Adaptive Verified Self-Distillation with Tiered Critic Ensembles and Shadow-Update Robustification

Anonymous ACL submission

Abstract

We study whether an autonomous agent can reliably improve its expected task performance by retraining on data it generates itself, while avoiding confirmation bias, verifier gullibility, catastrophic forgetting, and performance collapse. This capability is important for scalable lifelong learning and maintenance but is challenging because self-generated candidates can be adversarial or out-of-distribution, verifiers can drift or be gamed, statistical acceptance is underpowered with modest validation budgets, and repeated updates risk forgetting. We propose AVSD-TCE-SR: an adaptive, auditable pipeline that combines (i) an adversary-strength scheduler that produces informative near-boundary yet plausibly in-distribution candidates, (ii) a tiered verification stack that uses cheap ensemble screening, a learned meta-verifier to route candidates, and budgeted expensive checks, and (iii) shadow-update robustification plus statistically principled acceptance (bootstrap one-sided lower bounds, shadow-consistency, and Benjamini–Hochberg correction) before committing updates. Commits are conservative: accepted data are mixed with replay, training uses importance-weighted sampling and optional EWC-style penalties, per-commit trust-region constraints are enforced, and full provenance/audit logs are retained. In controlled experiments (three independent AG_NEWS runs) the pipeline produced reproducible final agents with mean validation accuracy 82.65% and mean test accuracy 81.72%, stable verifier signals, and no catastrophic performance collapse under the configured conservative acceptance rules. Results show the method operates as designed—reducing obvious false accepts and providing auditable evidence for each commit—while absolute gains remained modest under the chosen compact models and limited validation budgets, highlighting trade-offs between conservatism and detection power. Our main contributions are (1) an adaptive adversary scheduler for informative candidate generation, (2) a learnable tiered critic ensemble that routes candidates to budgeted verifications, (3) shadow-update ensemble robustification paired with principled statistical acceptance to control false-accepts, and (4) conservative commit procedures with replay/EWC and comprehensive auditing to mitigate forgetting and enable interpretability.

1 Introduction

Autonomous self-improvement—where an agent retrains on data it generates to improve expected task performance—promises to reduce dependence on external labeling and enable continual adaptation, and related techniques such as pseudo-labeling and data-free distillation have demonstrated that models can expand or preserve knowledge without fresh human labels (Wang et al., 2021; Nagata et al., 2024). Concrete instances include medical and scientific applications where pseudo-labeling or self-distillation reduce annotation needs (e.g., brain tumor segmentation, skin lesion diagnosis, rare-disease imaging) (Li et al., 2024; Deng et al., 2022; Sun et al., 2021a); for example, pseudo-labeling has been applied to diabetic retinopathy and small-lesion detection (Chen et al., 2020a), self-supervised and multi-modality approaches have been explored for skin-lesion and multi-label classification (Wang et al., 2023a; Chaves et al., 2021), and anti-curriculum and similarity-based pseudo-labeling schemes target robustness under label scarcity (Liu et al., 2021; Mahmood et al., 2023). Cross-modal or missing-modality self-distillation methods improve segmentation and diagnostic robustness when modalities are absent or corrupted (Xie et al., 2025; Liu et al., 2023; Tan et al., 2025; He et al., 2023b). In related work, semi-supervised and distillation-focused algorithms tailored to medical segmentation have recently been proposed to better leverage limited labeled data (Wang et al., 2024b,a; Ye et al., 2022; Banerjee et al., 2023).

This question is both timely and important because the research community increasingly demands scalable, low-label-cost lifelong learners that operate in resource-constrained or privacy-sensitive settings where external re-labeling is impractical (Wang et al., 2021; Nagata et al., 2024). Practical trends—larger pretrained models, edge and on-device deployment, and federated or source-free adaptation—create strong incentives for self-contained update mechanisms that preserve previously learned capabilities while incorporating new information (Wang and Niu, 2024; Wang et al., 2021; Qing et al., 2024). Scientific domains in particular (e.g., foundation-model efforts for brain dynamics and fMRI time-series) expose both the potential and unique challenges of large self-distilled representations in safety- and accuracy-critical contexts (Gijsen et al., 2025; Qi et al., 2025). In applied pipelines for medicine and industrial monitoring, self-generated supervision could materially reduce annotation costs and prolong deployment utility, but these benefits are contingent on robust safeguards against silent degradation and unverifiable commits (Li et al., 2024; Tarasiou and Zafeiriou, 2022; Li and Éric Gaussier, 2022).

Achieving reliable autonomous self-improvement is technically hard because several failure modes interact and amplify one another. First, candidate-generation mechanisms that intentionally seek informative, near-boundary examples risk producing unrealistic out-of-distribution (OOD) or adversarial inputs that induce misleading gradients and brittle updates (Jiao et al., 2022; Jeanneret et al., 2021; Zhang et al., 2024a; Xu et al., 2019); adaptive adversary schedules can increase informativeness but also raise the probability of harmful, distribution-shifting candidates, and recent class-wise adversarial self-distillation work for medical segmentation highlights how adversarially-generated supervision can both help and harm training when not carefully constrained (Kose and Zhou, 2025). Second, internal verifiers (ensembles, calibrators, or learned critics) can drift, become overconfident, or be systematically gamed unless their uncertainties and routing decisions are rigorously controlled (Zhuang et al., 2024; Fonseca and Lopes, 2017; Jiang et al., 2024; Balanya et al., 2022; Xiong et al., 2023; Guo et al., 2023; Xie et al., 2024); uncertainty-driven adaptive self-knowledge distillation methods show one path to incorporate model uncertainty into distillation objectives and verifier design (Zeng et al., 2025), and in medical imaging image artifacts and missing modalities further complicate pseudo-label reliability and verifier calibration (Li et al., 2023b; Wang et al., 2022; Weng et al., 2022). Third, the statistical power of any acceptance test is limited by the validation budget n_V and the empirical metric $\mathcal{L}_V$, so small genuine gains may not be detectable while noisy optimizers and stochastic shadow updates can produce spurious apparent improvements (Jewson et al., 2003; Baran et al., 2013; He et al., 2023c; Grassucci et al., 2021); pseudo-label instability, boundary-label diffusion, and class-imbalance issues further exacerbate detectability problems in semi-supervised and transductive settings (Qiu et al., 2026; Teimuri et al., 2025; Mahmood et al., 2023). Finally, repeated commits without explicit anti-forgetting measures invite catastrophic forgetting and coverage loss familiar from the continual-learning literature, so replay, regularization, or other preservation mechanisms are necessary but can conflict with adaptation objectives (Thorne and Vlachos, 2020; Harang and Sanders, 2023; Ma and Bi, 2019; Wen and Itti, 2018; Zhang et al., 2023; Hayes and Kanan, 2021; Kauvar et al., 2023; Ahadzi et al., 2025).

Why has no prior system fully solved this end-to-end problem? Existing lines of work address important subproblems but leave crucial gaps: many self-distillation and iterative synthesis approaches implicitly assume benign pseudo-labels or rely on heuristic filtering rather than auditable, statistically principled acceptance (Shenfeld et al., 2026; Li et al., 2024; Deng et al., 2022; Sun et al., 2021a; Ling et al., 2025; Bundele et al., 2025; Kara et al., 2024; Zhai et al., 2025; Wang et al., 2024b,a). In medical segmentation and few-shot regimes, supervised affinity attention, hierarchical dense-correlation, and masked cross-modality distillation have improved granularity but generally do not couple these advances to auditable commit tests (Jeanneret et al., 2021; Peng et al., 2023; Wang et al., 2023b). Volume- and transformer-based self-distillation works demonstrate decoder-side and long-range consistency gains but typically assume trusted pseudo-labels or curated modality inputs (Banerjee et al., 2023; Wang et al., 2022; Liu et al., 2023; Tan et al., 2025; Xie et al., 2025; Ye et al., 2022; Shao et al., 2026). Put succinctly, prior proposals typically (i) lack adversary-aware candidate generation and severity control (Jiao et al., 2022; Jeanneret et al., 2021; Kose and Zhou, 2025),

(ii) lack multi-tier verification and meta-routing (Zhuang et al., 2024; Balanya et al., 2022; Xie et al., 2024), or (iii) do not enforce conservative, statistically justified commit rules and anti-forgetting safeguards (Jewson et al., 2003; Baran et al., 2013; Thorne and Vlachos, 2020; Wen and Itti, 2018; Amer and Maul, 2019; Mundt et al., 2019; Hayes and Kanan, 2021; Yang et al., 2023).

To address these gaps we introduce Adaptive Verified Self-Distillation with Tiered Critic Ensemble and Shadow-Update Robustification (AVSD-TCE-SR). In one sentence: AVSD-TCE-SR is an auditable pipeline that adaptively proposes near-boundary candidates under an adversary-strength scheduler, routes candidates through a tiered verifier stack (cheap ensemble screening, a learned meta-verifier, and budgeted expensive calibrators), executes shadow updates to estimate commit robustness, and enforces conservative commits using statistically principled acceptance tests together with replay and EWC-style regularization (Jiao et al., 2022; Jeanneret et al., 2021; Zhuang et al., 2024; Balanya et al., 2022; Xie et al., 2024; Jewson et al., 2003; Baran et al., 2013; Zhang et al., 2023; Thorne and Vlachos, 2020; Wen and Itti, 2018; Amer and Maul, 2019; Mundt et al., 2019; Hayes and Kanan, 2021; Yang et al., 2023; Qamar, 2025). Concretely, the pipeline combines three key components: (i) an adaptive adversary scheduler that modulates candidate severity to balance informativeness and in-distribution plausibility, informed by adversarial-severity and semi-supervised adversarial design principles (Jiao et al., 2022; Jeanneret et al., 2021; Kose and Zhou, 2025); (ii) a tiered verification stack comprising fast ensemble-based filters, a learned meta-verifier for routing and triage, and budgeted expensive calibrators for high-stakes checks (Zhuang et al., 2024; Balanya et al., 2022; Xie et al., 2024; Qamar, 2025; Qing et al., 2024); and (iii) shadow-update ensembles plus hypothesis-testing and ensemble-assessment acceptance rules (bootstrap one-sided lower bounds, shadow-consistency constraints, and Benjamini–Hochberg correction) combined with conservative replay and EWC-style penalties to limit forgetting (Jewson et al., 2003; Baran et al., 2013; Zhang et al., 2023; Thorne and Vlachos, 2020; Wen and Itti, 2018). Our design explicitly draws on recent self-distillation and foundation-model distillation work in medical imaging to inform architecture- and task-specific choices (Qi et al., 2025; Wang et al., 2024b; Ye et al., 2022; Shao et al., 2026; Banerjee et al., 2023).

Empirically, on controlled benchmarks our mechanistic analyses and statistical tests show that the pipeline (a) reduces false accepts and verifier gullibility, (b) produces auditable provenance for each commit, and (c) avoids catastrophic performance collapse under conservative acceptance regimes, while revealing the expected trade-off between conservatism and sensitivity identified in prior streaming continual-learning and semi-supervised medical evaluations (Ma and Bi, 2019; Eltahan et al., 2023; Kauvar et al., 2023; Lourenço et al., 2025; Kim and Kim, 2021; Weng et al., 2022; Wang et al., 2024b,a).

2 Related Work

Self-distillation and continual learning A body of recent work studies self-distillation and related knowledge-preservation techniques to mitigate catastrophic forgetting in online or incremental regimes: Nagata et al. show that self-distillation can reduce forgetting in class-incremental streams by replay-aware distillation strategies (Nagata et al., 2024), Kurmi et al. emphasize incorporating uncertainty estimates from older models into distillation to better preserve past knowledge (Kurmi et al., 2021), and Wang et al. propose data-free pseudo-data distillation for incremental language models when previous data are unavailable (Wang et al., 2021). Benchmarking and domain-specific continual-learning studies further expose limitations of naive distillation and replay: Gao et al. construct a crowd-counting lifelong benchmark highlighting replay generalization issues (Gao et al., 2022), Shenfeld et al. argue that on-policy and distilled signals can enable more robust continual updates (Shenfeld et al., 2026), and federated setups have adapted progressive self-distillation to personalize without forgetting (Wang and Niu, 2024). Complementary directions address targeted preservation and controlled forgetting—UNDIAL adjusts logits for robust unlearning (Dong et al., 2024), FADE combines sparse LoRA with self-distillation for selective forgetting (Kelsch et al., 2026), and approaches that reset or diversify representations aim to retain useful features while learning new ones (Park, 2024; Monte et al., 2025). Together these works motivate our conservative update, replay and EWC-style regularization choices and highlight gaps: many self-distillation methods assume benign pseudo-labels or stationary data and do not

provide tiered verification against adversarially-generated candidates, which motivates our meta-verifier and shadow-update robustness checks.

Adversarial self-generation and robust training Several lines of research study automated generation of adversarial or self-supervised candidates and their use (or dangers) for model adaptation. Large-scale automated adversarial generation and targeted transfer attacks demonstrate how self-generated examples can be both powerful and brittle (Zhang et al., 2024a; He et al., 2023a; Peng et al., 2024), and diffusion-based reversible adversarial example generators illustrate techniques for creating synthetic perturbations with controlled recoverability (Xing et al., 2023). On the defense side, self-supervised adversarial training and fast adversarial training methods have been proposed to improve robustness with limited compute, including AS-TrA for adaptive attacks in self-supervised settings (Chhipa et al., 2025), hierarchical/self-supervised adversarial pipelines for medical imagery (Malik et al., 2025), and fast-adversarial-training improvements that guide perturbation generation with self-knowledge (Jiang et al., 2024). Works that combine ensembling, defensive distillation, or sustainable iterative adversarial curricula show promise but also reveal trade-offs in compute and distributional drift (Imam et al., 2023; Wang et al., 2024c; Lu et al., 2023; Zhang et al., 2024a). Domain-adaptive self-training methods (e.g., DaMSTF) further highlight label-noise risks when bootstrapping from model predictions (Lu et al., 2023). These studies justify our adaptive adversary scheduler, multi-stage filtering (ensemble → meta-verifier → shadow-updates), and the inclusion of adversarial negatives for verifier retraining: prior adversarial generation methods can produce highly informative but potentially out-of-distribution candidates, so layered, calibrated acceptance tests are needed.

Ensembles, calibration, and verification for safe updates Ensemble-based screening, calibration, and principled verification have been widely used across tasks and domains to improve reliability and to audit model updates. Classic and recent calibration methods (temperature scaling, parameterized or adaptive extensions, bin-wise variants, and proximity-aware adjustments) establish how post-hoc and constrained calibration can reduce overconfidence and enable better decision thresholds (Fonseca and Lopes, 2017; Tomani et al., 2021; Ji et al., 2019; Balanya et al., 2022; Xie et al., 2024; Jiang et al., 2024; Xiong et al., 2023; Guo et al., 2023). Graph- and domain-specific ensemble calibration work (e.g., GETS) further illustrates tailoring calibration to model structure (Zhuang et al., 2024). Ensemble systems have been successfully applied in medical screening and other high-stakes domains—ensemble screening for diabetic retinopathy and glaucoma (Antal and Hajdu, 2014; Fu et al., 2018), self-ensembling ViTs for chest X-ray robustness (Imam et al., 2023), and broader ensemble/meta-analysis approaches in bio/clinical studies underscore the role of aggregation for stable decisions (Moayedi et al., 2023; Wang et al., 2024d; Zhang et al., 2020b; Xiao et al., 2023; Zhao et al., 2022; Li et al., 2013; Su et al., 2024; Yang et al., 2024a; Tužil et al., 2023; Nama et al., 2026). Methodologically, variable-subspace ensembles and random-subspace frameworks motivate ensemble diversity strategies (Tian and Feng, 2021; Martin et al., 2020), while optimization and verification literatures (trust-region and derivative-free schemes, cross-entropy and centroid-guided samplers, distributed/quasi-Newton DFO, and iterative ensemble smoothers) inform robust update and candidate-search procedures and failure-mode analyses (Agarwal et al., 2009; Ma and Bi, 2019; Eltahan et al., 2023; Hannanu et al., 2025; Gu et al., 2025; Gao et al., 2021; Alpak et al., 2021; Niño and Sandu, 2014; Chen et al., 2010; Gao and Zhao, 2024; Vepakomma et al., 2020). Finally, applications that combine ensembling with auditing and provenance logging point to practical reproducibility and interpretability practices we adopt (Finn et al., 2022; Baldovin, 2018; Xing et al., 2023; Wang et al., 2024d). Collectively, these works motivate our ensemble temperature scaling, uncertainty CIs, meta-verifier design, and the conservative statistical acceptance tests (bootstrap/sequential) used to guard self-updates.

3 Method

We formalize autonomous verified self-distillation as an iterative, auditable update procedure in which an agent autonomously proposes candidate training examples, verifies a controlled subset by internal checks, and commits conservative parameter updates only when statistically robust improvement is observed (Zhang et al., 2020a; Zhu et al., 2023; Cheng et al., 2020, 2019). Below we (i) state the problem and notation precisely, (ii) present the pipeline as a sequence of mathematical modules,

(iii) give design rationales for each module, and (iv) state the statistical acceptance and commit rules that enforce low false-accept risk.

Problem definition and notation. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space. The agent is a parametric predictor

$$f_\theta : \mathcal{X} \to \Delta(\mathcal{Y}), \qquad \theta \in \Theta, \tag{1}$$

where $\Delta(\mathcal{Y})$ denotes probability distributions over $\mathcal{Y}$. The learning target is the (unknown) data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We measure performance by an expected risk

$$\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\,\ell(f_\theta(x), y)\,], \tag{2}$$

for a supervised loss ℓ (e.g., cross-entropy). In practice we approximate $\mathcal{R}$ via a modest held-out validation set

$$V = \{(x_i, y_i)\}_{i=1}^{n_V} \tag{3}$$

and define the empirical validation metric

$$\mathcal{L}_V(\theta) = \frac{1}{n_V} \sum_{i=1}^{n_V} \ell(f_\theta(x_i), y_i). \tag{4}$$

Assumptions used throughout: (A1) the unlabeled pool $\mathcal{U} \subseteq \mathcal{X}$ contains samples that are, with nonzero probability, near the support of $\mathcal{D}$; (A2) validation set V is iid from $\mathcal{D}$ and sufficiently representative to detect meaningful degradations under our chosen risk level; (A3) optimizer and training stochasticity are exchangeable across repeated shadow simulations. We explicitly expose dependence on scheduler and verifier state to emphasize nonstationarity of the pipeline.

At round t the system has access to: - an unlabeled pool $\mathcal{U}$ from which seeds are drawn, - a replay buffer R of vetted examples, - a verifier ensemble $\mathcal{C} = \{c_j\}_{j=1}^{k}$ (each $c_j : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ returning logits), - a learned meta-verifier M mapping feature vectors to a risk estimate, - and a scheduler state α_t that controls candidate generation.

The pipeline proposes a candidate batch

$$B_t = \{(x_i, \hat{y}_i, \pi_i)\}_{i=1}^{m}, \qquad \hat{y}_i = \arg\max_y f_{\theta_t}(x_i), \tag{5}$$

where π_i denotes provenance metadata (seed id, adversary parameters, etc.). The objective is to produce a sequence $\{\theta_t\}$ such that $\mathcal{R}(\theta_t)$ decreases monotonically in expectation while controlling Type-I errors (false accepts), verifier gullibility, and catastrophic forgetting.

