Supplementary Methods and Results

S1. Segmentation-assisted pose estimation and validation

To quantify pre-odor movement, we extracted a markerless skeletal representation of each dog from lateral treadmill video using a DeepLabCut-based pose estimation pipeline. To improve robustness under experimental conditions with structured background elements, we implemented a segmentation-assisted refinement procedure and evaluated its effect on pose prediction quality.

Pose estimation pipeline

A total of 29 anatomical landmarks were defined to capture cranial, axial, appendicular, and caudal body regions (Table S1). These included cranial reference points (e.g., tip of nose, top of head, axis), axial landmarks along the spine (e.g., withers, thoracolumbar junction, mid lumbar spine, tuber sacrale), tail landmarks (base and tip of tail), and paired limb landmarks covering forelimb and hindlimb joints (including shoulder, elbow, carpal, fetlock, tarsal, stifle, and toe tips).

The landmark set was designed to provide broad anatomical coverage while capturing movement features relevant to locomotion and odor-related behavior, including head posture, axial alignment, limb kinematics, and tail position.

A total of 120 video frames were manually annotated across dogs and sessions using the Computer Vision Annotation Tool (CVAT). A single DeepLabCut model was trained on this dataset and applied uniformly across all animals. For each frame, the model produced two-dimensional landmark coordinates together with likelihood values reflecting prediction confidence.

Table S1: Wilcoxon U test results for per-keypoint likelihoods comparing full-frame and segmentation-enhanced pose estimation. Statistically significant differences ($p < 0.05$) are marked with an asterisk (*).

ID	Keypoint	U Statistic	p-value	Direction / Significance
1	Tip of nose	5	3×10^{-5}	full < segmented*
2	Top of head	0	1×10^{-5}	full < segmented*
3	Axis (C2 vertebra)	2	1×10^{-5}	full < segmented*
4	Withers	51	0.04529	full < segmented*
5	Thoracolumbar junction	115	0.06197	full > segmented
6	Mid lumbar spine	125	0.02012	full > segmented*
7	Tuber sacrale	74	0.30404	full < segmented
8	Base of tail	72	0.26915	full < segmented
9	Tip of tail	60	0.1092	full < segmented
10	Left forepaw toe tip	20	0.00052	full < segmented*
11	Left fore fetlock joint	28	0.00204	full < segmented*
12	Left fore carpal joint	18	0.00036	full < segmented*
13	Left elbow joint	57	0.08309	full < segmented
14	Left shoulder joint	55	0.06848	full < segmented
15	Right forepaw toe tip	20	0.00052	full < segmented*
16	Right fore fetlock joint	12	0.00011	full < segmented*
17	Right fore carpal joint	50	0.04062	full < segmented*
18	Right elbow joint	80	0.60118	full > segmented
19	Right shoulder joint	87	0.45915	full > segmented
20	Left hindpaw toe tip	3	2×10^{-5}	full < segmented*
21	Left hind fetlock joint	24	0.00105	full < segmented*
22	Left tarsal joint	39	0.01051	full < segmented*
23	Left stifle joint	50	0.04062	full < segmented*
24	Left ischial tuberosity	91	0.37916	full > segmented
25	Right hindpaw toe tip	29	0.0024	full < segmented*
26	Right hind fetlock joint	33	0.00446	full < segmented*
27	Right tarsal joint	50	0.04062	full < segmented*
28	Right stifle joint	38	0.00916	full < segmented*
29	Right ischial tuberosity	156	0.00014	full > segmented*

Segmentation-assisted refinement

Although the experimental setup provided controlled lighting and a fixed camera perspective, the treadmill apparatus and surrounding structures introduced visually salient background elements that could interfere with pose

estimation, particularly for small or partially occluded landmarks.

To reduce this source of noise, a segmentation-assisted refinement step was introduced. For each frame, an initial pose prediction was obtained from the full image. Based on this prediction, a region of interest (ROI) was defined using stable torso landmarks, which exhibit relatively low positional variability across stride phases. This ROI was used to center a cropped view on the dog's body while excluding as much irrelevant background as possible.

Within this ROI, a MediaPipe-based segmentation model was applied to separate the dog from the background. The resulting segmentation mask was used to suppress non-dog pixels, and the masked image was subsequently passed through the pose estimation network a second time to obtain refined landmark predictions.

