# Metacognitive Sensitivity for Test-Time Dynamic Model Selection

**Le Tuan Minh Trinh**
University College London
London, United Kingdom
minh.trinh.23@ucl.ac.uk

**Le Minh Vu Pham**
FPT University
Hanoi, Vietnam
vuplmhe180526@fpt.edu.vn

**Thi Minh Anh Pham**
Queen Mary University of London
London, United Kingdom
t.pham@se22.qmul.ac.uk

**Duc An Nguyen**
University of Oxford
Oxford, United Kingdom
annguyen@robots.ox.ac.uk

## Abstract

A key aspect of human cognition is metacognition - the ability to assess one's own knowledge and judgment reliability. While deep learning models can express confidence in their predictions, they often suffer from poor calibration, a cognitive bias where expressed confidence does not reflect true competence. Do models truly know what they know? Drawing from human cognitive science, we propose a new framework for evaluating and leveraging AI metacognition. We introduce *meta-d'*, a psychologically-grounded measure of metacognitive sensitivity, to characterise how reliably a model's confidence predicts its own accuracy. We then use this dynamic sensitivity score as context for a bandit-based arbiter that performs test-time model selection, learning which of several expert models to trust for a given task. Our experiments across multiple datasets and deep learning model combinations (including CNNs and VLMs) demonstrate that this metacognitive approach improves joint-inference accuracy over constituent models. This work provides a novel behavioural account of AI models, recasting ensemble selection as a problem of evaluating both short-term signals (confidence prediction scores) and medium-term traits (metacognitive sensitivity).

## 1 Introduction

Deep learning research is characterised by increasing specialisation. Convolutional Neural Networks (CNNs) are widely used for perceptual tasks such as image recognition; Transformers and Large Language Models (LLMs) dominate natural language processing; and Vision-Language Models (VLMs) integrate across modalities Hu & Frank (2024); Gadetsky et al. (2025); McCoy et al. (2024); Steyvers et al. (2025). This specialisation improves performance on benchmark tasks but also reflects the "No Free Lunch" theorem: no single architecture is optimal across all problems Bigelow et al. (2023). As a result, systems rely on multiple specialised models, raising the arbitration problem - determining which model should be used for a given input. This is the focus of dynamic model selection, which develops methods for assigning tasks to the most appropriate model.

A central obstacle in this area is the unreliability of model confidence. Deep neural networks produce probabilistic outputs that are often miscalibrated, meaning confidence does not align with true prediction accuracy. Research has shown that this miscalibration arises from the same architectural and training practices that enable high accuracy Guo et al. (2017); Geirhos et al. (2018); Steyvers & Peters (2025); Song et al. (2025a).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

Cognitive science has long studied the analogous human ability, termed *metacognition*, providing theoretical and mathematical tools for assessing how well an agent can evaluate its own knowledge Rouault et al. (2019); Fleming (2021); Lee et al. (2025). Concepts such as metacognitive sensitivity are increasingly viewed as necessary for developing AI systems that can self-monitor, detect errors, and adjust behaviour based on uncertainty Yedetore et al. (2023); Kurvers et al. (2023); Kuribayashi et al. (2025); Nguyen et al. (2025). Recent work has applied the *meta-d'* framework to AI, introducing "AI metacognitive sensitivity" as a performance dimension. Findings suggest that in human-AI collaboration, an AI with lower accuracy but higher metacognitive sensitivity can be a more effective partner, as its confidence signals are more reliable for guiding when its advice should be followed.

While the literature has started to recognise the importance of AI metacognition as a diagnostic tool Gandhi et al. (2024); Ivanova et al. (2024); Gandhi et al. (2025); Song et al. (2025c); Murthy et al. (2025); Song et al. (2025b); Hu et al. (2025), the work reviewed here represents a critical next step: operationalising this concept as a functional component within an adaptive algorithm for a concrete engineering problem like dynamic model selection at test-time. It moves the idea of AI metacognition from a desirable property to be measured to a dynamic signal to be actively leveraged.

In this work, we embed metacognitive sensitivity into a bandit-based framework for dynamic model selection, demonstrating accuracy gains compared to individual constituent models. Beyond these improvements, our results suggest that leveraging metacognition as a functional signal can enable ensemble systems that adapt more flexibly and reliably.

## 2 Related Work

Given a pool of expert models, a system must decide how to aggregate their outputs to arrive at a final prediction. The paradigms for combining and selecting models are summarised as follows:

**Static Ensembles:** combine predictions from all models using fixed rules, such as majority voting for classification or averaging for regression Ganaie et al. (2022). Static ensembles reduce variance but are non-adaptive, treating all models as equally competent and using all of them for every input, which can be computationally inefficient when models have localised competence Ju et al. (2018); Su et al. (2021); Jiang et al. (2024).

**Dynamic Ensembles Selection:** selects a model or subset of models for each test instance, based on the idea that different classifiers perform best in different regions of the feature space Cruz et al. (2015); Liu et al. (2023); Piwko et al. (2025); Vasheghani & Sharifi (2025).

