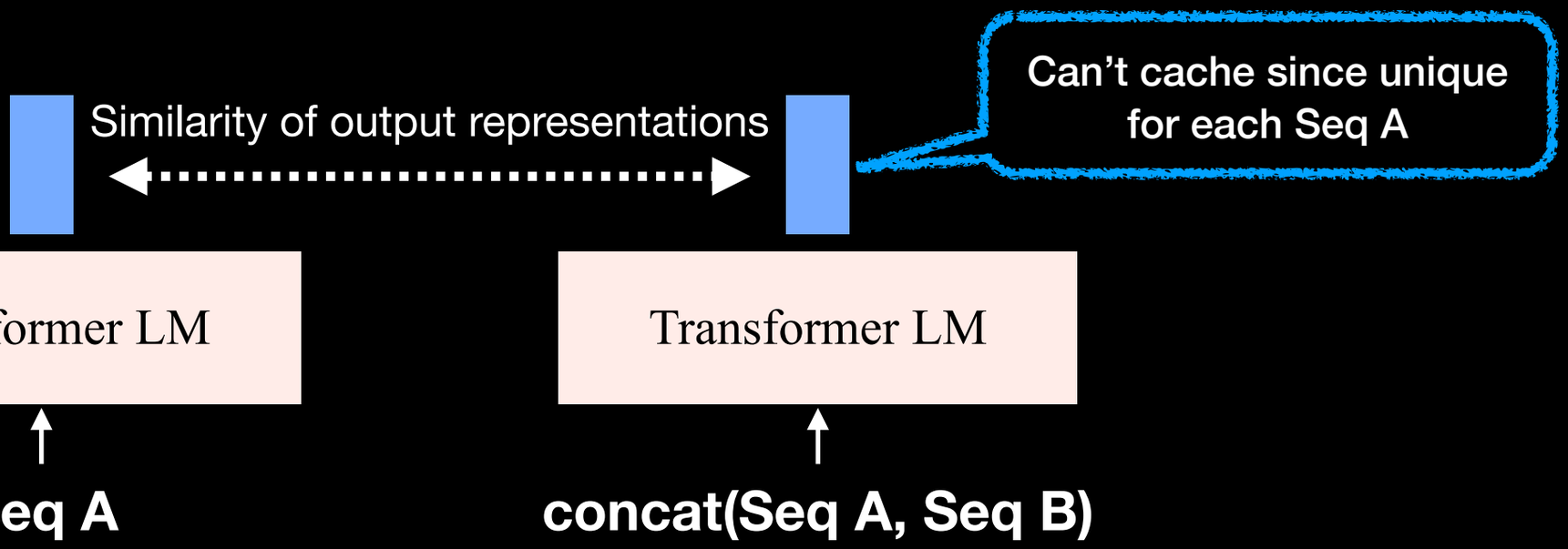


- Instead, resort to matching against representations from independent passes through the network **(bi-encoder)**

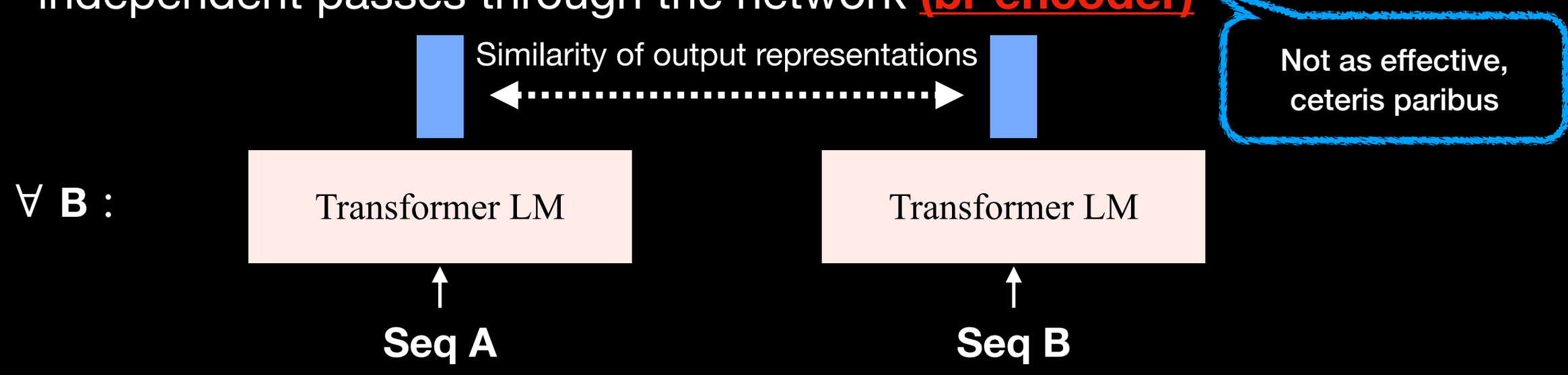


bi-encoder vs. cross-encoder dilemma

- Most effective to compose all relevant sequences in the input to the network **(cross-encoder)**

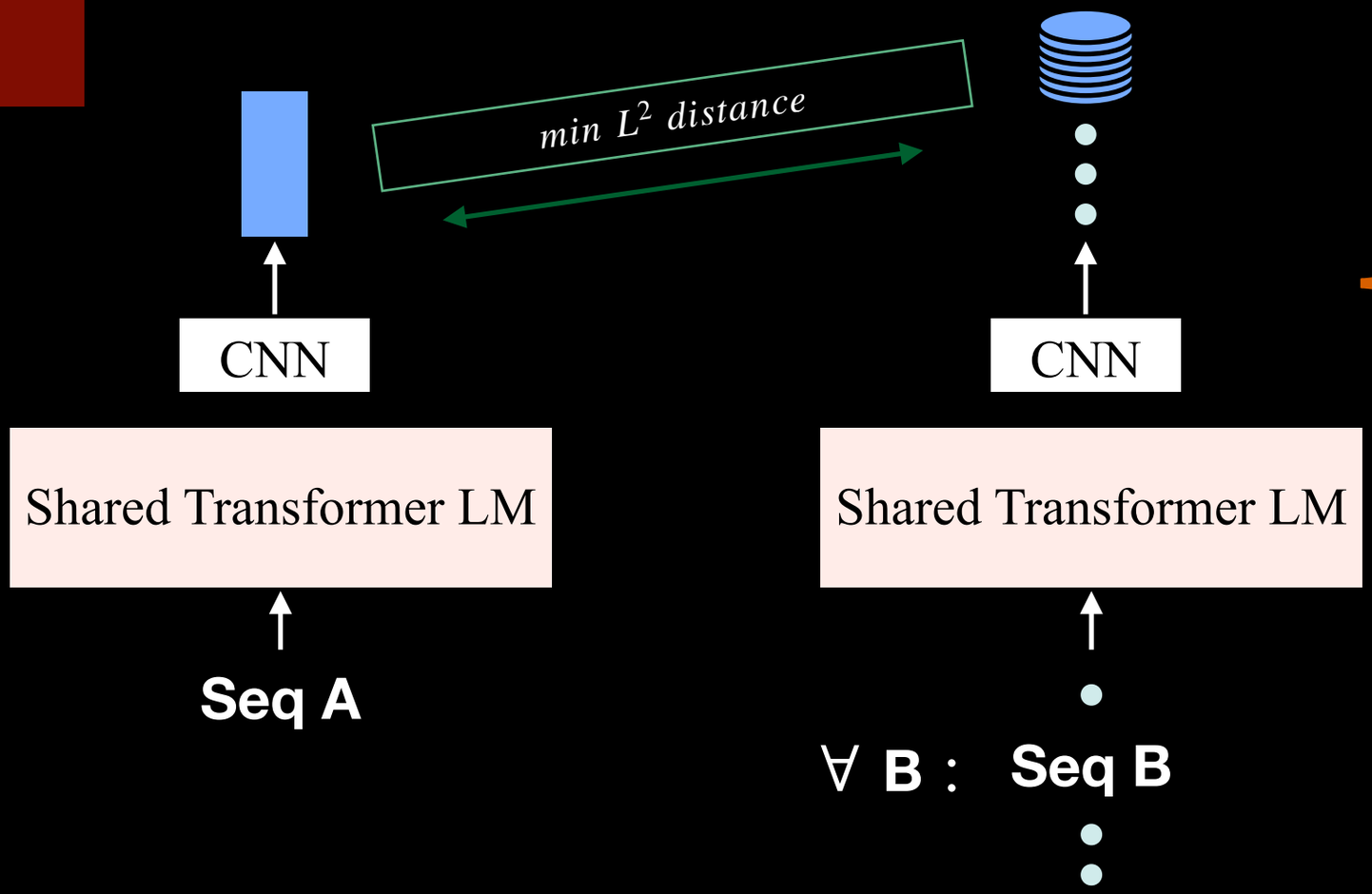


- Instead, resort to matching against representations from independent passes through the network **(bi-encoder)**



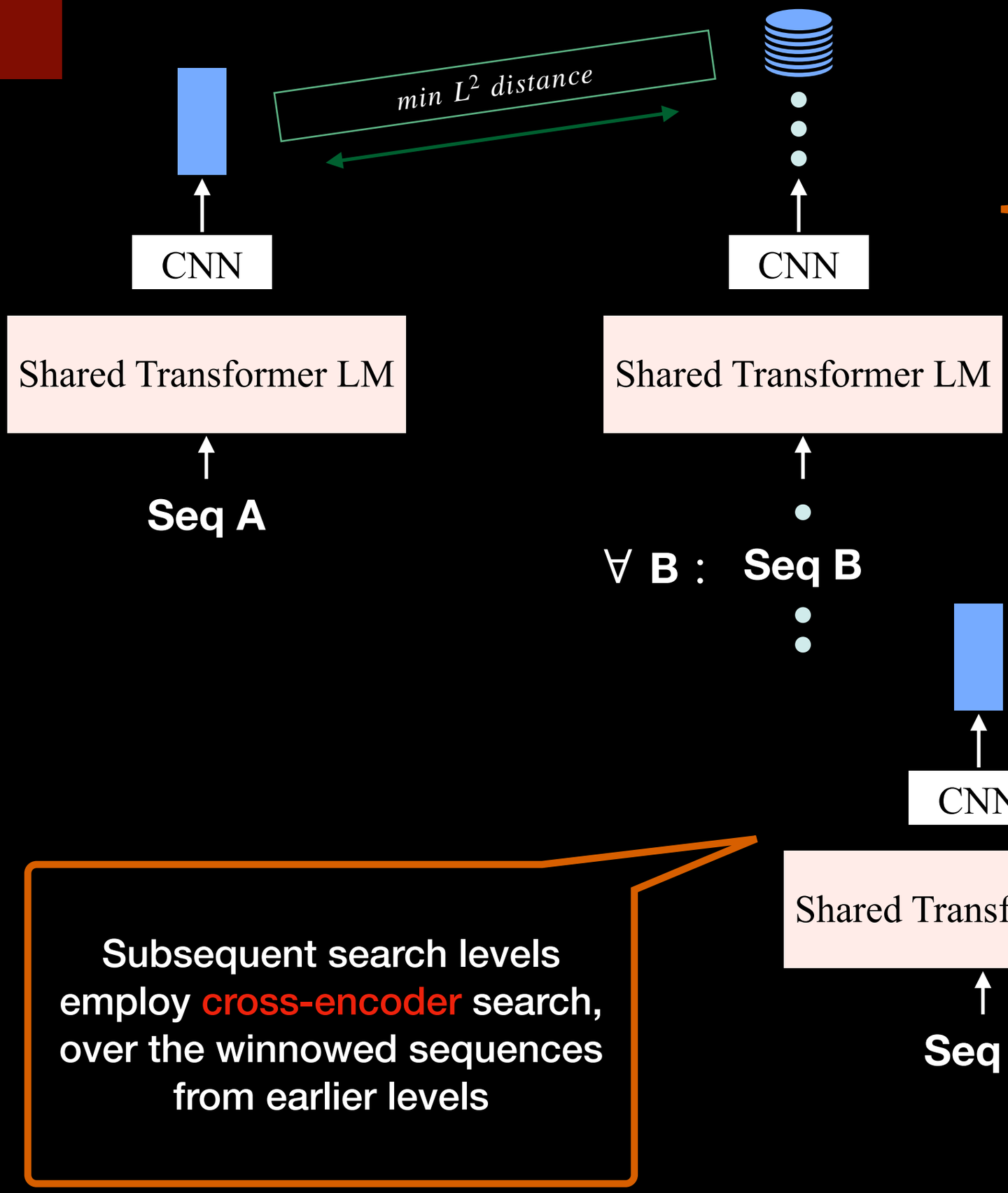
Solution: Add memory layers to Transformer & search over dense representations