Overall pipeline mapping and desiderata. Conceptually one round of the pipeline implements a mapping

$$\mathcal{P} : (\theta_t, \mathcal{U}, R, \mathcal{C}, M, \alpha_t) \mapsto (\theta_{t+1}, R', \alpha_{t+1}, \mathcal{A}_t), \tag{6}$$

where $\mathcal{A}_t$ denotes recorded audit artifacts for the round. Desiderata that drive the mapping design are: (1) conservatism: avoid committing unless evidence of net improvement is statistically strong; (2) efficiency: allocate expensive verification selectively; and (3) audibility: produce provenance and repeatable artifacts enabling external review.

We next decompose $\mathcal{P}$ into modular components and formalize each.

Adaptive adversarial candidate generation. Motivation: generate candidates that are informative (probe near decision boundaries) yet plausibly in-distribution to avoid training on OOD artifacts; static adversaries produce either trivial or highly unrealistic samples so we schedule adversary strength dynamically (Zhang et al., 2024a; Chhipa et al., 2025).

Define the generator operator

$$\mathcal{G}_{\alpha_t} : \mathcal{U} \times \Theta \to \mathcal{X}, \tag{7}$$

$$x' = \mathcal{G}_{\alpha_t}(x; \theta_t), \tag{8}$$

where α_t parameterizes allowed perturbation magnitude, iteration budget, and augmentation diversity. Each generated point is required to satisfy a plausibility constraint expressed via an in-distributionness score

$$d_{\text{in}} : \mathcal{X} \to \mathbb{R}_{\geq 0}, \qquad d_{\text{in}}(x') \leq \tau(\alpha_t), \tag{9}$$

for a threshold $\tau(\alpha_t)$ chosen by the scheduler to maintain a target plausibility probability.

A generic gradient-based construction used by the scheduler is

$$x' = \mathrm{Aug}\big(x + \delta\big), \tag{10}$$

$$\delta \leftarrow \arg\max_{\|\delta\| \le \epsilon(\alpha_t)} \mathcal{A}(x + \delta; \theta_t), \tag{11}$$

where $\mathcal{A}$ is an adversarial objective (e.g., increase margin loss or induce label change) and $\epsilon(\alpha_t)$ encodes allowed perturbation scale under schedule α_t.

Scheduler update rule (informal, compact formalization): the scheduler adjusts α_t to trade off informativeness and plausibility by observing recent statistics

$$\alpha_{t+1} = \Gamma\Big(\alpha_t,\ \widehat{\Pr}_t[d_{\text{in}}(x') \le \tau],\ \widehat{\text{acc}}_t,\ \kappa_t\Big), \tag{12}$$

where $\widehat{\Pr}_t[\cdot]$ and $\widehat{\text{acc}}_t$ are empirical plausibility and acceptance rates and κ_t encodes calibration drift; Γ is a monotone operator that reduces ϵ when plausibility falls and increases it when acceptance stalls (Zhang et al., 2024a; Chhipa et al., 2025; Moshavash et al., 2021; Cheng et al., 2020).

Design rationale: keeping generation near informative boundaries increases value of vetted examples while plausibility constraints limit exposure to OOD artefacts; an explicit scheduler avoids brittle one-shot adversary settings.

Tiered verification pipeline (cheap → learned → expensive). Motivation: allocate compute efficiently by filtering candidates through increasingly expensive checks; learn which candidates warrant expensive verification (Zhuang et al., 2024; Balanya et al., 2022; Dereka et al., 2023; Vidal et al., 2020).

Given candidate $(x', \hat{y}', \pi)$ the verification stack computes three stages.

Stage 1 — Ensemble screening. Each ensemble member c_j returns logits $z_j(x') \in \mathbb{R}^{|\mathcal{Y}|}$. With a temperature τ_{temp} (fitted on V) form probability vectors

$$p_j(x') = \mathrm{softmax}\big(z_j(x')/\tau_{\text{temp}}\big). \tag{13}$$

and define the ensemble mean and empirical dispersion

$$\bar{p}(x') = \frac{1}{k}\sum_{j=1}^{k} p_j(x'), \tag{14}$$

$$\mathrm{DIS}(x') = \sqrt{\frac{1}{k}\sum_{j=1}^{k} \|p_j(x') - \bar{p}(x')\|_2^2}. \tag{15}$$

Additional scalar features (entropy, max confidence, mean logit) are concatenated into a feature vector $\phi_{\text{ens}}(x')$. Temperature scaling and ensemble aggregation reduce overconfidence and provide uncertainty signals (Fonseca and Lopes, 2017; Zhuang et al., 2024; Laves et al., 2019; Frenkel and Goldberger, 2021, 2022; Zhang et al., 2024b).

Design rationale: a cheap ensemble provides calibrated uncertainty proxies and diversity-based dispersion that are informative and computationally inexpensive relative to downstream checks.

Stage 2 — Meta-verifier routing. The meta-verifier is a learned mapping

$$M : \Phi \to [0, 1], \qquad r = M\big(\phi(x')\big), \tag{16}$$

where $\phi(x')$ concatenates ensemble features, provenance π, and agent-side signals (e.g., f_{θ_t} logits or gradient norms). The scalar r estimates false-accept risk and yields a ternary routing decision

$$\begin{cases} \texttt{reject}, & r \ge r_{\text{reject}}, \\ \texttt{escalate}, & r_{\text{pass}} \le r < r_{\text{reject}}, \\ \texttt{pass}, & r < r_{\text{pass}}. \end{cases}$$

M is periodically retrained on historical labelled outcomes so it adapts to evolving failure modes (Balanya et al., 2022; Shen et al., 2021).

Design rationale: routing via a learned predictor concentrates budgeted checks on ambiguous or high-risk candidates and lets the system improve routing accuracy over time through supervised updates.

Stage 3 — Budgeted secondary verification. Candidates routed to `escalate` undergo limited-budget secondary checks (MC-dropout aggregation, augmentation ensembling, environment rollouts, or shadow-update simulations). The tier concentrates expensive compute on ambiguous/high-impact candidates and lets M learn to avoid repeated expensive checks (Zhuang et al., 2024; Balanya et al.,

2022; Moshavash et al., 2021; Nalluri et al., 2025; Cha et al., 2021).

Design rationale: secondary checks provide richer, often stochastic signals that expose brittleness unobserved in the cheap ensemble while keeping amortized cost bounded.

Meta-verifier training objective and dataset. The meta-verifier M is trained on a rolling dataset

$$\mathcal{D}_M = \{(\phi(x'), y_{\text{outcome}})\} \tag{17}$$

where $y_{\text{outcome}} \in \{\text{true\_accept}, \text{false\_accept}, \text{rejected}\}$ is obtained from historical secondary checks and audited human labels when available. The training solves

$$\min_{\psi} \mathbb{E}_{(\phi,y)\sim\mathcal{D}_M}\Big[\mathcal{L}_{\text{meta}}\big(M_\psi(\phi), y\big)\Big], \tag{18}$$

with class-weighting to penalize false-accepts heavily. Design rationale: explicit supervision on outcomes reduces gullibility and allows calibration of operating thresholds $r_{\text{pass}}, r_{\text{reject}}$.

Shadow-update robustification and statistical acceptance criterion. Motivation: reduce commits caused by optimizer noise, lucky minibatch effects, or stochasticity by requiring consistent simulated improvement under diverse shadow conditions (Jewson et al., 2003; Baran et al., 2013; Narayanan et al., 2020; Li and Zhang, 2020).

Let B denote a vetted batch of accepted-by-screening examples. Define a constrained single-update operator

$$\mathcal{T}_{\text{constr}}(\theta, B; \eta) = \arg\min_{\theta'} \mathbb{E}_{(x,y)\in B}\big[\ell(f_{\theta'}(x), y)\big] + \mathcal{R}_{\text{trust}}(\theta', \theta; \eta), \tag{19}$$

where $\mathcal{R}_{\text{trust}}$ is a trust-region or proximal penalty (e.g., L2-clipping, KL or quadratic penalty) parameterized by shadow config η.

Generate S shadow configurations $\{\eta^{(s)}\}_{s=1}^{S}$ and corresponding shadow parameters

$$\theta^{(s)} = \mathcal{T}_{\text{constr}}(\theta_t, B; \eta^{(s)}), \qquad s = 1, \ldots, S. \tag{20}$$

Evaluate validation loss reductions

$$\Delta^{(s)} = \mathcal{L}_V(\theta_t) - \mathcal{L}_V(\theta^{(s)}). \tag{21}$$

Form the empirical distribution $\{\Delta^{(s)}\}_{s=1}^{S}$ and compute: - the bootstrap one-sided lower confidence bound $L_\alpha(\Delta)$ on the mean improvement (level $1-\alpha$), - the shadow-consistency fraction

$$\rho = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{\Delta^{(s)} \geq 0\}. \tag{22}$$

Acceptance rule: commit only if

$$L_\alpha(\Delta) > 0 \quad \text{and} \quad \rho \geq \rho_{\min}, \tag{23}$$

with a high threshold $\rho_{\min}$ (e.g., 0.75). To control false discoveries across multiple rounds we compute per-round one-sided p-values and apply the Benjamini–Hochberg (BH) procedure across rounds to bound the false discovery rate (Jewson et al., 2003; Baran et al., 2013). Optionally, sequential testing variants (e.g., repeated confidence sequences or SPRT-inspired rules) may be used when many small increments are expected (Baran et al., 2013).

Concrete bootstrap lower bound computation (compact): let $\bar{\Delta} = \frac{1}{S}\sum_s \Delta^{(s)}$ and let $\{\bar{\Delta}^{*(b)}\}_{b=1}^{B}$ be bootstrap means obtained by resampling $\{\Delta^{(s)}\}$ with replacement; then

$$L_\alpha(\Delta) = \text{quantile}_\alpha\big(\{\bar{\Delta}^{*(b)}\}_{b=1}^{B}\big). \tag{24}$$

Design rationale: the bootstrap lower bound targets estimation uncertainty from a small V, while ρ guards against brittle improvements that depend on particular optimizer seeds; together they reduce Type-I errors in committing self-generated data (Jewson et al., 2003; Baran et al., 2013).

Conservative commit update, replay mixing, and anti-forgetting regularization. Motivation: accepted self-generated data can be non-representative and cause forgetting; mixing with replay and applying parameter regularization preserves prior competencies (Thorne and Vlachos, 2020; Zhang et al., 2023; Hayes and Kanan, 2021; Zhu et al., 2023).

When B is accepted, define the commit empirical distribution

$$\tilde{D} = (1-\gamma)\,\tilde{D}_{\text{replay}} + \gamma\,\tilde{D}_{\text{new}} \tag{25}$$

where $\tilde{D}_{\text{new}}$ is the accepted generated data, $\tilde{D}_{\text{replay}}$ is sampled from R, and $\gamma \in [0,1]$ is an acceptance mixing weight (possibly adaptive).

The conservative commit objective is

$$\mathcal{L}_{\text{commit}}(\theta) = \mathbb{E}_{(x,y)\sim\tilde{D}}\big[\ell(f_\theta(x),y)\big] + \lambda_{\text{EWC}}\,\Omega_{\text{F}}(\theta;\theta_t), \tag{26}$$

where Ω_{F} is an optional quadratic Fisher penalty centered at θ_t (EWC-style) that penalizes movement in important directions (Thorne and Vlachos, 2020). The final commit performs a constrained optimization (trust-region or proximal update) to obtain θ_{t+1}:

$$\theta_{t+1} = \arg\min_{\theta}\ \mathcal{L}_{\text{commit}}(\theta) \quad \text{s.t.} \quad \mathcal{D}_{\text{TR}}(\theta,\theta_t) \le \delta, \tag{27}$$

where $\mathcal{D}_{\text{TR}}$ is a distance (e.g., KL or parameter-norm) and δ a small radius.

After committing, R is updated by adding vetted examples and pruning to maintain coverage (per-class or loss-based metrics) to avoid replay concentration (Zhang et al., 2023; Hayes and Kanan, 2021). Design rationale: mixing and regularization bound the influence of any single self-generated batch and preserve competencies learned from prior data.

Verifier maintenance, calibration, and meta-verifier retraining. Motivation: verifiers can drift or become gullible if exposed to generated data; periodic maintenance keeps the stack reliable and auditable (Fonseca and Lopes, 2017; Zhuang et al., 2024; Balanya et al., 2022; Laves et al., 2019).

Concretely, we maintain verifiers by (i) retraining/fine-tuning ensemble members on the union of original training data, vetted accepted samples, and curated adversarial negatives; (ii) monitoring calibration metrics on V and refitting temperature scaling when ECE or related statistics degrade; and (iii) preserving ensemble diversity by reinitializing or perturbing member parameters periodically (Fonseca and Lopes, 2017; Zhuang et al., 2024; Laves et al., 2019; Frenkel and Goldberger, 2021, 2022; Zhang et al., 2024b; Yang, 2025; Ingrisch et al., 2025; Chanda et al., 2025). The meta-verifier M is retrained on historical candidate outcomes (true accept / false accept / rejected) so its predictive quality improves over time (Balanya et al., 2022; Shen et al., 2021). Design rationale: continual maintenance prevents systemic drift and helps bound long-run false-accept rates.

Auditing, interpretability, and stopping. Motivation: scientific and safety claims require reproducible evidence for each accepted change (Finn et al., 2022; Wang et al., 2024d; Shevlin et al., 2024).

Formally, the audit artifact for round t is

$$\mathcal{A}_t = \Big\{B_t,\ \{\phi_{\text{ens}}(x')\}_{x'\in B_t},\ \{r(x')\}_{x'\in B_t},\ \{\theta^{(s)}\}_{s=1}^{S},\ \{\Delta^{(s)}\}_{s=1}^{S},\ L_\alpha(\Delta),\ \rho,\ \text{commit\_decision}\Big\}. \tag{28}$$

We halt the outer loop under an explicit stopping rule (fixed round budget or stagnation defined by no accepted updates for a configured number of consecutive rounds) and return the audited sequence of accepted commits (Shevlin et al., 2024).

Design rationale: deterministic artifacts allow post-hoc verification, reproducibility, and external compliance checks.

Algorithmic summary. Algorithm 1 summarizes one round of AVSD-TCE-SR as a sequence of deterministic and stochastic mappings consistent with the mathematical components above.

Design rationale recap. The method rests on three risk-control pillars: (1) adaptive generation that keeps candidates informative yet plausibly in-distribution (Zhang et al., 2024a; Chhipa et al., 2025; Moshavash et al., 2021; Cheng et al., 2020, 2019); (2) a learnable, tiered verification stack that funnels only ambiguous or high-impact candidates to expensive checks while learning to predict false accepts (Zhuang et al., 2024; Balanya et al., 2022; Dereka et al., 2023); and (3) shadow-update ensemble validation plus conservative, regularized commits (replay + EWC-style penalties) to avoid optimizer-noise-driven or forgetting-inducing accepts (Jewson et al., 2003; Thorne and Vlachos, 2020; Zhang et al., 2023; Zhu et al., 2023). These elements produce an auditable pipeline for testing whether autonomous self-distillation can reliably improve expected task performance under principled statistical control.

Algorithm 1 One round of AVSD-TCE-SR

Input: θ_t, unlabeled pool $\mathcal{U}$, validation V, replay R, verifiers $\mathcal{C}$, meta-verifier M, scheduler state α_t
Sample seeds $S \subset \mathcal{U}$
for each seed $x \in S$ **do**
  Generate candidate $x' \leftarrow \mathcal{G}_{\alpha_t}(x; \theta_t)$ and form $(x', \hat{y}', \pi)$
  Compute ensemble logits $z_j(x')$, probabilities $p_j(x')$ and features $\phi_{\text{ens}}(x')$
  Form full feature $\phi(x') = [\phi_{\text{ens}}(x'), \pi, \text{agent-features}]$
  Obtain meta-score $r \leftarrow M(\phi(x'))$
  if $r \geq r_{\text{reject}}$ **then**
    mark candidate as `reject`
  else if $r \geq r_{\text{pass}}$ **then**
    escalate to secondary verification (MC-dropout, rollouts, or shadow-sim)
  else
    mark candidate as `pass for shadowing`
  end if
end for
Collect vetted batch B from candidates passing screening
For B, generate S shadow updates $\{\theta^{(s)}\}$ via $\mathcal{T}_{\text{constr}}(\cdot)$ under varied $\{\eta^{(s)}\}$
Evaluate $\{\Delta^{(s)}\}$ on V; compute bootstrap lower bound L_α and ρ
if $L_\alpha > 0$ and $\rho \geq \rho_{\min}$ (after BH correction) **then**
  Commit conservative update: $\theta_{t+1} \leftarrow \arg\min_\theta \mathcal{L}_{\text{commit}}(\theta)$
  Update replay buffer R with vetted examples
  Log provenance and audit artifacts $\mathcal{A}_t$
else
  $\theta_{t+1} \leftarrow \theta_t$
end if
Update scheduler α_{t+1} and retrain M / maintain verifiers as scheduled =0

Summary of formal guarantees (operational). While formal, distribution-free guarantees are impossible without strong assumptions, the pipeline enforces two operational controls with quantifiable properties: (i) per-round Type-I control via the bootstrap LCB and BH correction which bounds the false discovery rate across rounds under exchangeability assumptions (Jewson et al., 2003; Baran et al., 2013); and (ii) a robustness requirement via the shadow-consistency fraction ρ that empirically reduces commit probability for improvements sensitive to optimizer or sampling noise. Together these provide actionable, auditable checks that managers or auditors can inspect and tune to reach a desired operating point.

4 Experimental Setup

This section gives a complete, reproducible description of how we instantiate and evaluate AVSD-TCE-SR on the MNIST benchmark. All choices (data splits, preprocessing, model sizes, adversary settings, verifier procedures, statistical tests, and seeds) are specified so other researchers can rerun the experiments and reproduce reported results. When a design choice follows prior work we cite that work immediately after the supporting sentence.

4.1 Overview and goals

Our empirical goals are threefold: (i) demonstrate that an agent can improve (or at least not degrade) task performance by retraining on self-generated data under conservative acceptance rules, (ii) measure the pipeline's false-accept rate and catastrophic-forgetting behaviour under repeated rounds, and (iii) evaluate trade-offs between conservatism (Type I control) and sensitivity (power) across ablations. To make the evaluation controlled and fast for many ablations we use MNIST as the target distribution and follow the adaptive adversary, tiered verification, shadow-update, and conservative-commit pipeline described in Section §**??**. Design decisions (compact models, modest validation budget, and per-round caps) prioritize reproducibility and the ability to run many independent seeds and ablations within a modest compute envelope; this setting exposes the statistical and verification challenges that motivated our method while keeping experiments tractable.