**Mixture of Experts (MoE):** consists of expert networks and a gating network Chen et al. (2022). The gating network assigns weights to experts for each input, and the final prediction is a weighted sum of expert outputs. Unlike Dynamic Selection, which uses pre-trained models with an external selection mechanism, MoE is trained end-to-end, with experts and gating network co-adapting during training Shazeer et al. (2017); Li et al. (2024, 2025); Mu & Lin (2025).

The shift from static ensembles to dynamic selection, and MoE methods underscores the need for adaptive mechanisms. Existing methods rely on fixed heuristics like local accuracy rather than learned policies. Reinforcement learning, particularly contextual bandits, offers a framework for adaptive selection by balancing exploration and exploitation across models. However, current approaches lack metacognition-aware feedback. They either depend on miscalibrated raw confidence or on noisy, static heuristics such as local accuracy. None leverage measures of metacognitive sensitivity to guide adaptive selection.

## 3 Problem Formulation

The central problem is to perform **test-time dynamic model selection** for an image classification task. Given a pair of pre-trained models, $M = \{M_A, M_B\}$, and a sequence of images, $D = \{x_1, \dots, x_N\}$, the objective is to create a framework that, for each image $x_t$, selects the model whose prediction is most likely to be correct. Since models often vary in performance across domains, particularly on out-of-distribution data, a static choice is suboptimal. We propose a dynamic selection agent that chooses between models per sample, informed by both immediate confidence and a cognitively inspired, dynamically updated measure of **metacognitive sensitivity** Fleming & Daw (2017).

The goal is to learn a selection policy $\pi$ that maximises the cumulative reward, which is equivalent to maximising the total classification accuracy of the framework over the dataset:

$$\max_{\pi} \sum_{t=1}^{N} R_t = \max_{\pi} \sum_{t=1}^{N} \mathbb{I}(\hat{y}_{a_t,t} = y_t) \tag{1}$$

where $a_t = \pi(s_t)$ is the model selected at time $t$ based on context $s_t$, $\hat{y}_{a_t,t}$ is its prediction, and $y_t$ is the ground truth label.

## 4 Dynamic Model Selection Framework

**Algorithm 1** Test-Time Dynamic Model Selection

```
Require: Model set $M = \{M_A, M_B\}$, Dataset $D = \{x_1, \ldots, x_N\}$, Burn-in size $B = 100$,
  Window size $W = 100$, Update frequency $F = 50$.
Initialise: Contextual Bandit (LinUCB or LinTS).
Initialise: Total Reward $R_{total} \leftarrow 0$.
                                                        ▷ Burn-in Phase
Collect performance data (confidence, reward) $H_k = \{(c_{k,i}, r_{k,i})\}_{i=1}^{B}$ for each model $M_k \in M$
  on the first $B$ trials.
for each model $M_k \in M$ do
    $\mu_{k,B} \leftarrow$ meta-d'$(H_k)$               ▷ Compute initial metacognitive score
end for
                                                        ▷ Dynamic Selection Phase
for $t = B + 1$ to $N$ do
    for each model $M_k \in M$ do                       ▷ Context Formulation
        Get confidence $c_{k,t}$ on image $x_t$.
        if $(t - B) \pmod{F} = 1$ then
            Collect past performance $H_k \leftarrow \{(c_{k,i}, r_{k,i})\}_{i=t-W}^{t-1}$.
            Update metacognitive score: $\mu_{k,t} \leftarrow$ meta-d'$(H_k)$.
        else
            Keep previous score: $\mu_{k,t} \leftarrow \mu_{k,t-1}$.
        end if
    end for
    Construct context vector
        $s_t = [c_{A,t}, \mu_{A,t}, c_{B,t}, \mu_{B,t}]$        (2)
                                                        ▷ Bandit Action & Reward
    Select model $a_t$ using bandit policy on $s_t$.
    $a_t \leftarrow \arg\max_{k \in \{A,B\}} \pi_t(s_t, k)$   ▷ $\pi_t(s_t, k)$: a specific bandit selection in Appendix
    Get prediction $\hat{y}_{a_t,t}$ from model $M_{a_t}$.
    Observe true label $y_t$ and calculate reward $R_t = \mathbb{I}(\hat{y}_{a_t,t} = y_t)$.
    $R_{total} \leftarrow R_{total} + R_t$.
    Update bandit policy with $(s_t, a_t, R_t)$.          ▷ Learning
end for
return Overall Accuracy: $R_{total}/(N - B)$.
```

We propose a framework in which a contextual bandit serves as the dynamic selection agent. At each time step $t$, the bandit observes a context vector from two candidate models, selects one to act (the "arm"), receives a reward based on prediction correctness, and updates its policy. For each input $x_t$, models $M_A$ and $M_B$ produce predictions and confidence scores, which are combined into a 4-dimensional context vector as shown in Equation. 2.

