Performance History-Based Model Selection for Neonatal EEG Seizure Detection

Selecting models for unlabeled new patients

with the performance history of similar past cases

Data Engineering Project Proposal

Team: Team 9 | Seongbeom Park · Inkyu Lee · Minho Park · Minseong Kim

2026.10.22

Contents

01	Problem Definition	Why new patients need model selection	p. 3–5
02	Objectives	What the system will deliver	p. 6
03	Data	Neonate dataset, patient profile, and split	p. 7–8
04	Method	Detection models, history DB, case search, agent	p. 9–12
05	Architecture	How training, history, and selection connect	p. 13
06	Evaluation Plan	Baselines, metrics, and experiments	p. 14

Problem Definition ①

Which seizure detection model should a new patient get?

01



EEG-only seizures

Often no clear clinical signs, so EEG monitoring is essential

02



Review workload

Continuously reviewing long multichannel EEG burdens experts

03



Varies by patient

Accuracy and false alarms vary with patient traits and EEG state

04



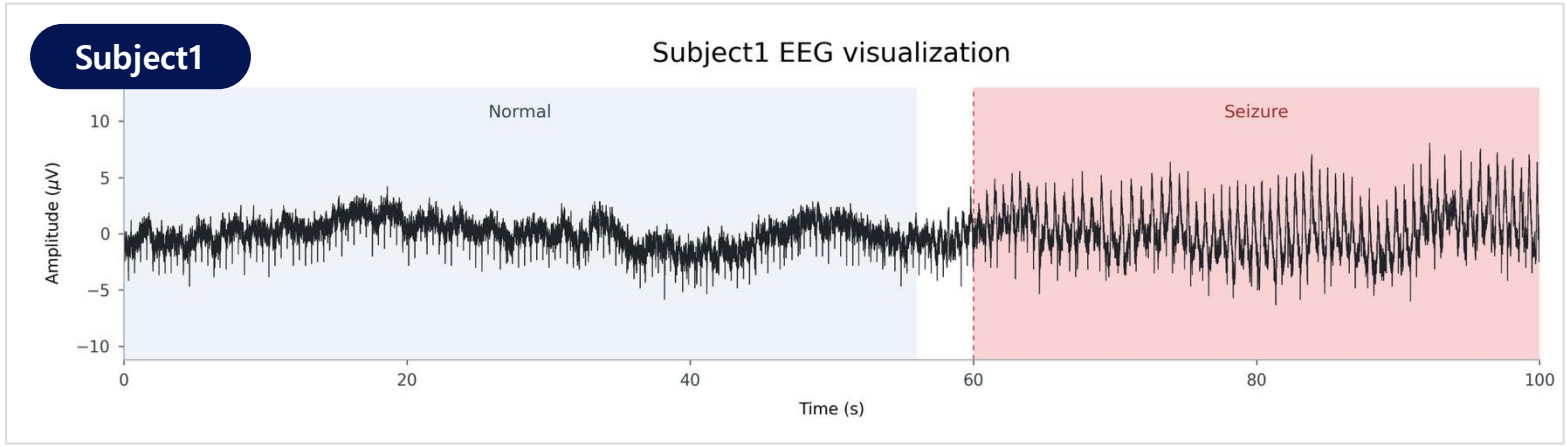
No labels at start

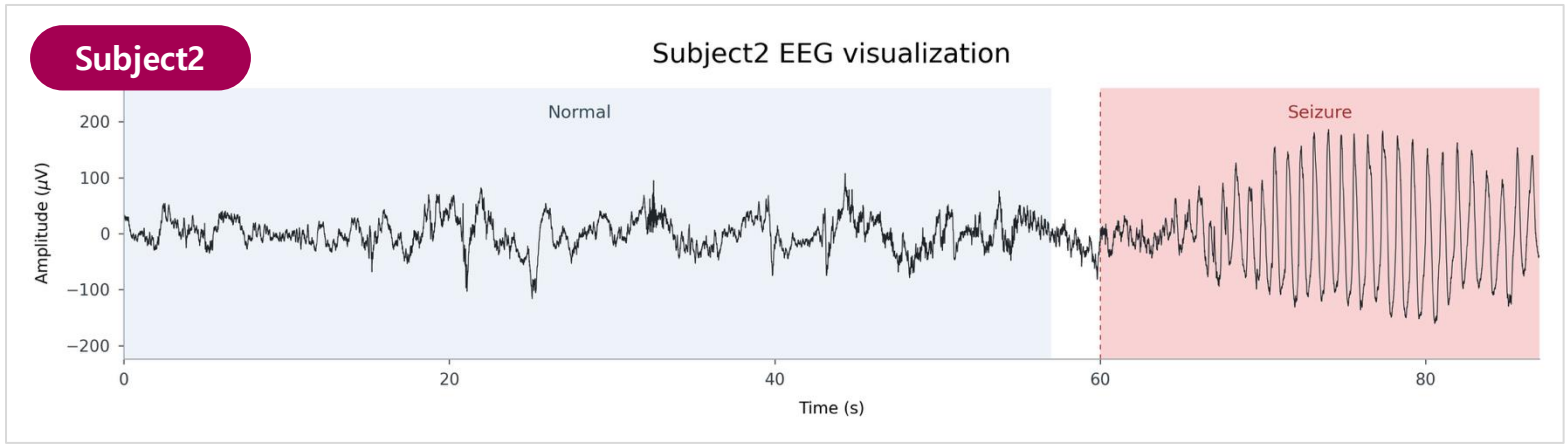
Patient-specific training needs labels, which are unavailable at the start

Question **Without extra labels or training, which existing model fits a new patient?**

Motivating Example

Same seizure label, different EEG waveforms across patients





※ This figure illustrates waveform differences; model performance differences will be tested separately

Observed differences

Amplitude range
Subject1 within ±10 μV
Subject2 about ±200 μV

Non-seizure segment
Subject1 low amp., fast activity
Subject2 high amp., slow waves

Seizure segment
Subject1 repetitive sharp spikes
Subject2 large regular rhythm

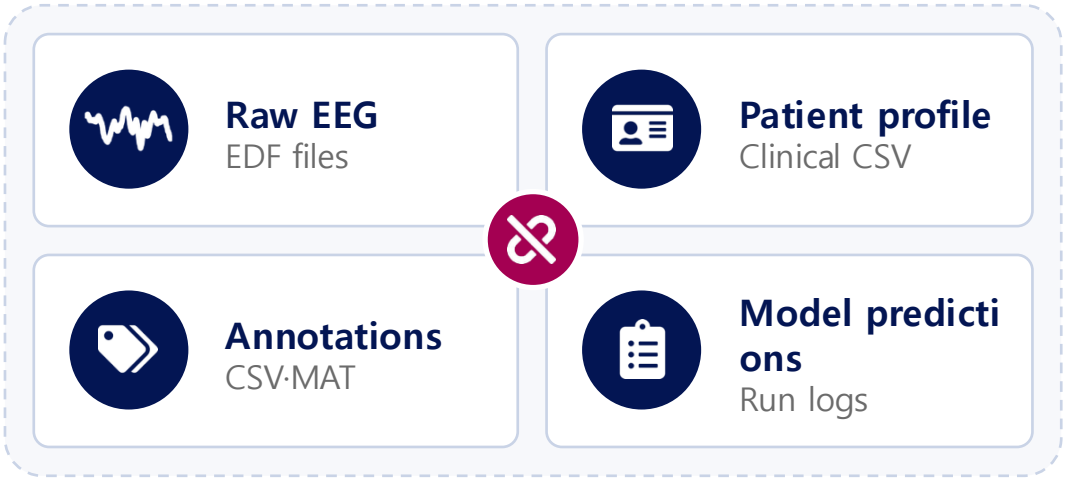
Key message
Waveforms can differ by patient even u nder the same seizure label

Next question
Given these differences, can we select t he right model for the current EEG seg ment?