Solution: Add memory layers to Transformer & search over dense representations

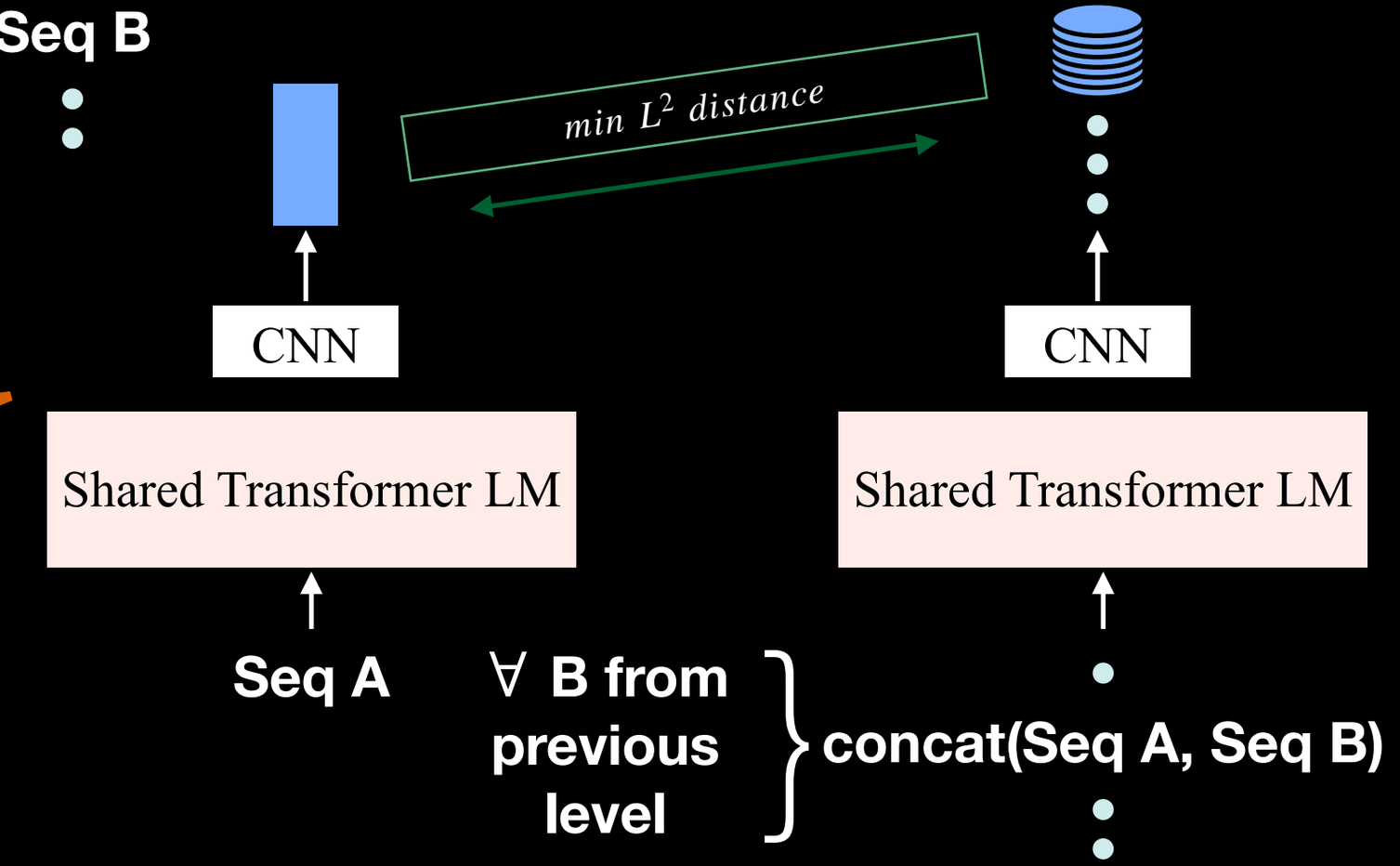


Initial, coarse bi-encoder search

Solution: Add memory layers to Transformer & search over dense representations



Initial, coarse bi-encoder search



Subsequent search levels employ cross-encoder search, over the winnowed sequences from earlier levels

Input Terminology

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.
  - **`Query sequence`**

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.
  - **`Query sequence`**
- Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.
  - **`Query sequence`**
- Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
  - **`Support sequence`**

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.
  - **`Query sequence`**
- Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
  - **`Support sequence`**
- During search, dynamically create **`Query`** & **`Support sequences`**

Input Terminology

- Claim: Charles de Gaulle was a leader in the French Resistance.
  - **`Query sequence`**
- Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
  - **`Support sequence`**
- During search, dynamically create **`Query`** & **`Support sequences`**
  - In final level, also concat CLASSIFICATION LABELS to the **`Support sequences`**

Coarse-to-Fine Search

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]

- Level 2 (cross-encoder retrieval)

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
    - Match [**claim**] to a [**Wikipedia sentence**]
- Level 2 (cross-encoder retrieval)
    - Match [**claim**] to a [**claim + Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]

- Level 2 (cross-encoder retrieval)
  - Match [**claim**] to a [**claim + Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
    - Match [**claim**] to a [**Wikipedia sentence**]
- Level 2 (cross-encoder retrieval)
    - Match [**claim**] to a [**claim + Wikipedia sentence**]

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
    - Match [**claim**] to a [**Wikipedia sentence**]
- Level 2 (cross-encoder retrieval)
    - Match [**claim**] to a [**claim + Wikipedia sentence**]
- Level 3 (cross-encoder classification)

Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
    - Match [**claim**] to a [**Wikipedia sentence**]
- Level 2 (cross-encoder retrieval)
    - Match [**claim**] to a [**claim + Wikipedia sentence**]
- Level 3 (cross-encoder classification)
    - Match [**claim**] to a [**LABEL + claim + Wikipedia sentences**]

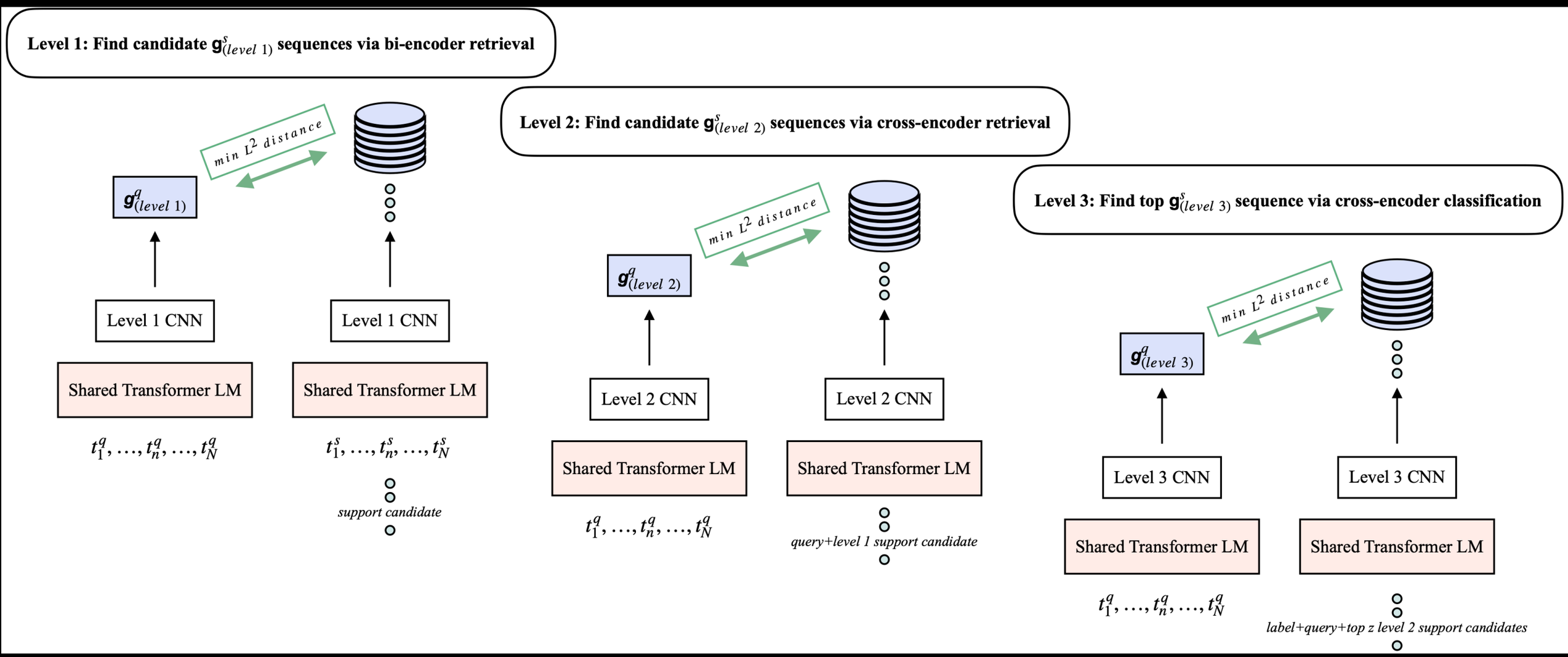
Coarse-to-Fine Search

- For each of 3 levels, we aim to minimize the distance between the **QUERY SEQUENCE** and the correct **SUPPORT SEQUENCE**