This procedure was applied sequentially across frames. For each new frame, the ROI was updated based on the landmark configuration from the preceding frame, allowing the method to follow the dog's movement without requiring a separate tracking module. All pose sequences used in downstream analyses were derived from these segmentation-refined predictions.

Evaluation of pose prediction quality

To quantify the effect of segmentation-assisted refinement, we compared landmark predictions obtained from the standard full-frame pipeline with those obtained after segmentation-based preprocessing. Two complementary metrics were evaluated: (i) landmark confidence, measured by DeepLabCut likelihood values, and (ii) spatial localization accuracy for lower confidence predictions, measured as pixel-wise error.

Across all evaluated frames and landmarks, segmentation-assisted refinement increased prediction confidence relative to the full-frame pipeline (Fig. S1), with the difference being statistically significant based on a Wilcoxon signed-rank test across keypoints.

In addition to increasing confidence, segmentation-assisted refinement improved spatial precision for more uncertain predictions. For keypoints with likelihood values at or below 0.6, the mean pixel-wise error decreased from 20.53 pixels in the full-frame pipeline to 7.11 pixels after segmentation-based refinement.

The magnitude of improvement varied across anatomical regions. Per-keypoint analysis (Table S1) showed the largest gains in cranial landmarks, axial landmarks in the cranial trunk region, and distal limb joints. These regions are particularly susceptible to background interference and partial

occlusion in the experimental setup, indicating that segmentation is most beneficial where visual ambiguity is highest.

Summary

Segmentation-assisted refinement improved markerless pose estimation by increasing landmark confidence and reducing localization error, particularly in anatomically and visually challenging regions. As all downstream analyses in this study rely on these pose representations, this preprocessing step improves the signal-to-noise ratio of the behavioral features used for subsequent modeling.

S2. Detailed model architecture and training procedure

Table S2: **Validation metrics at best epoch for each training run.** Bold indicates the best value across all runs per column (excluding epoch). Abbreviations: AUC = area under the ROC curve, Acc = accuracy, Loss = binary cross-entropy loss, Prec = precision, Rec = recall.

Trial	Epoch	AUC	Acc	Loss	Prec	Rec
Video only						
1	47	0.882	0.798	0.427	0.824	0.763
2	16	0.831	0.743	0.511	0.710	0.786
3	17	0.847	0.763	0.493	0.718	**0.840**
Video + physiology						
1	20	0.872	0.769	0.466	0.845	0.665
2	19	0.864	0.790	0.467	0.788	0.795
3	73	**0.953**	**0.870**	**0.290**	**0.899**	0.844

To classify whether a target-present trial would culminate in an indication or a miss, we used a spatiotemporal graph-based neural network that operates on markerless pose sequences extracted from the 4.5 s interval preceding odor onset. In a second version of the model, physiological measurements recorded over the same interval were incorporated through a parallel auxiliary branch and fused with the pose-derived representation prior to classification.

Model input representation

Each pose-based input sample consisted of 135 consecutive video frames sampled at 30 Hz, corresponding to the 4.5 s pre-odor window. For every frame, 29 anatomical landmarks were represented by three features: the two-dimensional landmark coordinates (x, y) and the corresponding pose-estimation likelihood. This yielded an input tensor of shape $(135, 29, 3)$.

To reduce sensitivity to absolute position on the treadmill and within the video frame, all landmark coordinates were translated frame-wise such that the withers landmark was located at the origin. The model therefore operated on relative posture and movement rather than on global body position.

For the multimodal model, pose input was complemented by three physiological inputs extracted from the same pre-odor interval: (i) a 5-element heart-rate vector, (ii) a 7-element vector of inter-beat interval (IBI) summary statistics, and (iii) a single scalar value representing core body temperature. The IBI feature vector comprised mean RR interval, standard deviation, minimum, maximum, root-mean-square of successive differences (RMSSD), coefficient of variation, and mean absolute successive difference. All pose and physiological features were z-scored per dog before model training.

Graph definition and spatiotemporal architecture

The pose sequence was represented as a fixed undirected graph with 29 nodes, one for each anatomical landmark. Edges were defined according to anatomical connectivity, including the axial chain along the head, neck, spine, and tail, the bilateral forelimb and hindlimb chains, and the corresponding limb-to-torso attachments. A self-loop was added to each node so that landmark-specific information was preserved during graph propagation.