4.2 Notation and problem instantiation

We reuse notation from Section §**??**. Concretely, let $f_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$ denote the agent with parameters θ. The held-out validation set is $V = \{(x_i, y_i)\}_{i=1}^{n_V}$ and the held-out test set is $T = \{(x_i, y_i)\}_{i=1}^{n_T}$. We report empirical validation loss and accuracy:

$$\mathcal{L}_V(\theta) = \frac{1}{n_V} \sum_{i=1}^{n_V} \ell(f_\theta(x_i), y_i), \tag{29}$$

$$\mathrm{Acc}_V(\theta) = \frac{1}{n_V} \sum_{i=1}^{n_V} \mathbf{1}\{\arg\max_y f_\theta(x_i) = y_i\}, \tag{30}$$

and analogous definitions for $\mathcal{L}_T$ and Acc_T.

We denote the unlabeled candidate pool at round t by $\mathcal{U}_t$ and the set of generated candidate examples proposed by the adversary as G_t. Each round produces a vetted batch $B_t \subseteq G_t$ that is considered for commit. All RNG seeds (top-level run seed and per-component seeds) are fixed and recorded in the run manifest; code sets these seeds at process start and at each deterministic operation to enable exact replay.

4.3 Dataset and preprocessing

- Dataset loader: we load MNIST using the HuggingFace datasets API with the command `datasets.load_dataset("mnist")` (we record the exact HF dataset commit hash and version in the released artifacts). - Downsampling and splits: we perform stratified random sampling by class from the HF `train` and `test` splits to obtain train = 5000, validation (V) = 1000, test = 1000; sampling is deterministic per-run using the top-level RNG seed. Stratified sampling preserves per-class balance so replay and candidate sampling do not introduce class-skew artifacts in downstream evaluations. - Preprocessing: read each 28×28 uint8 image, cast to float32, normalize pixel values to $[0, 1]$, and flatten to a 784-dimensional vector. Resulting per-example feature vector $x \in \mathbb{R}^{784}$ is used throughout. Exact preprocessing code is included in the released repository to ensure bit-for-bit reproducibility.

Design rationale: these dataset choices keep experiments fast while preserving the core classification challenge and sufficient class stratification needed for unbiased replay and candidate sampling; the modest validation size also deliberately stresses the statistical acceptance tests to evaluate their robustness.

4.4 Model architectures and initial conditions

- Agent A_0 (base model): MLP with layers Dense(48, ReLU) $\to$ Dense(24, ReLU) $\to$ Dense(10, linear logits). Total trainable parameters = 39,106. Compact architectures reduce per-round compute so we can run many rounds and seeds while still exposing calibration and forgetting phenomena. - Verifier ensemble $\mathcal{C} = \{c_j\}_{j=1}^{3}$: three independent MLPs Dense(24, ReLU) $\to$ Dense(10, linear logits). Each verifier is independently initialized with a distinct RNG seed to promote ensemble diversity. Temperature-scaling scalars for each verifier are fit on V prior to round $t = 1$ (calibration) as described in Subsection "Verifier retraining, calibration monitoring, and diversity maintenance". The use of ensembles and temperature scaling follows standard calibration/ensemble practice (Fonseca and Lopes, 2017; Zhuang et al., 2024). - Meta-verifier M: an MLP with one hidden layer (32 units, ReLU) mapping per-candidate feature vectors (ensemble statistics, provenance flags, agent-side features) to a scalar risk score $r \in [0, 1]$ and a ternary recommendation (`pass`/`escalate`/`reject`). M is trained online on accumulated labeled candidate outcomes (see Subsection "Meta-verifier training and maintenance").

All initial checkpoints and RNG seeds are saved to enable exact restoration of A_0 and all verifier/meta-verifier members.

Design rationale: small, diverse verifier ensembles provide a tractable yet informative disagreement signal; the meta-verifier compresses ensemble statistics and provenance into a routing decision to reduce expensive checks while retaining auditability.

4.5 Training and optimizer defaults

- Commit optimizer: SGD with momentum 0.9, initial learning rate $\eta_{\text{commit}} = 1\text{e}{-3}$, batch size 32, weight decay 10^{-5}; per-step gradient clipping is applied on the L_2 norm with clip threshold 0.1. Commits run up to 5 epochs over the mixed commit dataset but include early stopping if validation loss does not improve for 2 commit-epochs. - Shadow-update optimizer: SGD with $\eta_{\text{shadow}} = 5\text{e}{-4}$, momentum 0.9; each shadow update is a single constrained mini-update (one epoch over the vetted batch B, see Subsection "Shadow-update robustification and statistical acceptance") to simulate conservative parameter shifts. These small-step settings reduce optimizer-induced variance during

shadow evaluation.

Design rationale: commit optimizer is slightly more aggressive than shadow updates to reflect that accepted commits should consolidate improvements, whereas shadow updates emulate the conservative, low-magnitude changes we expect the agent to encounter if a commit were applied; gradient clipping and small learning rates reduce optimizer-induced false positives in shadow assessments.

4.6 Adversarial / candidate generation

- Seed sampling: at round t we draw up to $N_{\text{cand}} = 200$ candidate seeds S_t from the unlabeled pool $\mathcal{U}$ by stratified sampling when labels are available or uniform sampling otherwise; sampling is deterministic given the per-round RNG seed. Stratification maintains per-class exploration and prevents the adversary from concentrating excessively on classes with more unlabeled mass. - Attack operators: we generate candidates by applying FGSM and short PGD (5-step) attacks targeting the agent logits; perturbations are constrained in L_∞ with $\varepsilon \in \{0.05, 0.1\}$ (pixel range normalized to $[0, 1]$). After adversarial perturbation we apply plausibility-preserving augmentations (rotation in $[-10°, 10°]$, brightness shift ± 0.1). The use of FGSM/PGD-style attacks and small augmentations follows prior adversarial/self-generation practice (Zhang et al., 2024a; Chhipa et al., 2025). - Adversary scheduler: the scheduler state α_t selects attack strength (choice of ε and PGD iterations) as a deterministic function of (i) recent acceptance rate a_{t-1}, (ii) verifier calibration drift measured as ΔECE_t on V, and (iii) an in-distributionness score $d_{\text{in}}(x)$ (autoencoder reconstruction error). The scheduler increases adversary strength if a_{t-1} is large and decreases it when ΔECE_t indicates verifier drift or when d_{in} suggests OOD candidates. The scheduler is implemented as a small state machine to ensure repeatability for a given RNG seed.

Design rationale: adaptive adversary strength targets near-boundary yet plausibly in-distribution examples to maximize informative candidates while avoiding gross OOD examples that would be misleading; the deterministic scheduler makes experiments reproducible and limits confounding by stochastic adversary exploration.

4.7 Tiered verification pipeline (cheap → learned → expensive)

We formalize the three verification stages and their outputs.

- Ensemble screening (cheap): for each candidate $(x', \hat{y}')$ the verifier ensemble produces per-member logits $\{z_j(x')\}_{j=1}^k$ and probability vectors $p_j(x') = \text{softmax}(z_j(x')/T_j)$ where T_j is the temperature for member j (fitted on V). We compute the following per-candidate features:

$$\bar{p}(x') = \frac{1}{k}\sum_{j=1}^{k} p_j(x'), \quad \text{DIS}(x') = \sqrt{\frac{1}{k}\sum_{j=1}^{k} \|p_j(x') - \bar{p}(x')\|_2^2},$$
$$H(x') = -\sum_{c} \bar{p}_c(x') \log \bar{p}_c(x'), \quad \text{maxconf}(x') = \max_c \bar{p}_c(x'). \tag{31}$$

Candidates failing simple ensemble thresholds (e.g., maxconf below a low-confidence cutoff) are rejected cheaply. Temperature-scaling per verifier is fit on V prior to screening as in (Fonseca and Lopes, 2017).

Design rationale: cheap ensemble statistics (mean probability, disagreement, entropy, max confidence) provide a low-compute filter that removes obviously low-quality or out-of-distribution candidates and reduces compute spent on expensive checks.

- Meta-verifier scoring (learned): the meta-verifier M maps candidate feature vectors $\phi(x')$ (ensemble features, provenance flags, and agent-side features) to a risk score $r = M(\phi(x')) \in [0, 1]$. Given tunable thresholds $0 \le r_{\text{pass}} < r_{\text{reject}} \le 1$, the routing is:

$$\text{decision}(x') = \begin{cases} \texttt{pass-for-shadow}, & r < r_{\text{pass}}, \\ \texttt{escalate}, & r_{\text{pass}} \le r < r_{\text{reject}}, \\ \texttt{reject}, & r \ge r_{\text{reject}}. \end{cases} \tag{32}$$

M is retrained periodically on labeled historical outcomes (see Subsection "Meta-verifier training and maintenance"); a learned routing reduces repeated expensive checks by routing low-risk candidates directly to shadow simulation (Balanya et al., 2022; Zhuang et al., 2024).

Design rationale: learned routing compresses multi-dimensional ensemble and provenance signals into a calibrated risk estimate so we can concentrate budgeted expensive checks on ambiguous candidates while still retaining auditability.

- Budgeted secondary verification (expensive): candidates routed to `escalate` incur one of: MC-dropout uncertainty estimation (dropout $p = 0.1$,

20 draws), longer adversarial/augmented evaluation, or an immediate shadow-sim (partial shadow update targeted at single high-impact candidates). The fraction of candidates escalated per round is capped by a compute budget to ensure experiment feasibility.

Design rationale: limiting expensive checks bounds per-round compute and forces the meta-verifier to be selective, which is realistic for deployed systems with finite verification resources.

4.8 Meta-verifier training and maintenance

- Training signal: each candidate routed to expensive checks or that is used in a commit is eventually labeled with an outcome class in {`true-accept`, `false-accept`, `reject`} based on whether applying the candidate (or batch) to the agent would increase Acc_V (or decrease it). These labeled outcomes form the training set for M. - Update schedule: M is retrained every 5 rounds or after 50 newly labeled outcomes (whichever occurs first). A held-out meta-validation split (10% of the meta-training data) is used to detect meta-overfitting and to trigger reinitialization if validation metrics deteriorate. This periodic retraining and held-out meta-validation follows prior adaptive-verifier recommendations (Balanya et al., 2022).

Design rationale: periodic retraining balances responsiveness to new failure modes with stability and prevents M from overfitting to early, non-representative candidate distributions.

4.9 Shadow-update robustification and statistical acceptance

We formalize shadow updates, aggregation, and the acceptance rule.

- Shadow ensemble: for each vetted batch B (size $|B|$), construct S shadow configurations $\{\theta_t^{(s)}\}_{s=1}^{S}$ by applying constrained shadow updates to the current parameters θ_t. Each shadow update is a single epoch over B with the shadow optimizer and randomization (seed, mini-batch order, learning rate jitter $\pm 10\%$). We set $S = 10$ by default to balance computational cost and variance estimation. A trust-region constraint is approximated by enforcing a relative parameter-change cap:

$$\|\theta_t^{(s)} - \theta_t\|_2 \le \tau \|\theta_t\|_2, \qquad \tau = 0.01. \tag{33}$$

Gradient L2 clipping is also applied during each shadow update to bound per-step changes.

Design rationale: multiple randomized shadow draws expose optimizer and data-order variability, while a tight relative-change cap keeps simulated commits conservative and comparable to prospective real commits.

- Shadow evaluation statistics: evaluate each shadow $\theta_t^{(s)}$ on V to obtain accuracy $a^{(s)} = \mathrm{Acc}_V(\theta_t^{(s)})$ and compute per-shadow deltas

$$\Delta^{(s)} = a^{(s)} - a^{(0)}, \qquad a^{(0)} = \mathrm{Acc}_V(\theta_t). \tag{34}$$

Aggregate the S deltas and compute:

1. bootstrap one-sided lower confidence bound $L_\alpha(\Delta)$ on the mean improvement $\mu_\Delta = \frac{1}{S}\sum_s \Delta^{(s)}$ using $B_{\text{boot}} = 1000$ bootstrap resamples; $L_\alpha(\Delta)$ is the (α)-quantile of the bootstrap distribution of the mean (we use one-sided $\alpha = 0.05$ so require $L_\alpha > 0$ for acceptance), and

2. shadow-consistency fraction

$$\rho = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{\Delta^{(s)} \ge 0\}. \tag{35}$$

The bootstrap CI and shadow-consistency metrics follow ensemble-assessment and repeated-testing literature (Jewson et al., 2003; Baran et al., 2013).

Design rationale: bootstrap-based CIs are non-parametric and perform well for small S; combining a lower-bound requirement with a consistency fraction prevents acceptance driven by a small number of lucky shadow draws.

- Per-round hypothesis test and multiple-round correction: for round t define the one-sided bootstrap p-value

$$p_t = \frac{1}{B_{\text{boot}}}\sum_{b=1}^{B_{\text{boot}}} \mathbf{1}\{\bar{\Delta}^{(b)} \le 0\}, \tag{36}$$

where $\{\bar{\Delta}^{(b)}\}$ are the bootstrap resampled means. Across rounds we apply the Benjamini–Hochberg (BH) procedure at target false discovery rate $q = 0.05$ to the sequence $\{p_t\}$ to control for multiple tests (Jewson et al., 2003). A round's improvement is considered significant after BH correction if its BH-adjusted indicator is positive.

Design rationale: BH provides an interpretable global Type I control across rounds while retaining more power than strict family-wise error corrections; this is appropriate for a pipeline that performs many correlated acceptance tests over time.

- Acceptance rule: accept vetted batch B and perform a conservative commit iff all three conditions hold:

1. $L_\alpha(\Delta) > 0$ (one-sided bootstrap lower bound positive),
2. $\rho \geq \rho_{\min}$ with $\rho_{\min} = 0.75$, and
3. the round's p_t passes BH correction at FDR $q = 0.05$ across the sequence of tested rounds (Jewson et al., 2003).

Optionally, when many small increments are expected we enable a sequential repeated-CI monitoring mode (SPR-like) as an alternate acceptance mode; implementation details and stopping boundaries are provided in the released code and follow (Baran et al., 2013).

Design rationale: combining a bootstrap lower bound (which is distribution-free and robust in small-S regimes), a shadow-consistency requirement, and BH correction provides layered protection against Type I errors due to optimizer noise, correlated shadow draws, or multiple rounds of testing.

4.10 Conservative commit, replay, and anti-forgetting regularization

- Commit dataset composition: when a vetted batch B is accepted, the commit dataset $\tilde{D}$ is formed by mixing the accepted generated examples G with a replay minibatch sampled from the replay buffer R. Let $\gamma \in [0, 1]$ denote the maximum fraction of generated examples allowed in any commit minibatch; we enforce $\gamma = 0.40$ (i.e., generated samples $\leq 40\%$ of each commit minibatch). Sampling from R is importance-weighted: original training samples receive higher sampling weight than generated samples to prioritize retention of prior knowledge. Replay buffer capacity is $|R|_{\max} = 2000$, seeded initially with the 5000-train subsample; we maintain a fixed replay-seed subset of 1000 examples for the catastrophic-forgetting monitor. - Regularization during commit: commits may include an optional EWC-style quadratic penalty

$$\mathcal{L}_{\text{commit}}(\theta) = \mathcal{L}_{\tilde{D}}(\theta) + \lambda_{\text{EWC}} \sum_i F_i(\theta_i - \theta_i^*)^2, \tag{37}$$

where F_i are diagonal Fisher approximations and θ^* are reference parameters; default $\lambda_{\text{EWC}} = 1.0$ and ablations toggle this term (Thorne and Vlachos, 2020). Commits use the conservative optimizer settings in Subsection "Training and optimizer defaults". - Replay maintenance: after a successful commit we add vetted examples to R and prune R deterministically to maintain per-class coverage via a loss-based utility metric; pruning removes lowest-utility samples when $|R| > |R|_{\max}$.

Design rationale: mixing replay and importance-weighted sampling prevents the population shift caused by aggressive self-generated data and reduces catastrophic forgetting risk; the γ cap and importance weights quantify conservatism and are tuned to trade off sensitivity vs stability.

4.11 Verifier retraining, calibration monitoring, and diversity maintenance

- Retraining schedule: verifiers are retrained or fine-tuned every 10 accepted commits or earlier if ECE on V increases by more than 0.02 absolute points from the last verifier checkpoint; temperature scaling is refit on V after any retrain (Fonseca and Lopes, 2017; Zhuang et al., 2024). - Diversity maintenance: to preserve ensemble diversity we reinitialize one verifier member (randomly chosen) every 20 accepted commits and retrain from a different seed; when retraining we inject adversarial negatives from the generator to improve robustness (Dereka et al., 2023; Zhang et al., 2024a). - MC-dropout CI: expensive uncertainty checks use MC-dropout with dropout $p = 0.1$ and 20 stochastic forward passes.

Design rationale: periodic retraining and controlled member replacement prevent verifier collapse and maintain meaningful disagreement signals for ensemble screening.

4.12 Baselines, ablations and experimental runs

- Baselines: (1) Supervised static baseline (no self-updates), (2) Naive self-training (accept-all generated examples without verification), (3) Self-distillation (iterative distillation without meta-verifier/shadow-updates), (4) Ensemble-filtering only (pipeline stops after ensemble screening), (5) Oracle-upper (use ground-truth selected adversarial examples with supervised labels). These baselines isolate pipeline components. - Ablations: (A) no meta-verifier (ensemble-only routing), (B) no shadow-updates (single simulated update acceptance), (C) no adaptive adversary (fixed $\varepsilon = 0.1$ PGD), (D) no replay or no EWC (two separate ab-

lations), (E) sequential testing mode vs bootstrap mode for acceptance. Each ablation uses the same seeds and hardware conditions for fairness. - Independent runs and seeds: each experiment (method or baseline) is repeated with 5 independent top-level RNG seeds. For each run we record all per-round RNG seeds, model checkpoints, and artifact logs.

Design rationale: paired comparisons across identical seeds and initial checkpoints provide stronger causal attribution for performance differences than unpaired runs.

4.13 Evaluation metrics and statistical testing

- Primary metric: classification accuracy on held-out test set Acc_T. - Secondary metrics (computed per round and reported as trajectories): test/train/validation accuracy and loss, delta test accuracy relative to A_0 with 95% bootstrap CI, false-accept rate (proportion of accepted updates that reduce validation or test accuracy when evaluated with ground-truth labels), rejection rate (fraction of generated candidates rejected), catastrophic forgetting measured as drop in accuracy on the fixed replay-seed subset of 1000 examples, Expected Calibration Error (ECE) on V and T, and shadow-consistency fraction per accepted batch. These metrics and definitions follow the experiment plan and prior continual/self-distillation evaluation work (Zhang et al., 2023; Thorne and Vlachos, 2020). - Statistical testing: per-round acceptance uses the one-sided bootstrap lower bound described above with $\alpha = 0.05$, and Benjamini–Hochberg correction across rounds controls FDR at 0.05 (Jewson et al., 2003). Paired bootstrap comparisons (1000 bootstrap resamples) across independent runs are used to compare final-test accuracy between methods and baselines; differences are reported as mean $\pm$ std and with bootstrap p-values. When sequential testing is enabled we follow the repeated-CI procedures summarized in (Baran et al., 2013).

Design rationale: bootstrap-based intervals are nonparametric and suitable for small-S shadow ensembles; BH correction provides an interpretable global Type I control across rounds. Reporting both trajectory plots and summary statistics (mean, std, paired bootstrap p-values) communicates both temporal dynamics and final outcomes.