The **short-term signal** ($c_{k,t}$) is the model's raw *confidence* on the current image $x_t$, taken as the maximum value of its softmax output probability; and the **metacognitive sensitivity** ($\mu_{k,t}$) is the score representing the model's recent historical ability to know when it knows. This is a more stable, medium-term trait, which was operationalised using the **meta-d'** values.

Sensitivity scores are initialised using the first 100 trials and updated every 50 trials via a 100-trial sliding window. To address the computational demands of hierarchical Bayesian inference, we developed a GPU-parallelised package for efficient *meta-d'* estimation. For learning and decision-making, we employ the contextual bandit algorithms (Further bandit algorithmic details are provided in the Appendix).

## 5 Results

We evaluate our framework using four diverse pre-trained models: AlexNet, GoogleNet (classical CNN), EfficientNet (efficiently scaled CNN), and Vision Transformer (ViT) (transformer-based model). This selection captures both convolutional and attention-based models, enabling the assessment of the adaptability of our framework across fundamentally different design choices. We evaluate accuracy at three checkpoints corresponding to the number of trials indicates how many images have been processed up to that point.

Table 1: Accuracy of the best individual image models and the joint framework on the CIFAR10.

| Model Pair | 300 trials | | 700 trials | | 1000 trials | |
|---|---|---|---|---|---|---|
| | Model | Comb. | Model | Comb. | Model | Comb. |
| AlexNet-ViT | 62.4 | 69.5 (+7.1%) | 64.8 | 66.2 (+1.4%) | 62.4 | 65.9 (+3.5%) |
| AlexNet-GoogleNet | 62.7 | 70.6 (+7.9%) | 57.7 | 57.5 (-0.2%) | 56.8 | 58.4 (+1.6%) |
| EfficientNet-ViT | 67.7 | 75.9 (+8.2%) | 66.4 | 68.0 (+1.6%) | 66.4 | 67.8 (+1.4%) |
| EfficientNet-GoogleNet | 54.8 | 59.0 (+4.8%) | 53.6 | 55.8 (+2.2%) | 54.8 | 57.3 (+2.5%) |

Table 1 shows that the joint framework's accuracy is significantly improved in the early trials and then stabilises at 1.4%–3.5% higher than that of the individual models. At certain points, however, the joint framework underperforms relative to the best individual model in the pair. This effect can be attributed to correlated errors: performance is limited when both models misclassify the same sample, a pattern more common among architecturally similar models that share inductive biases from pre-training. For example, in one pairing (Figure. 1), the metacognitive sensitivity of AlexNet dropped (at trial 700), but the bandit quickly adapted by shifting selection toward GoogleNet. (Learning dynamics graphs are detailed in the Appendix A.)

Despite these challenges, architectural diversity enhances complementarity. Heterogeneous pairings, such as CNN-Transformer models, exhibit fewer correlated errors and achieve higher accuracy. This highlights the importance of inductive-bias diversity for effective dynamic model selection.

Table 2: Performance of the best individual VLM and the joint framework on the CIFAR10 - PACS.

| Model Pair | 1500 trials | | 2500 trials | | 4000 trials | |
|---|---|---|---|---|---|---|
| | Model | Comb. | Model | Comb. | Model | Comb. |
| MetaCLIP-SigLIP | 98.7 | 99.0 (+0.3%) | 98.7 | 98.6 (0.0%) | 98.4 | 98.5 (+0.1%) |
| CLIP-ALIGN | 94.2 | 96.0 (+1.8%) | 94.8 | 96.2 (+1.6%) | 94.8 | 95.8 (+1.0%) |

To further probe the framework's robustness, we evaluated it under domain shift using Vision-Language Models (CLIP, ALIGN, SigLIP and MetaCLIP). To process the OOD bias in VLMs evaluation, we augment PACS Xu et al. (2019) with CIFAR-10. PACS exhibits a 1:5 imbalance between photographic and non-photographic styles, which may skew the results. Adding CIFAR-10 introduces additional photographic diversity, yielding a more balanced and robust combined dataset for benchmarking. Table 2 shows that while individual model performance declines under more demanding conditions, the framework leverages complementary strengths across models. This results in an initial accuracy improvement of 0.3%–1.8%, though because VLMs are already so accurate on their own, the gains are noticeable but modest compared to what had been seen with image models. However these potential suggest that metacognition-aware selection holds promise for addressing challenging, test-time distributional shifts in prediction tasks.

## 6 Conclusion

In this work, we presented a metacognition-inspired framework for dynamic model selection, introducing *meta-d'* as a measure of how reliably a model's confidence predicts its accuracy. Our experiments show that embedding metacognitive sensitivity in a bandit-based framework for dynamic model selection yields accuracy gains compared to individual constituent models. Future directions include extending the framework to large language model ensembles and exploring richer reinforcement learning strategies for selection. Overall, this work provides a first step toward incorporating metacognition-inspired feedback into adaptive model selection, with the goal of creating more reliable and interpretable ensemble systems.