Problem Definition ②

Past model performance is hard to reuse for new patients

Now **Siloed data, hard to query together**



- Past EEG and model runs show which model worked w ell for which patient and segment
- Overall average performance cannot show how a mod el does on cases similar to a new patient

Approach **A data pipeline from similar-case search to model selection, execution, and logging**

1. **Profile · current EEG segment**
2. **Search similar past cases**
3. **Look up per-model history**
4. **Model selection · execution · logging**

Objectives

Build a performance-history-based model selection system



01

Integration

Link EEG, profiles, and annotations by patient ID and time



02

History DB

Store per-segment predictions and errors of 4 models in a DB



03

Case search

Top-k search with profile and EEG embeddings



04

Agent tools

Call search, selection, and inference tools; log every run



05

Evaluation

Compare accuracy and latency with single models and ensembles

Deliverable A pipeline that selects models without labels and tracks the rationale and results

Data ①

Neonate dataset and patient profile

79
NICU neonates with EEG recorded for suspected seizures

19 ch
EEG channels, standard 10–20 system

256 Hz
Sampling rate

74 min
Median recording length (one continuous session per patient)

3
Experts independently marked seizures at 1-s resolution

EDF·CSV
EEG: EDF
Annotations: CSV·MAT
Clinical info: CSV

Patient profile

Input
Gender · BW · GA · EEG to PMA
Sex · birth weight · gestational age · PMA at EEG (binned values)

Excluded
Number of Reviewers · Primary Localisation · Other
Directly tied to the seizure labels

Source: Stevenson et al., *A dataset of neonatal EEG recordings with seizure annotations*, Scientific Data, 2019

Data ②

Preprocessing and patient-level split

Link source data → Quality checks → Filtering · normalization → Segmentation → Labels · embeddings

Patient-level split all segments of a patient go to the same group

Seizure Detection Model Training	Model Selection Agent Training	Test
40%	40%	20%
Train and validate Model A–D	Measure per-model performance and build the history DB	Model selection and final evaluation on new patients

Example of patient-level assignment



Analysis units

Patient recording → **Segment** unit of selection & labels → **Window** unit of model input

Method ① Seizure Detection Model

Segment-level seizure detection and the candidate model pool

Input Each segment of a patient's EEG recording → **Process** Aggregate window-level predictions per segment → **Output** Seizure probability → Seizure / Non-seizure

Candidate model pool

each model looks at a different aspect of seizures

A Morphology

FCN (Fully Convolutional Net)
IEEE MLSP 2017
Neonatal-EEG CNN for local waveform shapes

B Rhythm

TimesNet
ICLR 2023
Rhythm via a period-based 2D transform

C Spatial

iTransformer
ICLR 2024
Captures spatial spread via inter-channel relations

D Pretrained

CBraMod
ICLR 2025
Pretrained on large-scale EEG data

Hypothesis The performance history of similar past segments improves model selection

Method ② Performance History DB

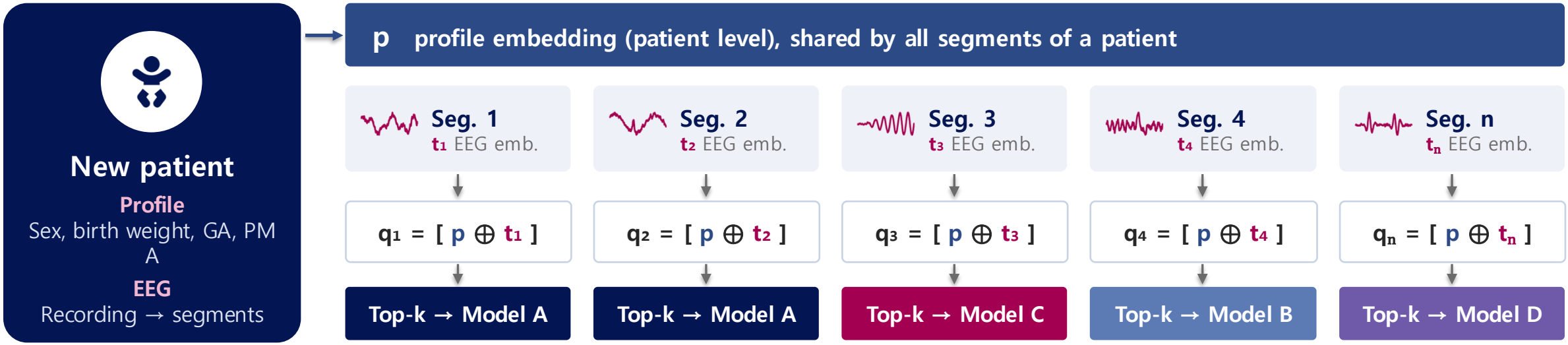
A searchable record of which model worked on which patient segment