- Level 1 (bi-encoder retrieval)
  - Match [**claim**] to a [**Wikipedia sentence**]
- Level 2 (cross-encoder retrieval)
  - Match [**claim**] to a [**claim + Wikipedia sentence**]
- Level 3 (cross-encoder classification)
  - Match [**claim**] to a [**LABEL + claim + Wikipedia sentences**]

QUERY SEQUENCE **SUPPORT SEQUENCE**

Full Model


Level 1: Find candidate $\mathbf{g}^s_{(level\ 1)}$ sequences via bi-encoder retrieval
$min\ L^2\ distance$
$\mathbf{g}^q_{(level\ 1)}$
Level 1 CNN
Level 1 CNN
Shared Transformer LM
Shared Transformer LM
$t^q_1, \ldots, t^q_n, \ldots, t^q_N$
$t^s_1, \ldots, t^s_n, \ldots, t^s_N$
support candidate
Level 2: Find candidate $\mathbf{g}^s_{(level\ 2)}$ sequences via cross-encoder retrieval
$min\ L^2\ distance$
$\mathbf{g}^q_{(level\ 2)}$
Level 2 CNN
Level 2 CNN
Shared Transformer LM
Shared Transformer LM
$t^q_1, \ldots, t^q_n, \ldots, t^q_N$
query+level 1 support candidate
Level 3: Find top $\mathbf{g}^s_{(level\ 3)}$ sequence via cross-encoder classification
$min\ L^2\ distance$
$\mathbf{g}^q_{(level\ 3)}$
Level 3 CNN
Level 3 CNN
Shared Transformer LM
Shared Transformer LM
$t^q_1, \ldots, t^q_n, \ldots, t^q_N$
label+query+top z level 2 support candidates

Training

Training

- Run coarse-to-fine search to find hard negatives and prediction sequences

Training

- Run coarse-to-fine search to find hard negatives and prediction sequences
- Unlike typical ML settings, training set is not static

Training

- Run coarse-to-fine search to find hard negatives and prediction sequences
  - Unlike typical ML settings, training set is not static
- Loss MINIMIZES distance to CORRECT MATCHES

Training

- Run coarse-to-fine search to find hard negatives and prediction sequences
  - Unlike typical ML settings, training set is not static
- Loss MINIMIZES distance to CORRECT MATCHES
- Loss MAXIMIZES distance to WRONG MATCHES

Loss

Minimize difference
to
correct matches

$$\delta_L = \left| g^q - g^s \right| \in \mathbb{R}^M$$

$g^q_{(level\ L)}$

$g^s_{(level\ L)}$

Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t^q_1, \ldots, t^q_n, \ldots, t^q_N$

$t^s_1, \ldots, t^s_n, \ldots, t^s_N$

Loss

Dense representation of query sequence:

$$\boldsymbol{g}^q \in \mathbb{R}^{1000} = \begin{bmatrix} g_1^q \\ g_2^q \\ \vdots \\ g_{1000}^q \end{bmatrix}$$

$\boldsymbol{g}^q_{(level\ L)}$

Minimize difference to correct matches

$$\boldsymbol{\delta}_L = \left| \boldsymbol{g}^q - \boldsymbol{g}^s \right| \in \mathbb{R}^M$$

$\boldsymbol{g}^s_{(level\ L)}$

Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t_1^q, \ldots, t_n^q, \ldots, t_N^q$

$t_1^s, \ldots, t_n^s, \ldots, t_N^s$

Loss

Dense representation of query sequence:

$$\boldsymbol{g}^q \in \mathbb{R}^{1000} = \begin{bmatrix} g_1^q \\ g_2^q \\ \vdots \\ g_{1000}^q \end{bmatrix}$$

$\boldsymbol{g}^q_{(level\ L)}$

Minimize difference to correct matches

$$\boldsymbol{\delta}_L = \left| \boldsymbol{g}^q - \boldsymbol{g}^s \right| \in \mathbb{R}^M$$

$\boldsymbol{g}^s_{(level\ L)}$

Dense representation of support sequence:

$$\boldsymbol{g}^s \in \mathbb{R}^{1000} = \begin{bmatrix} g_1^s \\ g_2^s \\ \vdots \\ g_{1000}^s \end{bmatrix}$$

Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t_1^q, \ldots, t_n^q, \ldots, t_N^q$

$t_1^s, \ldots, t_n^s, \ldots, t_N^s$

Loss

Maximize difference
to
incorrect matches

$\boldsymbol{g}^q_{(level\ L)}$

$\boldsymbol{g}^s_{(level\ L)}$

$\boldsymbol{\delta}_L = \left|\boldsymbol{g}^q - \boldsymbol{g}^s\right| \in \mathbb{R}^M$

Level L CNN

Level L CNN

Shared Transformer LM

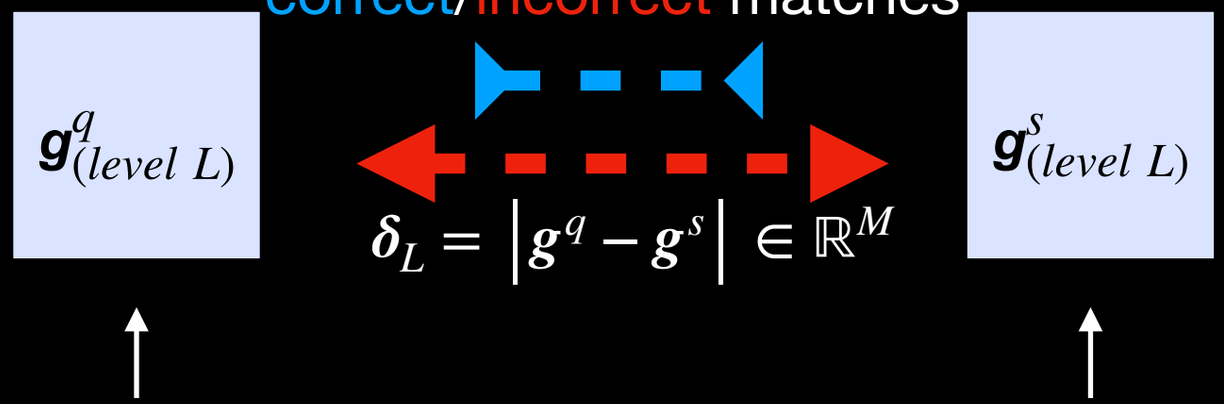
Shared Transformer LM

$t^q_1, \ldots, t^q_n, \ldots, t^q_N$

$t^s_1, \ldots, t^s_n, \ldots, t^s_N$

Loss

Minimize/maximize difference
to
correct/incorrect matches



Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t^q_1, \ldots, t^q_n, \ldots, t^q_N$

$t^s_1, \ldots, t^s_n, \ldots, t^s_N$

Loss

Minimize/maximize difference
to
correct/incorrect matches

$\boldsymbol{g}^q_{(level\ L)}$

$\boldsymbol{g}^s_{(level\ L)}$

$\boldsymbol{\delta}_L = \left|\boldsymbol{g}^q - \boldsymbol{g}^s\right| \in \mathbb{R}^M$