Spatial dependencies within each frame were modeled using two stacked Graph Attention Network layers. Each layer used four attention heads, concatenated to produce 64 output features per node, followed by rectified linear unit activation. These graph-attention layers were applied frame-wise using the same graph structure at every time step. The resulting node-wise embeddings were then averaged over joints to obtain one posture embedding per frame.

Temporal dynamics across the 135-frame sequence were modeled with a Temporal Convolutional Network composed of six one-dimensional convolutional layers with dilation factors of 1, 2, 4, 8, 16, and 32, kernel size 2, and 32 filters per layer. This design enabled the network to capture both short-

timescale movement fluctuations and broader temporal patterns across the full pre-odor interval. The temporally processed sequence representation was globally pooled across time and passed through a dense layer with 64 units and rectified linear unit activation, followed by a sigmoid output layer producing the predicted probability of an indication.

Physiological branch and multimodal fusion

To test whether internal physiological state improved prediction beyond movement alone, we extended the pose-based architecture with a parallel physiological branch. Heart rate and IBI features were each processed through two fully connected layers with 32 and 16 units, respectively, using rectified linear unit activation. Core temperature was processed through a separate dense layer with 8 units. The outputs of these three branches were concatenated and passed through an additional dense layer with 16 units to produce a compact physiological embedding.

This physiological embedding was concatenated with the pose-derived spatiotemporal representation immediately before the final classification head. This late-fusion design preserved the pose-processing backbone while allowing behavioral and physiological information to contribute jointly to the final prediction.

Training data generation, augmentation, and class balancing

Model development was restricted to target-present trials recorded at 8 km/h and containing complete pose and physiological data for the full pre-odor interval. For each retained trial, an initial 5 s clip preceding odor onset was extracted. From this clip, the final 4.5 s model input was derived.

For non-augmented samples, the final 135 frames immediately preceding odor onset were used. For augmented samples, a 4.5 s subwindow was selected with a random temporal shift of up to (± 15) frames within the original 5 s clip. This procedure reduced dependence on a fixed alignment within the pre-odor interval while preserving the strictly pre-stimulus character of the input.

Additional spatial augmentation was applied to pose coordinates only. These transformations included isotropic and anisotropic scaling, random rotation within ($\pm 10^{\circ}$), and additive Gaussian noise applied to landmark coordinates. Augmentation was performed only on the development data and never on the held-out test set.

Because the original dataset was imbalanced with fewer miss trials than

indication trials, augmentation and balancing were performed separately within classes. All augmented samples from the minority class were retained. A random subset of augmented samples from the majority class was then added until both classes were matched. The resulting balanced development set comprised 1,716 samples, with 858 samples per class.

Optimization, repeated training, and model selection

Both model variants were trained with the Adam optimizer using a learning rate of 10^{-3} and binary cross-entropy loss. Model performance was monitored using area under the receiver operating characteristic curve (AUC), accuracy, precision, and recall. AUC served as the primary criterion for model selection because it provides a threshold-independent measure of discrimination and is less sensitive to class prevalence than accuracy alone.

Within the development data, samples were randomly partitioned into training and validation subsets at the sample level. Early stopping was based on validation AUC, with training terminated when validation performance no longer improved. The best weights from each run were restored. To reduce the influence of random initialization and convergence instability, each architecture was trained independently three times with different random seeds. Final model selection was based on the run with the highest validation AUC.

A separate held-out test set, containing only original non-augmented samples, was excluded from augmentation, training, and model selection throughout.

Training stability and selected models

Repeated training revealed clear differences in convergence behavior between the pose-only and multimodal models. For the pose-only architecture, the three independent runs reached best validation AUC values of 0.882, 0.831, and 0.847, with the best-performing model converging at epoch 47. The corresponding validation accuracy, precision, and recall were 0.798, 0.824, and 0.763, respectively (Table S2).

Training dynamics of the selected pose-only model showed a progressive divergence between training and validation metrics in later epochs, with continued improvement in training loss and performance accompanied by increased variability in validation curves (Fig. S2). This pattern is consistent with the onset of overfitting and limited generalization stability.

In contrast, the multimodal architecture demonstrated both improved

peak performance and more stable convergence across runs. The three independent runs reached best validation AUC values of 0.872, 0.864, and 0.953, with the selected model converging at epoch 73. The corresponding validation accuracy, precision, and recall were 0.870, 0.899, and 0.844, respectively (Table S2).