## References

Eric J Bigelow, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, and Tomer D Ullman. In-context learning dynamics with random binary sequences. *arXiv preprint arXiv:2310.17639*, 2023.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35: 23049–23062, 2022.

Rafael MO Cruz, Robert Sabourin, and George DC Cavalcanti. Meta-des. h: A dynamic ensemble selection technique using meta-learning and a dynamic weighting approach. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2015.

Stephen M Fleming. Hmeta-d: hierarchical bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of consciousness*, 2017(1):nix007, 2017.

Stephen M Fleming. *Know thyself: The new science of self-awareness*. Hachette UK, 2021.

Stephen M. Fleming and Nathaniel D. Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91–114, 2017.

Stephen M Fleming and Hakwan C Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014.

Artyom Gadetsky, Andrei Atanov, Yulun Jiang, Zhitong Gao, Ghazal Hosseini Mighan, Amir Zamir, and Maria Brbic. Large (vision) language models are unsupervised in-context learners. In *International Conference on Learning Representations*, 2025.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C Ong, and Noah D Goodman. Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*, 2024.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jennifer Hu and Michael C Frank. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*, 2024.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating Theory of Mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380 (1932):20230499, August 2025. doi: 10.1098/rstb.2023.0499. URL `https://doi.org/10.1098/rstb.2023.0499`. Publisher: Royal Society.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models, 2024. URL `https://arxiv.org/abs/2405.09605`.

Kai Jiang, Zheli Xiong, Qichong Yang, Jianpeng Chen, and Gang Chen. An interpretable ensemble method for deep representation learning. *Engineering Reports*, 6(3):e12725, 2024.

Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of applied statistics*, 45 (15):2800–2818, 2018.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*, 2025.

Ralf HJM Kurvers, Andrea Giovanni Nuzzolese, Alessandro Russo, Gioele Barabucci, Stefan M Herzog, and Vito Trianni. Automating hybrid collective intelligence in open-ended medical diagnostics. *Proceedings of the National Academy of Sciences*, 120(34):e2221473120, 2023.

Doyeon Lee, Joseph Pruitt, Tianyu Zhou, Jing Du, and Brian Odegaard. Metacognitive sensitivity: The key to calibrating trust and optimal decision making with ai. *PNAS nexus*, 4(5):pgaf133, 2025.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:131224–131246, 2024.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Lihua Liu, Jibing Wu, Xuan Li, and Hongbin Huang. Dynamic ensemble selection with reinforcement learning. In *International Conference on Intelligent Computing*, pp. 629–640. Springer, 2023.

Neil A Macmillan. Signal detection theory. *Stevens' handbook of experimental psychology: Methodology in experimental psychology*, 3:43–90, 2002.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

Sonia K. Murthy, Rosie Zhao, Jennifer Hu, Sham Kakade, Markus Wulfmeier, Peng Qian, and Tomer Ullman. Inside you are many wolves: Using cognitive models to interpret value trade-offs in llms, 2025. URL `https://arxiv.org/abs/2506.20666`.

Duc-An Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, and Nick Hawes. Joint decision-making in robot teleoperation: When are two heads better than one? *arXiv preprint arXiv:2503.15510*, 2025.

Jakub Piwko, Jędrzej Ruciński, Dawid Płudowski, Antoni Zajko, Patryzja Żak, Mateusz Zacharecki, Anna Kozak, and Katarzyna Woźnica. Divide, specialize, and route: A new approach to efficient ensemble learning. *arXiv preprint arXiv:2506.20814*, 2025.

Marion Rouault, Peter Dayan, and Stephen M Fleming. Forming global estimates of self-performance from local confidence. *Nature communications*, 10(1):1141, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*, 2025a.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Proceedings of the Conference on Language Modeling*, 2025b. URL `https://arxiv.org/abs/2503.07513`.

Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. Privileged self-access matters for introspection in ai, 2025c. URL `https://arxiv.org/abs/2508.14802`.

Mark Steyvers and Megan AK Peters. Metacognition and uncertainty communication in humans and large language models. *arXiv preprint arXiv:2504.14045*, 2025.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Kaixiang Su, Jiao Wu, Dongxiao Gu, Shanlin Yang, Shuyuan Deng, and Aida K Khakimova. An adaptive deep ensemble learning method for dynamic evolving diagnostic task scenarios. *Diagnostics*, 11(12):2288, 2021.

Sanaz Vasheghani and Shayan Sharifi. Dynamic ensemble learning for robust image classification: A model-specific selection strategy. *Available at SSRN 5215134*, 2025.

Jiaolong Xu, Liang Xiao, and Antonio López. Self-supervised domain adaptation for computer vision tasks, 07 2019.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*, 2023.