Level L CNN

Level L CNN

Shared Transformer LM

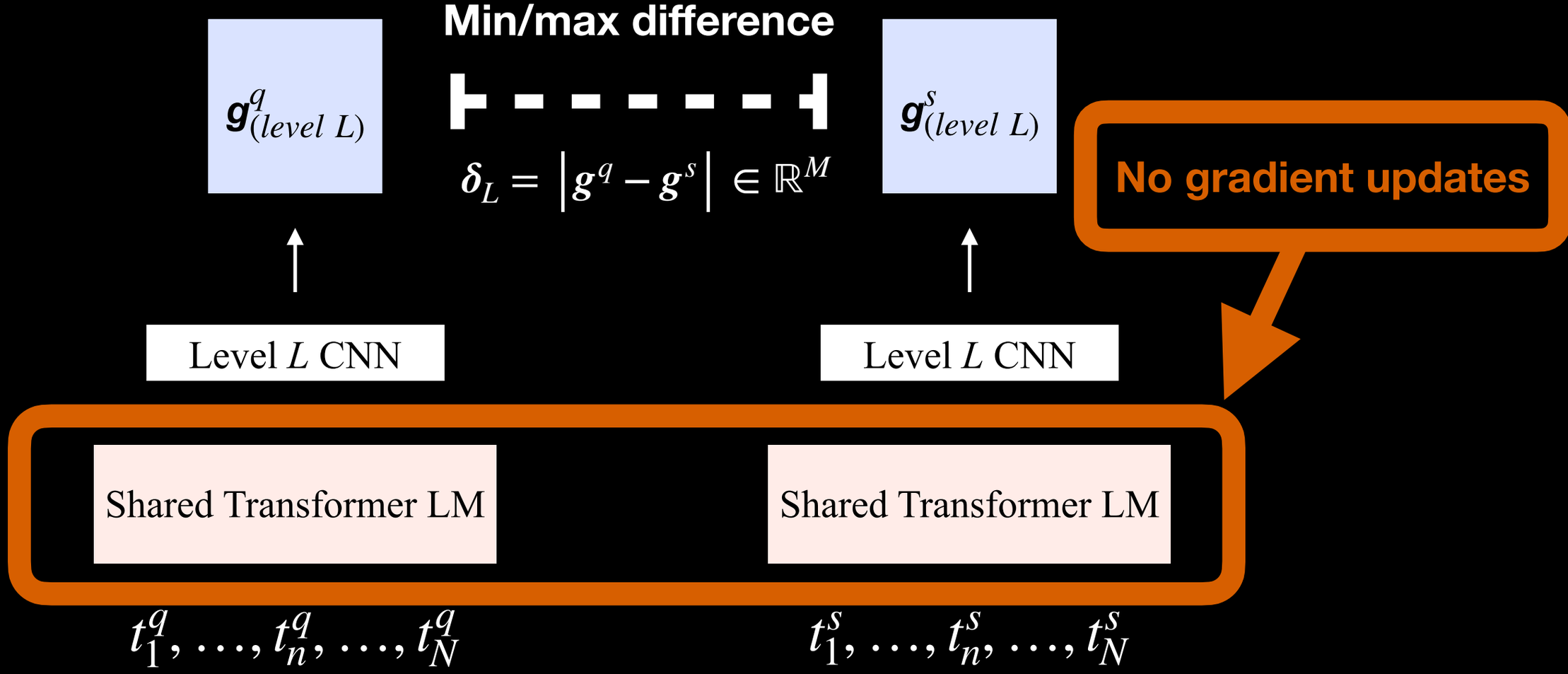
Shared Transformer LM

$t^q_1, \ldots, t^q_n, \ldots, t^q_N$

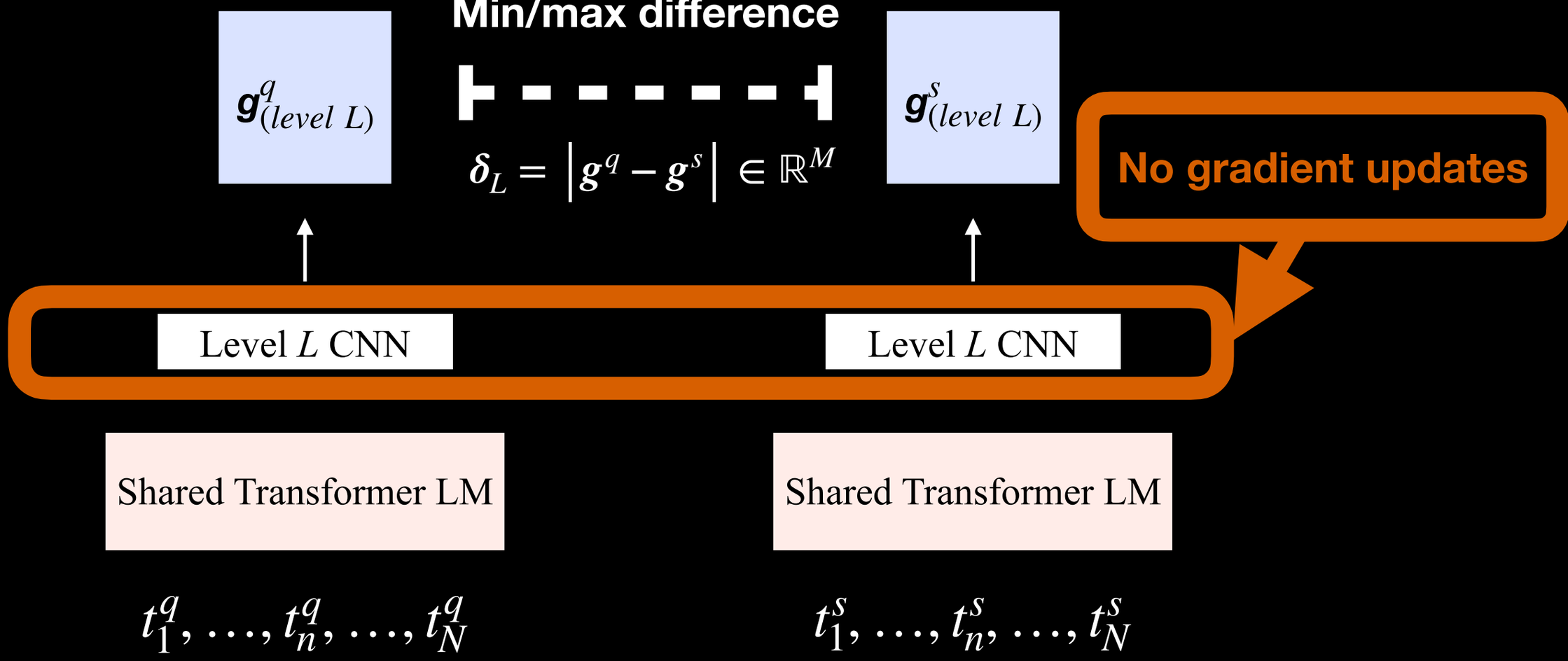
$t^s_1, \ldots, t^s_n, \ldots, t^s_N$

Backprop through
all 3 levels together

Iterative Freezing: Epoch mod 2==0



Iterative Freezing: Epoch mod 2==1



Inference

Inference

- Run coarse-to-fine search

Inference

- Run coarse-to-fine search

- Top level 3 support sequence contains the predicted classification label and evidence sentences

Example - Level 1

Level 1	
QUERY sequence	Claim: Charles de Gaulle was a leader in the French Resistance.
SUPPORT sequence, beam index 0	Evidence: French Resistance, sentence 0: The French Resistance (La Résistance) was the collection of French resistance movements that fought against the Nazi German occupation of France and against the collaborationist Vichy régime during the Second World War.
SUPPORT sequence, beam index 1	Evidence: Charles de Gaulle, sentence 1: He was the leader of Free France (1940 – 44) and the head of the Provisional Government of the French Republic (1944 – 46).
⋮	⋮
SUPPORT sequence, beam index 14	Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
⋮	⋮
SUPPORT sequence, beam index 99	Evidence: Resistance (EP), sentence 7: This EP or mini-album sold nearly all of its 200,000 copies.

The ground-truth evidence sentence is in the 15th beam position in level 1.

Example - Level 2

Level 2	
QUERY sequence	Consider: Claim: Charles de Gaulle was a leader in the French Resistance.
SUPPORT sequence, beam index 0	Consider: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 1: He was the leader of Free France (1940 – 44) and the head of the Provisional Government of the French Republic (1944 – 46).
SUPPORT sequence, beam index 1	Consider: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
SUPPORT sequence, beam index 2	Consider: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 0: Charles André Joseph Marie de Gaulle ([ʃaʁl də gol]; 22 November 1890 – 9 November 1970) was a French general and statesman.

The ground-truth evidence sentence rises to the 2nd beam position in level 2 (i.e., cross-encoding is important).

Example - Level 3

Level 3	
QUERY sequence	Predict: Claim: Charles de Gaulle was a leader in the French Resistance.
SUPPORT sequence, beam index 0	Supports: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 1: He was the leader of Free France (1940 – 44) and the head of the Provisional Government of the French Republic (1944 – 46). Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance. Evidence: Charles de Gaulle, sentence 0: Charles André Joseph Marie de Gaulle ([ʃaʁl də gol]; 22 November 1890 – 9 November 1970) was a French general and statesman.
SUPPORT sequence, beam index 1	Unverifiable: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 1: He was the leader of Free France (1940 – 44) and the head of the Provisional Government of the French Republic (1944 – 46). Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance. Evidence: Charles de Gaulle, sentence 0: Charles André Joseph Marie de Gaulle ([ʃaʁl də gol]; 22 November 1890 – 9 November 1970) was a French general and statesman.
SUPPORT sequence, beam index 2	Refutes: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 1: He was the leader of Free France (1940 – 44) and the head of the Provisional Government of the French Republic (1944 – 46). Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance. Evidence: Charles de Gaulle, sentence 0: Charles André Joseph Marie de Gaulle ([ʃaʁl də gol]; 22 November 1890 – 9 November 1970) was a French general and statesman.

In level 3, we include the classification label and the top 3 predicted evidence sentences from level 2. The final prediction is the top of the level 3 beam.

Example - Level 3 - training

Level 3 (training only)	
QUERY sequence	Reference: Claim: Charles de Gaulle was a leader in the French Resistance.
SUPPORT sequence, positive training instance	Supports: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
SUPPORT sequence, negative training instance	Refutes: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.
SUPPORT sequence, negative training instance	Unverifiable: Claim: Charles de Gaulle was a leader in the French Resistance. Evidence: Charles de Gaulle, sentence 12: Despite frosty relations with Britain and especially the United States, he emerged as the undisputed leader of the French resistance.

For training, in level 3 we also* create positive and negative instances by flipping the label on the ground-truth evidence sentences.

(*in addition to hard negatives and predictions from search)

Empirical Comparisons

Model	Acc.	FEV.	Pt.ct.	IR	Ling.
GEAR	71.60	67.10	> 204	•	•
DREAM	76.85	70.60	$\geq (373, 833]$	•	•
CompoundLabel	66.21	61.65	18	•	
NSMN	68.16	64.23	28	•	•
$\text{BERT}_{\text{LARGE}}$	38.	N/A	340		
$\text{BERT}_{\text{LARGE}}$+FT	57.	N/A	340		
$\text{BERT}_{\text{LARGE}}$+KBFeat	49.	N/A	> 340		•
RAG	72.5	N/A	626		
$\text{BERT}_{\text{BASE}}$+MemMatch	70.42	63.95	120		

FEVER hidden test results, with light-gray rows indicating end-to-end models
Accuracy (Acc.); FEVER score (FEV.); Parameter estimates, in millions (Pt.ct.); Non-neural IR features (IR); Linguistic tools (Ling.); Our model is **$\text{BERT}_{\text{BASE}}$ + MemMatch**

Analysis

Analysis Properties

- Level distances: Can use distances at each search level to analyze and constrain the model

- Exemplar auditing: Can create exemplar vectors via the differences between the query and support sequences

Level Distances

Euclidean distance for nearest predicted matches at each level (retrieval & classification)

$\boldsymbol{g}^{q}_{(level\ L)}$

$\boldsymbol{g}^{s}_{(level\ L)}$

Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t^{q}_{1}, \ldots, t^{q}_{n}, \ldots, t^{q}_{N}$

$t^{s}_{1}, \ldots, t^{s}_{n}, \ldots, t^{s}_{N}$

Level Distances

Euclidean distance for nearest predicted matches at each level (retrieval & classification)

$\boldsymbol{g}^{q}_{(level\ L)}$

$$distance(\boldsymbol{g}^{\mathbf{q}}, \boldsymbol{g}^{\mathbf{s}}) = \|\boldsymbol{g}^{\mathbf{q}} - \boldsymbol{g}^{\mathbf{s}}\|_2 \in \mathbb{R}^1$$

$\boldsymbol{g}^{s}_{(level\ L)}$

Level L CNN

Level L CNN

Shared Transformer LM

Shared Transformer LM

$t^q_1, \ldots, t^q_n, \ldots, t^q_N$

$t^s_1, \ldots, t^s_n, \ldots, t^s_N$

Analysis: Level Distances

- Smaller level distances associated with more reliable predictions on challenge set that modifies Wikipedia and claims