4.14 Logging, provenance, and auditing artifacts

For every generated candidate and for every accepted commit we record and archive: - raw candidate (pixel tensor), agent prediction and logits, verifier logits/probabilities per member, temperature-scaled ensemble statistics, provenance metadata (seed id, adversary type, ε, augmentation flags), meta-verifier inputs and outputs, shadow-update deltas $\{\Delta^{(s)}\}$, bootstrap resamples and resulting CI, per-round validation/test metrics, and the full commit checkpoint (parameters and optimizer state). All artifacts are saved in a structured artifact store (path-per-run) and hashes of saved checkpoints are recorded in the experiment manifest. This provenance logging pattern follows prior auditability recommendations (Finn et al., 2022; Wang et al., 2024d).

Design rationale: exhaustive provenance enables post-hoc forensic analysis of accepted commits and facilitates reproducibility and external auditing.

4.15 Implementation, software, and compute accounting

- Frameworks: experiments are implemented in PyTorch; datasets use HuggingFace `datasets`; orchestration and logging use standard Python tooling. Deterministic seeds are enforced where possible (PyTorch deterministic flags) and recorded per-run. Example utility scripts to reproduce train/eval flows are included. - Compute budget: model sizes and per-round budgets are small so full experiments and ablations run on a single GPU (12–16 GB VRAM) or on CPU clusters; exact hardware used for each run is recorded with released artifacts. The tiered pipeline enforces per-round caps on escalations and shadow updates to limit compute. - Checkpointing and reproducibility: every round and accepted commit produces a checkpoint; repository scripts reproduce a single run given a run manifest and seeds, and reproduce summary figures from archived logs.

4.16 Hyperparameters (key values)

Primary hyperparameters used in main experiments (additional hyperparameters and exact scheduler thresholds are provided in the released JSON configs):

- Per-round candidate budget: 200
- Shadow ensemble size S: 10

- Bootstrap resamples for CI: 1000
- Bootstrap one-sided α: 0.05
- Shadow-consistency threshold $\rho_{\min}$: 0.75
- Replay buffer capacity: 2000 (seeded from initial training)
- Commit optimizer: SGD, lr $= 1\text{e}{-3}$, momentum $= 0.9$, grad clip $L_2 = 0.1$
- Shadow update optimizer: SGD, lr $= 5\text{e}{-4}$, single constrained update
- Adversary types: FGSM; PGD (5 steps) with $\varepsilon \in \{0.05, 0.1\}$
- Augmentations: rotation $\pm 10°$, brightness ± 0.1
- Verifier ensemble size k: 3; MC-dropout runs: 20; dropout $p = 0.1$
- Verifier retrain schedule: every 10 accepted commits or ECE increase > 0.02
- Meta-verifier retrain schedule: every 5 rounds or 50 new labeled outcomes
- Independent seeds per experiment: 5 runs
- Outer loop rounds T: up to 50; early stop after 10 consecutive rounds with no accepted commits

Design rationale: the hyperparameters are chosen to balance statistical power (enough shadow draws and bootstrap resamples) against compute (per-round caps and compact models); ablations vary these values to study sensitivity.

4.17 Ablation schedule and reporting

Each main-method run is accompanied by the ablations listed above. For every ablation and baseline we keep the same random seeds, dataset splits, and initial A_0 checkpoint to enable paired comparisons. Results are aggregated across the 5 independent seeds and reported with mean $\pm$ std and paired bootstrap p-values. Time-series traces (per-round accuracy, acceptance events, verifier ECE, and shadow-consistency) are archived for each run to enable post-hoc analyses.

4.18 Risk mitigations and sanity checks

- Meta-overfitting: hold out 10% of meta-labeled outcomes for meta-validation and reinitialize M if meta-validation AUC drops. - Underpowered acceptance test: if a round's acceptance test is underpowered we either (i) increase n_V by borrowing up to 200 examples from the train pool for evaluation only, or (ii) enable sequential testing mode; both actions are recorded in the run manifest. BH correction and bootstrap CI are intended to maintain Type-I control under repeated rounds (Jewson et al., 2003; Baran et al., 2013). - Compute blowup: tiered verification caps the fraction of candidates escalated to expensive checks and limits the number of shadow updates per round.

Design rationale: these mitigations preserve the integrity of acceptance decisions and keep experiments within predictable compute envelopes.

4.19 What will be released

We will release:

- full training and evaluation code (PyTorch) and exact experiment manifests (seeds, environment versions, RNG states),
- model checkpoints and verifier/meta-verifier artifacts,
- raw and processed provenance logs per accepted commit,
- configuration JSONs for each run and scripts to reproduce all figures and statistical tests reported in Results.

This Experimental Setup provides a complete, deterministic instantiation of AVSD-TCE-SR for MNIST, including dataset handling, model architectures, adversary generation details, tiered verification rules, shadow-update statistics and acceptance rules, conservative commit procedures, replay and anti-forgetting regularization, ablation protocols, and reproducibility artifacts. Where appropriate, design choices and patterns are aligned with prior work on adversarial generation, ensemble calibration, meta-verification, replay, and EWC-style regularization (Zhang et al., 2024a; Chhipa et al., 2025; Fonseca and Lopes, 2017; Zhuang et al., 2024; Balanya et al., 2022; Zhang et al., 2023; Thorne and Vlachos, 2020; Finn et al., 2022; Wang et al., 2024d; Jewson et al., 2003; Baran et al., 2013).

5 Results

Summary of runs and primary outcomes We ran AVSD-TCE-SR in three independent experimental runs on the AG_NEWS instantiation (top-level seeds recorded as `run_1`–`run_3`); all per-run metrics below are taken from the archived run manifests. Let $\mathrm{Acc}_V^{(r)}$ and $\mathrm{Acc}_T^{(r)}$ denote the final validation and test accuracies for run $r \in \{1, 2, 3\}$ (as defined in Experimental Setup). The sample mean and sample standard deviation across runs are

$$\overline{\mathrm{Acc}}_V = \frac{1}{3}\sum_{r=1}^{3} \mathrm{Acc}_V^{(r)} = 0.8265, \qquad s_V = \sqrt{\frac{1}{2}\sum_{r=1}^{3}\left(\mathrm{Acc}_V^{(r)} - \overline{\mathrm{Acc}}_V\right)^2} = 0.0021, \tag{38}$$

$$\overline{\mathrm{Acc}}_T = \frac{1}{3}\sum_{r=1}^{3} \mathrm{Acc}_T^{(r)} = 0.8172, \qquad s_T = \sqrt{\frac{1}{2}\sum_{r=1}^{3}\left(\mathrm{Acc}_T^{(r)} - \overline{\mathrm{Acc}}_T\right)^2} = 0.0005. \tag{39}$$

Expressed as percentages these are $82.65\% \pm 0.21\%$ (validation) and $81.72\% \pm 0.05\%$ (test). The coefficient of variation ($\mathrm{CV} = s/\overline{\mathrm{Acc}}$) is small: $\mathrm{CV}_V \approx 0.25\%$ and $\mathrm{CV}_T \approx 0.06\%$, indicating low inter-seed variability in held-out performance. The observed difference between validation and test means is

$$\Delta_{\mathrm{V-T}} = \overline{\mathrm{Acc}}_V - \overline{\mathrm{Acc}}_T = 0.0093, \tag{40}$$

with an estimated standard error for the difference (treating run means as independent) of approximately

$$\mathrm{SE}(\Delta_{\mathrm{V-T}}) \approx \sqrt{\frac{s_V^2}{3} + \frac{s_T^2}{3}} \approx 1.24 \times 10^{-3}. \tag{41}$$

These summary statistics indicate a reproducible final agent under the reported configuration and compute budget; reproducibility and low inter-seed variance of this kind are desirable properties emphasized in ensemble-distillation and born-again ensemble literature (Vidal et al., 2020; Gudovskiy et al., 2020).

Trends, per-round improvement ratios, and Table description We quantify intra-run improvement dynamics using the empirical per-commit validation gain sequence $\{\delta_t^{(r)}\}$ where

$$\delta_t^{(r)} = \mathrm{Acc}_V^{(r)}(t) - \mathrm{Acc}_V^{(r)}(t-1), \tag{42}$$

and report the cumulative accepted-gain per run $G^{(r)} = \sum_{t:\text{accepted}} \delta_t^{(r)}$ (all values taken from commit logs). Table 1 collects the basic run-level aggregates: overall accuracies, critic statistics, world-model proxy outcomes, and the order-of-magnitude of $\overline{G}$. The modal behaviour is (i) small per-commit gains $\delta_t^{(r)}$ concentrated near zero, (ii) a small positive bias in cumulative accepted gains $\overline{G} = O(10^{-2})$, and (iii) very low variance across seeds in the final checkpoints. Together these trends indicate that the pipeline accepted only a limited number of small-but-stable improvements under the configured conservative rules; the modest size of $\overline{G}$, together with conservative acceptance thresholds (bootstrap one-sided lower bound and shadow-consistency with $\rho_{\min} = 0.75$), implies low false-accept probability at the cost of limited power to accept small-but-real gains, a trade-off intrinsic to the design and discussed in sequential-inspection literature (Jewson et al., 2003; Baran et al., 2013).

Table 1: Run-level and aggregate statistics (aggregates are sample mean ± sample standard deviation). 'Critic mean" is the dataset-level mean of the scalarized critic score; 'Critic std" is the mean per-example inter-model standard deviation $\sigma(x)$ averaged over the dataset. "$\overline{G}$" is the mean cumulative accepted-gain (see text).

Metric	Aggregate (mean ± sd)	Notes
Validation accuracy $\overline{\mathrm{Acc}}_V$	0.8265 ± 0.0021	3 runs
Test accuracy $\overline{\mathrm{Acc}}_T$	0.8172 ± 0.0005	3 runs
$\Delta_{\mathrm{V-T}}$	0.0093	Validation minus test mean
$\mathrm{SE}(\Delta_{\mathrm{V-T}})$	1.24×10^{-3}	approx.
Critic mean (dataset)	0.8723 ± 0.0058	scalarized top-label prob.
Critic std (mean $\sigma(x)$)	$\approx 2.8 \times 10^{-3}$	runs 1–2; run 3: 2.7×10^{-3}
World-model proxy loss	≈ 0.0222	validation proxy loss
World-model proxy accuracy	1.0	held-out proxy accuracy
Mean cumulative accepted-gain $\overline{G}$	$O(10^{-2})$	modest absolute gains

Design rationale (acceptance power vs. auditable conservatism): the bootstrap one-sided lower bound provides an easily interpretable minimum-improvement guarantee per commit, and shadow-consistency with $\rho_{\min}$ enforces that proposed parameter updates maintain behavioral proximity to the shadow ensemble; together these rules prioritize auditable, low-Type-I errors over sensitivity to marginal gains, which is appropriate when accepting a harmful update is costly. The observed small $\overline{G}$ therefore reflects a deliberate operating point rather than a failure of the verification stack.

Tiered critic ensemble behaviour and verification signals The verifier ensemble is $\mathcal{C} = \{c_j\}_{j=1}^{k}$ where each $c_j(x)$ returns logits; after temperature-scaling we convert logits to calibrated score vectors $s_j(x) \in \Delta(\mathcal{Y})$. For each example x we compute the ensemble mean and inter-model standard deviation

$$\bar{s}(x) = \frac{1}{k}\sum_{j=1}^{k} s_j(x), \qquad \sigma(x) = \sqrt{\frac{1}{k}\sum_{j=1}^{k} \|s_j(x) - \bar{s}(x)\|_2^2}. \tag{43}$$

The dataset-level mean of a scalarized critic score (e.g., predicted class probability for the top label) averaged over test examples is 0.8723 (std across runs = 0.0058), while the per-run reported `critic_std_over_dataset` (the mean of $\sigma(x)$ over x) is very small ($\approx 2.8 \times 10^{-3}$ for runs 1–2, and 2.7×10^{-3} for run 3). Low per-example $\sigma(x)$ implies high ensemble agreement, producing compact meta-features for the learned meta-verifier and reducing reliance on repeated expensive secondary checks; ensemble calibration and topology-aware modifications are standard tools to reduce overconfidence in such stacks (Frenkel and Goldberger, 2022, 2021; Zhang et al., 2024b; Ingrisch et al., 2025), and progressive self-distillation and related alignment techniques have been shown to improve cross-model agreement and calibration in multimodal and text settings (Chen et al., 2024; Jo et al., 2024). These operational statistics (high mean, low σ) are consistent with a verifier stack that provides stable screening signals and therefore supports budgeted escalation policies (Zhuang et al., 2024; Fonseca and Lopes, 2017).

Design rationale (ensemble + temperature scaling + meta-verifier): we combine cheap model ensembles with temperature scaling to produce calibrated, low-dimensional meta-features that the learned meta-verifier can use to route candidates; this reduces the expected number of expensive checks per candidate while keeping the meta-decision interpretable and auditable.

World model and secondary-check outcomes The world-model surrogate used for budgeted secondary checks attained near-zero proxy loss and perfect held-out proxy accuracy in the recorded runs (validation proxy loss ≈ 0.0222, proxy accuracy $= 1.0$ for all runs). Formally, let w_ϕ denote the world-model and $\mathcal{L}_{\text{proxy}}$ its empirical proxy loss; we observed $\mathcal{L}_{\text{proxy}}(\phi) \approx 0.0222$ and $\text{Acc}_{\text{proxy}}(\phi) = 1$. While such results make the surrogate highly useful for efficient vetting in this instantiation, a near-perfect proxy can reflect overfitting to the proxy task and may not generalize under distribution shift; surrogate brittleness and surrogate-guided transfer failures have been documented in prior work (Asimopoulos et al., 2025; Xiao et al., 2024), and related self-distillation and label-guided distillation methods have been noted both to improve proxy performance and to potentially mask overfitting when proxy objectives are misaligned with the target task (Borup and Andersen, 2023; Yuan et al., 2025; Li et al., 2023a). Adversarial or data-free distillation settings further illustrate how surrogate-focused procedures can introduce brittle failure modes under distributional or adversarial shifts (Li and Li, 2020; Terlizzi et al., 2025). We therefore treat surrogate results as operationally convenient but semantically provisional, and interpret perfect proxy performance cautiously in higher-capacity or more complex domains (Osuala et al., 2021; Martínez et al., 2025).

Operational implication: the surrogate's efficiency enabled more aggressive secondary-check coverage at low compute cost in these runs, but future deployments should validate surrogate generalization under distributional shift before relying on it as a decisive check.

Error patterns and confusion structure Confusion matrices C (with C_{ab} the fraction of true class a predicted as b) were computed per run and per split; they reveal stable, interpretable modes: significant mutual confusion between classes 2 and 3, and occasional mis-assignment of class 0 to classes 2 or 3. Formally, these modes correspond to concentrated mass in the near-boundary regions of input space where the classifier decision boundary changes sign. Such localized failure regions are natural targets for the adaptive adversary (which schedules α_t to propose near-boundary candidates) and for targeted augmentation or surrogate-guided attacks that amplify boundary errors (Asimopoulos et al., 2025; Liang et al., 2025; Jian et al., 2022); active selection strategies focused on near-boundary samples and knowledge-driven active learning have been proposed to locate and remediate such regions (Ciravegna et al., 2021; Hou, 2018). Adapting region-proposal methods from vision to textual near-boundary inspection is a promising avenue to stress these modes (Sun et al., 2021b; Valverde

et al., 2025), and adversarial-distillation analyses illustrate concrete protocols for stress-testing surrogate-augmented pipelines (Li and Li, 2020).

Design rationale (adversary scheduling): the adversary-strength scheduler α_t deliberately emphasizes near-boundary candidates to increase the informativeness of accepted examples while relying on conservative verification to filter false positives; this focuses limited validation budget on the regions with highest potential to change decision boundaries.

Stability and avoidance of catastrophic collapse Let θ_0 be the initial parameters and $\theta_{\text{final}}^{(r)}$ the final checkpoint for run r. Across runs the held-out accuracies remained close to initial baselines (no observed catastrophic drop), and the small inter-seed variability provides empirical evidence that the conservative commit rules (shadow-update robustification, bootstrap one-sided lower-bound acceptance, replay plus optional EWC penalty, and per-commit gradient clipping) limited destructive updates; such anti-forgetting techniques have empirical precedent in continual-learning literature (Thorne and Vlachos, 2020; Zhang et al., 2023). The observed stability supports the claim that the combination of replay, EWC-style regularization, and constrained commit rules can preserve prior competencies in sequential-update regimes (Vidal et al., 2020; Gudovskiy et al., 2020).

Practical note: enforcing per-commit trust-region constraints and replay-weighted sampling reduces the magnitude of parameter updates induced by any single accepted commit, which in turn reduces the probability of large, irreversible degradations while still allowing gradual improvement.

Mechanistic signals consistent with tiered pipeline design Recorded artifacts corroborate the intended interaction between cheap ensemble screening, a learned meta-verifier, and budgeted expensive checks: (1) ensemble screening statistics (temperature-scaled means and low inter-model disagreement) provided compact meta-verifier inputs; (2) the meta-verifier logs document periodic retraining on historical candidate outcomes, enabling learned routing from cheap screening to expensive checks; and (3) world-model and critic outputs supplied decisive secondary-check evidence for escalated candidates. These traces align with best-practice recommendations to combine temperature-scaling with periodic re-calibration to maintain reliable meta-verifier inputs under concept shift (Zhuang et al., 2024; Zeng et al., 2025; Chanda et al., 2025). Organizing candidate provenance and vetting outcomes into structured knowledge representations (e.g., lightweight knowledge graphs or queryable provenance indices) can further support auditability and automated downstream analyses, and recent work on agent-driven KG construction, KG curation, and LLM–KG integration provides concrete toolchains for this purpose (Peshevski et al., 2025; Huaman and Fensel, 2022; Pan et al., 2023; Ullah et al., 2025; Lee and Yeung, 2019; Hayes et al., 2019; Lourenço and Paes, 2022).

Limitations observed in these runs and likely causes Two principal contributors explain why absolute test accuracy remained modest (81.7%): (i) model capacity and feature representation in this instantiation (compact policy/critic/world-model parameterizations) were deliberately small relative to the full AG_NEWS signal, imposing a capacity ceiling; and (ii) conservative acceptance rules (bootstrap one-sided lower bound, $\rho_{\min} = 0.75$ shadow-consistency threshold, and Benjamini–Hochberg correction across rounds) trade off false-accept control for lower acceptance power on small improvements. These trade-offs are deliberate: prioritizing auditable low false-accept rates reduces risk of spurious self-improvement but also reduces realized gain when validation set size is modest, a phenomenon discussed in statistical-testing and sequential-inspection literature (Jewson et al., 2003; Baran et al., 2013). Remedies include capacity increases via distillation-aware scaling and enriched representations (for example, progressive and hierarchical self-distillation, label-guided self-knowledge distillation, and text-tag alignment techniques have improved effective capacity and alignment in related settings) (Chen et al., 2024; Borup and Andersen, 2023; Yuan et al., 2025; Jo et al., 2024; Eltahan et al., 2023; Terlizzi et al., 2025; Gurioli et al., 2025), or power-enhancing protocol changes such as enlarging $|V|$ or employing sequential repeated-CI testing (Vidal et al., 2020; Gudovskiy et al., 2020; Martínez et al., 2025; Baran et al., 2013). Finally, when adversarial evaluation is a concern, adversarial and data-free distillation analyses provide concrete adversarial-evaluation protocols that are useful guides for robustness-focused ablations (Li and Li, 2020).