Training and validation curves of the multimodal model remained closely aligned throughout training, with reduced fluctuations and no pronounced late-epoch divergence (Fig. S3). This indicates improved generalization behavior and suggests that the inclusion of physiological features stabilizes the learning process.

Taken together, these results show that predictive signal was consistently recoverable across independent runs, while multimodal integration of physiological data improved both peak validation performance and training stability relative to the pose-only baseline.

S3. Bayesian aggregation details

We modeled the underlying miss propensity of a dog as a latent probability $\theta \in [0, 1]$, representing the probability of a missed alert under current conditions. Model outputs for individual trials were treated as probabilistic evidence and aggregated using a Beta–Bernoulli updating scheme.

Posterior updating with soft evidence Let $p_t \in [0, 1]$ denote the predicted probability of a miss for trial t. Starting from a Beta prior

$$\theta \sim \mathrm{Beta}(\alpha_0, \beta_0), \tag{1}$$

we updated the posterior parameters using soft counts:

$$\alpha_t = \alpha_{t-1} + p_t, \quad \beta_t = \beta_{t-1} + (1 - p_t). \tag{2}$$

After T trials, the posterior is

$$\theta \mid \{p_t\}_{t=1}^{T} \sim \mathrm{Beta}(\alpha_T, \beta_T). \tag{3}$$

This formulation allows continuous model outputs to contribute proportionally, rather than requiring binarization.

Posterior summaries The posterior mean miss probability is given by

$$\mathbb{E}[\theta] = \frac{\alpha_T}{\alpha_T + \beta_T}. \tag{4}$$

Uncertainty is captured by the variance

$$\mathrm{Var}(\theta) = \frac{\alpha_T \beta_T}{(\alpha_T + \beta_T)^2(\alpha_T + \beta_T + 1)}. \tag{5}$$

Threshold-exceedance probability To assess operational risk, we computed the probability that the latent miss rate exceeds a predefined threshold θ^*:

$$P(\theta > \theta^*) = 1 - F_{\mathrm{Beta}}(\theta^* \mid \alpha_T, \beta_T), \tag{6}$$

where F_{Beta} is the cumulative distribution function of the Beta distribution.

This quantity provides a direct measure of confidence that the dog is operating above an acceptable miss-rate threshold.

Sequential updating The update was performed sequentially across trials, allowing real-time tracking of the posterior:

$$(\alpha_{t-1}, \beta_{t-1}) \to (\alpha_t, \beta_t). \tag{7}$$

This results in a dynamically evolving estimate of θ , reflecting recent evidence while retaining prior information.

Prior specification and sensitivity We evaluated prior sensitivity using Beta priors of varying concentration:

- weak prior: Beta$(1, 1)$,
- moderate prior: Beta$(2, 2)$,
- informative prior: Beta$(5, 5)$.

The prior mean was fixed at 0.5, while concentration $\alpha_0 + \beta_0$ controlled the influence of prior versus data. Lower concentrations allowed rapid adaptation to new evidence,whereas higher concentrations produced smoother, more conservative updates.

Prior sensitivity was assessed by comparing trajectories of $\mathbb{E}[\theta]$ and $P(\theta > \theta^*)$ across different prior specifications (Fig. SS4).

Implementation note All updates were computed in closed form without numerical approximation. The approach is computationally lightweight and suitable for real-time deployment.

Supplementary Videos

Supplementary Video S1. Segmentation-enhanced pose estimation.

The video shows a detection dog on a treadmill during markerless pose estimation. The first segment shows the raw video. Subsequently, joint predictions are overlaid on the original recording. In the final segment, the background is removed, corresponding to the segmented condition used to improve pose estimation.

Supplementary Video S2. Keypoint saliency prior to odor presentation.

The video shows markerless pose predictions for a detection dog on a treadmill. Each joint is color-coded according to its saliency, defined as its contribution to the model's prediction of a missed alert. Saliency is displayed dynamically over the 4.5,s interval preceding odor presentation, with red indicating high saliency and blue indicating low saliency. The visualization illustrates how the model shifts attention across body regions over time.