# A Appendix

## A.1 Metacognitive Sensitivity (*meta-d’*)

To implement the metacognitive sensitivity score $\mu_k$, we adopt the ‘meta-d’‘ framework developed by Fleming & Lau and detailed in Fleming & Daw (2017). Metacognition refers to the ability to monitor one’s own cognitive processes, and in this context, how well a model’s confidence relates to its actual success. *Meta-d’* is a metric derived from **Signal Detection Theory (SDT)** by Macmillan that quantifies this ability. Its primary advantage is that it provides a measure of metacognitive sensitivity that is independent of the model’s task performance and overall confidence bias. We calculates *meta-d’* by fitting a hierarchical Bayesian model to the observed distributions of confidence ratings for correct and incorrect trials Fleming (2017). For our framework, the score $\mu_{k,t}$ for a model $M_k$ at time $t$ is its calculated meta-d’ value.

## A.2 Dynamic Update Mechanism

A key feature of our framework is that the metacognitive sensitivity score, $\mu_k$, is not static but dynamically updated to reflect recent performance.

1. **Initialisation**: A burn-in set of the first 100 trials is used to compute the initial scores, $\mu_{A,0}$ and $\mu_{B,0}$.
2. **Sliding Window Update**: During the selection process, the *meta-d’* score for each model is recalculated every 50 trials using a **sliding window** of the 100 most recent trials.

This updating mechanism enables the framework to adapt to non-stationarity in model performance, such as when the underlying data distribution shifts.

## A.3 Contextual Bandit Selection

The learning and decision-making core of our framework is a contextual bandit. We adopt two well-established algorithms: Algorithm. 2: **LinUCB** (Linear Upper Confidence Bound) and Algorithm. 3: **LinTS** (Linear Thompson Sampling). The bandit’s interaction at each step $t$ is as follows:

1. **Observe Context**: Receives the context vector $s_t$.
2. **Select Action**: The algorithm’s policy selects an action $a_t \in \{A, B\}$.
3. **Receive Reward**: The framework executes the chosen model $M_{a_t}$. The reward $R_t$ is 1 if the model’s prediction is correct, and 0 otherwise.
4. **Update Policy**: The bandit uses the tuple $(s_t, a_t, R_t)$ to update its internal model, improving its policy for future decisions.

**Algorithm 2** Linear Contextual Upper Confidence Bound

```
Require: Number of arms $K$, observe context $s_t$ with dimension $d$, exploration parameter $\alpha$
Initialise $A_a \leftarrow I_d$ and $b_a \leftarrow \mathbf{0}_d$ for all $k \in \{A, B\}$
for each round $t = 1, 2, \dots$ do
    Observe context vector $s_t \in \mathbb{R}^d$
    for each arm $k \in \{A, B\}$ do
        Compute $\hat{\theta}_k \leftarrow A_k^{-1} b_k$
        Compute $\pi_t(s_t, k) \leftarrow \hat{\theta}_k^\top s_t + \alpha\sqrt{s_t^\top A_k^{-1} s_t}$
    end for
    Choose arm $k_t \leftarrow \arg\max_{k \in \{A,B\}} \pi_t(s_t, k)$
    Observe reward $r_t$
    Update: $A_{k_t} \leftarrow A_{k_t} + s_t s_t^\top, \quad b_{k_t} \leftarrow b_{k_t} + r_t s_t$
end for
```

**Algorithm 3** Linear Contextual Thompson Sampling

```
Require: Number of arms $K$, observe context $s_t$ with dimension $d$, prior variance parameter $\sigma$
Initialise $A_a \leftarrow I_d$, $b_a \leftarrow \mathbf{0}_d$ for all $k \in \{A, B\}$
for each round $t = 1, 2, \dots$ do
    Observe context vector $s_t \in \mathbb{R}^d$
    for each arm $k \in \{A, B\}$ do
        Compute $A_k^{-1} \leftarrow (A_k + \epsilon I_d)^{-1}$
        Compute posterior mean: $\hat{\theta}_k \leftarrow A_k^{-1} b_k$
        Sample $\tilde{\theta}_k \sim \mathcal{N}(\hat{\theta}_k, \sigma^2 A_k^{-1})$
        Compute sampled reward: $\pi_t(s_t, k) \leftarrow \tilde{\theta}_k s_t^\top$
    end for
    Choose arm $k_t \leftarrow \arg\max_{k \in \{A,B\}} \pi_t(s_t, k) \rightarrow$ Observe reward $r_t$
    Update: $A_{k_t} \leftarrow A_{k_t} + s_t s_t^\top$,   $b_{k_t} \leftarrow b_{k_t} + r_t s_t$
end for
```