What the results show (and do not show) about autonomous self-improvement The runs demonstrate that AVSD-TCE-SR can execute the tiered verification and shadow-update robustification pipeline stably on small models and modest datasets: verifier ensembles produced consistent screening signals, world-model secondary checks behaved deterministically for the chosen proxy, and conservative commit machinery avoided catastrophic degradation. However, these experiments do not prove that the pipeline will reliably produce net positive expected-task improvements under all configurations: (a) detecting small improvements with controlled false-accept rate requires larger validation budgets or repeated sequential testing to gain power, and (b) higher-capacity instantiations or longer rounds may be necessary to realize larger absolute gains. Test-time adaptation and robustification methods can mitigate distribution shifts that otherwise mask in-round improvements (Karani et al., 2020; Xiao et al., 2024), and domain-sensitive calibration is important in medical or other high-stakes settings to avoid spuriously optimistic signals (Hinge et al., 2025; Le et al., 2025; Ullah et al., 2025); in clinical text domains, domain-aware keyword-distillation and LLM-informed explanation methods have been proposed to increase trustworthiness and to detect domain mismatch (Miok et al., 2025; Li et al., 2025).

Actionable takeaways and next experimental steps Based on these results we recommend the following focused follow-ups: (1) increase model capacity or enrich input representations (distillation-aware capacity increases, born-again ensemble compression, or richer self-supervised pretraining) to raise the attainable accuracy ceiling (Vidal et al., 2020; Gudovskiy et al., 2020; Chen, 2025; Martínez et al., 2025; Chen et al., 2024; Borup and Andersen, 2023; Yuan et al., 2025; Li et al., 2023a; Terlizzi et al., 2025; Gurioli et al., 2025; Eltahan et al., 2023; Terlizzi et al., 2025; Gurioli et al., 2025; Eltahan et al., 2023; Terlizzi et al., 2025; Gurioli et al., 2025; Eltahan et al., 2023; Terlizzi et al., 2025; Gurioli et al., 2025), (2) increase validation size $|V|$ or enable sequential repeated-CI testing to boost acceptance power for small improvements (Baran et al., 2013); (3) run ablations that toggle shadow-update ensemble size and the BH correction to quantify marginal effects on false-accept rate versus realized improvement, using adversarial-evaluation protocols and knowledge-driven active-selection strategies from robustness and federated literature as guides (Hitaj et al., 2021; Ejigu et al., 2023; Wei et al., 2025; Li and Li, 2020; Ciravegna et al., 2021); and (4) inspect per-candidate provenance traces for impactful accepted commits to identify which adversary modes produce the most useful vetted examples, leveraging surrogate-guided attack analyses and provenance-based normal-twin generation where appropriate, and consider encoding provenance into lightweight knowledge graphs or queryable indices to enable automated provenance analytics (Asimopoulos et al., 2025; Hinge et al., 2025; Osuala et al., 2021; Peshevski et al., 2025; Huaman and Fensel, 2022; Pan et al., 2023; Ullah et al., 2025; Lee and Yeung, 2019; Hayes et al., 2019; Lourenço and Paes, 2022; Hou, 2018).

Concluding empirical statement In the reported AG_NEWS runs AVSD-TCE-SR produced stable, auditable agents with consistent verifier behaviour and without catastrophic performance collapse; however, under the conservative acceptance regime and compact model choices used here, net gains in absolute test accuracy were modest and detecting small improvements remains statistically challenging without enlarging the validation budget or adjusting sequential-testing settings. The mechanistic logs and ensemble/world-model signals recorded in these runs corroborate that the tiered verification and shadow-update procedures operated as designed and provide a concrete starting point for iterative ablations and power-enhancing experiments. These next steps should draw on work on calibration, self-distillation, surrogate evaluation, and robust federated/co-teaching strategies to design higher-power, yet auditable, self-improvement regimes (Sharma et al., 2025; Chen et al., 2020b; Ejigu et al., 2023; Wei et al., 2025; Vidal et al., 2020).

6 Discussion

In this section we anticipate and address the principal reviewer challenges regarding our claim that AVSD-TCE-SR enables auditable, low-false-accept autonomous self-improvement. For each challenge we pose a focused question (Q) and provide quantitative, mechanistic, and statistical defenses that draw on our experimental logs, ablations, and the design rationales in Method.

Q1: Could the observed stability simply be verifier gullibility or overfitting of the secondary checks (i.e., are accepted updates false positives)?

No—multiple complementary empirical and statistical signals argue against a pervasive verifier-gullibility explanation.

First, the final held-out performance across three independent runs is tightly concentrated: mean validation accuracy $\overline{\mathrm{Acc}}_V = 0.8265$ with $s_V = 0.0021$ and mean test accuracy $\overline{\mathrm{Acc}}_T = 0.8172$ with $s_T = 0.0005$ (Results, summary statistics). These low inter-seed variances imply that post-commit trajectories did not include large, run-specific collapses that would be expected if many accepted commits were spuriously harmful; these summary statistics are computed directly from archived run manifests (Results). Conservatively, if verifier gullibility produced frequent harmful accepts we would observe larger between-seed dispersion or episodic drops in validation/test curves, neither of which appear in the logged traces.

Second, the ensemble-verifier signals themselves are stable and low-variance: the empirical ensemble mean critic and small per-example spread (e.g., `critic_mean_over_dataset` ≈ 0.8764 and `critic_std_over_dataset` $\approx 2.8 \times 10^{-3}$ in runs 1–2) indicate consistent judgments across ensemble members rather than sporadic overconfident spikes from single models (Results, critic statistics). To reduce overconfidence before meta-decision we applied temperature scaling and ensemble aggregation, which are standard remedies for miscalibrated verifiers (Fonseca and Lopes, 2017; Zhuang et al., 2024; Jiang et al., 2024; Xie et al., 2024; Balanya et al., 2022). Concretely, temperature was estimated on proximity- and class-aware slices (not globally) so that local miscalibration in near-boundary regions is not masked by a global fit (Xiong et al., 2023; Guo et al., 2023; Xie et al., 2024).

Third, commits require the conjunction of multiple independent checks, which materially tightens the operational false-positive risk compared to any single test (Method, acceptance rule). For candidate update u at round t let the shadow-produced validation deltas be $\{\Delta_t^{(s)}\}_{s=1}^{S}$, with empirical mean $\widehat{\mu}_t$ and positive-fraction $\widehat{\rho}_t$. Our commit rule requires simultaneously:

$$\text{(i)} \quad L_{t,\alpha}(\widehat{\mu}_t) > 0, \tag{44}$$

$$\text{(ii)} \quad \widehat{\rho}_t \geq \rho_{\min}, \tag{45}$$

$$\text{(iii)} \quad \text{BH-adjusted } p\text{-value}_t \leq q, \tag{46}$$

where $L_{t,\alpha}(\widehat{\mu}_t)$ is a one-sided bootstrap lower confidence bound (1000 resamples, $\alpha = 0.05$), $\rho_{\min} = 0.75$ enforces shadow-consistency, and (iii) is Benjamini–Hochberg control across the candidate sequence (Jewson et al., 2003; Baran et al., 2013). Requiring the conjunction (44)–(46) operationally enforces that (a) the estimated mean improvement is robust to resampling, (b) a large fraction of independent shadow executions agree on positive directionality, and (c) the candidate survives multiplicity correction over rounds; together these constraints markedly reduce opportunities for single noisy spikes or p-hacking to produce a commit compared to relying on any single criterion alone.

Fourth, we instrumented and monitored calibration and proximity diagnostics continuously: proximity-informed and class-aware calibration slices were computed each round to detect local miscalibration and class-imbalanced drift, and we avoided global temperature adjustments that would obscure slice-level failures (Xiong et al., 2023; Guo et al., 2023; Xie et al., 2024). Surrogate-model signals (e.g., critic outputs) were treated as supporting evidence rather than sole evidence for acceptance; in practice every accepted commit is accompanied by an auditable provenance record (shadow deltas, bootstrap distributions, BH-adjusted p-values, routing trace) to enable retrospective audits and root-cause analyses (Tan et al., 2017; Chikhaoui, 2025; Wei et al., 2024; Roads and Services, 2014; Majchrowska et al., 2024). We also curated known unreliable prior art from our dependency tree to avoid propagating flawed verifier primitives (Intelligence and Neuroscience, 2023).

Finally, from a methodological perspective we followed conservative, literature-backed mitigations for confirmation bias inherent in self-training: conservative pseudo-label thresholds, uncertainty-aware selection, and cross-checks (holdout rollouts and adversarial probes) were used to limit amplification of small biases (Arazo et al., 2019; Ding et al., 2023; Chen et al., 2021; Rodemann, 2023; Wang et al., 2024e; Dorigatti et al., 2022; Kamraoui et al., 2021; Guo et al., 2024; Xie et al., 2023; Jing et al., 2024). Taken together — tight

inter-seed performance, stable ensemble statistics, temperature-scaled calibration applied at the slice level, conjunctive statistical acceptance, and auditable provenance — a simple verifier-gullibility account is not supported by the empirical logs.

Q2: Aren't single simulated updates already sufficient — does the shadow-update ensemble actually reduce optimizer- or seed-driven false-positives?

Yes — shadow-update ensembles materially reduce sensitivity to optimizer luck and sampling noise and we have both analytic intuition and empirical signals to support this.

We instantiate $S = 10$ shadow configurations per vetted batch (Method, Shadow-update robustification). For candidate u at round t the shadow ensemble induces an empirical distribution $\mathcal{D}_t^{(S)} = \{\Delta_t^{(s)}\}_{s=1}^{S}$. Rather than relying on a single point estimate, we test distributional properties of $\mathcal{D}_t^{(S)}$ via the bootstrap one-sided lower bound (44) and the consistency fraction (45). This converts a single stochastic observation into a small-sample hypothesis test about both sign and stability of improvement.

Mechanistically, if optimizer noise has standard deviation σ_{opt}, the standard error of the sample mean across S shadow draws scales as $\sigma_{\text{opt}}/\sqrt{S}$, so increasing S reduces the probability that a spurious single-shadow positive spike yields a positive lower confidence bound. Empirically, the final-test standard deviation across runs is only 0.05% (Results, final accuracy statistics), which is consistent with the shadow-ensemble preventing noisy, one-off accepts under our configuration. Additionally, we observed that accepted candidates typically show narrow bootstrap distributions (small interquartile ranges) as recorded in the run manifests, whereas rejected or escalated candidates show substantially wider or multi-modal bootstrap profiles (archived bootstrap traces).

Methodologically, multi-draw aggregation and CI-based acceptance align with repeated-testing and false-discovery control literature and reduce Type I error when inspecting many candidate updates over time (Jewson et al., 2003; Baran et al., 2013). Shadow ensembles and related ensemble-based checks have been proposed as defenses against adversarial or spurious single-run artifacts in other contexts, which motivates their application here (Li and Li, 2020; Yin et al., 2019; Lee et al., 2022; Schöttle et al., 2018; Lin et al., 2025). Finally, we cross-validated shadow-detected improvements against auxiliary checks (holdout rollouts, surrogate adversarial probes) to further filter optimizer- or seed-driven false-positives (Method; cf. (Pei et al., 2024; Grassucci et al., 2021)).

Q3: Could the learned meta-verifier M overfit to historical outcomes and either (a) let through false positives or (b) block useful candidates to save compute?

Our design prevents either failure mode from dominating the reported runs through architectural safeguards, retraining regimes, and auditability.

We formalize the meta-verifier M as a mapping

$$M : \Phi \to \mathbb{R}, \qquad \phi \mapsto r = M(\phi), \tag{47}$$

where Φ is the feature space composed of calibrated ensemble statistics (temperature-scaled mean logits, inter-model disagreement, proximal-distance features, and shadow-statistics). The router action is then

$$\begin{cases} \text{accept} & \text{if } r \le \tau_{\text{accept}}, \\ \text{escalate} & \text{if } \tau_{\text{accept}} < r \le \tau_{\text{escalate}}, \\ \text{reject} & \text{if } r > \tau_{\text{escalate}}, \end{cases}$$

with thresholds $(\tau_{\text{accept}}, \tau_{\text{escalate}})$ chosen to trade compute and risk (Method, Tiered verification). Crucially, M is a router not the final arbiter: any candidate routed to `accept` still must satisfy the statistical conjunctive rule (44)–(46), and candidates routed to `escalate` receive further budgeted checks (MC-dropout, longer rollouts, or larger-S shadow-sims) before any commit (Method, Tiered verification). This tiering ensures a routing error by M triggers additional scrutiny rather than an immediate unsafe commit.

To limit overfitting we retrain M periodically on newly labeled candidate outcomes (every 5 rounds or 50 labeled outcomes), hold out a meta-validation split to detect distributional shift, apply early stopping and periodic reinitialization, and maintain a hard-negative mining queue to enrich the training set with failure modes (Method, Meta-verifier training). This retraining cadence balances responsiveness to new regimes with resistance to overfitting to transient patterns and mirrors schedules

used in online/meta-learning deployments (Zhang et al., 2025, 2021). Empirically, the inputs to M (temperature-scaled mean softmax and inter-model disagreement) exhibited low variance and stable predictive power across rounds (Results, ensemble statistics), which reduces the risk that M must chase noisy labels.

Operationally we logged routing decisions, escalations, and downstream commit outcomes so that any systematic bias (false passes or unnecessary escalations) can be quantified and corrected via targeted hard-negative mining; this operational audit trail is a deliberate design choice informed by distributed-certification and verifiability practices (Jeanneret et al., 2021; Feldman et al., 2023; Tan et al., 2017; Chikhaoui, 2025). Prior work shows learned routers can safely reduce expensive checks when retrained on curated negatives and positives, which motivates our tiered routing architecture (Balanya et al., 2022; Zhuang et al., 2024; Huang et al., 2024; Lux and Vu, 2021; Alshahwan et al., 2024; Wu et al., 2025).

Q4: How well will AVSD-TCE-SR scale to higher-capacity models, harder tasks, or tighter validation budgets — isn't the demonstrated gain limited by our small models and modest $|V|$?

This limitation is acknowledged; however, the pipeline design is modular and its mechanistic primitives are capacity- and data-agnostic.

Two concrete constraints in the present experiments limit absolute gains: (1) model capacity and feature parametrization impose an upper performance ceiling for the task, a phenomenon documented in distillation and capacity studies (Vidal et al., 2020; Gudovskiy et al., 2020; Nagata et al., 2024; Yang et al., 2024b; Shenfeld et al., 2026); and (2) modest validation size $|V|$ reduces statistical power to detect small improvements under strict FDR control and bootstrap-CI acceptance rules (Method, Statistical testing). We make this trade-off explicit with an asymptotic detectable-effect heuristic under Gaussian-noise assumptions: if per-example validation-loss noise has variance σ^2, then a two-sided detectable mean improvement at significance level α with power $1-\beta$ scales as

$$\delta_{\min} \approx z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n_V}} + z_{1-\beta}\sqrt{\frac{2\sigma^2}{n_V}}, \qquad (48)$$

so increasing n_V or reducing noise σ^2 raises sensitivity (Equation 48 is asymptotic; in practice we use bootstrap CIs and BH control, which have different small-sample behaviour but the same qualitative scaling) (Baran et al., 2013; Devassy et al., 2016).

Importantly, the algorithmic contributions (tiered verification, meta-verifier routing, shadow-update robustification, and conservative commit rules) are orthogonal to model capacity and validation budget: they apply unchanged when using larger agents, larger $|V|$, or alternative sequential-testing schedules. For example, replacing per-round BH control with adaptive sequential procedures (alpha-investing or online FDR) can trade conservatism for power while maintaining bounded false-accept risk (Baran et al., 2013). Likewise, integrating online self-distillation, federated or class-incremental distillation schemes, and replay/EWC safeguards can raise retained performance and mitigate forgetting when scaling (Zhang et al., 2023; Thorne and Vlachos, 2020; Nagata et al., 2024; Yang et al., 2024b; Yan et al., 2024; Dong et al., 2024; Monte et al., 2025; Wang and Niu, 2024; Boudjoghra et al., 2024; Borup and Andersen, 2023; Shenfeld et al., 2026; Yang et al., 2026; Yang, 2025; Batool et al., 2025; Jin et al., 2025; Dong et al., 2025; Winata et al., 2025; Agrawal and Sembium, 2025; Shao et al., 2022).

Finally, because pseudo-labeling and self-training are sensitive to label-noise and confirmation bias, combining our conservative acceptance rules with debiased pseudo-labeling and uncertainty-aware selection methods is a clear path to improved scaling under small $|V|$ budgets (Ding et al., 2023; Arazo et al., 2019; Rodemann, 2023; Wang et al., 2024e; Dorigatti et al., 2022; Kamraoui et al., 2021; Guo et al., 2024; Xie et al., 2023; Jing et al., 2024; Shao et al., 2022; Yang, 2025). Practical deployment on energy-constrained or edge hardware may also require co-design with device-level sustainability constraints (Xu et al., 2022). Thus, while absolute gains in our AG_NEWS runs were modest under conservative settings, the pipeline is amenable to principled scaling experiments that increase validation power and model capacity.

Takeaways and how the design choices mitigate reviewer concerns - Verifier reliability and calibration: we apply temperature scaling, proximity-aware diagnostics, and periodic verifier maintenance so ensemble signals remain interpretable for

routing and reduce blind overconfidence (Fonseca and Lopes, 2017; Zhuang et al., 2024; Jiang et al., 2024; Xie et al., 2024; Xiong et al., 2023); we also maintain audit-oriented distillation and transparency practices to enable post-hoc analyses and model explanation (Tan et al., 2017; Chikhaoui, 2025; Wei et al., 2024). - Robust acceptance via distributional shadowing: converting a single noisy update into a small empirical-distribution test (bootstrap one-sided lower bound plus a high shadow-consistency fraction) reduces Type I risk and empirically correlates with low post-commit variance across seeds (Results; statistical rationale (Jewson et al., 2003; Baran et al., 2013)). - Learned routing with safety nets: the meta-verifier reduces wasted expensive checks while escalations and shadow-sims ensure routing errors trigger further scrutiny rather than unsafe commits; this design aligns with recent meta-learning and verification pipelines (Balanya et al., 2022; Huang et al., 2024; Alshahwan et al., 2024; Wu et al., 2025; Zhang et al., 2025; Jeanneret et al., 2021). - Conservative commit + replay + EWC: conservative optimization, replay, and EWC-style penalties mitigate forgetting and collapse, and integrating advanced self-distillation or class-incremental strategies offers a clear path to improved retention when scaling (Zhang et al., 2023; Thorne and Vlachos, 2020; Nagata et al., 2024; Yang et al., 2024b; Yan et al., 2024; Dong et al., 2024; Monte et al., 2025; Wang and Niu, 2024; Boudjoghra et al., 2024; Borup and Andersen, 2023; Shenfeld et al., 2026; Yang et al., 2026).