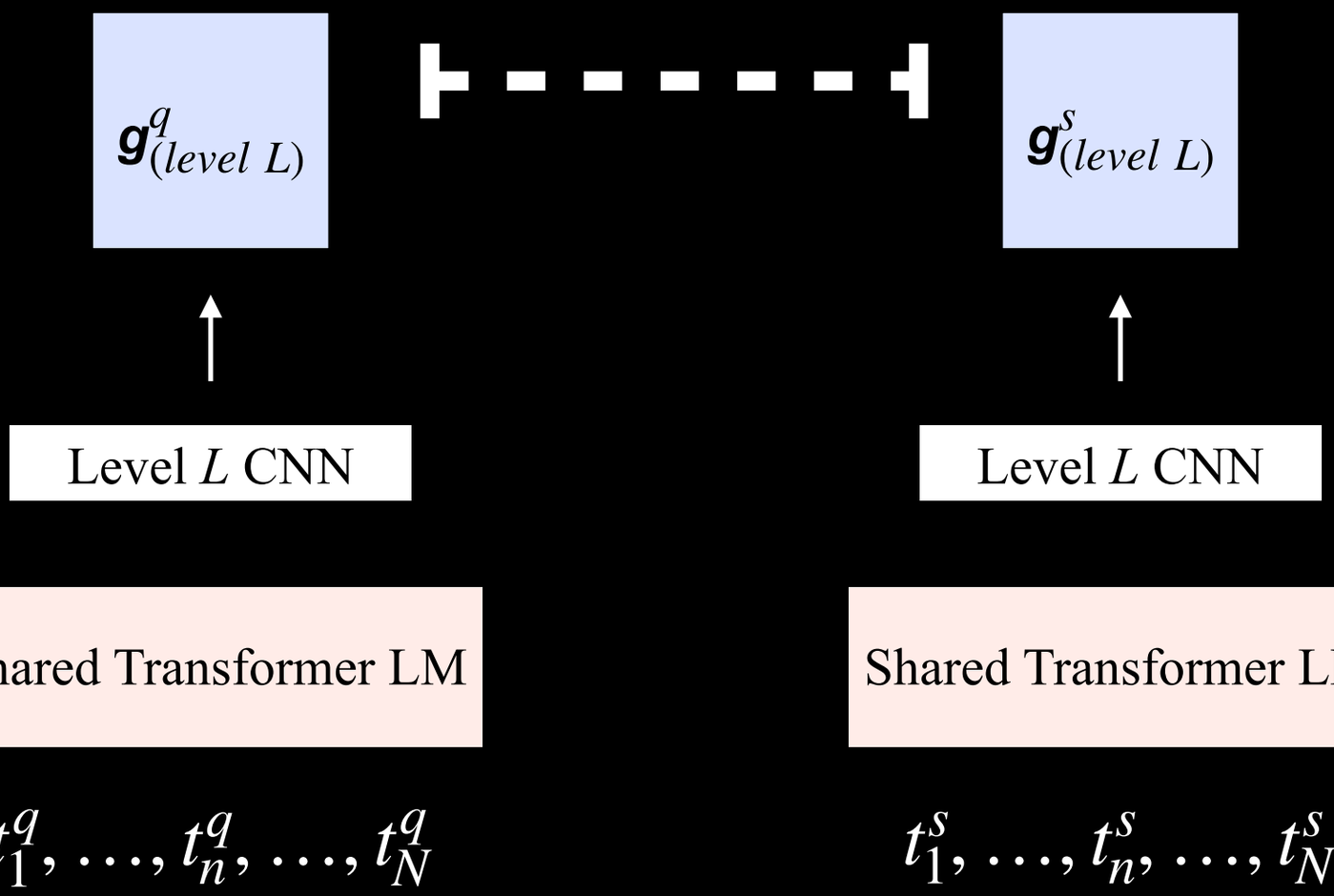
Analysis: Level Distances

- Smaller level distances associated with more reliable predictions on challenge set that modifies Wikipedia and claims

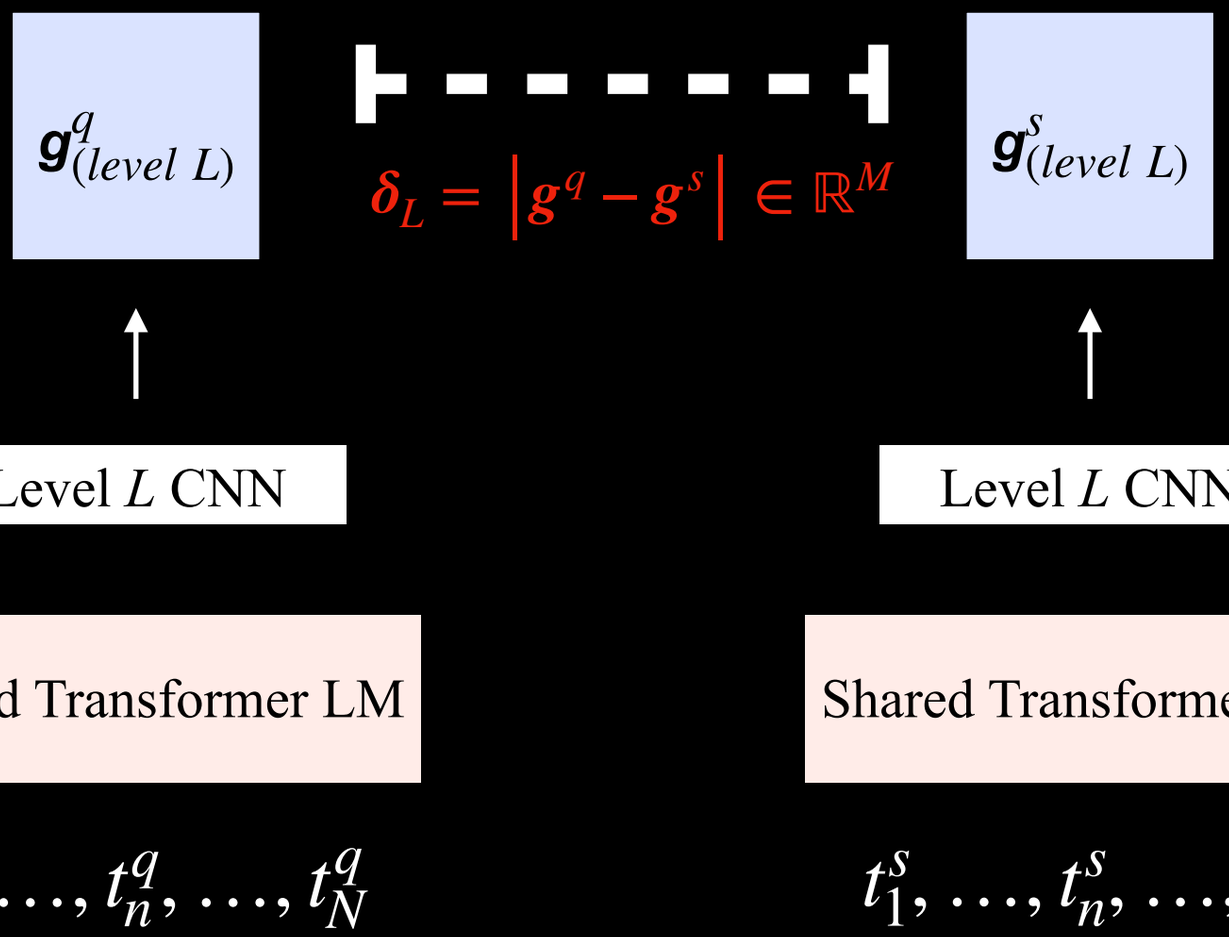
	Claim id 2024540000004 (reference label: FALSE)
$\text{SUPPORT}_{\text{LEVEL3}}$ sequence Level 2 distance: 200.89 Level 3 distance: 2.39	Supports: Claim: Tinker Tailor Soldier Spy is an espionage film. Evidence: Tinker Tailor Soldier Spy (film), sentence 0: Tinker Tailor Soldier Spy is a 2011 music video by Tomas Alfredson.
	Claim id 1390370000004 (reference label: TRUE)
$\text{SUPPORT}_{\text{LEVEL3}}$ sequence Level 2 distance: 0.05 Level 3 distance: 0.54	Supports: Claim: Star Trek: Discovery is an album. Evidence: Star Trek: Discovery, sentence 0: Star Trek: Discovery is an upcoming music album of Lady Gaga.

The mean level 2 distance from the training set at the top of the beam, given a correct retrieval, is 0.49 (+/- 4.75), and the mean level 3 distance, given a correct classification, is 0.92 (+/- 1.80). Incorrect classification labels and distances > mean are in red.