### A.4 Learning Dynamics Graphs

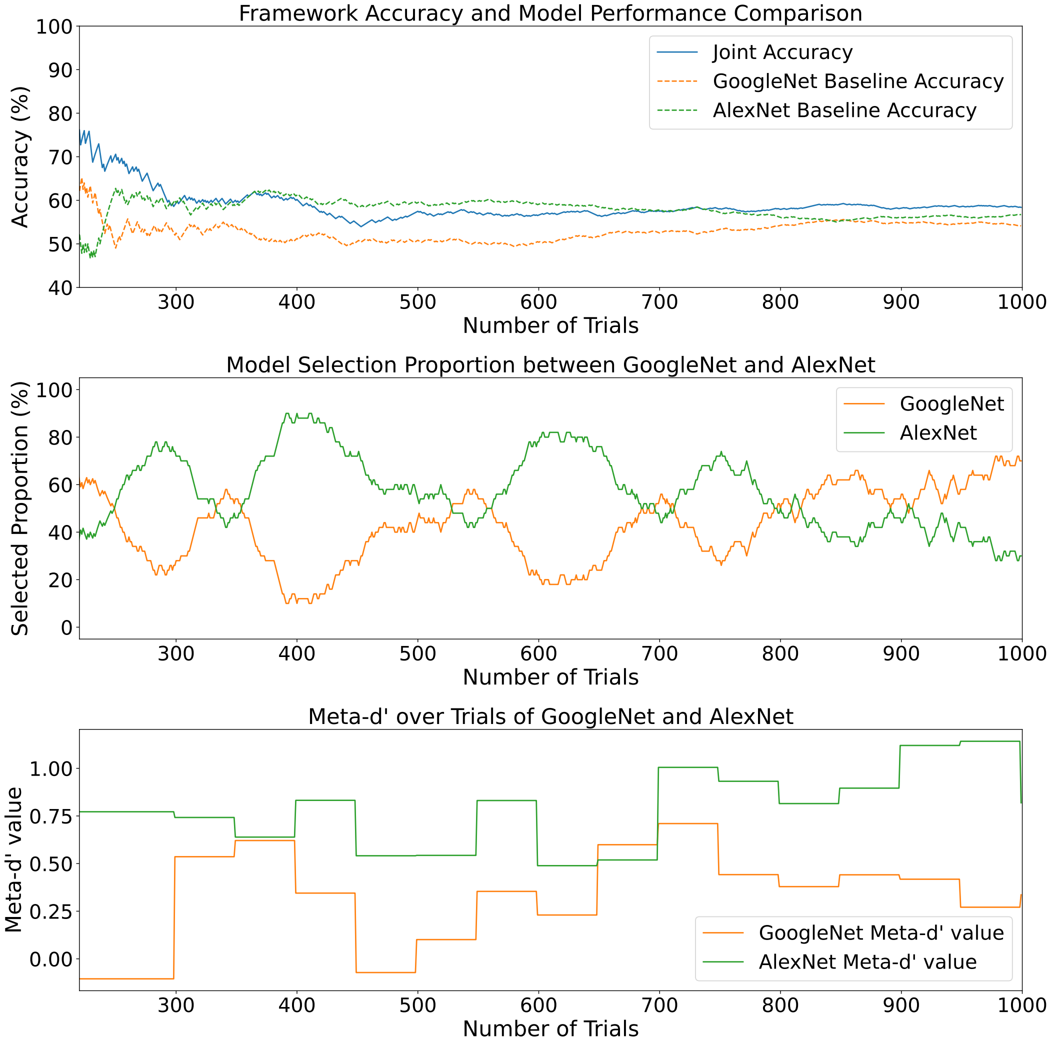

Figure 1: Figure of GoogleNet and AlexNet with Framework Accuracy (using LinTS with $\sigma = 0.5$)