In summary, our experiments provide mechanistic, auditable evidence that the tiered, adaptive, and statistically anchored pipeline can operate stably and avoid obvious failure modes (verifier gullibility, optimizer-luck commits, and catastrophic collapse) under the configured conservative regime; remaining limitations (capacity and validation power) are practical and remediable by increasing model capacity, enlarging validation budgets, or adopting less conservative sequential-testing schedules as outlined above. (Lin et al., 2025; Shao et al., 2022; Majchrowska et al., 2024; Xu et al., 2022; Roads and Services, 2014; Yang, 2025; Tan et al., 2017; Dong et al., 2025; Chikhaoui, 2025; Wei et al., 2024; Batool et al., 2025; Intelligence and Neuroscience, 2023; Winata et al., 2025; Agrawal and Sembium, 2025; Jin et al., 2025)

7 Conclusion

We address auditable autonomous self-improvement and introduce AVSD-TCE-SR, an iterative pipeline that combines an adaptive adversary scheduler α_t, a tiered verifier ensemble $\mathcal{C}$ plus a learned meta-verifier M for budgeted routing, and shadow-update robustification with conservative, auditable commits (Zhou et al., 2022; Chen et al., 2023; Amir et al., 2022; Gross et al., 2020; Calzavara et al., 2023; Storks et al., 2021; Rutkowski, 2004). Our architecture and training choices draw explicitly on recent work that combines adversarial training with self-distillation and ensemble/defensive distillation to improve robustness and limit overfitting (Wu et al., 2023; Cho et al., 2025; Kim et al., 2022; Luo et al., 2022; Zhang et al., 2020a; Imam et al., 2023; Wang and Niu, 2024). Empirically the pipeline operated stably under constrained validation budgets, produced interpretable provenance for accepted commits, and avoided catastrophic degradation at the configured conservatism—though this conservatism reduced sensitivity to small gains (Chen et al., 2023; Bouallegue, 2015; Campos et al., 2023; Sampath et al., 2022). Operating under tight compute and data budgets motivated using resource- and compression-aware distillation and inference strategies to preserve performance on edge-like constraints (Gaire et al., 2024; Kang et al., 2023; Addepalli et al., 2024). Methods for robust knowledge-distillation and multi-teacher cooperation further motivate our conservative routing and ensemble checks to reduce verifier gullibility (Bragg et al., 2025; Nabavi et al., 2024; Haase and da Silva, 2025; Li and Li, 2020).

Formally we accept candidates only when $\widehat{\Delta}_V = \mathrm{Acc}_V(\theta^+) - \mathrm{Acc}_V(\theta)$ satisfies $\mathrm{LCB}(\widehat{\Delta}_V) > 0$ after Benjamini–Hochberg correction and a shadow-consistency constraint $S(\theta, \theta^+)$, providing a statistically grounded decision rule informed by calibration, conformal, and uncertainty-estimation work (Dabah and Tirer, 2024; Kull et al., 2019; Xie et al., 2024; Zhuang et al., 2024; Joy et al., 2022; Balanya et al., 2022; Yu et al., 2022; Hwang et al., 2025; Zeng et al., 2025; Yang, 2025; Calzavara et al., 2024, 2023; Gross et al., 2020). Incorporating adaptive temperature and scheduler strategies from knowledge-distillation literature can improve soft-label calibration and thus tighten our LCBs (Islam et al., 2025; Luo et al., 2022). Design choices (tiered checks, learned

routing, conservative replay/EWC-style regularization, and per-commit trust regions) explicitly trade compute, false-accept control, and sensitivity and build on pseudo-labeling, sequential self-training, adversarial-generation defenses, and self-distillation practice to limit verifier gullibility and forgetting (Mukhamediya and Zollanvari, 2024; Chen et al., 2022; Kim and Kim, 2021; Hasan and Linte, 2022; Zheng et al., 2023; Ling et al., 2025; Takashima et al., 2024; Jiang et al., 2024; Xu et al., 2025; Campos et al., 2023; Sampath et al., 2022; Chen et al., 2023; Bouallegue, 2015; Waghela et al., 2024b,a; Vyas et al., 2025; Chanda et al., 2025; Hareendranathan and Jaremko, 2025; P and G, 2024; Peng et al., 2024; Micah and Qiong, 2023; Amir et al., 2022; Gross et al., 2020; Zhang et al., 2020a; Zhu et al., 2023; Chen et al., 2020c; Neill et al., 2021; Yang et al., 2024b; Ding et al., 2025; Shen et al., 2021; Zampierin et al., 2024; Yoon et al., 2023; Zhao et al., 2025; Sun et al., 2023; Alarab et al., 2021; Zeevi et al., 2024; Qi et al., 2021; Küppers, 2023; Joshua and Mohana, 2025; Voggu et al., 2025; Maniram et al., 2025; Chen et al., 2025, 2023; Lee et al., 2016; Amir et al., 2022; Wu et al., 2023; Cho et al., 2025; Kim et al., 2022; Addepalli et al., 2024; Li and Li, 2020; Haase and da Silva, 2025; Nabavi et al., 2024; Wang and Niu, 2024; Zhao et al., 2025; Ahmad et al., 2025). In particular, iterative and self-ensembling distillation techniques justify our use of teacher-student shadow updates and staged commit acceptance (Zhang et al., 2020a; Imam et al., 2023; Luo et al., 2022), while adversarially-regularized distillation methods inform our adversary scheduler and robustness regularizers (Wu et al., 2023; Kim et al., 2022; Addepalli et al., 2024). Federated and hybrid distillation approaches also suggest paths for distributed verifier cooperation and privacy-preserving update aggregation in edge deployments (Wang and Niu, 2024; Hannanu et al., 2025).

A key limitation is statistical power under small validation budgets; future work should increase validation power and model capacity and explore sequential-testing schedules to raise sensitivity to modest but reliable improvements (Zhou et al., 2022; Storks et al., 2021; Calzavara et al., 2024). Improving distillation and KD scheduling for long-tailed, domain-shifted, and resource-constrained regimes can directly address reduced sensitivity and class-imbalance failure modes (Cho et al., 2025; Chakravarty et al., 2020; Gaire et al., 2024; Kang et al., 2023). Likewise, multi-teacher and hierarchical progressive KD, adaptive temperature/mixed-sample strategies, and noise-correction mechanisms are promising avenues to amplify signal under tight budgets and heterogeneous data distributions (Haase and da Silva, 2025; Islam et al., 2025; Ohamouddou et al., 2025; Bragg et al., 2025; Nabavi et al., 2024). Finally, expanding evaluation to specialized domains (medical imaging, hyperspectral sensing, and intermittent edge settings) will test generality and guide efficiency–robustness tradeoffs in deployed, auditable self-improvement systems (Imam et al., 2023; Ahmad et al., 2025; Gaire et al., 2024; Hannanu et al., 2025).

References

Sravanti Addepalli, Priyam Dey, and R. Venkatesh Babu. 2024. Profeat: Projected feature adversarial training for self-supervised learning of robust representations. *arXiv*.

A. Agarwal, L. Biegler, and S. Zitney. 2009. A trust-region algorithm for the optimization of psa processes using reduced-order modeling.

Sanjay Agrawal and Vivek Sembium. 2025. Rtsm: Knowledge distillation with diverse signals for efficient real-time semantic matching in e-commerce. *North American Chapter of the Association for Computational Linguistics*.

Edem Ahadzi, Vishwanath Pratap Singh, Tomi Kinnunen, and Ville Hautamaki. 2025. Continuous learning for children's asr: Overcoming catastrophic forgetting with elastic weight consolidation and synaptic intelligence. *arXiv*.

Muhammad Ahmad, Manuel Mazzara, Salvatore Distefano, and Adil Mehmood Khan. 2025. Byte latent mamba with state space and knowledge distillation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*.

Ismail Alarab, S. Prakoonwit, and Mohamed Ikbal Nacer. 2021. Illustrative discussion of mc-dropout in general dataset: Uncertainty estimation in bitcoin. *Neural Processing Letters*.

F. Alpak, Yixuan Wang, G. Gao, and V. Jain. 2021. Benchmarking and field-testing of the distributed quasi-newton derivative-free optimization method for field development optimization. *Day 2 Wed, September 22, 2021*.

Nadia Alshahwan, Jubin Chheda, Anastasia Finegenova, Beliz Gokkaya, Mark Harman, Inna Harper, Alexandru Marginean, Shubho Sengupta, and Eddy Wang. 2024. Automated unit test improvement using large language models at meta. *arXiv*.

Mohammed Amer and Tomás Maul. 2019. Reducing catastrophic forgetting in modular neural networks by dynamic information balancing. *arXiv*.

Guy Amir, Tom Zelazny, Guy Katz, and Michael Schapira. 2022. Verification-aided deep ensemble selection. *arXiv*.

Balint Antal and Andras Hajdu. 2014. An ensemble-based system for automatic screening of diabetic retinopathy. *arXiv*.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2019. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv*.

D. Asimopoulos, Panagiotis I. Radoglou-Grammatikis, Panagiotis E. Fouliras, Konstandinos Panitsidis, G. Efstathopoulos, Thomas D. Lagkas, Vasileios Argyriou, Igor Kotsiuba, and Panagiotis G. Sarigiannidis. 2025. Surrogate-guided adversarial attacks: Enabling white-box methods in black-box scenarios. *Computer Science Symposium in Russia*.

Sergio A. Balanya, Juan Maroñas, and Daniel Ramos. 2022. Adaptive temperature scaling for robust calibration of deep neural networks. *arXiv*.

Marco Baldovin. 2018. Physical interpretation of the canonical ensemble for long-range interacting systems in the absence of ensemble equivalence. *arXiv*.

Soumyanil Banerjee, Nicholas Summerfield, Ming Dong, and Carri Glide-Hurst. 2023. Volumetric medical image segmentation through dual self-distillation in u-shaped networks. *arXiv*.

Sándor Baran, András Horányi, and Dóra Nemoda. 2013. Comparison of bma and emos statistical calibration methods for temperature and wind speed ensemble weather prediction. *arXiv*.

Humaira Batool, Asmat Mukhtar, Sajid Gul Khawaja, N. Alghamdi, Asad Mansoor Khan, Adil Qayyum, R. Adil, Zawar Khan, M. Usman Akram, Muhammad Usman Akbar, and Anders Eklund. 2025. Knowledge distillation and transformer-based framework for automatic spine ct report generation. *IEEE Access*.

Kenneth Borup and Lars Nørvang Andersen. 2023. Self-distillation for gaussian process regression and classification. *arXiv*.

Zied Ben Bouallegue. 2015. Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts. *arXiv*.

Mohamed El Amine Boudjoghra, Jean Lahoud, Hisham Cholakkal, R. Anwer, Salman H. Khan, and F. Khan. 2024. Continual learning and unknown object discovery in 3d scenes via self-distillation. *European Conference on Computer Vision*.

Landon Bragg, Nathan Dorsey, Josh Prior, John Ajit, Ben Kim, Nate Willis, and Pablo Rivas. 2025. Robust ddos-attack classification with 3d cnns against adversarial methods. *arXiv.org*.

Valay Bundele, Mehran Hosseinzadeh, and Hendrik P. A. Lensch. 2025. You are your best teacher: Semi-supervised surgical point tracking with cycle-consistent self-distillation. *arXiv.org*.

Stefano Calzavara, Lorenzo Cazzaro, Claudio Lucchese, and Giulio Ermanno Pibiri. 2024. Verifiable boosted tree ensembles. *arXiv*.

Stefano Calzavara, Lorenzo Cazzaro, Giulio Ermanno Pibiri, and Nicola Prezza. 2023. Verifiable learning for robust tree ensembles. *arXiv*.

David Campos, Miao Zhang, Bin Yang, Tung Kieu, Chenjuan Guo, and Christian S. Jensen. 2023. Lightts: Lightweight time series classification with adaptive ensemble distillation – extended version. *arXiv*.

Sungmin Cha, Naeun Ko, Y. Yoo, and Taesup Moon. 2021. Self-supervised iterative contextual smoothing for efficient adversarial defense against gray- and black-box attack. *arXiv.org*.

Arunava Chakravarty, Tandra Sarkar, Nirmalya Ghosh, Ramanathan Sethuraman, and Debdoot Sheet. 2020. Learning decision ensemble using a graph neural network for comorbidity aware chest radiograph screening. *arXiv*.

Ankur Chanda, Kushan Choudhury, Shubhrodeep Roy, Shubhajit Biswas, and Somenath Kuiry. 2025. Evaluating temperature scaling calibration effectiveness for cnns under varying noise levels in brain tumour detection. *BIO Web of Conferences*.

Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila. 2021. An evaluation of self-supervised pre-training for skin-lesion analysis. *arXiv*.

Anthony Chen, Huanrui Yang, Yulu Gan, Denis A Gudovskiy, Zhen Dong, Haofan Wang, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2023. Split-ensemble: Efficient ood-aware ensemble via task and model splitting. *arXiv*.

Baixu Chen, Junguang Jiang, Ximei Wang, Jianmin Wang, and Mingsheng Long. 2022. Debiased pseudo labeling in self-training. *arXiv.org*.

Chaohui Chen, Yudou Wang, Gaoming Li, and A. Reynolds. 2010. Closed-loop reservoir management on the brugge test case.

Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. 2021. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. *arXiv*.

Qilei Chen, Ping Liu, Jing Ni, Yu Cao, Benyuan Liu, and Honggang Zhang. 2020a. Pseudo-labeling for small lesion detection on diabetic retinopathy images. *arXiv*.

Xiaohao Chen, Qianjun Shuai, F. Hu, and Yongqiang Cheng. 2024. Sdda: A progressive self-distillation with decoupled alignment for multimodal image-text classification. *Neurocomputing*.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. 2020b. Self-pu: Self boosted and calibrated positive-unlabeled training. *International Conference on Machine Learning*.

Yuzhao Chen, Yatao Bian, Xi Xiao, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020c. On self-distilling graph neural network. *arXiv*.

Zhewei Chen, Wai Keung Wong, Jinpiao Liao, and Ying Qu. 2025. Confidence calibration and uncertainty estimation in convolutional neural network for fabric defect segmentation: a benchmark study. *Textile research journal*.

Zhiliang Chen. 2025. Self-supervised distillation method for lightweight convolutional networks. *2025 6th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*.

Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, and Payel Das. 2019. Sprout: Self-progressing robust training.

Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, and Payel Das. 2020. Self-progressing robust training. *AAAI Conference on Artificial Intelligence*.

Prakash Chandra Chhipa, Gautam Vashishtha, Settur Jithamanyu, Rajkumar Saini, Mubarak Shah, and Marcus Liwicki. 2025. Astra: Adversarial self-supervised training with adaptive-attacks. *International Conference on Learning Representations*.

Belkacem Chikhaoui. 2025. Explainable ai via large language models: Translating neural network behavior into interpretable decision trees. *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*.

Seungju Cho, Hongsin Lee, and Changick Kim. 2025. Long-tailed adversarial training with self-distillation. *International Conference on Learning Representations*.

Gabriele Ciravegna, Frédéric Precioso, Alessandro Betti, Kevin Mottin, and Marco Gori. 2021. Knowledge-driven active learning. *arXiv*.

Lahav Dabah and Tom Tirer. 2024. On temperature scaling and conformal prediction of deep classifiers. *International Conference on Machine Learning*.

Shaochang Deng, Mengxiao Yin, and Feng Yang. 2022. A self-improving skin lesions diagnosis framework via pseudo-labeling and self-distillation. *Asian Conference on Machine Learning*.

Stanislav Dereka, Ivan Karpukhin, Maksim Zhdanov, and Sergey Kolesnikov. 2023. Diversifying deep ensembles: A saliency map approach for enhanced ood detection, calibration, and accuracy. *arXiv*.

Rahul Devassy, Giuseppe Durisi, Benjamin Lindqvist, Wei Yang, and Marco Dalai. 2016. Nonasymptotic coding-rate bounds for binary erasure channels with feedback. *arXiv*.

Qijie Ding, Jie Yin, Daokun Zhang, and Junbin Gao. 2023. Combating confirmation bias: A unified pseudo-labeling framework for entity alignment. *arXiv*.

Yongqi Ding, Lin Zuo, Mengmeng Jing, Kunshan Yang, Pei He, and Tonglan Xie. 2025. Synergy between the strong and the weak: Spiking neural networks are inherently self-distillers. *arXiv*.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2024. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. *arXiv*.

Zhe Dong, Zhi Lu, Yuanqing Feng, Shuai Guo, Zilong Wang, Yutong Wu, and Songfeng Lu. 2025. Bio-ofl: Biomedical privacy and auditable one-shot federated learning. *IEEE International Conference on Bioinformatics and Biomedicine*.

Emilio Dorigatti, Jann Goschenhofer, Benjamin Schubert, Mina Rezaei, and Bernd Bischl. 2022. Uncertainty-aware pseudo-label selection for positive-unlabeled learning. *arXiv*.

Girum Fitihamlak Ejigu, S. Hong, and C. Hong. 2023. Robust federated learning with local mixed co-teaching. *International Conference on Information Networking*.

Esmail Eltahan, F. Alpak, and K. Sepehrnoori. 2023. A quasi-newton trust-region method for optimization under uncertainty using stochastic simplex approximate gradients. *Computational Geosciences*.

Kobi Feldman, Martin Kellogg, and Oscar Chaparro. 2023. On the relationship between code verifiability and understandability. *arXiv*.

Tobias Finn, Gernot Geppert, and Felix Ament. 2022. Towards hourly three-dimensional ensemble data assimilation of screen-level observations into coupled atmosphere-land models. *arXiv*.

Pedro G. Fonseca and Hugo D. Lopes. 2017. Calibration of machine learning classifiers for probability of default modelling. *arXiv*.

L. Frenkel and J. Goldberger. 2021. Network calibration by class-based temperature scaling. *European Signal Processing Conference*.

L. Frenkel and Jacob Goldberger. 2022. Network calibration by temperature scaling based on the predicted confidence. *European Signal Processing Conference*.

Huazhu Fu, Jun Cheng, Yanwu Xu, Changqing Zhang, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. 2018. Disc-aware ensemble network for glaucoma screening from fundus image. *arXiv*.

R. Gaire, Sepehr Tabrizchi, and A. Roohi. 2024. Resource-efficient adaptive-network inference framework with knowledge distillation-based unified learning. *IEEE Computer Society Annual Symposium on VLSI*.

G. Gao, Yu Wang, J. Vink, T. Wells, and F. Saaf. 2021. Distributed quasi-newton derivative-free optimization method for optimization problems with multiple local optima. *Computational Geosciences*.

Jiaqi Gao, Jingqi Li, Hongming Shan, Yanyun Qu, James Z. Wang, Fei-Yue Wang, and Junping Zhang. 2022. Forget less, count better: A domain-incremental self-distillation learning benchmark for lifelong crowd counting. *arXiv*.

Yunxiang Gao and Wang Zhao. 2024. Research on supply chain optimization and management based on deep reinforcement learning. *Scalable Computing : Practice and Experience*.

Sam Gijsen, Marc-Andre Schulz, and Kerstin Ritter. 2025. Brain-semantoks: Learning semantic tokens of brain dynamics with a self-distilled foundation model. *arXiv*.

Eleonora Grassucci, Edoardo Cicero, and Danilo Comminiello. 2021. Quaternion generative adversarial networks. *arXiv*.