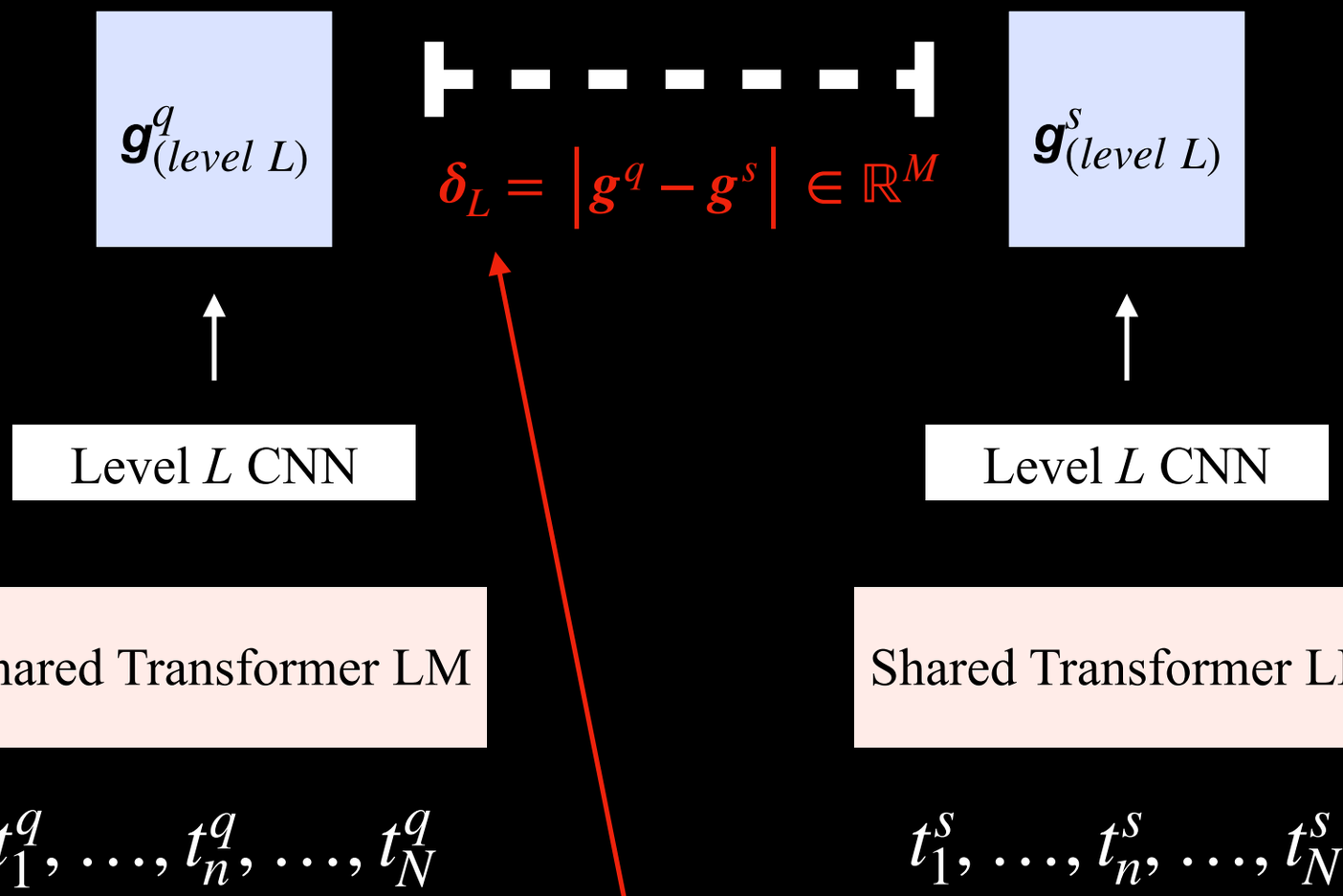
Exemplar Auditing



Exemplar Auditing

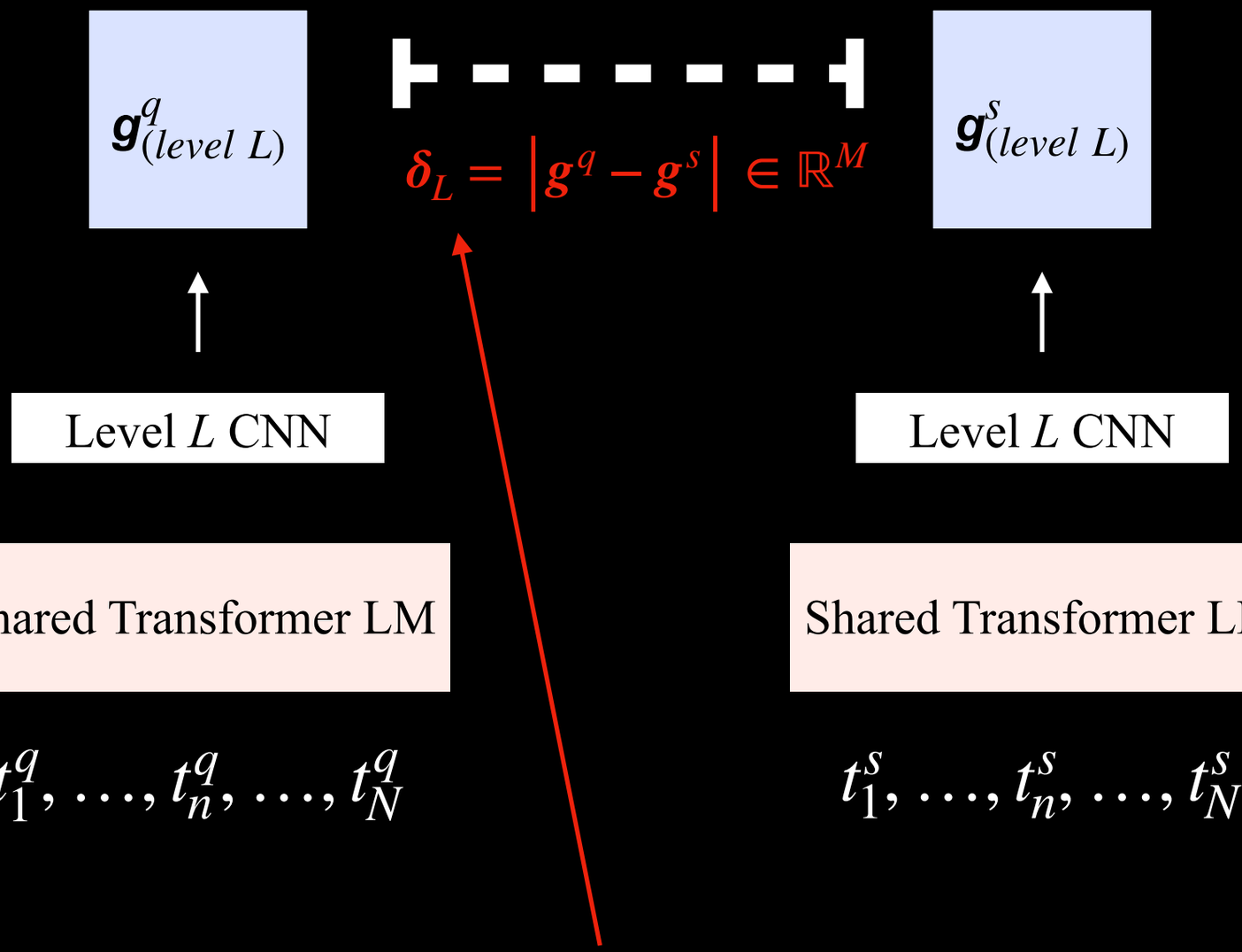


Exemplar Auditing

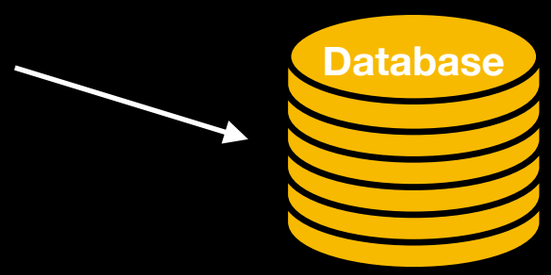


- Create exemplar vectors from difference vectors from final 2 levels:
  - concat(δ_2, δ_3)

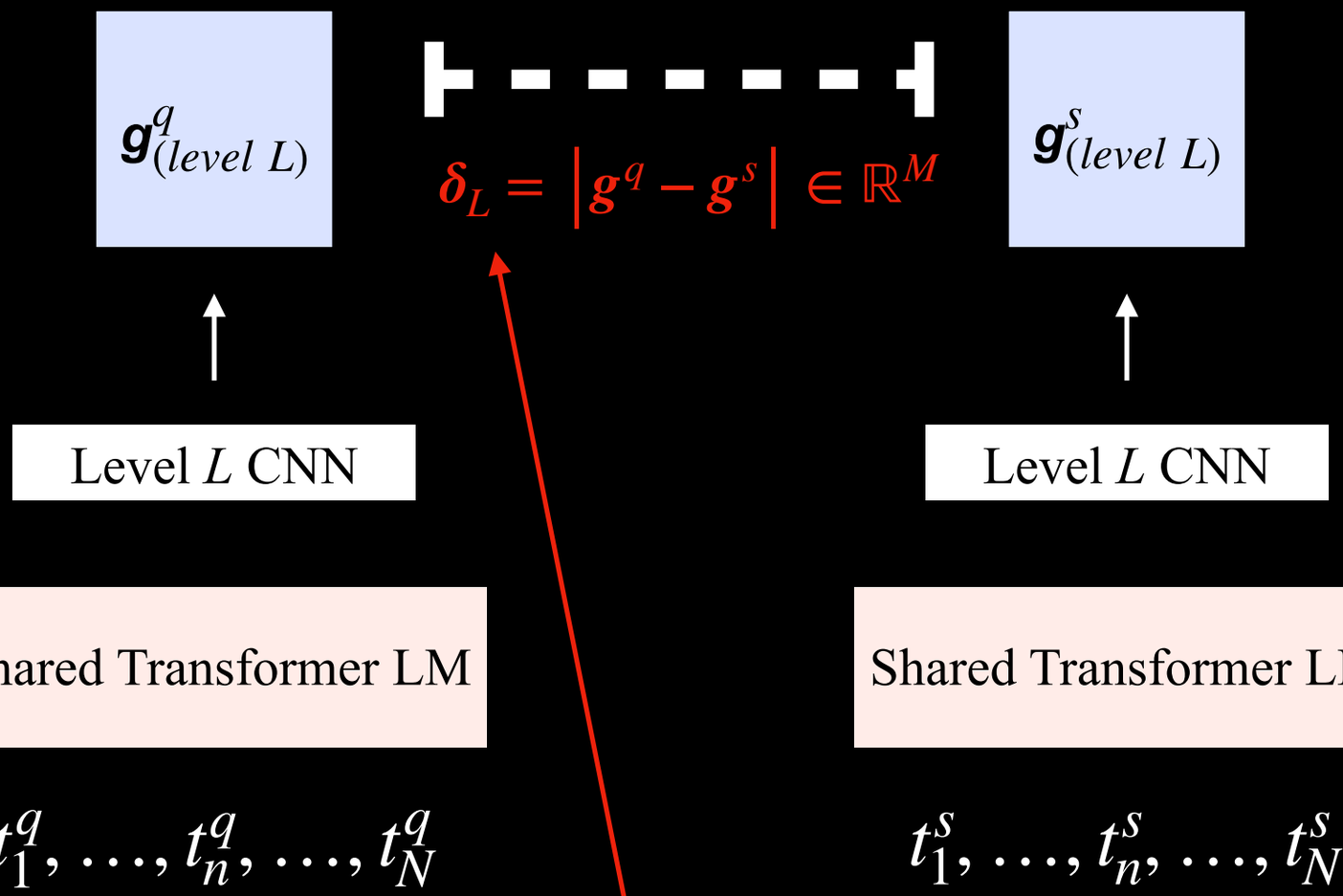
Exemplar Auditing



- Create exemplar vectors from difference vectors from final 2 levels:
  - concat(δ_2, δ_3)



Exemplar Auditing



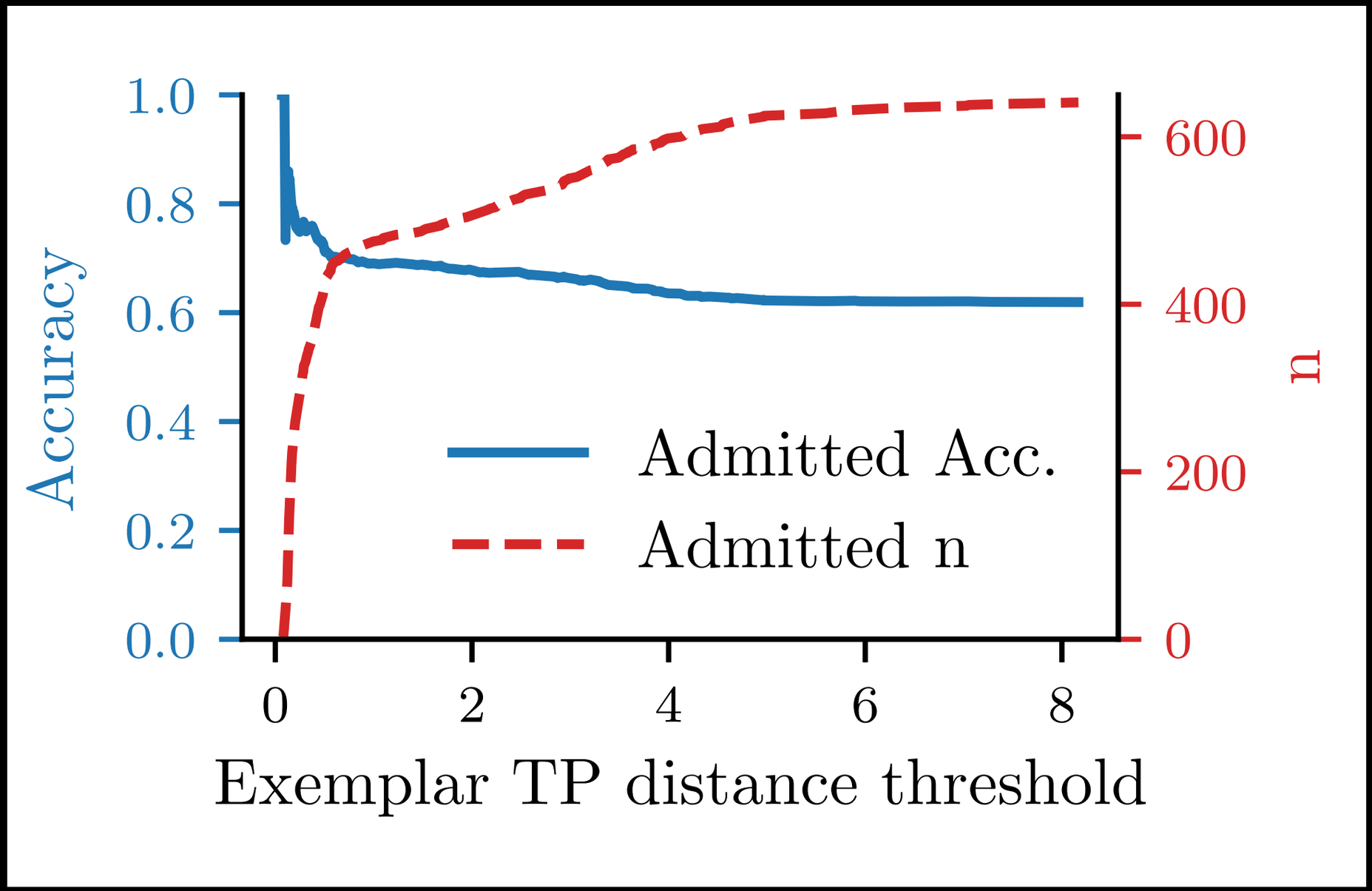
- Create exemplar vectors from difference vectors from final 2 levels:
  - concat(δ_2, δ_3)