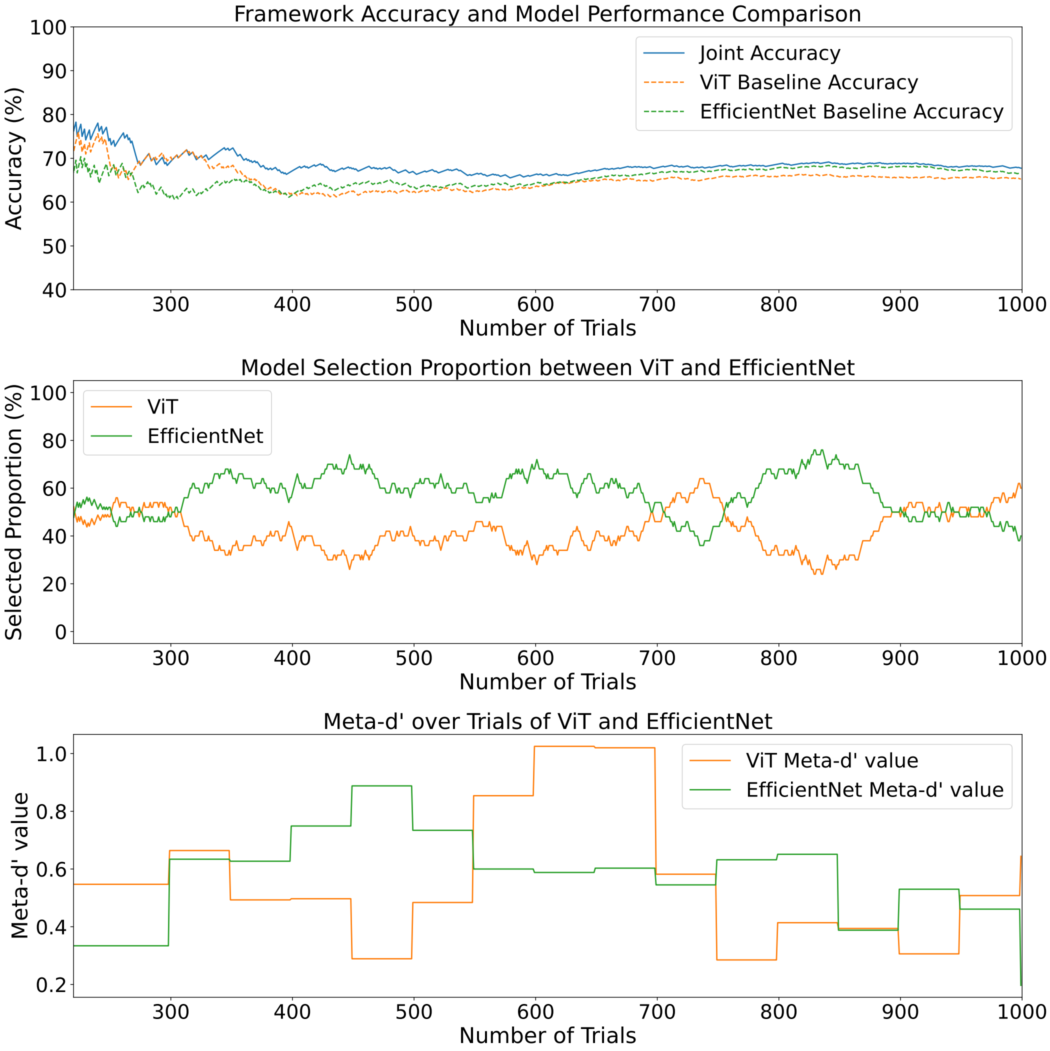

Figure 2: Figure of ViT and EfficientNet with Framework Accuracy (using LinTS with $\sigma$ = 1.0)