Performance history record logical unit: patients × segments

```
{
  "patient_id": "P-07",    # Unique patient ID
  "segment_id": 2,     # Segment index (recording split into 5 parts, 1–5)
  "profile_emb": [0.12, -0.38, 0.77, ..., 0.05], # Patient profile embedding via LLM
  "ts_emb":      [0.64, 0.21, -0.19, ..., -0.42], # EEG segment embedding (TS encoder)
  "query_vector": [ √α·profile_emb ⊕ √(1−α)·ts_emb], # Search vector, weighted by α
  "label": "seizure", # Ground truth (seizure / normal)
  "predictions": {  # Per-model predictions
    "model_A": { "prob": 0.87, "correct": true  }, # Model A: "prob" = P(seizure), "correct" = T/F
    "model_B": { "prob": 0.31, "correct": false },
    "model_C": { "prob": 0.35, "correct": false  },
    "model_D": { "prob": 0.34, "correct": false  }
  },
  "best_model": "A"  # Model that best fit this patient's segment
}
```

Method ③ Similar-Case Search

Find similar cases with patient-level profiles and segment-level waveforms



※ Example: segments with different waveforms can select different models, even within one patient

Similarity

Weighted combination of profile and EEG embeddings, then cosine si milarity

$q = [\ \sqrt{\alpha} \cdot \text{profile_emb} \ \oplus \ \sqrt{(1-\alpha)} \cdot \text{ts_emb}\ ]$