Note: Exemplar database is distinct from datastores created during the coarse-to-fine search

Exemplar Auditing

- Associate test with nearest exemplar vectors in the database, only admitting predictions for True Positive (TP) exemplars
- Instances matched to TP exemplars with closer distances are associated with more accurate model predictions on challenge set

Exemplar Auditing



Exemplar Auditing

	Claim id 147493 (reference label: TRUE)
Test $\text{SUPPORT}_{\text{LEVEL3}}$ sequence	`Supports: Claim: T2 Trainspotting is set in and around a city. Evidence: T2 Trainspotting, sentence 0: T2 Trainspotting is a 2017 British comedy drama film, set in and around Edinburgh, Scotland.`
Exemplar $\text{SUPPORT}_{\text{LEVEL3}}$ sequence Exemplar distance: 0.14	`Supports: Claim: All My Children is set in a fictional suburb of a city. Evidence: All My Children, sentence 1: Created by Agnes Nixon, All My Children is set in Pine Valley, Pennsylvania, a fictional suburb of Philadelphia, which is modeled on the actual Philadelphia suburb of Rosemont.`
	Claim id 166634 (reference label: FALSE)
Test $\text{SUPPORT}_{\text{LEVEL3}}$ sequence	`Refutes: Claim: Anne Rice was born in Japan. Evidence: Anne Rice, sentence 5: Born in New Orleans, Rice spent much of her early life there before moving to Texas, and later to San Francisco.`
Exemplar $\text{SUPPORT}_{\text{LEVEL3}}$ sequence Exemplar distance: 0.21	`Refutes: Claim: Emma Stone was born in Taiwan. Evidence: Emma Stone, sentence 5: Born and raised in Scottsdale, Arizona, Stone began acting as a child, in a theater production of The Wind in the Willows in 2000.`

The close distance mappings between the test and exemplar instances tend to exhibit similar abstract, relational patterns

Exemplar Auditing

- Can update the model behavior by modifying the labels and/or instances in the exemplar database
- We illustrate this behavior on the challenge set (see paper for details)

Discussion

Broader Implications

Broader Implications

- In principle, model can be frozen and substituted in for other settings using pre-trained Transformers, but with the new retrieval and analysis functionalities

Broader Implications

- In principle, model can be frozen and substituted in for other settings using pre-trained Transformers, but with the new retrieval and analysis functionalities

- Many real-world tasks can be re-cast to the retrieval-classification setting

- EHR data, Proteins, Medical QA, …

Broader Implications

- In principle, model can be frozen and substituted in for other settings using pre-trained Transformers, but with the new retrieval and analysis functionalities

- Many real-world tasks can be re-cast to the retrieval-classification setting

  - EHR data, Proteins, Medical QA, …

- Prospectively, multi-hop reasoning and even generation can be re-cast to this setting via a deeper search graph

  - Analysis advantages + offload model capacity to datastores

Summary: Full Resolution Language/Sequence Modeling

Summary: Full Resolution Language/Sequence Modeling

- Updatability via the retrieval datastore

Summary: Full Resolution Language/Sequence Modeling

- Updatability via the retrieval datastore
- Updatability via exemplar auditing: Both in terms of difference vectors and token-level memories in the case of BLADE/multiBLADE (earlier work)

Summary: Full Resolution Language/Sequence Modeling

- Updatability via the retrieval datastore
- Updatability via exemplar auditing: Both in terms of difference vectors and token-level memories in the case of BLADE/multiBLADE (earlier work)
- Visualization (and associated analysis use cases) of level alignments and token-level contributions

Summary: Full Resolution Language/Sequence Modeling

- Updatability via the retrieval datastore
- Updatability via exemplar auditing: Both in terms of difference vectors and token-level memories in the case of BLADE/multiBLADE (earlier work)
- Visualization (and associated analysis use cases) of level alignments and token-level contributions
    - With token-level contributions, flexibility to impose priors (e.g., associate one token with the final decision)

Summary: Full Resolution Language/Sequence Modeling

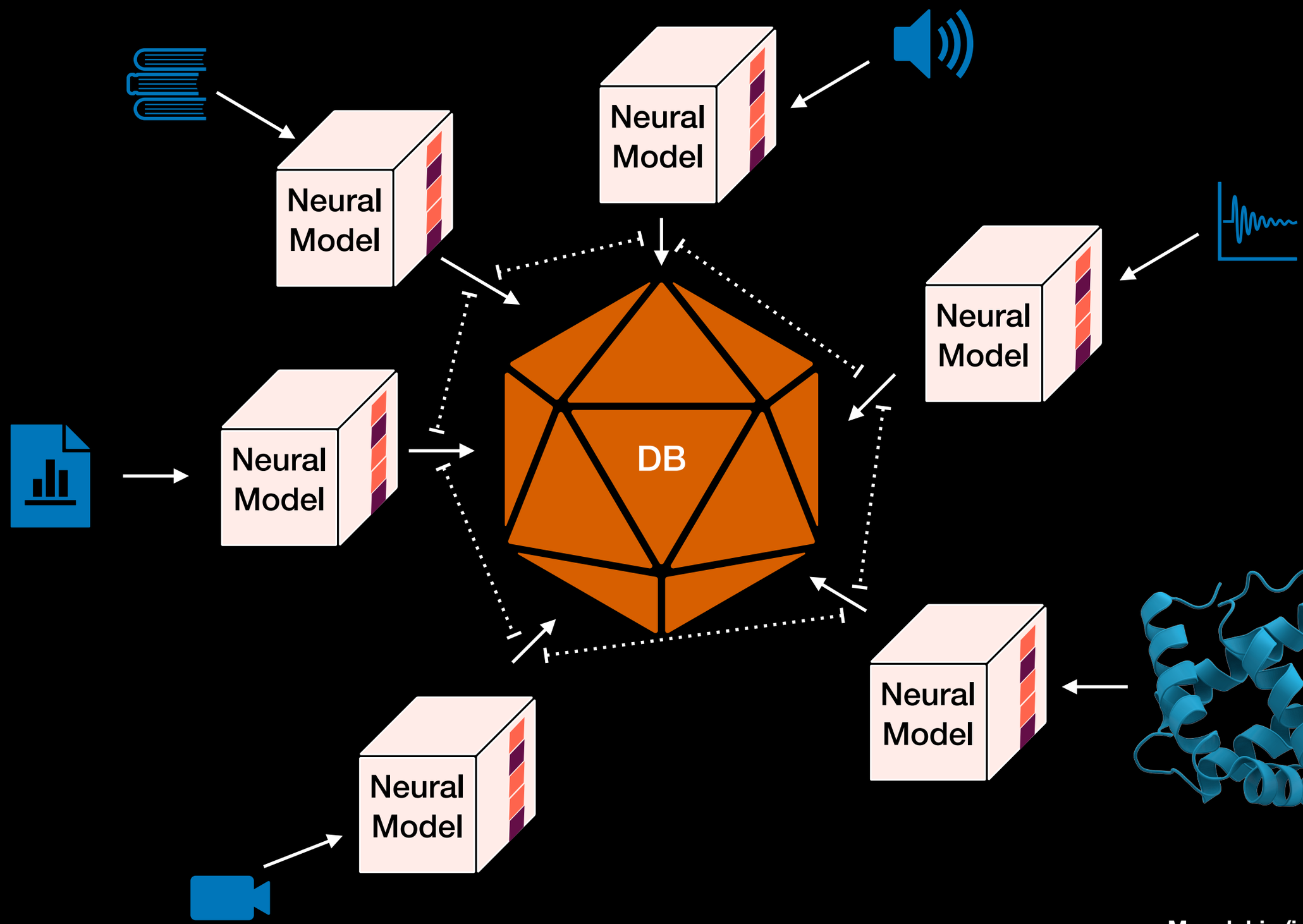
- Updatability via the retrieval datastore
- Updatability via exemplar auditing: Both in terms of difference vectors and token-level memories in the case of BLADE/multiBLADE (earlier work)
- Visualization (and associated analysis use cases) of level alignments and token-level contributions
    - With token-level contributions, flexibility to impose priors (e.g., associate one token with the final decision)
- Ability to constrain and analyze the model via level distances and exemplar distances

Summary: Full Resolution Language/Sequence Modeling

- Updatability via the retrieval datastore
- Updatability via exemplar auditing: Both in terms of difference vectors and token-level memories in the case of BLADE/multiBLADE (earlier work)
- Visualization (and associated analysis use cases) of level alignments and token-level contributions
    - With token-level contributions, flexibility to impose priors (e.g., associate one token with the final decision)
- Ability to constrain and analyze the model via level distances and exemplar distances
- Ability to analyze corpora/datasets via the above and the sequence feature weighting as demonstrated in the BLADE paper (i.e., defacto extractive, comparative summarization).