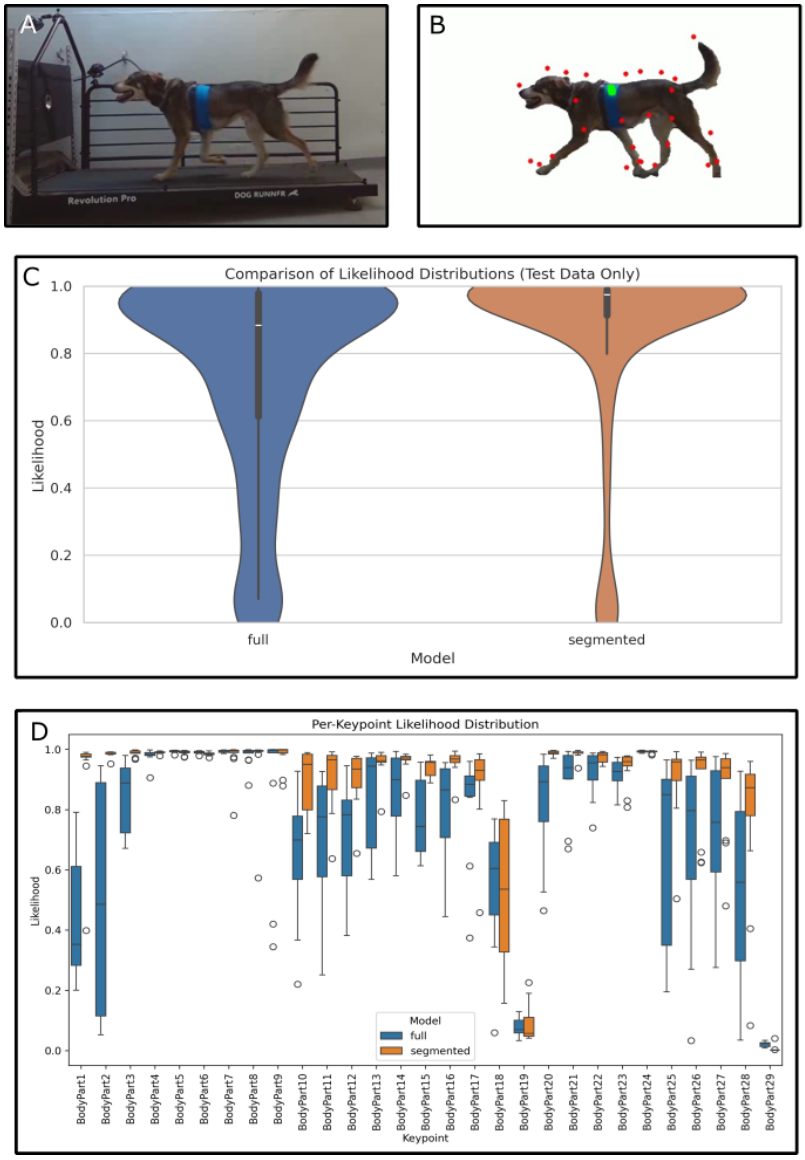

Figure S1: **Segmentation improves pose prediction quality by enhancing signal-to-noise ratio.** (**A**) Sample frame from the olfacto-treadmill setup during trotting trials. (**B**) Segmentation-enhanced pose prediction with joint markers overlaid; background elements are removed. (**C**) Violin plots comparing DeepLabCut likelihood distributions between full and segmented pipelines on the test set ($p < 0.00001$, Wilcoxon U test). (**D**) Per-joint likelihood distributions showing strongest gains in cranial, axial, and distal limb regions (see Table S1).

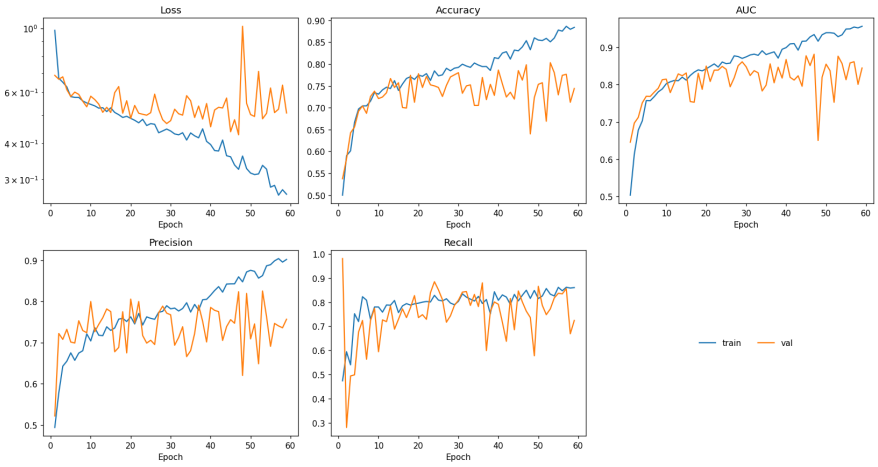

Figure S2: **Full training metrics for video-only model.** Shown are training and validation curves for the ST-GAT + TCN model trained on 2D joint trajectories extracted from video, without physiological inputs. Metrics include binary cross-entropy loss, accuracy, AUC, precision, and recall across 60 training epochs. Each curve corresponds to the best-performing training run based on peak validation AUC. While training performance improved steadily, validation metrics began to fluctuate beyond epoch 45, indicating potential onset of overfitting.