$\text{similarity} = \text{cosine}(q_test, q_DB)$

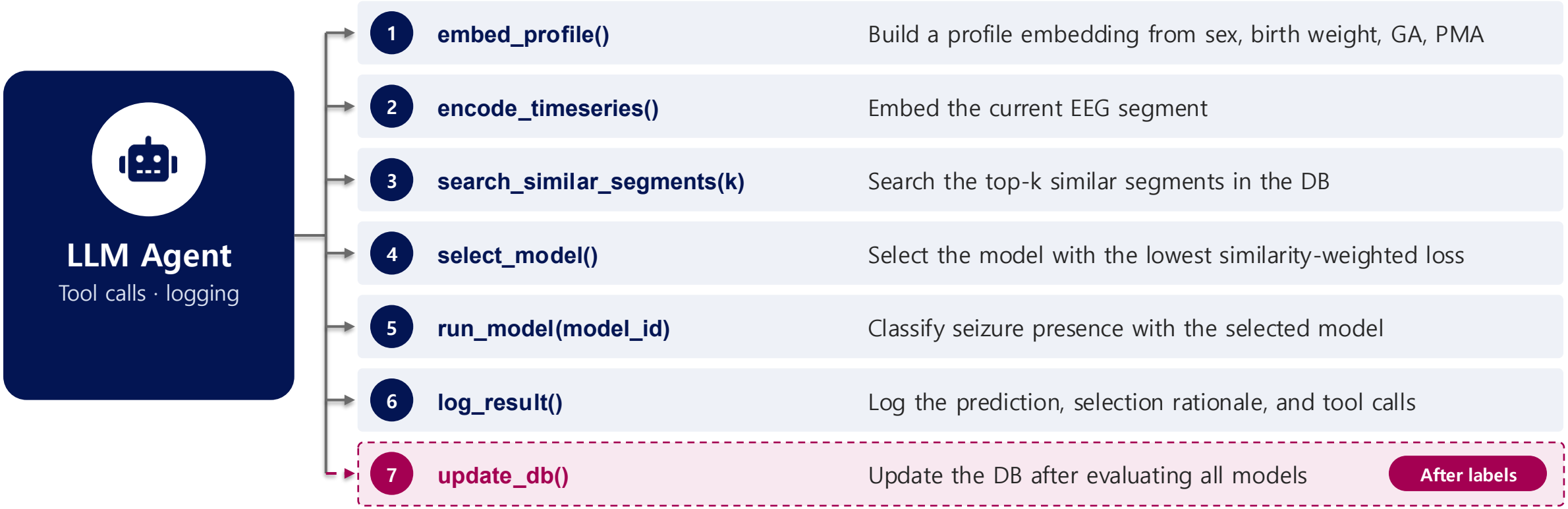
Role of α

α ↑ more profile weight → segments of a patient find similar cases

α ↓ more waveform weight → each segment finds its own cases and model

Method ④ Agent Tool Calling

The LLM Agent calls search, selection, and inference tools and logs each step



Selection policy

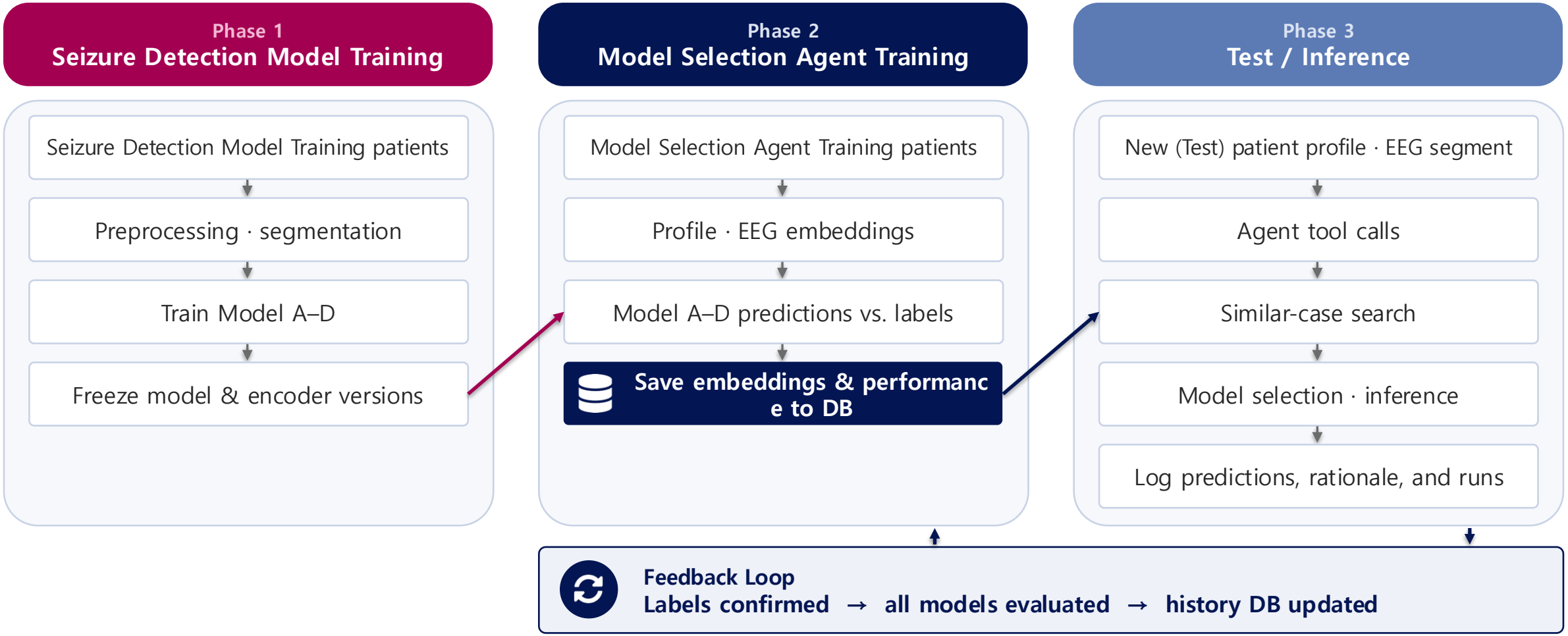
① Select by per-model performance on retrieved cases

② Too few similar cases → use the best model overall

③ Fix search weights and rules before testing

Architecture

Linking training, performance history, and model selection



Evaluation Plan

Agent-based model selection vs. single models and ensembles

Baselines

- **Model A–D alone**
- **Simple ensemble**
- **Agent-based model selection**

Metrics

- **Detection** PR-AUC · F1 · sensitivity
- **Efficiency** search and inference latency

Main

Fix the history DB and evaluate on Test patients

Extended

Sequential: update the DB as labels arrive

→ Earlier segments of the same patient become usable cases

Takeaway **Select models by their record on similar past cases, and log the rationale**

Thank you