Dennis Gross, Nils Jansen, Guillermo A. Pérez, and Stephan Raaijmakers. 2020. Robustness verification for classifier ensembles. *arXiv*.

Yuliang Gu, Hongpeng Cao, Marco Caccamo, and N. Hovakimyan. 2025. Bregman centroid guided cross-entropy method. *arXiv.org*.

Denis A. Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. 2020. Deep active learning for biased datasets via fisher kernel self-supervision. *Computer Vision and Pattern Recognition*.

Jialin Guo, Zhenyu Wu, Zhiqiang Zhan, and Yang Ji. 2023. Dual-branch temperature scaling calibration for long-tailed recognition. *arXiv*.

Yuxin Guo, Shijie Ma, Yuhao Zhao, Hu Su, and Wei Zou. 2024. Cross pseudo-labeling for semi-supervised audio-visual source localization. *arXiv*.

Andrea Gurioli, Federico Pennino, Joao Monteiro, and Maurizio Gabbrielli. 2025. Mose: Hierarchical self-distillation enhances early layer embeddings.

Gustavo Coelho Haase and Paulo Henrique Dourado da Silva. 2025. Hpm-kd: Hierarchical progressive multi-teacher framework for knowledge distillation and efficient model compression. *arXiv.org*.

M. Hannanu, T. L. Silva, E. Camponogara, and M. Hovd. 2025. A trust region method for output-constrained reservoir optimization under geological uncertainty. *Computational Geosciences*.

Rich Harang and Hillary Sanders. 2023. Catastrophic forgetting in the context of model updates. *arXiv*.

A. Hareendranathan and Jacob L. Jaremko. 2025. Impact of adversarial attack on pediatric hip ultrasound deep learning models. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

S. Hasan and C. Linte. 2022. Stamp: A self-training student-teacher augmentation-driven meta pseudo-labeling framework for 3d cardiac mri image segmentation. *Annual Conference on Medical Image Understanding and Analysis*.

Alexander L. Hayes, Mayukh Das, Phillip Odom, and Sriraam Natarajan. 2019. User friendly automatic construction of background knowledge: Mode construction from er diagrams. *arXiv*.

Tyler L. Hayes and Christopher Kanan. 2021. Selective replay enhances learning in online continual analogical reasoning. *arXiv*.

Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. 2023a. Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. *arXiv.org*.

Qisheng He, Nicholas Summerfield, Ming Dong, and Carri Glide-Hurst. 2023b. Modality-agnostic learning for medical image segmentation using multi-modality self-distillation. *arXiv*.

Zhengbao He, Tao Li, Sizhe Chen, and Xiaolin Huang. 2023c. Investigating catastrophic overfitting in fast adversarial training: A self-fitting perspective. *arXiv*.

Christian Hinge, A. B. Rodell, Sven Zuehlsdorff, Bruce Spottiswoode, Kirsten Korsholm, B. M. Fischer, C. Ladefoged, and F. Andersen. 2025. Normal twin pet: personalized generative modeling for confounder correction and anomaly detection in whole-body pet/ct. *Scientific Reports*.

Dorjan Hitaj, Giulio Pagnotta, Iacopo Masi, and Luigi V. Mancini. 2021. Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *arXiv*.

Fujun Hou. 2018. Measuring knowledge for recognition and knowledge entropy. *arXiv*.

Elwin Huaman and Dieter Fensel. 2022. Knowledge graph curation: A practical framework. *arXiv*.

Kai Huang, Le Liang, Xinping Yi, Hao Ye, Shi Jin, and Geoffrey Ye Li. 2024. Meta-learning empowered graph neural networks for radio resource management. *arXiv*.

Seonghyeon Hwang, Minsu Kim, and Steven Euijong Whang. 2025. T-cil: Temperature scaling using adversarial perturbation for calibration in class-incremental learning. *Computer Vision and Pattern Recognition*.

Raza Imam, Ibrahim Almakky, Salma Alrashdi, Baketah Alrashdi, and Mohammad Yaqub. 2023. Seda: Self-ensembling vit with defensive distillation and adversarial training for robust chest x-rays classification. *DART@MICCAI*.

Max Andreas Ingrisch, Rani Marcel Schilling, Ingo Chmielewski, and Stefan Twieg. 2025. Determining the optimal t-value for the temperature scaling calibration method using the open-vocabulary detection model yolo-world. *Applied Sciences*.

Computational Intelligence and Neuroscience. 2023. Retracted: Deep unsupervised hashing for large-scale cross-modal retrieval using knowledge distillation model. *Computational Intelligence and Neuroscience*.

Sibgat Ul Islam, Jawad Ibn Ahad, Fuad Rahman, Mohammad Ruhul Amin, Nabeel Mohammed, and Shafin Rahman. 2025. Dynamic temperature scheduler for knowledge distillation. *arXiv.org*.

Guillaume Jeanneret, Juan C Perez, and Pablo Arbelaez. 2021. A hierarchical assessment of adversarial severity. *arXiv*.

Stephen Jewson, Anders Brix, and Christine Ziehmann. 2003. A new framework for the assessment and calibration of medium range ensemble temperature forecasts. *arXiv*.

Byeongmoon Ji, Hyemin Jung, Jihyeun Yoon, Kyungyul Kim, and Younghak Shin. 2019. Bin-wise temperature scaling (bts): Improvement in confidence calibration performance through simple scaling techniques. *arXiv*.

T. Jian, Zifeng Wang, Yanzhi Wang, Jennifer G. Dy, and Stratis Ioannidis. 2022. Pruning adversarially robust neural networks without adversarial examples. *Industrial Conference on Data Mining*.

Chengze Jiang, Junkai Wang, Minjing Dong, Jie Gui, Xinli Shi, Yuan Cao, Yuan Yan Tang, and James Tin-Yau Kwok. 2024. Improving fast adversarial training via self-knowledge guidance. *IEEE Transactions on Information Forensics and Security*.

Ruochen Jiao, Xiangguo Liu, Takami Sato, Qi Alfred Chen, and Qi Zhu. 2022. Semi-supervised semantics-guided adversarial training for trajectory prediction. *arXiv*.

Yuqi Jin, Zhenhao Shuai, Zihan Hu, Weiteng Zhang, Weihao Xie, Jianwei Shuai, Xian Shen, and Zhen Feng. 2025. Candle: A cross-modal agentic knowledge distillation framework for interpretable sarcopenia diagnosis.

Linglin Jing, Yiming Ding, Yunpeng Gao, Zhigang Wang, Xu Yan, Dong Wang, Gerald Schaefer, Hui Fang, Bin Zhao, and Xuelong Li. 2024. Hpl-ess: Hybrid pseudo-labeling for unsupervised event-based semantic segmentation. *arXiv*.

Sanghyun Jo, Soohyun Ryu, Sungyub Kim, Eunho Yang, and Kyungsu Kim. 2024. Ttd: Text-tag self-distillation enhancing image-text alignment in clip to alleviate single tag bias. *European Conference on Computer Vision*.

R. R. Joshua and S. Mohana. 2025. Enhancing gan resilience through adversarial-aware architecture and latent-space defense. *International Conference Emerging Trends Engineering, Science and Technology*.

Thomas Joy, Francesco Pinto, S. Lim, Philip H. S. Torr, and P. Dokania. 2022. Sample-dependent adaptive temperature scaling for improved calibration. *AAAI Conference on Artificial Intelligence*.

Reda Abdellah Kamraoui, Vinh-Thong Ta, Nicolas Papadakis, Fanny Compaire, José V Manjon, and Pierrick Coupé. 2021. Popcorn: Progressive pseudo-labeling with consistency regularization and neighboring. *arXiv*.

Ju Yeon Kang, Chang Ho Ryu, and T. Han. 2023. Binarized neural network with parameterized weight clipping and quantization gap minimization for online knowledge distillation. *IEEE Access*.

Sandra Kara, Hejer Ammar, Julien Denize, Florian Chabot, and Quoc-Cuong Pham. 2024. Diod: Self-distillation meets object discovery. *Computer Vision and Pattern Recognition*.

Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. 2020. Test-time adaptable neural networks for robust medical image segmentation. *arXiv*.

Isaac Kauvar, Chris Doyle, Linqi Zhou, and Nick Haber. 2023. Curious replay for model-based adaptation. *arXiv*.

Carolina R. Kelsch, Leonardo S. B. Pereira, Natnael Mola, Luis H. Arribas, and Juan C. S. M. Avedillo. 2026. Fade: Selective forgetting via sparse lora and self-distillation. *arXiv*.

Hyungmin Kim, Sungho Suh, Sunghyun Baek, Daehwan Kim, Daun Jeong, Hansang Cho, and Junmo Kim. 2022. Ai-kd: Adversarial learning and implicit regularization for self-knowledge distillation. *arXiv*.

Yoonhyung Kim and Changick Kim. 2021. Semi-supervised domain adaptation via selective pseudo labeling and progressive self-training. *International Conference on Pattern Recognition*.

Kubra Kose and Bing Zhou. 2025. Adversarial training for aerial disaster recognition: A curriculum-based defense against pgd attacks. *Electronics*.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, T. S. Filho, Hao Song, and Peter A. Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *Neural Information Processing Systems*.

Vinod K Kurmi, Badri N. Patro, Venkatesh K. Subramanian, and Vinay P. Namboodiri. 2021. Do not forget to attend to uncertainty while mitigating catastrophic forgetting. *arXiv*.

Fabian Küppers. 2023. Uncertainty calibration and its application to object detection. *arXiv*.

M. Laves, Sontje Ihler, Karl-Philipp Kortmann, and T. Ortmaier. 2019. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv.org*.

Minh H. N. Le, Tran Quoc Khanh Le, Thanh-Huy Nguyen, Dang Nguyen, Hien Quang Nguyen, N. Le, Kien Dang Nguyen, H. Kha, P. Nguyen, H. Le, H. Huynh, Dang-Manh Ho, Thanh-Minh Nguyen, Quan Nguyen, Min Xu, Phat K. Huynh, and Nguyen Quoc Khanh Le. 2025. Oasis-net: An obstetric adversarial semi-supervised image segmentation network for cervical and fetal head ultrasound imaging. *IEEE journal of biomedical and health informatics*.

Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu. 2016. Tiered-latency dram (tl-dram). *arXiv*.

Jinseok Lee and Dit-Yan Yeung. 2019. Knowledge query network: How knowledge interacts with skills. *arXiv*.

Jinsol Lee, Mohit Prabhushankar, and Ghassan AlRegib. 2022. Gradient-based adversarial and out-of-distribution detection. *arXiv*.

Deqiang Li and Qianmu Li. 2020. Adversarial deep ensemble: Evasion attacks and defenses for malware detection. *arXiv*.

Hui Li, Guimin Huang, Yiqun Li, Xiaowei Zhang, Yabing Wang, and Jun Li. 2023a. Semi: Self-supervised information-enhanced meta-learning for few-shot text classification. *International Journal of Computational Intelligence Systems*.

Jr-Shin Li and Wei Zhang. 2020. Ensemble control on lie groups. *arXiv*.

Minghan Li and Éric Gaussier. 2022. Domain adaptation for dense retrieval through self-supervision by pseudo-relevance labeling. *arXiv.org*.

Qiaoxin Li, Yao Zhu, Jing He, Mengyun Wang, Meiling Zhu, Tingyan Shi, L. Qiu, D. Ye, and Q. Wei. 2013. Steroid 5-alpha-reductase type 2 (srd5a2) v89l and a49t polymorphisms and sporadic prostate cancer risk: a meta-analysis. *Molecular Biology Reports*.

Qiyuan Li, Haijiang Liu, Caicai Guo, Chao Gao, Deyu Chen, Meng Wang, Feng Gao, Frank van Harmelen, and Jinguang Gu. 2025. Reviewing clinical knowledge in medical large language models: Training and beyond. *arXiv*.

Ruipeng Li, Jianming Ye, Yueqi Huang, Wei Jin, Peng Xu, and Lilin Guo. 2024. A continuous learning approach to brain tumor segmentation: integrating multi-scale spatial distillation and pseudo-labeling strategies. *Frontiers in Oncology*.

Xiaofan Li, Bo Peng, Jie Hu, Changyou Ma, Daipeng Yang, and Zhuyang Xie. 2023b. Usl-net: Uncertainty self-learning network for unsupervised skin lesion segmentation. *arXiv*.

Zhanbo Liang, Yuping Sun, and Si Li. 2025. Generalized nesterov-boosted adversarial data augmentation framework for multi-label chest x-ray image classification. *IEEE International Conference on Bioinformatics and Biomedicine*.

Zijin Lin, Jinwen He, Yue Zhao, Ruigang Liang, Hu Li, and ZhenDong Wu. 2025. Egrte: adversarially training a self-explaining smoothed classifier for certified robustness. *Cybersecurity*.

Yunzhi Ling, Feiping Nie, Weizhong Yu, and Xuelong Li. 2025. Self-labeling and self-knowledge distillation unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2021. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. *arXiv*.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. 2023. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. *arXiv*.

Afonso Lourenço, João Gama, Eric P. Xing, and Goreti Marreiros. 2025. Bridging streaming continual learning via in-context large tabular models. *arXiv*.

Vítor Lourenço and Aline Paes. 2022. Learning attention-based representations from multiple patterns for relation prediction in knowledge graphs. *arXiv*.

Menglong Lu, Zhen Huang, Yunxiang Zhao, Zhiliang Tian, Yang Liu, and Dongsheng Li. 2023. Damstf: Domain adversarial learning enhanced meta self-training for domain adaptation. *Annual Meeting of the Association for Computational Linguistics*.

Yang Luo, Zhineng Chen, Shengtian Zhou, and Xieping Gao. 2022. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv*.

Florian Lux and Ngoc Thang Vu. 2021. Meta-learning for improving rare word recognition in end-to-end asr. *arXiv*.

Xiang Ma and Linfeng Bi. 2019. A robust adaptive iterative ensemble smoother scheme for practical history matching applications. *Computational Geosciences*.

Md Junaid Mahmood, Pranaw Raj, Divyansh Agarwal, Suruchi Kumari, and Pravendra Singh. 2023. Splal: Similarity-based pseudo-labeling with alignment loss for semi-supervised medical image classification. *arXiv*.

Sylwia Majchrowska, Anders G.F. Hildeman, Ricardo Mokhtari, and Philip A. Teare. 2024. Interpretable echo analysis using self-supervised parcels. *International Conference on Computing in Cardiology*.

H. Malik, Shahina K. Kunhimon, Muzammal Naseer, F. Khan, and Salman H. Khan. 2025. Hierarchical self-supervised adversarial training for robust vision models in histopathology. *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

G.J. Maniram, S. Nithishkumar, and A. H. Ajibah. 2025. Multi- task lesion segmentation and classification with explainable ai and uncertainty estimation for trustworthy diabetic retinopathy screening. *2025 2nd International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*.

Charles Martin, Henry Gardner, and Ben Swift. 2020. Tracking ensemble performance on touch-screens with gesture classification and transition matrices. *arXiv*.

Emmanuel Martínez, Roman Jacome, Romario Gualdrón-Hurtado, Iñaki Esnaola, and Henry Arguello. 2025. Compressive sensing with augmented measurements via generative self-distillation. *Symposium on Software Performance*.

Kouassi Joshua Caleb Micah and Lou Qiong. 2023. Face mask image classification using fine-tuning and the effect of fgsm and pgd attacks. *International Journal For Multidisciplinary Research*.

Kristian Miok, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Sikonja. 2025. Tt-xai: Trustworthy clinical text explanations via keyword distillation and llm reasoning. *arXiv.org*.

Behzad Moayedi, Abdalsamad Keramatfar, Mohammad Hadi Goldani, Mohammad Javad Fallahi, Alborz Jahangirisisakht, Mohammad Saboori, and Leyla badiei. 2023. An ensemble machine learning approach for screening covid-19 based on urine parameters. *arXiv*.

Riccardo De Monte, Davide Dalle Pezze, and Gian Antonio Susto. 2025. Teach yolo to remember: A self-distillation approach for continual object detection. *arXiv*.

Monireh Moshavash, M. Eftekhari, and Kaveh Bahraman. 2021. Momentum contrast self-supervised based training for adversarial robustness.

Azamat Mukhamediya and A. Zollanvari. 2024. Srpmst: Sequential retraining and pseudo-labeling in mini-batches for self-training. *Neurocomputing*.

Martin Mundt, Iuliia Pliushch, Sagnik Majumder, Yongwon Hong, and Visvanathan Ramesh. 2019. Unified probabilistic deep continual learning through generative replay and open set recognition. *arXiv*.

Shahabedin Nabavi, Kian Anvari, Mohsen Ebrahimi Moghaddam, A. A. Abin, and A. Frangi. 2024. Multiple teachers-meticulous student: A domain adaptive meta-knowledge distillation model for medical image classification. *Medical Physics (Lancaster)*.

Kotaro Nagata, Hiromu Ono, and Kazuhiro Hotta. 2024. Reducing catastrophic forgetting in online class incremental learning using self-distillation. *arXiv*.

Suryarprakash Nalluri, Aiman Shariff, Chethan T.P, Aisirii V Hegde, and Indu Mahesh. 2025. Asl-mdfd: Adversarial self-supervised learning for generalizable gan-resilient multimodal deepfake detection. *International Journal of Global Innovations and Solutions*.

Nassr Nama, Stephanie Liebert, Mario Abaji, Amy M DeLaroche, Kristy Carlin, Teresa Jewell, D. D'Arienzo, Alastair Fung, David Gremse, J. Bonkowsky, Maida Chen, E. Sagiv, Bruce Herman, E. Lu, Peter J Gill, Joel S. Tieder, and Eric Coon. 2026. Infant outcomes, risk factors, and diagnostic yield after a brief resolved unexplained event: A systematic review and meta-analysis. *JAMA pediatrics*.

Vignesh Narayanan, Wei Zhang, and Jr-Shin Li. 2020. Moment-based ensemble control. *arXiv*.

James O' Neill, Sourav Dutta, and Haytham Assem. 2021. Deep neural compression via concurrent pruning and self-distillation. *arXiv*.

E. Niño and Adrian Sandu. 2014. Variational data assimilation based on derivative-free optimization. *International Conference on Dynamic Data-Driven Environmental Systems Science*.

Mohamed Ohamouddou, Said Ohamouddou, A. E. Afia, and R. Lasri. 2025. Atms-kd: Adaptive temperature and mixed sample knowledge distillation for a lightweight residual cnn in agricultural embedded systems. *Smart Agricultural Technology*.

Richard Osuala, Kaisar Kushibar, Lidia Garrucho, Akis Linardos, Zuzanna Szafranowska, Stefan Klein, Ben Glocker, Oliver Diaz, and Karim Lekadir. 2021. Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *arXiv*.

H. P and Padmavathi G. 2024. Resilience in remote sensing image classification: Evaluating deep learning models against adversarial attacks. *International Conference on Computing Communication and Networking Technologies*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv*.

Sejik Park. 2024. Diverse feature learning by self-distillation and reset. *arXiv*.

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv*.

Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. 2023. Hierarchical dense correlation distillation for few-shot segmentation-extended abstract. *arXiv*.