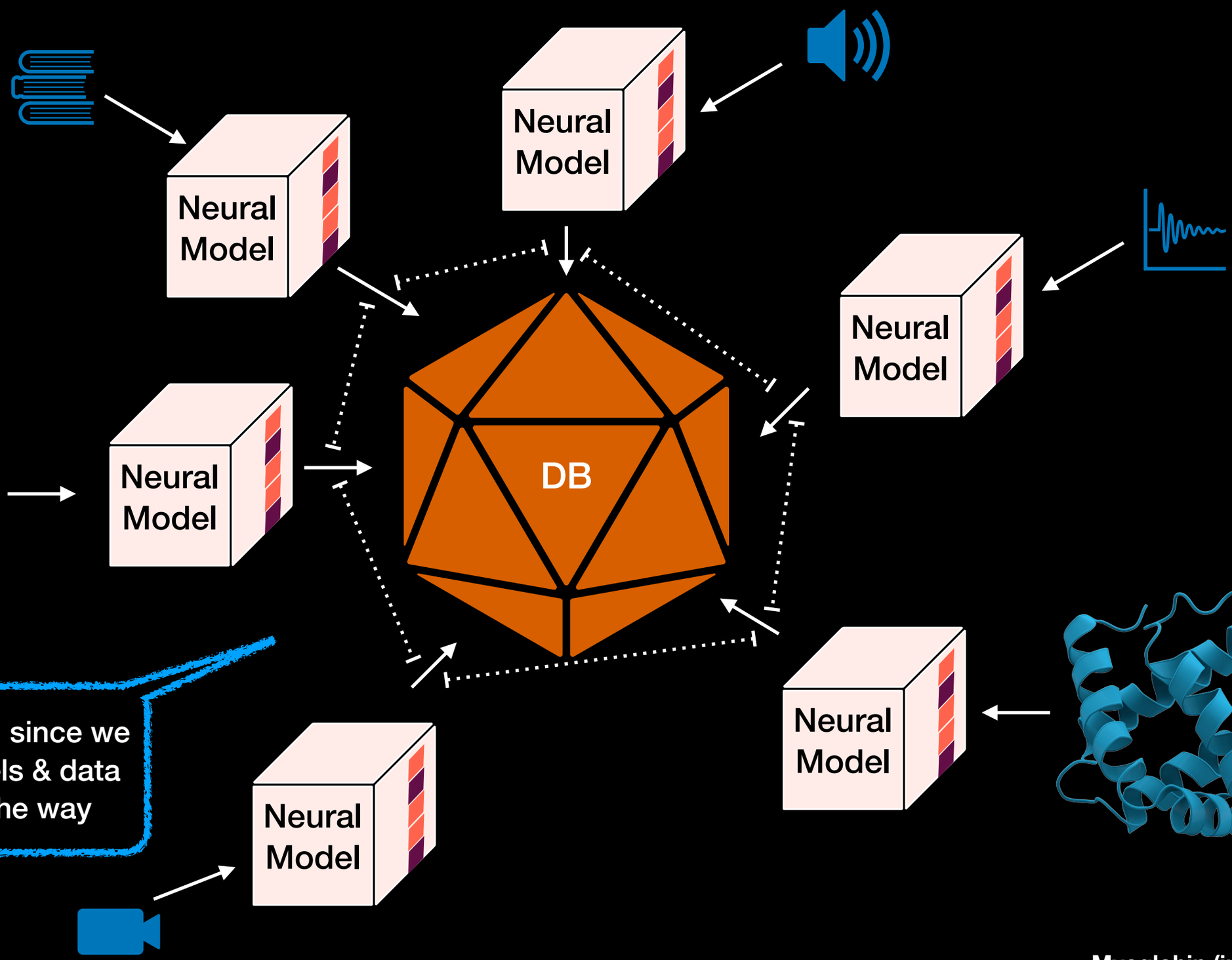
Framework for Future Work

Prospective: Interlocking distance constraints across input modalities/tasks with a single model...



Myoglobin (image from Wikipedia)

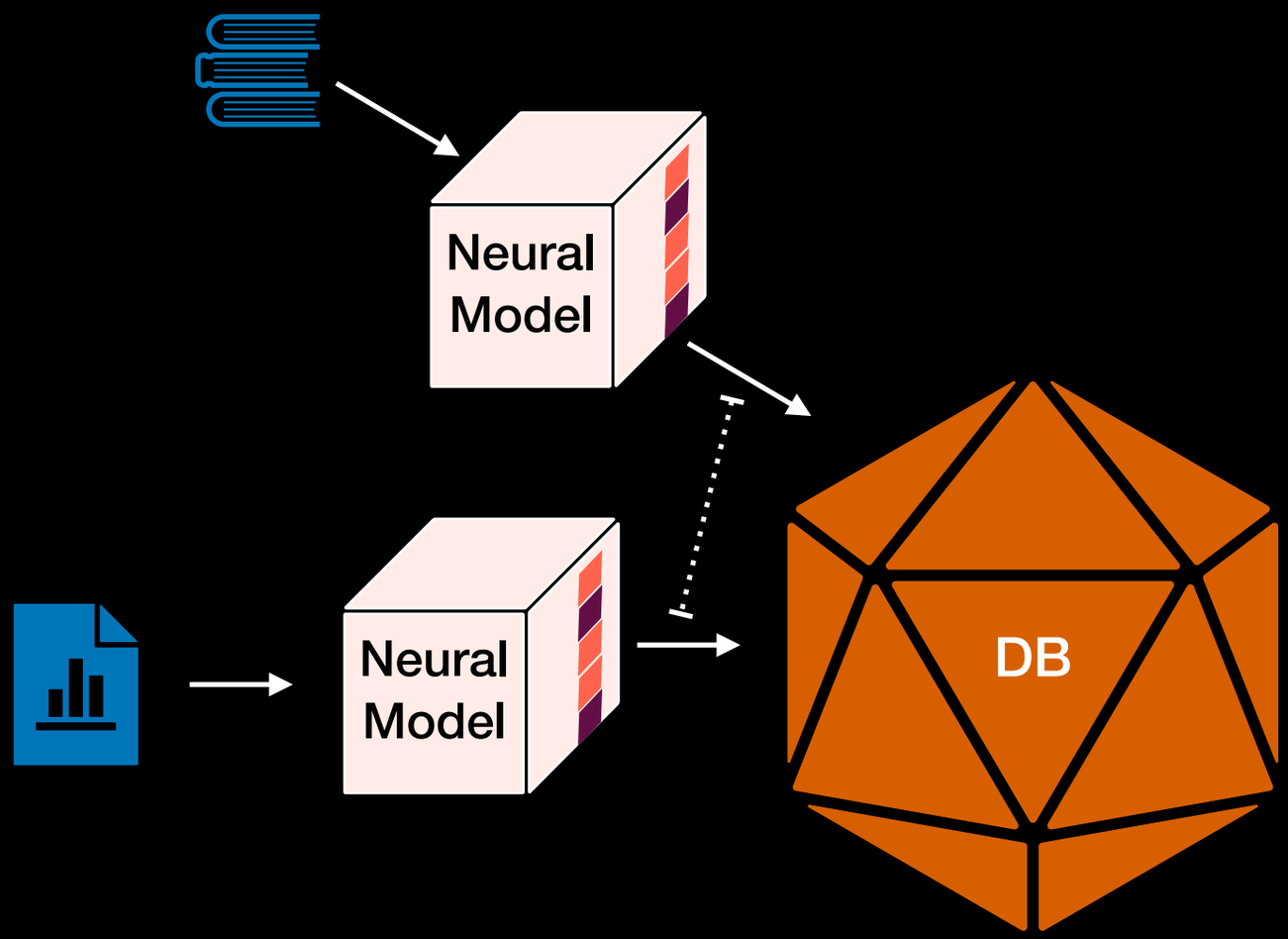
Prospective: Interlocking distance constraints across input modalities/tasks with a single model...



Productive outlook, since we get practical models & data analyses along the way

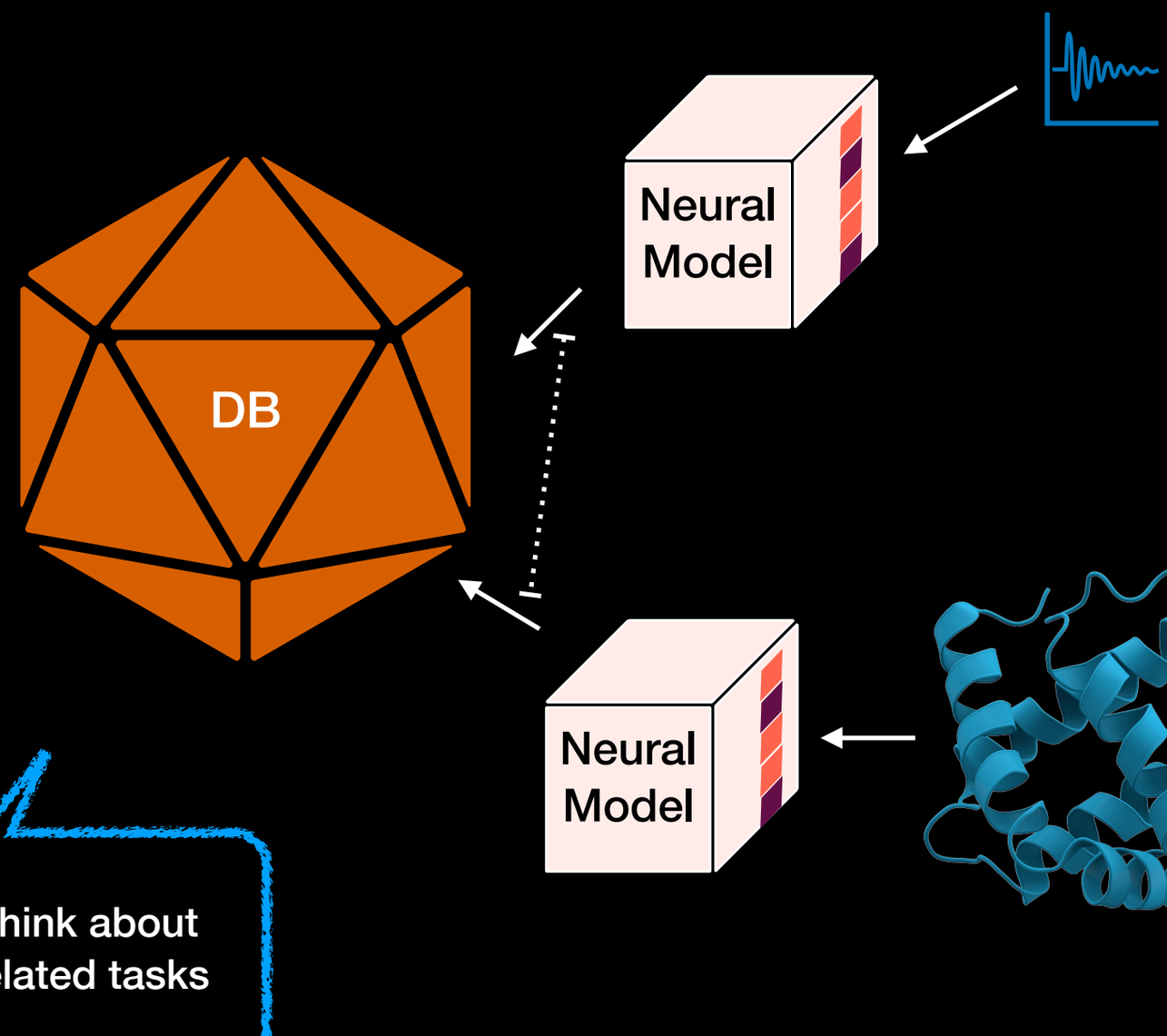
Myoglobin (image from Wikipedia)

Prospective: Interlocking distance constraints across input modalities/tasks with a single model...



In near term, can think about grouping closely related tasks

Prospective: Interlocking distance constraints across input modalities/tasks with a single model...



In near term, can think about grouping closely related tasks

Myoglobin (image from Wikipedia)

https://arxiv.org/abs/2012.02287