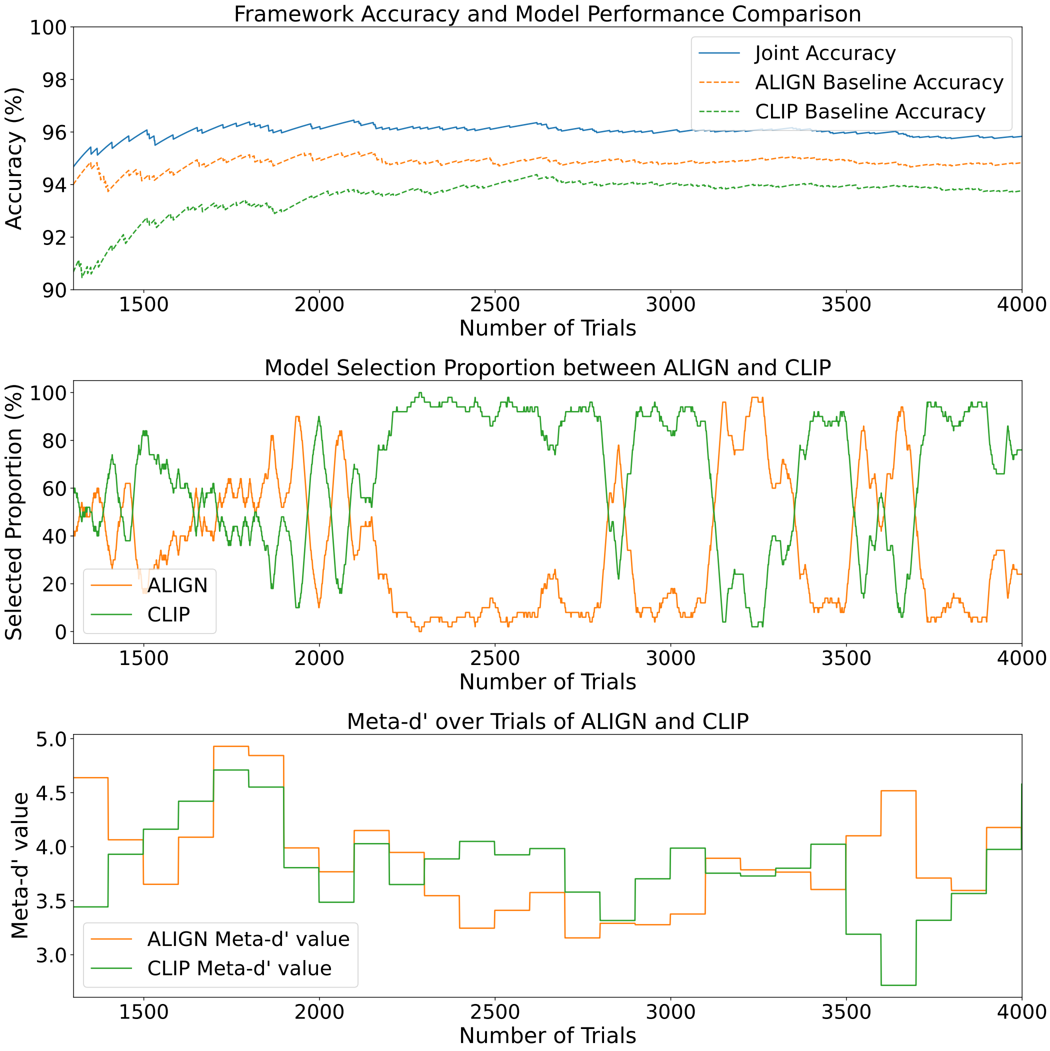

Figure 3: Figure of ALIGN and CLIP with Framework Accuracy (using LinUCB with $\alpha = 1.0$)

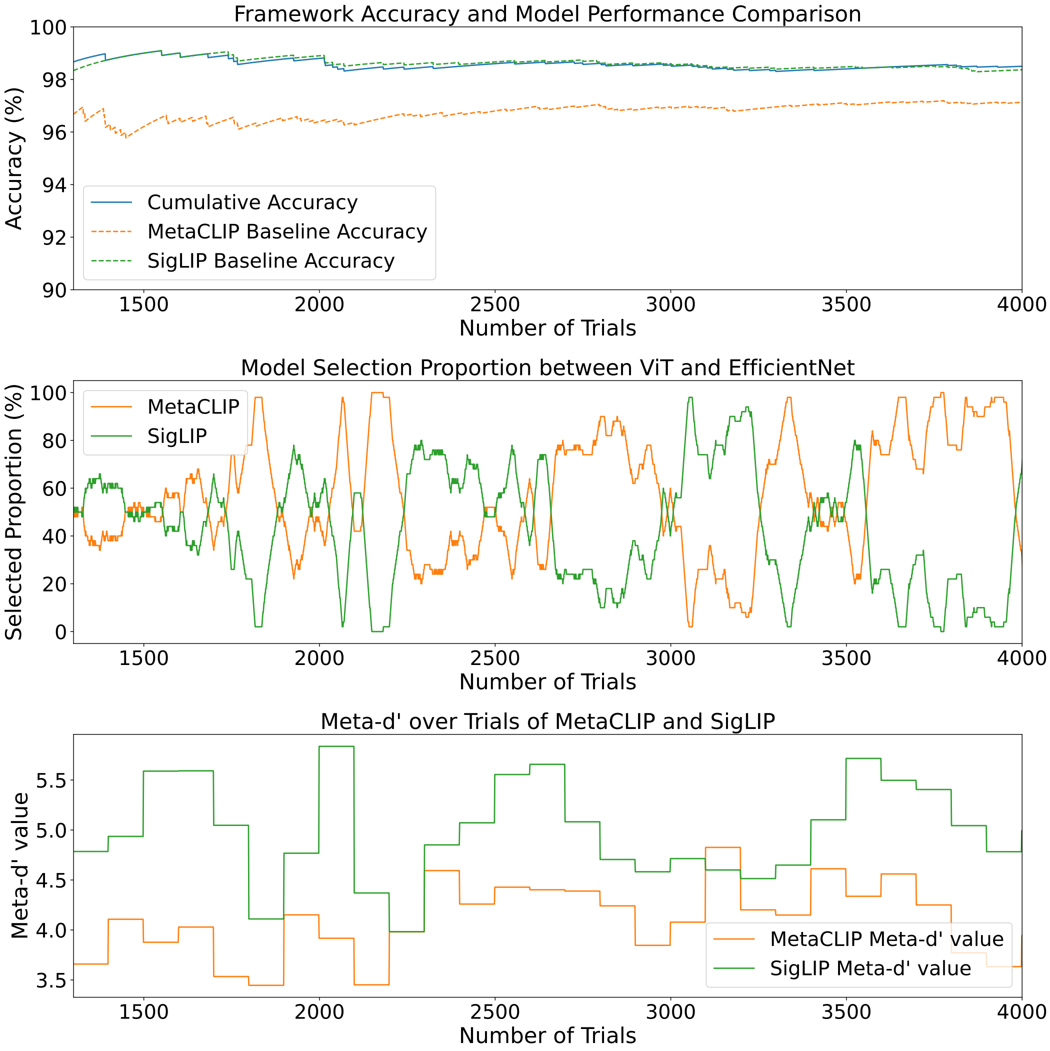

Figure 4: Figure of MetaCLIP and SigLIP with Framework Accuracy (using LinUCB with $\alpha = 0.5$)