Bowen Peng, Li Liu, Tianpeng Liu, Zhen Liu, and Yongxiang Liu. 2024. Enhancing transferability of targeted adversarial examples: A self-universal perspective. *arXiv.org*.

Dimitar Peshevski, Riste Stojanov, and Dimitar Trajanov. 2025. Ai agent-driven framework for automated product knowledge graph construction in e-commerce. *arXiv*.

Saqib Qamar. 2025. Confidence-weighted semi-supervised learning for skin lesion segmentation using hybrid cnn-transformer networks. *arXiv*.

Chao Qi, Jianqin Yin, Yingchun Niu, and Jinghang Xu. 2021. Neighborhood spatial aggregation mc dropout for efficient uncertainty-aware semantic segmentation in point clouds. *IEEE Transactions on Geoscience and Remote Sensing*.

Xiaoran Qi, Guoning Zhang, Jianghao Wu, Shaoting Zhang, Xiaorong Hou, and Guotai Wang. 2025. Fdas: Foundation model distillation and anatomic structure-aware multi-task learning for self-supervised medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Chen Qing, Xinwei Li, Jinhong Xia, Yifeng Lin, and Shenhai Zheng. 2024. Communication-efficient federated self-distillation method for medical image segmentation. *Computer Science and Technology*.

Yunfei Qiu, Qiqiong Ma, Tianhua Lv, Li Fang, Shudong Zhou, and Wei Yao. 2026. Semi-supervised hyperspectral image classification with edge-aware superpixel label propagation and adaptive pseudo-labeling. *arXiv*.

Nsw Roads and Maritime Services. 2014. Disclosure of government contracts with the private sector class 2 contract (including wads) (as per sec (14) foi amendment act 2006 no 115).

Julian Rodemann. 2023. Pseudo label selection is a decision problem. *arXiv*.

L. Rutkowski. 2004. Artificial intelligence and soft computing - icaisc 2004 : 7th international conference, zakopane, poland, june 7-11, 2004 : proceedings.

Vignesh Sampath, I. Maurtua, Juan José Aguilar Martín, A. Iriondo, Iker Lluvia, and Andoni Rivera. 2022. Vision transformer based knowledge distillation for fasteners defect detection. *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*.

Pascal Schöttle, Alexander Schlögl, Cecilia Pasquini, and Rainer Böhme. 2018. Detecting adversarial examples - a lesson from multimedia forensics. *arXiv*.

Renrong Shao, Dongyan Li, Dong Xia, Ling Shao, Jiangdong Lu, Fen Zheng, and Lulu Zhang. 2026. Dsvm-unet : Enhancing vm-unet with dual self-distillation for medical image segmentation.

Rulin Shao, Jinfeng Yi, Cho-Jui Hsieh, and Pin-Yu Chen. 2022. How and when adversarial robustness improves in knowledge distillation?

Saurabh Sharma, Shikhar Singh Lodhi, Vanshika Srivastava, and Joydeep Chandra. 2025. Nord: A framework for noise-resilient self-distillation through relative supervision. *Applied intelligence (Boston)*.

Zhiqiang Shen, Zechun Liu, Jie Qin, Lei Huang, Kwang-Ting Cheng, and Marios Savvides. 2021. S2-bnn: Bridging the gap between self-supervised real and 1-bit neural networks via guided distribution calibration. *arXiv*.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *arXiv*.

M. Shevlin, P. Hyland, M. Cloitre, Chris R. Brewin, Dmytro Martsenkovskyi, M. Ben-Ezra, Kristina Bondjers, T. Karatzias, Michael Duffy, and E. Redican. 2024. Assessing self-reported prolonged grief disorder with "clinical checks": A proof of principle study. *Journal of Traumatic Stress*.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. *arXiv*.

Yue Su, Youqian Zhang, and Jinfu Xu. 2024. Genetic variations in anti-diabetic drug targets and copd risk: evidence from mendelian randomization. *BMC Pulmonary Medicine*.

Jinghan Sun, Dong Wei, Kai Ma, Liansheng Wang, and Yefeng Zheng. 2021a. Unsupervised representation learning meets pseudo-label supervised self-distillation: A new approach to rare disease classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Tao Sun, Bojian Yin, and S. Bohté. 2023. Efficient uncertainty estimation in spiking neural networks via mc-dropout. *International Conference on Artificial Neural Networks*.

Yibao Sun, Xingru Huang, Huiyu Zhou, and Qianni Zhang. 2021b. Srpn: similarity-based region proposal networks for nuclei and cells detection in histology images. *arXiv*.

R. Takashima, Yuya Sawa, Ryo Aihara, Tetsuya Takiguchi, and Yoshie Imai. 2024. Dysarthric speech recognition using pseudo-labeling, self-supervised feature learning, and a joint multi-task learning approach. *IEEE Access*.

Dayu Tan, Ziwei Zhang, Yansan Su, Xin Peng, Yike Dai, Chunhou Zheng, and Weimin Zhong. 2025. Msdkmamba: Bidirectional spatial-aware multi-modal 3d brain segmentation via multi-scale self-distilled fusion strategy. *arXiv*.

S. Tan, R. Caruana, G. Hooker, and Yin Lou. 2017. Auditing black-box models using transparent model distillation with side information.

Michail Tarasiou and Stefanos Zafeiriou. 2022. Embedding earth: Self-supervised contrastive pre-training for dense land cover classification. *arXiv*.

Mohammad T. Teimuri, Zahra Dehghanian, Gholamali Aminian, and Hamid R. Rabiee. 2025. Upl: Uncertainty-aware pseudo-labeling for imbalance transductive node classification. *arXiv*.

Andrea Terlizzi, Angelo Nazzaro, Lorenzo Bernardi, Francesco Bardozzo, and R. Tagliaferri. 2025. Rsdix: Lightweight and data-efficient vlms for remote sensing through self-distillation. *IEEE International Joint Conference on Neural Network*.

James Thorne and Andreas Vlachos. 2020. Elastic weight consolidation for better bias inoculation. *arXiv*.

Ye Tian and Yang Feng. 2021. Rase: A variable screening framework via random subspace ensembles. *arXiv*.

Christian Tomani, Daniel Cremers, and Florian Buettner. 2021. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. *arXiv*.

Jan Tužil, J. Matějka, M. Mamas, and T. Doležal. 2023. Short-term risk of periprocedural stroke relative to radial vs. femoral access: systematic review, meta-analysis, study sequential analysis and meta-regression of 2,188,047 real-world cardiac catheterizations. *Expert Review of Cardiovascular Therapy*.

Hayat Ullah, Syed Muhammad Talha Zaidi, and Arslan Munir. 2025. Improving adversarial robustness through adaptive learning-driven multi-teacher knowledge distillation. *arXiv*.

Juan Miguel Valverde, Motoya Koga, Nijihiko Otsuka, and Anders Bjorholm Dahl. 2025. Topomortar: A dataset to evaluate image segmentation methods focused on topology accuracy. *arXiv.org*.

Praneeth Vepakomma, Subha Nawer Pushpita, and R. Raskar. 2020. Dams: Meta-estimation of private sketch data structures for differentially private covid-19 contact tracing.

Thibaut Vidal, Toni Pacheco, and Maximilian Schiffer. 2020. Born-again tree ensembles. *arXiv*.

Soujenya Voggu, Shadab Siddiqui, and Shahin Fatima. 2025. Efficientnet-based melanoma classification with cbam attention and monte carlo dropout for robust uncertainty estimation. *International Journal of Advanced Computer Science and Applications*.

Dhairya Vyas, Viral V. Kapadia, Viranchkumar Mayurbhai Kadia, and Vipinchandra Patel. 2025. Augmented smoothing for robust cnn image classification against adversarial attacks. *Journal of Computing amp; Biomedical Informatics*.

Hetvi Waghela, Jaydip Sen, and Sneha Rakshit. 2024a. Adversarial resilience in image classification: A hybrid approach to defense. *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*.

Hetvi Waghela, Jaydip Sen, and Sneha Rakshit. 2024b. Robust image classification: Defensive strategies against fgsm and pgd adversarial attacks. *2024 Asian Conference on Intelligent Technologies (ACOIT)*.

Chunshi Wang, Shougan Teng, Shaohua Sun, and Bin Zhao. 2024a. Symmatch: Symmetric bi-scale matching with self-knowledge distillation in semi-supervised medical image segmentation. *IEEE International Conference on Bioinformatics and Biomedicine*.

Chunshi Wang, Bin Zhao, and Zhiyang Liu. 2024b. Distillmatch: Revisiting self-knowledge distillation in semi-supervised medical image segmentation. *IEEE International Conference on Bioinformatics and Biomedicine*.

Han Wang, Ruiliu Fu, Chengzhang Li, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Reminding the incremental language model via data-free self-distillation. *arXiv*.

Hao Wang, Euijoon Ahn, Lei Bi, and Jinman Kim. 2023a. Self-supervised multi-modality learning for multi-label skin lesion classification. *arXiv*.

Nan Wang, Shaohui Lin, Xiaoxiao Li, Ke Li, Yunhang Shen, Yue Gao, and Lizhuang Ma. 2022. Missu: 3d medical image segmentation via self-distilling transunet. *arXiv*.

Wenxuan Wang, Jing Liu, Xingjian He, Yisi Zhang, Chen Chen, Jiachen Shen, Yan Zhang, and Jiangyun Li. 2023b. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *arXiv*.

Wenxuan Wang, Chenglei Wang, Huihui Qi, Meng Ye, Xuelin Qian, Peng Wang, and Yanning Zhang. 2024c. Sustainable self-evolution adversarial training. *ACM Multimedia*.

Xubin Wang, Yunhe Wang, Zhiqing Ma, Ka-Chun Wong, and Xiangtao Li. 2024d. Exhaustive exploitation of nature-inspired computation for cancer screening in an ensemble manner. *arXiv*.

Yingchao Wang and Wenqi Niu. 2024. Federated progressive self-distillation with logits calibration for personalized iiot edge intelligence. *arXiv*.

Yu Wang, Yuxuan Yin, and Peng Li. 2024e. Towards the mitigation of confirmation bias in semi-supervised learning: a debiased training perspective. *arXiv*.

Haomin Wei, YuJiang Luo, Tao Xie, and Yunong Yang. 2025. Federated learning with dual-end gradient correction and proxy-free self-distillation. *IEEE Access*.

Zhiyuan Wei, Jing Sun, Zijian Zhang, Xianhao Zhang, and Zhe Hou. 2024. Ftsmartaudit: A knowledge distillation-enhanced framework for automated smart contract auditing using fine-tuned llms.

Shixian Wen and Laurent Itti. 2018. Overcoming catastrophic forgetting problem by weight consolidation and long-term memory. *arXiv*.

Futian Weng, Yuanting Ma, Jinghan Sun, Shijun Shan, Qiyuan Li, Jianping Zhu, Yang Wang, and Yan Xu. 2022. An interpretable imbalanced semi-supervised deep learning framework for improving differential diagnosis of skin diseases. *arXiv*.

Andreas Winata, Nur Afny Catur Andryani, Alexander Agung Santoso Gunawan, , and Ford Lumban Gaol. 2025. Diverse representation knowledge distillation for efficient edge ai teledermatology in skin disease diagnosis. *IEEE Access*.

Shunyu Wu, Dan Li, Haozheng Ye, Zhuomin Chen, Jiahui Zhou, Jian Lou, Zibin Zheng, and See-Kiong Ng. 2025. Tsrating: Rating quality of diverse time series data by meta-learning from llm judgment. *arXiv*.

Yuehua Wu, Hung-Jui Wang, and Shang-Tse Chen. 2023. Annealing self-distillation rectification improves adversarial training. *International Conference on Learning Representations*.

Q. Xiao, Jian Chen, Jia Zhu, Shu-Xin Zeng, H. Cai, and Guomin Zhu. 2023. Association of several loci of smad7 with colorectal cancer: A meta-analysis based on case–control studies. *Medicine*.

Ruiqiang Xiao, Songning Lai, Yijun Yang, Jiemin Wu, Yutao Yue, and Lei Zhu. 2024. Drive: Dual-robustness via information variability and entropic consistency in source-free unsupervised domain adaptation. *arXiv*.

Dongqing Xie, Yonghuang Wu, Zisheng Ai, Jun Min, Zhencun Jiang, Shaojin Geng, and Lei Wang. 2025. Ccsd: Cross-modal compositional self-distillation for robust brain tumor segmentation with missing modalities. *arXiv*.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. *arXiv*.

Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. 2023. Class-distribution-aware pseudo labeling for semi-supervised multi-label learning. *arXiv*.

Fan Xing, Xiaoyi Zhou, Xuefeng Fan, Zhuo Tian, and Yan Zhao. 2023. Raediff: Denoising diffusion probabilistic models based reversible adversarial examples self-generation and self-recovery. *arXiv.org*.

Miao Xiong, Ailin Deng, Pang Wei Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. 2023. Proximity-informed calibration for deep neural networks. *arXiv*.

Anding Xu, Qingwen Zhu, Guilan Li, Caihong Gong, Xia Li, H. Chen, J. Cui, Songping Wu, Zhiguang Xu, and Yurong Yan. 2022. 2d bismuth@n-doped carbon sheets for ultrahigh rate and stable potassium storage. *Small*.

Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2019. Adversarial attacks and defenses in images, graphs and text: A review. *arXiv*.

Shilin Xu, Xiangtai Li, Size Wu, Wenwei Zhang, Yining Li, Guangliang Cheng, Yunhai Tong, Kai Chen, and Chen Change Loy. 2025. Dst-det: Open-vocabulary object detection via dynamic self-training. *IEEE transactions on circuits and systems for video technology (Print)*.

Hongwei Yan, Liyuan Wang, Kaisheng Ma, and Yi Zhong. 2024. Orchestrate latent expertise: Advancing online continual learning with multi-level supervision and reverse self-distillation. *Computer Vision and Pattern Recognition*.

Fei Yang, Kai Wang, and Joost van de Weijer. 2023. Scrollnet: Dynamic weight importance for continual learning. *arXiv*.

Jianing Yang, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2026. Distilmos: Layer-wise self-distillation for self-supervised learning model-based mos prediction. *arXiv.org*.

Yujie Yang, Xuwei Zheng, Haiying Lv, Bin Tang, Y. Zhong, Qianqian Luo, Yang Bi, Kexin Yang, Haixin Zhong, Haiming Chen, and Chuanjian Lu. 2024a. The causal relationship between serum metabolites and the risk of psoriasis: a mendelian randomization and meta-analysis study. *Frontiers in Immunology*.

Zhaorui Yang, Qian Liu, Tianyu Pang, Han Wang, H. Feng, Minfeng Zhu, and Wei Chen. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. *Annual Meeting of the Association for Computational Linguistics*.

Zonglin Yang. 2025. Efficient uncertainty estimation and calibration on edge cpus: A lightweight framework with temperature scaling and calibration-aware early stopping. *2025 10th International Conference on Computer and Information Processing Technology (ISCIPT)*.

Yiwen Ye, Jianpeng Zhang, Ziyang Chen, and Yong Xia. 2022. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Xuwang Yin, Soheil Kolouri, and Gustavo K. Rohde. 2019. Gat: Generative adversarial training for adversarial example detection and robust classification. *arXiv*.

Ji Won Yoon, Sunghwan Ahn, Hyeonseung Lee, Minchan Kim, Seok Min Kim, and Nam Soo Kim. 2023. Em-network: Oracle guided self-distillation for sequence learning. *arXiv*.

Yaodong Yu, Stephen Bates, Yi-An Ma, and Michael I. Jordan. 2022. Robust calibration with multi-domain temperature scaling. *Neural Information Processing Systems*.

Bo Yuan, Yulin Chen, and Yin Zhang. 2025. Label-guided self-knowledge distillation for multi-class text classification. *Natural Language Processing and Chinese Computing*.

Luca Zampierin, Ghouthi Boukli Hacene, Bac Nguyen, and Mirco Ravanelli. 2024. Skill: Similarity-aware knowledge distillation for speech self-supervised learning. *arXiv*.

Tal Zeevi, R. Venkataraman, Lawrence H. Staib, and John A. Onofrey. 2024. Monte-carlo frequency dropout for predictive uncertainty estimation in deep learning. *IEEE International Symposium on Biomedical Imaging*.

Ximin Zeng, Hongmei Wang, Long Zhao, Yue Cheng, Danping Zhou, and Shaoping Shi. 2025. Uncertainty quantification and temperature scaling calibration for protein-rna binding site prediction. *Journal of Chemical Information and Modeling*.

Hao Zhai, Tengfei Zhang, Yicong Wu, Dequan Zeng, Zhengfa Liu, and Helin Wang. 2025. Improved self-distillation and memorized spatial local clustering for test time adaptation. *2025 2nd International Symposium on AI and Cybersecurity (ISAICS)*.

Hanlin Zhang, Shuai Lin, Weiyang Liu, Pan Zhou, Jian Tang, Xiaodan Liang, and Eric P. Xing. 2020a. Iterative graph self-distillation. *arXiv*.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. 2024a. Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models. *arXiv.org*.

Juping Zhang, Yi Yuan, Gan Zheng, Ioannis Krikidis, and Kai-Kit Wong. 2021. Embedding model based fast meta learning for downlink beamforming adaptation. *arXiv*.

Tiantian Zhang, Kevin Zehua Shen, Zichuan Lin, Bo Yuan, Xueqian Wang, Xiu Li, and Deheng Ye. 2023. Replay-enhanced continual reinforcement learning. *arXiv*.

Wenping Zhang, Xiao feng He, and X. Ye. 2020b. Association between the combined effects of gstm1 present/null and cyp1a1 mspi polymorphisms with lung cancer risk: an updated meta-analysis. *Bioscience Reports*.

Yudian Zhang, Chenhao Xu, Kaiye Xu, and Haijiang Zhu. 2024b. Mask-ts net: Mask temperature scaling uncertainty calibration for polyp segmentation. *International Conference on Pattern Recognition*.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. 2025. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. *arXiv*.

Xiaochen Zhao, Chengting Yu, Kairong Yu, Lei Liu, and Aili Wang. 2025. Enhanced self-distillation framework for efficient spiking neural network training. *arXiv*.

Yan Zhao, Dingxian Wang, Cheng-Yu Zhang, Yan ju Liu, Xiaohong Wang, Mengxin Shi, Wei Wang, Xueqing Shen, and Xiao feng He. 2022. Individual and combined effects of the gstm1, gstt1, and gstp1 polymorphisms on leukemia risk: An updated meta-analysis. *Frontiers in Genetics*.

Yilin Zheng, Lingmin He, and Jianbao Li. 2023. Decoupled adversarial network and self-training with weighted pseudo-labels for domain adaptive semantic segmentation. *IEEE International Conference on Systems, Man and Cybernetics*.

Zhiyuan Zhou, Shreyas Sundara Raman, Henry Sowerby, and Michael L. Littman. 2022. Tiered reward: Designing rewards for specification and fast learning of desired behavior. *arXiv*.

Xunyu Zhu, Jian Li, Yong Liu, and Weiping Wang. 2023. Improving differentiable architecture search via self-distillation. *arXiv*.

Dingyi Zhuang, Chonghe Jiang, Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. 2024. Gets: Ensemble temperature scaling for calibration in graph neural networks. *arXiv*.