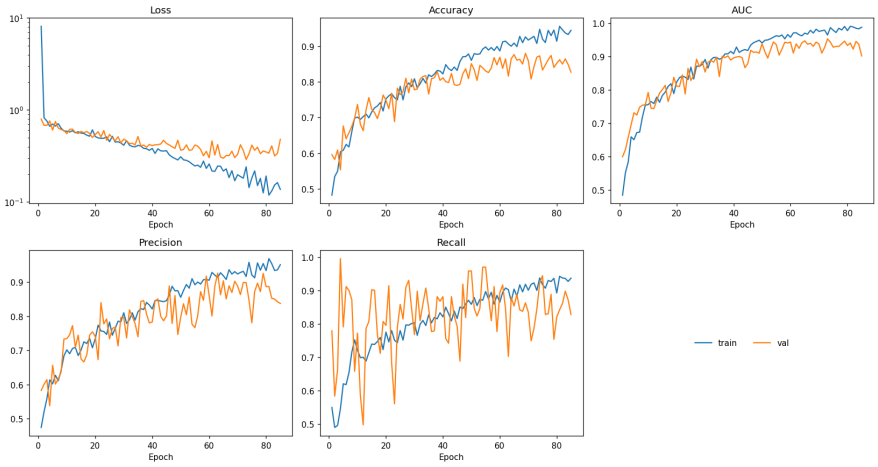

Figure S3: **Full training metrics for multimodal model (video + physiology).** Shown are training and validation curves for the ST-GAT + TCN model trained on 2D joint trajectories combined with physiological inputs (heart rate, inter-beat interval, and core body temperature). Metrics include binary cross-entropy loss, accuracy, AUC, precision, and recall across 90 training epochs. Training and validation curves remained well aligned throughout, indicating stable learning and improved generalization compared to the video-only model.

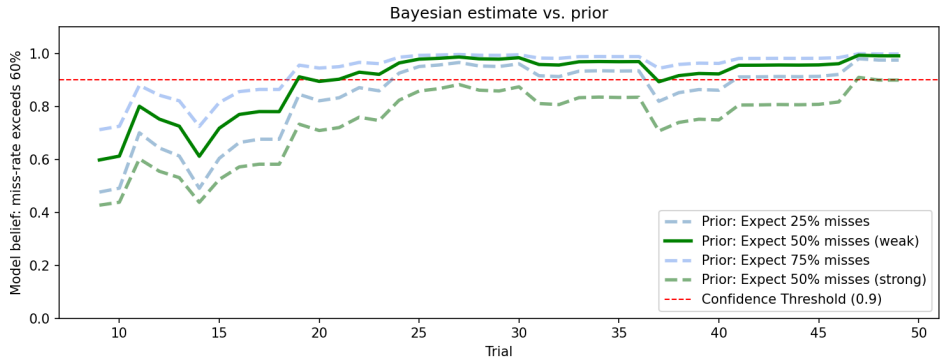

Figure S4: **Prior sensitivity of Bayesian aggregation of trial-level predictions.** Sequential Bayesian updating of the latent miss probability θ across trials, shown for different prior assumptions. Model outputs (trial-level miss probabilities) were treated as soft evidence and integrated using a Beta–Bernoulli framework. The plot shows the posterior probability that the miss rate exceeds a predefined threshold ($P(\theta > \theta^*)$ with $\theta^* = 0.6$) as a function of trial number.
Different priors are compared, including low, moderate, and high prior expectations of miss rate, as well as weak and strong prior concentrations. While early estimates reflect prior assumptions, trajectories converge as evidence accumulates, indicating that the influence of the prior diminishes with increasing data. The dashed horizontal line indicates a decision threshold (0.9) for concluding that the underlying miss rate exceeds the specified level.